HAL
open science

The gradient discretisation method

Jérôme Droniou, Robert Eymard, Thierry Gallouët, Cindy Guichard, Raphaele Herbin

To cite this version:

Jérôme Droniou, Robert Eymard, Thierry Gallouët, Cindy Guichard, Raphaele Herbin. The gradient discretisation method . 2018. hal-01382358v7

HAL Id: hal-01382358

https://hal.science/hal-01382358v7

Preprint submitted on 15 Mar 2018 (v7), last revised 9 Jul 2018 (v8)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

J. Droniou, R. Eymard, T. Gallouët,
C. Guichard and R. Herbin

The gradient discretisation method

March 15, 2018

Preface

This monograph is dedicated to the presentation of the gradient discretisation method (GDM) and to some of its applications. It is intended for masters students, researchers and experts in the field of the numerical analysis of partial differential equations.

The GDM is a framework which contains classical and recent discretisation schemes for diffusion problems of different kinds: linear or non-linear, steady-state or time-dependent. The schemes may be conforming or non-conforming, low or high order, and may be built on very general meshes.

In this monograph, the core properties that are required to prove the convergence of a GDM are stressed, and the analysis of the method is performed on a series of elliptic and parabolic problems. As a result, for these models, any scheme entering the GDM framework is known to converge. A key feature of this monograph is the presentation of techniques and results which enable a complete convergence analysis of the GDM on fully non-linear, and sometimes degenerate, models. The scope of some of these techniques and results goes beyond the GDM, and makes them potentially applicable to numerical schemes not (yet) known to fit into this framework.

Appropriate tools are also provided to easily check whether a given scheme satisfies the core properties of a GDM. Using these tools, it is shown that a number of methods are GDMs; some of these methods are classical, such as the conforming finite elements, the non-conforming finite elements, and the mixed finite elements. Others are more recent, such as the discontinuous Galerkin methods, the hybrid mimetic mixed or nodal mimetic finite differences methods, some discrete duality finite volume schemes, and some multi-point flux approximation schemes.

Marseille, Melbourne, Paris
the authors, 2017

Contents

Part I Elliptic problems

1 Motivation and basic ideas . . . 3
- 1.1 Some well-known approximations of linear elliptic problems . . . 3
  - 1.1.1 Galerkin methods . . . 3
  - 1.1.2 Non-conforming $\mathbb{P}_1$ finite elements . . . 5
  - 1.1.3 Two-point flux approximation finite volumes on Cartesian meshes . . . 7
- 1.2 Towards the gradient discretisation method . . . 10
- 1.3 Generalisation to non-linear problems . . . 13

2 Dirichlet boundary conditions . . . 17
- 2.1 Homogeneous Dirichlet boundary conditions . . . 18
  - 2.1.1 Gradient discretisations . . . 18
  - 2.1.2 Gradient schemes for linear problems . . . 32
  - 2.1.3 On the notions of consistency and stability . . . 38
  - 2.1.4 Gradient schemes for quasi-linear problems . . . 41
  - 2.1.5 Gradient schemes for p-Laplace type problems: $p \in (1, +\infty)$ . . . 45
- 2.2 Non-homogeneous Dirichlet boundary conditions . . . 58
  - 2.2.1 Gradient discretisations . . . 58
  - 2.2.2 Gradient schemes for linear problems . . . 61
  - 2.2.3 Gradient schemes for quasi-linear problems . . . 63

3 Neumann, Fourier and mixed boundary conditions . . . 65
- 3.1 Neumann boundary conditions . . . 65
  - 3.1.1 Gradient discretisations . . . 65
  - 3.1.2 Complements on trace operators . . . 74
  - 3.1.3 Gradient schemes for linear problems . . . 77
  - 3.1.4 Gradient schemes for quasi-linear problems . . . 80
- 3.2 Fourier boundary conditions . . . 85

3.2.1 Gradient discretisations 85
3.2.2 Gradient schemes for quasi-linear problems 87
3.3 Mixed boundary conditions 89
3.3.1 Gradient discretisations 89
3.3.2 Gradient schemes for linear problems 93

Part II Parabolic problems

4 Time-dependent problems 97
4.1 Space–time gradient discretisation 97
4.2 Compactness results for space-time gradient discretisations . . . 106
4.2.1 Averaged-in-time compactness for space-time GDs 106
4.2.2 Uniform-in-time compactness for space-time GDs 112

5 Parabolic problems 117
5.1 The gradient discretisation method for a quasi-linear parabolic problem 117
5.1.1 The continuous problem 118
5.1.2 The gradient scheme 119
5.1.3 Error estimate in the linear case 120
5.1.4 Convergence analysis in the non-linear case 124
5.2 Non-conservative problems 133
5.2.1 The continuous problem 133
5.2.2 Fully implicit scheme 135
5.2.3 Semi-implicit scheme 142
5.3 Non-linear time-dependent Leray–Lions type problems 147
5.3.1 Model 147
5.3.2 Gradient scheme and main results 148
5.3.3 *A priori* estimates 150
5.3.4 Proof of the convergence results 152

6 Degenerate parabolic problems 161
6.1 The continuous problem 162
6.1.1 Hypotheses and notion of solution 162
6.1.2 A maximal monotone operator viewpoint 163
6.2 Gradient scheme 165
6.3 Estimates on the approximate solution 165
6.4 A first convergence theorem 170
6.5 Uniform-in-time, strong L^2 convergence results 174
6.6 Auxiliary results 180
6.7 Proof of the uniqueness of the solution to the model 187

Part III Examples of gradient discretisation methods

7 Analysis tools for gradient discretisations 201
7.1 Polytopal tools 202
7.1.1 Polytopal meshes 202
7.1.2 Operators, norm and regularity factors associated with a polytopal mesh 205
7.2 Control of a GD by a polytopal toolbox 207
7.2.1 Dirichlet boundary conditions 207
7.2.2 Neumann and Fourier boundary conditions 212
7.2.3 Mixed boundary conditions 215
7.3 Local linearly exact GDs 217
7.3.1 $\mathbb{P}_0$-exact and $\mathbb{P}_1$-exact reconstructions 217
7.3.2 Definition and consistency of local linearly exact GDs for Dirichlet boundary conditions 222
7.3.3 From local to global basis functions, and matrix assembly 227
7.3.4 Barycentric condensation 228
7.3.5 Mass lumping 234
7.3.6 Non-homogeneous Dirichlet, Neumann and Fourier boundary conditions 238
7.4 $W^{2,p}$ estimate of $S_{\mathcal{D}}$ for local linearly exact GDs 243
7.4.1 Functional estimates in $W^{2,p}$ 243
7.4.2 Application to local linearly exact GDs 249
7.5 Further topics on LLE GDs 255
7.5.1 LLE GDs with generalised discrete unknowns 255
7.5.2 Non-linearly exact barycentric condensation 256

8 Conforming approximations 259
8.1 Conforming Galerkin methods 259
8.1.1 Homogeneous Dirichlet boundary conditions 259
8.1.2 Non-homogeneous Neumann boundary conditions 260
8.2 $\mathbb{P}_k$ finite elements for homogeneous Dirichlet boundary conditions 261
8.2.1 Definition of $\mathbb{P}_k$ gradient discretisations 261
8.2.2 Properties of $\mathbb{P}_k$ gradient discretisations 265
8.3 $\mathbb{P}_k$ FE for non-homogeneous Dirichlet, Neumann and Fourier BCs 268
8.3.1 Non-homogeneous Dirichlet conditions 268
8.3.2 Neumann boundary conditions 269
8.3.3 Fourier conditions 270
8.4 Mass-lumped $\mathbb{P}_1$ finite elements 270
8.5 Vertex approximate gradient (VAG) methods 273

9 Non-conforming finite element methods 281
9.1 Non-conforming finite element methods for homogeneous Dirichlet BCs 281
9.1.1 Abstract framework 281
9.1.2 GDM formulation of abstract non-conforming finite element methods 286
9.2 $\mathbb{P}_k$ NCFE GDs for homogeneous Dirichlet BCs 287
9.3 Non-conforming $\mathbb{P}_1$ FE for homogeneous Dirichlet BCs 291
9.3.1 Definition of the non-conforming $\mathbb{P}_1$ gradient discretisation 291
9.3.2 Estimate on $S_{\mathcal{D}}$ for the non-conforming $\mathbb{P}_1$ LLE gradient discretisation 293
9.4 Non-conforming $\mathbb{P}_1$ methods for Neumann and Fourier BCs . . . 295
9.4.1 Neumann boundary conditions 295
9.4.2 Fourier boundary conditions 297
9.5 Non-conforming $\mathbb{P}_1$ FE for non-homogeneous Dirichlet BCs . . . 297
9.6 Mass-lumped non-conforming $\mathbb{P}_1$ gradient discretisation 298

10 H_{div} conforming GDs from mixed finite element methods . . 301
10.1 Mixed finite element schemes 301
10.1.1 Presentation and error estimate 301
10.1.2 Construction and analysis of a mixed finite element GD 302
10.2 The particular case of the $\mathbb{RT}_k$ mixed finite element 305
10.2.1 $\mathbb{RT}_k$ mixed finite element gradient discretisation 305
10.2.2 Gradient discretisation from the $\mathbb{RT}_k$ hybrid mixed finite element formulation 313

11 Discontinuous Galerkin methods 319
11.1 Discontinuous Galerkin gradient discretisation 319
11.1.1 Definition of the DGGD 319
11.1.2 Link with the symmetric interior penalty discontinuous Galerkin (SIPG) method 322
11.2 Mathematical properties of DG gradient discretisations 325
11.2.1 Preliminary results 325
11.2.2 Properties of DGGDs 330
11.3 Average DG gradient discretisation 334

12 The multi-point flux approximation MPFA-O scheme 337
12.1 MPFA methods, Dirichlet boundary conditions 337
12.1.1 Definition of the MPFA gradient discretisation 337
12.1.2 Preliminary lemmas 340
12.1.3 Properties of the MPFA-O gradient discretisation 343
12.2 MPFA-O methods, Neumann and Fourier boundary conditions 344
12.2.1 Neumann boundary conditions 344
12.2.2 Fourier boundary conditions 345

13 Hybrid mimetic mixed schemes . . . 347
13.1 HMM methods for Dirichlet boundary conditions . . . 348
13.1.1 Definition of HMM gradient discretisations . . . 348
13.1.2 Preliminary lemmas . . . 354
13.1.3 Properties of HMM gradient discretisations . . . 357
13.2 HMM methods, Neumann and Fourier boundary conditions . . . 360
13.2.1 Neumann boundary conditions . . . 360
13.2.2 Fourier boundary conditions . . . 361
13.3 HMM fluxes and link with the two-point finite volume method 361
13.4 A cell-centred variant of HMM schemes on Δ-admissible meshes 364
13.5 The harmonic averaging points for heterogeneous domains . . . 364
13.5.1 Harmonic interpolation coefficients . . . 365
13.5.2 Construction of interpolation families . . . 366

14 Nodal mimetic finite difference methods . . . 371
14.1 Definition and properties of nMFD gradient discretisations . . . 371
14.1.1 Preliminary lemmas . . . 379
14.1.2 Properties of nMFD gradient discretisations . . . 383
14.2 Link with discrete duality finite volume methods (DDFV) . . . 384

Part IV Appendix

A Gradient discretisations – abstract setting . . . 393
A.1 Continuous abstract setting . . . 393
A.1.1 Homogeneous Dirichlet BCs . . . 395
A.1.2 Homogeneous Neumann BCs . . . 395
A.1.3 Non-homogeneous Neumann BCs . . . 396
A.1.4 Fourier BCs . . . 396
A.2 Gradient discretisation in the abstract setting . . . 397

B Discrete functional analysis . . . 405
B.1 Preliminary results . . . 405
B.1.1 Geometrical properties of cells . . . 406
B.1.2 Interpolant on $X_{\mathfrak{T}}$ . . . 410
B.1.3 Approximation properties of $\overline{\nabla}_{\mathfrak{T}}$ . . . 415
B.2 Discrete functional analysis for Dirichlet boundary conditions . 418
B.2.1 Discrete Sobolev embeddings . . . 418
B.2.2 Compactness of $\Pi_{\mathfrak{T}}$ . . . 422
B.3 Discrete functional analysis for Neumann and Fourier BCs . . . 424
B.3.1 Estimates involving the reconstructed trace . . . 424
B.3.2 Discrete Sobolev embeddings . . . 431
B.3.3 Compactness of $\Pi_{\mathfrak{T}}$ and $\mathbb{T}_{\mathfrak{T}}$ . . . 433
B.4 Discrete functional analysis for mixed boundary condition . . . 435
B.4.1 Discrete Sobolev embeddings . . . 435

B.4.2 Compactness of $\Pi_{\mathfrak{T}}$ and $\mathbb{T}_{\mathfrak{T}}$ 438

C Discrete functional analysis for time-dependent problems . . 439
C.1 Averaged-in-time compactness results 439
C.2 Uniform-in-time compactness 449

D Technical results 457
D.1 Standard notations, inequalities and relations 457
D.1.1 $\mathbb{R}^d$ and measures 457
D.1.2 Lebesgue and Sobolev spaces 457
D.1.3 Hölder inequalities 458
D.1.4 Young inequality 459
D.1.5 Jensen inequality 460
D.1.6 Power of sums inequality 460
D.1.7 Discrete integration-by-parts (summation-by-parts) 460
D.2 Topological degree 462
D.3 Derivation and convergence in the sense of distributions 462
D.4 Weak and strong convergence results 464
D.5 Minty trick and convexity inequality 465

E Some numerical examples 469
E.1 A 3D elliptic problem 469
E.2 ADGGD for the $p-$Laplace problem 472
E.2.1 The one-dimensional case 472
E.2.2 The two-dimensional case 473
E.3 An example of the application of the GDM to a degenerate parabolic problem 475

References 481

Abbreviations 490

Notations 492

Introduction

The purpose of this book is the study of the gradient discretisation method (GDM), which includes a large family of conforming or non-conforming numerical methods for elliptic and parabolic partial differential equations (PDEs). A gradient discretisation method is based on the choice of a set of discrete spaces and operators, referred to as a "gradient discretisation" (GD). Replacing, in the weak formulation of a diffusion problem, the continuous space and operators by the discrete elements provided by a particular GD yields a numerical scheme called a "gradient scheme".

Considering here the case of homogeneous Dirichlet boundary conditions, the stationary linear and non-linear diffusion problems under study can be written under the form:

$$\begin{aligned} -\mathrm{div}\ \boldsymbol{a}(\boldsymbol{x}, \overline{u}, \nabla\overline{u}) &= f \quad \text{in } \Omega, \\ \overline{u} &= 0 \quad \text{on } \partial\Omega, \end{aligned}$$

where

- Ω is an open bounded connected subset of $\mathbb{R}^d$, $d \in \mathbb{N}^*$, with boundary $\partial\Omega = \overline{\Omega} \setminus \Omega$,
- $\boldsymbol{a}$ is a function from $\mathbb{R}^d \times \mathbb{R} \times \mathbb{R}^d$ to $\mathbb{R}^d$.

The function $\boldsymbol{a}$ may be a general anisotropic heterogeneous linear operator, that is $\boldsymbol{a}(\boldsymbol{x}, s, \boldsymbol{\xi}) = \Lambda(\boldsymbol{x})\boldsymbol{\xi}$, which yields a linear diffusion problem. Non-linear problems with a solution-dependent diffusion matrix can be considered by setting $\boldsymbol{a}(\boldsymbol{x}, s, \boldsymbol{\xi}) = \Lambda(\boldsymbol{x}, s)\boldsymbol{\xi}$. Another possible choice for $\boldsymbol{a}$ is a Leray–Lions operator such as the p-Laplacian $\boldsymbol{a}(\boldsymbol{x}, s, \boldsymbol{\xi}) = |\boldsymbol{\xi}|^{p-2}\boldsymbol{\xi}$ with $p > 1$, which yields a non-linear diffusion problem involving the gradient in the non-linearity, instead of just the function.
Standard diffusion evolution problems of the form

$$\begin{aligned} \partial_t\overline{u} - \mathrm{div}\, \boldsymbol{a}(\boldsymbol{x}, \overline{u}, \nabla\overline{u}) &= f && \text{in } \Omega \times (0, T), \\ \overline{u}(\cdot, 0) &= u_{\mathrm{ini}} && \text{in } \Omega, \\ \overline{u} &= 0 && \text{on } \partial\Omega \times (0, T), \end{aligned}$$

are treated, as well as degenerate evolution problems, such as the following:

$$\begin{aligned} \partial_t \beta(\overline{u}) - \Delta\zeta(\overline{u}) &= f && \text{in } \Omega \times (0,T), \\ \beta(\overline{u}(\cdot,0)) &= \beta(u_{\text{ini}}) && \text{in } \Omega, \\ \zeta(\overline{u}) &= 0 && \text{on } \partial\Omega \times (0,T), \end{aligned}$$

where the functions ζ and β are assumed to be Lipschitz-continuous and non-decreasing. This latter model includes both the Stefan problem, modelling a melting material, and the Richards problem, modelling a two phase flow in a porous medium under the assumption that the pressure of one of the phases is given.

The above problems arise in various frameworks, such as underground engineering (oil recovery, hydrology, nuclear waste disposals, etc.) and image processing. In the case of underground engineering, numerical simulations have to be performed on meshes adapted to the geological layers; complex geometries such as faults, vanishing layers, inclined wells must therefore be accounted for, along with highly heterogeneous permeability fields. Locally non-conforming refined meshes are thus often used. Furthermore, the problems to be solved can involve a coupled set of equations of various types (including coupling by convection–reaction terms, changes of phase, or algebraic equations modelling a chemical reaction). A large number of discretisation methods have been developed in the past 30 years to deal with these problems, for which conventional methods such as conforming finite elements are not well adapted. One of the first schemes in this direction is probably the nine point finite volume scheme [105] developed at Institut Français du Pétrole; this scheme is conservative and features consistent fluxes, but is unfortunately non-symmetric. A "diamond scheme" using a reconstructed gradient on diamond-shaped cells was proven to converge [54] under restrictive conditions on the mesh. A number of so-called "multi point flux approximation" (MPFA) schemes were also developed, and some of them were shown to be convergent, again under restrictive conditions on the mesh [1, 2, 3, 4, 5, 6, 8, 9]. Most of these schemes are non-symmetric.

The seeds of the GDM as presented here originated when trying to find a finite volume scheme with consistent fluxes for anisotropic problems on so-called Δ-admissible meshes, *i.e.*, meshes for which there exists a set of points – one point per cell – such that the lines joining the points of two neighbouring cells are orthogonal to the interface between these cells; examples of Δ-admissible meshes include triangles, rectangles, and Voronoï cells – see [91, chapter 3]. Denoting by K a cell of a Δ-admissible mesh, by $\mathcal{F}_K$ the set of its edges, and by $|\sigma|$, $\overline{\boldsymbol{x}}_\sigma$ and $\boldsymbol{n}_{K,\sigma}$, respectively, the measure, centre of mass and outward unit normal to $\sigma \in \mathcal{F}_K$, a discrete gradient was constructed [92, 93] by noting that for any vector $\boldsymbol{v} \in \mathbb{R}^2$, one has $|K|\boldsymbol{v} = \sum_{\sigma\in\mathcal{F}_K} |\sigma|\boldsymbol{v}\cdot\boldsymbol{n}_{K,\sigma}(\overline{\boldsymbol{x}}_\sigma - \boldsymbol{x}_K)$, where $\boldsymbol{x}_K$ is any point of the triangle K. This can also be written as the matrix identity

$$\sum_{\sigma\in\mathcal{F}_K} |\sigma|(\overline{\boldsymbol{x}}_\sigma - \boldsymbol{x}_K)\boldsymbol{n}_{K,\sigma}^T = |K|\mathrm{Id}. \tag{0.1}$$

Stabilising this gradient by using it together with the two-point flux approximation of [115] for an isotropic part of the diffusion matrix led to a consistent and stable scheme (this idea is used in Section 13.4 to construct a GD on Δ-admissible meshes, see Definition 13.40). However, the Δ-admissibility property for the meshes was too restrictive to handle the variety of meshes used in industry for the numerical simulation of flows in heterogeneous and anisotropic porous media. A large gap was filled in the direction of general polyhedral meshes thanks to the idea of transposing the formula (0.1), leading to

$$\sum_{\sigma\in\mathcal{F}_K} |\sigma|\boldsymbol{n}_{K,\sigma}(\overline{\boldsymbol{x}}_\sigma - \boldsymbol{x}_K)^T = |K|\mathrm{Id}.$$

Indeed, this new matrix identity (proven in Lemma B.3) is the key to the construction of a consistent gradient on general polytopal meshes as introduced in [94] and developed in [96]; it is also a main tool of the "polytopal toolbox" that is used in the present book to analyse the GDM, see Chapter 7. This idea was simultaneously and independently used by the teams at the origin of the mimetic family of schemes [130, 36, 35, 129, 34, 128] to construct the normal fluxes at each face of the mesh by solving a local linear system (for the method in [129], the invertibility of the local system is conditional to some mesh properties). Inspired by the finite volume ideas, other methods were devised and studied simultaneously. Among them, let us mention the discrete duality finite volume (DDFV) schemes introduced by Hermeline [117, 118, 119, 120, 121, 122], and studied in the linear case [64], and in different non-linear cases, see e.g. [28, 15, 14, 13, 44]. The DDFV scheme constructs a discrete gradient by using two meshes (in 2D) and writing a finite volume on each mesh. We refer to [69] for an introduction and review of all these finite volume schemes.

The profusion of new methods for anisotropic problems and distorted meshes led to the organisation of a benchmark whose results were presented at the FVCA conference of 2008 in Aussois [116]. It was clear that some methods, in particular the SUSHI scheme and the mixed-hybrid mimetic finite difference methods produced extremely similar outputs. This triggered a closer analysis which showed that these methods are in fact algebraically equivalent [75]. One disadvantage of these schemes is the use of interface unknowns in the construction of the discrete gradient, which is quite costly; their elimination is possible, as in the SUCCES scheme [96], but leads to wide stencils. This motivated the construction of a vertex approximated gradient (VAG); the resulting scheme was proved to be convergent [99] by identifying general abstract properties which, when satisfied by a scheme, ensure its convergence. These properties are called the "core properties" of GDs in this book, and were later shown to yield the right tools for the study of a larger variety of

problems, including non-linear models [76]. The general theory of GDMs was then set up [77].

We show in this book that several of the above mentioned recent methods as well as a number of classical methods are GDMs, in particular:

1. the conforming finite element methods, including mass lumped versions,
2. the non-conforming finite element methods, again including mass lumped versions,
3. the mixed finite element methods and in particular the Raviart–Thomas ones,
4. the symmetric interior penalty Galerkin version of the discontinuous Galerkin (DG) method,
5. the multi-point flux approximation (MPFA) schemes and the discrete duality finite volume (DDFV) schemes on particular grids,
6. the hybrid mimetic mixed (HMM) family of schemes, which includes the hybrid mimetic finite difference schemes, the SUSHI scheme and the mixed finite volume scheme,
7. the nodal mimetic finite difference scheme.

This already long list is not exhaustive. For example, research is ongoing to include recent high order methods in the GDM framework meshes, such as high order mimetic finite difference (MFD) methods [127], virtual element methods (VEM) [21, 38] and hybrid high order (HHO) methods [62]. It is shown in [59] that the HHO method and the non-conforming versions of MFD and VEM are gradient discretisation methods. Future work will certainly lead to view other low or high order numerical methods as GDMs.
Recent research on numerical schemes for elliptic and parabolic problems has led to other unification frameworks, like the Compatible Discrete Operator schemes [26, 27, 40]. Contrary to the GDM, this framework relies on a specific choice of unknowns on a mesh.

Organisation of the book.
Part I is an introduction to the GDM and its usage for elliptic equations.
Chapter 1 describes the basic concepts underlying this method.
The GDM is then formally introduced for elliptic PDEs with Dirichlet boundary conditions in Chapter 2. For linear equations, error estimates are obtained. For non-linear equations (including a Leray–Lions type model with a non-local dependency of the operator), convergence is obtained by compactness techniques.
The case of Neumann, Fourier and mixed Dirichlet/Neumann boundary conditions is covered in Chapter 3.

Part II is devoted to the study of the GDM for linear and non-linear parabolic problems.
Chapter 4 covers the definitions and main compactness results that are used to analyse the GDM for non-linear parabolic problems.

Chapter 5 deals with the linear and quasi-linear parabolic heat equations, with a non-conservative parabolic problem comin g from image processing, and finally with a transient Leray–Lions type problem.
Chapter 6 covers the study of the GDM applied to degenerate parabolic problems, including the Stefan and Richards problems.

In Part III, examples of schemes that fit into the GDM framework are presented.
Some discrete analysis tools are first introduced in Chapter 7; these are used later to establish that particular gradient discretisations satisfy the required properties for the convergence analysis of Parts I and II to hold.
Chapters 8 to 14 then list some important examples of GDMs. It is shown that the standard finite element method, the non-conforming $\mathbb{P}_k$ finite element method, the mixed finite element method and the discontinuous Galerkin method (in its symmetric interior penalty version) are GDMs. We then analyse, in the framework of GDM, some particular finite volume methods (multipoint flux approximation), the hybrid mimetic mixed family, and the nodal mimetic finite difference method.

An appendix gathers four chapters of useful tools for the analysis in the other parts of the book.
Chapter A provides an abstract setting covering a variety of boundary conditions. The generic properties proved in this chapter are used in Chapter 3 for Neumann, Fourier and mixed boundary conditions. They could also be used in the case of Dirichlet boundary conditions in Chapter 2, but detailed direct proofs were written in this initial chapter for pedagogical reasons.
The two next chapters are concerned with *discrete functional analysis*, a mathematical setting for the convergence analysis of numerical schemes. Chapter B establishes some discrete functional analysis tools for space discretisations based on polytopal meshes. These results are the discrete adaptations of the standard Sobolev and Rellich embedding theorems, with also considerations on discrete traces. Chapter C presents compactness results for time-dependent functions with abstract co-domains, including a discrete Aubin–Simon theorem and a discontinuous Ascoli–Arzelà theorem. It is worth noticing that the discrete functional analysis for space- and/or time-dependent functions developed in Appendices B and C is potentially applicable outside the GDM, to schemes which are not necessarily written in the form of gradient schemes; this is also the case for the convergence analysis tools for fully non-linear and degenerate models developed in Section 2.1.5 and Chapters 5 and 6.
In Chapter D, classical notations and results are gathered for the reader's convenience. Finally, in Chapter E, some numerical examples are proposed in order to illustrate the theoretical notions and schemes previously presented.

User guide:
This book is written assuming that the reader is familiar with Sobolev spaces and weak formulations of elliptic and parabolic partial differential equations

(PDEs). We refer to [33] for an introduction on this topic. The reader should also have some notions of numerical analysis, in particular of the discretisation of elliptic and parabolic PDEs – for example the knowledge of one of the aforementioned methods, such as the conforming $\mathbb{P}_1$ finite elements on triangles.
Parts of this book are of easy access, others require more work. We recommend that students or newcomers to the field follow the discovery track below, while more advanced or expert researchers can follow the eponymous tracks.

- **Discovery track:** Read Chapter 1. Then read Section 2.1.1 in Chapter 2, focusing more on the definitions and their explanations than on the proofs of the lemmas. Then read Sections 2.1.2 and 2.1.4, and the case of non-conforming finite elements in Chapter 9, referring when needed to the definitions and statements of Chapter 7. Appendix D presents some notations and classical results used throughout the book; initially skimming through this appendix and then coming back for specific details might prove beneficial.
- **Advanced track:** In addition to the preceding track, read Chapter 5, referring when needed to the definitions and results of Chapter 4. Consider the case of mass-lumped conforming finite elements in Chapter 8, referring to the definitions and statements of Chapter 7.
- **Expert track:** Welcome to the whole book.

Remark 0.1 (Shaded remarks)
Shaded remarks such as the present one contain notions, comments or results that can be somewhat technical or very specific. In a first reading, these remarks can be skipped.

Part I

Elliptic problems

1

Motivation and basic ideas

Throughout this book, Ω – the physical domain over which PDEs are considered – is a connected open bounded subset of $\mathbb{R}^d$, $d \in \mathbb{N}^\star$ is the space dimension, and $p \in (1, +\infty)$ denotes a regularity index of the sought solution. For linear and quasi-linear problems, we take $p = 2$. In some abstract theorems, p might be allowed to take the value 1.

1.1 Some well-known approximations of linear elliptic problems

Let us consider the following simple elliptic problem:

$$\begin{cases} -\Delta \overline{u} = f & \text{in } \Omega, \\ \overline{u} = 0 & \text{on } \partial\Omega, \end{cases} \tag{1.1}$$

where $f \in L^2(\Omega)$. The weak formulation of (1.1) is:

$$\begin{aligned} &\text{Find } \overline{u} \in H_0^1(\Omega) \text{ such that, for all } v \in H_0^1(\Omega), \\ &\int_\Omega \nabla \overline{u}(\boldsymbol{x}) \cdot \nabla v(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} = \int_\Omega f(\boldsymbol{x}) v(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}. \end{aligned} \tag{1.2}$$

1.1.1 Galerkin methods

A classical family of numerical methods to approximate this problem is given by conforming Galerkin methods. Their main idea is to seek the approximate solution in an approximation space which is a finite dimensional subspace V_h of $H_0^1(\Omega)$. This is for example the case for the well-known $\mathbb{P}_1$ finite element method, in which a partition of Ω into simplices (*e.g.* triangles in dimension $d = 2$) is chosen and the approximation space V_h is made of the piecewise linear functions on this partition, which are continuous over Ω and have a zero value on $\partial\Omega$. In such a case, the index h denotes the mesh size, see *e.g.* [48] for more on finite element approximations.

Once a finite dimensional subspace V_h of $H_0^1(\Omega)$ has been chosen, the Galerkin approximation of (1.2) is

$$\begin{aligned}&\text{Find } u_h \in V_h \text{ such that, for all } v_h \in V_h,\\ &\int_\Omega \nabla u_h(\boldsymbol{x}) \cdot \nabla v_h(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} = \int_\Omega f(\boldsymbol{x}) v_h(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}.\end{aligned} \tag{1.3}$$

It is then easy to establish an error bound between the weak solution $\overline{u}$ to (1.1) and the approximate solution u_h. Using a generic $v = v_h \in V_h \subset H_0^1(\Omega)$ as a test function in (1.2) and subtracting (1.3) we see that

$$\int_\Omega \nabla(\overline{u} - u_h)(\boldsymbol{x}) \cdot \nabla v_h(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} = 0. \tag{1.4}$$

Taking $v_h = w_h - u_h$, where w_h is any function in V_h, and writing $v_h = w_h - \overline{u} + \overline{u} - u_h$ gives

$$\int_\Omega \nabla(\overline{u} - u_h)(\boldsymbol{x}) \cdot \nabla(\overline{u} - u_h)(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} = \int_\Omega \nabla(\overline{u} - u_h)(\boldsymbol{x}) \cdot \nabla(\overline{u} - w_h)(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}.$$

Using the Cauchy–Schwarz inequality (that is, (D.5) with $p = p' = 2$) in the right-hand side and recalling that $\|\varphi\|_{H_0^1(\Omega)}^2 = \int_\Omega |\nabla\varphi(\boldsymbol{x})|^2 \mathrm{d}\boldsymbol{x}$, it is inferred that

$$\|\overline{u} - u_h\|_{H_0^1(\Omega)}^2 \leq \|\overline{u} - u_h\|_{H_0^1(\Omega)} \|\overline{u} - w_h\|_{H_0^1(\Omega)}.$$

Finally, since this estimate is valid for any $w_h \in V_h$,

$$\|\overline{u} - u_h\|_{H_0^1(\Omega)} \leq \min_{w_h \in V_h} \|w_h - \overline{u}\|_{H_0^1(\Omega)}. \tag{1.5}$$

This result, which may be generalised to other problems than (1.1), is known as Céa's lemma [49, Theorem 2.4.1].
Assume that a family of subspaces $(V_h)_{h>0}$ is "ultimately dense" in $H_0^1(\Omega)$ as $h \to 0$, *i.e.*, for all $\varphi \in H_0^1(\Omega)$,

$$\min_{w_h \in V_h} \|w_h - \varphi\|_{H_0^1(\Omega)} \to 0 \text{ as } h \to 0.$$

Then Estimate (1.5) shows that $u_h \to \overline{u}$ in $H_0^1(\Omega)$ as $h \to 0$.

The beauty of this analysis lies in its simplicity. It is however limited to methods for which the approximation space V_h is included in the space in which the continuous solution lives. These methods are referred to as "conforming". Numerous numerical schemes for elliptic equations are "non-conforming" in the sense that the provided approximate solutions do not belong to $H_0^1(\Omega)$. This happens for instance in the case of the non-conforming $\mathbb{P}_1$ finite element, which yields a piecewise affine approximation that is not necessarily continuous across the edges, and in the case of cell-centred finite volume schemes, which yield a piecewise constant approximation.

1.1.2 Non-conforming $\mathbb{P}_1$ finite elements

Another well known method is the non-conforming finite element method, a simple version of which is given here – see Chapter 9 for a more detailed and general presentation. Let $\mathcal{M}$ be a conforming triangular mesh of $\Omega \subset \mathbb{R}^2$, that is, a mesh made of triangles and such that no edge of any triangle contains a vertex other than its two endpoints. Let $\mathcal{F}$ be the finite set of the edges of the mesh, $\mathcal{F}_{\text{ext}}$ be the set of all $\sigma \in \mathcal{F}$ such that $\sigma \subset \partial\Omega$, and $\mathcal{F}_{\text{int}} = \mathcal{F} \setminus \mathcal{F}_{\text{ext}}$ be the set of interior edges. For any $\sigma \in \mathcal{F}$, $\overline{\boldsymbol{x}}_\sigma$ is the centre of mass of σ. The approximation space V_h of the non-conforming $\mathbb{P}_1$ finite element method is the set of piecewise affine functions on the triangles of the mesh such that, for all $\sigma \in \mathcal{F}_{\text{int}}$ between two cells K and L, denoting by $\mathrm{d}\gamma$ the measure on σ,

$$\int_\sigma (u_h)_{|K} \mathrm{d}\gamma(\boldsymbol{x}) = \int_\sigma (u_h)_{|L} \mathrm{d}\gamma(\boldsymbol{x}) \tag{1.6}$$

and, for all $\sigma \in \mathcal{F}_{\text{ext}}$, if K is the cell whose σ is an edge,

$$\int_\sigma (u_h)_{|K} \mathrm{d}\gamma(\boldsymbol{x}) = 0. \tag{1.7}$$

A part of such a function is depicted in Figure 1.1. The space V_h is spanned by the basis $(\varphi_\sigma)_{\sigma \in \mathcal{F}_{\text{int}}}$, where φ_σ is the piecewise affine function such that $\varphi_\sigma(\overline{\boldsymbol{x}}_\sigma) = 1$ and $\varphi_\sigma(\overline{\boldsymbol{x}}_{\sigma'}) = 0$ for all $\sigma' \in \mathcal{F} \setminus \{\sigma\}$. The space V_h is clearly

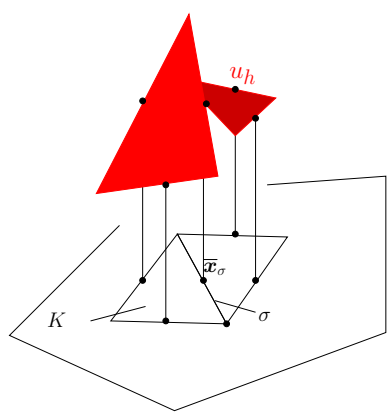


Fig. 1.1. Non-conforming $\mathbb{P}_1$ finite element.

not a subspace of $H_0^1(\Omega)$; however, the restriction to any cell of a function of V_h is piecewise affine, so its gradient is well defined and constant in each cell. For $K \in \mathcal{M}$, let $\nabla_K \varphi_\sigma$ be the constant value of the gradient of the function φ_σ, $\sigma \in \mathcal{F}$, on K (note that $\nabla_K \varphi_\sigma = 0$ if σ is not an edge of K). It is remarkable that (1.3) still makes sense if the gradient operator ∇ in this formula is replaced by the "broken" gradient operator $\nabla_{\mathcal{M}}$ defined by

$$\begin{aligned}&\text{For any } u_h = \sum_{\sigma\in\mathcal{F}_{\rm int}} u_\sigma \varphi_\sigma \in V_h,\\ &\forall K\in\mathcal{M}\,, \forall \boldsymbol{x}\in K\,,\ \nabla_{\mathcal{M}} u_h(\boldsymbol{x}) = \sum_{\sigma\in\mathcal{F}_{\rm int}} u_\sigma \nabla_K \varphi_\sigma\end{aligned} \tag{1.8}$$

(in other words, the gradients are computed without taking into account the jumps along the edges). Then the following norm is defined on $H_0^1(\Omega)+V_h$:

$$\forall u_h \in H_0^1(\Omega)+V_h\,,\ \|u_h\|_h^2 = \sum_{K\in\mathcal{M}} \int_K |\nabla u_h(\boldsymbol{x})|^2 \mathrm{d}\boldsymbol{x}.$$

Let us check that $\|\cdot\|_h$ is indeed a norm. It is clearly a semi-norm and, if $\|u_h\|_h = 0$ then u_h is constant in each cell K. Since $u_h \in H_0^1(\Omega)+V_h$, by (1.6) (also valid for functions in $H_0^1(\Omega)$) $(u_h)_{|K} = (u_h)_{|L}$ whenever K and L are neighbouring cells. Working from neighbour to neighbour, u_h is then seen to be constant over the connected domain Ω; the relation (1.7) written for one boundary edge finally shows that $u_h = 0$ in Ω.
Let $a_h\ :\ (H_0^1(\Omega)+V_h)^2 \to \mathbb{R}$ be the bilinear form defined by

$$\forall (u_h, v_h) \in\ (H_0^1(\Omega)+V_h)^2\,,\ a_h(u_h,v_h) = \sum_{K\in\mathcal{M}} \int_K \nabla u_h(\boldsymbol{x})\cdot\nabla v_h(\boldsymbol{x})\mathrm{d}\boldsymbol{x}.$$

The approximate problem to Problem (1.1) is defined by

$$\text{Find } u_h\in V_h \text{ such that, } \forall v_h\in V_h,\ a_h(u_h,v_h) = \int_\Omega f(\boldsymbol{x}) v_h(\boldsymbol{x})\mathrm{d}\boldsymbol{x}. \tag{1.9}$$

There exists one and only one solution to (1.9), and it satisfies the following error estimate [49, Theorem 4.2.2], based on the second Strang Lemma (see [132] and [133, Section 4.2]): for some $C>0$, depending only on the regularity of $\mathcal{M}$ but not on h,

$$\begin{aligned}&\|\overline{u}-u_h\|_h\\ &\le C\left[\inf_{v_h\in V_h}\|\overline{u}-v_h\|_h + \sup_{w_h\in V_h\setminus\{0\}} \frac{\left|a_h(\overline{u},w_h) - \displaystyle\int_\Omega f(\boldsymbol{x})w_h(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\right|}{\|w_h\|_h}\right].\end{aligned} \tag{1.10}$$

This estimate can be written as

$$\|\overline{u}-u_h\|_h \le C(S_{\mathcal{M}}(\overline{u}) + W_{\mathcal{M}}(\nabla\overline{u})), \tag{1.11}$$

where $S_{\mathcal{M}}(\varphi)$ is defined, for any $\varphi\in H_0^1(\Omega)$, by

$$S_{\mathcal{M}}(\varphi) = \inf_{v_h\in V_h}\|\varphi - v_h\|_h, \tag{1.12}$$

and $W_{\mathcal{M}}(\boldsymbol{\varphi})$ is defined, for any sufficiently regular function $\boldsymbol{\varphi}:\Omega\to\mathbb{R}^d$, by

$$W_{\mathcal{M}}(\boldsymbol{\varphi}) = \sup_{w_h \in V_h \setminus \{0\}} \frac{\left| \int_\Omega (\boldsymbol{\varphi}(\boldsymbol{x}) \cdot \nabla_{\mathcal{M}} w_h(\boldsymbol{x}) + \mathrm{div}\boldsymbol{\varphi}(\boldsymbol{x}) w_h(\boldsymbol{x})) \mathrm{d}\boldsymbol{x} \right|}{\|w_h\|_h}. \tag{1.13}$$

To relate $W_{\mathcal{M}}(\nabla \overline{u})$ in (1.11) with the last term in (1.10), notice simply that $\mathrm{div}(\nabla \overline{u}) = \Delta \overline{u} = -f$. Under regularity assumptions on the mesh, the quantities $S_{\mathcal{M}}(\varphi)$ and $W_{\mathcal{M}}(\boldsymbol{\varphi})$ tend to zero as the size of the mesh tends to zero [48, 84].

1.1.3 Two-point flux approximation finite volumes on Cartesian meshes

A second example of non-conforming scheme is given by the "two-point flux approximation" (TPFA) finite volume scheme [91]. The TPFA scheme is widely used in petroleum engineering: constant values are considered in control volumes over which a discrete mass balance of the various components is established. The particular case of the TPFA scheme for Cartesian grids is considered here and denoted by TPFA-CG. Let $\mathcal{M}$ be a rectangular mesh of a rectangle $\Omega \subset \mathbb{R}^2$. In addition to the notations K, σ and $\overline{\boldsymbol{x}}_\sigma$ defined in Section 1.1.2, we introduce the following (see Figure 1.2):

- for each $K \in \mathcal{M}$, $\boldsymbol{x}_K$ is the intersection of the bisectors of the edges of K (since K is a rectangle, $\boldsymbol{x}_K$ is also the centre of mass of K) and $\mathcal{F}_K$ is the set of edges of K,
- $\mathcal{V}$ is the set of vertices of the mesh and, for $K \in \mathcal{M}$, $\mathcal{V}_K$ is the set of vertices of K,
- for each $K \in \mathcal{M}$ and each $\boldsymbol{s} \in \mathcal{V}_K$, $V_{K,\boldsymbol{s}}$ is the rectangle defined by $\overline{\boldsymbol{x}}_\sigma$, $\boldsymbol{s}$, $\overline{\boldsymbol{x}}_{\sigma'}$ and $\boldsymbol{x}_K$, where σ and σ' are the edges of K touching $\boldsymbol{s}$,
- u_K (resp. u_σ) represents an approximate value of the unknown u at $\boldsymbol{x}_K$ (resp. $\overline{\boldsymbol{x}}_\sigma$).

The idea of finite volume schemes consists in finding approximate values $F_{K,\sigma}$ of the exact fluxes $-\int_\sigma \nabla \overline{u} \cdot \boldsymbol{n}_{K,\sigma} \mathrm{d}\gamma(\boldsymbol{x})$ ($\boldsymbol{n}_{K,\sigma}$ is the normal to σ outward K), and in writing the following discrete flux balance in each cell

$$\forall K \in \mathcal{M}\,, \sum_{\sigma \in \mathcal{F}_K} F_{K,\sigma} = \int_K f(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}, \tag{1.14}$$

and the flux conservativity across each interior edge:

$$\forall \sigma \in \mathcal{F}_{\mathrm{int}} \text{ common face of } K \text{ and } L\,,\ F_{K,\sigma} + F_{L,\sigma} = 0. \tag{1.15}$$

Relation (1.14) simply mimicks the Stokes formula applied to the continuous problem (1.1):

$$-\sum_{\sigma \in \mathcal{F}_K} \int_\sigma \nabla u \cdot \boldsymbol{n}_{K,\sigma} \mathrm{d}\gamma(\boldsymbol{x}) = \int_K f(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}.$$

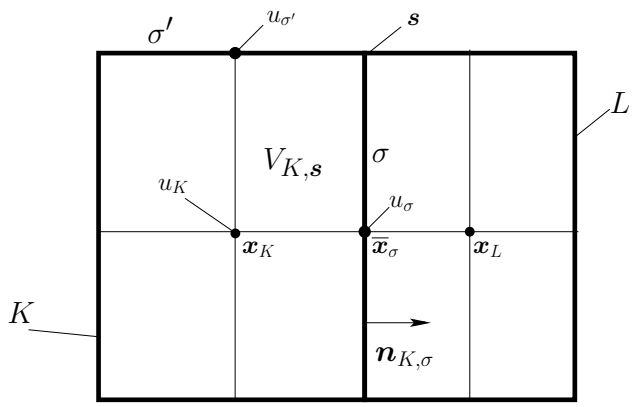


Fig. 1.2. Notation for a rectangular mesh.

The TPFA-CG finite volume scheme consists in substituting, in the previous equations,

$$F_{K,\sigma} = -|\sigma| \frac{u_\sigma - u_K}{\mathrm{dist}(\overline{\boldsymbol{x}}_\sigma, \boldsymbol{x}_K)}. \tag{1.16}$$

The boundary condition is imposed by setting

$$u_\sigma = 0 \text{ if } \sigma \subset \partial\Omega. \tag{1.17}$$

There is no clear way to see the TPFA-CG scheme method as a non-conforming finite element method. However, it can be recast into a variational form. Consider a family $((v_K)_{K\in\mathcal{M}}, (v_\sigma)_{\sigma\in\mathcal{F}})$ such that $v_\sigma = 0$ if $\sigma \subset \partial\Omega$. Multiplying (1.14) by v_K and summing over $K \in \mathcal{M}$ yields

$$\sum_{K\in\mathcal{M}} \sum_{\sigma\in\mathcal{F}_K} v_K F_{K,\sigma} = \sum_{K\in\mathcal{M}} v_K \int_K f(\boldsymbol{x})\mathrm{d}\boldsymbol{x}. \tag{1.18}$$

Notice then that

$$\sum_{K\in\mathcal{M}} \sum_{\sigma\in\mathcal{F}_K} v_K F_{K,\sigma} = \sum_{K\in\mathcal{M}} \sum_{\sigma\in\mathcal{F}_K} (v_K - v_\sigma) F_{K,\sigma}. \tag{1.19}$$

Indeed, if $\sigma \subset \partial\Omega$, then $(v_K - v_\sigma)F_{K,\sigma} = v_K F_{K,\sigma}$ and, if σ is the common face between two control volumes K and L, then v_σ is multiplied in (1.19) by $F_{K,\sigma} + F_{L,\sigma}$, which vanishes thanks to (1.15). Thus, using (1.19) into (1.18) and invoking (1.16) leads to

$$\sum_{K\in\mathcal{M}} \sum_{\sigma\in\mathcal{F}_K} \frac{|\sigma|}{\mathrm{dist}(\overline{\boldsymbol{x}}_\sigma, \boldsymbol{x}_K)} (v_\sigma - v_K)(u_\sigma - u_K) = \sum_{K\in\mathcal{M}} v_K \int_K f(\boldsymbol{x})\mathrm{d}\boldsymbol{x}. \tag{1.20}$$

Conversely, assuming the boundary conditions (1.17) and the expression (1.16) for the fluxes, Relations (1.14) and (1.15) can be deduced from (1.20) with

appropriate choices of the family $((v_K)_{K\in\mathcal{M}}, (v_\sigma)_{\sigma\in\mathcal{F}})$ (namely, selecting only one value equal to 1 and all the other values equal to 0). Moreover, Relation (1.20) can be expressed in terms of reconstructed functions and gradients, using the discrete values defined on K and σ.

- Define $X_{\mathcal{D},0}$ as the space of real families $u_\mathcal{D} = ((u_K)_{K\in\mathcal{M}}, (u_\sigma)_{\sigma\in\mathcal{F}})$ satisfying the boundary conditions (1.17).
- For $u_\mathcal{D} \in X_{\mathcal{D},0}$, let $\Pi_\mathcal{D} u_\mathcal{D}$ be the piecewise constant function equal to u_K on the cell K.
- If $K \in \mathcal{M}$ and $\boldsymbol{s} \in \mathcal{V}_K$ is such that σ and σ' are the edges of K sharing the vertex $\boldsymbol{s}$, reconstruct a gradient by setting

$$\nabla_{K,\boldsymbol{s}} u_\mathcal{D} = \frac{u_\sigma - u_K}{\mathrm{dist}(\overline{\boldsymbol{x}}_\sigma, \boldsymbol{x}_K)}\, \boldsymbol{n}_{K,\sigma} + \frac{u_{\sigma'} - u_K}{\mathrm{dist}(\overline{\boldsymbol{x}}_{\sigma'}, \boldsymbol{x}_K)}\, \boldsymbol{n}_{K,\sigma'}.$$

  Denote then by $\nabla_\mathcal{D} u_\mathcal{D}$ the piecewise constant function equal to $\nabla_{K,\boldsymbol{s}} u_\mathcal{D}$ on $V_{K,\boldsymbol{s}}$, for any cell K and any vertex $\boldsymbol{s} \in \mathcal{V}_K$.

The following properties arise, for $(u_\mathcal{D}, v_\mathcal{D}) \in X^2_{\mathcal{D},0}$:

$$\sum_{K\in\mathcal{M}} v_K \int_K f(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = \int_\Omega f(\boldsymbol{x}) \Pi_\mathcal{D} v_\mathcal{D}(\boldsymbol{x})\mathrm{d}\boldsymbol{x},$$

and, using the orthogonality of $\boldsymbol{n}_{K,\sigma}$ and $\boldsymbol{n}_{K,\sigma'}$ when σ and σ' are two edges of K sharing a vertex $\boldsymbol{s}$,

$$\sum_{K\in\mathcal{M}} \sum_{\sigma\in\mathcal{F}_K} \frac{|\sigma|}{\mathrm{dist}(\overline{\boldsymbol{x}}_\sigma, \boldsymbol{x}_K)} (v_\sigma - v_K)(u_\sigma - u_K) = \int_\Omega \nabla_\mathcal{D} u_\mathcal{D}(\boldsymbol{x}) \cdot \nabla_\mathcal{D} v_\mathcal{D}(\boldsymbol{x})\mathrm{d}\boldsymbol{x}.$$

As a result, (1.20) can be recast in the form of a discrete variational problem:

$$\begin{aligned} &\text{Find } u_\mathcal{D} \in X_{\mathcal{D},0} \text{ such that, for all } v_\mathcal{D} \in X_{\mathcal{D},0}, \\ &\int_\Omega \nabla_\mathcal{D} u_\mathcal{D}(\boldsymbol{x}) \cdot \nabla_\mathcal{D} v_\mathcal{D}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = \int_\Omega f(\boldsymbol{x}) \Pi_\mathcal{D} v_\mathcal{D}(\boldsymbol{x})\mathrm{d}\boldsymbol{x}. \end{aligned} \tag{1.21}$$

The study of the TPFA-CG scheme was performed in [91] using finite volume techniques, and the following results were obtained: if the size of the mesh tends to 0 while the ratio height/width of each cell remains bounded above and below, then $\Pi_\mathcal{D} u_\mathcal{D}$ converges to $\overline{u}$ in $L^2(\Omega)$ and an error estimate holds, which depends on the regularity of $\overline{u}$.

Remark 1.1 (TPFA on unstructured meshes)
The TPFA scheme can also be analysed on unstructured meshes, provided an orthogonality condition holds (see [91, Definition 9.1]). However, in the unstructured case, it does not seem in general possible to write the scheme under the form (1.21), and the TPFA scheme is not known to be a GDM as described in this book (but it can be included in an *asymmetric* version of the GDM [73]).

One case, though, where the TPFA scheme can be written as (1.21) is that of "super-admissible" meshes, that is, when the point $\boldsymbol{x}_K$ in each cell K is the intersection of the orthogonal bisectors of the faces of K. Rectangular and acute triangular meshes are examples of super-admissible meshes. In this case, the HMM method described in Chapter 13 contains the TPFA scheme, see Section 13.3.

Now, could the TPFA-CG scheme be studied using the non-conforming techniques of Section 1.1.2? There are unfortunately a series of objections to this approach:

1. Comparing the right-hand sides of (1.9) and (1.21), the natural space V_h would be
$$V_h = \{\Pi_{\mathcal{D}} v_{\mathcal{D}} \,:\, v_{\mathcal{D}} \in X_{\mathcal{D},0}\}.$$
However, this space "forgets" about the edge unknowns $(v_\sigma)_{\sigma\in\mathcal{F}}$ of $v_{\mathcal{D}} \in X_{\mathcal{D},0}$, and there is therefore no way to compute $\nabla_{\mathcal{D}} v_{\mathcal{D}}$ solely from $\Pi_{\mathcal{D}} v_{\mathcal{D}}$.
2. Partially as a consequence of the previous item, there does not seem to exist any bilinear form a_h, defined on $V_h + H_0^1(\Omega)$, which would be equal to $\int_\Omega \nabla_{\mathcal{D}} u_{\mathcal{D}}(\boldsymbol{x}) \cdot \nabla_{\mathcal{D}} v_{\mathcal{D}}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}$ for $(u_{\mathcal{D}}, v_{\mathcal{D}}) \in V_h^2$, and to $\int_\Omega \nabla \overline{u}(\boldsymbol{x}) \cdot \nabla v(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}$ for $(\overline{u}, v) \in H_0^1(\Omega)^2$.
3. The same problem arises for the definition of the norm $\|\cdot\|_h$.

Although the technique from non-conforming finite elements schemes cannot be directly used on the TPFA-CG scheme, there is however a way of merging these two kinds of schemes into on common framework, which also covers conforming finite element methods. The next section presents an introduction to this framework.

1.2 Towards the gradient discretisation method

What does it take to design a unified convergence analysis framework covering the preceding three examples, as well as other conforming and non-conforming methods?

A numerical method obviously starts from selecting a finite number of discrete unknowns describing the finite dimensional space in which the approximate solution is sought. This finite dimensional space was called $X_{\mathcal{D},0}$ in the previous section ("$\mathcal{D}$" for "discretisation", and the 0 to indicate that, in some way, this space accounts for the homogeneous boundary condition in (1.1)). The two linear operators $\Pi_{\mathcal{D}}$ and $\nabla_{\mathcal{D}}$, which respectively reconstruct, from the discrete unknowns, a function on Ω and its "gradient", are such that

$$\Pi_{\mathcal{D}} : X_{\mathcal{D},0} \to L^2(\Omega) \quad \text{and} \quad \nabla_{\mathcal{D}} : X_{\mathcal{D},0} \to L^2(\Omega)^d.$$

All the schemes presented in the previous section can be written as

$$\text{Find } u_{\mathcal{D}} \in X_{\mathcal{D},0} \text{ such that, for all } v_{\mathcal{D}} \in X_{\mathcal{D},0},$$
$$\int_\Omega \nabla_{\mathcal{D}} u_{\mathcal{D}}(\boldsymbol{x}) \cdot \nabla_{\mathcal{D}} v_{\mathcal{D}}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} = \int_\Omega f(\boldsymbol{x}) \Pi_{\mathcal{D}} v_{\mathcal{D}}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \tag{1.22}$$

for suitable choices of $(X_{\mathcal{D},0}, \Pi_{\mathcal{D}}, \nabla_{\mathcal{D}})$. Indeed,

- For conforming $\mathbb{P}_1$ finite elements, each $v_{\mathcal{D}} \in X_{\mathcal{D},0}$ is a vector of values at the vertices of the mesh, $\Pi_{\mathcal{D}} v_{\mathcal{D}} \in C(\overline{\Omega})$ is the piecewise linear function on the mesh that takes these values at the vertices, and $\nabla_{\mathcal{D}} v_{\mathcal{D}} = \nabla(\Pi_{\mathcal{D}} v_{\mathcal{D}})$.
- For non-conforming $\mathbb{P}_1$ elements, each $v_{\mathcal{D}} \in X_{\mathcal{D},0}$ is a vector of values at the centres of mass of the edges, $\Pi_{\mathcal{D}} v_{\mathcal{D}}$ is the piecewise linear function on the mesh which takes these values at these centres of mass, and $\nabla_{\mathcal{D}} v_{\mathcal{D}} = \nabla_{\mathcal{M}}(\Pi_{\mathcal{D}} v_{\mathcal{D}})$ is the broken gradient defined in (1.8).
- The space and operators for the TPFA-CG scheme have already been given under the form $(X_{\mathcal{D},0}, \Pi_{\mathcal{D}}, \nabla_{\mathcal{D}})$ in Section 1.1.3.

The question now is to understand which properties the triplet $(X_{\mathcal{D},0}, \Pi_{\mathcal{D}}, \nabla_{\mathcal{D}})$ must satisfy to enable some error estimates between the solution $\overline{u}$ to (1.2) and the solution $u_{\mathcal{D}}$ to (1.22) (assuming for the time being that it exists). The main issue is that, contrary to Problem (1.2) and its conforming discretisation (1.3), Problem (1.2) and its general discretisation (1.22) do not appear to have any common test functions. Hence, no equation equivalent to (1.4) seems attainable. There is however a way to write an approximate version of this relation in the same spirit as in the analysis of the non-conforming finite element method. Contrary to the non-conforming finite element where the broken gradient is computed from the approximate function u_h, in the general GDM framework, the approximate function $\Pi_{\mathcal{D}} u$ does not always allow the computation of $\nabla_{\mathcal{D}} u$ (indeed, $\Pi_{\mathcal{D}} u$ does not necessarily involve all the components of the discrete unknown vector u). For instance, in the simple case of the Laplace operator (1.1), the left hand side of the numerical scheme involves only $\nabla_{\mathcal{D}} u$ while the right-hand-side involves only $\Pi_{\mathcal{D}} u$. The operators $\nabla_{\mathcal{D}}$ and $\Pi_{\mathcal{D}}$ are not deduced from one another, but they are not completely independent: a compatibility condition between the operators $\Pi_{\mathcal{D}}$ and $\nabla_{\mathcal{D}}$ is enforced through a so-called *limit-conformity* relation, see (1.28)-(P3) below. As already mentioned, this is mandatory for the TPFA-CG scheme to be part of this framework.

By noticing that (1.2) implies that $-\Delta \overline{u} = f$ in the sense of distributions, we get from (1.22) that, for any $v_{\mathcal{D}} \in X_{\mathcal{D},0}$,

$$\int_\Omega \nabla_{\mathcal{D}} u_{\mathcal{D}}(\boldsymbol{x}) \cdot \nabla_{\mathcal{D}} v_{\mathcal{D}}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} = \int_\Omega -\Delta \overline{u}(\boldsymbol{x}) \Pi_{\mathcal{D}} v_{\mathcal{D}}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}. \tag{1.23}$$

If $\Pi_{\mathcal{D}} v_{\mathcal{D}}$ were a classical regular function, the Stokes formula would allow us to replace the integrand in the right-hand side with $\nabla \overline{u}(\boldsymbol{x}) \cdot \nabla(\Pi_{\mathcal{D}} v_{\mathcal{D}})(\boldsymbol{x})$. Except in some particular cases, the discrete operators $\Pi_{\mathcal{D}}$, $\nabla_{\mathcal{D}}$ of a numerical scheme do not satisfy an exact discrete Stokes formula, only an approximate

one. We measure the resulting defect of conformity of the method, in the spirit of (1.13), by a function $W_{\mathcal{D}}(\boldsymbol{\varphi})$ such that, for any sufficiently regular function $\boldsymbol{\varphi} : \Omega \to \mathbb{R}^d$,

$$W_{\mathcal{D}}(\boldsymbol{\varphi}) = \sup_{v_{\mathcal{D}} \in X_{\mathcal{D},0} \setminus \{0\}} \frac{\left| \int_\Omega (\boldsymbol{\varphi}(\boldsymbol{x}) \cdot \nabla_{\mathcal{D}} v_{\mathcal{D}}(\boldsymbol{x}) + \mathrm{div}\boldsymbol{\varphi}(x) \Pi_{\mathcal{D}} v_{\mathcal{D}}(\boldsymbol{x})) \mathrm{d}\boldsymbol{x} \right|}{\|\nabla_{\mathcal{D}} v_{\mathcal{D}}\|_{L^2(\Omega)^d}}. \quad (1.24)$$

Here, we assume that $\|\nabla_{\mathcal{D}} v_{\mathcal{D}}\|_{L^2(\Omega)^d} \neq 0$ if $v_{\mathcal{D}} \neq 0$, which is somewhat natural given the homogeneous boundary conditions. The quantity $W_{\mathcal{D}}(\boldsymbol{\varphi})$ is expected to be small if the discretisation is "fine enough" (*e.g.*, the underlying mesh size is small). Then, considering $\boldsymbol{\varphi} = \nabla \overline{u}$ in (1.24) and using (1.23) to compute $\int_\Omega \mathrm{div}(\boldsymbol{\varphi})(\boldsymbol{x}) \Pi_{\mathcal{D}} v_{\mathcal{D}}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} = \int_\Omega \Delta \overline{u}(\boldsymbol{x}) \Pi_{\mathcal{D}} v_{\mathcal{D}}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}$, an approximate version of (1.4) is obtained:

$$\int_\Omega (\nabla \overline{u}(\boldsymbol{x}) - \nabla_{\mathcal{D}} u_{\mathcal{D}}(\boldsymbol{x})) \cdot \nabla_{\mathcal{D}} v_{\mathcal{D}}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \leq \|\nabla_{\mathcal{D}} v_{\mathcal{D}}\|_{L^2(\Omega)^d} W_{\mathcal{D}}(\nabla \overline{u}).$$

Take now a generic $w_{\mathcal{D}} \in X_{\mathcal{D},0}$, apply this estimate to $v_{\mathcal{D}} = w_{\mathcal{D}} - u_{\mathcal{D}}$, and write $\nabla \overline{u} - \nabla_{\mathcal{D}} u_{\mathcal{D}} = \nabla \overline{u} - \nabla_{\mathcal{D}} w_{\mathcal{D}} + \nabla_{\mathcal{D}} w_{\mathcal{D}} - \nabla_{\mathcal{D}} u_{\mathcal{D}}$ to find

$$\begin{aligned} \int_\Omega (\nabla_{\mathcal{D}} w_{\mathcal{D}}(\boldsymbol{x}) - \nabla_{\mathcal{D}} u_{\mathcal{D}}(\boldsymbol{x})) \cdot (\nabla_{\mathcal{D}} w_{\mathcal{D}}(\boldsymbol{x}) - \nabla_{\mathcal{D}} u_{\mathcal{D}}(\boldsymbol{x})) \mathrm{d}\boldsymbol{x} \\ \leq \int_\Omega (\nabla_{\mathcal{D}} w_{\mathcal{D}}(\boldsymbol{x}) - \nabla \overline{u}(\boldsymbol{x})) \cdot (\nabla_{\mathcal{D}} w_{\mathcal{D}}(\boldsymbol{x}) - \nabla_{\mathcal{D}} u_{\mathcal{D}}(\boldsymbol{x})) \mathrm{d}\boldsymbol{x} \\ + \|\nabla_{\mathcal{D}}(w_{\mathcal{D}} - u_{\mathcal{D}})\|_{L^2(\Omega)^d} W_{\mathcal{D}}(\nabla \overline{u}). \end{aligned}$$

Using the Cauchy–Schwarz inequality on the first term in the right-hand side, we infer

$$\|\nabla_{\mathcal{D}} w_{\mathcal{D}} - \nabla_{\mathcal{D}} u_{\mathcal{D}}\|_{L^2(\Omega)^d} \leq \|\nabla_{\mathcal{D}} w_{\mathcal{D}} - \nabla \overline{u}\|_{L^2(\Omega)^d} + W_{\mathcal{D}}(\nabla \overline{u}). \quad (1.25)$$

Define the "best interpolation error" (in the spirit of (1.12)) by

$$S_{\mathcal{D}}(\overline{u}) := \min_{w_{\mathcal{D}} \in X_{\mathcal{D},0}} \left(\|\Pi_{\mathcal{D}} w_{\mathcal{D}} - \overline{u}\|_{L^2(\Omega)} + \|\nabla_{\mathcal{D}} w_{\mathcal{D}} - \nabla \overline{u}\|_{L^2(\Omega)^d} \right)$$

and pick $w_{\mathcal{D}}$ that realises this minimum. Since

$$\begin{aligned} \|\nabla_{\mathcal{D}} u_{\mathcal{D}} - \nabla \overline{u}\|_{L^2(\Omega)^d} &\leq \|\nabla_{\mathcal{D}} u_{\mathcal{D}} - \nabla_{\mathcal{D}} w_{\mathcal{D}}\|_{L^2(\Omega)^d} + \|\nabla_{\mathcal{D}} w_{\mathcal{D}} - \nabla \overline{u}\|_{L^2(\Omega)^d} \\ &\leq \|\nabla_{\mathcal{D}} u_{\mathcal{D}} - \nabla_{\mathcal{D}} w_{\mathcal{D}}\|_{L^2(\Omega)^d} + S_{\mathcal{D}}(\overline{u}), \end{aligned}$$

Equation (1.25) gives

$$\|\nabla_{\mathcal{D}} u_{\mathcal{D}} - \nabla \overline{u}\|_{L^2(\Omega)^d} \leq 2 S_{\mathcal{D}}(\overline{u}) + W_{\mathcal{D}}(\nabla \overline{u}). \quad (1.26)$$

The question is now to check how $\Pi_{\mathcal{D}} u_{\mathcal{D}}$ approximates $\overline{u}$. Assume the following discrete Poincaré inequality:

There exists $C_{\mathcal{D}} > 0$ such that, $\forall v_{\mathcal{D}} \in X_{\mathcal{D},0}$,
$\|\Pi_{\mathcal{D}} v_{\mathcal{D}}\|_{L^2(\Omega)} \leq C_{\mathcal{D}} \|\nabla_{\mathcal{D}} v_{\mathcal{D}}\|_{L^2(\Omega)^d}$.

Then, with the same $w_{\mathcal{D}}$ selected above,

$$\begin{aligned}\|\Pi_{\mathcal{D}} u_{\mathcal{D}} - \overline{u}\|_{L^2(\Omega)} &\leq \|\Pi_{\mathcal{D}} u_{\mathcal{D}} - \Pi_{\mathcal{D}} w_{\mathcal{D}}\|_{L^2(\Omega)} + \|\Pi_{\mathcal{D}} w_{\mathcal{D}} - \overline{u}\|_{L^2(\Omega)} \\ &\leq C_{\mathcal{D}} \|\nabla_{\mathcal{D}} u_{\mathcal{D}} - \nabla_{\mathcal{D}} w_{\mathcal{D}}\|_{L^2(\Omega)^d} + S_{\mathcal{D}}(\overline{u}).\end{aligned}$$

Estimate (1.25) then shows that

$$\|\Pi_{\mathcal{D}} u_{\mathcal{D}} - \overline{u}\|_{L^2(\Omega)} \leq (C_{\mathcal{D}} + 1) S_{\mathcal{D}}(\overline{u}) + C_{\mathcal{D}} W_{\mathcal{D}}(\nabla \overline{u}). \tag{1.27}$$

Equations (1.26) and (1.27) are error estimates between $\overline{u}$ and $\Pi_{\mathcal{D}} u_{\mathcal{D}}$ and between $\nabla \overline{u}$ and $\nabla_{\mathcal{D}} u_{\mathcal{D}}$.
In particular, if a sequence $(X_{\mathcal{D}_m,0}, \Pi_{\mathcal{D}_m}, \nabla_{\mathcal{D}_m})_{m\in\mathbb{N}}$ is selected such that

$$\begin{array}{ll} \text{(P1)} & (C_{\mathcal{D}_m})_{m\in\mathbb{N}} \text{ is bounded,} \\ \text{(P2)} & S_{\mathcal{D}_m}(\overline{u}) \to 0 \text{ as } m \to \infty, \\ \text{(P3)} & W_{\mathcal{D}_m}(\nabla \overline{u}) \to 0 \text{ as } m \to \infty, \end{array} \tag{1.28}$$

then (1.26) and (1.27) show that, as $m \to \infty$, $\Pi_{\mathcal{D}_m} u_{\mathcal{D}_m} \to \overline{u}$ in $L^2(\Omega)$ and that $\nabla_{\mathcal{D}_m} u_{\mathcal{D}_m} \to \nabla \overline{u}$ in $L^2(\Omega)^d$.

Properties (P1-(P3) are thus the core properties that $(X_{\mathcal{D},0}, \Pi_{\mathcal{D}}, \nabla_{\mathcal{D}})$ must satisfy to provide a proper approximation of (1.1) under the form (1.22).

- Property (P1) is related to some *coercivity* property of this triplet, since this uniform Poincaré inequality is also what ensures an estimate of the form $\|\nabla_{\mathcal{D}} u_{\mathcal{D}}\|_{L^2(\Omega)} \leq C \|f\|_{L^2(\Omega)}$ if $u_{\mathcal{D}}$ is a solution to (1.22).
- Property (P2) states that $\Pi_{\mathcal{D}}$ and $\nabla_{\mathcal{D}}$ are *consistent* reconstructions of functions and their gradient; it enables the approximation of $\overline{u}$ and its gradient by using elements in $X_{\mathcal{D},0}$.
- As already discussed, $W_{\mathcal{D}}$ measures the error in the discrete Stokes formula and (P3) therefore relates to the *limit-conformity* of $(\Pi_{\mathcal{D}}, \nabla_{\mathcal{D}})$, stating that these two operators should, in the limit, satisfy the exact Stokes formula (as in the conforming case). Note that, in fact, the limit-conformity property (P3) implies the coercivity property (P1) (see e.g. Lemma 2.6 below).

1.3 Generalisation to non-linear problems

Non-linear equations are ubiquitous in real-world modelling, and a framework for the convergence analysis of numerical schemes should be able to handle such equations. Consider here the following example of a non-linear problem:

$$\begin{cases} \beta(\overline{u}) - \Delta \overline{u} = f & \text{in } \Omega, \\ \overline{u} = 0 & \text{on } \partial\Omega, \end{cases} \tag{1.29}$$

with the same notations as in Section 1.1, and where the function $\beta : \mathbb{R} \to \mathbb{R}$ is continuous, $s\beta(s) \geq 0$ and $|\beta(s)| \leq C(1 + |s|)$ for all $s \in \mathbb{R}$, where C does not depend on s. The weak formulation of (1.29) is:

$$\begin{aligned}&\text{Find } \overline{u} \in H_0^1(\Omega) \text{ such that, for all } v \in H_0^1(\Omega),\\ &\int_\Omega (\beta(\overline{u}(\boldsymbol{x}))v(\boldsymbol{x}) + \nabla\overline{u}(\boldsymbol{x}) \cdot \nabla v(\boldsymbol{x}))\mathrm{d}\boldsymbol{x} = \int_\Omega f(\boldsymbol{x})v(\boldsymbol{x})\mathrm{d}\boldsymbol{x}.\end{aligned} \tag{1.30}$$

It can be shown that there exists at least one solution to (1.30). Using, say, the conforming $\mathbb{P}_1$ finite element method and denoting by V_h the space of continuous piecewise linear functions on a triangular mesh of Ω, an approximation of (1.30) is

$$\begin{aligned}&\text{Find } u_h \in V_h \text{ such that, for all } v_h \in V_h,\\ &\int_\Omega (\beta(u_h(\boldsymbol{x}))v_h(\boldsymbol{x}) + \nabla u_h(\boldsymbol{x}) \cdot \nabla v_h(\boldsymbol{x}))\mathrm{d}\boldsymbol{x} = \int_\Omega f(\boldsymbol{x})v_h(\boldsymbol{x})\mathrm{d}\boldsymbol{x}.\end{aligned} \tag{1.31}$$

Although this approximate problem has at least one solution, its analysis presents three major difficulties. A first difficulty lies in the computation, using the expansion $u_h = \sum_{\boldsymbol{s}' \in \mathcal{V}_{\mathrm{int}}} u_{\boldsymbol{s}'}\varphi_{\boldsymbol{s}'}$, of the integral

$$\int_\Omega \beta\left(\sum_{\boldsymbol{s}' \in \mathcal{V}_{\mathrm{int}}} u_{\boldsymbol{s}'}\varphi_{\boldsymbol{s}'}(\boldsymbol{x})\right)\varphi_{\boldsymbol{s}}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \tag{1.32}$$

related to a given interior vertex $\boldsymbol{s}$ of the mesh; due to the non-linearity of β, the integrand may not be a piecewise polynomial and thus exact quadrature rules may not exist for this integral term. A second difficulty is to define an algorithm to approximate the solution of the non-linear system of equations (1.31). A third diffulty is to prove that the numerical method converges to the solution of the initial problem.

A classical answer to the first issue is to use the so-called "mass-lumping" method. This method consists in replacing, in (1.31) with $v_h = \varphi_{\boldsymbol{s}}$, the term (1.32) with $\omega_{\boldsymbol{s}}\beta(u_{\boldsymbol{s}})$ where $\omega_{\boldsymbol{s}}$ is some weight to be defined. The GDM framework provides a natural way of analysing the stability and convergence of this mass-lumped scheme, with weights defined as the measure of some "dual cells" denoted by $K_{\boldsymbol{s}}$ (see Figure 1.3). The set of discrete unknowns $X_{\mathcal{D},0}$ is, as for the conforming $\mathbb{P}_1$ method, the space of vectors with one component per interior vertex of the mesh. For $u \in X_{\mathcal{D},0}$, the reconstructed function $\Pi_{\mathcal{D}}u$ and gradient $\nabla_{\mathcal{D}}u$ are defined by

$$\begin{aligned}\Pi_{\mathcal{D}}u &= \sum_{\boldsymbol{s} \in \mathcal{V}_{\mathrm{int}}} u_{\boldsymbol{s}}\mathbf{1}_{K_{\boldsymbol{s}}} \quad &&\text{(piecewise constant reconstruction)},\\ \nabla_{\mathcal{D}}u &= \sum_{\boldsymbol{s} \in \mathcal{V}_{\mathrm{int}}} u_{\boldsymbol{s}}\nabla\varphi_{\boldsymbol{s}} \quad &&\text{(as for conforming } \mathbb{P}_1 \text{ finite elements)},\end{aligned} \tag{1.33}$$

where $\mathbf{1}_{K_{\boldsymbol{s}}}$ is the characteristic function of $K_{\boldsymbol{s}}$. Then the scheme (1.31) is replaced with

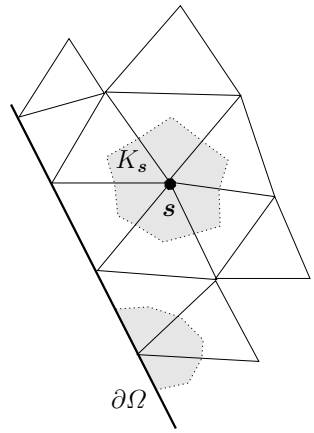


Fig. 1.3. Definition of $K_{\boldsymbol{s}}$

$$\begin{aligned}&\text{Find } u \in X_{\mathcal{D},0} \text{ such that, for all } v \in X_{\mathcal{D},0},\\ &\int_\Omega (\beta(\Pi_{\mathcal{D}} u(\boldsymbol{x}))\Pi_{\mathcal{D}} v(\boldsymbol{x}) + \nabla_{\mathcal{D}} u(\boldsymbol{x}) \cdot \nabla_{\mathcal{D}} v(\boldsymbol{x}))\mathrm{d}\boldsymbol{x}\\ &\qquad = \int_\Omega f(\boldsymbol{x})\Pi_{\mathcal{D}} v(\boldsymbol{x})\mathrm{d}\boldsymbol{x}.\end{aligned} \tag{1.34}$$

Since $\Pi_{\mathcal{D}} u$ and $\Pi_{\mathcal{D}} v$ are piecewise constant, all integral terms here are very easy to compute, which facilitates the implementation of the scheme. Another major interest of dealing with these piecewise constant reconstructions $\Pi_{\mathcal{D}}$ is that they satisfy

$$\beta(\Pi_{\mathcal{D}} u) = \Pi_{\mathcal{D}}(\beta(u)) = \sum_{\boldsymbol{s}\in\mathcal{V}_{\mathrm{int}}} \beta(u_{\boldsymbol{s}})\mathbf{1}_{K_{\boldsymbol{s}}}.$$

This property is sometimes crucial to obtain *a priori* estimates (see Section 6.3).

If $(X_{\mathcal{D}_m,0}, \Pi_{\mathcal{D}_m}, \nabla_{\mathcal{D}_m})_{m\in\mathbb{N}}$ is a sequence of spaces and operators associated as above with mass-lumped $\mathbb{P}_1$ finite elements on regular meshes whose size tends to zero, one can show that properties (P1), (P2) and (P3) hold. Then, only using these properties, it is possible to prove that:

1. The scheme (1.34) has at least one solution, denoted by $u_m \in X_{\mathcal{D}_m,0}$,
2. Up to a subsequence, as $m \to \infty$, $\Pi_{\mathcal{D}_m} u_m$ converges weakly in $L^2(\Omega)$ to some function $\overline{u} \in H_0^1(\Omega)$, $\nabla_{\mathcal{D}_m} u_m$ converges weakly in $L^2(\Omega)^d$ to $\nabla\overline{u}$, and $\beta(\Pi_{\mathcal{D}_m} u_m)$ converges weakly in $L^2(\Omega)$ to some function $\overline{\beta}$.

It is however not possible in general to deduce from Properties (P1)–(P3) that $\overline{\beta} = \beta(\overline{u})$. An additional *compactness* property of the sequence $(X_{\mathcal{D}_m,0}, \Pi_{\mathcal{D}_m}, \nabla_{\mathcal{D}_m})_{m\in\mathbb{N}}$ is required to establish this equality, and thus complete the convergence analysis of the scheme. This property reads

(P4) for any $u_m \in X_{\mathcal{D}_m,0}$ such that $(\nabla_{\mathcal{D}_m} u_m)_{m\in\mathbb{N}}$ is bounded in $L^2(\Omega)^d$, $(\Pi_{\mathcal{D}_m} u_m)_{m\in\mathbb{N}}$ is relatively compact in $L^2(\Omega)$.

(1.35)

This property can be established for spaces and operators coming from the mass-lumped $\mathbb{P}_1$ finite element method.

The discrete elements $(X_{\mathcal{D},0}, \Pi_{\mathcal{D}}, \nabla_{\mathcal{D}})$ and Properties (P1), (P2), (P3), (P4) are at the core of the definition and properties of the *gradient discretisation method* (GDM) (with (P3) in fact implying (P1)). A piecewise constant reconstruction $\Pi_{\mathcal{D}}$, as in (1.33), is a fifth property that is instrumental for some non-linear problems.

2

Dirichlet boundary conditions

The gradient discretisation method (GDM) is a design and analysis framework for numerical schemes for elliptic and parabolic partial differential equations. As suggested by its name, the GDM relies on a gradient discretisation (GD), denoted by $\mathcal{D}$, which contains at least the three following discrete entities:

- a **discrete space of unknowns** $X_{\mathcal{D}}$, which is a finite dimensional space of discrete unknowns – *e.g.*, the values at the nodes of a mesh (as in the conforming $\mathbb{P}_1$ finite element method), at particular point in the mesh cells (as in the TPFA-CG scheme), or at particular points on the mesh faces (as in the non-conforming $\mathbb{P}_1$ finite element method),
- a **function reconstruction** operator $\Pi_{\mathcal{D}}$, which creates from an element of $X_{\mathcal{D}}$ a function defined a.e. on the physical domain Ω.
- a **gradient reconstruction** operator $\nabla_{\mathcal{D}}$, which builds a "discrete gradient" (vector-valued function) defined a.e. on Ω from the discrete unknowns.

The idea of the GDM is to construct a scheme by replacing, in the weak formulation of the problem to be solved, the continuous space and operators by discrete ones coming from a GD. The scheme thus obtained is called a gradient scheme (GS).

The convergence analysis of the GDM depends, of course, on the nature of the PDE to be solved. The definition of the GD, on the other hand, depends to a large extent only on the boundary conditions (but a common abstract framework can be designed to cover various boundary conditions (BC), see Appendix A). The present chapter deals with Dirichlet boundary conditions, and is split in two sections. Section 2.1 covers homogeneous Dirichlet boundary conditions, and Section 2.2 considers non-homogeneous Dirichlet boundary conditions. In each section, the concept of gradient discretisation is defined, along with a list of the properties of the spaces and mappings that are important for the convergence analysis of the GDM. The corresponding GSs for linear and some non-linear elliptic PDEs are then described, and their

convergence analysis is performed. Error estimates are provided in the linear case. In the non-linear case, the convergence is proved thanks to compactness arguments.
The GDM for boundary conditions other than Dirichlet BCs is detailed in Chapter 3, and Part II analyses the GDM for linear, non-linear and degenerate parabolic problems.
In this chapter, we let $p \in (1, +\infty)$ be given.

2.1 Homogeneous Dirichlet boundary conditions

This section is devoted to the notion and the various required properties of a gradient discretisation for homogeneous Dirichlet boundary conditions. Then, the corresponding gradient schemes for linear and non-linear elliptic PDEs are presented and their convergence is analysed.

2.1.1 Gradient discretisations

Definition 2.1 (GD, homogeneous Dirichlet BCs). *A gradient discretisation $\mathcal{D}$ for homogeneous Dirichlet conditions is defined by $\mathcal{D} = (X_{\mathcal{D},0}, \Pi_{\mathcal{D}}, \nabla_{\mathcal{D}})$, where:*

1. *the set of discrete unknowns $X_{\mathcal{D},0}$ is a finite dimensional real vector space,*
2. *the function reconstruction $\Pi_{\mathcal{D}} : X_{\mathcal{D},0} \to L^p(\Omega)$ is a linear mapping that reconstructs, from an element of $X_{\mathcal{D},0}$, a function over Ω,*
3. *the gradient reconstruction $\nabla_{\mathcal{D}} : X_{\mathcal{D},0} \to L^p(\Omega)^d$ is a linear mapping which reconstructs, from an element of $X_{\mathcal{D},0}$, a "gradient" (vector-valued function) over Ω,*
4. *the gradient reconstruction is such that $\|\cdot\|_{\mathcal{D}} := \|\nabla_{\mathcal{D}} \cdot\|_{L^p(\Omega)^d}$ is a norm on $X_{\mathcal{D},0}$.*

The following sections present gradient schemes for several problems, starting from such a gradient discretisation. In order to show the convergence of the scheme, we use some properties of consistency and stability. As in the finite element method framework, stability is obtained through some uniform coercivity of the discrete operator which relies on a discrete Poincaré inequality.

Definition 2.2 (Coercivity, Dirichlet BCs)

If $\mathcal{D}$ is a gradient discretisation in the sense of Definition 2.1, define $C_{\mathcal{D}}$ as the norm of the linear mapping $\Pi_{\mathcal{D}}$:

$$C_{\mathcal{D}} = \max_{v \in X_{\mathcal{D},0} \setminus \{0\}} \frac{\|\Pi_{\mathcal{D}} v\|_{L^p(\Omega)}}{\|v\|_{\mathcal{D}}}. \tag{2.1}$$

A sequence $(\mathcal{D}_m)_{m\in\mathbb{N}}$ of gradient discretisations in the sense of Definition 2.1 is **coercive** if there exists $C_P \in \mathbb{R}_+$ such that $C_{\mathcal{D}_m} \le C_P$ for all $m \in \mathbb{N}$.

Remark 2.3 (Discrete Poincaré inequality). Equation (2.1) yields the discrete Poincaré inequality $\|\Pi_{\mathcal{D}} v\|_{L^p(\Omega)} \le C_{\mathcal{D}} \|\nabla_{\mathcal{D}} v\|_{L^p(\Omega)^d}$ for all $v \in X_{\mathcal{D},0}$.

The consistency properties that we need indicate how a regular function (and its gradient) are more or less well approximated by some function and gradient which are reconstructed from the space $X_{\mathcal{D},0}$. The function $S_{\mathcal{D}}$ which we introduce hereafter is often called "interpolation error" in the framework of finite elements.

Definition 2.4 (GD-consistency, homogeneous Dirichlet BCs)

If $\mathcal{D}$ is a gradient discretisation in the sense of Definition 2.1, define $S_{\mathcal{D}} : W_0^{1,p}(\Omega) \to [0, +\infty)$ by

$$\begin{aligned} &\forall \varphi \in W_0^{1,p}(\Omega)\,, \\ &S_{\mathcal{D}}(\varphi) = \min_{v\in X_{\mathcal{D},0}} \left(\|\Pi_{\mathcal{D}} v - \varphi\|_{L^p(\Omega)} + \|\nabla_{\mathcal{D}} v - \nabla\varphi\|_{L^p(\Omega)^d} \right). \end{aligned} \tag{2.2}$$

A sequence $(\mathcal{D}_m)_{m\in\mathbb{N}}$ of gradient discretisations in the sense of Definition 2.1 is **GD-consistent**, or consistent for short, if

$$\forall \varphi \in W_0^{1,p}(\Omega), \lim_{m\to\infty} S_{\mathcal{D}_m}(\varphi) = 0. \tag{2.3}$$

Note that the definition (2.2) of $S_{\mathcal{D}}(\varphi)$ makes sense; indeed, since the $L^p(\Omega)$ and $L^p(\Omega)^d$ norms are strictly convex, for each $\varphi \in W_0^{1,p}(\Omega)$ there is a unique $I_{\mathcal{D}}\varphi \in X_{\mathcal{D},0}$ realizing the minimum in $S_{\mathcal{D}}(\varphi)$, that is, such that

$$S_{\mathcal{D}}(\varphi) = \|\Pi_{\mathcal{D}} I_{\mathcal{D}}\varphi - \varphi\|_{L^p(\Omega)} + \|\nabla_{\mathcal{D}} I_{\mathcal{D}}\varphi - \nabla\varphi\|_{L^p(\Omega)^d}\,.$$

Hence the interpolant $I_{\mathcal{D}}\varphi$ can be defined as

$$I_{\mathcal{D}}\varphi = \operatorname*{argmin}_{v\in X_{\mathcal{D},0}} \left(\|\Pi_{\mathcal{D}} v - \varphi\|_{L^p(\Omega)} + \|\nabla_{\mathcal{D}} v - \nabla\varphi\|_{L^p(\Omega)^d} \right).$$

Note that $I_{\mathcal{D}}$, even though uniquely defined, is not necessarily a linear map. In the case $p = 2$, a *linear* interpolant $I_{\mathcal{D}}^{(2)} : W_0^{1,2}(\Omega)(= H_0^1(\Omega)) \to X_{\mathcal{D},0}$ can be defined by setting

$$I_{\mathcal{D}}^{(2)}\varphi = \operatorname*{argmin}_{v\in X_{\mathcal{D},0}} \left(\|\Pi_{\mathcal{D}} v - \varphi\|^2_{L^2(\Omega)} + \|\nabla_{\mathcal{D}} v - \nabla\varphi\|^2_{L^2(\Omega)^d} \right)^{1/2}.$$

This interpolant will be used to establish error estimates for linear parabolic equations: see the proof of Theorem 5.3, which also contains a proof of the linearity of $I_{\mathcal{D}}^{(2)}$ and of its approximation properties.
The next important notion of the GDM framework is the limit-conformity of the gradient and divergence operators. A well-known property of the gradient operator in H_0^1 is the so-called grad-div duality; the Stokes formula gives:

$$\int_\Omega (\nabla u \cdot \boldsymbol{\varphi} + u \ \mathrm{div}\boldsymbol{\varphi})\mathrm{d}\boldsymbol{x} = 0, \forall u \in H_0^1(\Omega), \forall \boldsymbol{\varphi} \in H_{\mathrm{div}}(\Omega). \tag{2.4}$$

where $H_{\mathrm{div}}(\Omega) = \{\boldsymbol{\varphi} \in L^2(\Omega)^d \ : \ \mathrm{div}\boldsymbol{\varphi} \in L^2(\Omega)\}$. The Stokes formula is still valid at the discrete level when using a conforming method such as the linear $\mathbb{P}_1$ finite element. However, when dealing with non-conforming methods, this property is no longer exact at the discrete level. The concept of limit-conformity states that the discrete gradient and divergence operator satisfy this property asymptotically. Since non-linear problems are also considered in this book, adequate functional spaces need to be introduced. For any $q \in (1, +\infty)$, the space $W_{\mathrm{div}}^q(\Omega)$ of functions in $(L^q(\Omega))^d$ with divergence in $L^q(\Omega)$:

$$W_{\mathrm{div}}^q(\Omega) = \{\boldsymbol{\varphi} \in L^q(\Omega)^d \ : \ \mathrm{div}\boldsymbol{\varphi} \in L^q(\Omega)\}. \tag{2.5}$$

We recall that the space $W_0^{1,2}(\Omega)$ is commonly denoted by $H_0^1(\Omega)$ and that $W_{\mathrm{div}}^2(\Omega) = H_{\mathrm{div}}(\Omega)$ (see notations for Sobolev spaces in Section D.1.2).

Definition 2.5 (Limit-conformity, Dirichlet BCs)

If $\mathcal{D}$ is a gradient discretisation in the sense of Definition 2.1, let $p' = \frac{p}{p-1}$ and define $W_{\mathcal{D}}$: $W_{\mathrm{div}}^{p'}(\Omega) \to [0, +\infty)$ by

$$\forall \boldsymbol{\varphi} \in W_{\mathrm{div}}^{p'}(\Omega)\,,$$
$$W_{\mathcal{D}}(\boldsymbol{\varphi}) = \sup_{u \in X_{\mathcal{D},0} \setminus \{0\}} \frac{\left| \int_\Omega (\nabla_{\mathcal{D}} u(\boldsymbol{x}) \cdot \boldsymbol{\varphi}(\boldsymbol{x}) + \Pi_{\mathcal{D}} u(\boldsymbol{x}) \mathrm{div}\boldsymbol{\varphi}(\boldsymbol{x}))\, \mathrm{d}\boldsymbol{x} \right|}{\|u\|_{\mathcal{D}}}. \tag{2.6}$$

A sequence $(\mathcal{D}_m)_{m \in \mathbb{N}}$ of gradient discretisations is **limit-conforming** if

$$\forall \boldsymbol{\varphi} \in W_{\mathrm{div}}^{p'}(\Omega), \ \lim_{m \to \infty} W_{\mathcal{D}_m}(\boldsymbol{\varphi}) = 0. \tag{2.7}$$

It is clear from its definition that the quantity $W_{\mathcal{D}}$ measures how well the reconstructed function and gradient satisfy the divergence (Stokes) formula (2.4). If the method is *conforming* in the sense that $\Pi_{\mathcal{D}}(X_{\mathcal{D},0}) \subset W_0^{1,p}(\Omega)$ and $\nabla_{\mathcal{D}} u = \nabla(\Pi_{\mathcal{D}} u)$ for all $u \in X_{\mathcal{D},0}$, then $W_{\mathcal{D}} \equiv 0$. In general, $W_{\mathcal{D}}$ measures the defect of conformity of the method, and must vanish in the limit – hence the name "limit-conformity" for the above property.

The following lemma shows that the coercivity is actually a consequence of the limit-conformity.

Lemma 2.6 (Limit-conformity implies coercivity, Dirichlet BCs). *Any sequence of gradient discretisations that is limit-conforming in the sense of Definition 2.5 is also coercive in the sense of Definition 2.2.*

Proof. Let $(\mathcal{D}_m)_{m\in\mathbb{N}}$ be a limit-conforming sequence of GDs, and define

$$E = \left\{ \frac{\Pi_{\mathcal{D}_m} v}{\|v\|_{\mathcal{D}_m}} \in L^p(\Omega) \,:\, m \in \mathbb{N}\,,\ v \in X_{\mathcal{D}_m,0}\backslash\{0\} \right\}.$$

Proving the coercivity of $(\mathcal{D}_m)_{m\in\mathbb{N}}$ consists in proving that E is bounded in $L^p(\Omega)$. Let $\ell \in (L^p(\Omega))'$. There is $f \in L^{p'}(\Omega)$ such that, for all $w \in L^p(\Omega)$, $\ell(w) = \int_\Omega f(\boldsymbol{x})w(\boldsymbol{x})\mathrm{d}\boldsymbol{x}$. Let $\boldsymbol{\varphi} \in W^{p'}_{\mathrm{div}}(\Omega)$ be such that $\mathrm{div}\boldsymbol{\varphi} = f$ (for example, take $\boldsymbol{\varphi} = -|\nabla\overline{u}|^{p-2}\nabla\overline{u}$ where $\overline{u}$ is the solution in $W^{1,p}_0(\Omega)$ of $-\mathrm{div}(|\nabla\overline{u}|^{p-2}\nabla\overline{u}) = f$). For $z \in E$, take $m \in \mathbb{N}$ and $v \in X_{\mathcal{D}_m,0}\backslash\{0\}$ such that $z = \frac{\Pi_{\mathcal{D}_m} v}{\|v\|_{\mathcal{D}_m}}$ and write

$$\begin{aligned}
|\ell(z)| &= \frac{1}{\|v\|_{\mathcal{D}_m}} \left| \int_\Omega \Pi_{\mathcal{D}_m} v(\boldsymbol{x}) f(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \right| \\
&\le \frac{1}{\|v\|_{\mathcal{D}_m}} \left| \int_\Omega \left(\Pi_{\mathcal{D}_m} v(\boldsymbol{x})\mathrm{div}\boldsymbol{\varphi}(\boldsymbol{x}) + \nabla_{\mathcal{D}_m} v(\boldsymbol{x}) \cdot \boldsymbol{\varphi}(\boldsymbol{x})\right) \mathrm{d}\boldsymbol{x} \right| \\
&\quad + \frac{1}{\|\nabla_{\mathcal{D}_m} v\|_{L^p(\Omega)^d}} \left| \int_\Omega \nabla_{\mathcal{D}_m} v(\boldsymbol{x}) \cdot \boldsymbol{\varphi}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \right| \\
&\le W_{\mathcal{D}_m}(\boldsymbol{\varphi}) + \|\boldsymbol{\varphi}\|_{L^{p'}(\Omega)^d}\,.
\end{aligned} \tag{2.8}$$

In the last line, Hölder's inequality (D.5) was used. Since $(\mathcal{D}_m)_{m\in\mathbb{N}}$ is limit-conforming, $(W_{\mathcal{D}_m}(\boldsymbol{\varphi}))_{m\in\mathbb{N}}$ converges to 0, and is therefore bounded. Estimate (2.8) thus shows that $\{\ell(z) \,:\, z \in E\}$ is bounded by some constant depending on ℓ. Since this is valid for any $\ell \in (L^p(\Omega))'$, the Banach–Steinhaus theorem [33, Theorem 2.2] (sometimes called "Uniform Boundedness Principle") enables us to conclude that E is bounded in $L^p(\Omega)$. ∎

The following equivalent condition for the limit-conformity property facilitates the proof of the regularity of a possible limit (Lemma 2.15 below).

Lemma 2.7 (On limit-conformity, Dirichlet BCs). *Let $\mathcal{D}$ be a gradient discretisation in the sense of Definition 2.1. Set $p' = \frac{p}{p-1}$ and define $\widetilde{W}_{\mathcal{D}}$: $W^{p'}_{\mathrm{div}}(\Omega) \times X_{\mathcal{D},0} \to [0,+\infty)$ by*

$$\begin{aligned}
&\forall(\boldsymbol{\varphi}, u) \in W^{p'}_{\mathrm{div}}(\Omega) \times X_{\mathcal{D},0}\,, \\
&\widetilde{W}_{\mathcal{D}}(\boldsymbol{\varphi}, u) = \int_\Omega \left(\nabla_{\mathcal{D}} u(\boldsymbol{x}) \cdot \boldsymbol{\varphi}(\boldsymbol{x}) + \Pi_{\mathcal{D}} u(\boldsymbol{x})\mathrm{div}\boldsymbol{\varphi}(\boldsymbol{x})\right) \mathrm{d}\boldsymbol{x}.
\end{aligned} \tag{2.9}$$

A sequence $(\mathcal{D}_m)_{m\in\mathbb{N}}$ of gradient discretisations in the sense of Definition 2.1 is limit-conforming if and only if, for any sequence $u_m \in X_{\mathcal{D}_m,0}$ such that $(\|u_m\|_{\mathcal{D}_m})_{m\in\mathbb{N}}$ is bounded,

$$\forall \boldsymbol{\varphi} \in W^{p'}_{\mathrm{div}}(\Omega),\ \lim_{m\to\infty} \widetilde{W}_{\mathcal{D}_m}(\boldsymbol{\varphi}, u_m) = 0. \tag{2.10}$$

Proof. Let us remark that

$$W_{\mathcal{D}}(\boldsymbol{\varphi}) = \sup_{u\in X_{\mathcal{D},0}\setminus\{0\}} \frac{|\widetilde{W}_{\mathcal{D}}(\boldsymbol{\varphi}, u)|}{\|u\|_{\mathcal{D}}}.$$

The proof that (2.7) implies (2.10) is then straightforward, since $|\widetilde{W}_{\mathcal{D}}(\boldsymbol{\varphi}, u)| \le \|u\|_{\mathcal{D}} W_{\mathcal{D}}(\boldsymbol{\varphi})$. Let us prove the converse by way of contradiction. If (2.7) does not hold then there exists $\boldsymbol{\varphi} \in W^{p'}_{\mathrm{div}}(\Omega)$, $\varepsilon > 0$ and a subsequence of $(\mathcal{D}_m)_{m\in\mathbb{N}}$, still denoted by $(\mathcal{D}_m)_{m\in\mathbb{N}}$, such that $W_{\mathcal{D}_m}(\boldsymbol{\varphi}) \ge \varepsilon$ for all $m \in \mathbb{N}$. We can then find $u_m \in X_{\mathcal{D}_m,0} \setminus \{0\}$ such that

$$|\widetilde{W}_{\mathcal{D}}(\boldsymbol{\varphi}, u_m)| \ge \frac{1}{2}\varepsilon \|u_m\|_{\mathcal{D}_m}.$$

Considering the bounded sequence $(u_m/\|u_m\|_{\mathcal{D}_m})_{m\in\mathbb{N}}$, we get a contradiction with (2.10). ∎

Dealing with generic non-linearity often requires additional compactness properties on the scheme.

Definition 2.8 (Compactness, Dirichlet BCs)

A sequence $(\mathcal{D}_m)_{m\in\mathbb{N}}$ of gradient discretisations in the sense of Definition 2.1 is **compact** if, for any sequence $u_m \in X_{\mathcal{D}_m,0}$ such that $(\|u_m\|_{\mathcal{D}_m})_{m\in\mathbb{N}}$ is bounded, the sequence $(\Pi_{\mathcal{D}_m} u_m)_{m\in\mathbb{N}}$ is relatively compact in $L^p(\Omega)$.

Remark 2.9 (Compactly embedded sequence). Let $(\mathcal{D}_m)_{m\in\mathbb{N}}$ be a sequence of gradient discretisations in the sense of Definition 2.1, and define the space $X_m = \Pi_{\mathcal{D}_m}(X_{\mathcal{D}_m,0})$ with norm

$$\|w\|_{X_m} = \min\{\|u\|_{\mathcal{D}_m} \,:\, u \in X_{\mathcal{D}_m,0} \text{ such that } \Pi_{\mathcal{D}_m} u = w\}.$$

Then the sequence $(\mathcal{D}_m)_{m\in\mathbb{N}}$ is compact in the sense of Definition 2.8 if and only if the sequence $(X_m)_{m\in\mathbb{N}}$ is compactly embedded in $L^p(\Omega)$ in the sense of Definition C.4.

Compactness is stronger than coercivity, as stated in the following lemma; in fact, coercivity is required in linear problems, whereas compactness is not (see Corollary 2.31 and Remark 2.32).

Lemma 2.10 (Compactness implies coercivity, Dirichlet BCs). *Any sequence of gradient discretisations that is compact in the sense of Definition 2.8 is also coercive in the sense of Definition 2.2.*

Proof. Let $(\mathcal{D}_m)_{m\in\mathbb{N}}$ be a compact sequence of GDs, and assume that it is not coercive. Then there exists a subsequence of $(\mathcal{D}_m)_{m\in\mathbb{N}}$ (denoted in the same way) such that, for all $m \in \mathbb{N}$, there exists $u_m \in X_{\mathcal{D}_m,0} \setminus \{0\}$ with

$$\lim_{m\to\infty} \frac{\|\Pi_{\mathcal{D}_m} u_m\|_{L^p(\Omega)}}{\|u_m\|_{\mathcal{D}_m}} = +\infty.$$

Setting $v_m = u_m/\|u_m\|_{\mathcal{D}_m}$, this gives $\lim_{m\to\infty} \|\Pi_{\mathcal{D}_m} v_m\|_{L^p(\Omega)} = +\infty$. But $\|v_m\|_{\mathcal{D}_m} = 1$ and the compactness of the sequence of discretisations therefore implies that the sequence $(\Pi_{\mathcal{D}_m} v_m)_{m\in\mathbb{N}}$ is relatively compact in $L^p(\Omega)$. This gives a contradiction. ∎

Remark 2.11 (Existence of GD-consistent, limit-conforming and compact sequence of GDs). Part III provides examples of GD-consistent, limit-conforming and compact sequence of GDs. Simple ones are for instance the Galerkin gradient discretisations (see Section 8.1), which only use the existence of a countable dense family of elements in $W_0^{1,p}(\Omega)$.

Let us turn to a property that we shall often require on the function reconstruction $\Pi_{\mathcal{D}}$. Indeed, it is very often handy to obtain piecewise constant functions as approximate functions, the reason being that piecewise constant functions commute with any non-linearity. This is a key argument for non-linear degenerate parabolic problems.

Definition 2.12 (Piecewise constant reconstruction)

Let $\mathcal{D} = (X_{\mathcal{D},0}, \Pi_{\mathcal{D}}, \nabla_{\mathcal{D}})$ be a gradient discretisation in the sense of Definition 2.1. The operator $\Pi_{\mathcal{D}} : X_{\mathcal{D},0} \to L^p(\Omega)$ is a piecewise constant reconstruction if there exists a basis $(e_i)_{i\in B}$ of $X_{\mathcal{D},0}$ and a family of disjoint subsets $(\Omega_i)_{i\in B}$ of Ω such that $\Pi_{\mathcal{D}} u = \sum_{i\in B} u_i \mathbf{1}_{\Omega_i}$ for all $u = \sum_{i\in B} u_i e_i \in X_{\mathcal{D},0}$, where $\mathbf{1}_{\Omega_i}$ is the characteristic function of Ω_i. In other words, $\Pi_{\mathcal{D}} u$ is the piecewise constant function equal to u_i on Ω_i, for all $i \in B$.

The set B is usually the natural set of (geometrical entities attached to the) discrete unknowns of the scheme. Moreover, $\|\Pi_{\mathcal{D}} \cdot\|_{L^p(\Omega)}$ is not requested to be a norm on $X_{\mathcal{D},0}$. Indeed, all unknowns are involved in the definition of the reconstructed gradients, but in several examples some unknowns are not used to reconstruct the functions itself. Hence some of the subsets Ω_i may be empty, which prevents $\|\Pi_{\mathcal{D}} \cdot\|_{L^p(\Omega)}$ from being a norm.

Remark 2.13. If $\Pi_{\mathcal{D}}$ is a piecewise constant reconstruction and $g : \mathbb{R} \mapsto \mathbb{R}$ we have

$$g(\Pi_{\mathcal{D}} u(\boldsymbol{x})) = \Pi_{\mathcal{D}} g(u)(\boldsymbol{x}) \text{ for a.e. } \boldsymbol{x} \in \Omega\,,\ \forall u \in X_{\mathcal{D},0}$$

where, for $u = \sum_{i\in B} u_i e_i$, we set $g(u) = \sum_{i\in B} g(u)_i e_i \in X_{\mathcal{D},0}$ with $g(u)_i = g(u_i)$. We also have

$$\Pi_{\mathcal{D}} u(\boldsymbol{x}) \Pi_{\mathcal{D}} v(\boldsymbol{x}) = \Pi_{\mathcal{D}}(uv)(\boldsymbol{x}) \text{ for a.e. } \boldsymbol{x} \in \Omega\,,\ \forall u, v \in X_{\mathcal{D},0},$$

where $uv \in X_{\mathcal{D},0}$ is defined by $(uv)_i = u_i v_i$ for all $i \in B$.
Note that these definitions of $g(u)$ or uv depend on the choice of the discrete unknowns B in $X_{\mathcal{D},0}$. We should therefore denote $g^B(u)$ or $(uv)^B$ to emphasise the dependency on B but, in practice, we will remove this superscript B as the discrete unknowns are usually canonically chosen and fixed throughout the whole study of a gradient scheme.

Remark 2.14 (Independence of the notions of GD-consistency, limit-conformity and compactness)

Lemmata 2.6 and 2.10 show that the coercivity property is a consequence of the limit-conformity and of the compactness property. This raises the questions of further links between the properties of gradient discretisations. The following examples show that no such general link exists between the limit-conformity, the compactness and the GD-consistency.

- *Limit-conforming and compact, but not GD-consistent.* Let E be a finite-dimensional subspace of $W_0^{1,p}(\Omega)$, and consider the sequence $(\mathcal{D}_m)_{m\in\mathbb{N}}$ of GDs defined, for all $m \in \mathbb{N}$, by: $X_{\mathcal{D}_m,0} = E$, $\Pi_{\mathcal{D}_m} u = u$ and $\nabla_{\mathcal{D}_m} u = \nabla u$. Functions in E satisfy the Stokes formula, so $W_{\mathcal{D}_m}(\boldsymbol{\varphi}) = 0$ for all m. Since E is finite dimensional, the compactness of $(\mathcal{D}_m)_{m\in\mathbb{N}}$ is trivial. However, no function outside E can be approximated by elements of $X_{\mathcal{D}_m,0}$, so $\lim_{m\to\infty} S_{\mathcal{D}_m}(\varphi) \neq 0$ if $\varphi \notin E$ and $(\mathcal{D}_m)_{m\in\mathbb{N}}$ is not consistent.
- *GD-consistent and limit-conforming, but not compact.* Consider $m \in \mathbb{N}^\star$, $h = 1/(2m)$, and the $\mathbb{P}_1$ finite element basis $(\varphi_i)_{i=1,\dots,2m-1}$ associated with the nodes $(ih)_{i=1,\dots,2m-1}$. Let $X_{\mathcal{D}_m,0} = \{u = (u_i)_{i=1,\dots,2m-1}\}$, $\widehat{u} = \sum_{i=1}^{2m-1} u_i \varphi_i$, $\nabla_{\mathcal{D}_m} u = \widehat{u}'$, and $\Pi_{\mathcal{D}_m} u(x) = \widehat{u}(x) + \widehat{u}'(x) - \widehat{u}'(x + \varepsilon(x))$, with

$$\forall k = 0, \dots, m-1\,, \begin{cases} \varepsilon(x) = h & \text{if } x \in (2kh, (2k+1)h), \\ \varepsilon(x) = -h & \text{if } x \in ((2k+1)h, (2k+2)h). \end{cases}$$

  To see that $(\mathcal{D}_m)_{m\in\mathbb{N}}$ is limit-conforming, write, for $\psi \in W^{p'}_{\mathrm{div}}(\Omega)$,

$$\int_0^1 \Pi_{\mathcal{D}_m} u(x)\psi'(x)\mathrm{d}\boldsymbol{x} = \int_0^1 \widehat{u}(x)\psi'(x)\mathrm{d}\boldsymbol{x} + \int_0^1 \widehat{u}'(x)(\psi'(x) - \psi'(x + \varepsilon(x)))\mathrm{d}\boldsymbol{x},$$

  and notice that the second term tends to 0 by continuity in means of ψ' in $L^{p'}((0,1))$. The standard interpolation of a regular function φ, defined by $u_i = \varphi(ih)$, gives an element of $W_0^{1,p}((0,1))$ that converges in this space to φ; this shows the GD-consistency. Considering the sequence $(u^m)_{m\in\mathbb{N}}$ defined by $u^m_{2k} = 0$ and $u^m_{2k+1} = h$ for all $k = 0, \dots, m-1$, we obtain a sequence such that $\|u^m\|_{\mathcal{D}_m}$ is constant, but $\Pi_{\mathcal{D}_m} u^m - \widehat{u}^m$ oscillates between 2 and -2 with $\widehat{u}^m$ uniformly converging to 0, showing that $(\mathcal{D}_m)_{m\in\mathbb{N}}$ is not compact.

- *GD-consistent and compact, but not limit-conforming.* Consider $m \in \mathbb{N}^\star$, $h = 1/m$, and the $\mathbb{P}_1$ finite element basis $(\varphi_i)_{i=1,\dots,m}$ associated with the nodes $(ih)_{i=1,\dots,m}$. Set $X_{\mathcal{D}_m,0} = \{u = (u_i)_{i=1,\dots,m}\}$, $\Pi_{\mathcal{D}_m} u = \sum_{i=1}^m u_i\varphi_i$ and $\nabla_{\mathcal{D}_m} u = (\Pi_{\mathcal{D}_m} u)'$. Note that only the left Dirichlet boundary condition is satisfied. The sequence $(\mathcal{D}_m)_{m\in\mathbb{N}}$ is GD-consistent and compact, but not limit-conforming in $W_0^{1,p}((0,1))$, since $\int_0^1 (\Pi_{\mathcal{D}} u(x)\varphi'(x) + \nabla_{\mathcal{D}} u(x)\varphi(x))\mathrm{d}x = \varphi(1)u_N$.

The convergence analysis of sequences of approximate solutions to a partial differential equation (PDE) usually starts by finding *a priori* estimates on the solutions to the schemes. In the GDM framework, this means proving that $\|u_{\mathcal{D}}\|_{\mathcal{D}}$ remains bounded. Lemma 2.15 below states that, if such a bound holds, we can find a weak limit to the reconstructed functions and their gradients. Combined if necessary with the compactness of the gradient discretisations, this opens the way to the last stage of the convergence analysis, which consists in showing that this limit is a solution to the PDE.

Lemma 2.15 (Regularity of the limit, homogeneous Dirichlet BCs). *Let $(\mathcal{D}_m)_{m\in\mathbb{N}}$ be a limit-conforming sequence of gradient discretisations, in the sense of Definition 2.5. Let $u_m \in X_{\mathcal{D}_m,0}$ be such that $(\|u_m\|_{\mathcal{D}_m})_{m\in\mathbb{N}}$ is bounded. Then there exist a subsequence of $(\mathcal{D}_m, u_m)_{m\in\mathbb{N}}$, denoted in the same way, and $u \in W_0^{1,p}(\Omega)$ such that $\Pi_{\mathcal{D}_m} u_m$ converges weakly in $L^p(\Omega)$ to u and $\nabla_{\mathcal{D}_m} u_m$ converges weakly in $L^p(\Omega)^d$ to ∇u.*

Proof. Owing to Lemma 2.6, $(\mathcal{D}_m)_{m\in\mathbb{N}}$ is coercive and thus the sequence $(\Pi_{\mathcal{D}_m} u_m)_{m\in\mathbb{N}}$ is bounded in $L^p(\Omega)$. Therefore, there exists a subsequence of $(\mathcal{D}_m, u_m)_{m\in\mathbb{N}}$, denoted in the same way, and there exist $u \in L^p(\Omega)$ and $\boldsymbol{v} \in L^p(\Omega)^d$ such that $\Pi_{\mathcal{D}_m} u_m$ converges weakly in $L^p(\Omega)$ to u and $\nabla_{\mathcal{D}_m} u_m$ converges weakly in $L^p(\Omega)^d$ to $\boldsymbol{v}$. Extend $\Pi_{\mathcal{D}_m} u_m$, u, $\nabla_{\mathcal{D}_m} u_m$ and $\boldsymbol{v}$ by 0 outside Ω; the previous convergence results hold respectively in $L^p(\mathbb{R}^d)$ and $L^p(\mathbb{R}^d)^d$. Using the limit-conformity of $(\mathcal{D}_m)_{m\in\mathbb{N}}$ and the bound on $\|u_m\|_{\mathcal{D}_m}$, passing to the limit in (2.10) gives

$$\forall \boldsymbol{\varphi} \in W_{\mathrm{div}}^{p'}(\mathbb{R}^d)\,,\ \int_{\mathbb{R}^d} (\boldsymbol{v}(\boldsymbol{x})\cdot\boldsymbol{\varphi}(\boldsymbol{x}) + u(\boldsymbol{x})\mathrm{div}\boldsymbol{\varphi}(\boldsymbol{x}))\,\mathrm{d}\boldsymbol{x} = 0.$$

Being valid for any $\boldsymbol{\varphi} \in C_c^\infty(\mathbb{R}^d)^d$, this relation proves both that $\boldsymbol{v} = \nabla u$ and that $u \in W_0^{1,p}(\Omega)$. ∎

Let us now present some equivalent or sufficient conditions for the GD-consistency, limit-conformity and compactness of a sequence of gradient discretisations.

Lemma 2.16 (Equivalent condition for GD-consistency, homogeneous Dirichlet BCs). *A sequence $(\mathcal{D}_m)_{m\in\mathbb{N}}$ of gradient discretisations is GD-consistent in the sense of Definition 2.4 if and only if there exists a dense subset W_s in $W_0^{1,p}(\Omega)$ such that*

$$\forall \psi \in W_s\,,\ \lim_{m\to\infty} S_{\mathcal{D}_m}(\psi) = 0. \tag{2.11}$$

Proof. Let us assume that (2.11) holds and let us prove (2.3) (the converse is straightforward, take $W_s = W_0^{1,p}(\Omega)$). Let $\varphi \in W_0^{1,p}(\Omega)$ and $\varepsilon > 0$. Take $\psi \in W_s$ such that $\|\varphi - \psi\|_{W_0^{1,p}(\Omega)} \le \varepsilon$. For $v \in X_{\mathcal{D},0}$, the triangle inequality yields

$$\begin{aligned}\|\Pi_{\mathcal{D}} v - \varphi\|_{L^p(\Omega)} + \|\nabla_{\mathcal{D}} v - \nabla\varphi\|_{L^p(\Omega)^d} \\ \le \|\Pi_{\mathcal{D}} v - \psi\|_{L^p(\Omega)} + \|\nabla_{\mathcal{D}} v - \nabla\psi\|_{L^p(\Omega)^d} \\ + \|\varphi - \psi\|_{L^p(\Omega)} + \|\nabla\varphi - \nabla\psi\|_{L^p(\Omega)^d}\,.\end{aligned}$$

Hence, $S_{\mathcal{D}_m}(\varphi) \le S_{\mathcal{D}_m}(\psi) + \|\varphi - \psi\|_{W_0^{1,p}(\Omega)} \le S_{\mathcal{D}_m}(\psi) + \varepsilon$, we get from (2.11) that $\limsup_{m\to\infty} S_{\mathcal{D}_m}(\varphi) \le \varepsilon$. The proof is then completed by letting $\varepsilon \to 0$. ∎

Lemma 2.17 (Equivalent condition for limit-conformity, Dirichlet BCs). *Let $(\mathcal{D}_m)_{m\in\mathbb{N}}$ be a sequence of gradient discretisations. Then $(\mathcal{D}_m)_{m\in\mathbb{N}}$ is limit-conforming in the sense of Definition 2.5 if and only if it is coercive in the sense of Definition 2.2, and there exists a dense subset $\boldsymbol{W}_w$ in $W^{p'}_{\mathrm{div}}(\Omega)$ (endowed with the norm $\|\boldsymbol{\varphi}\|_{W^{p'}_{\mathrm{div}}(\Omega)} = \|\boldsymbol{\varphi}\|_{L^{p'}(\Omega)^d} + \|\mathrm{div}\boldsymbol{\varphi}\|_{L^{p'}(\Omega)}$) such that*

$$\forall \boldsymbol{\psi} \in \boldsymbol{W}_w,\ \lim_{m\to\infty} W_{\mathcal{D}_m}(\boldsymbol{\psi}) = 0. \tag{2.12}$$

Remark 2.18. If Ω is locally star-shaped, then $\boldsymbol{W}_w = C_c^\infty(\mathbb{R}^d)^d$ is dense in $W^{p'}_{\mathrm{div}}(\Omega)$ (this can be established by following the technique in the proof of [134, Theorem 1.1] or [74, Lemma A.1]) and can therefore be used in Lemma 2.17.
We recall that Ω is locally star-shaped if, for any $\boldsymbol{x} \in \partial\Omega$, there is a neighbourhood $O_{\boldsymbol{x}}$ of $\boldsymbol{x}$ such that $O_{\boldsymbol{x}} \cap \Omega$ is star-shaped with respect to some $\boldsymbol{y_x}$ (i.e. for all $\boldsymbol{z} \in O_{\boldsymbol{x}} \cap \Omega$, $[\boldsymbol{y}_x, \boldsymbol{z}] \subset O_{\boldsymbol{x}} \cap \Omega$). In particular, polytopal open sets as in Section 7.1.1 are locally star-shaped.

Proof. Let us first assume that $(\mathcal{D}_m)_{m\in\mathbb{N}}$ is limit-conforming. Lemma 2.6 shows that $(\mathcal{D}_m)_{m\in\mathbb{N}}$ is coercive, and (2.12) is implied by (2.7). Reciprocally, let us assume that $(\mathcal{D}_m)_{m\in\mathbb{N}}$ is coercive (that is, there is $C_P \in \mathbb{R}_+$ such that $C_{\mathcal{D}_m} \le C_P$) and that (2.12) holds. To prove (2.7), let $\boldsymbol{\varphi} \in W^{p'}_{\mathrm{div}}(\Omega)$, $\varepsilon > 0$ and take $\boldsymbol{\psi} \in \boldsymbol{W}_w$ such that $\|\boldsymbol{\varphi} - \boldsymbol{\psi}\|_{W^{p'}_{\mathrm{div}}(\Omega)} \le \varepsilon$. This means that $\|\boldsymbol{\varphi} - \boldsymbol{\psi}\|_{L^{p'}(\Omega)^d} \le \varepsilon$ and $\|\mathrm{div}\boldsymbol{\varphi} - \mathrm{div}\boldsymbol{\psi}\|_{L^{p'}(\Omega)} \le \varepsilon$. We have

$$\begin{aligned}W_{\mathcal{D}_m}(\boldsymbol{\varphi}) &\le W_{\mathcal{D}_m}(\boldsymbol{\psi}) + \|\boldsymbol{\varphi} - \boldsymbol{\psi}\|_{L^{p'}(\Omega)^d} + C_P\,\|\mathrm{div}\boldsymbol{\varphi} - \mathrm{div}\boldsymbol{\psi}\|_{L^{p'}(\Omega)} \\ &\le W_{\mathcal{D}_m}(\boldsymbol{\psi}) + (1 + C_P)\varepsilon.\end{aligned}$$

Using (2.12) we deduce that $\limsup_{m\to\infty} W_{\mathcal{D}_m}(\boldsymbol{\varphi}) \le (1 + C_P)\varepsilon$ and the proof is completed by letting $\varepsilon \to 0$. ∎

Remark 2.19 (Condition (2.12) alone does not imply limit-conformity)
Without the coercivity property, (2.12) alone is not sufficient to ensure the limit-conformity. This is illustrated by the following example, a modification of one of the examples in Remark 2.14.
Let $\Omega = (0,1)$ and take $m \in \mathbb{N}^\star$, $h = 1/(2m)$, and $(\varphi_i)_{i=1,\dots,2m-1}$ the $\mathbb{P}_1$ finite element basis associated with the nodes $(ih)_{i=1,\dots,2m-1}$. Let $X_{\mathcal{D}_m,0} = \{u = (u_i)_{i=1,\dots,2m-1}\ , u_i \in \mathbb{R}\}$, $\widehat{u} = \sum_{i=1}^{2m-1} u_i\varphi_i$, $\nabla_{\mathcal{D}_m} u = \widehat{u}'$, and $\Pi_{\mathcal{D}_m} u(x) = \widehat{u}(x) + \frac{1}{\sqrt{h}}(\widehat{u}'(x) - \widehat{u}'(x+\varepsilon(x)))$ for all $x \in (0,1)$, with

$$\forall k = 0,\dots,m-1\,, \begin{cases} \varepsilon(x) = h & \text{if } x \in (2kh,(2k+1)h), \\ \varepsilon(x) = -h & \text{if } x \in ((2k+1)h,(2k+2)h). \end{cases}$$

Then, on one hand, $(\mathcal{D}_m)_{m\in\mathbb{N}}$ is GD-consistent (consider the natural $\mathbb{P}_1$ interpolation of functions belonging to $W_s = C^2([0,1]) \cap W_0^{1,p}(\Omega)$, and apply Lemma 2.16) and, on the other hand, it satisfies (2.12) with $\boldsymbol{W}_w = C^2([0,1])$. Indeed, if $\psi \in C^2([0,1])$ and $u \in X_{\mathcal{D}_m,0}$ then

$$\begin{aligned} &\int_0^1 \Pi_{\mathcal{D}_m} u(x)\psi'(x)\mathrm{d}x + \int_0^1 \nabla_{\mathcal{D}_m} u(x)\psi(x)\mathrm{d}x \\ &= \int_0^1 \widehat{u}(x)\psi'(x)\mathrm{d}x + \int_0^1 \widehat{u}'(x)\psi(x)\mathrm{d}x + \int_0^1 \widehat{u}'(x)\frac{1}{\sqrt{h}}(\psi'(x) - \psi'(x+\varepsilon(x)))\mathrm{d}x. \end{aligned}$$

Since $\widehat{u} \in W_0^{1,p}(\Omega)$, the first two integrals in the right-hand side cancel out. Moreover, $\frac{1}{\sqrt{h}}|\psi'(x) - \psi'(x+\varepsilon(x))| \le \sqrt{h}\,\|\psi''\|_{L^\infty(\Omega)}$ and thus $W_{\mathcal{D}_m}(\psi) \le \sqrt{h}\,\|\psi''\|_{L^\infty(\Omega)} \to 0$ as $m \to \infty$.
The sequence $(\mathcal{D}_m)_{m\in\mathbb{N}}$ is however not coercive and thus, by Lemma 2.6, not limit-conforming. To see this, define $u^m \in X_{\mathcal{D}_m,0}$ by $u^m_{2k} = 0$ and $u^m_{2k+1} = h$ for all $k = 0,\dots,m-1$. We have $\|\nabla_{\mathcal{D}_m} u^m\|_{L^p(\Omega)} = 1$ for all $m \in \mathbb{N}$. However, $|\Pi_{\mathcal{D}_m} u^m - \widehat{u}^m| = 2/\sqrt{h} \to \infty$ and $|\widehat{u}^m| \to 0$, which shows that $(\Pi_{\mathcal{D}_m} u^m)_{m\in\mathbb{N}}$ is not bounded in $L^p(\Omega)$ and thus that the coercivity property is not satisfied by $(\mathcal{D}_m)_{m\in\mathbb{N}}$.

Remark 2.20 (Abstract setting for GDs)
In Appendix A, an abstract setting allows the simultaneous analysis of all boundary conditions in the GDM; abstract spaces and operators are introduced (see Section A.1.1), enabling to recover some of the definitions and results given above, namely: Definitions 2.2, 2.4, 2.5 and 2.8 (of GDs and of their coercivity, consistency, limit-conformity and compactness) and Lemmas 2.6, 2.7, 2.10, 2.15, 2.16 and 2.17. However, for the sake of legibility, the theory of GDs for homogeneous Dirichlet boundary conditions is fully developed here without reference to this abstract setting.

Lemma 2.21 (Equivalent condition for compactness, Dirichlet BCs). *Let $\mathcal{D}$ be a gradient discretisation in the sense of Definition 2.1 and let $T_{\mathcal{D}} : \mathbb{R}^d \to \mathbb{R}^+$ be defined by*

$$\forall \boldsymbol{\xi} \in \mathbb{R}^d\,, \quad T_{\mathcal{D}}(\boldsymbol{\xi}) = \max_{v \in X_{\mathcal{D},0}\setminus\{0\}} \frac{\|\Pi_{\mathcal{D}} v(\cdot + \boldsymbol{\xi}) - \Pi_{\mathcal{D}} v\|_{L^p(\mathbb{R}^d)}}{\|v\|_{\mathcal{D}}}, \tag{2.13}$$

where $\Pi_{\mathcal{D}} v$ has been extended by 0 *outside* Ω.
The sequence $(\mathcal{D}_m)_{m\in\mathbb{N}}$ *is compact in the sense of Definition 2.8 if and only if*

$$\lim_{|\boldsymbol{\xi}|\to 0} \sup_{m\in\mathbb{N}} T_{\mathcal{D}_m}(\boldsymbol{\xi}) = 0.$$

Proof.
This lemma is a consequence of Kolmogorov's compactness theorem in Lebesgue spaces.

Step 1: we prove that the compactness of $(\mathcal{D}_m)_{m\in\mathbb{N}}$ in the sense of Definition 2.8 is equivalent to the relative compactness in $L^p(\Omega)$ of the set $A = \cup_{m\in\mathbb{N}} A_m$, where

$$A_m = \Pi_{\mathcal{D}_m}\left(\{u \in X_{\mathcal{D}_m,0}, \|u\|_{\mathcal{D}_m} = 1\}\right).$$

Indeed, any sequence in A is either contained in a finite union of A_m, which means that it remains bounded in a finite dimensional space, or has a subsequence which can be written $\Pi_{\mathcal{D}_{m(k)}} u_{m(k)}$ for some increasing sequence $(m(k))_{k\in\mathbb{N}} \subset \mathbb{N}$ and some $u_{m(k)} \in X_{\mathcal{D}_{m(k)},0}$ with $\left\|u_{m(k)}\right\|_{\mathcal{D}_{m(k)}} = 1$. Hence, the compactness of $(\mathcal{D}_m)_{m\in\mathbb{N}}$ gives the relative compactness of A in $L^p(\Omega)$.
Moreover, any sequence $u_m \in X_{\mathcal{D}_m,0}$ such that $\|u_m\|_{\mathcal{D}_m}$ is bounded can be written $u_m = \lambda_m u'_m$ with $(\lambda_m)_{m\in\mathbb{N}}$ bounded and $\|u'_m\|_{\mathcal{D}_m} = 1$. We have then $\Pi_{\mathcal{D}_m} u_m = \lambda_m v_m$ for some $v_m \in A$ and the relative compactness of A in $L^p(\Omega)$ therefore shows that $(\mathcal{D}_m)_{m\in\mathbb{N}}$ is compact in the sense of Definition 2.8.

Step 2: a statement of Kolmogorov's theorem.
Let $\widetilde{w} \in L^p(\mathbb{R}^d)$ be the extension of $w \in L^p(\Omega)$ by 0 outside Ω. A classical statement of Kolmogorov's compactness theorem is: A is relatively compact in $L^p(\Omega)$ if and only it is bounded in $L^p(\Omega)$ and if

$$\tau_A(\boldsymbol{\xi}) := \sup_{w\in A} \|\widetilde{w}(\cdot + \boldsymbol{\xi}) - \widetilde{w}\|_{L^p(\mathbb{R}^d)} \to 0 \text{ as } |\boldsymbol{\xi}| \to 0.$$

But τ_A is sub-additive. Indeed, for all $\boldsymbol{\xi}, \boldsymbol{\xi}' \in \mathbb{R}^d$ we have

$$\begin{aligned}
&\|\widetilde{w}(\cdot + \boldsymbol{\xi} + \boldsymbol{\xi}') - \widetilde{w}\|_{L^p(\mathbb{R}^d)} \\
&\le \|\widetilde{w}(\cdot + \boldsymbol{\xi} + \boldsymbol{\xi}') - \widetilde{w}(\cdot + \boldsymbol{\xi}')\|_{L^p(\mathbb{R}^d)} + \|\widetilde{w}(\cdot + \boldsymbol{\xi}') - \widetilde{w}\|_{L^p(\mathbb{R}^d)} \\
&= \|\widetilde{w}(\cdot + \boldsymbol{\xi}) - \widetilde{w}\|_{L^p(\mathbb{R}^d)} + \|\widetilde{w}(\cdot + \boldsymbol{\xi}') - \widetilde{w}\|_{L^p(\mathbb{R}^d)},
\end{aligned}$$

and therefore $\tau_A(\boldsymbol{\xi} + \boldsymbol{\xi}') \le \tau_A(\boldsymbol{\xi}) + \tau_A(\boldsymbol{\xi}')$. Hence, if $\lim_{|\boldsymbol{\xi}|\to 0} \tau_A(\boldsymbol{\xi}) = 0$, then τ_A is finite on a neighbourhood of 0 in $\mathbb{R}^d$ and its sub-additivity shows that it is in fact finite on $\mathbb{R}^d$.
Now, taking $\boldsymbol{\xi}_0 \in \mathbb{R}^d$ such that $|\boldsymbol{\xi}_0| > \operatorname{diam}(\Omega)$, for all $w \in A$ we see that $\widetilde{w}(\cdot + \boldsymbol{\xi})$ and $\widetilde{w}$ have disjoint supports and therefore

$$\tau_A(\boldsymbol{\xi}_0) \ge \left(\int_{\mathbb{R}^d} |\widetilde{w}(x + \boldsymbol{\xi}_0)|^p \, dx + \int_{\mathbb{R}^d} |\widetilde{w}(x)|^p \, dx\right)^{1/p} = 2^{1/p} \|w\|_{L^p(\Omega)}.$$

The finiteness of $\tau_A(\boldsymbol{\xi}_0)$ then ensures that A is bounded in $L^p(\Omega)$. Kolmogorov's theorem can therefore be re-stated as: A is relatively compact in $L^p(\Omega)$ if and only if $\lim_{|\boldsymbol{\xi}|\to 0} \tau_A(\boldsymbol{\xi}) = 0$.

Step 3: conclusion.
We have

$$A = \bigcup_{m\in\mathbb{N}} \Pi_{\mathcal{D}_m}\left(\left\{\frac{v}{\|v\|_{\mathcal{D}_m}}, v \in X_{\mathcal{D}_m,0} \setminus \{0\}\right\}\right)$$

and thus, since

$$T_{\mathcal{D}_m}(\boldsymbol{\xi}) = \max_{v\in X_{\mathcal{D}_m,0}\setminus\{0\}} \left\| \Pi_{\mathcal{D}_m}\left(\frac{v}{\|v\|_{\mathcal{D}_m}}\right)(\cdot + \boldsymbol{\xi}) - \Pi_{\mathcal{D}_m}\left(\frac{v}{\|v\|_{\mathcal{D}_m}}\right)\right\|_{L^p(\mathbb{R}^d)}$$

(the functions being extended by 0 outside Ω), we deduce $\sup_{m\in\mathbb{N}} T_{\mathcal{D}_m}(\boldsymbol{\xi}) = \tau_A(\boldsymbol{\xi})$. The conclusion then follows from Steps 1 and 2. ∎

When used to establish error estimates in the GDM framework, the consistency measure $S_{\mathcal{D}}$ is applied to the solution of the PDE, and the limit-conformity measure $W_{\mathcal{D}}$ is applied to some function of this solution's gradient. Oftentimes, specific regularity properties are known on this solution, and $S_{\mathcal{D}}$ and $W_{\mathcal{D}}$ are thus of particular interest on subspaces W_s and $\boldsymbol{W}_w$ of $W_0^{1,p}(\Omega)$ and $W_{\mathrm{div}}^{p'}(\Omega)$, respectively. Then W_s is continuously embedded in $W_0^{1,p}(\Omega)$, which means that $\|\cdot\|_{W^{1,p}(\Omega)} \le C_s \|\cdot\|_{W_s}$ and thus, for $\varphi \in W_s$,

$$\begin{aligned} S_{\mathcal{D}}(\varphi) &\le \|\Pi_{\mathcal{D}} 0 - \varphi\|_{L^p(\Omega)} + \|\nabla_{\mathcal{D}} 0 - \nabla\varphi\|_{L^p(\Omega)^d} \\ &\le 2\|\varphi\|_{W^{1,p}(\Omega)} \le 2C_s \|\varphi\|_{W_s}. \end{aligned}$$

Similarly, $\boldsymbol{W}_w$ is continuously embedded in $W_{\mathrm{div}}^{p'}(\Omega)$, so that $\|\cdot\|_{W_{\mathrm{div}}^{p'}(\Omega)} \le C_w \|\cdot\|_{\boldsymbol{W}_w}$ and, for $\boldsymbol{\varphi} \in \boldsymbol{W}_w$, using Hölder's inequality (D.5) and the definition (2.1) of $C_{\mathcal{D}}$,

$$\begin{aligned} W_{\mathcal{D}}(\boldsymbol{\varphi}) &= \sup_{v\in X_{\mathcal{D},0}\setminus\{0\}} \frac{1}{\|\nabla_{\mathcal{D}} v\|_{L^p(\Omega)^d}} \int_\Omega (\mathrm{div}\boldsymbol{\varphi}(\boldsymbol{x})\Pi_{\mathcal{D}} v(\boldsymbol{x}) + \boldsymbol{\varphi}(\boldsymbol{x})\cdot\nabla_{\mathcal{D}} v(\boldsymbol{x}))\mathrm{d}\boldsymbol{x} \\ &\le \sup_{v\in X_{\mathcal{D},0}\setminus\{0\}} \frac{1}{\|\nabla_{\mathcal{D}} v\|_{L^p(\Omega)^d}} \left((C_{\mathcal{D}}+1)\|\nabla_{\mathcal{D}} v\|_{L^p(\Omega)^d} \|\boldsymbol{\varphi}\|_{W_{\mathrm{div}}^{p'}(\Omega)}\right) \\ &\le (C_{\mathcal{D}}+1) C_w \|\boldsymbol{\varphi}\|_{\boldsymbol{W}_w}. \end{aligned}$$

These considerations justify the following notion of space size of a GD.

Definition 2.22 (Space size of a GD with respect to continuously embedded spaces). *Let W_s (resp. $\boldsymbol{W}_w$) be a Banach space continuously embedded in $W_0^{1,p}(\Omega)$ (resp. $W_{\mathrm{div}}^{p'}(\Omega)$). Let $\mathcal{D}$ be a gradient discretisation in the sense of Definition 2.1, and let $s_{\mathcal{D}}(W_s) \ge 0$ and $w_{\mathcal{D}}(\boldsymbol{W}_w) \ge 0$ be respectively defined by*

$$s_{\mathcal{D}}(W_s) = \sup\left\{\frac{S_{\mathcal{D}}(\varphi)}{\|\varphi\|_{W_s}} \,:\, \varphi \in W_s \setminus \{0\}\right\}, \tag{2.14}$$

$$w_{\mathcal{D}}(\boldsymbol{W}_w) = \sup\left\{\frac{W_{\mathcal{D}}(\boldsymbol{\varphi})}{\|\boldsymbol{\varphi}\|_{\boldsymbol{W}_w}} \,:\, \boldsymbol{\varphi} \in \boldsymbol{W}_w \setminus \{0\}\right\}. \tag{2.15}$$

The space size of $\mathcal{D}$ with respect to W_s and $\boldsymbol{W}_w$ is then defined by

$$h_{\mathcal{D}}(W_s; \boldsymbol{W}_w) = \max(s_{\mathcal{D}}(W_s), w_{\mathcal{D}}(\boldsymbol{W}_w)). \tag{2.16}$$

Remark 2.23. Under the hypotheses of Definition 2.22, if $\widehat{W}_s$ (resp. $\widehat{\boldsymbol{W}}_w$) is a Banach space continuously embedded in W_s (resp. $\boldsymbol{W}_w$) with $\|\cdot\|_{W_s} \leq \widehat{C}_s \|\cdot\|_{\widehat{W}_s}$ and $\|\cdot\|_{\boldsymbol{W}_w} \leq \widehat{C}_w \|\cdot\|_{\widehat{\boldsymbol{W}}_w}$, then

$$h_{\mathcal{D}}(\widehat{W}_s; \widehat{\boldsymbol{W}}_w) \leq \max(\widehat{C}_s, \widehat{C}_w) h_{\mathcal{D}}(W_s; \boldsymbol{W}_w)$$

Remark 2.24 (Link between $h_{\mathcal{D}}(W_s; \boldsymbol{W}_w)$ and the mesh size for mesh-based gradient discretisations)
For the mesh-based examples of GDs given in Part III, the following results are established (here, $h_{\mathcal{M}}$ is the mesh size defined by (7.6)):

- There exists $C > 0$ depending only on regularity factors of the mesh such that
$$h_{\mathcal{D}}(W^{2,\infty}(\Omega) \cap W_0^{1,p}(\Omega); W^{1,\infty}(\Omega)^d) \leq C h_{\mathcal{M}}.$$
- In the case $p > d/2$, there exists $C > 0$ depending only on regularity factors of the mesh such that
$$h_{\mathcal{D}}(W^{2,p}(\Omega) \cap W_0^{1,p}(\Omega); W^{1,p'}(\Omega)^d) \leq C h_{\mathcal{M}}.$$
- For some particular GDs of higher order (see Chapters 8, 9, 10 and 11), there exists $C > 0$ depending only on regularity factors of the mesh such that, for some $k \geq 2$,
$$h_{\mathcal{D}}(C^{k+1}(\overline{\Omega}) \cap W_0^{1,p}(\Omega); C^k(\overline{\Omega})^d) \leq C h_{\mathcal{M}}^k.$$

Lemma 2.25 (Necessary and sufficient conditions for consistency and limit-conformity, homogeneous Dirichlet BCs). *Let W_s be a Banach space compactly and densely embedded in $W_0^{1,p}(\Omega)$, and let $\boldsymbol{W}_w$ be a Banach space continuously and densely embedded in $W_{\mathrm{div}}^{p'}(\Omega)$, such that $\boldsymbol{W}_w$ is compactly embedded in $L^{p'}(\Omega)^d$.*
Assume that $(\mathcal{D}_m)_{m\in\mathbb{N}}$ is a compact sequence of gradient discretisations in the sense of Definition 2.8. Then $(\mathcal{D}_m)_{m\in\mathbb{N}}$ is GD-consistent and limit-conforming if and only if $h_{\mathcal{D}_m}(W_s; \boldsymbol{W}_w) \to 0$ as $m \to \infty$, where $h_{\mathcal{D}_m}(W_s; \boldsymbol{W}_w)$ is defined by (2.16).

Remark 2.26 (Classical W_s and $\boldsymbol{W}_w$ satisfy the hypotheses of the lemma)
The hypotheses on W_s and $\boldsymbol{W}_w$ in Lemma 2.25 hold in the cases in Remark 2.24, for which, respectively,

- $W_s = W^{2,\infty}(\Omega) \cap W_0^{1,p}(\Omega)$ and $\boldsymbol{W}_w = W^{1,\infty}(\Omega)^d$,
- $W_s = W^{2,p}(\Omega) \cap W_0^{1,p}(\Omega)$ and $\boldsymbol{W}_w = W^{1,p'}(\Omega)^d$,
- $W_s = C^{k+1}(\overline{\Omega}) \cap W_0^{1,p}(\Omega)$ and $\boldsymbol{W}_w = C^k(\overline{\Omega})^d$.

In all these examples, Ω is a polytopal open set, so the Rellich compactness embeddings apply.

Proof. Let us first consider a GD-consistent sequence $(\mathcal{D}_m)_{m\in\mathbb{N}}$ and let us prove, by way of contradiction, that $s_{\mathcal{D}_m}(W_s)$ defined by (2.14) tends to 0 as $m \to \infty$. Assume therefore that there exists $C > 0$ and a subsequence of $(\mathcal{D}_m)_{m\in\mathbb{N}}$, again denoted by $(\mathcal{D}_m)_{m\in\mathbb{N}}$, such that for all $m \in \mathbb{N}$, we can find $\varphi_m \in W_s \setminus \{0\}$ satisfying $S_{\mathcal{D}_m}(\varphi_m) \geq C \|\varphi_m\|_{W_s}$. Hence, for all $v \in X_{\mathcal{D}_m,0}$,

$$\|\Pi_{\mathcal{D}_m} v - \varphi_m\|_{L^p(\Omega)} + \|\nabla_{\mathcal{D}_m} v - \nabla\varphi_m\|_{L^p(\Omega)^d} \geq C \|\varphi_m\|_{W_s}.$$

Divide this relation by $\|\varphi_m\|_{W_s}$ and set $\widetilde{\varphi}_m = \varphi_m / \|\varphi_m\|_{W_s}$ and $w = v / \|\varphi_m\|_{W_s}$. The vector v being any vector in $X_{\mathcal{D}_m,0}$, w is also free to be any vector in this space. Hence, for all $w \in X_{\mathcal{D},0}$,

$$\|\Pi_{\mathcal{D}_m} w - \widetilde{\varphi}_m\|_{L^p(\Omega)} + \|\nabla_{\mathcal{D}_m} w - \nabla\widetilde{\varphi}_m\|_{L^p(\Omega)^d} \geq C.$$

Since $\|\widetilde{\varphi}_m\|_{W_s} = 1$, the compact embedding of W_s in $W_0^{1,p}(\Omega)$ gives the existence of a subsequence of $(\mathcal{D}_m, \varphi_m)_{m\in\mathbb{N}}$, again denoted by $(\mathcal{D}_m, \varphi_m)_{m\in\mathbb{N}}$, such that $\widetilde{\varphi}_m \to \psi$ in $W^{1,p}(\Omega)$ for some $\psi \in W_0^{1,p}(\Omega)$. The triangle inequality then yields, for all $w \in X_{\mathcal{D}_m,0}$,

$$\begin{aligned}\|\Pi_{\mathcal{D}_m} w - \psi\|_{L^p(\Omega)} &+ \|\nabla_{\mathcal{D}_m} w - \nabla\psi\|_{L^p(\Omega)^d} \\ &\geq C - (\|\psi - \widetilde{\varphi}_m\|_{L^p(\Omega)} + \|\nabla\psi - \nabla\widetilde{\varphi}_m\|_{L^p(\Omega)^d}),\end{aligned} \tag{2.17}$$

which implies

$$S_{\mathcal{D}_m}(\psi) \geq C - (\|\psi - \widetilde{\varphi}_m\|_{L^p(\Omega)} + \|\nabla\psi - \nabla\widetilde{\varphi}_m\|_{L^p(\Omega)^d}).$$

This shows that $\liminf_{m\to\infty} S_{\mathcal{D}_m}(\psi) \geq C$, which is a contradiction with the GD-consistency of $(\mathcal{D}_m)_{m\in\mathbb{N}}$ that gives $S_{\mathcal{D}_m}(\psi) \to 0$ as $m \to \infty$.

We now take a limit-conforming sequence $(\mathcal{D}_m)_{m\in\mathbb{N}}$ and prove, still by contradiction, that $w_{\mathcal{D}_m}(\boldsymbol{W}_w)$ defined by (2.15) tends to 0 as $m \to \infty$. Assume that there exists $C > 0$ and a subsequence of $(\mathcal{D}_m)_{m\in\mathbb{N}}$, again denoted by $(\mathcal{D}_m)_{m\in\mathbb{N}}$, such that, for all $m \in \mathbb{N}$, we can find $\boldsymbol{\varphi}_m \in \boldsymbol{W}_w \setminus \{0\}$ satisfying $W_{\mathcal{D}}(\boldsymbol{\varphi}_m) \geq C \|\boldsymbol{\varphi}_m\|_{\boldsymbol{W}_w}$. Setting $\widetilde{\boldsymbol{\varphi}}_m = \boldsymbol{\varphi}_m / \|\boldsymbol{\varphi}_m\|_{\boldsymbol{W}_w}$, we then have

$$W_{\mathcal{D}}(\widetilde{\boldsymbol{\varphi}}_m) \geq C. \tag{2.18}$$

Since $\|\widetilde{\boldsymbol{\varphi}}_m\|_{\boldsymbol{W}_w} = 1$ and $\boldsymbol{W}_w$ is compactly embedded in $L^{p'}(\Omega)^d$, there exists $\boldsymbol{\psi} \in \boldsymbol{W}_w$ such that, up to a subsequence denoted the same way, $\widetilde{\boldsymbol{\varphi}}_m \to \boldsymbol{\psi}$ strongly in $L^{p'}(\Omega)^d$ and $\mathrm{div}\widetilde{\boldsymbol{\varphi}}_m \to \mathrm{div}\boldsymbol{\psi}$ weakly in $L^{p'}(\Omega)$. Relation (2.18) gives $v_m \in X_{\mathcal{D}_m,0} \setminus \{0\}$ such that

$$\int_\Omega (\mathrm{div}\widetilde{\boldsymbol{\varphi}}_m(\boldsymbol{x})\Pi_{\mathcal{D}_m} v_m(\boldsymbol{x}) + \widetilde{\boldsymbol{\varphi}}_m(\boldsymbol{x}) \cdot \nabla_{\mathcal{D}_m} v_m(\boldsymbol{x}))\mathrm{d}\boldsymbol{x} \geq \frac{C}{2} \|\nabla_{\mathcal{D}_m} v_m\|_{L^p(\Omega)^d}.$$

The rescaling $\widetilde{v}_m = v_m / \|\nabla_{\mathcal{D}_m} v_m\|_{L^p(\Omega)^d}$ then enables us to write

$$\int_\Omega (\mathrm{div}\widetilde{\boldsymbol{\varphi}}_m(\boldsymbol{x})\Pi_{\mathcal{D}_m} \widetilde{v}_m(\boldsymbol{x}) + \widetilde{\boldsymbol{\varphi}}_m(\boldsymbol{x}) \cdot \nabla_{\mathcal{D}_m} \widetilde{v}_m(\boldsymbol{x}))\mathrm{d}\boldsymbol{x} \geq \frac{C}{2}. \tag{2.19}$$

Since $\|\widetilde{v}_m\|_{\mathcal{D}_m} = 1$, Lemma 2.15 and the compactness hypothesis on $(\mathcal{D}_m)_{m\in\mathbb{N}}$ give a subsequence of $(\mathcal{D}_m, \boldsymbol{\varphi}_m, v_m)_{m\in\mathbb{N}}$, again denoted by $(\mathcal{D}_m, \boldsymbol{\varphi}_m, v_m)_{m\in\mathbb{N}}$, such that $\Pi_{\mathcal{D}_m}\widetilde{v}_m \to v$ strongly in $L^p(\Omega)$ and $\nabla_{\mathcal{D}_m}\widetilde{v}_m \to \nabla v$ weakly in $L^p(\Omega)^d$, for some $v \in W_0^{1,p}(\Omega)$. Using the weak–strong convergence property (Lemma D.8) and Stokes' formula between functions in $W_0^{1,p}(\Omega)$ and $W_{\mathrm{div}}^{p'}(\Omega)$, we infer

$$\begin{aligned}\lim_{m\to\infty} \int_\Omega (\mathrm{div}\widetilde{\boldsymbol{\varphi}}_m(\boldsymbol{x})\Pi_{\mathcal{D}_m} \widetilde{v}_m(\boldsymbol{x}) + \widetilde{\boldsymbol{\varphi}}_m(\boldsymbol{x}) \cdot \nabla_{\mathcal{D}_m} \widetilde{v}_m(\boldsymbol{x}))\mathrm{d}\boldsymbol{x} \\ = \int_\Omega (\mathrm{div}\boldsymbol{\psi}(\boldsymbol{x})v(\boldsymbol{x}) + \boldsymbol{\psi}(\boldsymbol{x}) \cdot \nabla v(\boldsymbol{x}))\mathrm{d}\boldsymbol{x} = 0.\end{aligned}$$

This is a contradiction with (2.19), and the proof that $h_{\mathcal{D}_m}(W_s; \boldsymbol{W}_w) \to 0$ as $m \to \infty$ is complete.

Finally, if we assume that $h_{\mathcal{D}_m}(W_s; \boldsymbol{W}_w) \to 0$ as $m \to \infty$, then the GD-consistency and limit-conformity follow directly from Lemmas 2.16 and 2.17.

■

2.1.2 Gradient schemes for linear problems

In this section, linear problems are considered, so that $p = 2$ is chosen in all the definitions of Section 2.1.1. We consider the following problem:

$$-\mathrm{div}(\Lambda\nabla\overline{u}) = f + \mathrm{div}(\boldsymbol{F}) \text{ in } \Omega, \tag{2.20a}$$

with boundary conditions

$$\overline{u} = 0 \text{ on } \partial\Omega, \tag{2.20b}$$

under the following assumptions:

- Ω is an open bounded connected subset of $\mathbb{R}^d$ ($d \in \mathbb{N}^\star$), (2.21a)
- Λ is a measurable function from Ω to the set of $d \times d$

symmetric matrices and there exists $\underline{\lambda}, \overline{\lambda} > 0$ such that,

for a.e. $\boldsymbol{x} \in \Omega$, $\Lambda(\boldsymbol{x})$ has eigenvalues in $[\underline{\lambda}, \overline{\lambda}]$, (2.21b)

- $f \in L^2(\Omega)\,,\ \boldsymbol{F} \in L^2(\Omega)^d.$ (2.21c)

The assumptions on f and $\boldsymbol{F}$ include both the case of a right hand side of (2.20a) in $L^2(\Omega)$ (taking $\boldsymbol{F} = 0$) and the case of a right hand side in $H^{-1}(\Omega)$, since $H^{-1}(\Omega) = \{\mathrm{div}\boldsymbol{v} \,:\, \boldsymbol{v} \in L^2(\Omega)^d\}$. Note also that the symmetry assumption on $\Lambda(\boldsymbol{x})$ is not mandatory to study the convergence of the GDM for (2.20a), but it is commonly satisfied in applications. Under these hypotheses, the weak solution of (2.20) is the unique function $\overline{u}$ satisfying:

$$\begin{aligned}&\overline{u} \in H_0^1(\Omega),\ \forall v \in H_0^1(\Omega),\\ &\int_\Omega \Lambda(\boldsymbol{x})\nabla\overline{u}(\boldsymbol{x})\cdot\nabla v(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = \int_\Omega f(\boldsymbol{x})v(\boldsymbol{x})\mathrm{d}\boldsymbol{x} - \int_\Omega \boldsymbol{F}(\boldsymbol{x})\cdot\nabla v(\boldsymbol{x})\mathrm{d}\boldsymbol{x}.\end{aligned} \tag{2.22}$$

Let us now introduce an approximation of Problem (2.22) by the GDM.

Definition 2.27 (GS, homogeneous Dirichlet BCs).
If $\mathcal{D} = (X_{\mathcal{D},0}, \Pi_{\mathcal{D}}, \nabla_{\mathcal{D}})$ is a GD in the sense of Definition 2.1, then the related gradient scheme!linear problem *for Problem (2.22) is defined by*

$$\begin{aligned}&\textit{Find } u \in X_{\mathcal{D},0} \textit{ such that for any } v \in X_{\mathcal{D},0},\\ &\int_\Omega \Lambda(\boldsymbol{x})\nabla_{\mathcal{D}}u(\boldsymbol{x})\cdot\nabla_{\mathcal{D}}v(\boldsymbol{x})\mathrm{d}\boldsymbol{x} =\\ &\qquad\qquad \int_\Omega f(\boldsymbol{x})\Pi_{\mathcal{D}}v(\boldsymbol{x})\mathrm{d}\boldsymbol{x} - \int_\Omega \boldsymbol{F}(\boldsymbol{x})\cdot\nabla_{\mathcal{D}}v(\boldsymbol{x})\mathrm{d}\boldsymbol{x}.\end{aligned} \tag{2.23}$$

Let $(\xi^{(i)})_{i=1,\dots,N}$ be a basis of the space $X_{\mathcal{D},0}$; the scheme (2.23) is equivalent to solving the linear square system $AU = B$, where

$$\begin{aligned}u &= \sum_{j=1}^N U_j\xi^{(j)},\\ A_{ij} &= \int_\Omega \Lambda(\boldsymbol{x})\nabla_{\mathcal{D}}\xi^{(j)}(\boldsymbol{x})\cdot\nabla_{\mathcal{D}}\xi^{(i)}(\boldsymbol{x})\mathrm{d}\boldsymbol{x},\\ B_i &= \int_\Omega f(\boldsymbol{x})\Pi_{\mathcal{D}}\xi^{(i)}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} - \int_\Omega \boldsymbol{F}(\boldsymbol{x})\cdot\nabla_{\mathcal{D}}\xi^{(i)}(\boldsymbol{x})\mathrm{d}\boldsymbol{x}.\end{aligned} \tag{2.24}$$

The following theorem was first proved in [99] in the case $\boldsymbol{F} = 0$. It provides an upper bound for the approximation error of Problem (2.22) in terms of the measures of GD-consistency and conformity defect, in the spirit of the second Strang lemma [132] . This theorem also shows the optimality of this upper bound as, up to a multiplicative constant, it is also a lower bound of the approximation error.

Theorem 2.28 (Error estimate, homogeneous Dirichlet BCs). *Under Assumptions (2.21), let $\overline{u} \in H_0^1(\Omega)$ be the solution of Problem (2.22) (which implies that in the distribution sense $-\mathrm{div}(\Lambda\nabla\overline{u} + \boldsymbol{F}) = f \in L^2(\Omega)$ and therefore $\Lambda\nabla\overline{u} + \boldsymbol{F} \in H_{\mathrm{div}}(\Omega)$). Let $\mathcal{D}$ be a GD in the sense of Definition 2.1. Then there exists one and only one $u_\mathcal{D} \in X_{\mathcal{D},0}$ solution to the GS (2.23); this solution satisfies the following inequalities:*

$$\|\nabla\overline{u} - \nabla_\mathcal{D} u_\mathcal{D}\|_{L^2(\Omega)^d} \leq \frac{1}{\underline{\lambda}} \left[W_\mathcal{D}(\Lambda\nabla\overline{u} + \boldsymbol{F}) + (\overline{\lambda} + \underline{\lambda})S_\mathcal{D}(\overline{u})\right], \tag{2.25}$$

$$\|\overline{u} - \Pi_\mathcal{D} u_\mathcal{D}\|_{L^2(\Omega)} \leq \frac{1}{\underline{\lambda}} \left[C_\mathcal{D} W_\mathcal{D}(\Lambda\nabla\overline{u} + \boldsymbol{F}) + (C_\mathcal{D}\overline{\lambda} + \underline{\lambda})S_\mathcal{D}(\overline{u})\right], \tag{2.26}$$

where $C_\mathcal{D}$, $S_\mathcal{D}$ and $W_\mathcal{D}$ are respectively the norm of the reconstruction operator $\Pi_\mathcal{D}$, the GD-consistency defect and the conformity defect, defined by (2.1)–(2.6).
Moreover, the following inequalities hold

$$W_\mathcal{D}(\Lambda\nabla\overline{u} + \boldsymbol{F}) \leq \overline{\lambda}\|\nabla\overline{u} - \nabla_\mathcal{D} u_\mathcal{D}\|_{L^2(\Omega)^d}, \tag{2.27}$$

$$S_\mathcal{D}(\overline{u}) \leq \|\overline{u} - \Pi_\mathcal{D} u_\mathcal{D}\|_{L^2(\Omega)} + \|\nabla\overline{u} - \nabla_\mathcal{D} u_\mathcal{D}\|_{L^2(\Omega)^d}, \tag{2.28}$$

which show the existence of $C_1 > 0$ and $C_2 > 0$, only depending on $\overline{\lambda}$ and $\underline{\lambda}$, such that

$$\begin{aligned} C_1 &\left[S_\mathcal{D}(\overline{u}) + W_\mathcal{D}(\Lambda\nabla\overline{u} + \boldsymbol{F})\right] \\ &\leq \|\overline{u} - \Pi_\mathcal{D} u_\mathcal{D}\|_{L^2(\Omega)} + \|\nabla\overline{u} - \nabla_\mathcal{D} u_\mathcal{D}\|_{L^2(\Omega)^d} \\ &\leq C_2(1 + C_\mathcal{D})\left[S_\mathcal{D}(\overline{u}) + W_\mathcal{D}(\Lambda\nabla\overline{u} + \boldsymbol{F})\right]. \end{aligned} \tag{2.29}$$

Remark 2.29 (Error estimate with respect to the space size of the GD). Under the hypotheses of Theorem 2.28, assume moreover that the coefficients of Λ are Lipschitz-continuous, that $\boldsymbol{F} \in H^1(\Omega)^d$ and that $\overline{u} \in H^2(\Omega) \cap H_0^1(\Omega)$ (this regularity on $\overline{u}$ follows from the assumptions on Λ and $\boldsymbol{F}$ if $d \leq 3$ and Ω is convex or has a regular boundary). Let $h_\mathcal{D} := h_\mathcal{D}(H^2(\Omega) \cap H_0^1(\Omega); H^1(\Omega)^d)$ be the space size of the GD, given by Definition 2.22. Then Theorem 2.28 gives the existence of C_3, depending only on $C_P \geq C_\mathcal{D}$, $\overline{\lambda}$ and $\underline{\lambda}$, such that

$$\begin{aligned} &\|\overline{u} - \Pi_\mathcal{D} u_\mathcal{D}\|_{L^2(\Omega)} + \|\nabla\overline{u} - \nabla_\mathcal{D} u_\mathcal{D}\|_{L^2(\Omega)^d} \\ &\qquad \leq C_3(\|\Lambda\nabla\overline{u} + \boldsymbol{F}\|_{H^1(\Omega)^d} + \|\overline{u}\|_{H^2(\Omega)})h_\mathcal{D}. \end{aligned} \tag{2.30}$$

For all the mesh-based GDs presented in this book, this latter inequality provides an $O(h_\mathcal{M})$ error estimate, where $h_\mathcal{M}$ is the maximum of the diameters of the mesh cells.
Assume now that, for some $k \in \mathbb{N}^\star$, $\Lambda \in C^k(\overline{\Omega})^{d\times d}$, $\boldsymbol{F} \in C^k(\overline{\Omega})^d$ and $\overline{u} \in C^{k+1}(\overline{\Omega}) \cap H_0^1(\Omega)$. Let $h_{\mathcal{D},k} := h_\mathcal{D}(C^{k+1}(\overline{\Omega}) \cap H_0^1(\Omega); C^k(\overline{\Omega})^d)$. Then, by Theorem 2.28, there exists C_4, depending only on $C_P \geq C_\mathcal{D}$, $\overline{\lambda}$ and $\underline{\lambda}$, such that

$$\|\overline{u} - \Pi_{\mathcal{D}} u_{\mathcal{D}}\|_{L^2(\Omega)} + \|\nabla \overline{u} - \nabla_{\mathcal{D}} u_{\mathcal{D}}\|_{L^2(\Omega)^d} \\ \leq C_4(\|\Lambda \nabla \overline{u} + \boldsymbol{F}\|_{C^k(\overline{\Omega})^d} + \|\overline{u}\|_{C^{k+1}(\overline{\Omega})}) h_{\mathcal{D},k}. \quad (2.31)$$

Some GDs described in Part III are of higher order . This is the case of the $\mathbb{P}_k$ conforming method of Chapter 8, the $\mathbb{P}_k$ non-conforming methods of Chapter 9, the $\mathbb{RT}_k$ mixed methods of Chapter 10, and the $\mathbb{P}_k$ discontinuous Galerkin methods of Chapter 11. For these methods, it is shown that there exists some $C > 0$ depending only on the regularity factors of the mesh, such that $h_{\mathcal{D},k} \leq C h_{\mathcal{M}}^k$, in which case (2.31) yields an $\mathcal{O}(h_{\mathcal{M}}^k)$ error estimate.
Other higher order methods are known to be GDMs, see for example the recent work [59] concerning non-conforming virtual element methods [19, 38], non-conforming mimetic finite differences [127], and hybrid high order schemes [62].

Remark 2.30 (Super-convergence)
As noticed in Remark 2.29 above, the L^2 estimate in Theorem 2.28 only provides an $\mathcal{O}(h_{\mathcal{M}})$ rate of convergence for low-order schemes. It is well known that several of these schemes, *e.g.*, conforming and non-conforming $\mathbb{P}_1$ finite elements, enjoy a higher rate of convergence in L^2 norm [84, 32, 30]. This phenomenon is known as super-convergence. Thanks to the GDM framework, it is possible to establish an improved L^2 estimate that provides such super-convergence results for various schemes, including some for which super-convergence was previously not proved [80].

Proof of Theorem2.28. Let us first prove that, if (2.25) holds for any solution $u_{\mathcal{D}} \in X_{\mathcal{D},0}$ to Scheme (2.23), then the solution to Scheme (2.23) exists and is unique. Indeed, let us prove that, assuming (2.25), the matrix denoted by A of the linear system (2.24) is non-singular. This will be completed if we prove $AU = 0$ implies $U = 0$. Thus, we consider the particular case where $f = 0$ and $\boldsymbol{F} = 0$ which gives a zero right-hand side. In this case the solution $\overline{u}$ of (2.22) is equal a.e. to zero. Then from (2.25), we get that any solution to the scheme satisfies $\|u_{\mathcal{D}}\|_{\mathcal{D}} = 0$. Since $\|\cdot\|_{\mathcal{D}}$ is a norm on $X_{\mathcal{D},0}$ this leads to $u_{\mathcal{D}} = 0$. Therefore (2.24) (as well as (2.23)) has a unique solution for any right-hand side f and $\boldsymbol{F}$.
Let us now prove that any solution $u_{\mathcal{D}} \in X_{\mathcal{D},0}$ to Scheme (2.23) satisfies (2.25) and (2.26). As noticed in the statement of the theorem, we can take $\boldsymbol{\varphi} = \Lambda \nabla \overline{u} + \boldsymbol{F} \in H_{\mathrm{div}}(\Omega)$ in the definition (2.6) of $W_{\mathcal{D}}$. We then obtain, for a given $v \in X_{\mathcal{D},0}$,

$$\left| \int_\Omega \nabla_{\mathcal{D}} v(\boldsymbol{x}) \cdot (\Lambda(\boldsymbol{x}) \nabla \overline{u}(\boldsymbol{x}) + \boldsymbol{F}(\boldsymbol{x})) + \Pi_{\mathcal{D}} v(\boldsymbol{x}) \mathrm{div}(\Lambda \nabla \overline{u} + \boldsymbol{F})(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \right| \\ \leq \|v\|_{\mathcal{D}} \, W_{\mathcal{D}}(\Lambda \nabla \overline{u} + \boldsymbol{F}),$$

which leads, since $f = -\mathrm{div}(\Lambda \nabla \overline{u} + \boldsymbol{F})$ a.e., to

$$\left|\int_\Omega \nabla_{\mathcal{D}} v(\boldsymbol{x}) \cdot (\Lambda(\boldsymbol{x})\nabla\overline{u}(\boldsymbol{x}) + \boldsymbol{F}(\boldsymbol{x})) - \Pi_{\mathcal{D}} v(\boldsymbol{x}) f(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}\right| \\ \leq \|v\|_{\mathcal{D}}\ W_{\mathcal{D}}(\Lambda\nabla\overline{u} + \boldsymbol{F}). \quad (2.32)$$

Since $u_{\mathcal{D}}$ is a solution to (2.23), we get

$$\left|\int_\Omega \Lambda\nabla_{\mathcal{D}} v(\boldsymbol{x}) \cdot (\nabla\overline{u}(\boldsymbol{x}) - \nabla_{\mathcal{D}} u_{\mathcal{D}}(\boldsymbol{x}))\mathrm{d}\boldsymbol{x}\right| \leq \|v\|_{\mathcal{D}}\ W_{\mathcal{D}}(\Lambda\nabla\overline{u} + \boldsymbol{F}). \quad (2.33)$$

Define

$$I_{\mathcal{D}}\overline{u} = \operatorname*{argmin}_{w\in X_{\mathcal{D},0}} \left(\|\Pi_{\mathcal{D}} w - \overline{u}\|_{L^2(\Omega)} + \|\nabla_{\mathcal{D}} w - \nabla\overline{u}\|_{L^2(\Omega)^d}\right) \quad (2.34)$$

and notice that, by definition (2.2) of $S_{\mathcal{D}}$,

$$\|\Pi_{\mathcal{D}} I_{\mathcal{D}}\overline{u} - \overline{u}\|_{L^2(\Omega)} + \|\nabla_{\mathcal{D}} I_{\mathcal{D}}\overline{u} - \nabla\overline{u}\|_{L^2(\Omega)^d} = S_{\mathcal{D}}(\overline{u}). \quad (2.35)$$

Recalling the definition of $\|\cdot\|_{\mathcal{D}}$ in Definition 2.1, by (2.33) we get

$$\begin{aligned} &\left|\int_\Omega \Lambda\nabla_{\mathcal{D}} v(\boldsymbol{x}) \cdot (\nabla_{\mathcal{D}} I_{\mathcal{D}}\overline{u}(\boldsymbol{x}) - \nabla_{\mathcal{D}} u_{\mathcal{D}}(\boldsymbol{x}))\mathrm{d}\boldsymbol{x}\right| \\ &\quad\leq \|\nabla_{\mathcal{D}} v\|_{L^2(\Omega)^d}\ W_{\mathcal{D}}(\Lambda\nabla\overline{u} + \boldsymbol{F}) \\ &\qquad + \left|\int_\Omega \Lambda(\boldsymbol{x})\nabla_{\mathcal{D}} v(\boldsymbol{x}) \cdot (\nabla_{\mathcal{D}} I_{\mathcal{D}}\overline{u}(\boldsymbol{x}) - \nabla\overline{u}(\boldsymbol{x}))\mathrm{d}\boldsymbol{x}\right| \\ &\quad\leq \|\nabla_{\mathcal{D}} v\|_{L^2(\Omega)^d}\left(W_{\mathcal{D}}(\Lambda\nabla\overline{u} + \boldsymbol{F}) + \overline{\lambda}\,\|\nabla_{\mathcal{D}} I_{\mathcal{D}}\overline{u} - \nabla\overline{u}\|_{L^2(\Omega)^d}\right) \\ &\quad\leq \|\nabla_{\mathcal{D}} v\|_{L^2(\Omega)^d}\left(W_{\mathcal{D}}(\Lambda\nabla\overline{u} + \boldsymbol{F}) + \overline{\lambda} S_{\mathcal{D}}(\overline{u})\right). \end{aligned} \quad (2.36)$$

Choosing $v = I_{\mathcal{D}}\overline{u} - u_{\mathcal{D}}$ yields

$$\underline{\lambda}\|\nabla_{\mathcal{D}}(I_{\mathcal{D}}\overline{u} - u_{\mathcal{D}})\|_{L^2(\Omega)^d} \leq W_{\mathcal{D}}(\Lambda\nabla\overline{u} + \boldsymbol{F}) + \overline{\lambda} S_{\mathcal{D}}(\overline{u}) \quad (2.37)$$

and (2.25) follows by writing

$$\begin{aligned} &\|\nabla\overline{u} - \nabla_{\mathcal{D}} u_{\mathcal{D}}\|_{L^2(\Omega)^d} \\ &\qquad\leq \|\nabla\overline{u} - \nabla_{\mathcal{D}} I_{\mathcal{D}}\overline{u}\|_{L^2(\Omega)^d} + \|\nabla_{\mathcal{D}}(I_{\mathcal{D}}\overline{u} - u_{\mathcal{D}})\|_{L^2(\Omega)^d} \\ &\qquad\leq S_{\mathcal{D}}(\overline{u}) + \frac{1}{\underline{\lambda}}\left(W_{\mathcal{D}}(\Lambda\nabla\overline{u} + \boldsymbol{F}) + \overline{\lambda} S_{\mathcal{D}}(\overline{u})\right). \end{aligned} \quad (2.38)$$

Using (2.1) and (2.37), we get

$$\underline{\lambda}\,\|\Pi_{\mathcal{D}} I_{\mathcal{D}}\overline{u} - \Pi_{\mathcal{D}} u_{\mathcal{D}}\|_{L^2(\Omega)} \leq C_{\mathcal{D}}(W_{\mathcal{D}}(\Lambda\nabla\overline{u} + \boldsymbol{F}) + \overline{\lambda} S_{\mathcal{D}}(\overline{u})), \quad (2.39)$$

which yields (2.26) by using, as in (2.38), a triangle inequality and the estimate $\|\overline{u} - \Pi_{\mathcal{D}} I_{\mathcal{D}}\overline{u}\|_{L^2(\Omega)^d} \leq S_{\mathcal{D}}(\overline{u})$.

Let us now turn to the proof of (2.27). From (2.23), we get for any $v \in X_{\mathcal{D},0} \setminus \{0\}$,

$$\int_\Omega f(\boldsymbol{x})\Pi_{\mathcal{D}} v(\boldsymbol{x})\mathrm{d}\boldsymbol{x} - \int_\Omega (\boldsymbol{F}(\boldsymbol{x}) + \Lambda(\boldsymbol{x})\nabla\overline{u}(\boldsymbol{x})) \cdot \nabla_{\mathcal{D}} v(\boldsymbol{x})\mathrm{d}\boldsymbol{x}$$
$$= \int_\Omega \Lambda(\boldsymbol{x})(\nabla_{\mathcal{D}} u(\boldsymbol{x}) - \nabla\overline{u}(\boldsymbol{x})) \cdot \nabla_{\mathcal{D}} v(\boldsymbol{x})\mathrm{d}\boldsymbol{x}.$$

The Cauchy–Schwarz inequality then implies

$$\frac{\left|\int_\Omega f(\boldsymbol{x})\Pi_{\mathcal{D}} v(\boldsymbol{x})\mathrm{d}\boldsymbol{x} - \int_\Omega (\boldsymbol{F}(\boldsymbol{x}) + \Lambda(\boldsymbol{x})\nabla\overline{u}) \cdot \nabla_{\mathcal{D}} v(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\right|}{\|\nabla_{\mathcal{D}} v\|_{L^2(\Omega)^d}}$$
$$\leq \overline{\lambda}\, \|\nabla_{\mathcal{D}} u - \nabla\overline{u}\|_{L^2(\Omega)^d}.$$

Taking the supremum over v on the left hand side yields (2.27). Inequality (2.28) is an immediate consequence of the definition of $S_{\mathcal{D}}(\overline{u})$. ∎

Let us conclude this section by stating the convergence of the GS, which follows easily from Theorem 2.28 (notice that, by Lemma 2.6, the sequence of GDs considered in the corollary is also coercive in the sense of Definition 2.2).

Corollary 2.31 (Convergence, homogeneous Dirichlet BCs). *Under Hypotheses (2.21), let $(\mathcal{D}_m)_{m\in\mathbb{N}}$ be a sequence of GDs in the sense of Definition 2.1, which is GD-consistent and limit-conforming in the sense of Definitions 2.4 and 2.5.*
Then, for any $m \in \mathbb{N}$, there exists a unique solution $u_m \in X_{\mathcal{D}_m,0}$ to the gradient scheme (2.23) and, if $\overline{u}$ is the solution of (2.22) then, as $m \to \infty$, $\Pi_{\mathcal{D}_m} u_m$ converges to $\overline{u}$ in $L^2(\Omega)$ and $\nabla_{\mathcal{D}_m} u_m$ converges to $\nabla\overline{u}$ in $L^2(\Omega)^d$.

Remark 2.32 (On the compactness assumption). Note that, in the linear case, the compactness of the sequence of GDs is not required to obtain the convergence. This compactness assumption is in general only needed for some non-linear problems (see also Remark 2.46).

Remark 2.33 (GD-consistency and limit-conformity are necessary conditions)
We state here a kind of reciprocal property to Corollary 2.31. Let us assume that, under Assumptions (2.21a)–(2.21b), a sequence $(\mathcal{D}_m)_{m\in\mathbb{N}}$ of GDs is such that, for all $f \in L^2(\Omega)$ and $\boldsymbol{F} \in L^2(\Omega)^d$ and for all $m \in \mathbb{N}$, there exists $u_m \in X_{\mathcal{D}_m,0}$ which is solution to the gradient scheme (2.23) and which satisfies that $\Pi_{\mathcal{D}_m} u_m$ (resp. $\nabla_{\mathcal{D}_m} u_m$) converges in $L^2(\Omega)$ to the solution $\overline{u}$ of (2.22) (resp. in $L^2(\Omega)^d$ to $\nabla\overline{u}$). Then $(\mathcal{D}_m)_{m\in\mathbb{N}}$ is GD-consistent and limit-conforming in the sense of Definitions 2.4 and 2.5.
Indeed, for $\varphi \in H_0^1(\Omega)$, let us consider $f = 0$ and $\mathbf{F} = -\Lambda\nabla\varphi$ in (2.22). Since in this case, $\overline{u} = \varphi$, the assumption that $\Pi_{\mathcal{D}_m} u_m$ (resp. $\nabla_{\mathcal{D}_m} u_m$) converges in $L^2(\Omega)$

to the solution φ of (2.22) (resp. converges in $L^2(\Omega)^d$ to $\nabla\varphi$), inequality (2.28) proves that $S_{\mathcal{D}_m}(\varphi)$ tends to 0 as $m \to \infty$, and therefore the sequence $(\mathcal{D}_m)_{m\in\mathbb{N}}$ is GD-consistent.
For $\boldsymbol{\varphi} \in H_{\mathrm{div}}(\Omega)$, let us set $f = \mathrm{div}\boldsymbol{\varphi}$ and $\boldsymbol{F} = -\boldsymbol{\varphi}$ in (2.22). In this case, the solution $\overline{u}$ is equal to 0 a.e., since the right-hand side of (2.22) vanishes for any $v \in H_0^1(\Omega)$. Then inequality (2.27) implies

$$W_{\mathcal{D}_m}(\boldsymbol{\varphi}) \leq \overline{\lambda}\|\nabla_{\mathcal{D}_m} u_m\|_{L^2(\Omega)^d} \to 0 \text{ as } m \to 0,$$

hence concluding that the sequence $(\mathcal{D}_m)_{m\in\mathbb{N}}$ is limit-conforming.
Note that, if we now assume that $\nabla_{\mathcal{D}_m} u_m$ converges only weakly (instead of strongly) in $L^2(\Omega)^d$ to $\nabla\overline{u}$, the same conclusion holds. Indeed, the other hypotheses on $(\mathcal{D}_m)_{m\in\mathbb{N}}$ are sufficient to prove that $\nabla_{\mathcal{D}_m} u_m$ actually converges strongly in $L^2(\Omega)^d$ to $\nabla\overline{u}$. It suffices to observe that

$$\lim_{m\to\infty} \int_\Omega (f(\boldsymbol{x})\Pi_{\mathcal{D}_m} u_m(\boldsymbol{x}) - \boldsymbol{F}(\boldsymbol{x}) \cdot \nabla_{\mathcal{D}_m} u_m(\boldsymbol{x}))\mathrm{d}\boldsymbol{x} \\ = \int_\Omega (f(\boldsymbol{x})\overline{u}(\boldsymbol{x}) - \boldsymbol{F}(\boldsymbol{x}) \cdot \nabla\overline{u}(\boldsymbol{x}))\mathrm{d}\boldsymbol{x}.$$

Then we take $v = \overline{u}$ in (2.22) and $v = u_m$ in (2.23), this leads to

$$\lim_{m\to\infty} \int_\Omega \Lambda(\boldsymbol{x})\nabla_{\mathcal{D}_m} u_m(\boldsymbol{x}) \cdot \nabla_{\mathcal{D}_m} u_m(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = \int_\Omega \Lambda(\boldsymbol{x})\nabla\overline{u}(\boldsymbol{x}) \cdot \nabla\overline{u}(\boldsymbol{x})\mathrm{d}\boldsymbol{x}.$$

In addition to the assumed weak convergence property of $\nabla_{\mathcal{D}_m} u_m$, this proves

$$\lim_{m\to\infty} \int_\Omega \Lambda(\boldsymbol{x})(\nabla_{\mathcal{D}_m} u_m(\boldsymbol{x}) - \nabla\overline{u}(\boldsymbol{x})) \cdot (\nabla_{\mathcal{D}_m} u_m(\boldsymbol{x}) - \nabla\overline{u}(\boldsymbol{x}))\mathrm{d}\boldsymbol{x} = 0,$$

and the convergence of $\nabla_{\mathcal{D}_m} u_m$ to $\nabla\overline{u}$ in $L^2(\Omega)^d$ follows.

2.1.3 On the notions of consistency and stability

Theorem 2.28 gives a control of the approximation error thanks to the GD-consistency indicator $S_{\mathcal{D}}$, the limit-conformity indicator $W_{\mathcal{D}}$ and the coercivity indicator $C_{\mathcal{D}}$. This theorem yields the convergence of the GDM for sequences of GDs that are GD-consistent and limit-conforming, as stated in Corollary 2.31 below. Can this be re-stated as the usual

$$\textit{Consistency} \text{ and } \textit{Stability} \Longrightarrow \textit{Convergence} \tag{2.40}$$

statement, well-known in the context of finite difference schemes? The answer to this question is yes, provided a correct definition of consistency is chosen. In the classical finite difference setting, the consistency error measures (roughly speaking) how well the exact solution "fits" into the scheme. Formally, assume that the equation to be discretised is written under the form $Lu = f$, and that

the scheme is under the form $L_h u_h = f_h = \Pi_h f$ where h denotes the discretisation step and, for a given function g, $\Pi_h g$ is the vector whose components are the values $(g(\boldsymbol{x}_i))_{i=1,\dots,N}$ of g at the discretisation points $(\boldsymbol{x}_i)_{i=1,\dots,N}$. Then the consistency error for the finite difference scheme is defined by

$$c_h = L_h \Pi_h u - f_h = L_h \Pi_h u - \Pi_h(Lu).$$

In this context, it is well-known that (2.40) holds: indeed, consistency (*i.e.* $c_h \to 0$ as $h \to 0$) and stability (*i.e.* L_h^{-1} bounded) imply convergence (*i.e.* $\max_{i=1,\dots,N} |\Pi_h(u) - u_h|(\boldsymbol{x}_i) \to 0$ as $h \to 0$).
In the finite element context, (2.40) no longer holds under these terms, although the spirit remains the same. The reason for the failure of (2.40) is that consistency no longer refers to how the exact solution fits into the complete scheme, but only into the discrete equation of the scheme. To be more explicit, consider the following elliptic problem:

$$u \in V, \tag{2.41}$$

$$a(u,v) = (f,v)\,, \quad \forall v \in V, \tag{2.42}$$

where $V = H_0^1(\Omega)$, $f \in L^2(\Omega)$, and a is a continuous coercive bilinear form on V. Consider a finite element scheme for the discretisation of Problem (2.42), which reads

$$u_h \in V_h, \tag{2.43}$$

$$a_h(u_h,v) = (f,v)\,, \quad \forall v \in V_h, \tag{2.44}$$

where V_h is a finite dimensional space. In order to measure "how well the exact solution fits into the scheme", the consistency error should measure

(i) how far V_h is from V,
(ii) how far κ_h is from 0, with

$$\kappa_h = \max_{v \in V_h \setminus \{0\}} \frac{|a_h(\Pi_h u, v) - (f,v)|}{\|v\|_V}, \tag{2.45}$$

where $\Pi_h u$ is either u itself, or some kind of interpolant of u. In most finite element textbooks, these two notions have been separated: Property (i) is measured by the so-called interpolation error, while the term "consistency" (or asymptotic consistency) only refers to the fact that $\kappa_h = 0$ (or $\kappa_h \to 0$ as $h \to 0$). We shall call this latter property "FEM-consistency" for the sake of clarity. For the conforming $\mathbb{P}_1$ finite element for instance, $a_h = a$, $\Pi_h u$ can be taken equal to u, and $\kappa_h = 0$, in which case the finite element scheme is said to be consistent. However, there are cases where the solution to the PDE itself cannot be plugged into the scheme's equation (for instance when using numerical quadrature), but when an *interpolant of this solution* needs to be used; for more on this, see, *e.g.*, [49, Chapter 4] or [87, Chapter 20]. Hence in the FEM context, (2.40) still holds provided

$$\textit{Consistency} = \textit{FEM-consistency} \text{ and } \textit{interpolation error control}. \tag{2.46}$$

For the stability issue in the FEM context, we also refer to [87].
Let us now view a finite element method as a GDM. In this context, the GD-consistency (see Definition 2.4) together with the limit-conformity (see Definition 2.2) is sufficient to ensure the consistency of the scheme in sense (2.46). Indeed, in the context of the GDM, the equivalent of the term κ_h defined by (2.45) reads (for $\boldsymbol{F} = 0$)

$$\kappa_{\mathcal{D}} = \underbrace{\frac{\displaystyle\int_\Omega \Lambda(\boldsymbol{x})\nabla_{\mathcal{D}}(I_{\mathcal{D}}\overline{u})(\boldsymbol{x}) \cdot \nabla_{\mathcal{D}} v(\boldsymbol{x})\mathrm{d}\boldsymbol{x} - \int_\Omega \Lambda(\boldsymbol{x})\nabla\overline{u}(\boldsymbol{x}) \cdot \nabla_{\mathcal{D}} v(\boldsymbol{x})\mathrm{d}\boldsymbol{x}}{\|\nabla_{\mathcal{D}} v\|_{L^2(\Omega)^d}}}_{\text{Controlled by } S_{\mathcal{D}}(\overline{u}), \text{ see } (2.36)}$$
$$+ \underbrace{\frac{\displaystyle\int_\Omega \Lambda(\boldsymbol{x})\nabla\overline{u}(\boldsymbol{x}) \cdot \nabla_{\mathcal{D}} v(\boldsymbol{x})\mathrm{d}\boldsymbol{x} - \int_\Omega f(\boldsymbol{x})\Pi_{\mathcal{D}} v(\boldsymbol{x})\mathrm{d}\boldsymbol{x}}{\|\nabla_{\mathcal{D}} v\|_{L^2(\Omega)^d}}}_{\text{Controlled by } W_{\mathcal{D}}(\lambda\nabla\overline{u}), \text{ see } (2.32)}.$$

Note that GD-consistency and stability (or coercivity) are not sufficient to prove the convergence of a general GDM. It is clear that limit-conformity is inherent in all conforming finite element methods; for non-conforming methods, it is needed to ensure that the discrete function reconstruction and the discrete gradient reconstruction are chosen in a coherent way. Hence for the GDM, we may also write

$$\textit{Consistency} \text{ and } \textit{Stability} \Longrightarrow \textit{Convergence}, \tag{2.47}$$

provided

$$\textit{Consistency} = \textit{GD-consistency} \text{ and } \textit{Limit-conformity}. \tag{2.48}$$

There is an additional twist. In (2.47), the *stability* of a sequence of GDs amounts to its coercivity, since a bound on $C_{\mathcal{D}}$ and the *consistency* (as defined by (2.48)) are sufficient to prove the convergence by Theorem 2.28. However, by Lemma 2.6 the limit-conformity implies the coercivity and thus the stability. Hence, (2.47) can actually be recast, in the context of the GDM and with the definition (2.48),

$$\textit{Consistency} \Longrightarrow \textit{Convergence}.$$

The trick, that does not seem to be explicit in most FEM books, is that the consistency hides a stability requirement.

2.1.4 Gradient schemes for quasi-linear problems

Here, and in the other sections on quasi-linear problems, we consider the quasi-linear operator[1] $u \mapsto -\mathrm{div}(\Lambda(\boldsymbol{x}, \overline{u}(\boldsymbol{x}))\nabla\overline{u})$, which is often used to model non-linear heterogeneous materials. For such an operator, we remain in the functional framework of the linear case, taking again $p = 2$ in the definitions and results of Section 2.1.1.

We consider the following problem:

$$-\mathrm{div}(\Lambda(\boldsymbol{x}, \overline{u})\nabla\overline{u}) = f \text{ in } \Omega, \tag{2.49a}$$

with boundary conditions

$$\overline{u} = 0 \text{ on } \partial\Omega, \tag{2.49b}$$

under the following assumptions:

- Ω is an open bounded connected subset of $\mathbb{R}^d$, $d \in \mathbb{N}^\star$, (2.50a)
- Λ is a Caratheodory function from $\Omega \times \mathbb{R}$ to $\mathcal{M}_d(\mathbb{R})$,
  (*i.e.* $\Lambda(\boldsymbol{x}, s)$ is measurable w.r.t. $\boldsymbol{x}$ and continuous w.r.t. s),
  there exists $\underline{\lambda}, \overline{\lambda} > 0$ such that, for a.e. $\boldsymbol{x} \in \Omega$, for all $s \in \mathbb{R}$,
  $\Lambda(\boldsymbol{x}, s)$ is symmetric with eigenvalues in $[\underline{\lambda}, \overline{\lambda}]$, (2.50b)
- $f \in L^2(\Omega)$. (2.50c)

Under these hypotheses, a weak solution of (2.49a) is a function $\overline{u}$ (not necessarily unique) satisfying:

$$\begin{aligned} &\overline{u} \in H_0^1(\Omega),\ \forall v \in H_0^1(\Omega), \\ &\int_\Omega \Lambda(\boldsymbol{x}, \overline{u}(\boldsymbol{x}))\nabla\overline{u}(\boldsymbol{x}) \cdot \nabla v(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = \int_\Omega f(\boldsymbol{x})v(\boldsymbol{x})\mathrm{d}\boldsymbol{x}. \end{aligned} \tag{2.51}$$

Then Problem (2.51) is approximated under Assumptions (2.50) by the following gradient scheme.

Definition 2.34 (GS, quasi-linear problem, homogeneous Dirichlet BCs). *If $\mathcal{D} = (X_{\mathcal{D},0}, \Pi_\mathcal{D}, \nabla_\mathcal{D})$ is a GD in the sense of Definition 2.1, then we define the related* gradient scheme *for* (2.22) *by*

$$\begin{aligned} &\textit{Find } u \in X_{\mathcal{D},0} \textit{ such that, for any } v \in X_{\mathcal{D},0}, \\ &\int_\Omega \Lambda(\boldsymbol{x}, \Pi_\mathcal{D} u(\boldsymbol{x}))\nabla_\mathcal{D} u(\boldsymbol{x}) \cdot \nabla_\mathcal{D} v(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = \int_\Omega f(\boldsymbol{x})\Pi_\mathcal{D} v(\boldsymbol{x})\mathrm{d}\boldsymbol{x}. \end{aligned} \tag{2.52}$$

[1] Recall that a partial differential operator is said to be quasi-linear if it is linear with respect to all the highest order derivatives of the unknown function.

Note that, considering a basis $(\xi^{(i)})_{i=1,\dots,N}$ of the space $X_{\mathcal{D},0}$, Scheme (2.52) is equivalent to solving the system of N non-linear equations with N unknowns $A(u)U = B$ with

$$\begin{aligned} u &= \sum_{j=1}^{N} U_j \xi^{(j)}, \\ A_{ij}(u) &= \int_\Omega \Lambda(\boldsymbol{x}, \Pi_{\mathcal{D}} u(\boldsymbol{x})) \nabla_{\mathcal{D}} \xi^{(j)}(\boldsymbol{x}) \cdot \nabla_{\mathcal{D}} \xi^{(i)}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}, \\ B_i &= \int_\Omega f(\boldsymbol{x}) \Pi_{\mathcal{D}} \xi^{(i)}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}. \end{aligned} \tag{2.53}$$

Standard methods for the approximation of a solution of this system can be considered, such as the fixed point method $A(u^{(k)})U^{(k+1)} = B$ or the Newton method.

An error estimate between an approximate solution and a weak solution to (2.52) cannot be stated, since the uniqueness of the solution to neither (2.53) nor (2.52) is known in the general case. We can nonetheless establish a convergence result on the GDM for quasi-linear models.

Theorem 2.35 (Convergence, quasi-linear problem, homogeneous Dirichlet BCs). *Under assumptions (2.50), take a sequence $(\mathcal{D}_m)_{m\in\mathbb{N}}$ of GDs in the sense of Definition 2.1, which is GD-consistent, limit-conforming and compact in the sense of Definitions 2.4, 2.5 and 2.8.*
Then, for any $m \in \mathbb{N}$, there exists at least one $u_m \in X_{\mathcal{D}_m,0}$ solution to the gradient scheme (2.52) and, up to a subsequence, $\Pi_{\mathcal{D}_m} u_m$ converges strongly in $L^2(\Omega)$ to a solution $\overline{u}$ of (2.51) and $\nabla_{\mathcal{D}_m} u_m$ converges strongly in $L^2(\Omega)^d$ to $\nabla\overline{u}$ as $m \to \infty$.
In the case where the solution $\overline{u}$ of (2.51) is unique, then the whole sequence converges to $\overline{u}$ as $m \to \infty$ in the senses above.

Proof.
By Lemma 2.10, $(\mathcal{D}_m)_{m\in\mathbb{N}}$ is coercive in the sense of Definition 2.2.
Step 1: existence of a solution to the scheme.
Let $\mathcal{D} = (X_{\mathcal{D},0}, \Pi_{\mathcal{D}}, \nabla_{\mathcal{D}})$ be a GD in the sense of Definition 2.1. Let $w \in X_{\mathcal{D},0}$ be given, and let $u \in X_{\mathcal{D},0}$ be such that

$$\begin{aligned} &\text{Find } u \in X_{\mathcal{D},0} \text{ such that, } \forall v \in X_{\mathcal{D},0}, \\ &\int_\Omega \Lambda(\boldsymbol{x}, \Pi_{\mathcal{D}} w(\boldsymbol{x})) \nabla_{\mathcal{D}} u(\boldsymbol{x}) \cdot \nabla_{\mathcal{D}} v(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} = \int_\Omega f(\boldsymbol{x}) \Pi_{\mathcal{D}} v(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}. \end{aligned} \tag{2.54}$$

Therefore, u is solution to the square linear system $A(w)U = B$, where $A(w)$ and B are defined in (2.53). Let us prove that the matrix $A(w)$ is invertible. Letting $v = u$ in (2.54), and applying the Cauchy–Schwarz inequality and Hypothesis (2.50b), we get

$$\underline{\lambda} \|\nabla_{\mathcal{D}} u\|^2_{L^2(\Omega)^d} \le \|f\|_{L^2(\Omega)} \|\Pi_{\mathcal{D}} u\|_{L^2(\Omega)} \le C_{\mathcal{D}} \|f\|_{L^2(\Omega)} \|\nabla_{\mathcal{D}} u\|_{L^2(\Omega)^d},$$

where $C_{\mathcal{D}}$ is defined by (2.1) in Definition 2.2. This shows that

$$\|\nabla_{\mathcal{D}} u\|_{L^2(\Omega)^d} \leq \frac{C_{\mathcal{D}}}{\underline{\lambda}} \|f\|_{L^2(\Omega)}. \tag{2.55}$$

This completes the proof that $A(w)$ is invertible, since (2.55) shows that $A(w)U = 0$ implies $U = 0$. We then can define the mapping $F : \mathbb{R}^N \to \mathbb{R}^N$, by $F(w) = U$, with U is the solution of the linear system $A(w)U = B$. This mapping is continuous, thanks to the continuity of the coefficients of the inverse of a matrix with respect to its coefficients. Moreover, we get from (2.55) that some norm of U remains bounded, which means that F maps $\mathbb{R}^N$ into some closed ball B if $\mathbb{R}^N$. Therefore the Brouwer fixed point theorem D.2 proves that the equation $F(U) = U$ has at least one solution. This proves the existence of at least one discrete solution to (2.52).
We note that the previous estimates easily show that any solution to this scheme satisfies (2.55).

Step 2: convergence of $\Pi_{\mathcal{D}_m} u_m$ and $\nabla_{\mathcal{D}_m} u_m$.
Thanks to the coercivity hypothesis and (2.55), we have

$$\|\nabla_{\mathcal{D}_m} u_m\|_{L^2(\Omega)^d} \leq \frac{C_P}{\underline{\lambda}} \|f\|_{L^2(\Omega)}. \tag{2.56}$$

We may then apply Lemma 2.15, which states that there exists a subsequence of $(\mathcal{D}_m, u_m)_{m\in\mathbb{N}}$, denoted in the same way, and there exists $\overline{u} \in H_0^1(\Omega)$ such that $\nabla_{\mathcal{D}_m} u_m$ converges weakly in $L^2(\Omega)^d$ to $\nabla\overline{u}$ and $\Pi_{\mathcal{D}_m} u_m$ converges weakly in $L^2(\Omega)$ to $\overline{u}$. Thanks to the compactness hypothesis, once again there exists a subsequence of the latter one, denoted in the same way, such that $\Pi_{\mathcal{D}_m} u_m$ converges in $L^2(\Omega)$ to $\overline{u}$.

Step 3: proof that $\overline{u}$ is a solution to Problem (2.51).
This proof is done by passing to the limit in the gradient scheme (2.52), considering as test function the following interpolation of a given function $\varphi \in H_0^1(\Omega)$.
Let us define, for a given GD $\mathcal{D}$, $I_{\mathcal{D}} : H_0^1(\Omega) \to X_{\mathcal{D},0}$ by

$$I_{\mathcal{D}}\varphi = \operatorname*{argmin}_{v\in X_{\mathcal{D},0}} \left(\|\Pi_{\mathcal{D}} v - \varphi\|_{L^2(\Omega)} + \|\nabla_{\mathcal{D}} v - \nabla\varphi\|_{L^2(\Omega)^d} \right).$$

We have

$$\|\Pi_{\mathcal{D}}(I_{\mathcal{D}}\varphi) - \varphi\|_{L^2(\Omega)} + \|\nabla_{\mathcal{D}}(I_{\mathcal{D}}\varphi) - \nabla\varphi\|_{L^2(\Omega)^d} \leq S_{\mathcal{D}}(\varphi)$$

and therefore, by GD-consistency of the sequence $(\mathcal{D}_m)_{m\in\mathbb{N}}$, $\Pi_{\mathcal{D}_m}(I_{\mathcal{D}_m}\varphi) \to \varphi$ strongly in $L^2(\Omega)$ and $\nabla_{\mathcal{D}_m}(I_{\mathcal{D}_m}\varphi) \to \nabla\varphi$ strongly in $L^2(\Omega)^d$.
Using Lemma D.9 page 465 (non-linear strong convergence), we infer that $\Lambda(\cdot, \Pi_{\mathcal{D}_m} u_m)\nabla_{\mathcal{D}_m}(I_{\mathcal{D}_m}\varphi) \to \Lambda(\cdot, \overline{u})\nabla\varphi$ strongly in $L^2(\Omega)^d$. By symmetry of Λ and the weak-strong convergence property (Lemma D.8), this shows that

$$\begin{aligned}
&\int_\Omega \Lambda(\boldsymbol{x}, \Pi_{\mathcal{D}_m} u_m(\boldsymbol{x}))\nabla_{\mathcal{D}_m} u(\boldsymbol{x}) \cdot \nabla_{\mathcal{D}_m}(I_{\mathcal{D}_m}\varphi)(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \\
&\qquad = \int_\Omega \nabla_{\mathcal{D}_m} u(\boldsymbol{x}) \cdot [\Lambda(\boldsymbol{x}, \Pi_{\mathcal{D}_m} u_m(\boldsymbol{x}))\nabla_{\mathcal{D}_m}(I_{\mathcal{D}_m}\varphi)(\boldsymbol{x})]\,\mathrm{d}\boldsymbol{x} \\
&\qquad \to \int_\Omega \nabla\overline{u}(\boldsymbol{x}) \cdot [\Lambda(\boldsymbol{x}, \overline{u}(\boldsymbol{x}))\nabla\varphi(\boldsymbol{x})]\,\mathrm{d}\boldsymbol{x} \quad \text{as } m \to \infty \\
&\qquad = \int_\Omega \Lambda(\boldsymbol{x}, \overline{u}(\boldsymbol{x}))\nabla\overline{u}(\boldsymbol{x}) \cdot \nabla\varphi(\boldsymbol{x})\mathrm{d}\boldsymbol{x}. \qquad (2.57)
\end{aligned}$$

Moreover, since $\Pi_{\mathcal{D}_m}(I_{\mathcal{D}_m}\varphi) \to \varphi$ in $L^2(\Omega)$ as $m \to \infty$,

$$\int_\Omega f(\boldsymbol{x})\Pi_{\mathcal{D}_m}(I_{\mathcal{D}_m}\varphi)(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \to \int_\Omega f(\boldsymbol{x})\varphi(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \quad \text{as } m \to \infty. \qquad (2.58)$$

Letting $v = I_{\mathcal{D}_m}\varphi$ in (2.52), we can use (2.57) and (2.58) to pass to the limit and see that $\overline{u}$ is a solution to (2.51).

Step 4: strong convergence of $\nabla_{\mathcal{D}_m} u_m$.
Let now prove that $\nabla_{\mathcal{D}_m} u_m$ converges to $\nabla\overline{u}$ in $L^2(\Omega)^d$. We let $v = u_m$ in (2.52) and we pass to the limit in the right-hand side. Since $\overline{u}$ is a solution to (2.51), we obtain

$$\begin{aligned}
&\lim_{m\to\infty} \int_\Omega \Lambda(\boldsymbol{x}, \Pi_{\mathcal{D}_m} u_m(\boldsymbol{x}))\nabla_{\mathcal{D}_m} u_m(\boldsymbol{x}) \cdot \nabla_{\mathcal{D}_m} u_m(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \\
&\qquad = \int_\Omega f(\boldsymbol{x})\overline{u}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = \int_\Omega \Lambda(\boldsymbol{x}, \overline{u}(\boldsymbol{x}))\nabla\overline{u}(\boldsymbol{x}) \cdot \nabla\overline{u}(\boldsymbol{x})\mathrm{d}\boldsymbol{x}. \qquad (2.59)
\end{aligned}$$

We have

$$\begin{aligned}
&\int_\Omega \Lambda(\boldsymbol{x}, \Pi_{\mathcal{D}_m} u_m(\boldsymbol{x}))(\nabla_{\mathcal{D}_m} u_m(\boldsymbol{x}) - \nabla\overline{u}(\boldsymbol{x})) \cdot (\nabla_{\mathcal{D}_m} u_m(\boldsymbol{x}) - \nabla\overline{u}(\boldsymbol{x}))\mathrm{d}\boldsymbol{x} \\
&\quad = \int_\Omega \Lambda(\boldsymbol{x}, \Pi_{\mathcal{D}_m} u_m(\boldsymbol{x}))\nabla_{\mathcal{D}_m} u_m(\boldsymbol{x}) \cdot \nabla_{\mathcal{D}_m} u_m(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \\
&\qquad - \int_\Omega \Lambda(\boldsymbol{x}, \Pi_{\mathcal{D}_m} u_m(\boldsymbol{x}))\nabla_{\mathcal{D}_m} u_m(\boldsymbol{x}) \cdot \nabla\overline{u}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \\
&\qquad - \int_\Omega \Lambda(\boldsymbol{x}, \Pi_{\mathcal{D}_m} u_m(\boldsymbol{x}))\nabla\overline{u}(\boldsymbol{x}) \cdot (\nabla_{\mathcal{D}_m} u_m(\boldsymbol{x}) - \nabla\overline{u}(\boldsymbol{x}))\mathrm{d}\boldsymbol{x}. \qquad (2.60)
\end{aligned}$$

By (2.59), the weak convergence of $\nabla_{\mathcal{D}_m} u_m$, the strong convergence of $\Lambda(\cdot, \Pi_{\mathcal{D}_m} u_m)\nabla\overline{u}$ (obtained by non-linear strong convergence property, see Lemma D.9 page 465), and the weak-strong convergence lemma (Lemma D.8), we infer that

$$\begin{aligned}
&\int_\Omega \Lambda(\boldsymbol{x}, \Pi_{\mathcal{D}_m} u_m(\boldsymbol{x}))(\nabla_{\mathcal{D}_m} u_m(\boldsymbol{x}) - \nabla\overline{u}(\boldsymbol{x})) \cdot (\nabla_{\mathcal{D}_m} u_m(\boldsymbol{x}) - \nabla\overline{u}(\boldsymbol{x}))\mathrm{d}\boldsymbol{x} \\
&\qquad \to 0 \quad \text{as } m \to \infty.
\end{aligned}$$

The coercivity of Λ shows that the left-hand side is larger than

$$\underline{\lambda} \int_\Omega |\nabla_{\mathcal{D}_m} u_m(\boldsymbol{x}) - \nabla \overline{u}(\boldsymbol{x}))|^2 \mathrm{d}\boldsymbol{x}.$$

This quantity therefore converges to 0 and the proof of the strong $L^2(\Omega)$ convergence of the gradients is complete. ∎

2.1.5 Gradient schemes for p-Laplace type problems: $p \in (1, +\infty)$

This section is concerned with PDEs involving non-linearities with respect to the gradient of the unknown function, not just the unknown function itself as in Section 2.1.4. The p-Laplace problem is first considered, for which error estimates in terms of $W_{\mathcal{D}}$ and $S_{\mathcal{D}}$ are established in the same way as error estimates are provided in Section 2.1.2. Then, in Section 2.1.5, the case of a Leray–Lions type operator is considered, and the convergence of the GDM is proved for this model.
As mentioned in the introduction of this monograph, some of the techniques developed for the GDM are applicable outside this framework. As an illustration of this, the techniques used here to analyse the GDM for p-Laplace type problems and the discrete functional analysis ofB have been adapted in [57, 58] to a higher order numerical scheme which is not presented under the form of a gradient scheme.

An error estimate for the p-Laplace problem

We consider in this section the so-called p-Laplace equation:

$$-\operatorname{div}(|\nabla \overline{u}|^{p-2} \nabla \overline{u}) = f + \operatorname{div}(\boldsymbol{F}) \text{ in } \Omega, \tag{2.61a}$$

with boundary conditions

$$\overline{u} = 0 \text{ on } \partial\Omega, \tag{2.61b}$$

under the following assumptions:

- Ω is an open bounded connected subset of $\mathbb{R}^d$ ($d \in \mathbb{N}^\star$), (2.62a)
- $p \in (1, +\infty)$ (2.62b)
- $f \in L^{p'}(\Omega)$ and $\boldsymbol{F} \in L^{p'}(\Omega)^d$ with $p' = \dfrac{p}{p-1}$. (2.62c)

Under these hypotheses, the weak solution of (2.61) is the unique function $\overline{u}$ satisfying:

$$\begin{aligned} &\overline{u} \in W_0^{1,p}(\Omega) \text{ and, for all } v \in W_0^{1,p}(\Omega), \\ &\int_\Omega |\nabla \overline{u}|^{p-2} \nabla \overline{u}(\boldsymbol{x}) \cdot \nabla v(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \\ &\qquad = \int_\Omega f(\boldsymbol{x}) v(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} - \int_\Omega \boldsymbol{F}(\boldsymbol{x}) \cdot \nabla v(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}. \end{aligned} \tag{2.63}$$

Definition 2.36 (GS, p-Laplace problem). *Let $\mathcal{D} = (X_{\mathcal{D},0}, \Pi_{\mathcal{D}}, \nabla_{\mathcal{D}})$ be a GD in the sense of Definition 2.1. The corresponding gradient scheme for Problem (2.63) is defined by*

$$\begin{aligned}&\text{Find } u \in X_{\mathcal{D},0} \text{ such that, for any } v \in X_{\mathcal{D},0},\\ &\int_\Omega |\nabla_{\mathcal{D}} u(\boldsymbol{x})|^{p-2} \nabla_{\mathcal{D}} u(\boldsymbol{x}) \cdot \nabla_{\mathcal{D}} v(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}\\ &\qquad = \int_\Omega f(\boldsymbol{x}) \Pi_{\mathcal{D}} v(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} - \int_\Omega \boldsymbol{F}(\boldsymbol{x}) \cdot \nabla_{\mathcal{D}} v(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}.\end{aligned} \tag{2.64}$$

The following lemma establishes the existence and uniqueness of the solutions to (2.63) and (2.64), as well as estimates on these solutions.

Lemma 2.37. *Under Hypotheses (2.62), there exists one and only one solution to each of the problems (2.63) and (2.64). These solutions moreover satisfy*

$$\|\nabla \overline{u}\|_{L^p(\Omega)^d} \le (C_{P,p} \|f\|_{L^{p'}(\Omega)} + \|\boldsymbol{F}\|_{L^{p'}(\Omega)^d})^{\frac{1}{p-1}} \tag{2.65}$$

and

$$\|\nabla_{\mathcal{D}} u_{\mathcal{D}}\|_{L^p(\Omega)^d} \le (C_{\mathcal{D}} \|f\|_{L^{p'}(\Omega)} + \|\boldsymbol{F}\|_{L^{p'}(\Omega)^d})^{\frac{1}{p-1}}, \tag{2.66}$$

where $C_{P,p}$ is the continuous Poincaré's constant in $W_0^{1,p}(\Omega)$, and $C_{\mathcal{D}}$ is defined by (2.1).

Proof. The existence and uniqueness of $\overline{u}$ and $u_{\mathcal{D}}$ are obtained by noticing that (2.63) and (2.64) are respectively equivalent to the minimisation problems

$$\overline{u} \in \operatorname*{argmin}_{v \in W_0^{1,p}(\Omega)} \left(\frac{1}{p} \int_\Omega |\nabla v|^p \mathrm{d}\boldsymbol{x} - \int_\Omega f(\boldsymbol{x}) v(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} + \int_\Omega \boldsymbol{F}(\boldsymbol{x}) \cdot \nabla v(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \right) \tag{2.67}$$

and

$$\begin{aligned}u_{\mathcal{D}} \in \operatorname*{argmin}_{v \in X_{\mathcal{D},0}} \Big(&\frac{1}{p} \int_\Omega |\nabla_{\mathcal{D}} v(\boldsymbol{x})|^p \mathrm{d}\boldsymbol{x} - \\ &\int_\Omega f(\boldsymbol{x}) \Pi_{\mathcal{D}} v(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} + \int_\Omega \boldsymbol{F}(\boldsymbol{x}) \cdot \nabla_{\mathcal{D}} v(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \Big).\end{aligned} \tag{2.68}$$

This equivalence is a consequence of the inequality

$$\forall \chi, \xi \in \mathbb{R}^d,\ |\chi + \xi|^p - |\chi|^p - p|\chi|^{p-2} \chi \cdot \xi \ge 0,$$

which follows by writing that the convex mapping $H : \zeta \mapsto |\zeta|^p$ lies above its tangent at χ, and by noting that $\nabla H(\chi) = p|\chi|^{p-2}\chi$. The existence and uniqueness of the solutions to (2.67) and (2.68) are classical consequence of standard convex minimisation theorems, see, *e.g.*, [18].
Then inequalities (2.65) and (2.66) follow by taking, in each corresponding problem, the solution itself as a test function. ∎

Theorem 2.38 (Error estimate, p-Laplace problem). *Under Hypotheses (2.62), let $\overline{u} \in W_0^{1,p}(\Omega)$ be the solution of Problem (2.63), let $\mathcal{D}$ be a GD in the sense of Definition 2.1, and let $u_\mathcal{D} \in X_{\mathcal{D},0}$ be the solution to the gradient scheme (2.64). Then there exists $C_5 > 0$, depending only on p such that:*

1. *If $p \in (1,2]$,*

$$\|\nabla\overline{u} - \nabla_\mathcal{D} u_\mathcal{D}\|_{L^p(\Omega)^d} \le S_\mathcal{D}(\overline{u}) + C_5\big[W_\mathcal{D}(|\nabla\overline{u}|^{p-2}\nabla\overline{u} + \boldsymbol{F}) + S_\mathcal{D}(\overline{u})^{p-1}\big] \times \Big[S_\mathcal{D}(\overline{u})^p + \big[(C_\mathcal{D} + C_{P,p})\|f\|_{L^{p'}(\Omega)} + \|\boldsymbol{F}\|_{L^{p'}(\Omega)^d}\big]^{\frac{p}{p-1}}\Big]^{\frac{2-p}{2}}. \quad (2.69)$$

2. *If $p \in (2,+\infty)$,*

$$\|\nabla\overline{u} - \nabla_\mathcal{D} u_\mathcal{D}\|_{L^p(\Omega)^d} \le S_\mathcal{D}(\overline{u}) + C_5\Big[W_\mathcal{D}(|\nabla\overline{u}|^{p-2}\nabla\overline{u} + \boldsymbol{F}) + S_\mathcal{D}(\overline{u})\big[(C_{P,p}\|f\|_{L^{p'}(\Omega)} + \|\boldsymbol{F}\|_{L^{p'}(\Omega)^d})^{\frac{1}{p-1}} + S_\mathcal{D}(\overline{u})\big]^{p-2}\Big]^{\frac{1}{p-1}}. \quad (2.70)$$

As a consequence of (2.69)–(2.70), we have the following error estimate:

$$\|\overline{u} - \Pi_\mathcal{D} u_\mathcal{D}\|_{L^p(\Omega)} \le S_\mathcal{D}(\overline{u}) + C_\mathcal{D}(S_\mathcal{D}(\overline{u}) + \|\nabla\overline{u} - \nabla_\mathcal{D} u_\mathcal{D}\|_{L^p(\Omega)^d}). \quad (2.71)$$

Remark 2.39 (Link with space size of gradient schemes and lower bound). As in Remark 2.29 (for the case $p = 2$ and $d \le 3$), if $\overline{u} \in W^{2,p}(\Omega)$ and $|\nabla\overline{u}|^{p-2}\nabla\overline{u} + \boldsymbol{F} \in W^{1,p'}(\Omega)^d$, Theorem 2.38 gives an error estimate of the form

$$\begin{array}{ll} \mathcal{O}(h_\mathcal{D}^{p-1}) & \text{if } p \in (1,2], \\ \mathcal{O}(h_\mathcal{D}^{\frac{1}{p-1}}) & \text{if } p > 2, \end{array}$$

where $h_\mathcal{D} := h_\mathcal{D}(W^{2,p}(\Omega) \cap W_0^{1,p}(\Omega); W^{1,p'}(\Omega)^d)$ is the space size of the GD in the sense of Definition 2.22. Note that a lower bound in the spirit of (2.29) can also be derived.

Proof. Notice that, since $f \in L^{p'}(\Omega)$, the equation (2.61a) in the sense of distributions (*i.e.* taking $v \in C_c^\infty(\Omega)$ in (2.63)) shows that $\boldsymbol{\varphi} = |\nabla\overline{u}|^{p-2}\nabla\overline{u} + \boldsymbol{F}$ belongs to $W_{\mathrm{div}}^{p'}(\Omega)$ defined by (2.5), with $\mathrm{div}\boldsymbol{\varphi} = -f$. The function $\boldsymbol{\varphi}$ is therefore valid in the definition (2.6) of $W_\mathcal{D}$ and we obtain, for any $v \in X_{\mathcal{D},0}$, setting $\overline{W} = W_\mathcal{D}(|\nabla\overline{u}|^{p-2}\nabla\overline{u} + \boldsymbol{F})$,

$$\left|\int_\Omega \nabla_\mathcal{D} v(\boldsymbol{x}) \cdot (|\nabla\overline{u}(\boldsymbol{x})|^{p-2}\nabla\overline{u}(\boldsymbol{x}) + \boldsymbol{F}(\boldsymbol{x})) - \Pi_\mathcal{D} v(\boldsymbol{x}) f(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}\right| \le \|\nabla_\mathcal{D} v\|_{L^p(\Omega)^d}\, \overline{W}.$$

Use the fact that $u_\mathcal{D}$ satisfies (2.64) to replace the term $\Pi_\mathcal{D} v\, f$:

$$\left|\int_\Omega \nabla_\mathcal{D} v(\boldsymbol{x}) \cdot \big[|\nabla\overline{u}(\boldsymbol{x})|^{p-2}\nabla\overline{u}(\boldsymbol{x}) - |\nabla_\mathcal{D} u(\boldsymbol{x})|^{p-2}\nabla_\mathcal{D} u(\boldsymbol{x})\big] \mathrm{d}\boldsymbol{x}\right|$$

$$\leq \|\nabla_{\mathcal{D}} v\|_{L^p(\Omega)^d}\, \overline{W}.$$

Defining

$$I_{\mathcal{D}}\overline{u} = \operatorname*{argmin}_{w\in X_{\mathcal{D},0}} (\|\Pi_{\mathcal{D}} w - \overline{u}\|_{L^p(\Omega)} + \|\nabla_{\mathcal{D}} w - \nabla\overline{u}\|_{L^p(\Omega)^d}),$$

we obtain

$$\begin{aligned}
&A(v):= \\
&\quad \left|\int_\Omega \nabla_{\mathcal{D}} v(\boldsymbol{x}) \cdot \big[|\nabla_{\mathcal{D}} I_{\mathcal{D}}\overline{u}(\boldsymbol{x})|^{p-2}\nabla_{\mathcal{D}} I_{\mathcal{D}}\overline{u}(\boldsymbol{x}) - |\nabla_{\mathcal{D}} u(\boldsymbol{x})|^{p-2}\nabla_{\mathcal{D}} u_{\mathcal{D}}(\boldsymbol{x})\big]\mathrm{d}\boldsymbol{x}\right| \\
&\quad \leq \|\nabla_{\mathcal{D}} v\|_{L^p(\Omega)^d}\, \overline{W} \\
&\qquad + \left|\int_\Omega \nabla_{\mathcal{D}} v(\boldsymbol{x}) \cdot \big[|\nabla_{\mathcal{D}} I_{\mathcal{D}}\overline{u}(\boldsymbol{x})|^{p-2}\nabla_{\mathcal{D}} I_{\mathcal{D}}\overline{u}(\boldsymbol{x}) - |\nabla\overline{u}(\boldsymbol{x})|^{p-2}\nabla\overline{u}(\boldsymbol{x})\big]\mathrm{d}\boldsymbol{x}\right| \\
&\quad \leq \|\nabla_{\mathcal{D}} v\|_{L^p(\Omega)^d}\Big[\overline{W} + \|\,|\nabla_{\mathcal{D}} I_{\mathcal{D}}\overline{u}|^{p-2}\nabla_{\mathcal{D}} I_{\mathcal{D}}\overline{u} - |\nabla\overline{u}|^{p-2}\nabla\overline{u}\|_{L^{p'}(\Omega)}\Big].
\end{aligned}$$

Case $p \in (1,2]$.
Thanks to (2.75) in Lemma 2.40 below, we get the existence of C_6 depending only on p such that

$$\|\,|\nabla_{\mathcal{D}} I_{\mathcal{D}}\overline{u}|^{p-2}\nabla_{\mathcal{D}} I_{\mathcal{D}}\overline{u} - |\nabla\overline{u}|^{p-2}\nabla\overline{u}\|^{p'}_{L^{p'}(\Omega)^d} \leq C_6\|\nabla_{\mathcal{D}} I_{\mathcal{D}}\overline{u} - \nabla\overline{u}\|^p_{L^p(\Omega)^d}.$$

Setting $\overline{S} = S_{\mathcal{D}}(\overline{u})$ defined by (2.2), this leads to

$$A(v) \leq \|\nabla_{\mathcal{D}} v\|_{L^p(\Omega)^d}\Big[\overline{W} + C_6^{p-1}\overline{S}^{p-1}\Big]. \tag{2.72}$$

We then apply (2.77) in Lemma 2.40 with $\xi = \nabla_{\mathcal{D}} I_{\mathcal{D}}\overline{u}$ and $\chi = \nabla_{\mathcal{D}} u_{\mathcal{D}}$, and use Hölder's inequality with exponents $2/p$ and $2/(2-p)$. Taking $v = I_{\mathcal{D}}\overline{u} - u_{\mathcal{D}}$, we get C_7 depending only on p such that

$$\begin{aligned}
&\|\nabla_{\mathcal{D}} I_{\mathcal{D}}\overline{u} - \nabla_{\mathcal{D}} u_{\mathcal{D}}\|^p_{L^p(\Omega)^d} \\
&\qquad\qquad \leq C_7 A(I_{\mathcal{D}}\overline{u} - u_{\mathcal{D}})^{\frac{p}{2}}(\|\nabla_{\mathcal{D}} I_{\mathcal{D}}\overline{u}\|^p_{L^p(\Omega)^d} + \|\nabla_{\mathcal{D}} u_{\mathcal{D}}\|^p_{L^p(\Omega)^d})^{\frac{2-p}{2}},
\end{aligned}$$

and thus

$$\begin{aligned}
&\|\nabla_{\mathcal{D}} I_{\mathcal{D}}\overline{u} - \nabla_{\mathcal{D}} u_{\mathcal{D}}\|^2_{L^p(\Omega)^d} \\
&\qquad\qquad \leq C_7^{\frac{2}{p}} A(I_{\mathcal{D}}\overline{u} - u_{\mathcal{D}})(\|\nabla_{\mathcal{D}} I_{\mathcal{D}}\overline{u}\|^p_{L^p(\Omega)^d} + \|\nabla_{\mathcal{D}} u_{\mathcal{D}}\|^p_{L^p(\Omega)^d})^{\frac{2-p}{p}}.
\end{aligned}$$

Plugging (2.72) into this estimate gives C_8 depending only on p such that

$$\|\nabla_{\mathcal{D}} I_{\mathcal{D}}\overline{u} - \nabla_{\mathcal{D}} u_{\mathcal{D}}\|_{L^p(\Omega)^d} \leq C_8\big[\overline{W} + \overline{S}^{p-1}\big]$$

$$\times \left[\|\nabla_{\mathcal{D}} I_{\mathcal{D}} \overline{u}\|^p_{L^p(\Omega)^d} + \|\nabla_{\mathcal{D}} u_{\mathcal{D}}\|^p_{L^p(\Omega)^d}\right]^{\frac{2-p}{2}}.$$

We have $\|\nabla \overline{u} - \nabla_{\mathcal{D}} u_{\mathcal{D}}\|_{L^p(\Omega)^d} \leq \overline{S} + \|\nabla_{\mathcal{D}} I_{\mathcal{D}} \overline{u} - \nabla_{\mathcal{D}} u_{\mathcal{D}}\|_{L^p(\Omega)^d}$, and Estimate (2.69) therefore follows from (2.65) and (2.66).

Case $p \in (2, +\infty)$.
We use (2.74) in Lemma 2.40, Hölder's inequality with exponents $p/p' = p-1$ and $\frac{p-1}{p-2}$, and $(a+b)^\theta \leq 2^\theta(a^\theta + b^\theta)$ with $\theta = \frac{p}{p-2}$, $a = |\nabla_{\mathcal{D}} I_{\mathcal{D}} \overline{u}|^{p-2}$ and $b = |\nabla \overline{u}|^{p-2}$. This gives C_9 depending only on p such that

$$\begin{aligned}
&\| \, |\nabla_{\mathcal{D}} I_{\mathcal{D}} \overline{u}|^{p-2} \nabla_{\mathcal{D}} I_{\mathcal{D}} \overline{u} - |\nabla \overline{u}|^{p-2} \nabla \overline{u}\|_{L^{p'}(\Omega)^d} \\
&\qquad \leq C_9 \|\nabla_{\mathcal{D}} I_{\mathcal{D}} \overline{u} - \nabla \overline{u}\|_{L^p(\Omega)^d} (\|\nabla_{\mathcal{D}} I_{\mathcal{D}} \overline{u}\|_{L^p(\Omega)^d} + \|\nabla \overline{u}\|_{L^p(\Omega)^d})^{p-2}.
\end{aligned}$$

This leads to

$$\begin{aligned}
&A(v) \leq \\
&\qquad \|\nabla_{\mathcal{D}} v\|_{L^p(\Omega)^d} \Big[\overline{W} + C_9 \overline{S} \big[\|\nabla_{\mathcal{D}} I_{\mathcal{D}} \overline{u}\|_{L^p(\Omega)^d} + \|\nabla \overline{u}\|_{L^p(\Omega)^d}\big]^{p-2}\Big]. \qquad (2.73)
\end{aligned}$$

As before, we take $v = I_{\mathcal{D}} \overline{u} - u_{\mathcal{D}}$. Thanks to (2.78) in Lemma 2.40, we get the existence of C_{10} depending only on p such that

$$\|\nabla_{\mathcal{D}} I_{\mathcal{D}} \overline{u} - \nabla_{\mathcal{D}} u_{\mathcal{D}}\|^p_{L^p(\Omega)^d} \leq C_{10} A(I_{\mathcal{D}} \overline{u} - u_{\mathcal{D}}).$$

Using (2.73) we infer

$$\|\nabla_{\mathcal{D}} I_{\mathcal{D}} \overline{u} - \nabla_{\mathcal{D}} u_{\mathcal{D}}\|^{p-1}_{L^p(\Omega)^d} \leq C_{10} \Big[\overline{W} + C_9 \overline{S} \big[\|\nabla \overline{u}\|_{L^p(\Omega)^d} + \overline{S}\big]^{p-2}\Big],$$

and the proof of (2.70) is complete by invoking (2.65). ∎

In the following lemma, we gather a few useful estimates.

Lemma 2.40. *Let $p \in (1, +\infty)$ and $d \in \mathbb{N}^\star$. Then*

$$\begin{aligned}
&\forall \xi, \chi \in \mathbb{R}^d, \\
&\qquad \big||\xi|^{p-2}\xi - |\chi|^{p-2}\chi\big| \leq \max(1, p-1)|\xi - \chi| \; (|\xi|^{p-2} + |\chi|^{p-2}), \qquad (2.74)
\end{aligned}$$

which implies

$$\forall p \in (1, 2], \; \forall \xi, \chi \in \mathbb{R}^d, \; \big||\xi|^{p-2}\xi - |\chi|^{p-2}\chi\big| \leq 5|\xi - \chi|^{p-1}. \qquad (2.75)$$

Moreover, setting $C_0(p) = \frac{2}{p-1}$ for $p \in (1, 2]$ and $C_0(p) = 2^{p-1}$ for $p > 2$, there holds

$$\begin{aligned}
&\forall \xi, \chi \in \mathbb{R}^d, \\
&\qquad C_0(p)(|\xi|^{p-2}\xi - |\chi|^{p-2}\chi) \cdot (\xi - \chi) \geq |\xi - \chi|^2 \; (|\xi| + |\chi|)^{p-2}, \qquad (2.76)
\end{aligned}$$

which implies

$$\forall p \in (1,2],\ \forall \xi, \chi \in \mathbb{R}^d,$$
$$|\xi - \chi|^p \leq \left(\frac{2}{p-1}(|\xi|^{p-2}\xi - |\chi|^{p-2}\chi)\cdot(\xi-\chi)\right)^{\frac{p}{2}} \times (2^{p-1}(|\xi|^p + |\chi|^p))^{\frac{2-p}{2}}, \tag{2.77}$$

and

$$\forall p \geq 2,\ \forall \xi, \chi \in \mathbb{R}^d,\ |\xi - \chi|^p \leq 2^{p-1}(|\xi|^{p-2}\xi - |\chi|^{p-2}\chi)\cdot(\xi-\chi). \tag{2.78}$$

Proof. Estimates (2.74) and (2.75) originally appeared in [20]. Let $H(\xi) = |\xi|^{p-2}\xi$. If $p \geq 2$ then $H \in C^1(\mathbb{R}^d)^d$ and $|DH(\xi)| \leq (p-1)|\xi|^{p-2}$ (where DH is the differential of H and $|DH|$ the norm induced by the Euclidean norm). Hence, for all $\xi, \chi \in \mathbb{R}^d$,

$$|H(\xi) - H(\chi)| \leq |\xi - \chi|(p-1)\max_{\zeta\in[\xi,\chi]}|\zeta|^{p-2}. \tag{2.79}$$

The proof of (2.74) is complete in the case $p \geq 2$ since the mapping $s \mapsto s^{p-2}$ is non-decreasing, and thus $\max_{\zeta\in[\xi,\chi]}|\zeta|^{p-2} = \max(|\xi|^{p-2}, |\chi|^{p-2}) \leq (|\xi|^{p-2} + |\chi|^{p-2})$.
If $p < 2$, (2.79) remains valid but does directly lead to (2.74). Without loss of generality, we can assume that $0 < |\chi| \leq |\xi|$ (the case where $\chi = 0$ is trivial since the right-hand side of (2.74) is then equal to $+\infty$). Let $\widetilde{\xi}$ be the point in $\mathbb{R}^d$ at the intersection of the segment (ξ, χ) and of the ball of centre 0 and radius $|\chi|$ (see Figure 2.1). Since $|\widetilde{\xi}| = |\chi|$ we have $|H(\widetilde{\xi}) - H(\chi)| = |\chi|^{p-2}|\widetilde{\xi} - \chi|$. Hence, by the triangle inequality and (2.79) between ξ and $\widetilde{\xi}$,

$$\begin{aligned}|H(\xi) - H(\chi)| &\leq |H(\xi) - H(\widetilde{\xi})| + |\chi|^{p-2}|\widetilde{\xi} - \chi| \\ &\leq |\xi - \widetilde{\xi}|(p-1)\max_{\zeta\in[\xi,\widetilde{\xi}]}|\zeta|^{p-2} + |\chi|^{p-2}|\widetilde{\xi} - \chi|.\end{aligned}$$

Since $p < 2$, $\max_{\zeta\in[\xi,\widetilde{\xi}]}|\zeta|^{p-2} = |\widetilde{\xi}|^{p-2} = |\chi|^{p-2}$ and therefore

$$|H(\xi) - H(\chi)| \leq \left[|\xi - \widetilde{\xi}| + |\widetilde{\xi} - \chi|\right]|\chi|^{p-2}.$$

The proof of (2.74) in the case $p < 2$ is complete by noticing that $|\xi - \widetilde{\xi}| + |\widetilde{\xi} - \chi| = |\xi - \chi|$.
Let us now prove (2.75). Let $\eta > 0$. If $|\xi|$ and $|\chi|$ belong to $[\eta, +\infty)$, by (2.74) we have

$$\left||\xi|^{p-2}\xi - |\chi|^{p-2}\chi\right| \leq 2\eta^{p-2}|\xi - \chi|. \tag{2.80}$$

Otherwise, assume that $|\chi| \in (0, \eta]$. We have

$$\left||\xi|^{p-2}\xi - |\chi|^{p-2}\chi\right| \leq |\xi|^{p-1} + \eta^{p-1} \leq (|\xi - \chi| + \eta)^{p-1} + \eta^{p-1}. \tag{2.81}$$

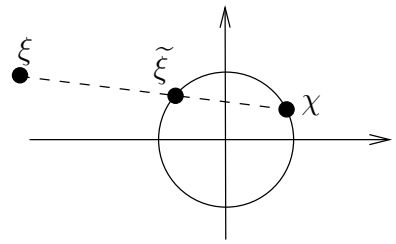


Fig. 2.1. Definition of the additional variable $\widetilde{\xi}$ for the case $p \in (1,2)$.

Combining (2.80) and (2.81) we see that, for all $\xi, \chi \in \mathbb{R}^d$ and all $\eta > 0$,

$$\big||\xi|^{p-2}\xi - |\chi|^{p-2}\chi\big| \le 2\eta^{p-2}|\xi - \chi| + (|\xi - \chi| + \eta)^{p-1} + \eta^{p-1}.$$

Estimate (2.75) follows by choosing $\eta = |\xi - \chi|$.
We now turn to the proof of (2.76). Set $A = (|\xi|^{p-2}\xi - |\chi|^{p-2}\chi) \cdot (\xi - \chi)$. By developing both sides we see that

$$A = (|\xi|^{p-1} - |\chi|^{p-1})(|\xi| - |\chi|) + (|\xi|^{p-2} + |\chi|^{p-2})(|\xi|\,|\chi| - \xi \cdot \chi). \tag{2.82}$$

Let us prove that the function $f(x) = x^{p-1} - y^{p-1} - \frac{2}{C_0(p)}(x+y)^{p-2}(x-y)$ satisfies $f'(x) \ge 0$ for all $x \ge y \ge 0$. We have

$$f'(x) = (p-1)x^{p-2} - \frac{2}{C_0(p)}(p-2)(x+y)^{p-3}(x-y) - \frac{2}{C_0(p)}(x+y)^{p-2}.$$

- If $1 < p \le 2$, we write, since $C_0(p) = 2/(p-1)$,

$$f'(x) \ge (p-1)x^{p-2} - \frac{2}{C_0(p)}x^{p-2} = 0.$$

- If $p > 2$,

$$f'(x) \ge (p-1)x^{p-2} - \frac{2}{C_0(p)}(p-2)(x+y)^{p-3}(x+y) - \frac{2}{C_0(p)}(x+y)^{p-2},$$

  and therefore, since $x \ge y$ and $C_0(p) = 2^{p-1}$,

$$f'(x) \ge (p-1)(x^{p-2} - \frac{2}{C_0(p)}(2x)^{p-2}) = 0.$$

In either case, since $f(y) = 0$, this shows that, if $x \ge y \ge 0$,

$$x^{p-1} - y^{p-1} \ge \frac{2}{C_0(p)}(x+y)^{p-2}(x-y).$$

Assuming (without loss of generality) that $|\xi| \ge |\chi|$ and applying the previous inequality to $x = |\xi|$ and $y = |\chi|$ gives

$$(|\xi|^{p-1} - |\chi|^{p-1})(|\xi| - |\chi|) \geq \frac{1}{C_0(p)}(|\xi| + |\chi|)^{p-2}(|\xi| - |\chi|)^2. \tag{2.83}$$

Let us again take generic numbers $x \geq y \geq 0$. If $1 < p \leq 2$ we can write

$$x^{p-2} + y^{p-2} \geq y^{p-2} \geq (x+y)^{p-2} \geq (p-1)(x+y)^{p-2} = \frac{2}{C_0(p)}(x+y)^{p-2}.$$

If $p > 2$ we have

$$x^{p-2} + y^{p-2} \geq x^{p-2} \geq 2^{2-p}(x+y)^{p-2} = \frac{2}{C_0(p)}(x+y)^{p-2}.$$

Applying these inequalities with $x = |\xi|$ and $y = |\chi|$, plugging the result in (2.82) and using (2.83) leads to

$$A \geq \frac{1}{C_0(p)}(|\xi| + |\chi|)^{p-2}\left[(|\xi| - |\chi|)^2 + 2(|\xi|\,|\chi| - \xi \cdot \chi)\right].$$

The proof of (2.76) is complete by writing

$$\begin{aligned}(|\xi| - |\chi|)^2 + 2(|\xi|\,|\chi| - \xi \cdot \chi) &= |\xi|^2 - 2|\xi|\,|\chi| + |\chi|^2 + 2|\xi|\,|\chi| - 2\xi \cdot \chi \\ &= |\xi|^2 - 2\xi \cdot \chi + |\chi|^2 = |\xi - \chi|^2.\end{aligned}$$

Estimate (2.77) is obtained by raising (2.76) to the power $p/2$ and by using $(|\xi| + |\chi|)^p \leq 2^{p-1}(|\xi|^p + |\chi|^p)$. Estimate (2.78) follows by writing $|\xi - \chi|^p = |\xi - \chi|^2|\xi - \chi|^{p-2} \leq |\xi - \chi|^2(|\xi| + |\chi|)^{p-2}$ and by using (2.76). ∎

Convergence of gradient schemes for Leray–Lions type problems

We now study the convergence of gradient schemes for the non-linear problem

$$\begin{aligned}&-\mathrm{div}\ \boldsymbol{a}(\boldsymbol{x}, \overline{u}, \nabla\overline{u}) = f \text{ in } \Omega, \\ &\overline{u} = 0 \text{ on } \partial\Omega,\end{aligned} \tag{2.84}$$

under the following assumptions:

- $p \in (1, \infty)$ and $\boldsymbol{a} : \Omega \times L^p(\Omega) \times \mathbb{R}^d \to \mathbb{R}^d$ is a Caratheodory function (2.85a)

(*i.e.* for a.e. $\boldsymbol{x} \in \Omega$ the function $(u, \boldsymbol{\xi}) \mapsto \boldsymbol{a}(\boldsymbol{x}, u, \boldsymbol{\xi})$ is continuous, and for any $(u, \boldsymbol{\xi}) \in L^p(\Omega) \times \mathbb{R}^d$ the function $\boldsymbol{x} \mapsto \boldsymbol{a}(\boldsymbol{x}, u, \boldsymbol{\xi})$ is measurable),

- $\exists \underline{a} \in (0, +\infty)$ such that $\boldsymbol{a}(\boldsymbol{x}, u, \boldsymbol{\xi}) \cdot \boldsymbol{\xi} \geq \underline{a}|\boldsymbol{\xi}|^p$ for a.e. $\boldsymbol{x} \in \Omega$, $\forall u \in L^p(\Omega),\ \forall \boldsymbol{\xi} \in \mathbb{R}^d$, (2.85b)
- $(\boldsymbol{a}(\boldsymbol{x}, u, \boldsymbol{\xi}) - \boldsymbol{a}(\boldsymbol{x}, u, \boldsymbol{\chi})) \cdot (\boldsymbol{\xi} - \boldsymbol{\chi}) \geq 0$ for a.e. $\boldsymbol{x} \in \Omega$,

$$\forall u \in L^p(\Omega),\ \forall \boldsymbol{\xi}, \boldsymbol{\chi} \in \mathbb{R}^d, \tag{2.85c}$$

- $\exists \overline{a} \in L^{p'}(\Omega)$, $\exists \mu \in (0, +\infty)$ such that $|\boldsymbol{a}(\boldsymbol{x}, u, \boldsymbol{\xi})| \leq \overline{a}(\boldsymbol{x}) + \mu |\boldsymbol{\xi}|^{p-1}$ for a.e. $\boldsymbol{x} \in \Omega$, $\forall u \in L^p(\Omega)$, $\forall \boldsymbol{\xi} \in \mathbb{R}^d$, (2.85d)
- $f \in L^{p'}(\Omega)$, where $p' = \dfrac{p}{p-1}$. (2.85e)

Remark 2.41 (Leray–Lions operator [126]). Note that the dependence of $\boldsymbol{a}$ on u is assumed to be non-local: $\boldsymbol{a}(\boldsymbol{x}, u, \cdot)$ depends on all the values of $u \in L^p(\Omega)$, not only on $u(\boldsymbol{x})$. These assumptions cover for example the case where $\boldsymbol{a}(\boldsymbol{x}, u, \nabla u(\boldsymbol{x})) = \Lambda[u](\boldsymbol{x}) \nabla u(\boldsymbol{x})$ with $\Lambda : L^p(\Omega) \to L^\infty(\Omega; \mathcal{S}_d(\mathbb{R}))$ as in [42, 65, 138].
These assumptions (in particular (2.85a)) do not cover the usual local dependencies $\boldsymbol{a}(\boldsymbol{x}, u(\boldsymbol{x}), \nabla u(\boldsymbol{x}))$ as in the non-monotone operators studied in [126]. However, the adaptation of the following results to this case is quite easy and more classical, see, *e.g.*, [68, 15, 29, 95, 57] for adaptations of the original Leray–Lions techniques to the convergence analysis of various schemes (some of them based the HMM method of Chapter 13, or the DDFV method presented in Section 14.2) for local non-monotone operators.

If a function $\boldsymbol{a}$ satisfies (2.85), then the mapping $u \mapsto -\mathrm{div}\, \boldsymbol{a}(\cdot, u, \nabla u)$ is called a Leray–Lions type operator. Such operators contain the p-Laplace operator as a particular case, obtained by setting $\boldsymbol{a}(\boldsymbol{x}, u, \boldsymbol{\xi}) = |\boldsymbol{\xi}|^{p-2}\boldsymbol{\xi}$. The existence of at least one solution to (2.84) is shown in [126] under hypotheses (2.85) in the case where $\boldsymbol{a}$ does not depend on u. In our framework, we say that a function $\overline{u}$ is a weak solution to (2.84) if:

$$\begin{aligned} &\overline{u} \in W_0^{1,p}(\Omega),\ \forall \overline{v} \in W_0^{1,p}(\Omega), \\ &\int_\Omega \boldsymbol{a}(\boldsymbol{x}, \overline{u}, \nabla \overline{u}(\boldsymbol{x})) \cdot \nabla \overline{v}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} = \int_\Omega f(\boldsymbol{x}) \overline{v}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}. \end{aligned} \tag{2.86}$$

Remark 2.42. Note that, even if $\boldsymbol{a}$ does not depend on $u \in L^p(\Omega)$, the solution to (2.86) is not necessarily unique. Consider the case where $d = 1$, $\Omega = (-1, 2)$, $f(x) = 0$ for $x \in (-1, 0) \cup (1, 2)$, $f(x) = 2$ for $x \in (0, 1)$ and

$$\boldsymbol{a}(\boldsymbol{x}, u, \boldsymbol{\xi}) = (\min(|\boldsymbol{\xi}|, 1) + \max(|\boldsymbol{\xi}| - 2, 0)) \frac{\boldsymbol{\xi}}{|\boldsymbol{\xi}|},\ \forall \boldsymbol{\xi} \in \mathbb{R},\ \forall u \in L^2(\Omega).$$

Then (2.85b) is satisfied with $\underline{a} = \frac{1}{2}$, (2.85c) is satisfied since $\boldsymbol{a}$ is non-decreasing with respect to $\boldsymbol{\xi}$ and (2.85d) is satisfied with $\overline{a}(\boldsymbol{x}) = 0$ and $\mu = 1$. Then the function $u(x) = \alpha(x+1)$ for $x \in (-1, 0)$, $\alpha + x(1-x)$ for $x \in (0, 1)$, $\alpha(2-x)$ for $x \in (1, 2)$ is solution to (2.86) for any value $\alpha \in [1, 2]$.

The hypothesis that $\boldsymbol{a}$ is strictly monotone, which may be expressed by

$$\begin{aligned} &(\boldsymbol{a}(\boldsymbol{x}, u, \boldsymbol{\xi}) - \boldsymbol{a}(\boldsymbol{x}, u, \boldsymbol{\chi})) \cdot (\boldsymbol{\xi} - \boldsymbol{\chi}) > 0, \\ &\text{for a.e. } \boldsymbol{x} \in \Omega,\ \forall u \in L^p(\Omega),\ \forall \boldsymbol{\xi}, \boldsymbol{\chi} \in \mathbb{R}^d \text{ with } \boldsymbol{\xi} \neq \boldsymbol{\chi}, \end{aligned} \tag{2.87}$$

is only used to prove the strong convergence of the reconstructed gradient (see theorem below). We now define the gradient scheme for Problem (2.84) .

Definition 2.43 (GS, Leray–Lions type problems). *If $\mathcal{D} = (X_{\mathcal{D},0}, \Pi_{\mathcal{D}}, \nabla_{\mathcal{D}})$ is a GD, then we define the related* gradient scheme *for* (2.84) *by*

$$\begin{aligned}&\textit{Find } u \in X_{\mathcal{D},0} \textit{ such that, } \forall v \in X_{\mathcal{D},0},\\ &\int_\Omega \boldsymbol{a}(\boldsymbol{x}, \Pi_{\mathcal{D}} u, \nabla_{\mathcal{D}} u(\boldsymbol{x})) \cdot \nabla_{\mathcal{D}} v(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} = \int_\Omega f(\boldsymbol{x}) \Pi_{\mathcal{D}} v(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}.\end{aligned} \tag{2.88}$$

Theorem 2.44 (Convergence, Leray–Lions type problems). *Under Assumptions* (2.85)*, take a sequence $(\mathcal{D}_m)_{m\in\mathbb{N}}$ of GDs in the sense of Definition 2.1, which is GD-consistent, limit-conforming and compact in the sense of Definitions 2.4, 2.5 and 2.8.*
Then, for any $m \in \mathbb{N}$, there exists at least one $u_{\mathcal{D}_m} \in X_{\mathcal{D}_m,0}$ solution to the gradient scheme (2.88)*. Moreover, there exists $\overline{u}$ solution of* (2.86) *such that, up to a subsequence, $\Pi_{\mathcal{D}_m} u_{\mathcal{D}_m}$ converges strongly in $L^p(\Omega)$ to $\overline{u}$ and $\nabla_{\mathcal{D}_m} u_{\mathcal{D}_m}$ converges weakly in $L^p(\Omega)^d$ to $\nabla\overline{u}$ as $m \to \infty$.*
If we assume that the Leray–Lions type operator $\boldsymbol{a}$ is strictly monotone in the sense of (2.87)*, then $\nabla_{\mathcal{D}_m} u_{\mathcal{D}_m}$ converges strongly in $L^p(\Omega)^d$ to $\nabla\overline{u}$ as $m \to \infty$.*
In the case where the solution $\overline{u}$ of (2.86) *is unique, then the whole sequence converges to $\overline{u}$ as $m \to \infty$ in the above senses.*

Remark 2.45 (Existence of a solution to (2.86)*).* Since there exists at least one sequence of GDs which satisfies the assumptions above (see Remark 2.11), a by-product of this theorem is the existence of a solution $\overline{u}$ to (2.86).

Remark 2.46 (Non-linearity without a lower order term)
In the case where $\boldsymbol{a}$ does not depend on $u \in L^p(\Omega)$, the proof of the weak convergence of $\Pi_{\mathcal{D}_m} u$ to a solution of (2.86) does not require the compactness of the sequence of GDs. In this case the strong convergence results from (2.87) (which gives the strong convergence of the reconstructed gradients) and from the coercivity and the GD-consistency of the sequence $(\mathcal{D}_m)_{m\in\mathbb{N}}$.

Proof.
This proof follows the same ideas as in [68, 95]. We start by noticing that, by Lemma 2.10, $(\mathcal{D}_m)_{m\in\mathbb{N}}$ is coercive in the sense of Definition 2.2.

Step 1: existence of a solution to the scheme.

Let $\mathcal{D}$ be a GD in the sense of Definition 2.1. We endow the finite dimensional space $X_{\mathcal{D},0}$ with an inner product $\langle\ ,\ \rangle$ and we denote by $|\cdot|$ its related norm. We define $F : X_{\mathcal{D},0} \to X_{\mathcal{D},0}$ as the function such that, if $u \in X_{\mathcal{D},0}$, $F(u)$ is the unique element in $X_{\mathcal{D},0}$ which satisfies

$$\forall v \in X_{\mathcal{D},0}\,, \quad \langle F(u), v\rangle = \int_\Omega \boldsymbol{a}(\boldsymbol{x}, \Pi_{\mathcal{D}} u, \nabla_{\mathcal{D}} u(\boldsymbol{x})) \cdot \nabla_{\mathcal{D}} v(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}.$$

Likewise, we denote by $w \in X_{\mathcal{D},0}$ the unique element such that

$$\forall v \in X_{\mathcal{D},0}\,, \quad \langle w, v\rangle = \int_\Omega f(\boldsymbol{x}) \Pi_{\mathcal{D}} v(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}.$$

The assumptions on $\boldsymbol{a}$ show that F is continuous and that, for all $u \in X_{\mathcal{D},0}$, $\langle F(u), u\rangle \geq \underline{a}\, \|\nabla_{\mathcal{D}} u\|^p_{L^p(\Omega)^d}$. By equivalence of the norms $|\cdot|$ and $\|\nabla_{\mathcal{D}}\cdot\|_{L^p(\Omega)^d}$ on $X_{\mathcal{D},0}$, we deduce that $\langle F(u), u\rangle \geq C_{11}|u|^p$ with C_{11} not depending on u. This shows that $\lim_{|u|\to\infty} \frac{\langle F(u),u\rangle}{|u|} = +\infty$ and thus that F is surjective (see [126] or [56, Theorem 3.3, page 19]). Note that we could as well use Theorem D.1, consequence of the topological degree. There exists therefore $u_{\mathcal{D}} \in X_{\mathcal{D},0}$ such that $F(u_{\mathcal{D}}) = w$, and this $u_{\mathcal{D}}$ is a solution to (2.88).

Step 2: convergence to a solution of the continuous problem.

Letting $v = u_{\mathcal{D}_m}$ in (2.88) with $\mathcal{D} = \mathcal{D}_m$ and using (2.1) and Hypothesis (2.85b), we get

$$\underline{a}\, \|\nabla_{\mathcal{D}_m} u_{\mathcal{D}_m}\|^{p-1}_{L^p(\Omega)^d} \leq C_{\mathcal{D}_m} \|f\|_{L^{p'}(\Omega)}.$$

Thanks to the coercivity of the sequence of GDs, this provides an estimate on $\nabla_{\mathcal{D}_m} u_{\mathcal{D}_m}$ in $L^p(\Omega)^d$ and on $\Pi_{\mathcal{D}_m} u_{\mathcal{D}_m}$ in $L^p(\Omega)$. Lemma 2.15 then gives $\overline{u} \in W_0^{1,p}(\Omega)$ such that, up to a subsequence, $\Pi_{\mathcal{D}_m} u_{\mathcal{D}_m} \to \overline{u}$ weakly in $L^p(\Omega)$ and $\nabla_{\mathcal{D}_m} u_{\mathcal{D}_m} \to \nabla\overline{u}$ weakly in $L^p(\Omega)^d$. By compactness of the sequence of GDs, we can also assume that the convergence of $\Pi_{\mathcal{D}_m} u_{\mathcal{D}_m}$ to $\overline{u}$ is strong in $L^p(\Omega)$ (this strong convergence property is only necessary for coping with the dependence of $\boldsymbol{a}$ with respect to u).
By Hypothesis (2.85d), the sequence of functions

$$\mathcal{A}_{\mathcal{D}_m}(\boldsymbol{x}) = \boldsymbol{a}(\boldsymbol{x}, \Pi_{\mathcal{D}_m} u_{\mathcal{D}_m}, \nabla_{\mathcal{D}_m} u_{\mathcal{D}_m}(\boldsymbol{x}))$$

remains bounded in $L^{p'}(\Omega)^d$ and converges therefore, up to a subsequence, to some $\boldsymbol{A}$ weakly in $L^{p'}(\Omega)^d$, as $m \to \infty$.
Let us now show that $\overline{u}$ is solution to (2.86), using the well-known Minty trick [131]. For a given $\varphi \in W_0^{1,p}(\Omega)$ and for any GD $\mathcal{D}$ belonging to the sequence $(\mathcal{D}_m)_{m\in\mathbb{N}}$, we introduce

$$I_{\mathcal{D}}\varphi = \operatorname*{argmin}_{v\in X_{\mathcal{D},0}} \left(\|\Pi_{\mathcal{D}} v - \varphi\|_{L^p(\Omega)} + \|\nabla_{\mathcal{D}} v - \nabla\varphi\|_{L^p(\Omega)^d}\right)$$

as a test function in (2.88). By the GD-consistency of $(\mathcal{D}_m)_{m\in\mathbb{N}}$, letting $m \to \infty$ we get

$$\int_\Omega \boldsymbol{A}(\boldsymbol{x}) \cdot \nabla\varphi(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} = \int_\Omega f(\boldsymbol{x})\varphi(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}, \quad \forall \varphi \in W_0^{1,p}(\Omega). \tag{2.89}$$

On the other hand, we may let $m \to \infty$ in (2.88) with $u_{\mathcal{D}_m}$ as a test function. Using (2.89) with $\varphi = \overline{u}$, this leads to

$$\begin{aligned}&\lim_{m\to\infty}\int_\Omega \boldsymbol{a}(\boldsymbol{x},\Pi_{\mathcal{D}_m}u_{\mathcal{D}_m},\nabla_{\mathcal{D}_m}u_{\mathcal{D}_m}(\boldsymbol{x}))\cdot\nabla_{\mathcal{D}_m}u_{\mathcal{D}_m}(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\\&=\int_\Omega f(\boldsymbol{x})\overline{u}(\boldsymbol{x})\mathrm{d}\boldsymbol{x}=\int_\Omega \boldsymbol{A}(\boldsymbol{x})\cdot\nabla\overline{u}(\boldsymbol{x})\mathrm{d}\boldsymbol{x}.\end{aligned}\tag{2.90}$$

Hypothesis (2.85c) gives, for any $\boldsymbol{G}\in L^p(\Omega)^d$,

$$\int_\Omega(\boldsymbol{a}(\boldsymbol{x},\Pi_{\mathcal{D}_m}u_{\mathcal{D}_m},\nabla_{\mathcal{D}_m}u_{\mathcal{D}_m}(\boldsymbol{x}))-\boldsymbol{a}(\boldsymbol{x},\Pi_{\mathcal{D}_m}u_{\mathcal{D}_m},\boldsymbol{G}(\boldsymbol{x})))\\ \cdot(\nabla_{\mathcal{D}_m}u_{\mathcal{D}_m}(\boldsymbol{x})-\boldsymbol{G}(\boldsymbol{x}))\mathrm{d}\boldsymbol{x}\geq 0.$$

Developing the preceding inequality, using Lemma D.8 for the weak-strong convergences and (2.90) for the convergence of the sole term involving a product of two weak convergences, we may let $m\to\infty$ and we get

$$\int_\Omega(\boldsymbol{A}(\boldsymbol{x})-\boldsymbol{a}(\boldsymbol{x},\overline{u},\boldsymbol{G}(\boldsymbol{x})))\cdot(\nabla\overline{u}(\boldsymbol{x})-\boldsymbol{G}(\boldsymbol{x}))\mathrm{d}\boldsymbol{x}\geq 0,\ \forall\boldsymbol{G}\in L^p(\Omega)^d.$$

We then set $\boldsymbol{G}=\nabla\overline{u}+\alpha\boldsymbol{\varphi}$ in the preceding inequality, where $\boldsymbol{\varphi}\in C_c^\infty(\Omega)^d$ and $\alpha>0$. Dividing by α, we get

$$-\int_\Omega(\boldsymbol{A}(\boldsymbol{x})-\boldsymbol{a}(\boldsymbol{x},\overline{u},\nabla\overline{u}(\boldsymbol{x})+\alpha\boldsymbol{\varphi}(\boldsymbol{x})))\cdot\boldsymbol{\varphi}(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\geq 0,\ \forall\boldsymbol{\varphi}\in C_c^\infty(\Omega)^d,\ \forall\alpha>0.$$

We then let $\alpha\to 0$ and use the dominated convergence theorem, which leads to

$$-\int_\Omega(\boldsymbol{A}(\boldsymbol{x})-\boldsymbol{a}(\boldsymbol{x},\overline{u},\nabla\overline{u}(\boldsymbol{x})))\cdot\boldsymbol{\varphi}(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\geq 0,\ \forall\boldsymbol{\varphi}\in C_c^\infty(\Omega)^d.$$

Changing $\boldsymbol{\varphi}$ into $-\boldsymbol{\varphi}$, we deduce that

$$\int_\Omega(\boldsymbol{A}(\boldsymbol{x})-\boldsymbol{a}(\boldsymbol{x},\overline{u},\nabla\overline{u}(\boldsymbol{x})))\cdot\boldsymbol{\varphi}(\boldsymbol{x})\mathrm{d}\boldsymbol{x}=0,\ \forall\boldsymbol{\varphi}\in C_c^\infty(\Omega)^d,$$

and therefore that

$$\boldsymbol{A}(\boldsymbol{x})=\boldsymbol{a}(\boldsymbol{x},\overline{u},\nabla\overline{u}(\boldsymbol{x})),\ \text{for a.e. }\boldsymbol{x}\in\Omega.\tag{2.91}$$

In addition to (2.89), this shows that $\overline{u}$ is a solution to (2.86). This concludes the proof of the convergence of $\Pi_{\mathcal{D}_m}u_{\mathcal{D}_m}$ to $\overline{u}$ in $L^p(\Omega)$ and of $\nabla_{\mathcal{D}_m}u_{\mathcal{D}_m}$ to $\nabla\overline{u}$ weakly in $L^p(\Omega)^d$ as $m\to\infty$.

Step 3: Assuming now Hypothesis (2.87), strong convergence of the reconstructed gradients.

We follow here the ideas of [126]. Thanks to (2.90) and (2.91), we get

$$\lim_{m\to\infty}\int_\Omega\Big[\boldsymbol{a}(\boldsymbol{x},\Pi_{\mathcal{D}_m}u_{\mathcal{D}_m}(\boldsymbol{x}),\nabla_{\mathcal{D}_m}u_{\mathcal{D}_m}(\boldsymbol{x}))-\boldsymbol{a}(\boldsymbol{x},\Pi_{\mathcal{D}_m}u_{\mathcal{D}_m}(\boldsymbol{x}),\nabla\overline{u}(\boldsymbol{x}))\Big]\\ \cdot\Big[\nabla_{\mathcal{D}_m}u_{\mathcal{D}_m}(\boldsymbol{x})-\nabla\overline{u}(\boldsymbol{x})\Big]\mathrm{d}\boldsymbol{x}=0.$$

Since the integrand is non-negative, this shows that

$$[\boldsymbol{a}(\cdot, \Pi_{\mathcal{D}_m} u_{\mathcal{D}_m}, \nabla_{\mathcal{D}_m} u_{\mathcal{D}_m}) - \boldsymbol{a}(\cdot, \Pi_{\mathcal{D}_m} u_{\mathcal{D}_m}, \nabla \overline{u})] \cdot [\nabla_{\mathcal{D}_m} u_{\mathcal{D}_m} - \nabla \overline{u}] \to 0 \text{ in } L^1(\Omega), \quad (2.92)$$

and therefore a.e. for a sub-sequence. Then, thanks to the strict monotonicity assumption (2.87), we may use Lemma 2.47 given below to show that $\nabla_{\mathcal{D}_m} u_{\mathcal{D}_m} \to \nabla \overline{u}$ a.e. as $m \to \infty$, at least for the same sub-sequence. This shows the a.e. convergence of $\boldsymbol{a}(\cdot, \Pi_{\mathcal{D}_m} u_{\mathcal{D}_m}, \nabla_{\mathcal{D}_m} u_{\mathcal{D}_m}) \cdot \nabla_{\mathcal{D}} u_{\mathcal{D}}$ to $\boldsymbol{a}(\cdot, \overline{u}, \nabla \overline{u}) \cdot \nabla \overline{u}$. We next recall that, by (2.90) and (2.91),

$$\lim_{m\to\infty} \int_\Omega \boldsymbol{a}(\boldsymbol{x}, \Pi_{\mathcal{D}_m} u_{\mathcal{D}_m}, \nabla_{\mathcal{D}_m} u_{\mathcal{D}_m}(\boldsymbol{x})) \cdot \nabla_{\mathcal{D}_m} u_{\mathcal{D}_m}(\boldsymbol{x}) d\boldsymbol{x} = \int_\Omega \boldsymbol{a}(\boldsymbol{x}, \overline{u}, \nabla \overline{u}(\boldsymbol{x})) \cdot \nabla \overline{u}(\boldsymbol{x}) d\boldsymbol{x}. \quad (2.93)$$

Since $\boldsymbol{a}(\cdot, \Pi_{\mathcal{D}_m} u_{\mathcal{D}_m}, \nabla_{\mathcal{D}_m} u_{\mathcal{D}_m}) \cdot \nabla_{\mathcal{D}_m} u_{\mathcal{D}_m} \geq 0$, we can apply Lemma 2.48 to get $\boldsymbol{a}(\cdot, \Pi_{\mathcal{D}_m} u_{\mathcal{D}_m}, \nabla_{\mathcal{D}_m} u_{\mathcal{D}_m}) \cdot \nabla_{\mathcal{D}_m} u_{\mathcal{D}_m} \to \boldsymbol{a}(\cdot, \overline{u}, \nabla \overline{u}) \cdot \nabla \overline{u}$ in $L^1(\Omega)$ as $m \to \infty$. This L^1-convergence gives the equi-integrability of the sequence of functions $\boldsymbol{a}(\cdot, \Pi_{\mathcal{D}_m} u_{\mathcal{D}_m}, \nabla_{\mathcal{D}_m} u_{\mathcal{D}_m}) \cdot \nabla_{\mathcal{D}_m} u_{\mathcal{D}_m}$, which gives in turn, thanks to (2.85b), the equi-integrability of $(|\nabla_{\mathcal{D}_m} u_{\mathcal{D}_m}|^p)_{m\in\mathbb{N}}$. The strong convergence of $\nabla_{\mathcal{D}_m} u_{\mathcal{D}_m}$ to $\nabla \overline{u}$ in $L^p(\Omega)^d$ is then a consequence of Vitali's theorem. ∎

Lemma 2.47. *Let B be a metric space, let $\boldsymbol{b}$ be a continuous function from $B \times \mathbb{R}^d$ to $\mathbb{R}^d$ such that*

$$(\boldsymbol{b}(u, \boldsymbol{\delta}) - \boldsymbol{b}(u, \boldsymbol{\gamma})) \cdot (\boldsymbol{\delta} - \boldsymbol{\gamma}) > 0, \ \forall \boldsymbol{\delta} \neq \boldsymbol{\gamma} \in \mathbb{R}^d, \ \forall u \in B.$$

Let $(u_m, \boldsymbol{\beta}_m)_{n\in\mathbb{N}}$ be a sequence in $B \times \mathbb{R}^d$ and $(u, \boldsymbol{\beta}) \in B \times \mathbb{R}^d$ be such that $(\boldsymbol{b}(u_m, \boldsymbol{\beta}_m) - \boldsymbol{b}(u_m, \boldsymbol{\beta})) \cdot (\boldsymbol{\beta}_m - \boldsymbol{\beta}) \to 0$ and $u_m \to u$ as $m \to \infty$. Then, $\boldsymbol{\beta}_m \to \boldsymbol{\beta}$ as $m \to \infty$.

Proof. We begin the proof with a preliminary remark. Let $\boldsymbol{\delta} \in \mathbb{R}^d \backslash \{0\}$. We define, for all $m \in \mathbb{N}$, the function $h_{\boldsymbol{\delta},m}$ from $\mathbb{R}$ to $\mathbb{R}$ by $h_{\boldsymbol{\delta},m}(s) = (\boldsymbol{b}(u_m, \boldsymbol{\beta} + s\boldsymbol{\delta}) - \boldsymbol{b}(u_m, \boldsymbol{\beta})) \cdot \boldsymbol{\delta}$. The hypothesis on $\boldsymbol{b}$ gives that $h_{\boldsymbol{\delta},m}$ is an increasing function since, for $s > s'$,

$$h_{\boldsymbol{\delta},m}(s) - h_{\boldsymbol{\delta},m}(s') = (\boldsymbol{b}(u_m, \boldsymbol{\beta} + s\boldsymbol{\delta}) - \boldsymbol{b}(u_m, \boldsymbol{\beta} + s'\boldsymbol{\delta})) \cdot \boldsymbol{\delta} > 0.$$

We prove now, by contradiction, that $\lim_{m\to\infty} \boldsymbol{\beta}_m = \boldsymbol{\beta}$. If the sequence $(\boldsymbol{\beta}_m)_{m\in\mathbb{N}}$ does not converge to $\boldsymbol{\beta}$, there exists $\varepsilon > 0$ and a subsequence, still denoted by $(\boldsymbol{\beta}_m)_{m\in\mathbb{N}}$, such that $s_m := |\boldsymbol{\beta}_m - \boldsymbol{\beta}| \geq \varepsilon$, for all $m \in \mathbb{N}$. Set $\boldsymbol{\delta}_m = \frac{\boldsymbol{\beta}_m - \boldsymbol{\beta}}{|\boldsymbol{\beta}_m - \boldsymbol{\beta}|}$. We can assume that, up to a subsequence, that $\boldsymbol{\delta}_m \to \boldsymbol{\delta}$ as $m \to \infty$, for some $\boldsymbol{\delta} \in \mathbb{R}^d$ with $|\boldsymbol{\delta}| = 1$. Then, since $s_m \geq \varepsilon$,

$$(\boldsymbol{b}(u_m, \boldsymbol{\beta}_m) - \boldsymbol{b}(u_m, \boldsymbol{\beta})) \cdot \frac{\boldsymbol{\beta}_m - \boldsymbol{\beta}}{s_m} = h_{\boldsymbol{\delta}_m,m}(s_m) \geq h_{\boldsymbol{\delta}_m,m}(\varepsilon)$$

$$= (\boldsymbol{b}(u_m, \boldsymbol{\beta} + \varepsilon \boldsymbol{\delta}_m) - \boldsymbol{b}(u_m, \boldsymbol{\beta})) \cdot \boldsymbol{\delta}_m.$$

Then, passing to the limit as $m \to \infty$,

$$\begin{aligned} 0 &= \lim_{m\to\infty} \frac{1}{s_m} (\boldsymbol{b}(u_m, \boldsymbol{\beta}_m) - \boldsymbol{b}(u_m, \boldsymbol{\beta})) \cdot (\boldsymbol{\beta}_m - \boldsymbol{\beta}) \\ &\geq (\boldsymbol{b}(u, \boldsymbol{\beta} + \varepsilon \boldsymbol{\delta}) - \boldsymbol{b}(u, \boldsymbol{\beta})) \cdot \boldsymbol{\delta} > 0, \end{aligned}$$

which is impossible. ■

The following result is classical (see [126]). Its proof is given for the sake of completeness.

Lemma 2.48. *Let $(F_m)_{m\in\mathbb{N}}$ be a sequence non-negative functions in $L^1(\Omega)$. Let $F \in L^1(\Omega)$ be such that $F_m \to F$ a.e. in Ω and $\int_\Omega F_m(\boldsymbol{x})d\boldsymbol{x} \to \int_\Omega F(\boldsymbol{x})d\boldsymbol{x}$, as $m \to \infty$. Then, $F_m \to F$ in $L^1(\Omega)$ as $m \to \infty$.*

Proof. The proof of this lemma is very classical. Applying the Dominated Convergence Theorem to the sequence $(F - F_m)^+$ leads to $\int_\Omega (F(\boldsymbol{x}) - F_m(\boldsymbol{x}))^+ d\boldsymbol{x} \to 0$ as $m \to \infty$. Then, since $|F - F_m| = 2(F - F_m)^+ - (F - F_m)$, we conclude that $F_m \to F$ in $L^1(\Omega)$ as $m \to \infty$. ■

2.2 Non-homogeneous Dirichlet boundary conditions

We now present GDs and GSs for non-homogeneous Dirichlet boundary conditions. To handle non-homogeneous boundary conditions, we need the concept of trace of functions in $W^{1,p}(\Omega)$, for $p \in (1, +\infty)$. The existence of the trace operator $\gamma : W^{1,p}(\Omega) \to W^{1-\frac{1}{p},p}(\partial\Omega)$ requires more regularity on Ω than in Section 2.1, and this open set is therefore assumed here to have a Lipschitz boundary.

2.2.1 Gradient discretisations

Definition 2.49 (GD, non-homogeneous Dirichlet BCs).
A gradient discretisation $\mathcal{D}$ for non-homogeneous Dirichlet conditions is defined by $\mathcal{D} = (X_\mathcal{D}, \mathcal{I}_{\mathcal{D},\partial}, \Pi_\mathcal{D}, \nabla_\mathcal{D})$ where:

1. *the set of discrete unknowns $X_\mathcal{D} = X_{\mathcal{D},0} \oplus X_{\mathcal{D},\partial}$ is the direct sum of two finite dimensional spaces on $\mathbb{R}$, corresponding respectively to the interior unknowns and to the boundary unknowns,*
2. *the linear mapping $\mathcal{I}_{\mathcal{D},\partial} : W^{1-\frac{1}{p},p}(\partial\Omega) \to X_{\mathcal{D},\partial}$ is an interpolation operator for the traces γu of the elements $u \in W^{1,p}(\Omega)$,*
3. *the function reconstruction $\Pi_\mathcal{D} : X_\mathcal{D} \to L^p(\Omega)$ is linear,*
4. *the gradient reconstruction $\nabla_\mathcal{D} : X_\mathcal{D} \to L^p(\Omega)^d$ is linear,*
5. *the operator $\nabla_\mathcal{D}$ is such that $\|\cdot\|_\mathcal{D} := \|\nabla_\mathcal{D}\cdot\|_{L^p(\Omega)^d}$ is a norm on $X_{\mathcal{D},0}$.*

Remark 2.50 (Domain of $\mathcal{I}_{\mathcal{D},\partial}$). The interpolation operator $\mathcal{I}_{\mathcal{D},\partial}$ does not necessarily need to be defined on the whole space $W^{1-\frac{1}{p},p}(\partial\Omega)$. If g is the boundary condition of the considered problem (*e.g.*, in (2.98b)), we only need to define $\mathcal{I}_{\mathcal{D},\partial}g$. Hence, if g has a better regularity than $W^{1-\frac{1}{p},p}(\partial\Omega)$, we can take advantage of this to find a simpler definition of $\mathcal{I}_{\mathcal{D},\partial}g$, see for example Remark 13.4.
In that case, the GD-consistency (Definition 2.51) is required only for functions $\varphi \in W^{1,p}(\Omega)$ such that $\gamma\varphi$ has the additional regularity supposed when constructing $\mathcal{I}_{\mathcal{D},\partial}$.

Coercivity, **limit-conformity**, **compactness** and **piecewise constant reconstructions** are defined as in the homogeneous case, by considering Definitions 2.2, 2.5, 2.8 and 2.12 on the spaces $X_{\mathcal{D},0}$. The definition of GD-consistency needs to be modified and implicitly imposes assumptions on the interpolation operator.

Definition 2.51 (GD-consistency, non-homogeneous Dirichlet BCs)

If $\mathcal{D}$ is a gradient discretisation in the sense of Definition 2.49, define $S_{\mathcal{D}} : W^{1,p}(\Omega) \to [0, +\infty)$ by

$$\begin{aligned} &\forall \varphi \in W^{1,p}(\Omega), \\ &S_{\mathcal{D}}(\varphi) = \min\Big\{ \|\Pi_{\mathcal{D}} v - \varphi\|_{L^p(\Omega)} + \|\nabla_{\mathcal{D}} v - \nabla\varphi\|_{L^p(\Omega)^d} : \\ &\qquad\qquad v \in X_{\mathcal{D}} \text{ such that } v - \mathcal{I}_{\mathcal{D},\partial}\gamma\varphi \in X_{\mathcal{D},0} \Big\}. \end{aligned} \tag{2.94}$$

A sequence $(\mathcal{D}_m)_{m\in\mathbb{N}}$ of gradient discretisations in the sense of Definition 2.49 is **GD-consistent** if

$$\forall \varphi \in W^{1,p}(\Omega), \lim_{m\to\infty} S_{\mathcal{D}_m}(\varphi) = 0. \tag{2.95}$$

Since coercivity, limit-conformity and compactness are the same as for homogeneous Dirichlet conditions, the characterisation Lemmas 2.7, 2.17 and 2.21 may also be used in the context of non-homogeneous Dirichlet conditions. Likewise, the limit-conformity or compactness of a sequence of GDs implies its coercivity, by Lemmata 2.6 and 2.10.
It will be useful, as in the homogeneous case, to also have a characterisation of the GD-consistency using dense subsets of $W^{1,p}(\Omega)$. This characterisation however requires an additional assumption on the trace interpolation operator, stating that for any given trace on $\partial\Omega$, we can find elements in $X_{\mathcal{D}}$ which interpolate this trace and have a norm controlled by this trace.

Lemma 2.52 (Equivalent condition for GD-consistency, non-homogeneous Dirichlet BCs). *Let $(\mathcal{D}_m)_{m\in\mathbb{N}}$ be a sequence of gradient discreti-*

sations in the sense of Definition 2.49. We assume that there exists C_{12} such that, for any $m \in \mathbb{N}$ and any $\varphi \in W^{1,p}(\Omega)$,

$$\begin{aligned}\min\{\|\Pi_{\mathcal{D}_m} v\|_{L^p(\Omega)} + \|\nabla_{\mathcal{D}_m} v\|_{L^p(\Omega)^d} \ &: v - \mathcal{I}_{\mathcal{D}_m,\partial}\gamma\varphi \in X_{\mathcal{D},0}\} \\ &\leq C_{12} \|\varphi\|_{W^{1,p}(\Omega)} .\end{aligned} \tag{2.96}$$

Then $(\mathcal{D}_m)_{m\in\mathbb{N}}$ is GD-consistent in the sense of Definition 2.51 if and only if there exists a dense subset W_s in $W^{1,p}(\Omega)$ such that

$$\forall \psi \in W_s \,, \ \lim_{m\to\infty} S_{\mathcal{D}_m}(\psi) = 0. \tag{2.97}$$

Remark 2.53. Note that (2.96) is almost a requirement to GD-consistency in the sense of Definition 2.51. Indeed, for any $\varphi \in W^{1,p}(\Omega)$, taking and element in $X_{\mathcal{D}} + \mathcal{I}_{\mathcal{D},\partial}\gamma\varphi \in X_{\mathcal{D},0}$ which realises the minimum $S_{\mathcal{D}}(\varphi)$, we see that

$$\begin{aligned}\min\{\|\Pi_{\mathcal{D}} v\|_{L^p(\Omega)} + \|\nabla_{\mathcal{D}} v\|_{L^p(\Omega)^d} \ , \ v - \mathcal{I}_{\mathcal{D},\partial}\gamma\varphi \in X_{\mathcal{D},0}\} \\ \leq S_{\mathcal{D}}(\varphi) + \|\varphi\|_{W^{1,p}(\Omega)} .\end{aligned}$$

Hence, if $(\mathcal{D}_m)_{m\in\mathbb{N}}$ is GD-consistent in the sense of Definition 2.51, estimate (2.96) is asymptotically true as $m \to \infty$ since $S_{\mathcal{D}_m}(\varphi) \to 0$.

Proof. The proof is very similar to the proof of Lemma 2.16 and we obviously only have to prove the "if" direction (the "only if" holds with $W_s = W^{1,p}(\Omega)$). Let $\varphi \in W^{1,p}(\Omega)$ and $\varepsilon > 0$. Take $\psi \in W_s$ such that $\|\varphi - \psi\|_{W^{1,p}(\Omega)} \leq \varepsilon$. Let $v \in X_{\mathcal{D}_m,0} + \mathcal{I}_{\mathcal{D}_m,\partial}\gamma\psi$ which realises the minimum in $S_{\mathcal{D}_m}(\psi)$ and let $w \in X_{\mathcal{D}_m,0} + \mathcal{I}_{\mathcal{D}_m,\partial}\gamma(\varphi - \psi)$ which realises the minimum for $\varphi - \psi$ in the left-hand side of (2.96). Then $v + w \in X_{\mathcal{D}_m,0} + \mathcal{I}_{\mathcal{D}_m,\partial}\gamma\varphi$ and, therefore,

$$\begin{aligned}S_{\mathcal{D}_m}(\varphi) &\leq \|\Pi_{\mathcal{D}_m}(v+w) - \varphi\|_{L^p(\Omega)} + \|\nabla_{\mathcal{D}_m}(v+w) - \nabla\varphi\|_{L^p(\Omega)^d} \\ &\leq \|\Pi_{\mathcal{D}_m} v - \psi\|_{L^p(\Omega)} + \|\nabla_{\mathcal{D}_m} v - \nabla\psi\|_{L^p(\Omega)^d} \\ &\quad + \|\Pi_{\mathcal{D}_m} w\|_{L^p(\Omega)} + \|\nabla_{\mathcal{D}_m} w\|_{L^p(\Omega)^d} \\ &\quad + \|\varphi - \psi\|_{L^p(\Omega)} + \|\nabla\varphi - \nabla\psi\|_{L^p(\Omega)^d} \\ &\leq S_{\mathcal{D}_m}(\psi) + (C_{12}+1)\|\varphi - \psi\|_{W^{1,p}(\Omega)} \\ &\leq S_{\mathcal{D}_m}(\psi) + (C_{12}+1)\epsilon.\end{aligned}$$

The conclusion follows as in the proof of Lemma 2.16. ∎

The convergence properties imposed on the interpolant $\mathcal{I}_{\mathcal{D},\partial}$ are somewhat hidden in the definition of GD-consistency. The following lemma shows that the formulation (2.101) of gradient schemes for non-homogeneous Dirichlet conditions make sense (for linear as well as non-linear problems): sequences of solutions to the gradient schemes indeed converge, up to a subsequence, to a function that has the required trace on the boundary of Ω.

Lemma 2.54 (Regularity of the limit, non-homogeneous Dirichlet BCs). *Let $(\mathcal{D}_m)_{m\in\mathbb{N}}$ be a sequence of gradient discretisations in the sense of Definition 2.49, that is limit-conforming (Definition 2.5) and GD-consistent (Definition 2.51). Let $g \in W^{1-\frac{1}{p},p}(\partial\Omega)$. Let $u_m \in X_{\mathcal{D}_m}$ be such that $u_m - \mathcal{I}_{\mathcal{D}_m,\partial}g \in X_{\mathcal{D}_m,0}$ and $(\|\nabla_{\mathcal{D}_m}u_m\|_{L^p(\Omega)^d})_{m\in\mathbb{N}}$ remains bounded. Then there exist a subsequence of $(\mathcal{D}_m, u_m)_{m\in\mathbb{N}}$, denoted in the same way, and $u \in W^{1,p}(\Omega)$ such that $\gamma u = g$ and, as $m \to \infty$, $\Pi_{\mathcal{D}_m}u_m$ converges weakly in $L^p(\Omega)$ to u and $\nabla_{\mathcal{D}_m}u_m$ converges weakly in $L^p(\Omega)^d$ to ∇u.*

Proof. Notice first that, by Lemma 2.6, $(\mathcal{D}_m)_{m\in\mathbb{N}}$ is coercive in the sense of Definition 2.2.
Let $\widetilde{g} \in W^{1,p}(\Omega)$ such that $\gamma\widetilde{g} = g$. By GD-consistency of $(\mathcal{D}_m)_{m\in\mathbb{N}}$, we can find $v_m \in X_{\mathcal{D}_m,0}+\mathcal{I}_{\mathcal{D}_m,\partial}g$ such that $\Pi_{\mathcal{D}_m}v_m \to \widetilde{g}$ in $L^p(\Omega)$ and $\nabla_{\mathcal{D}_m}v_m \to \nabla\widetilde{g}$ in $L^p(\Omega)^d$.
By assumption, $u_m - v_m = (u_m - \mathcal{I}_{\mathcal{D}_m,\partial}g) + (\mathcal{I}_{\mathcal{D}_m,\partial}g - v_m)$ belongs to $X_{\mathcal{D}_m,0}$, and $\|\nabla_{\mathcal{D}_m}(u_m - v_m)\|_{L^p(\Omega)^d}$ remains bounded. Hence, by recalling that the coercivity and limit-conformity of $(\mathcal{D}_m)_{m\in\mathbb{N}}$ are identical to the coercivity and limit-conformity of the underlying gradient discretisations for homogeneous Dirichlet conditions (*i.e.* with $X_{\mathcal{D}_m,0}$ instead of $X_{\mathcal{D}_m}$), Lemma 2.15 shows that, up to a subsequence, $\Pi_{\mathcal{D}_m}(u_m - v_m) \to \widetilde{u}$ weakly in $L^p(\Omega)$ and $\nabla_{\mathcal{D}_m}(u_m - v_m) \to \nabla\widetilde{u}$ weakly in $L^p(\Omega)^d$, where $\widetilde{u} \in W_0^{1,p}(\Omega)$.
The properties of $(v_m)_{m\in\mathbb{N}}$ then show that $\Pi_{\mathcal{D}_m}u_m = \Pi_{\mathcal{D}_m}(u_m - v_m) + \Pi_{\mathcal{D}_m}v_m \to \widetilde{u} + \widetilde{g} =: u$ in $L^p(\Omega)$, and $\nabla_{\mathcal{D}_m}u_m = \nabla_{\mathcal{D}_m}(u_m - v_m) + \nabla_{\mathcal{D}_m}v_m \to \nabla\widetilde{u} + \nabla\widetilde{g} = \nabla u$ in $L^p(\Omega)^d$. The function $u = \widetilde{u} + \widetilde{g}$ belongs to $W^{1,p}(\Omega)$ and has trace $\gamma u = \gamma\widetilde{u} + \gamma\widetilde{g} = 0 + g = g$. ∎

2.2.2 Gradient schemes for linear problems

Here, we take $p = 2$ in all the definitions of the previous section. We consider the linear problem defined in its strong form by:

$$-\operatorname{div}(\Lambda\nabla\overline{u}) = f + \operatorname{div}(\boldsymbol{F}) \text{ in } \Omega, \tag{2.98a}$$

with boundary conditions

$$\overline{u} = g \text{ on } \partial\Omega, \tag{2.98b}$$

under similar assumptions as in Section 2.1.2, adapted to the non-homogeneous BCs:

- Ω is an open bounded connected subset of $\mathbb{R}^d$ ($d \in \mathbb{N}^\star$), with Lipschitz boundary, (2.99a)
- Λ is a measurable function from Ω to the set of $d \times d$ symmetric matrices and there exists $\underline{\lambda}, \overline{\lambda} > 0$ such that, for a.e. $\boldsymbol{x} \in \Omega$, $\Lambda(\boldsymbol{x})$ has eigenvalues in $[\underline{\lambda}, \overline{\lambda}]$, (2.99b)

$$\bullet\ f \in L^2(\Omega)\,,\ \boldsymbol{F} \in L^2(\Omega)^d\,,\ g \in H^{1/2}(\partial\Omega). \tag{2.99c}$$

Under these hypotheses, the weak solution of (2.98) is the unique function $\overline{u}$ satisfying:

$$\begin{aligned}&\overline{u} \in \{w \in H^1(\Omega), \gamma(w) = g\},\ \forall v \in H_0^1(\Omega),\\ &\int_\Omega \Lambda(\boldsymbol{x})\nabla\overline{u}(\boldsymbol{x})\cdot\nabla v(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = \int_\Omega f(\boldsymbol{x})v(\boldsymbol{x})\mathrm{d}\boldsymbol{x} - \int_\Omega \boldsymbol{F}(\boldsymbol{x})\cdot\nabla v(\boldsymbol{x})\mathrm{d}\boldsymbol{x}.\end{aligned} \tag{2.100}$$

The GDM applied to Problem (2.100) yields the following gradient scheme .

Definition 2.55 (GS, non-homogeneous Dirichlet BCs). *If $\mathcal{D} = (X_\mathcal{D} = X_{\mathcal{D},0} \oplus X_{\mathcal{D},\partial}, \mathcal{I}_{\mathcal{D},\partial}, \Pi_\mathcal{D}, \nabla_\mathcal{D})$ is a gradient discretisation in the sense of Definition 2.49, then we define the related* gradient scheme *for* (2.100) *by*

$$\begin{aligned}&\textit{Find } u \in \mathcal{I}_{\mathcal{D},\partial}g + X_{\mathcal{D},0} \textit{ such that, for any } v \in X_{\mathcal{D},0},\\ &\int_\Omega \Lambda(\boldsymbol{x})\nabla_\mathcal{D}u(\boldsymbol{x})\cdot\nabla_\mathcal{D}v(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\\ &\qquad = \int_\Omega f(\boldsymbol{x})\Pi_\mathcal{D}v(\boldsymbol{x})\mathrm{d}\boldsymbol{x} - \int_\Omega \boldsymbol{F}(\boldsymbol{x})\cdot\nabla_\mathcal{D}v(\boldsymbol{x})\mathrm{d}\boldsymbol{x}.\end{aligned} \tag{2.101}$$

The following theorem states error estimates for the GS for non-homogeneous Dirichlet boundary conditions. This theorem yields a convergence result (not explicitly stated) similar to Corollary 2.31.

Theorem 2.56 (Error estimate, non-homogeneous Dirichlet BCs). *Under Hypotheses* (2.99)*, let $\overline{u} \in H^1(\Omega)$ be the solution of* (2.100) *(remark that since $f \in L^2(\Omega)$, one has $\Lambda\nabla\overline{u} + \boldsymbol{F} \in H_{\mathrm{div}}(\Omega)$).*
Let $\mathcal{D}$ be a GD in the sense of Definition 2.49. Then there exists one and only one $u_\mathcal{D} \in X_\mathcal{D}$ solution to the GS (2.101)*, and it satisfies the following inequalities:*

$$\|\nabla\overline{u} - \nabla_\mathcal{D}u_\mathcal{D}\|_{L^2(\Omega)^d} \le \frac{1}{\underline{\lambda}}\left[W_\mathcal{D}(\Lambda\nabla\overline{u} + \boldsymbol{F}) + (\overline{\lambda} + \underline{\lambda})S_\mathcal{D}(\overline{u})\right], \tag{2.102}$$

$$\|\overline{u} - \Pi_\mathcal{D}u_\mathcal{D}\|_{L^2(\Omega)} \le \frac{1}{\underline{\lambda}}\left[C_\mathcal{D}W_\mathcal{D}(\Lambda\nabla\overline{u} + \boldsymbol{F}) + (C_\mathcal{D}\overline{\lambda} + \underline{\lambda})S_\mathcal{D}(\overline{u})\right], \tag{2.103}$$

where $C_\mathcal{D}$, $S_\mathcal{D}$ and $W_\mathcal{D}$ are defined by Definitions 2.2, 2.51 and 2.5.

Remark 2.57. A lower bound in the spirit of (2.29) could also be stated.

Proof. Reasoning as in the proof of Theorem 2.28, we arrive at (2.33) for any $v \in X_{\mathcal{D},0}$. We then define

$$I_\mathcal{D}\overline{u} = \operatorname*{argmin}_{w \in \mathcal{I}_{\mathcal{D},\partial}g + X_{\mathcal{D},0}} (\|\Pi_\mathcal{D}w - \overline{u}\|_{L^2(\Omega)} + \|\nabla_\mathcal{D}w - \nabla\overline{u}\|_{L^2(\Omega)^d}),$$

and we notice that, by definition (2.94) of $S_\mathcal{D}$, (2.35) is still valid. Moreover, the vector $v = I_\mathcal{D}\overline{u} - u_\mathcal{D}$ belongs to $\mathcal{I}_{\mathcal{D},\partial}g + X_{\mathcal{D},0} + (-\mathcal{I}_{\mathcal{D},\partial}g + X_{\mathcal{D},0}) = X_{\mathcal{D},0}$ and can therefore be used in (2.33). The rest of the proof is then exactly as in the proof of Theorem 2.28. ∎

2.2.3 Gradient schemes for quasi-linear problems

We still consider $p = 2$ in the definitions and results of Section 2.2.1, and we deal with (2.49a) with non-homogeneous Dirichlet boundary conditions, that is:

$$\begin{aligned} -\mathrm{div}(\Lambda(\boldsymbol{x}, \overline{u})\nabla\overline{u}) &= f \text{ in } \Omega, \\ \overline{u} &= g \text{ on } \partial\Omega. \end{aligned} \tag{2.104}$$

We consider the same assumptions as in Section 2.1.4, adapted to the non-homogeneous boundary conditions:

- Ω is an open bounded connected subset of $\mathbb{R}^d$ ($d \in \mathbb{N}^\star$), with Lipschitz boundary, (2.105a)
- Λ is a Caratheodory function from $\Omega \times \mathbb{R}$ to $\mathcal{M}_d(\mathbb{R})$, (*i.e.* $\Lambda(\boldsymbol{x}, s)$ is measurable w.r.t. $\boldsymbol{x}$ and continuous w.r.t. s), there exists $\underline{\lambda}, \overline{\lambda} > 0$ such that, for a.e. $\boldsymbol{x} \in \Omega$, for all $s \in \mathbb{R}$, $\Lambda(\boldsymbol{x}, s)$ is symmetric with eigenvalues in $[\underline{\lambda}, \overline{\lambda}]$, (2.105b)
- $f \in L^2(\Omega)\,,\ g \in H^{1/2}(\partial\Omega)$. (2.105c)

Under these hypotheses, a weak solution to (2.104) is a function $\overline{u}$ (not necessarily unique) satisfying:

$$\begin{aligned} &\overline{u} \in \{w \in H^1(\Omega), \gamma(w) = g\},\ \forall v \in H^1_0(\Omega), \\ &\int_\Omega \Lambda(\boldsymbol{x}, \overline{u}(\boldsymbol{x}))\nabla\overline{u}(\boldsymbol{x}) \cdot \nabla v(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = \int_\Omega f(\boldsymbol{x})v(\boldsymbol{x})\mathrm{d}\boldsymbol{x}. \end{aligned} \tag{2.106}$$

This problem is approximated by the following gradient scheme.

Definition 2.58 (GS, quasi-linear problem, non-homogeneous Dirichlet BCs). *If $\mathcal{D} = (X_\mathcal{D} = X_{\mathcal{D},0} \oplus X_{\mathcal{D},\partial}, \mathcal{I}_{\mathcal{D},\partial}, \Pi_\mathcal{D}, \nabla_\mathcal{D})$ is a GD in the sense of Definition 2.49, then we define the related gradient scheme for (2.106) by*

$$\begin{aligned} &\textit{Find } u \in \mathcal{I}_{\mathcal{D},\partial}g + X_{\mathcal{D},0} \textit{ such that, for any } v \in X_{\mathcal{D},0}, \\ &\int_\Omega \Lambda(\boldsymbol{x}, \Pi_\mathcal{D}u(\boldsymbol{x}))\nabla_\mathcal{D}u(\boldsymbol{x}) \cdot \nabla_\mathcal{D}v(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = \int_\Omega f(\boldsymbol{x})\Pi_\mathcal{D}v(\boldsymbol{x})\mathrm{d}\boldsymbol{x}. \end{aligned} \tag{2.107}$$

This scheme leads to a non-linear system of equations under the form $A(u)U = B$, similar to (2.53). We then have the following convergence result.

Theorem 2.59 (Convergence, quasi-linear problem, non-homogeneous Dirichlet BCs). *Under Assumptions (2.105), let $(\mathcal{D}_m)_{m\in\mathbb{N}}$ be a sequence of GDs in the sense of Definition 2.49, which is GD-consistent, limit-conforming and compact in the sense of Definitions 2.51, 2.5 and 2.8.*
Then, for any $m \in \mathbb{N}$, there exists at least one $u_m \in X_{\mathcal{D}_m}$ solution to the gradient scheme (2.107) and, up to a subsequence, $\Pi_{\mathcal{D}_m}u_m$ converges strongly

in $L^2(\Omega)$ to a solution $\overline{u}$ of (2.106) and $\nabla_{\mathcal{D}_m} u_m$ converges strongly in $L^2(\Omega)^d$ to $\nabla\overline{u}$ as $m \to \infty$.
In the case where the solution $\overline{u}$ of (2.106) is unique, then the whole sequence converges to $\overline{u}$ as $m \to \infty$ in the senses above.

Proof. Take a lifting $\bar{g} \in H^1(\Omega)$ of g, that is, $\bar{g}$ is such that $\gamma\bar{g} = g$. For a gradient discretisation $\mathcal{D}$ in the sense of Definition 2.49, define

$$I_{\mathcal{D}}\bar{g} = \underset{v \in \mathcal{I}_{\mathcal{D},\partial g} + X_{\mathcal{D},0}}{\operatorname{argmin}} \left(\|\Pi_{\mathcal{D}} v - \bar{g}\|_{L^2(\Omega)} + \|\nabla_{\mathcal{D}} v - \nabla\bar{g}\|_{L^2(\Omega)^d} \right).$$

Thanks to Definition 2.51, as $m \to \infty$, $\Pi_{\mathcal{D}_m} I_{\mathcal{D}_m}\bar{g}$ converges strongly in $L^2(\Omega)$ to $\bar{g}$ and $\nabla_{\mathcal{D}_m} I_{\mathcal{D}_m}\bar{g}$ converges strongly in $L^2(\Omega)^d$ to $\nabla\bar{g}$. Then, for any solution u to (2.107), writing $w = u - I_{\mathcal{D}}\bar{g} \in X_{\mathcal{D},0}$, we have

$$\begin{aligned}
&\forall v \in X_{\mathcal{D},0},\\
&\int_\Omega \Lambda(\boldsymbol{x}, \Pi_{\mathcal{D}}(w + I_{\mathcal{D}}\bar{g})(\boldsymbol{x}))\nabla_{\mathcal{D}} w(\boldsymbol{x}) \cdot \nabla_{\mathcal{D}} v(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\\
&= \int_\Omega f(\boldsymbol{x})\Pi_{\mathcal{D}} v(\boldsymbol{x})\mathrm{d}\boldsymbol{x} - \int_\Omega \Lambda(\boldsymbol{x}, \Pi_{\mathcal{D}}(w + I_{\mathcal{D}}\bar{g})(\boldsymbol{x}))\nabla_{\mathcal{D}} I_{\mathcal{D}}\bar{g}(\boldsymbol{x}) \cdot \nabla_{\mathcal{D}} v(\boldsymbol{x})\mathrm{d}\boldsymbol{x}.
\end{aligned}$$

The remaining of the proof is then similar to that of Theorem 2.35, reasoning on w instead of u. ∎

3

Neumann, Fourier and mixed boundary conditions

The previous chapter presents the concepts of GDM for Dirichlet boundary conditions. Here, we show how these concepts must be modified and/or expanded to deal with other kinds of boundary conditions, namely Neumann, Fourier (or Robin), and mixed Dirichlet/Neumann. Each section is concerned with one particular form of boundary condition, and its structure is similar to the presentation adopted for Dirichlet boundary conditions: the notion of GD for the considered boundary conditions is first presented; it is followed by an analysis of the convergence of the corresponding GSs for some (linear or quasi-linear) elliptic PDEs.
The particular choices of definitions, and some of the resulting lemmas, are consequences of the abstract setting presented in Appendix A. Assimilating this appendix beforehand is however not mandatory, unless the reader wants complete proofs of all the results presented here. Since these proofs follow the same ideas as for homogeneous Dirichlet boundary conditions, the interested reader can alternatively attempt to write self-contained proofs, without reference to the results in Appendix A.

In all this chapter, Ω is assumed to have a Lipschitz-continuous boundary, and $p \in (1, +\infty)$.

3.1 Neumann boundary conditions

3.1.1 Gradient discretisations

Homogeneous Neumann boundary conditions

With the choice of spaces and operators described in Section A.1.2, the definitions and results of Section A.2 lead, with $\mathrm{P}_{\mathcal{D}} = \Pi_{\mathcal{D}}$ and $\mathrm{G}_{\mathcal{D}} = \nabla_{\mathcal{D}}$, to the following concept of GDs for homogeneous Neumann boundary conditions.

Definition 3.1 (GD, homogeneous Neumann BCs). *A gradient discretisation $\mathcal{D}$ for homogeneous Neumann conditions is $\mathcal{D} = (X_{\mathcal{D}}, \Pi_{\mathcal{D}}, \nabla_{\mathcal{D}})$ where:*

1. *the set of discrete unknowns $X_{\mathcal{D}}$ is a finite dimensional vector space on $\mathbb{R}$,*
2. *the function reconstruction $\Pi_{\mathcal{D}} : X_{\mathcal{D}} \to L^p(\Omega)$ is linear,*
3. *the gradient reconstruction $\nabla_{\mathcal{D}} : X_{\mathcal{D}} \to L^p(\Omega)^d$ is linear,*
4. *the operators $\nabla_{\mathcal{D}}$ and $\Pi_{\mathcal{D}}$ are such that*

$$\|v\|_{\mathcal{D}} := \left(\|\nabla_{\mathcal{D}} v\|^p_{L^p(\Omega)^d} + \left| \int_\Omega \Pi_{\mathcal{D}} v(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \right|^p \right)^{1/p} \tag{3.1}$$

is a norm on $X_{\mathcal{D}}$.

Remark 3.2. The choice of $\|v\|_{\mathcal{D}}$, involving the integral of $\Pi_{\mathcal{D}} v$ rather that its $L^p(\Omega)$ norm, is justified by the way GSs for Neumann problems are written, and by the *a priori* estimates that can be established on the approximate solution (cf Section 3.1.3).

The following discrete properties of gradient discretisations for Neumann problems, ensure the convergence of the associated gradient schemes.

Definition 3.3 (Coercivity, homogeneous Neumann BCs)

If $\mathcal{D}$ is a gradient discretisation in the sense of Definition 3.1, define

$$C_{\mathcal{D}} = \max_{v \in X_{\mathcal{D}} \setminus \{0\}} \frac{\|\Pi_{\mathcal{D}} v\|_{L^p(\Omega)}}{\|v\|_{\mathcal{D}}}. \tag{3.2}$$

A sequence $(\mathcal{D}_m)_{m \in \mathbb{N}}$ of gradient discretisations in the sense of Definition 3.1 is **coercive** if there exists $C_P \in \mathbb{R}_+$ such that $C_{\mathcal{D}_m} \leq C_P$ for all $m \in \mathbb{N}$.

Definition 3.4 (GD-consistency, Neumann BCs)

If $\mathcal{D}$ is a gradient discretisation in the sense of Definition 3.1, define $S_{\mathcal{D}} : W^{1,p}(\Omega) \to [0, +\infty)$ by

$$\begin{aligned} &\forall \varphi \in W^{1,p}(\Omega)\,, \\ &S_{\mathcal{D}}(\varphi) = \min_{v \in X_{\mathcal{D}}} \left(\|\Pi_{\mathcal{D}} v - \varphi\|_{L^p(\Omega)} + \|\nabla_{\mathcal{D}} v - \nabla \varphi\|_{L^p(\Omega)^d} \right). \end{aligned} \tag{3.3}$$

A sequence $(\mathcal{D}_m)_{m \in \mathbb{N}}$ of gradient discretisations in the sense of Definition 3.1 is **GD-consistent** if

$$\forall \varphi \in W^{1,p}(\Omega), \lim_{m \to \infty} S_{\mathcal{D}_m}(\varphi) = 0. \tag{3.4}$$

Definition 3.5 (Limit-conformity, – of the schemes)

For $p \in (1, +\infty)$, let $p' = \frac{p}{p-1}$ and

$$W^{p'}_{\mathrm{div},0}(\Omega) = \{\boldsymbol{\varphi} \in L^{p'}(\Omega)^d \,:\, \mathrm{div}\boldsymbol{\varphi} \in L^{p'}(\Omega)\,,\ \gamma_{\mathbf{n}}(\boldsymbol{\varphi}) = 0\},$$

where $\gamma_{\mathbf{n}}(\boldsymbol{\varphi})$ is the normal trace of $\boldsymbol{\varphi}$ on $\partial\Omega$ (see Section 3.1.2). If $\mathcal{D}$ be a gradient discretisation in the sense of Definition 3.1, define $W_{\mathcal{D}}$: $W^{p'}_{\mathrm{div},0}(\Omega) \to [0, +\infty)$ by

$$\begin{aligned} &\forall \boldsymbol{\varphi} \in W^{p'}_{\mathrm{div},0}(\Omega)\,, \\ &W_{\mathcal{D}}(\boldsymbol{\varphi}) = \max_{v \in X_{\mathcal{D}} \setminus \{0\}} \frac{1}{\|v\|_{\mathcal{D}}} \left| \int_\Omega \nabla_{\mathcal{D}} v(\boldsymbol{x}) \cdot \boldsymbol{\varphi}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \right. \\ &\qquad\qquad \left. + \int_\Omega \Pi_{\mathcal{D}} v(\boldsymbol{x}) \mathrm{div}\boldsymbol{\varphi}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \right|. \end{aligned} \tag{3.5}$$

A sequence $(\mathcal{D}_m)_{m\in\mathbb{N}}$ of gradient discretisations in the sense of Definition 3.1 is **limit-conforming** if

$$\forall \boldsymbol{\varphi} \in W^{p'}_{\mathrm{div},0}(\Omega),\ \lim_{m\to\infty} W_{\mathcal{D}_m}(\boldsymbol{\varphi}) = 0. \tag{3.6}$$

Definition 3.6 (Compactness, homogeneous Neumann BCs)

A sequence $(\mathcal{D}_m)_{m\in\mathbb{N}}$ of gradient discretisations in the sense of Definition 3.1 is **compact** if, for any sequence $u_m \in X_{\mathcal{D}_m}$ such that $(\|u_m\|_{\mathcal{D}_m})_{m\in\mathbb{N}}$ is bounded, the sequence $(\Pi_{\mathcal{D}_m} u_m)_{m\in\mathbb{N}}$ is relatively compact in $L^p(\Omega)$.

Note that the definition of **piecewise constant reconstruction** for a gradient discretisation for homogeneous Neumann boundary conditions is the same as Definition 2.12, replacing the space $X_{\mathcal{D},0}$ by $X_{\mathcal{D}}$.

As in the case of Dirichlet boundary conditions (see Lemma 2.16), the GD-consistency (resp. the limit-conformity) of sequences of gradient discretisations in the case of homogeneous Neumann BCs needs only be checked on a dense subset of $W^{1,p}(\Omega)$ (resp. $W^{1,p'}_0(\Omega)$).

Lemma 3.7 (Equivalent condition for GD-consistency, Neumann BCs). *A sequence $(\mathcal{D}_m)_{m\in\mathbb{N}}$ of gradient discretisations is GD-consistent in the sense of Definition 3.4 if and only if there exists a dense subset W_s in $W^{1,p}(\Omega)$ such that*

$$\forall \varphi \in W_s\,,\ \lim_{m\to\infty} S_{\mathcal{D}_m}(\varphi) = 0.$$

Proof. Apply Lemma A.13 with the choices in Section A.1.2. ∎

As for Dirichlet BCs, the limit-conformity or compactness of a sequence of gradient discretisations imply its coercivity.

Lemma 3.8 (Limit-conformity implies coercivity, homogeneous Neumann BCs). *Any sequence of gradient discretisations that is limit-conforming in the sense of Definition 3.5 is also coercive in the sense of Definition 3.3.*

Proof. Apply Lemma A.8 with the choices in Section A.1.2. ∎

Lemma 3.9 (Compactness implies coercivity, homogeneous Neumann BCs). *Any sequence of gradient discretisations that is compact in the sense of Definition 3.6 is also coercive in the sense of Definition 3.3.*

Proof. Apply Lemma A.16 with the choices in Section A.1.2. ∎

Lemma 2.21 states a compactness criterion in the case of Dirichlet boundary conditions; we state below a similar criterion for the case of Neumann boundary conditions, which holds under an additional regularity property on the domain Ω. Contrary to the case of Dirichlet boundary conditions, we may no longer extend the functions by 0 outside of Ω, and therefore the criterion can only involve "interior" translations

Lemma 3.10 (A criterion for compactness). *Let Ω be an open subset of $\mathbb{R}^d$ satisfying the "segment condition": there exist open sets $(U_i)_{i=1,\dots,k}$ and non-zero vectors $(\boldsymbol{\xi}_i)_{i=1,\dots,k}$ such that $\partial\Omega \subset \cup_{i=1}^k U_i$ and, for all $i = 1,\dots,k$ and all $t \in (0,1]$, $\overline{\Omega} \cap \overline{U_i} + t\boldsymbol{\xi}_i \subset \Omega$.*
Let $p \geq 1$ be given and $(u_m)_{m\in\mathbb{N}}$ be a bounded sequence in $L^p(\Omega)$ such that

$$\begin{array}{l} \lim_{|\boldsymbol{\xi}|\to 0} \sup_{m\in\mathbb{N}} \|u_m(\cdot+\boldsymbol{\xi}) - u_m\|_{L^p(\Omega_{\boldsymbol{\xi}})} = 0 \\ \textit{(where } \Omega_{\boldsymbol{\xi}} = \{\boldsymbol{x} \in \Omega, [\boldsymbol{x}, \boldsymbol{x}+\boldsymbol{\xi}] \subset \Omega\}\textit{).} \end{array} \tag{3.7}$$

Then $(u_m)_{m\in\mathbb{N}}$ is relatively compact in $L^p(\Omega)$.

Proof.
Let us first notice that, for any ω relatively compact in Ω, we have $\omega \subset \Omega_{\boldsymbol{\xi}}$ for $|\boldsymbol{\xi}|$ small enough. Hence, by the classical Kolmogorov compactness theorem, there exists $u \in L^p_{\text{loc}}(\Omega)$ and a subsequence, still denoted by $(u_m)_{m\in\mathbb{N}}$, such that $u_m \to u$ in $L^p_{\text{loc}}(\Omega)$. Since $(u_m)_{m\in\mathbb{N}}$ is bounded in $L^p(\Omega)$, Fatou's lemma shows that u belongs in fact to $L^p(\Omega)$. We infer that $(u_m - u)_{m\in\mathbb{N}}$ is bounded in $L^p(\Omega)$ and satisfies (3.7). Reasoning on $u_m - u$ rather than u, we can therefore assume that $u = 0$ and we have to prove that $u_m \to 0$ in $L^p(\Omega)$. The main issue is of course to estimate this convergence on a neighbourhood of $\partial\Omega$.
Let $(U_i)_{i=1,\dots,k}$ and $(\boldsymbol{\xi}_i)_{i=1,\dots,k}$ be given by the segment condition for Ω. For any $i \in \{1,\dots,k\}$ and any $r \in (0,1]$, $K_{i,r} = \overline{\Omega} \cap \overline{U_i} + r\boldsymbol{\xi}_i$ is a compact subset of Ω. Moreover, for any $m \in \mathbb{N}$, by the change of variable $\boldsymbol{y} = \boldsymbol{x} + r\boldsymbol{\xi}_i$, we get

$$\begin{aligned}\int_{\Omega\cap U_i}|u_m(\boldsymbol{x})|^p\mathrm{d}\boldsymbol{x} &\le \int_{\Omega\cap U_i+r\boldsymbol{\xi}_i}|u_m(\boldsymbol{y}-r\boldsymbol{\xi}_i)|^p\mathrm{d}\boldsymbol{y}\\ &\le 2^{p-1}\int_{\Omega\cap U_i+r\boldsymbol{\xi}_i}|u_m(\boldsymbol{y}-r\boldsymbol{\xi}_i)-u_m(\boldsymbol{y})|^p\mathrm{d}\boldsymbol{y}\\ &\quad +2^{p-1}\int_{K_{i,r}}|u_m(\boldsymbol{y})|^p\mathrm{d}\boldsymbol{y}.\end{aligned}$$

For any $z\in\Omega\cap U_i$ and any $s\in[0,1]$ we have $(z+r\boldsymbol{\xi}_i)-sr\boldsymbol{\xi}_i=z+(1-s)r\boldsymbol{\xi}_i\in\Omega$, by definition of z if $s=1$ and by definition of $\boldsymbol{\xi}_i$ if $s<1$. Hence, $\Omega\cap U_i+r\boldsymbol{\xi}_i\subset\Omega_{-r\boldsymbol{\xi}_i}$ and the preceding inequality gives

$$\int_{\Omega\cap U_i}|u_m(\boldsymbol{x})|^p\mathrm{d}\boldsymbol{x}\le 2^{p-1}\eta(-r\boldsymbol{\xi}_i)+2^{p-1}\int_{K_{i,r}}|u_m(\boldsymbol{y})|^p\mathrm{d}\boldsymbol{y}$$

where $\eta(\boldsymbol{\xi})=\sup_{m\in\mathbb{N}}\|u_m(\cdot+\boldsymbol{\xi})-u_m\|^p_{L^p(\Omega_{\boldsymbol{\xi}})}$ tends to 0 as $|\boldsymbol{\xi}|\to 0$. Summing all these inequalities on $i=1,\ldots,k$ and defining the open set $U=\cup_{i=1}^k U_i$, neighbourhood of $\partial\Omega$ in $\mathbb{R}^d$, we obtain

$$\int_{\Omega\cap U}|u_m(\boldsymbol{x})|^p\mathrm{d}\boldsymbol{x}\le 2^{p-1}\sum_{i=1}^k\eta(-r\boldsymbol{\xi}_i)+2^{p-1}\sum_{i=1}^k\int_{K_{i,r}}|u_m(\boldsymbol{y})|^p\mathrm{d}\boldsymbol{y}.$$

Let us now take $\varepsilon>0$ and fix $r\in(0,1]$ such that, for all $i=1,\ldots,k$, $\eta(-r\boldsymbol{\xi}_i)\le\varepsilon$. Since, for any i, $K_{i,r}$ is a compact subset of Ω we have $\int_{K_{i,r}}|u_m(\boldsymbol{y})|^p\mathrm{d}\boldsymbol{y}\to 0$ as $m\to\infty$ and therefore

$$\limsup_{m\to\infty}\int_{\Omega\cap U}|u_m(\boldsymbol{x})|^p\mathrm{d}\boldsymbol{x}\le 2^{p-1}k\varepsilon.$$

The proof is completed by letting $\varepsilon\to 0$. ∎

Non-homogeneous Neumann boundary conditions

The framework of gradient discretisations for diffusion problems with non-homogeneous Neumann boundary conditions is now presented. It corresponds to the abstract setting of Appendix A with the choice of spaces and operators described in Section A.1.3, provided that $\mathrm{P}_{\mathcal{D}}u=(\Pi_{\mathcal{D}}u,\mathbb{T}_{\mathcal{D}}u)$ and $\mathrm{G}_{\mathcal{D}}u=\nabla_{\mathcal{D}}u$.
We recall that Ω is a connected open bounded subset of $\mathbb{R}^d$ with Lipschitz boundary, and that $p\in(1,+\infty)$.

Definition 3.11 (GD, non-homogeneous Neumann BCs). *A gradient discretisation $\mathcal{D}$ for non-homogeneous Neumann BCs $\mathcal{D}$ is defined by $\mathcal{D}=(X_{\mathcal{D}},\Pi_{\mathcal{D}},\mathbb{T}_{\mathcal{D}},\nabla_{\mathcal{D}})$ where:*

1. *the set of discrete unknowns $X_{\mathcal{D}}$ is a finite dimensional vector space on $\mathbb{R}$,*

2. *the function reconstruction* $\Pi_{\mathcal{D}} : X_{\mathcal{D}} \to L^p(\Omega)$ *is linear,*
3. *the trace reconstruction* $\mathbb{T}_{\mathcal{D}} : X_{\mathcal{D}} \to L^p(\partial\Omega)$ *is linear; it provides, from an element of* $X_{\mathcal{D}}$*, a function over* $\partial\Omega$*,*
4. *the gradient reconstruction* $\nabla_{\mathcal{D}} : X_{\mathcal{D}} \to L^p(\Omega)^d$ *is linear,*
5. *the operators* $\nabla_{\mathcal{D}}$ *and* $\Pi_{\mathcal{D}}$ *are such that*

$$\|v\|_{\mathcal{D}} := \left(\|\nabla_{\mathcal{D}} v\|^p_{L^p(\Omega)^d} + \left| \int_\Omega \Pi_{\mathcal{D}} v(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \right|^p \right)^{1/p} \tag{3.8}$$

is a norm on $X_{\mathcal{D}}$*.*

The discrete properties of gradient discretisations for Neumann problems, that ensures the the convergence of the associated gradient schemes, are the following. The **GD-consistency** and **piecewise constant reconstruction** are still defined by Definitions 3.4 and 2.12 (replacing $X_{\mathcal{D},0}$ with $X_{\mathcal{D}}$ in the latter definition).

Remark 3.12 (Variant of the GD-consistency for non-homogeneous Neumann BCs) Following the abstract setting in Appendix A, with the choices made in Section A.1.3 the definition A.12 of GD-consistency would lead us to consider a different $S_{\mathcal{D}}$ than in Definition 3.4; namely, $S_{\mathcal{D}}$ would be defined by (3.52), as for Fourier BCs. Keeping the same definition (3.3) of $S_{\mathcal{D}}$ for non-homogeneous Neumann BCs as for homogeneous Neumann BCs offers a variant to the choice made in Appendix A. Note that Lemma A.13 (checking the GD-consistency using a dense subset of $W^{1,p}(\Omega)$) is still valid with this variant, upon minor modifications in the proof.
If approximating the trace is not a necessity, this definition (3.3) of $S_{\mathcal{D}}$ is sufficient to analyse gradient schemes for non-homogeneous Neumann problems. To recover a strong approximation of the trace, $S_{\mathcal{D}}$ needs to be changed into (3.52) – see Proposition 3.24.

Definition 3.13 (Coercivity, non-homogeneous Neumann BCs)

If $\mathcal{D}$ is a gradient discretisation in the sense of Definition 3.11, define

$$C_{\mathcal{D}} = \max_{v \in X_{\mathcal{D}} \setminus \{0\}} \left(\max \left\{ \frac{\|\Pi_{\mathcal{D}} v\|_{L^p(\Omega)}}{\|v\|_{\mathcal{D}}}, \frac{\|\mathbb{T}_{\mathcal{D}} v\|_{L^p(\partial\Omega)}}{\|v\|_{\mathcal{D}}} \right\} \right). \tag{3.9}$$

A sequence $(\mathcal{D}_m)_{m\in\mathbb{N}}$ of gradient discretisations in the sense of Definition 3.11 is **coercive** if there exists $C_P \in \mathbb{R}_+$ such that $C_{\mathcal{D}_m} \le C_P$ for all $m \in \mathbb{N}$.

Definition 3.14 (limit-conformity)

For $p \in (1, +\infty)$, let $p' = \frac{p}{p-1}$ and

$$W^{p'}_{\mathrm{div},\partial}(\Omega) = \{\boldsymbol{\varphi} \in L^{p'}(\Omega)^d \,:\, \mathrm{div}\boldsymbol{\varphi} \in L^{p'}(\Omega), \gamma_{\mathbf{n}}(\boldsymbol{\varphi}) \in L^{p'}(\partial\Omega)\}, \tag{3.10}$$

where $\gamma_{\mathbf{n}}(\boldsymbol{\varphi})$ is the normal trace of $\boldsymbol{\varphi}$ on $\partial\Omega$ (see Section 3.1.2). If $\mathcal{D}$ is a gradient discretisation in the sense of Definition 3.11, define $W_{\mathcal{D}}$: $W^{p'}_{\mathrm{div},\partial}(\Omega) \to [0, +\infty)$ by

$$\forall \boldsymbol{\varphi} \in W^{p'}_{\mathrm{div},\partial}(\Omega),$$
$$W_{\mathcal{D}}(\boldsymbol{\varphi}) = \max_{v \in X_{\mathcal{D}} \setminus \{0\}} \frac{1}{\|v\|_{\mathcal{D}}} \left| \int_\Omega (\nabla_{\mathcal{D}} v(\boldsymbol{x}) \cdot \boldsymbol{\varphi}(\boldsymbol{x}) + \Pi_{\mathcal{D}} v(\boldsymbol{x}) \mathrm{div}\boldsymbol{\varphi}(\boldsymbol{x}))\, \mathrm{d}\boldsymbol{x} - \int_{\partial\Omega} \mathbb{T}_{\mathcal{D}} v(\boldsymbol{x}) \gamma_{\mathbf{n}}(\boldsymbol{\varphi})(\boldsymbol{x}) \mathrm{d}\gamma(\boldsymbol{x}) \right|. \tag{3.11}$$

A sequence $(\mathcal{D}_m)_{m\in\mathbb{N}}$ of gradient discretisations in the sense of Definition 3.11 is **limit-conforming** if

$$\forall \boldsymbol{\varphi} \in W^{p'}_{\mathrm{div},\partial}(\Omega), \quad \lim_{m\to\infty} W_{\mathcal{D}_m}(\boldsymbol{\varphi}) = 0. \tag{3.12}$$

Note that in the case of Dirichlet boundary conditions, the definition of limit-conformity ensures that the dual operator to $\nabla_{\mathcal{D}_m}$ approximates the continuous divergence operator, whereas in the case of non-homogeneous Neumann boundary conditions, it ensures both this duality and the fact that $\mathbb{T}_{\mathcal{D}_m}$ approximates the continuous trace operator (see also Lemma 3.21 below).

Definition 3.15 (Compactness, non-homogeneous Neumann BCs)

A sequence $(\mathcal{D}_m)_{m\in\mathbb{N}}$ of gradient discretisations in the sense of Definition 3.11 is **compact** if, for any sequence $u_m \in X_{\mathcal{D}_m}$ such that $(\|u_m\|_{\mathcal{D}_m})_{m\in\mathbb{N}}$ is bounded, the sequences $(\Pi_{\mathcal{D}_m} u_m)_{m\in\mathbb{N}}$ and $(\mathbb{T}_{\mathcal{D}_m} u_m)_{m\in\mathbb{N}}$ are relatively compact in $L^p(\Omega)$ and $L^p(\partial\Omega)$, respectively.

Remark 3.16 (Alternative definition of compactness)
As for Dirichlet problems, compactness of the sequence of gradient discretisations is only useful when dealing with non-linearities in the PDE. If these non-linearities do not involve the trace of the solution, then the compactness property can be relaxed by only requesting the *weak* relative compactness of $(\mathbb{T}_{\mathcal{D}_m} u_m)_{m\in\mathbb{N}}$ in $L^p(\partial\Omega)$; see, *e.g.*, the proof of Theorem 3.35.

As for other boundary conditions, the limit-conformity or compactness imply the coercivity in the case of non-homogeneous Neumann BCs.

Lemma 3.17 (Limit-conformity implies coercivity, non-homogeneous Neumann BCs). *Any sequence of gradient discretisations that is limit-conforming in the sense of Definition 3.14 is also coercive in the sense of Definition 3.13.*

Proof. Apply Lemma A.8 with the choices in Section A.1.3. ∎

Lemma 3.18 (Compactness implies coercivity, non-homogeneous Neumann BCs). *Any sequence of gradient discretisations that is compact in the sense of Definition 3.15 is also coercive in the sense of Definition 3.13.*

Proof. Apply Lemma A.16 with the choices in Section A.1.3. ∎

Lemma 3.19 (Equivalent condition for limit-conformity, non-homogeneous Neumann BCs). *Let $(\mathcal{D}_m)$ be a sequence of sequence of gradient discretisations. Then $(\mathcal{D}_m)_{m\in\mathbb{N}}$ is limit-conforming in the sense of Definition 3.14 if and only if it is coercive in the sense of Definition 3.13, and there exists a dense subset $\boldsymbol{W}_w$ in $W^{p'}_{\mathrm{div},\partial}(\Omega)$ (endowed with the norm $\|\boldsymbol{\varphi}\|_{W^{p'}_{\mathrm{div},\partial}(\Omega)} = \|\boldsymbol{\varphi}\|_{L^{p'}(\Omega)^d} + \|\mathrm{div}\boldsymbol{\varphi}\|_{L^{p'}(\Omega)} + \|\gamma_{\mathbf{n}}(\boldsymbol{\varphi})\|_{L^{p'}(\partial\Omega)}$) such that*

$$\forall \boldsymbol{\varphi} \in \boldsymbol{W}_w,\ \lim_{m\to\infty} W_{\mathcal{D}_m}(\boldsymbol{\varphi}) = 0. \tag{3.13}$$

Proof. Apply Lemma A.9 with the choices in Section A.1.3. ∎

Remark 3.20 (A possible choice for the dense subset $\boldsymbol{W}_w$). Lemma 3.27 in Section 3.1.2 below shows that the set $\boldsymbol{W}_w = C^\infty(\mathbb{R}^d)^d$ is dense in $W^{p'}_{\mathrm{div},\partial}(\Omega)$ and is therefore an interesting choice to prove the limit-conformity.

The following lemma is the equivalent of Lemma 2.15 for non-homogeneous Neumann conditions.

Lemma 3.21 (Regularity of the limit, non-homogeneous Neumann BCs). *Let $(\mathcal{D}_m)_{m\in\mathbb{N}}$ be a limit-conforming sequence of gradient discretisations in the sense of Definition 3.14. Let $u_m \in X_{\mathcal{D}_m}$ be such that $(\|u_m\|_{\mathcal{D}_m})_{m\in\mathbb{N}}$ is bounded. Then there exists $u \in W^{1,p}(\Omega)$ such that, up to a subsequence,*

$$\Pi_{\mathcal{D}_m} u_m \to u \text{ weakly in } L^p(\Omega), \tag{3.14}$$

$$\mathbb{T}_{\mathcal{D}_m} u_m \to \gamma u \text{ weakly in } L^p(\partial\Omega), \tag{3.15}$$

$$\nabla_{\mathcal{D}_m} u_m \to \nabla u \text{ weakly in } L^p(\Omega)^d. \tag{3.16}$$

Proof. Apply Lemma A.11 with the choices in Section A.1.3.

■

Remark 3.22 (Convergence of the trace). This lemma shows in particular that, for a sequence $(\mathcal{D}_m)_{m\in\mathbb{N}}$ of limit-conforming GDs, if $u_m \in X_{\mathcal{D}_m}$ is such that, for some $u \in W^{1,p}(\Omega)$, $\Pi_{\mathcal{D}_m} u_m \to u$ weakly in $L^p(\Omega)$ and $\nabla_{\mathcal{D}_m} u_m \to \nabla u$ weakly in $L^p(\Omega)^d$, then $\mathbb{T}_{\mathcal{D}_m} u_m \to \gamma u$ weakly in $L^p(\partial\Omega)$.

Remark 3.23 (Homogeneous Neumann BCs). In the case of a gradient discretisation for homogeneous Neumann conditions, a trace reconstruction is not needed, and Lemma 3.21 is still valid with (3.15) removed.

We complete this section by stating an approximation property of $\mathbb{T}_{\mathcal{D}}$. This property is useful to deduce error estimates on the traces of gradient scheme approximations to linear elliptic problems (see Remark 3.31) .

Proposition 3.24 (Approximation property of $\mathbb{T}_{\mathcal{D}}$ – Neumann BCs). *Let $\mathcal{D}$ be a gradient discretisation in the sense of Definition 3.11. We define, for $\varphi \in W^{1,p}(\Omega)$,*

$$\overline{S}_{\mathcal{D}}(\varphi) = \min_{w\in X_{\mathcal{D}}} \Big(\|\Pi_{\mathcal{D}} w - \varphi\|_{L^p(\Omega)} + \|\mathbb{T}_{\mathcal{D}} w - \gamma\varphi\|_{L^p(\partial\Omega)} + \|\nabla_{\mathcal{D}} w - \nabla\varphi\|_{L^p(\Omega)^d} \Big). \tag{3.17}$$

Then, for any $v \in X_{\mathcal{D}}$ and any $\varphi \in W^{1,p}(\Omega)$,

$$\|\mathbb{T}_{\mathcal{D}} v - \gamma\varphi\|_{L^p(\partial\Omega)} \le C_{\mathcal{D}} \left(|\Omega|^{\frac{1}{p'}} \|\Pi_{\mathcal{D}} v - \varphi\|_{L^p(\Omega)} + \|\nabla_{\mathcal{D}} v - \nabla\varphi\|_{L^p(\Omega)^d} \right) + \max\left(1, C_{\mathcal{D}}, C_{\mathcal{D}}|\Omega|^{\frac{1}{p'}}\right) \overline{S}_{\mathcal{D}}(\varphi).$$

Remark 3.25. The quantity $\overline{S}_{\mathcal{D}}$ in (3.17) is actually the measure of the GD-consistency for Fourier boundary conditions, see (3.52).

Proof. We introduce

$$I_{\mathcal{D}}\varphi = \operatorname*{argmin}_{w\in X_{\mathcal{D}}} \Big(\|\Pi_{\mathcal{D}} w - \varphi\|_{L^p(\Omega)} + \|\mathbb{T}_{\mathcal{D}} w - \gamma\varphi\|_{L^p(\partial\Omega)} + \|\nabla_{\mathcal{D}} w - \nabla\varphi\|_{L^p(\Omega)^d} \Big)$$

and we notice that

$$\|\Pi_{\mathcal{D}} I_{\mathcal{D}}\varphi - \varphi\|_{L^p(\Omega)} + \|\mathbb{T}_{\mathcal{D}} I_{\mathcal{D}}\varphi - \gamma\varphi\|_{L^p(\partial\Omega)} + \|\nabla_{\mathcal{D}} I_{\mathcal{D}}\varphi - \nabla\varphi\|_{L^p(\Omega)^d} \le \overline{S}_{\mathcal{D}}(\varphi). \tag{3.18}$$

By definition of $C_{\mathcal{D}}$ and of $\|\cdot\|_{\mathcal{D}}$, Hölder's inequality gives, for all $w \in X_{\mathcal{D}}$,

$$\|\mathbb{T}_{\mathcal{D}} w\|_{L^p(\partial\Omega)} \leq C_{\mathcal{D}} \left(\|\nabla_{\mathcal{D}} w\|_{L^p(\Omega)^d} + |\Omega|^{\frac{1}{p'}} \|\Pi_{\mathcal{D}} w\|_{L^p(\Omega)} \right). \tag{3.19}$$

A triangle inequality therefore provides

$$\begin{aligned}
&\|\mathbb{T}_{\mathcal{D}} v - \gamma\varphi\|_{L^p(\partial\Omega)} \\
&\quad \leq \|\mathbb{T}_{\mathcal{D}}(v - I_{\mathcal{D}}\varphi)\|_{L^p(\partial\Omega)} + \|\mathbb{T}_{\mathcal{D}} I_{\mathcal{D}}\varphi - \gamma\varphi\|_{L^p(\partial\Omega)} \\
&\quad \leq C_{\mathcal{D}} \left(|\Omega|^{\frac{1}{p'}} \|\Pi_{\mathcal{D}} v - \Pi_{\mathcal{D}} I_{\mathcal{D}}\varphi\|_{L^p(\Omega)} + \|\nabla_{\mathcal{D}} v - \nabla_{\mathcal{D}} I_{\mathcal{D}}\varphi\|_{L^p(\Omega)^d} \right) \\
&\qquad + \|\mathbb{T}_{\mathcal{D}} I_{\mathcal{D}}\varphi - \gamma\varphi\|_{L^p(\partial\Omega)}.
\end{aligned}$$

We then use the triangle inequality again in the first two terms in the right-hand side, to introduce φ in the first one and $\nabla\varphi$ in the second one, and we apply (3.18) to conclude the proof. ∎

3.1.2 Complements on trace operators

Dealing with Neumann or Fourier boundary conditions often requires a specific knowledge of the trace operators in Sobolev spaces. We recall here the main properties used in this monograph.

Let Ω be a bounded open subset of $\mathbb{R}^d$ with Lipschitz boundary. The trace operator $\gamma : W^{1,p}(\Omega) \to W^{1-\frac{1}{p},p}(\partial\Omega)$ is well defined and surjective, and there exists a linear continuous lifting operator $\mathcal{L}_\partial : W^{1-\frac{1}{p},p}(\partial\Omega) \to W^{1,p}(\Omega)$ such that $\gamma\mathcal{L}_\partial g = g$ for any $g \in W^{1-\frac{1}{p},p}(\partial\Omega)$. Recall that in the case $p = 2$, $W^{\frac{1}{2},2}(\partial\Omega)$ is generally denoted by $H^{\frac{1}{2}}(\partial\Omega)$ and the set $W^{1,2}(\Omega)$ is denoted $H^1(\Omega)$.

We can then define the normal trace $\gamma_{\mathbf{n}}(\boldsymbol{\varphi}) \in (W^{1-\frac{1}{p},p}(\partial\Omega))'$ of $\boldsymbol{\varphi} \in W^{p'}_{\mathrm{div}}(\Omega)$ (where $p' = \frac{p}{p-1}$) the following way. Denoting by $\langle\cdot,\cdot\rangle_\partial$ the duality product between $(W^{1-\frac{1}{p},p}(\partial\Omega))'$ and $W^{1-\frac{1}{p},p}(\partial\Omega)$, we let, for any $g \in W^{1-\frac{1}{p},p}(\partial\Omega)$,

$$\langle \gamma_{\mathbf{n}}(\boldsymbol{\varphi}), g\rangle_\partial = \int_\Omega (\boldsymbol{\varphi}(\boldsymbol{x}) \cdot \nabla\mathcal{L}_\partial g(\boldsymbol{x}) + \mathcal{L}_\partial g(\boldsymbol{x}) \mathrm{div}\boldsymbol{\varphi}(\boldsymbol{x}))\, \mathrm{d}\boldsymbol{x}. \tag{3.20}$$

The linearity and continuity of $\mathcal{L}_\partial$ ensure that $\gamma_{\mathbf{n}}(\boldsymbol{\varphi})$ is indeed an element of $(W^{1-\frac{1}{p},p}(\partial\Omega))'$. Moreover, for any $\varphi \in W^{1,p}(\Omega)$ such that $\gamma\varphi = g$ we have $\mathcal{L}_\partial g - \varphi \in W^{1,p}_0(\Omega)$ and thus, by Stokes' formula,

$$\int_\Omega (\nabla(\mathcal{L}_\partial g - \varphi)(\boldsymbol{x}) \cdot \boldsymbol{\varphi}(\boldsymbol{x}) + (\mathcal{L}_\partial g - \varphi)(\boldsymbol{x}) \mathrm{div}\boldsymbol{\varphi}(\boldsymbol{x}))\, \mathrm{d}\boldsymbol{x} = 0.$$

This shows that (3.20) is also valid if we replace $\mathcal{L}_\partial g$ with any $\varphi \in W^{1,p}(\Omega)$ having trace g on $\partial\Omega$.

Lemma 3.26 (Surjectivity of the normal trace). *The normal trace $\gamma_{\mathbf{n}} : W^{p'}_{\mathrm{div}}(\Omega) \to (W^{1-\frac{1}{p},p}(\partial\Omega))'$ is linear continuous surjective. More precisely, there exists $C_\partial > 0$ depending only on Ω and p such that, for any $l \in (W^{1-\frac{1}{p},p}(\partial\Omega))'$, there exists $\boldsymbol{\varphi} \in W^{p'}_{\mathrm{div}}(\Omega)$ satisfying $\gamma_{\mathbf{n}}(\boldsymbol{\varphi}) = l$ and*

$$\|\boldsymbol{\varphi}\|_{W^{p'}_{\mathrm{div}}(\Omega)} \leq C_\partial \, \|l\|_{(W^{1-\frac{1}{p},p}(\partial\Omega))'}\,, \tag{3.21}$$

where we recall that $\|\boldsymbol{\varphi}\|_{W^{p'}_{\mathrm{div}}(\Omega)} = \|\boldsymbol{\varphi}\|_{L^{p'}(\Omega)^d} + \|\mathrm{div}\boldsymbol{\varphi}\|_{L^{p'}(\Omega)}$.

Proof. Since $\gamma : W^{1,p}(\Omega) \to W^{1-\frac{1}{p},p}(\partial\Omega)$, for any $l \in (W^{1-\frac{1}{p},p}(\partial\Omega))'$ we have $\gamma^* l \in (W^{1,p}(\Omega))'$. There exists thus $(h, \boldsymbol{\varphi}) \in L^{p'}(\Omega) \times L^{p'}(\Omega)^d$ such that, for all $\varphi \in W^{1,p}(\Omega)$,

$$\begin{aligned} \langle l, \gamma\varphi\rangle_\partial &= \langle \gamma^* l, \varphi\rangle_{(W^{1,p}(\Omega))', W^{1,p}(\Omega)} \\ &= \int_\Omega (\boldsymbol{\varphi}(\boldsymbol{x}) \cdot \nabla\varphi(\boldsymbol{x}) + h(\boldsymbol{x})\varphi(\boldsymbol{x}))\, \mathrm{d}\boldsymbol{x} \end{aligned} \tag{3.22}$$

and

$$\|\boldsymbol{\varphi}\|_{L^{p'}(\Omega)^d} + \|h\|_{L^{p'}(\Omega)} \leq \|\gamma^* l\|_{(W^{1,p}(\Omega))'} \leq C_\partial \, \|l\|_{(W^{1-\frac{1}{p},p}(\partial\Omega))'} \tag{3.23}$$

where C_∂ is the norm of γ (it is also the norm of γ^*). Testing (3.22) with functions φ in $C^\infty_c(\Omega)$ shows that $h = \mathrm{div}\boldsymbol{\varphi}$ and, therefore, that $\boldsymbol{\varphi} \in W^{p'}_{\mathrm{div}}(\Omega)$. Taking then a generic $g \in W^{1-\frac{1}{p},p}(\partial\Omega)$ and applying (3.22) with $\varphi = \mathcal{L}_\partial g$ gives $\gamma_{\mathbf{n}}(\boldsymbol{\varphi}) = l$ and Estimate (3.23) gives (3.21). ∎

Lemma 3.27 (Density of smooth functions in $W^{p'}_{\mathrm{div},\partial}(\Omega)$). *Let Ω be a polytopal open set (see Definition 7.2) and $p \in (1, +\infty)$. The space $W^{p'}_{\mathrm{div},\partial}(\Omega)$ is defined by (3.10) and endowed with the norm $\|\boldsymbol{\varphi}\|_{W^{p'}_{\mathrm{div},\partial}(\Omega)} = \|\boldsymbol{\varphi}\|_{L^{p'}(\Omega)^d} + \|\mathrm{div}\boldsymbol{\varphi}\|_{L^{p'}(\Omega)} + \|\gamma_{\mathbf{n}}(\boldsymbol{\varphi})\|_{L^{p'}(\partial\Omega)}$. Then*

1. *$C^\infty_c(\Omega)^d$ is dense in $W^{p'}_{\mathrm{div},0}(\Omega) = \{\boldsymbol{\varphi} \in W^{p'}_{\mathrm{div},\partial}(\Omega) \,:\, \gamma_{\mathbf{n}}(\boldsymbol{\varphi}) = 0\}$,*
2. *$C^\infty_c(\mathbb{R}^d)^d$ is dense in $W^{p'}_{\mathrm{div},\partial}(\Omega)$.*

Remark 3.28. We only state the lemma for polytopal open sets Ω, but the proof shows that the result is more general than this; in particular, it applies to open sets with piecewise $C^{1,1}$ boundary – since the normal $\boldsymbol{n}$ is then Lipschitz continuous outside a set of zero $(d-1)$-dimensional measure. An alternative approach to prove this lemma, that would give the density for Lipschitz domains, consists in using the mollification technique of [85].

Proof.
Item 1: using the localisation techniques of [135, Ch. 1, Theorem 1.1, (iii)], we can reduce the study to the case where Ω is strictly star-shaped, say with respect to 0. This means that, for any $\lambda \in (0,1)$, $\lambda\overline{\Omega} \subset \Omega$.
Let $\boldsymbol{\varphi} \in W^{p'}_{\mathrm{div},\partial}(\Omega)$ such that $\gamma_{\mathbf{n}}(\boldsymbol{\varphi}) = 0$. Then the extension $\widetilde{\boldsymbol{\varphi}}$ of $\boldsymbol{\varphi}$ to $\mathbb{R}^d$ by 0 outside Ω belongs to $W^{p'}_{\mathrm{div}}(\mathbb{R}^d)$, since the normal traces of $\widetilde{\boldsymbol{\varphi}}$ is continuous through $\partial\Omega$. Let $\lambda \in (0,1)$ and define $\widetilde{\boldsymbol{\varphi}}_\lambda : \boldsymbol{x} \mapsto \widetilde{\boldsymbol{\varphi}}(\boldsymbol{x}/\lambda)$. As $\lambda \to 1$, we have $\widetilde{\boldsymbol{\varphi}}_\lambda \to \widetilde{\boldsymbol{\varphi}}$ in $L^{p'}(\Omega)^d$ and $\mathrm{div}(\widetilde{\boldsymbol{\varphi}}_\lambda) = \lambda^{-1}(\mathrm{div}\widetilde{\boldsymbol{\varphi}})(\cdot/\lambda) \to \mathrm{div}\widetilde{\boldsymbol{\varphi}}$ in $L^{p'}(\Omega)$.

Moreover, the support of $\widetilde{\boldsymbol{\varphi}}_\lambda$ is contained in $\lambda\overline{\Omega}$, and is therefore compact in Ω.
Let $(\rho_\epsilon)_{\epsilon>0}$ be a smoothing kernel. For ϵ small enough, $\widetilde{\boldsymbol{\varphi}}_\lambda * \rho_\epsilon$ belongs to $C_c^\infty(\Omega)^d$. As $\epsilon \to 0$, we also have $\widetilde{\boldsymbol{\varphi}}_\lambda * \rho_\epsilon \to \widetilde{\boldsymbol{\varphi}}_\lambda$ in $L^{p'}(\Omega)^d$ and $\operatorname{div}(\widetilde{\boldsymbol{\varphi}}_\lambda * \rho_\epsilon) = \operatorname{div}(\widetilde{\boldsymbol{\varphi}}_\lambda) * \rho_\epsilon \to \operatorname{div}(\widetilde{\boldsymbol{\varphi}}_\lambda)$ in $L^{p'}(\Omega)$.
Hence, letting in that order $\lambda \to 1$ and $\epsilon \to 0$, the functions $(\widetilde{\boldsymbol{\varphi}}_\lambda * \rho_\epsilon)_{|\Omega}$ give an approximation in $W^{p'}_{\mathrm{div},\partial}(\Omega)$ of $\boldsymbol{\varphi}$ by functions in $C_c^\infty(\Omega)^d$, and the proof of Item 1 is complete.

Item 2: let $\boldsymbol{\varphi} \in W^{p'}_{\mathrm{div},\partial}(\Omega)$ and $\varepsilon > 0$. In the following, C denotes a generic constant, independent on ε but whose value can change from one occurrence to the other.
Since Ω is polytopal, $\boldsymbol{n}$ is piecewise constant and thus smooth outside a set S of 0 measure in $\partial\Omega$. We can therefore find a function ψ_ε that is C^∞-smooth, vanishes on a neighbourhood of S, and such that $\|\gamma_{\mathbf{n}}(\boldsymbol{\varphi}) - \psi_\varepsilon\|_{L^{p'}(\partial\Omega)} \le \varepsilon$. Since ψ_ε vanishes on a neighbourhood of the singularities S of $\boldsymbol{n}$, we can find a function $\boldsymbol{\psi}_\varepsilon \in C_c^\infty(\mathbb{R}^d)^d$ such that $\gamma_{\mathbf{n}}(\boldsymbol{\psi}_\varepsilon) = \psi_\varepsilon$ (simply extend, on a neighbourhood U of $\partial\Omega$, the smooth function $\psi_\varepsilon \boldsymbol{n}$ into a function that does not depend on the coordinate orthogonal to $\boldsymbol{n}$, and multiply this extension by a function in $C_c^\infty(U)$ equal to 1 on a neighbourhood of $\partial\Omega$).
Let us consider the function $\boldsymbol{\varphi} - \boldsymbol{\psi}_\varepsilon \in W^{p'}_{\mathrm{div},\partial}(\Omega)$. We have

$$\|\gamma_{\mathbf{n}}(\boldsymbol{\varphi} - \boldsymbol{\psi}_\varepsilon)\|_{L^{p'}(\partial\Omega)} = \|\gamma_{\mathbf{n}}(\boldsymbol{\varphi}) - \psi_\varepsilon\|_{L^{p'}(\partial\Omega)} \le \varepsilon. \tag{3.24}$$

By Lemma 3.26 and since $L^{p'}(\partial\Omega)$ is embedded in $(W^{1-\frac{1}{p},p}(\partial\Omega))'$, we can find $\boldsymbol{\zeta}_\varepsilon \in W^{p'}_{\mathrm{div}}(\Omega)$ such that

$$\gamma_{\mathbf{n}}(\boldsymbol{\zeta}_\varepsilon) = \gamma_{\mathbf{n}}(\boldsymbol{\varphi} - \boldsymbol{\psi}_\varepsilon) \tag{3.25}$$

and

$$\|\boldsymbol{\zeta}_\varepsilon\|_{W^{p'}_{\mathrm{div}}(\Omega)} \le C\,\|\gamma_{\mathbf{n}}(\boldsymbol{\varphi} - \boldsymbol{\psi}_\varepsilon)\|_{L^{p'}(\partial\Omega)} \le C\varepsilon. \tag{3.26}$$

Property (3.25) shows that $\boldsymbol{\zeta}_\varepsilon \in W^{p'}_{\mathrm{div},\partial}(\Omega)$ and, combined with (3.24) and (3.26), that

$$\|\boldsymbol{\zeta}_\varepsilon\|_{W^{p'}_{\mathrm{div},\partial}(\Omega)} = \|\boldsymbol{\zeta}_\varepsilon\|_{W^{p'}_{\mathrm{div}}(\Omega)} + \|\gamma_{\mathbf{n}}(\boldsymbol{\zeta}_\varepsilon)\|_{L^{p'}(\partial\Omega)} \le C\varepsilon. \tag{3.27}$$

The function $\boldsymbol{\varphi} - \boldsymbol{\psi}_\varepsilon - \boldsymbol{\zeta}_\varepsilon$ therefore belongs to $W^{p'}_{\mathrm{div},\partial}(\Omega)$ and satisfies $\gamma_{\mathbf{n}}(\varphi - \boldsymbol{\psi}_\varepsilon - \boldsymbol{\zeta}_\varepsilon) = 0$. By Item 1 we can find $\boldsymbol{\xi}_\varepsilon \in C_c^\infty(\Omega)^d$ such that

$$\|(\boldsymbol{\varphi} - \boldsymbol{\psi}_\varepsilon - \boldsymbol{\zeta}_\varepsilon) - \boldsymbol{\xi}_\varepsilon\|_{W^{p'}_{\mathrm{div},\partial}(\Omega)} \le \varepsilon. \tag{3.28}$$

We now set $\boldsymbol{\varphi}_\varepsilon = \boldsymbol{\psi}_\varepsilon + \boldsymbol{\xi}_\varepsilon$. This function belongs to $C_c^\infty(\mathbb{R}^d)^d$ and

$$\boldsymbol{\varphi} - \boldsymbol{\varphi}_\varepsilon = (\boldsymbol{\varphi} - \boldsymbol{\psi}_\varepsilon - \boldsymbol{\zeta}_\varepsilon - \boldsymbol{\xi}_\varepsilon) + \boldsymbol{\zeta}_\varepsilon$$

so, by (3.27) and (3.28), $\|\boldsymbol{\varphi} - \boldsymbol{\varphi}_\varepsilon\|_{W^{p'}_{\mathrm{div},\partial}(\Omega)} \le C\varepsilon$. ∎

3.1.3 Gradient schemes for linear problems

We consider here a linear elliptic problem with non-homogeneous Neumann boundary conditions

$$\begin{aligned} &-\mathrm{div}(\Lambda\nabla\overline{u}) = f + \mathrm{div}(\boldsymbol{F}) \text{ in } \Omega, \\ &\Lambda\nabla\overline{u}\cdot\mathbf{n} + \boldsymbol{F}\cdot\mathbf{n} = h \text{ on } \partial\Omega, \end{aligned} \tag{3.29}$$

where $\mathbf{n}$ is the unit normal outward Ω to $\partial\Omega$. We assume that

- Ω is an open bounded connected subset of $\mathbb{R}^d$ ($d \in \mathbb{N}^\star$), with a Lipschitz boundary, (3.30a)
- Λ is a measurable function from Ω to the set of $d \times d$ symmetric matrices and there exists $\underline{\lambda}, \overline{\lambda} > 0$ such that, for a.e. $\boldsymbol{x} \in \Omega$, $\Lambda(\boldsymbol{x})$ has eigenvalues in $[\underline{\lambda}, \overline{\lambda}]$, (3.30b)
- $f \in L^2(\Omega)\,,\ \boldsymbol{F} \in L^2(\Omega)^d\,,\ h \in L^2(\partial\Omega)$, and $\displaystyle\int_\Omega f(\boldsymbol{x})\mathrm{d}\boldsymbol{x} + \int_{\partial\Omega} h(\boldsymbol{x})\mathrm{d}\gamma(\boldsymbol{x}) = 0.$ (3.30c)

Under these hypotheses and defining

$$H^1_\star(\Omega) = \left\{\varphi \in H^1(\Omega)\,,\ \int_\Omega \varphi(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = 0\right\},$$

the weak formulation of (3.29) is

$$\begin{aligned} &\overline{u} \in H^1_\star(\Omega),\ \forall v \in H^1_\star(\Omega), \\ &\int_\Omega \Lambda(\boldsymbol{x})\nabla\overline{u}(\boldsymbol{x})\cdot\nabla v(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \\ &= \int_\Omega f(\boldsymbol{x})v(\boldsymbol{x})\mathrm{d}\boldsymbol{x} - \int_\Omega \boldsymbol{F}(\boldsymbol{x})\cdot\nabla v(\boldsymbol{x})\mathrm{d}\boldsymbol{x} + \int_{\partial\Omega} h(\boldsymbol{x})\gamma(v)(\boldsymbol{x})\mathrm{d}\gamma(\boldsymbol{x}). \end{aligned} \tag{3.31}$$

Owing to Hypothesis (3.30c), Problem (3.31) is equivalent to

$$\begin{aligned} &\overline{u} \in H^1(\Omega),\ \forall v \in H^1(\Omega), \\ &\int_\Omega \Lambda(\boldsymbol{x})\nabla\overline{u}(\boldsymbol{x})\cdot\nabla v(\boldsymbol{x})\mathrm{d}\boldsymbol{x} + \left(\int_\Omega \overline{u}(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\right)\left(\int_\Omega v(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\right) \\ &= \int_\Omega f(\boldsymbol{x})v(\boldsymbol{x})\mathrm{d}\boldsymbol{x} - \int_\Omega \boldsymbol{F}(\boldsymbol{x})\cdot\nabla v(\boldsymbol{x})\mathrm{d}\boldsymbol{x} + \int_{\partial\Omega} h(\boldsymbol{x})\gamma(v)(\boldsymbol{x})\mathrm{d}\gamma(\boldsymbol{x}). \end{aligned} \tag{3.32}$$

Indeed, letting $v \equiv 1$ in (3.32) implies that $\int_\Omega \overline{u}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = 0$. The approximation of Problem (3.31) by the GDM is described now.

Definition 3.29 (GS, Neumann BCs). *If $\mathcal{D} = (X_\mathcal{D}, \Pi_\mathcal{D}, \mathbb{T}_\mathcal{D}, \nabla_\mathcal{D})$ is a GD for Neumann problems in the sense of Definition 3.11, then we define the related* gradient scheme!linear elliptic problem *for* (3.31) *by*

Find $u \in X_{\mathcal{D}}$ such that, for any $v \in X_{\mathcal{D}}$,

$$\begin{aligned}
&\int_\Omega \Lambda(\boldsymbol{x})\nabla_{\mathcal{D}} u(\boldsymbol{x}) \cdot \nabla_{\mathcal{D}} v(\boldsymbol{x})\mathrm{d}\boldsymbol{x} + \left(\int_\Omega \Pi_{\mathcal{D}} u(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\right)\left(\int_\Omega \Pi_{\mathcal{D}} v(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\right) \\
&= \int_\Omega f(\boldsymbol{x})\Pi_{\mathcal{D}} v(\boldsymbol{x})\mathrm{d}\boldsymbol{x} - \int_\Omega \boldsymbol{F}(\boldsymbol{x}) \cdot \nabla_{\mathcal{D}} v(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \\
&\quad + \int_{\partial\Omega} h(\boldsymbol{x})\mathbb{T}_{\mathcal{D}} v(\boldsymbol{x})\mathrm{d}\gamma(\boldsymbol{x}).
\end{aligned} \tag{3.33}$$

Fixing a basis $(\xi^{(i)})_{i=1,\dots,N}$ of the space $X_{\mathcal{D}}$, Scheme (3.33) is equivalent to solving the linear square system $AU = B$ with

$$\begin{aligned}
u &= \sum_{j=1}^N U_j \xi^{(j)}, \\
A_{ij} &= \int_\Omega \Lambda(\boldsymbol{x})\nabla_{\mathcal{D}} \xi^{(j)}(\boldsymbol{x}) \cdot \nabla_{\mathcal{D}} \xi^{(i)}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \\
&\qquad + \left(\int_\Omega \Pi_{\mathcal{D}} \xi^{(j)}(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\right)\left(\int_\Omega \Pi_{\mathcal{D}} \xi^{(i)}(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\right), \\
B_i &= \int_\Omega f(\boldsymbol{x})\Pi_{\mathcal{D}} \xi^{(i)}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} - \int_\Omega \boldsymbol{F}(\boldsymbol{x}) \cdot \nabla_{\mathcal{D}} \xi^{(i)}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \\
&\qquad + \int_{\partial\Omega} h(\boldsymbol{x})\mathbb{T}_{\mathcal{D}} \xi^{(i)}(\boldsymbol{x})\mathrm{d}\gamma(\boldsymbol{x}).
\end{aligned}$$

Contrary to the case of Dirichlet boundary conditions, the matrix A here is full, due to the second term in the expression of A_{ij}. There exist different modifications of this system, aiming at recovering a sparse matrix. A standard one consists in adding a new unknown $\mu = \int_\Omega \Pi_{\mathcal{D}} u(\boldsymbol{x})\mathrm{d}\boldsymbol{x}$, and to rewrite (3.33) as

Find $u \in X_{\mathcal{D}}$ and $\mu \in \mathbb{R}$ such that, for any $v \in X_{\mathcal{D}}$,

$$\begin{aligned}
&\int_\Omega \Lambda(\boldsymbol{x})\nabla_{\mathcal{D}} u(\boldsymbol{x}) \cdot \nabla_{\mathcal{D}} v(\boldsymbol{x})\mathrm{d}\boldsymbol{x} + \mu \int_\Omega \Pi_{\mathcal{D}} v(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \\
&\qquad = \int_\Omega f(\boldsymbol{x})\Pi_{\mathcal{D}} v(\boldsymbol{x})\mathrm{d}\boldsymbol{x} - \int_\Omega \boldsymbol{F}(\boldsymbol{x}) \cdot \nabla_{\mathcal{D}} v(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \\
&\qquad\quad + \int_{\partial\Omega} h(\boldsymbol{x})\mathbb{T}_{\mathcal{D}} v(\boldsymbol{x})\mathrm{d}\gamma(\boldsymbol{x}), \\
&\int_\Omega \Pi_{\mathcal{D}} u(\boldsymbol{x})\mathrm{d}\boldsymbol{x} - \mu = 0.
\end{aligned} \tag{3.34}$$

Assuming that the diameter of the supports of $(\nabla_{\mathcal{D}} \xi^{(i)})_{i=1,\dots,N}$ is small compared to $\mathrm{diam}(\Omega)$, which is often the case for standard choices of spaces and bases, the new linear system deduced from this reformulation is sparse, symmetric and invertible (though not positive).

The error estimates for Neumann boundary conditions are stated in the next theorem. We do not explicitly state the convergence result, similar to Corollary 2.31, that stems from these error estimates.

Theorem 3.30 (Error estimate, Neumann BCs). *Under Hypotheses (3.30), let $\overline{u} \in H^1_\star(\Omega)$ be the solution of (3.31) (remark that $f \in L^2(\Omega)$ and $h \in L^2(\partial\Omega)$ imply that in $\Lambda\nabla\overline{u} + \boldsymbol{F} \in W^2_{\mathrm{div},\partial}(\Omega)$, see (3.10)).*
Let $\mathcal{D}$ be a GD for a Neumann problem in the sense of Definition 3.11.
Then there exists one and only one $u_\mathcal{D} \in X_\mathcal{D}$ solution to the GS (3.33), and this element satisfies the following inequalities:

$$\|\nabla\overline{u} - \nabla_\mathcal{D} u_\mathcal{D}\|_{L^2(\Omega)^d} \leq \mathrm{Err} + S_\mathcal{D}(\overline{u}), \tag{3.35}$$

$$\|\overline{u} - \Pi_\mathcal{D} u_\mathcal{D}\|_{L^2(\Omega)} \leq C_\mathcal{D}\mathrm{Err} + S_\mathcal{D}(\overline{u}), \tag{3.36}$$

$$\textit{with } \mathrm{Err} := \frac{1}{\min(\underline{\lambda}, 1)}\left[W_\mathcal{D}(\Lambda\nabla\overline{u} + \boldsymbol{F}) + (\overline{\lambda} + |\Omega|^{1/2}C_\mathcal{D})S_\mathcal{D}(\overline{u})\right],$$

where $C_\mathcal{D}$, $S_\mathcal{D}$ and $W_\mathcal{D}$ are defined by (3.9), (3.3) and (3.11).

Remark 3.31 (Error estimate on the traces and lower bound). If we let $\overline{S}_\mathcal{D}$ be the measure of GD-consistency for Fourier boundary conditions (*i.e.* (3.17)), then Proposition 3.24 and Theorem 3.30 show that

$$\|\gamma(\overline{u}) - \mathbb{T}_\mathcal{D} u_\mathcal{D}\|_{L^2(\partial\Omega)} \leq C_1\left(W_\mathcal{D}(\Lambda\nabla\overline{u} + \boldsymbol{F}) + \overline{S}_\mathcal{D}(\overline{u})\right),$$

where C_1 depends only on $\underline{\lambda}$, $\overline{\lambda}$, $|\Omega|$ and an upper bound of $C_\mathcal{D}$. Note that a lower bound in the spirit of (2.29) could also be derived.

Proof. Proving (3.35) for any solution $u_\mathcal{D} \in X_\mathcal{D}$ to Scheme (3.33) is sufficient to establish the existence and uniqueness of this solution. Indeed, this estimate shows that, whenever $f = 0$, $h = 0$ and $\boldsymbol{F} = 0$ (in which case $\overline{u} = 0$), $\nabla_\mathcal{D} u_\mathcal{D}$ vanishes a.e.; then, letting $v = u_\mathcal{D}$ in (3.33) implies $\int_\Omega \Pi_\mathcal{D} u(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = 0$, and therefore $u_\mathcal{D} = 0$ since (3.8) is a norm on $X_\mathcal{D}$.
To prove the estimates, we take $\boldsymbol{\varphi} = \Lambda\nabla\overline{u} + \boldsymbol{F} \in W^2_{\mathrm{div},\partial}(\Omega)$ in the definition (3.11) of $W_\mathcal{D}$ and using that $\overline{u}$ is the solution to (3.31). We then have, for any $v \in X_\mathcal{D}$,

$$\Big|\int_\Omega \left[\nabla_\mathcal{D} v(\boldsymbol{x}) \cdot (\Lambda(\boldsymbol{x})\nabla\overline{u}(\boldsymbol{x}) + \boldsymbol{F}(\boldsymbol{x})) - \Pi_\mathcal{D} v(\boldsymbol{x}) f(\boldsymbol{x})\right]\mathrm{d}\boldsymbol{x} \\ - \int_{\partial\Omega} h(\boldsymbol{x})\mathbb{T}_\mathcal{D} v(\boldsymbol{x})\mathrm{d}\gamma(\boldsymbol{x})\Big| \leq \|v\|_\mathcal{D}\, W_\mathcal{D}(\Lambda\nabla\overline{u} + \boldsymbol{F}).$$

Therefore, since $u_\mathcal{D}$ is a solution to (3.33), we get

$$\Big|\int_\Omega \Lambda\nabla_\mathcal{D} v(\boldsymbol{x}) \cdot (\nabla\overline{u}(\boldsymbol{x}) - \nabla_\mathcal{D} u_\mathcal{D}(\boldsymbol{x}))\mathrm{d}\boldsymbol{x} \\ - \int_\Omega \Pi_\mathcal{D} u_\mathcal{D}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \int_\Omega \Pi_\mathcal{D} v(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\Big| \leq \|v\|_\mathcal{D}\, W_\mathcal{D}(\Lambda\nabla\overline{u} + \boldsymbol{F}).$$

We then introduce

$$I_{\mathcal{D}}\overline{u} = \underset{w\in X_{\mathcal{D}}}{\operatorname{argmin}} \left(\|\Pi_{\mathcal{D}} w - \overline{u}\|_{L^2(\Omega)} + \|\nabla_{\mathcal{D}} w - \nabla\overline{u}\|_{L^2(\Omega)^d}\right) \tag{3.37}$$

in the above inequality. It leads to

$$\begin{aligned}
&\left| \int_\Omega \Lambda\nabla_{\mathcal{D}} v(\boldsymbol{x}) \cdot (\nabla_{\mathcal{D}} I_{\mathcal{D}}\overline{u}(\boldsymbol{x}) - \nabla_{\mathcal{D}} u_{\mathcal{D}}(\boldsymbol{x}))\mathrm{d}\boldsymbol{x} \right.\\
&\qquad \left. + \int_\Omega (\Pi_{\mathcal{D}} I_{\mathcal{D}}\overline{u}(\boldsymbol{x}) - \Pi_{\mathcal{D}} u_{\mathcal{D}}(\boldsymbol{x}))\mathrm{d}\boldsymbol{x} \int_\Omega \Pi_{\mathcal{D}} v(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \right|\\
&\le \|v\|_{\mathcal{D}}\, W_{\mathcal{D}}(\Lambda\nabla\overline{u} + \boldsymbol{F}) + \left| \int_\Omega \Lambda\nabla_{\mathcal{D}} v(\boldsymbol{x}) \cdot (\nabla_{\mathcal{D}} I_{\mathcal{D}}\overline{u}(\boldsymbol{x}) - \nabla\overline{u}(\boldsymbol{x}))\mathrm{d}\boldsymbol{x} \right.\\
&\qquad\qquad \left. + \int_\Omega \Pi_{\mathcal{D}} I_{\mathcal{D}}\overline{u}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \int_\Omega \Pi_{\mathcal{D}} v(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \right|.
\end{aligned}$$

Observing that

$$\begin{aligned}
\left| \int_\Omega \Pi_{\mathcal{D}} I_{\mathcal{D}}\overline{u}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \right| &= \left| \int_\Omega (\Pi_{\mathcal{D}} I_{\mathcal{D}}\overline{u}(\boldsymbol{x}) - \overline{u}(\boldsymbol{x}))\mathrm{d}\boldsymbol{x} \right|\\
&\le |\Omega|^{1/2} \left\|\Pi_{\mathcal{D}} I_{\mathcal{D}}\overline{u} - \overline{u}\right\|_{L^2(\Omega)},
\end{aligned}$$

we can write, using the definition (3.9) of $C_{\mathcal{D}}$,

$$\begin{aligned}
&\left| \int_\Omega \Lambda\nabla_{\mathcal{D}} v(\boldsymbol{x}) \cdot (\nabla_{\mathcal{D}} I_{\mathcal{D}}\overline{u}(\boldsymbol{x}) - \nabla_{\mathcal{D}} u_{\mathcal{D}}(\boldsymbol{x}))\mathrm{d}\boldsymbol{x} \right.\\
&\qquad \left. + \int_\Omega (\Pi_{\mathcal{D}} I_{\mathcal{D}}\overline{u}(\boldsymbol{x}) - \Pi_{\mathcal{D}} u_{\mathcal{D}}(\boldsymbol{x}))\mathrm{d}\boldsymbol{x} \int_\Omega \Pi_{\mathcal{D}} v(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \right|\\
&\qquad\qquad \le \|v\|_{\mathcal{D}} \left[W_{\mathcal{D}}(\Lambda\nabla\overline{u} + \boldsymbol{F}) + (\overline{\lambda} + |\Omega|^{1/2} C_{\mathcal{D}}) S_{\mathcal{D}}(\overline{u}) \right].
\end{aligned}$$

We now let $v = I_{\mathcal{D}}\overline{u} - u_{\mathcal{D}}$. Recalling the definition (3.8) of $\|\cdot\|_{\mathcal{D}}$, we obtain

$$\|I_{\mathcal{D}}\overline{u} - u_{\mathcal{D}}\|_{\mathcal{D}} \le \frac{W_{\mathcal{D}}(\Lambda\nabla\overline{u} + \boldsymbol{F}) + (\overline{\lambda} + |\Omega|^{1/2} C_{\mathcal{D}}) S_{\mathcal{D}}(\overline{u})}{\min(\underline{\lambda}, 1)}.$$

The conclusion follows as in the proof of Theorem 2.28. ∎

3.1.4 Gradient schemes for quasi-linear problems

Homogeneous Neumann BCs

We consider Problem (2.49a), replacing the homogeneous Dirichlet BCs by homogeneous Neumann BCs:

$$\begin{aligned} -\mathrm{div}(\Lambda(\boldsymbol{x},\overline{u})\nabla\overline{u}) &= f \text{ in } \Omega, \\ \Lambda(\boldsymbol{x},\overline{u})\nabla\overline{u}\cdot\boldsymbol{n} &= 0 \text{ on } \partial\Omega. \end{aligned} \tag{3.38}$$

We take usual assumptions:

- Ω is an open bounded connected subset of $\mathbb{R}^d$ ($d \in \mathbb{N}^\star$), with Lipschitz boundary, (3.39a)
- Λ is a Caratheodory function from $\Omega \times \mathbb{R}$ to $\mathcal{M}_d(\mathbb{R})$, (*i.e.* $\Lambda(\boldsymbol{x}, s)$ is measurable w.r.t. $\boldsymbol{x}$ and continuous w.r.t. s), there exists $\underline{\lambda}, \overline{\lambda} > 0$ such that, for a.e. $\boldsymbol{x} \in \Omega$, for all $s \in \mathbb{R}$, $\Lambda(\boldsymbol{x}, s)$ is symmetric with eigenvalues in $[\underline{\lambda}, \overline{\lambda}]$, (3.39b)
- $f \in L^2(\Omega)$ such that $\displaystyle\int_\Omega f(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = 0$. (3.39c)

Under these hypotheses, again defining $H^1_\star(\Omega) = \{\varphi \in H^1(\Omega) \,:\, \int_\Omega \varphi(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = 0\}$, a weak solution to (3.38) is a function $\overline{u}$ (not necessarily unique) satisfying:

$$\begin{aligned} &\overline{u} \in H^1_\star(\Omega),\ \forall v \in H^1_\star(\Omega), \\ &\int_\Omega \Lambda(\boldsymbol{x},\overline{u}(\boldsymbol{x}))\nabla\overline{u}(\boldsymbol{x})\cdot\nabla v(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = \int_\Omega f(\boldsymbol{x})v(\boldsymbol{x})\mathrm{d}\boldsymbol{x}. \end{aligned} \tag{3.40}$$

This weak formulation is approximated by the following gradient scheme.

Definition 3.32 (GS, quasi-linear problem, homogeneous Neumann BCs). *If $\mathcal{D} = (X_\mathcal{D}, \Pi_\mathcal{D}, \nabla_\mathcal{D})$ is a GD in the sense of Definition 3.1, then we define the related* gradient scheme *for* (3.40) *by*

$$\begin{aligned} &\textit{Find } u \in X_\mathcal{D} \textit{ such that for any } v \in X_\mathcal{D}, \\ &\int_\Omega \Lambda(\boldsymbol{x},\Pi_\mathcal{D}u(\boldsymbol{x}))\nabla_\mathcal{D}u(\boldsymbol{x})\cdot\nabla_\mathcal{D}v(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \\ &\quad + \left(\int_\Omega \Pi_\mathcal{D}u(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\right)\left(\int_\Omega \Pi_\mathcal{D}v(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\right) = \int_\Omega f(\boldsymbol{x})\Pi_\mathcal{D}v(\boldsymbol{x})\mathrm{d}\boldsymbol{x}. \end{aligned} \tag{3.41}$$

This scheme leads to a non-linear system of equations of the form $A(u)U = B$, similar to the system (2.53) obtained in the case of Dirichlet BCs. As in the linear case, the matrix $A(u)$ corresponding to (3.41) is in general not sparse, but algebraic methods can be used to compute the solution of (3.41) by solving linear systems with better sparsity properties (see (3.34)).

The next theorem states a convergence result for the gradient scheme (3.41).

Theorem 3.33 (Convergence, quasi-linear problem, homogeneous Neumann BCs). *Assume* (3.39)*, and let $(\mathcal{D}_m)_{m\in\mathbb{N}}$ be a sequence of GDs in the sense of Definition 3.1, which is GD-consistent, limit-conforming and compact in the sense of Definitions 3.4, 3.5 and 3.6.*

Then, for any $m \in \mathbb{N}$, there exists at least one $u_m \in X_{\mathcal{D}_m}$ solution to the gradient scheme (3.41) and, up to a subsequence, $\Pi_{\mathcal{D}_m} u_m$ converges strongly in $L^2(\Omega)$ to a solution $\overline{u}$ of (3.40) and $\nabla_{\mathcal{D}_m} u_m$ converges strongly in $L^2(\Omega)^d$ to $\nabla \overline{u}$ as $m \to \infty$.
In the case where the solution $\overline{u}$ of (3.40) is unique, then the whole sequence converges to $\overline{u}$ as $m \to \infty$ in the senses above.

Proof. We proceed as in the proof of Theorem 2.35.
For any GD $\mathcal{D}$ in the sense of Definition 3.11, let $w \in X_{\mathcal{D}}$ be given, and let $u \in X_{\mathcal{D}}$ be such that

$$\begin{aligned} &\int_\Omega \Lambda(\boldsymbol{x}, \Pi_{\mathcal{D}} w(\boldsymbol{x})) \nabla_{\mathcal{D}} u(\boldsymbol{x}) \cdot \nabla_{\mathcal{D}} v(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \\ &+ \int_\Omega \Pi_{\mathcal{D}} u(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \int_\Omega \Pi_{\mathcal{D}} v(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} = \int_\Omega f(\boldsymbol{x}) \Pi_{\mathcal{D}} v(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}\,, \quad \forall v \in X_{\mathcal{D}}. \end{aligned} \tag{3.42}$$

Then, letting $v = u$ in (3.42), and applying the Cauchy–Schwarz inequality and the coercivity property in the sense of Definition 3.3, we get

$$\min(\underline{\lambda}, 1) \|u\|_{\mathcal{D}}^2 \le \|f\|_{L^2(\Omega)} \|\Pi_{\mathcal{D}} u\|_{L^2(\Omega)} \le C_{\mathcal{D}} \|f\|_{L^2(\Omega)} \|u\|_{\mathcal{D}}.$$

This shows that

$$\|u\|_{\mathcal{D}} \le \frac{C_{\mathcal{D}}}{\min(\underline{\lambda}, 1)} \|f\|_{L^2(\Omega)}. \tag{3.43}$$

Therefore, u is obtained by the resolution of an invertible square linear system (since a null right hand side implies $u = 0$). The mapping $w \to u$ is continuous, by continuity of the coefficients of the inverse of a matrix with respect to its coefficients. The Brouwer theorem D.2 yields that this mapping $w \to u$ has at least one fixed point. This shows the existence of at least one discrete solution to (3.41). It is clear that any solution to this scheme satisfies (3.43).
We denote by $u_m \in X_{\mathcal{D}_m}$ such a solution for $\mathcal{D} = \mathcal{D}_m$. The estimate (3.43) shows that $(\|u_m\|_{\mathcal{D}_m})_{m \in \mathbb{N}}$ is bounded and thus, up to a subsequence still denoted by $(u_m)_{m \in \mathbb{N}}$, we find $\overline{u} \in H^1(\Omega)$ such that $\Pi_{\mathcal{D}_m} u_m$ converges strongly in $L^2(\Omega)$ and a.e. to $\overline{u}$ and $\nabla_{\mathcal{D}_m} u_m$ converges weakly in $L^2(\Omega)^d$ to $\nabla \overline{u}$. We used here Remark 3.23 and the compactness of the sequence of GDs.
We define $I_{\mathcal{D}} : H^1(\Omega) \to X_{\mathcal{D}}$ by

$$I_{\mathcal{D}} \varphi = \operatorname*{argmin}_{v \in X_{\mathcal{D}}} \left(\|\Pi_{\mathcal{D}} v - \varphi\|_{L^2(\Omega)} + \|\nabla_{\mathcal{D}} v - \nabla \varphi\|_{L^2(\Omega)^d} \right). \tag{3.44}$$

By GD-consistency of the sequence of GDs, for any $\varphi \in H^1(\Omega)$ we have $\Pi_{\mathcal{D}_m}(I_{\mathcal{D}_m} \varphi) \to \varphi$ strongly in $L^2(\Omega)$ and $\nabla_{\mathcal{D}_m}(I_{\mathcal{D}_m} \varphi) \to \nabla \varphi$ strongly in $L^2(\Omega)^d$.
Since $\mathbf{1}_\Omega$ (the characteristic function of Ω) belongs to $H^1(\Omega)$, we can take $v = I_{\mathcal{D}_m} \mathbf{1}_\Omega$ in (3.41) and pass to the limit. We get, thanks to Hypothesis (3.39c), that

$$
\begin{aligned}
0 &= \lim_{m\to\infty} \left(\int_\Omega \Pi_{\mathcal{D}_m} u_m(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\right)\left(\int_\Omega \Pi_{\mathcal{D}_m}(I_{\mathcal{D}_m}\varphi)(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\right)\\
&= \left(\int_\Omega \overline{u}(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\right)|\Omega|.
\end{aligned}
$$

This shows that $\overline{u} \in H^1_\star(\Omega)$ and that

$$
\lim_{m\to\infty} \int_\Omega \Pi_{\mathcal{D}_m} u_m(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = 0. \tag{3.45}
$$

Let $\varphi \in H^1_\star(\Omega)$ be given. Using the non-linear strong convergence property of Lemma D.9 page 465, $\Lambda(\cdot, \Pi_{\mathcal{D}_m} u_m)\nabla_{\mathcal{D}_m}(I_{\mathcal{D}_m}\varphi) \to \Lambda(\cdot, u)\nabla\varphi$ strongly in $L^2(\Omega)^d$. Lemma D.8 (weak-strong convergence property) enables us to pass to the limit in (3.41) with $v = I_{\mathcal{D}_m}\varphi$, which proves that $\overline{u}$ is a solution to (3.40).
By passing to the limit in the left-hand side of (3.41) with $v = u_m$ and using (3.45), we get

$$
\begin{aligned}
&\lim_{m\to\infty} \int_\Omega \Lambda(\boldsymbol{x}, \Pi_{\mathcal{D}_m} u_m(\boldsymbol{x}))\nabla_{\mathcal{D}_m} u_m(\boldsymbol{x}) \cdot \nabla_{\mathcal{D}_m} u_m(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\\
&= \int_\Omega f(\boldsymbol{x})\overline{u}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = \int_\Omega \Lambda(\boldsymbol{x}, \overline{u}(\boldsymbol{x}))\nabla\overline{u}(\boldsymbol{x}) \cdot \nabla\overline{u}(\boldsymbol{x})\mathrm{d}\boldsymbol{x}
\end{aligned}
$$

and the strong convergence of $\nabla_{\mathcal{D}_m} u_m$ to $\nabla\overline{u}$ follows from this as in the proof of Theorem 2.35. ■

Non-homogeneous Neumann boundary conditions

We consider Problem (3.38) in which the boundary condition is replaced with a non-homogeneous condition:

$$
\Lambda(\boldsymbol{x}, \overline{u})\nabla\overline{u} \cdot \boldsymbol{n} = h \text{ on } \partial\Omega,
$$

under the assumptions (3.39a)–(3.39b) and

$$
f \in L^2(\Omega)\,,\ h \in L^2(\partial\Omega),\ \int_\Omega f(\boldsymbol{x})\mathrm{d}\boldsymbol{x} + \int_{\partial\Omega} h(\boldsymbol{x})\mathrm{d}\gamma(\boldsymbol{x}) = 0. \tag{3.46}
$$

Under these hypotheses, recalling that

$$
H^1_\star(\Omega) = \left\{\varphi \in H^1(\Omega)\ :\ \int_\Omega \varphi(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = 0\right\},
$$

a weak solution of this problem is a function $\overline{u}$ (not necessarily unique) satisfying:

$$\begin{aligned}&\overline{u} \in H^1_\star(\Omega),\ \forall v \in H^1_\star(\Omega),\\&\int_\Omega \Lambda(\boldsymbol{x}, \overline{u}(\boldsymbol{x}))\nabla\overline{u}(\boldsymbol{x}) \cdot \nabla v(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\\&\qquad = \int_\Omega f(\boldsymbol{x})v(\boldsymbol{x})\mathrm{d}\boldsymbol{x} + \int_{\partial\Omega} h(\boldsymbol{x})\gamma(v)(\boldsymbol{x})\mathrm{d}\gamma(\boldsymbol{x}).\end{aligned} \tag{3.47}$$

As for homogeneous Neumann BCs, owing to Hypothesis (3.46), Problem (3.47) is equivalent to

$$\begin{aligned}&\overline{u} \in H^1(\Omega),\ \forall v \in H^1(\Omega),\\&\int_\Omega \Lambda(\boldsymbol{x}, \overline{u}(\boldsymbol{x}))\nabla\overline{u}(\boldsymbol{x}) \cdot \nabla v(\boldsymbol{x})\mathrm{d}\boldsymbol{x} + \left(\int_\Omega \overline{u}(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\right)\left(\int_\Omega v(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\right)\\&= \int_\Omega f(\boldsymbol{x})v(\boldsymbol{x})\mathrm{d}\boldsymbol{x} + \int_{\partial\Omega} h(\boldsymbol{x})\gamma(v)(\boldsymbol{x})\mathrm{d}\gamma(\boldsymbol{x}).\end{aligned} \tag{3.48}$$

This problem is therefore approximated by the following gradient scheme.

Definition 3.34 (GS, quasi-linear problem, non-homogeneous Neumann BCs). *If $\mathcal{D} = (X_\mathcal{D}, \Pi_\mathcal{D}, \mathbb{T}_\mathcal{D}, \nabla_\mathcal{D})$ is a GD in the sense of Definition 3.11, then we define the related* gradient scheme *for* (3.47) *by*

Find $u \in X_\mathcal{D}$ such that, for any $v \in X_\mathcal{D}$,

$$\begin{aligned}&\int_\Omega \Lambda(\boldsymbol{x}, \Pi_\mathcal{D} u(\boldsymbol{x}))\nabla_\mathcal{D} u(\boldsymbol{x}) \cdot \nabla_\mathcal{D} v(\boldsymbol{x})\mathrm{d}\boldsymbol{x} + \left(\int_\Omega \Pi_\mathcal{D} u(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\right)\left(\int_\Omega \Pi_\mathcal{D} v(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\right)\\&\qquad = \int_\Omega f(\boldsymbol{x})\Pi_\mathcal{D} v(\boldsymbol{x})\mathrm{d}\boldsymbol{x} + \int_{\partial\Omega} h(\boldsymbol{x})\mathbb{T}_\mathcal{D}(v)(\boldsymbol{x})\mathrm{d}\gamma(\boldsymbol{x}).\end{aligned} \tag{3.49}$$

We then have the following convergence result.

Theorem 3.35 (Convergence, quasi-linear problem, non-homogeneous Neumann BCs). *Under Assumptions* (3.39a)–(3.39b) *and* (3.46)*, let $(\mathcal{D}_m)_{m\in\mathbb{N}}$ be a sequence of GDs in the sense of Definition 3.11, which is GD-consistent, limit-conforming and compact in the sense of Definitions 3.4, 3.14 and 3.15.*
Then, for any $m \in \mathbb{N}$, there exists at least one $u_m \in X_{\mathcal{D}_m}$ solution to the gradient scheme (3.49) *and, up to a subsequence as $m \to \infty$, $\Pi_{\mathcal{D}_m} u_m$ converges strongly in $L^2(\Omega)$ to a solution $\overline{u}$ of* (3.47)*, $\mathbb{T}_{\mathcal{D}_m} u_m$ converges strongly in $L^2(\partial\Omega)$ to $\gamma\overline{u}$, and $\nabla_{\mathcal{D}_m} u_m$ converges strongly in $L^2(\Omega)^d$ to $\nabla\overline{u}$.*
In the case where the solution $\overline{u}$ of (3.47) *is unique, then the whole sequence converges to $\overline{u}$ as $m \to \infty$ in the senses above.*

Proof. We follow the same ideas as in the proof of Theorem 3.33. Note that, by Lemma 3.18, $(\mathcal{D}_m)_{m\in\mathbb{N}}$ is coercive.
Brouwer's fixed point theorem yields the existence of at least one discrete solution $u_m \in X_{\mathcal{D}_m}$ to (3.49) with $\mathcal{D} = \mathcal{D}_m$. Plug $v = u_m$ in this scheme. The coercivity property (Definition 3.13), which involves the discrete trace, then yields

$$\|u_m\|_{\mathcal{D}_m} \leq \frac{C_P}{\min(\underline{\lambda}, 1)}(\|f\|_{L^2(\Omega)} + \|h\|_{L^2(\partial\Omega)}). \tag{3.50}$$

Then the compactness of $(\mathcal{D}_m)_{m\in\mathbb{N}}$ (Definition 3.15) and similar arguments as in the proof of Theorem 3.33 give the existence of $\overline{u} \in H^1_\star(\Omega)$ and of a subsequence of $(\mathcal{D}_m, u_m)_{m\in\mathbb{N}}$, denoted the same way, such that $\nabla_{\mathcal{D}_m} u_m$ converges weakly in $L^2(\Omega)^d$ to $\nabla\overline{u}$, $\Pi_{\mathcal{D}_m} u_m$ converges strongly in $L^2(\Omega)$ to $\overline{u}$ and $\mathbb{T}_{\mathcal{D}_m} u_m$ converges strongly in $L^2(\partial\Omega)$ to $\gamma\overline{u}$. Moreover, if $\varphi \in H^1_\star(\Omega)$ and $v_m = I_{\mathcal{D}_m}\varphi$ is the interpolant defined by (3.44), Remark 3.22 shows that $\mathbb{T}_{\mathcal{D}_m} v_m \to \gamma\varphi$ weakly in $L^2(\partial\Omega)$ (actually, by the compactness property of $(\mathcal{D}_m)_{m\in\mathbb{N}}$, this convergence is strong, but this is not required to pass to the limit). The conclusion then follows as for Theorem 3.33. ∎

3.2 Fourier boundary conditions

The general case of non-homogeneous Fourier (or Robin) (also known as Robin) boundary conditions is considered here; it also covers, of course, homogeneous Fourier BCs. The notions and results in this section correspond to those in Appendix A with the choices in Section A.1.4.
Here Ω is again a connected open bounded subset of $\mathbb{R}^d$ with Lipschitz boundary, and $p \in (1, +\infty)$.

3.2.1 Gradient discretisations

Except for the choice of the norm $\|\cdot\|_{\mathcal{D}}$, the definition of a gradient discretisation for Fourier boundary conditions is the same as for Neumann boundary conditions.

Definition 3.36 (GD, non-homogeneous Fourier BCs). *A gradient discretisation $\mathcal{D}$ for non-homogeneous Fourier conditions $\mathcal{D}$ is defined by $\mathcal{D} = (X_\mathcal{D}, \Pi_\mathcal{D}, \mathbb{T}_\mathcal{D}, \nabla_\mathcal{D})$ where:*

1. *the set of discrete unknowns $X_\mathcal{D}$ is a finite dimensional vector space on $\mathbb{R}$,*
2. *the function reconstruction $\Pi_\mathcal{D} : X_\mathcal{D} \to L^p(\Omega)$ is linear,*
3. *the trace reconstruction $\mathbb{T}_\mathcal{D} : X_\mathcal{D} \to L^p(\partial\Omega)$ is linear,*
4. *the gradient reconstruction $\nabla_\mathcal{D} : X_\mathcal{D} \to L^p(\Omega)^d$ is linear,*
5. *the operators $\nabla_\mathcal{D}$ and $\mathbb{T}_\mathcal{D}$ are such that*

$$\|v\|_\mathcal{D} := \left(\|\nabla_\mathcal{D} v\|^p_{L^p(\Omega)^d} + \|\mathbb{T}_\mathcal{D} v\|^p_{L^p(\partial\Omega)}\right)^{1/p} \tag{3.51}$$

is a norm on $X_\mathcal{D}$.

The **coercivity**, **limit-conformity**, **compactness** and **piecewise constant reconstruction** for gradient discretisations for non-homogeneous Fourier conditions are defined exactly as for non-homogeneous Neumann conditions, *i.e.* Definitions 3.13, 3.14, 3.15 and 2.12 (with $X_{\mathcal{D},0}$ replaced by $X_{\mathcal{D}}$ in the latter, and using the norm (3.51) in all these definitions). The GD-consistency, however, must take into account the trace reconstruction.

Definition 3.37 (GD-consistency, non-homogeneous Fourier BCs)

If $\mathcal{D}$ is a gradient discretisation in the sense of Definition 3.36, define $S_{\mathcal{D}} : W^{1,p}(\Omega) \to [0, +\infty)$ by

$$\begin{aligned} &\forall \varphi \in W^{1,p}(\Omega)\,, \\ &S_{\mathcal{D}}(\varphi) = \min_{v \in X_{\mathcal{D}}} \big(\|\Pi_{\mathcal{D}} v - \varphi\|_{L^p(\Omega)} + \| \mathbb{T}_{\mathcal{D}} v - \gamma\varphi\|_{L^p(\partial\Omega)} \\ &\qquad\qquad + \|\nabla_{\mathcal{D}} v - \nabla\varphi\|_{L^p(\Omega)^d} \big). \end{aligned} \tag{3.52}$$

A sequence $(\mathcal{D}_m)_{m\in\mathbb{N}}$ of gradient discretisations in the sense of Definition 3.36 is **GD-consistent** if

$$\forall \varphi \in W^{1,p}(\Omega), \lim_{m\to\infty} S_{\mathcal{D}_m}(\varphi) = 0. \tag{3.53}$$

We notice that Lemma 3.7 (characterisation of GD-consistency using a dense set of $W^{1,p}(\Omega)$), Lemma 3.19 (characterisation of limit-conformity using a dense subset of $W^{p'}_{\mathrm{div},\partial}(\Omega)$) and Lemma 3.21 (regularity of the limit) still hold in the framework of gradient schemes for non-homogeneous Fourier boundary conditions. Similarly, the equivalent of Lemma 3.17 (limit-conformity implies coercivity) and Lemma 3.18 (compactness implies coercivity) are satisfied for sequences of GDs for Fourier boundary conditions.

For Fourier boundary conditions, the reconstructed trace has been included in the definition of $S_{\mathcal{D}}$, and we can therefore expect an approximation property as in Proposition 3.24. However, the norm is different and actually already includes the reconstructed trace. For this reason, an additional assumption must be introduced which states that the reconstructed trace can be controlled by the reconstructed function and gradients, see (3.54). In practice, for many gradient discretisation this assumption is easy to check by using Lemma B.21 and the notion of control by a polytopal toolbox (cf. Section 7.2.2).
The proof of this proposition is identical to the proof of Proposition 3.24, the assumption (3.54) playing the role of (3.19).

Proposition 3.38 (Approximation property of $\mathbb{T}_{\mathcal{D}}$ – Fourier BCs). *Let $\mathcal{D}$ be a gradient discretisation in the sense of Definition 3.36. We assume that there exists $\theta > 0$ such that*

$$\forall v \in X_{\mathcal{D}} \,:\, \|\mathbb{T}_{\mathcal{D}} v\|_{L^p(\partial\Omega)} \le \theta \left(\|\Pi_{\mathcal{D}} v\|_{L^p(\Omega)} + \|\nabla_{\mathcal{D}} v\|_{L^p(\Omega)^d} \right). \tag{3.54}$$

Then, for any $v \in X_{\mathcal{D}}$ and any $\varphi \in W^{1,p}(\Omega)$,

$$\|\mathbb{T}_{\mathcal{D}} v - \gamma\varphi\|_{L^p(\partial\Omega)} \leq \theta \left(\|\Pi_{\mathcal{D}} v - \varphi\|_{L^p(\Omega)} + \|\nabla_{\mathcal{D}} v - \nabla\varphi\|_{L^p(\Omega)^d} \right) + \max(1,\theta) S_{\mathcal{D}}(\varphi).$$

3.2.2 Gradient schemes for quasi-linear problems

We consider Problem (2.49a) with Fourier boundary conditions, that is

$$\begin{aligned} -\mathrm{div}(\Lambda(\boldsymbol{x},\overline{u})\nabla\overline{u}) &= f \text{ in } \Omega, \\ \Lambda(\boldsymbol{x},\overline{u})\nabla\overline{u}\cdot\boldsymbol{n} + b\overline{u} &= h \text{ on } \partial\Omega, \end{aligned} \tag{3.55}$$

- Ω is an open bounded connected subset of $\mathbb{R}^d$, $d \in \mathbb{N}^\star$, (3.56a)
- Λ is a Caratheodory function from $\Omega \times \mathbb{R}$ to $\mathcal{M}_d(\mathbb{R})$, (*i.e.* $\Lambda(\boldsymbol{x}, s)$ is measurable w.r.t. $\boldsymbol{x}$ and continuous w.r.t. s), there exists $\underline{\lambda}, \overline{\lambda} > 0$ such that, for a.e. $\boldsymbol{x} \in \Omega$, for all $s \in \mathbb{R}$, $\Lambda(\boldsymbol{x}, s)$ is symmetric with eigenvalues in $[\underline{\lambda}, \overline{\lambda}]$, (3.56b)
- $f \in L^2(\Omega)$, $h \in L^2(\partial\Omega)$, $b \in L^\infty(\partial\Omega)$ and there exists $\underline{b} > 0$ such that $b(\boldsymbol{x}) \geq \underline{b}$ for a.e. $\boldsymbol{x} \in \partial\Omega$. (3.56c)

A weak solution of this problem is:

$$\begin{aligned} &\overline{u} \in H^1(\Omega),\ \forall v \in H^1(\Omega), \\ &\int_\Omega \Lambda(\boldsymbol{x},\overline{u}(\boldsymbol{x}))\nabla\overline{u}(\boldsymbol{x})\cdot\nabla v(\boldsymbol{x})\mathrm{d}\boldsymbol{x} + \int_{\partial\Omega} b(\boldsymbol{x})\gamma(\overline{u})(\boldsymbol{x})\gamma(v)(\boldsymbol{x})\mathrm{d}\gamma(\boldsymbol{x}) \\ &\quad = \int_\Omega f(\boldsymbol{x})v(\boldsymbol{x})\mathrm{d}\boldsymbol{x} + \int_{\partial\Omega} h(\boldsymbol{x})\gamma(v)(\boldsymbol{x})\mathrm{d}\gamma(\boldsymbol{x}). \end{aligned} \tag{3.57}$$

The gradient scheme for this model is given in the following definition.

Definition 3.39 (GS, quasi-linear problem, Fourier BCs). *If $\mathcal{D} = (X_{\mathcal{D}}, \Pi_{\mathcal{D}}, \mathbb{T}_{\mathcal{D}}, \nabla_{\mathcal{D}})$ is a GD in the sense of Definition 3.36, then we define the related* gradient scheme *for* (3.57) *by*

Find $u \in X_{\mathcal{D}}$ such that, for any $v \in X_{\mathcal{D}}$,

$$\begin{aligned} &\int_\Omega \Lambda(\boldsymbol{x},\Pi_{\mathcal{D}} u(\boldsymbol{x}))\nabla_{\mathcal{D}} u(\boldsymbol{x})\cdot\nabla_{\mathcal{D}} v(\boldsymbol{x})\mathrm{d}\boldsymbol{x} + \int_{\partial\Omega} b(\boldsymbol{x})\mathbb{T}_{\mathcal{D}} u(\boldsymbol{x})\mathbb{T}_{\mathcal{D}} v(\boldsymbol{x})\mathrm{d}\gamma(\boldsymbol{x}) \\ &\quad = \int_\Omega f(\boldsymbol{x})\Pi_{\mathcal{D}} v(\boldsymbol{x})\mathrm{d}\boldsymbol{x} + \int_{\partial\Omega} h(\boldsymbol{x})\mathbb{T}_{\mathcal{D}}(v)(\boldsymbol{x})\mathrm{d}\gamma(\boldsymbol{x}). \end{aligned} \tag{3.58}$$

The convergence result is similar to the previous ones.

Theorem 3.40 (Convergence, quasi-linear problem, Fourier BCs). *Under assumptions (3.56), let $(\mathcal{D}_m)_{m\in\mathbb{N}}$ be a sequence of GDs in the sense of Definition 3.36, which is GD-consistent, limit-conforming and compact in the sense of Definitions 3.37, 3.14 and 3.15.*
Then, for any $m \in \mathbb{N}$, there exists at least one $u_m \in X_{\mathcal{D}_m}$ solution to the gradient scheme (3.58) and, up to a subsequence, $\Pi_{\mathcal{D}_m} u_m$ converges strongly in $L^2(\Omega)$ to a solution $\overline{u}$ of (3.57), $\nabla_{\mathcal{D}_m} u_m$ converges strongly in $L^2(\Omega)^d$ to $\nabla\overline{u}$, and $\mathbb{T}_{\mathcal{D}_m} u_m$ converges strongly in $L^2(\partial\Omega)$ to $\gamma\overline{u}$ as $m \to \infty$.
In the case where the solution $\overline{u}$ of (3.57) is unique, then the whole sequence converges to $\overline{u}$ as $m \to \infty$ in the senses above.

Proof. The proof is very similar to the proofs of Theorems 2.35 and 3.33, we only indicate here the elements which differ.
Letting $u = v$ in (3.58), by assumption on Λ and b and Definition 3.51 of $\|\cdot\|_{\mathcal{D}}$ we obtain

$$\begin{aligned}
\min(\underline{\lambda}, \underline{b})\, \|u\|_{\mathcal{D}}^2 &\leq \int_\Omega \Lambda(\boldsymbol{x}, \Pi_{\mathcal{D}} u(\boldsymbol{x})) \nabla_{\mathcal{D}} u(\boldsymbol{x}) \cdot \nabla_{\mathcal{D}} u(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \\
&\quad + \int_{\partial\Omega} b(\boldsymbol{x}) \mathbb{T}_{\mathcal{D}} u(\boldsymbol{x}) \mathbb{T}_{\mathcal{D}} u(\boldsymbol{x}) \mathrm{d}\gamma(\boldsymbol{x}) \\
&= \int_\Omega f(\boldsymbol{x}) \Pi_{\mathcal{D}} u(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} + \int_{\partial\Omega} h(\boldsymbol{x}) \mathbb{T}_{\mathcal{D}}(u)(\boldsymbol{x}) \mathrm{d}\gamma(\boldsymbol{x}) \\
&\leq \|f\|_{L^2(\Omega)} \|\Pi_{\mathcal{D}} u\|_{L^2(\Omega)} + \|h\|_{L^2(\partial\Omega)} \|\mathbb{T}_{\mathcal{D}}(u)\|_{L^2(\partial\Omega)} \\
&\leq C_{\mathcal{D}} (\|f\|_{L^2(\Omega)} + \|h\|_{L^2(\partial\Omega)}) \|u\|_{\mathcal{D}} .
\end{aligned}$$

This gives an estimate on $\|u\|_{\mathcal{D}}$ which allows us, as in the proof of Theorem 2.35, to use Brouwer's fixed point theorem to prove the existence of a solution to (3.58).
Noticing that $(\mathcal{D}_m)_{m\in\mathbb{N}}$ is coercive (Lemma 3.18), this estimate also shows that the solution u_m for $\mathcal{D} = \mathcal{D}_m$ is such that $\|u_m\|_{\mathcal{D}_m}$ remains bounded and therefore, using Lemma 3.21 and the compactness of the sequence of GDs, that, for some $\overline{u} \in H^1(\Omega)$,

$$\begin{aligned}
&\Pi_{\mathcal{D}_m} u_m \to \overline{u} \text{ strongly in } L^2(\Omega) \text{ and a.e.,} \\
&\mathbb{T}_{\mathcal{D}_m} u_m \to \gamma\overline{u} \text{ strongly in } L^2(\partial\Omega) \text{ and} \\
&\nabla_{\mathcal{D}_m} u_m \to \nabla\overline{u} \text{ weakly in } L^2(\Omega)^d.
\end{aligned} \tag{3.59}$$

Defining then $I_{\mathcal{D}} : H^1(\Omega) \to X_{\mathcal{D}}$ by

$$\begin{aligned}
I_{\mathcal{D}}\varphi = \operatorname*{argmin}_{v\in X_{\mathcal{D}}} \ &\big(\|\Pi_{\mathcal{D}} w - \varphi\|_{L^2(\Omega)} + \|\nabla_{\mathcal{D}} v - \nabla\varphi\|_{L^2(\Omega)^d} \\
&+ \|\mathbb{T}_{\mathcal{D}} v - \gamma\varphi\|_{L^2(\Omega)^d}\big),
\end{aligned}$$

the GD-consistency of the sequence of GDs shows that, for any $\varphi \in H^1(\Omega)$, $\Pi_{\mathcal{D}_m}(I_{\mathcal{D}_m}\varphi) \to \varphi$ strongly in $L^2(\Omega)$, $\nabla_{\mathcal{D}_m}(I_{\mathcal{D}_m}\varphi) \to \nabla\varphi$ strongly in $L^2(\Omega)^d$ and $\mathbb{T}_{\mathcal{D}_m}(I_{\mathcal{D}_m}\varphi) \to \gamma\varphi$ strongly in $L^2(\partial\Omega)$.

We can then, as in the proof of Theorem 2.35, use $v = I_{\mathcal{D}_m}\varphi$ in (3.58) and pass to the limit, thanks to these strong convergences and to (3.59), to see that $\overline{u}$ is a solution to (3.57).
We then take $v = u_m$ in (3.58) and pass to the limit, using the convergences (3.59) and the fact that $\overline{u}$ satisfies (3.57), to obtain

$$\begin{aligned}\lim_{m\to\infty}\int_\Omega &\Lambda(\boldsymbol{x}, \Pi_{\mathcal{D}_m}u_m(\boldsymbol{x}))\nabla_{\mathcal{D}_m}u_m(\boldsymbol{x})\cdot\nabla_{\mathcal{D}_m}u_m(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\\ &= \int_\Omega f(\boldsymbol{x})\overline{u}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} + \int_{\partial\Omega} h(\boldsymbol{x})\gamma\overline{u}(\boldsymbol{x})\mathrm{d}\gamma(\boldsymbol{x}) - \int_{\partial\Omega} b(\boldsymbol{x})\gamma\overline{u}(\boldsymbol{x})^2\mathrm{d}\gamma(\boldsymbol{x})\\ &= \int_\Omega \Lambda(\boldsymbol{x}, \overline{u}(\boldsymbol{x}))\nabla\overline{u}(\boldsymbol{x})\cdot\nabla\overline{u}(\boldsymbol{x})\mathrm{d}\boldsymbol{x}.\end{aligned}$$

By developing the following expression, this limit and (3.59) yield

$$\int_\Omega \Lambda(\boldsymbol{x}, \Pi_{\mathcal{D}_m}u_m(\boldsymbol{x}))(\nabla_{\mathcal{D}_m}u_m(\boldsymbol{x}) - \nabla\overline{u}(\boldsymbol{x}))\cdot(\nabla_{\mathcal{D}_m}u_m(\boldsymbol{x}) - \nabla\overline{u}(\boldsymbol{x}))\mathrm{d}\boldsymbol{x} \to 0.$$

By Assumptions (2.50b) on Λ, the left-hand side of this limit is greater than $\underline{\lambda}\,\|\nabla_{\mathcal{D}_m}u_m - \nabla\overline{u}\|^2_{L^2(\Omega)^d}$ and the strong convergence of the reconstructed gradients therefore follows. ∎

Remark 3.41. In the linear case (Λ independent of u), error estimates for (3.58) similar to the ones in Theorem 2.28 can be obtained, with an additional error estimate on the the traces.

3.3 Mixed boundary conditions

3.3.1 Gradient discretisations

From the framework of non-homogeneous Dirichlet and Neumann boundary conditions, it is very easy to construct a gradient scheme discretisation for mixed boundary conditions. We consider here $p \in (1, \infty)$, Ω a connected open bounded subset of $\mathbb{R}^d$ with Lipschitz boundary and we assume that

$$\begin{aligned}&\Gamma_d, \Gamma_n \text{ are two disjoint relatively open subsets of } \partial\Omega\\ &\text{such that } |\partial\Omega\backslash(\Gamma_d\cup\Gamma_n)| = 0 \text{ and } |\Gamma_d| > 0\end{aligned} \tag{3.60}$$

($|\cdot|$ denotes here the $(d-1)$-dimensional measure).

Definition 3.42 (GD, mixed BCs). *Under Assumption (3.60), a gradient discretisation $\mathcal{D}$ for mixed boundary conditions $\mathcal{D}$ is defined by $\mathcal{D} = (X_\mathcal{D}, \mathcal{I}_{\mathcal{D},\Gamma_d}, \Pi_\mathcal{D}, \mathbb{T}_{\mathcal{D},\Gamma_n}, \nabla_\mathcal{D})$ where:*

1. *the set of discrete unknowns $X_\mathcal{D} = X_{\mathcal{D},\Omega,\Gamma_n} \oplus X_{\mathcal{D},\Gamma_d}$ is the direct sum of two finite dimensional vector spaces on $\mathbb{R}$, corresponding respectively to the unknowns in Ω and on Γ_n and to the unknowns on Γ_d,*

2. *the linear mapping* $\mathcal{I}_{\mathcal{D},\Gamma_d} : W^{1-\frac{1}{p},p}(\partial\Omega) \to X_{\mathcal{D},\Gamma_d}$ *is an interpolation operator for the restrictions* $(\gamma u)_{|\Gamma_d}$ *of traces of elements* $u \in W^{1,p}(\Omega)$,
3. *the function reconstruction* $\Pi_{\mathcal{D}} : X_{\mathcal{D}} \to L^p(\Omega)$ *is linear,*
4. *the trace reconstruction* $\mathbb{T}_{\mathcal{D},\Gamma_n} : X_{\mathcal{D}} \to L^p(\Gamma_n)$ *is linear, and reconstructs from an element of* $X_{\mathcal{D}}$ *a function over* Γ_n,
5. *the gradient reconstruction* $\nabla_{\mathcal{D}} : X_{\mathcal{D}} \to L^p(\Omega)^d$ *is linear,*
6. *the operator* $\nabla_{\mathcal{D}}$ *is such that* $\|v\|_{\mathcal{D}} := \|\nabla_{\mathcal{D}} v\|_{L^p(\Omega)^d}$ *is a norm on* $X_{\mathcal{D},\Omega,\Gamma_n}$.

Definition 3.43 (Coercivity, mixed BCs)

Under Assumption (3.60), if $\mathcal{D}$ is a gradient discretisation in the sense of Definition 3.42, define

$$C_{\mathcal{D}} = \max_{v \in X_{\mathcal{D},\Omega,\Gamma_n}\setminus\{0\}} \left(\max \left\{ \frac{\|\Pi_{\mathcal{D}} v\|_{L^p(\Omega)}}{\|v\|_{\mathcal{D}}}, \frac{\|\mathbb{T}_{\mathcal{D},\Gamma_n} v\|_{L^p(\Gamma_n)}}{\|v\|_{\mathcal{D}}} \right\} \right). \tag{3.61}$$

A sequence $(\mathcal{D}_m)_{m\in\mathbb{N}}$ of gradient discretisations in the sense of Definition 3.42 is **coercive** if there exists $C_P \in \mathbb{R}_+$ such that $C_{\mathcal{D}_m} \le C_P$ for all $m \in \mathbb{N}$.

Definition 3.44 (GD-consistency, mixed BCs)

Under Assumption (3.60), if $\mathcal{D}$ is a gradient discretisation in the sense of Definition 3.42, define $S_{\mathcal{D}} : W^{1,p}(\Omega) \to [0, +\infty)$ by

$$\begin{aligned} &\forall \varphi \in W^{1,p}(\Omega)\,, \\ &S_{\mathcal{D}}(\varphi) = \min \Big\{ \|\Pi_{\mathcal{D}} v - \varphi\|_{L^p(\Omega)} + \|\nabla_{\mathcal{D}} v - \nabla\varphi\|_{L^p(\Omega)^d} : \\ &\qquad\qquad v \in X_{\mathcal{D}} \text{ such that } v - \mathcal{I}_{\mathcal{D},\Gamma_d}\gamma\varphi \in X_{\mathcal{D},\Omega,\Gamma_n} \Big\}. \end{aligned} \tag{3.62}$$

A sequence $(\mathcal{D}_m)_{m\in\mathbb{N}}$ of gradient discretisations in the sense of Definition 3.42 is **GD-consistent** if

$$\forall \varphi \in W^{1,p}(\Omega), \lim_{m\to\infty} S_{\mathcal{D}_m}(\varphi) = 0. \tag{3.63}$$

Definition 3.45 (Limit-conformity, mixed BCs)

For $p \in (1, +\infty)$, let $p' = \frac{p}{p-1}$ and

$$W^{p'}_{\mathrm{div},\Gamma_n}(\Omega) = \{\boldsymbol{\varphi} \in L^{p'}(\Omega)^d : \mathrm{div}\boldsymbol{\varphi} \in L^{p'}(\Omega)\,,\ \gamma_{\mathbf{n}}(\boldsymbol{\varphi}) \in L^{p'}(\Gamma_n)\}, \tag{3.64}$$

where $\gamma_{\mathbf{n}}(\boldsymbol{\varphi})$ is the normal trace of $\boldsymbol{\varphi}$ on $\partial\Omega$. Under Assumption (3.60), if $\mathcal{D}$ is a gradient discretisation in the sense of Definition 3.42, define $W_{\mathcal{D}}\colon W^{p'}_{\mathrm{div},\Gamma_n}(\Omega) \to [0,+\infty)$ by

$$\forall \boldsymbol{\varphi} \in W^{p'}_{\mathrm{div},\Gamma_n}(\Omega)\,,$$
$$W_{\mathcal{D}}(\boldsymbol{\varphi}) = \max_{v\in X_{\mathcal{D},\Omega,\Gamma_n}\setminus\{0\}} \frac{1}{\|v\|_{\mathcal{D}}} \left| \int_\Omega (\nabla_{\mathcal{D}} v(\boldsymbol{x})\cdot\boldsymbol{\varphi}(\boldsymbol{x}) + \Pi_{\mathcal{D}} v(\boldsymbol{x})\mathrm{div}\boldsymbol{\varphi}(\boldsymbol{x}))\,\mathrm{d}\boldsymbol{x} - \int_{\Gamma_n} \mathbb{T}_{\mathcal{D},\Gamma_n} v(\boldsymbol{x})\gamma_{\mathbf{n}}(\boldsymbol{\varphi})(\boldsymbol{x})\mathrm{d}\gamma(\boldsymbol{x}) \right|. \tag{3.65}$$

A sequence $(\mathcal{D}_m)_{m\in\mathbb{N}}$ of gradient discretisations in the sense of Definition 3.42 is **limit-conforming** if

$$\forall \boldsymbol{\varphi} \in W^{p'}_{\mathrm{div},\Gamma_n}(\Omega),\ \lim_{m\to\infty} W_{\mathcal{D}_m}(\boldsymbol{\varphi}) = 0. \tag{3.66}$$

Remark 3.46. Note that "$\gamma_{\mathbf{n}}(\boldsymbol{\varphi}) \in L^{p'}(\Gamma_n)$" makes sense because Γ_n is a relatively open subset of $\partial\Omega$. Indeed, when $\boldsymbol{\varphi} \in L^{p'}(\Omega)^d$ and $\mathrm{div}\boldsymbol{\varphi} \in L^{p'}(\Omega)$ then $\gamma_{\mathbf{n}}(\boldsymbol{\varphi}) \in (W^{1-\frac{1}{p},p}(\partial\Omega))'$ and saying that this linear form belongs to $L^{p'}(\Gamma_n)$ means by definition that there exists $g \in L^{p'}(\Gamma_n)$ such that, for any $w \in W^{1-\frac{1}{p},p}(\partial\Omega)$ with support in Γ_n,

$$\langle \gamma_{\mathbf{n}}(\boldsymbol{\varphi}), w\rangle_{(W^{1-\frac{1}{p},p}(\partial\Omega))',W^{1-\frac{1}{p},p}(\partial\Omega)} = \int_{\Gamma_n} g(\boldsymbol{x})w(\boldsymbol{x})\mathrm{d}\gamma(\boldsymbol{x}).$$

Definition 3.47 (Compactness, mixed BCs)

Under Assumption (3.60), a sequence $(\mathcal{D}_m)_{m\in\mathbb{N}}$ of gradient discretisations in the sense of Definition 3.42 is **compact** if, for any sequence $u_m \in X_{\mathcal{D}_m}$ such that $(\|u_m\|_{\mathcal{D}_m})_{m\in\mathbb{N}}$ is bounded, the sequences $(\Pi_{\mathcal{D}_m} u_m)_{m\in\mathbb{N}}$ and $(\mathbb{T}_{\mathcal{D}_m,\Gamma_n} u_m)_{m\in\mathbb{N}}$ are relatively compact in $L^p(\Omega)$ and $L^p(\Gamma_n)$, respectively.

The definition of **piecewise constant reconstruction** for a gradient discretisation for mixed boundary conditions is the same as Definition 2.12, replacing the space $X_{\mathcal{D},0}$ by $X_{\mathcal{D}}$.

The equivalent of Lemmas 2.15, 2.54 and 3.21 is the following lemma.

Lemma 3.48 (Regularity of the limit, mixed BCs). *Under Assumption (3.60), let $(\mathcal{D}_m)_{m\in\mathbb{N}}$ be a coercive, GD-consistent and limit-conforming sequence of gradient discretisations in the sense of Definitions 3.43, 3.44 and*

3.45. Let $g \in W^{1-\frac{1}{p},p}(\partial\Omega)$ and, for $m \in \mathbb{N}$, let $u_m \in X_{\mathcal{D}_m}$ be such that $u_m - \mathcal{I}_{\mathcal{D}_m,\Gamma_d} g \in X_{\mathcal{D},\Omega,\Gamma_n}$ and $(\|\nabla_{\mathcal{D}_m} u_m\|_{L^p(\Omega)^d})_{m\in\mathbb{N}}$ is bounded. Then, there exists $u \in W^{1,p}(\Omega)$ such that $\gamma u = g$ on Γ_d and, up to a subsequence, as $m \to \infty$,

$$\begin{aligned} &\Pi_{\mathcal{D}_m} u_m \to u \text{ weakly in } L^p(\Omega), \\ &\nabla_{\mathcal{D}_m} u_m \to \nabla u \text{ weakly in } L^p(\Omega)^d. \end{aligned} \tag{3.67}$$

If we assume moreover that $g = 0$, or that there exists $\varphi_g \in W^{1,p}(\Omega)$ such that $\gamma\varphi_g = g$ and, as $m \to \infty$,

$$\begin{aligned} \min \big\{ &\|\Pi_{\mathcal{D}_m} v - \varphi_g\|_{L^p(\Omega)} + \|\mathbb{T}_{\mathcal{D}_m,\Gamma_n} v - \gamma\varphi_g\|_{L^p(\Gamma_n)} \\ &+ \|\nabla_{\mathcal{D}_m} v - \nabla\varphi_g\|_{L^p(\Omega)^d} \,:\, v - \mathcal{I}_{\mathcal{D}_m,\Gamma_d}\gamma\varphi_g \in X_{\mathcal{D}_m,\Omega,\Gamma_n} \big\} \to 0, \end{aligned} \tag{3.68}$$

then we also have

$$\mathbb{T}_{\mathcal{D}_m,\Gamma_n} u_m \to (\gamma u)_{|\Gamma_n} \text{ weakly in } L^p(\Gamma_n). \tag{3.69}$$

Remark 3.49. Assumption (3.68), if satisfied for all $\varphi_g \in W^{1,p}(\Omega)$, is similar to the GD-consistency of $(\mathcal{D}_m)_{m\in\mathbb{N}}$ in the sense of Fourier boundary conditions (but with a trace reconstruction only on Γ_n).

Proof.
Step 1: we suppose that $g = 0$.
Since $u_m = u_m - \mathcal{I}_{\mathcal{D}_m,\Gamma_d} g \in X_{\mathcal{D}_m,\Omega,\Gamma_n}$, by coercivity of $(\mathcal{D}_m)_{m\in\mathbb{N}}$ the bound on $\|u_m\|_{\mathcal{D}_m}$ shows that $\|\Pi_{\mathcal{D}_m} u_m\|_{L^p(\Omega)}$, $\|\mathbb{T}_{\mathcal{D}_m,\Gamma_n} u_m\|_{L^p(\Gamma_n)}$ and $\|\nabla_{\mathcal{D}_m} u_m\|_{L^p(\Omega)^d}$ are bounded. There exists therefore $u \in L^p(\Omega)$, $w \in L^p(\Gamma_n)$ and $\mathbf{v} \in L^p(\Omega)^d$ such that, up to a subsequence as $m \to \infty$, $\Pi_{\mathcal{D}_m} u_m \to u$ weakly in $L^p(\Omega)$, $\mathbb{T}_{\mathcal{D}_m,\Gamma_n} u_m \to w$ weakly in $L^p(\Gamma_n)$, and $\nabla_{\mathcal{D}_m} u_m \to \mathbf{v}$ weakly in $L^p(\Omega)^d$.
Let $\boldsymbol{\varphi} \in W^{p'}_{\mathrm{div},\Gamma_n}(\Omega)$. By Definition (3.65) of $W_{\mathcal{D}}$, passing to the limit in

$$\begin{aligned} ||u_m||_{\mathcal{D}_m} W_{\mathcal{D}_m}(\boldsymbol{\varphi}) \geq \Bigg| &\int_\Omega (\nabla_{\mathcal{D}_m} u_m(\boldsymbol{x}) \cdot \boldsymbol{\varphi}(\boldsymbol{x}) + \Pi_{\mathcal{D}_m} u(\boldsymbol{x}) \mathrm{div}\boldsymbol{\varphi}(\boldsymbol{x})) \mathrm{d}\boldsymbol{x} \\ &- \int_{\Gamma_n} \mathbb{T}_{\mathcal{D}_m,\Gamma_n} u_m(\boldsymbol{x}) \gamma_{\mathbf{n}}(\boldsymbol{\varphi})(\boldsymbol{x}) \mathrm{d}\gamma(\boldsymbol{x}) \Bigg| \end{aligned}$$

yields

$$\int_\Omega (\mathbf{v}(\boldsymbol{x}) \cdot \boldsymbol{\varphi}(\boldsymbol{x}) + u(\boldsymbol{x}) \mathrm{div}\boldsymbol{\varphi}(\boldsymbol{x})) \mathrm{d}\boldsymbol{x} - \int_{\Gamma_n} w(\boldsymbol{x}) \gamma_{\boldsymbol{n}}(\boldsymbol{\varphi})(\boldsymbol{x}) \mathrm{d}\gamma(\boldsymbol{x}) = 0.$$

Selecting $\boldsymbol{\varphi} \in C_c^\infty(\Omega)^d$ gives $\mathbf{v} = \nabla u$, and therefore $u \in W^{1,p}(\Omega)$. Taking then $\boldsymbol{\varphi}$ smooth that does not vanish on $\partial\Omega$ and using an integration-by-parts, we obtain

$$\int_{\partial\Omega} \gamma_{\mathbf{n}}(\boldsymbol{\varphi})(\boldsymbol{x}) \gamma u(\boldsymbol{x}) \mathrm{d}\gamma(\boldsymbol{x}) = \int_{\Gamma_n} w(\boldsymbol{x}) \gamma_{\mathbf{n}}(\boldsymbol{\varphi})(\boldsymbol{x}) \mathrm{d}\gamma(\boldsymbol{x}).$$

This shows that $\gamma u = w$ on Γ_n and that $\gamma u = 0$ on Γ_d, which concludes the proof of (3.67) and (3.69) if $g = 0$.

Step 2: we consider a general $g \in W^{1-\frac{1}{p},p}(\partial\Omega)$.
As in the proof of Lemma 2.54, we take an extension $\widetilde{g} \in W^{1,p}(\Omega)$ of g and we use the GD-consistency to find $v_m \in X_{\mathcal{D}_m}$ such that $v_m - \mathcal{I}_{\mathcal{D}_m,\Gamma_d} g \in X_{\mathcal{D}_m,\Omega,\Gamma_n}$, $\Pi_{\mathcal{D}_m} v_m \to \widetilde{g}$ in $L^p(\Omega)$ and $\nabla_{\mathcal{D}_m} v_m \to \nabla\widetilde{g}$ in $L^p(\Omega)^d$. Then $u_m - v_m \in X_{\mathcal{D}_m,\Omega,\Gamma_n}$ and we can apply the reasoning in Step 1 to this function. We therefore find $U \in W^{1,p}(\Omega)$ such that $\gamma U = 0$ on Γ_d and, up to a subsequence,

$$\begin{aligned}
&\Pi_{\mathcal{D}_m}(u_m - v_m) \to U \text{ weakly in } L^p(\Omega)\,,\\
&\nabla_{\mathcal{D}_m}(u_m - v_m) \to \nabla U \text{ weakly in } L^p(\Omega)^d\,,\\
&\mathbb{T}_{\mathcal{D}_m,\Gamma_n}(u_m - v_m) \to \gamma U \text{ weakly in } L^p(\Gamma_n).
\end{aligned} \tag{3.70}$$

We then let $u = U + \widetilde{g} \in W^{1,p}(\Omega)$, so that $\gamma u = g$ on Γ_d. The convergence properties of $(v_m)_{m\in\mathbb{N}}$ and (3.70) then show that (3.67) holds.

Step 3: we consider a general $g \in W^{1-\frac{1}{p},p}(\partial\Omega)$, and we assume that (3.68) holds.
Then we can take $v_m \in X_{\mathcal{D}_m}$ such that $v_m - \mathcal{I}_{\mathcal{D}_m,\Gamma_d} g \in X_{\mathcal{D}_m,\Omega,\Gamma_n}$, $\Pi_{\mathcal{D}_m} v_m \to \widetilde{g}$ in $L^p(\Omega)$, $\mathbb{T}_{\mathcal{D}_m,\Gamma_n} v_m \to \gamma\widetilde{g} = g$ in $L^p(\Gamma_n)$, and $\nabla_{\mathcal{D}_m} v_m \to \nabla\widetilde{g}$ in $L^p(\Omega)^d$. We can then reproduce Step 2 with this v_m. Since $\mathbb{T}_{\mathcal{D}_m,\Gamma_n} v_m \to g$ in $L^p(\Omega)$, the convergence $\mathbb{T}_{\mathcal{D}_m,\Gamma_n}(u_m - v_m) \to \gamma U = \gamma u - g$ in $L^p(\Gamma_n)$-weak (see (3.70)) ensures that (3.69) holds. ∎

3.3.2 Gradient schemes for linear problems

We only present gradient schemes for mixed BCs on the linear model

$$\begin{aligned}
-\mathrm{div}(\Lambda\nabla\overline{u}) &= f + \mathrm{div}(\boldsymbol{F}) \text{ in } \Omega,\\
\overline{u} &= g \text{ on } \Gamma_d,\\
\Lambda\nabla\overline{u}\cdot\mathbf{n} + \boldsymbol{F}\cdot\mathbf{n} &= h \text{ on } \Gamma_n,
\end{aligned} \tag{3.71}$$

under Assumption (3.60) and

- Ω is an open bounded connected subset of $\mathbb{R}^d$ ($d \in \mathbb{N}^\star$), with Lipschitz boundary, (3.72a)
- Λ is a measurable function from Ω to the set of $d \times d$ symmetric matrices and there exists $\underline{\lambda}, \overline{\lambda} > 0$ such that, for a.e. $\boldsymbol{x} \in \Omega$, $\Lambda(\boldsymbol{x})$ has eigenvalues in $[\underline{\lambda}, \overline{\lambda}]$, (3.72b)
- $f \in L^2(\Omega)\,,\ \boldsymbol{F} \in L^2(\Omega)^d\,,\ g \in H^{1/2}(\partial\Omega)\,,\ h \in L^2(\Gamma_n)$. (3.72c)

Denoting by $H^1_{\Gamma_d}(\Omega)$ the set of functions in $H^1(\Omega)$ whose trace on Γ_d vanishes, the weak formulation of (3.71) is

$$\begin{aligned}
&\overline{u} \in \{w \in H^1(\Omega) \,:\, \gamma(w) = g \text{ on } \Gamma_d\},\ \forall v \in H^1_{\Gamma_d}(\Omega),\\
&\int_\Omega \Lambda(\boldsymbol{x})\nabla\overline{u}(\boldsymbol{x})\cdot\nabla v(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = \int_\Omega f(\boldsymbol{x})v(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\\
&\qquad - \int_\Omega \boldsymbol{F}(\boldsymbol{x})\cdot\nabla v(\boldsymbol{x})\mathrm{d}\boldsymbol{x} + \int_{\Gamma_n} h(\boldsymbol{x})\gamma(v)(\boldsymbol{x})\mathrm{d}\gamma(\boldsymbol{x}).
\end{aligned} \tag{3.73}$$

The GDM applied to this mixed problem yields the following scheme.

Definition 3.50 (GS, mixed BCs). *If $\mathcal{D} = (X_\mathcal{D}, \mathcal{I}_{\mathcal{D},\Gamma_d}, \Pi_\mathcal{D}, \mathbb{T}_{\mathcal{D},\Gamma_n}, \nabla_\mathcal{D})$ is a GD for mixed problems in the sense of Definition 3.42, then the related* gradient scheme *for (3.73) is defined by:*

$$\begin{aligned}
&\textit{Find } u \in \mathcal{I}_{\mathcal{D},\Gamma_d} g + X_{\mathcal{D},\Omega,\Gamma_n} \textit{ such that, for any } v \in X_{\mathcal{D},\Omega,\Gamma_n},\\
&\int_\Omega \Lambda(\boldsymbol{x})\nabla_\mathcal{D} u(\boldsymbol{x})\cdot\nabla_\mathcal{D} v(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = \int_\Omega f(\boldsymbol{x})\Pi_\mathcal{D} v(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\\
&\qquad - \int_\Omega \boldsymbol{F}(\boldsymbol{x})\cdot\nabla_\mathcal{D} v(\boldsymbol{x})\mathrm{d}\boldsymbol{x} + \int_{\Gamma_n} h(\boldsymbol{x})\mathbb{T}_{\mathcal{D},\Gamma_n} v(\boldsymbol{x})\mathrm{d}\gamma(\boldsymbol{x}).
\end{aligned} \tag{3.74}$$

The proof of the following error estimates for mixed boundary conditions is similar to the proofs made in the case of other boundary conditions (see Theorems 2.28, 2.56 and 3.30). Likewise, a convergence result similar to Corollary 2.31 follows from these error estimates.

Theorem 3.51 (Error estimate, mixed BCs). *Under Hypotheses (3.72), let $\overline{u} \in H^1(\Omega)$ be the solution of (3.73) (remark that since $f \in L^2(\Omega)$ and $h \in L^2(\Gamma_n)$, we have $\Lambda\nabla\overline{u} + \boldsymbol{F} \in W^2_{\mathrm{div},\Gamma_n}(\Omega)$, see (3.64)).*
Let $\mathcal{D}$ be a GD for mixed boundary conditions in the sense of Definition 3.42. Then there exists one and only one $u_\mathcal{D} \in X_\mathcal{D}$ solution to the gradient scheme (3.74), and this element satisfies the following inequalities:

$$\begin{aligned}
\|\nabla\overline{u} - \nabla_\mathcal{D} u_\mathcal{D}\|_{L^2(\Omega)^d} &\le \frac{1}{\underline{\lambda}}\left[W_\mathcal{D}(\Lambda\nabla\overline{u} + \boldsymbol{F}) + (\overline{\lambda} + \underline{\lambda})S_\mathcal{D}(\overline{u})\right],\\
\|\overline{u} - \Pi_\mathcal{D} u_\mathcal{D}\|_{L^2(\Omega)} &\le \frac{1}{\underline{\lambda}}\left[C_\mathcal{D} W_\mathcal{D}(\Lambda\nabla\overline{u} + \boldsymbol{F}) + (C_\mathcal{D}\overline{\lambda} + \underline{\lambda})S_\mathcal{D}(\overline{u})\right],
\end{aligned}$$

where $C_\mathcal{D}$, $S_\mathcal{D}$ and $W_\mathcal{D}$ are defined by (3.61), (3.62) and (3.65).

Remark 3.52. In this case again, a lower bound in the spirit of (2.29) could be derived.

Remark 3.53 (Other linear and non-linear models)
In [79], the notion of GD for mixed boundary conditions is extended to cover linear and non-linear elasticity models. It is also proved that several schemes designed for these models, including schemes built to be accurate in the near-incompressible limit, are GDMs. The time-dependent poro-elasticity model is covered in [125].
As other examples of models with mixed BCs that are covered by the GDM, we can mention linear and non-linear variational inequalities [10, 11].

Part II

Parabolic problems

4

Time-dependent problems

In this chapter, the definition of gradient discretisations (GDs) for time-dependent problems is first given; it is followed by compactness results for the analysis of such problems. These results include discrete Arzelà–Ascoli and Aubin–Simon theorems, and are presented first in a general setting, before their consequences for gradient discretisations are discussed.
To deal with all kinds of boundary conditions at once, the notation $X_{\mathcal{D},\bullet}$ stands for $X_{\mathcal{D},0}$ in the case of homogeneous Dirichlet boundary conditions, and for $X_{\mathcal{D}}$ in the case of other boundary conditions. Similarly, we write $W^{1,p}_{\bullet}(\Omega)$ for $W^{1,p}_0(\Omega)$ in the case of homogeneous Dirichlet boundary conditions, and $W^{1,p}(\Omega)$ in the case of other boundary conditions.

4.1 Space–time gradient discretisation

To fix ideas, let us consider a general time-dependent problem under the form $\partial_t u + A(u) = f$, over a domain $\Omega \times (0,T)$ with $T > 0$. Adequate boundary conditions and initial conditions are also assumed. If $\theta \in [0,1]$ and $t^{(0)} = 0 < t^{(1)} < \cdots < t^{(N)} = T$ is a set of time points, then the θ-scheme reads: for all $n = 0, \ldots, N-1$,

$$\frac{u^{(n+1)} - u^{(n)}}{t^{(n+1)} - t^{(n)}} + A(\theta u^{(n+1)} + (1-\theta)u^{(n)}) = f^{(n)}. \tag{4.1}$$

For $\theta = 1$, the scheme is Euler implicit (or "backward"), for $\theta = 0$ it is Euler explicit (or "forward"); both choices are of order 1 in time. For $\theta = \frac{1}{2}$ and A linear, $A(\theta u^{(n+1)} + (1-\theta)u^{(n)}) = \frac{A(u^{(n+1)}) + A(u^{(n)})}{2}$ and we recover the Crank-Nicolson time-stepping, which is of order 2; numerical experiments show that, for many non-linear models, the scheme (4.1) with $\theta = \frac{1}{2}$ is usually more accurate than with $\theta = 1$ or $\theta = 0$. Values $\theta \in [\frac{1}{2}, 1]$ yield an L^2 stability for several non-linear models, as shown in our analysis; they are the most frequently considered in this book due to the parabolic nature of the equations under study.

Definition 4.1 (GD for time-dependent problems). *Let $p \in (1, +\infty)$, Ω be an open subset of $\mathbb{R}^d$ (with $d \in \mathbb{N}^\star$), $T > 0$ and $\theta \in [0, 1]$. We say that $\mathcal{D}_T = (\mathcal{D}, \mathcal{I}_\mathcal{D}, (t^{(n)})_{n=0,\dots,N})$ is a space–time gradient discretisation if*

- *$\mathcal{D} = (X_{\mathcal{D},\bullet}, \Pi_\mathcal{D}, \nabla_\mathcal{D}, \dots)$ is a GD in the sense of Definition 2.1 (resp. Definition 2.49, 3.1, 3.11, 3.36 or 3.42 – depending on the considered boundary conditions), which satisfies $\Pi_\mathcal{D}(X_{\mathcal{D},\bullet}) \subset L^{\max(p,2)}(\Omega)$,*
- *$\mathcal{I}_\mathcal{D} : L^2(\Omega) \to X_{\mathcal{D},\bullet}$ is an interpolation operator,*
- *$t^{(0)} = 0 < t^{(1)} \dots < t^{(N)} = T$.*

The gradient discretisation $\mathcal{D}$ is called the underlying spatial discretisation *of $\mathcal{D}_T$. We then set $\eth t^{(n+\frac{1}{2})} = t^{(n+1)} - t^{(n)}$, for $n = 0, \dots, N-1$, and $\eth t_\mathcal{D} = \max_{n=0,\dots,N-1} \eth t^{(n+\frac{1}{2})}$. To a family $v = (v^{(n)})_{n=0,\dots,N} \in X_{\mathcal{D},\bullet}^{N+1}$ we associate the piecewise-constant-in-time functions $v_\theta \in L^\infty(0, T; X_{\mathcal{D},\bullet})$, $\Pi_\mathcal{D}^{(\theta)} v \in L^\infty(0, T; L^{\max(p,2)}(\Omega))$, $\nabla_\mathcal{D}^{(\theta)} v \in L^\infty(0, T; L^p(\Omega)^d)$ and $\mathbb{T}_\mathcal{D}^{(\theta)} v \in L^\infty(0, T; L^p(\partial\Omega))$ defined by:*

$$\begin{aligned}
&\forall n = 0, \dots, N-1, \text{ for all } t \in (t^{(n)}, t^{(n+1)}], \\
&\quad v_\theta(t) = v^{(n+\theta)} := \theta v^{(n+1)} + (1-\theta) v^{(n)} \text{ and, for a.e. } \boldsymbol{x} \in \Omega, \\
&\quad \Pi_\mathcal{D}^{(\theta)} v(\boldsymbol{x}, t) = \Pi_\mathcal{D}[v_\theta(t)](\boldsymbol{x}), \ \nabla_\mathcal{D}^{(\theta)} v(\boldsymbol{x}, t) = \nabla_\mathcal{D}[v_\theta(t)](\boldsymbol{x}) \text{ and} \\
&\quad \mathbb{T}_\mathcal{D}^{(\theta)} v(\boldsymbol{x}, t) = \mathbb{T}_\mathcal{D}[v_\theta(t)](\boldsymbol{x}).
\end{aligned} \tag{4.2}$$

To state uniform-in-time convergence results, we also need to extend the definition of $\Pi_\mathcal{D}^{(\theta)} v$ up to $t = 0$:

$$\text{For a.e. } \boldsymbol{x} \in \Omega, \ \Pi_\mathcal{D}^{(\theta)} v(\boldsymbol{x}, 0) = \Pi_\mathcal{D} v^{(0)}(\boldsymbol{x}). \tag{4.3}$$

If $v \in X_{\mathcal{D},\bullet}^{N+1}$, we define $\delta_\mathcal{D} v \in L^\infty(0, T; L^{\max(p,2)}(\Omega))$ by

$$\begin{aligned}
&\forall n = 0, \dots, N-1, \text{ for a.e. } t \in (t^{(n)}, t^{(n+1)}), \\
&\quad \delta_\mathcal{D} v(t) = \delta_\mathcal{D}^{(n+\frac{1}{2})} v := \frac{\Pi_\mathcal{D} v^{(n+1)} - \Pi_\mathcal{D} v^{(n)}}{\eth t^{(n+\frac{1}{2})}}.
\end{aligned} \tag{4.4}$$

Remark 4.2. The iterative definition (4.1) requires the initialisation step, a way to compute $u^{(0)}$. The interpolation operator $\mathcal{I}_\mathcal{D}$ applied to the initial condition describes this initialisation of $u^{(0)}$ (cf., *e.g.*, (5.5) in Section 5.1).

Definition 4.3 (Space–time-consistency for space–time GD)

For $T > 0$ and $\theta \in [0, 1]$, if $\mathcal{D}_T$ is a space–time GD in the sense of Definition 4.1, we define $\widehat{S}_\mathcal{D}$ by (2.2), (2.94), (3.3), (3.52) or (3.62) (depending on the considered boundary conditions), where $W_\bullet^{1,p}(\Omega)$ is

replaced with $W^{1,p}_\bullet(\Omega) \cap L^2(\Omega)$ and $\|\Pi_\mathcal{D} v - \varphi\|_{L^p(\Omega)}$ is replaced with $\|\Pi_\mathcal{D} v - \varphi\|_{L^{\max(p,2)}(\Omega)}$.
A sequence $((\mathcal{D}_T)_m)_{m\in\mathbb{N}}$ of space–time GDs in the sense of Definition 4.1, with underlying spatial discretisations $(\mathcal{D}_m)_{m\in\mathbb{N}}$, is said to be **space–time-consistent** if

$$\begin{aligned}
&\forall \varphi \in W^{1,p}_\bullet(\Omega) \cap L^2(\Omega),\ \lim_{m\to\infty} \widehat{S}_{\mathcal{D}_m}(\varphi) = 0,\\
&\forall u \in L^2(\Omega),\ \lim_{m\to\infty} \|u - \Pi_{\mathcal{D}_m}\mathcal{I}_{\mathcal{D}_m} u\|_{L^2(\Omega)} = 0,\\
&\delta t_{\mathcal{D}_m} \to 0 \text{ as } m \to \infty.
\end{aligned} \tag{4.5}$$

Remark 4.4 (A generic definition of $\mathcal{I}_\mathcal{D}$)
Given a spatial gradient discretisation $\mathcal{D}$ such that $\Pi_\mathcal{D}(X_{\mathcal{D},\bullet}) \subset L^2(\Omega)$, we can define an interpolator $\mathcal{I}_\mathcal{D} : L^2(\Omega) \to X_{\mathcal{D},\bullet}$ by

$$\begin{aligned}
&\forall u \in L^2(\Omega)\,,\\
&\mathcal{I}_\mathcal{D} u = \operatorname{argmin}\,\{\|v\|_\mathcal{D} : v \in X_{\mathcal{D},\bullet}\,,\ \Pi_\mathcal{D} v = \mathrm{Pr}_{\Pi_\mathcal{D}(X_{\mathcal{D},\bullet})} u\},
\end{aligned} \tag{4.6}$$

where $\mathrm{Pr}_{\Pi_\mathcal{D}(X_{\mathcal{D},\bullet})} : L^2(\Omega) \to \Pi_\mathcal{D}(X_{\mathcal{D},\bullet})$ is the $L^2(\Omega)$-orthogonal projector on $\Pi_\mathcal{D}(X_{\mathcal{D},\bullet})$. Since the norm $\|\cdot\|_\mathcal{D}$ is uniformly convex (see the definitions in Chapters 2 and 3, depending on the various boundary conditions), the argmin in (4.6) is indeed unique, and we can even check that $\mathcal{I}_\mathcal{D} : L^2(\Omega) \to X_{\mathcal{D},\bullet}$ is linear continuous (although this is not required in Definition 4.1).
Consider now a sequence $(\mathcal{D}_m)_{m\in\mathbb{N}}$ of spatial GDs, such that, as $m \to \infty$, $\widehat{S}_{\mathcal{D}_m}(u) \to 0$ for all $u \in W^{1,p}_\bullet(\Omega) \cap L^2(\Omega)$ (this is an improved consistency property of $(\mathcal{D}_m)_{m\in\mathbb{N}}$). This shows that, for such an u, there exists $u_m \in X_{\mathcal{D}_m,\bullet}$ such that $\|\Pi_{\mathcal{D}_m} u_m - u\|_{L^2(\Omega)} \to 0$. The definition (4.6) yields $\Pi_{\mathcal{D}_m}\mathcal{I}_{\mathcal{D}_m} = \mathrm{Pr}_{\Pi_{\mathcal{D}_m}(X_{\mathcal{D}_m,\bullet})}$, and thus, by the properties of the orthogonal projector,

$$\begin{aligned}
\|u - \Pi_{\mathcal{D}_m}\mathcal{I}_{\mathcal{D}_m} u\|_{L^2(\Omega)} &= \left\|u - \mathrm{Pr}_{\Pi_{\mathcal{D}_m}(X_{\mathcal{D}_m,\bullet})} u\right\|_{L^2(\Omega)}\\
&\le \|u - \Pi_{\mathcal{D}_m} u_m\|_{L^2(\Omega)} \to 0 \text{ as } m \to \infty.
\end{aligned}$$

Hence, (4.5) holds for $u \in W^{1,p}_\bullet(\Omega) \cap L^2(\Omega)$. Since the mapping $\Pi_{\mathcal{D}_m}\mathcal{I}_{\mathcal{D}_m} = \mathrm{Pr}_{\Pi_{\mathcal{D}_m}(X_{\mathcal{D}_m,\bullet})} : L^2(\Omega) \to L^2(\Omega)$ has norm 1, reasoning by density of $W^{1,p}_\bullet(\Omega) \cap L^2(\Omega)$ into $L^2(\Omega)$ shows that (4.5) actually holds for all $u \in L^2(\Omega)$.

Remark 4.5. To illustrate the definition of $\widehat{S}_\mathcal{D}$, here is how it looks for Fourier boundary conditions:

$$\begin{aligned}
&\forall \varphi \in W^{1,p}(\Omega) \cap L^2(\Omega)\,,\\
&\widehat{S}_\mathcal{D}(\varphi) = \min_{v\in X_\mathcal{D}} \big(\|\Pi_\mathcal{D} v - \varphi\|_{L^{\max(p,2)}(\Omega)} + \|\mathbb{T}_\mathcal{D} v - \gamma\varphi\|_{L^p(\partial\Omega)}\\
&\qquad\qquad + \|\nabla_\mathcal{D} v - \nabla\varphi\|_{L^p(\Omega)^d}\big).
\end{aligned}$$

The notions of coercivity, limit-conformity and compactness for sequences of space–time GDs boil down to the corresponding notion for the sequence of underlying spatial discretisations.

Definition 4.6 (Coercivity, limit-conformity and compactness for space–time GDs)

Let $T > 0$ and $((\mathcal{D}_T)_m)_{m\in\mathbb{N}}$ be a sequence of space–time GDs in the sense of Definition 4.1, with underlying spatial discretisations $(\mathcal{D}_m)_{m\in\mathbb{N}}$.
The sequence $((\mathcal{D}_T)_m)_{m\in\mathbb{N}}$ is coercive (resp. limit-conforming, resp. compact) if the sequence $(\mathcal{D}_m)_{m\in\mathbb{N}}$ is coercive (resp. limit-conforming, resp. compact) for the corresponding boundary conditions.

Remark 4.7 (One GD per time step)
In some instances, such as when using time-varying meshes or for coupled systems (see, e.g., [47, Section 4.1.1]), it might be required to consider one spatial GD for each time step. In this case, a space-time GD is $\mathcal{D}_T = (\mathcal{D}^S, \mathcal{I}_\mathcal{D}, (t^{(n)})_{n=0,\dots,N})$ where $\mathcal{D}^S = (\mathcal{D}^n)_{n=1,\dots,N}$ is a family of spatial GDs for the considered boundary conditions. The coercivity, GD-consistency and limit-conformity of a sequence $(\mathcal{D}_m^S)_{m\in\mathbb{N}} = ((\mathcal{D}_m^n)_{n=1,\dots,N_m})_{m\in\mathbb{N}}$ of such families of GDs should then be defined by using

$$C_{\mathcal{D}_m^S} = \max_{n=1,\dots,N_m} C_{\mathcal{D}_m^n}\,, \quad S_{\mathcal{D}_m^S} = \max_{n=1,\dots,N_m} S_{\mathcal{D}_m^n} \quad \text{and} \quad W_{\mathcal{D}_m^S} = \max_{n=1,\dots,N_m} W_{\mathcal{D}_m^n}\,.$$

The compactness of $(\mathcal{D}_m^S)_{m\in\mathbb{N}}$ is defined as the compactness of the function reconstructions (and possibly the trace reconstructions, depending on the boundary conditions) of any family $((u_m^n)_{n=1,\dots,N_m})_{m\in\mathbb{N}}$ with $u_m^n \in \mathcal{D}_m^n$ such that, for some C not depending on n or m, $\|u_m^n\|_{\mathcal{D}_m^n} \le C$.
After modifying the time-space reconstructions (4.2)–(4.4) by using the GD appropriate to each time step, the analysis carried out here and in Chapters 5 and 6 can easily be adapted to this setting of space–time GDs with time-varying spatial GDs.

The following lemma is the counterpart of Lemma 2.15 and Lemma 3.21. We could as easily state counterparts of the regularity of the limit for non-homogeneous Dirichlet boundary conditions or mixed boundary conditions (as in Lemma 2.54 or Lemma 3.48).

Lemma 4.8 (Regularity of the limit, space–time problems). *Let $p \in (1,\infty)$ and $((\mathcal{D}_T)_m)_{m\in\mathbb{N}}$ be a limit-conforming sequence of space–time GDs, in the sense of Definition 4.6, for homogeneous Dirichlet or non-homogeneous Neumann boundary conditions. Let $\theta \in [0,1]$, $q \in (1,+\infty)$ and take, for any $m \in \mathbb{N}$, $u_m \in X_{\mathcal{D}_m,\bullet}^{N_m+1}$ such that $(\|(u_m)_\theta\|_{L^q(0,T;X_{\mathcal{D}_m,\bullet})})_{m\in\mathbb{N}}$ is bounded.*

Then there exists $u \in L^q(0,T;W^{1,p}_\bullet(\Omega))$ *such that, up to a subsequence,* $\Pi^{(\theta)}_{\mathcal{D}_m} u_m \to u$ *weakly in* $L^q(0,T;L^p(\Omega))$ *and* $\nabla^{(\theta)}_{\mathcal{D}_m} u_m \to \nabla u$ *weakly in* $L^q(0,T;L^p(\Omega))^d$.
In the case of non-homogeneous Neumann boundary conditions, we also have, up to a subsequence, $\mathbb{T}^{(\theta)}_{\mathcal{D}_m} u_m \to \gamma u$ *weakly in* $L^q(0,T;L^p(\partial\Omega))$.
The same conclusions hold in the case $q = +\infty$*, provided that the weak convergences are replaced with weak-*$*$ *convergences.*

Remark 4.9. Note that each space $X_{\mathcal{D}_m,\bullet}$ is endowed with its natural norm $\|\cdot\|_{\mathcal{D}_m}$, *i.e.* $\|\nabla_{\mathcal{D}_m}\cdot\|_{L^p(\Omega)^d}$ for Dirichlet boundary conditions and (3.1) for Neumann boundary conditions. For $q < +\infty$, a bound on $\|(u_m)_\theta\|_{L^q(0,T;X_{\mathcal{D}_m,\bullet})}$ is therefore a bound on

$$\left(\int_0^T \|(u_m)_\theta(t)\|^q_{\mathcal{D}_m}\,\mathrm{d}t\right)^{1/q}.$$

Proof of Lemma 4.8. We only consider the case of homogeneous Dirichlet boundary conditions; the case of non homogeneous Neumann boundary conditions can handled similarly by following the proof of Lemma 3.21.
By Lemma 2.6 or 3.17 (depending on the considered boundary conditions), the sequence of space–time GDs $((\mathcal{D}_T)_m)_{m\in\mathbb{N}}$ is coercive and thus

$$\begin{aligned}\|(u_m)_\theta\|_{L^q(0,T;X_{\mathcal{D}_m,\bullet})} &= \left\|\nabla^{(\theta)}_{\mathcal{D}_m} u_m\right\|_{L^q(0,T;L^p(\Omega))^d}\\ &\geq \frac{1}{C_P}\left\|\Pi^{(\theta)}_{\mathcal{D}_m} u_m\right\|_{L^q(0,T;L^p(\Omega))}.\end{aligned}$$

The sequences $(\Pi^{(\theta)}_{\mathcal{D}_m} u_m)_{m\in\mathbb{N}}$ and $(\nabla^{(\theta)}_{\mathcal{D}_m} u_m)_{m\in\mathbb{N}}$ are therefore bounded in $L^q(0,T;L^p(\Omega))$ and $L^q(0,T;L^p(\Omega))^d$, respectively. Up to a subsequence, they converge weakly (or weakly-$*$ if $q = +\infty$) in these spaces towards u and $\boldsymbol{v}$, respectively. Extending all the functions by 0 outside Ω, these convergences still hold weakly in $L^q(0,T;L^p(\mathbb{R}^d))$ and $L^q(0,T;L^p(\mathbb{R}^d))^d$. The proof is complete if we show that $\boldsymbol{v} = \nabla u$ in the sense of distributions on $\mathbb{R}^d \times (0,T)$. To this purpose, let $\boldsymbol{\varphi} \in C^\infty_c(\mathbb{R}^d)^d$ and $\psi \in C^\infty_c(0,T)$. Dropping the indices m for the sake of legibility, we have, for $t \in (0,T)$, by definition (2.6) of $W_{\mathcal{D}}$,

$$\left|\int_\Omega \left[\nabla_{\mathcal{D}}[u_\theta(t)](\boldsymbol{x})\cdot\boldsymbol{\varphi}(\boldsymbol{x}) + \Pi_{\mathcal{D}}[u_\theta(t)](\boldsymbol{x})\mathrm{div}\boldsymbol{\varphi}(\boldsymbol{x})\right]\mathrm{d}\boldsymbol{x}\right| \leq \|u_\theta(t)\|_{\mathcal{D}}\, W_{\mathcal{D}}(\boldsymbol{\varphi}).$$

Hence, using Hölder's inequality:

$$\begin{aligned}&\left|\int_0^T\int_\Omega \left[\nabla^{(\theta)}_{\mathcal{D}} u(\boldsymbol{x},t)\cdot(\psi(t)\boldsymbol{\varphi}(\boldsymbol{x})) + \Pi^{(\theta)}_{\mathcal{D}} u(\boldsymbol{x},t)\mathrm{div}(\psi(t)\boldsymbol{\varphi})(\boldsymbol{x})\right]\mathrm{d}\boldsymbol{x}\mathrm{d}t\right|\\ &\quad\leq \int_0^T\left|\psi(t)\int_\Omega\left[\nabla^{(\theta)}_{\mathcal{D}} u(\boldsymbol{x},t)\cdot\boldsymbol{\varphi}(\boldsymbol{x}) + \Pi^{(\theta)}_{\mathcal{D}} u(\boldsymbol{x},t)\mathrm{div}\boldsymbol{\varphi}(\boldsymbol{x})\right]\mathrm{d}\boldsymbol{x}\right|\mathrm{d}t\end{aligned}$$

$$\leq W_{\mathcal{D}}(\boldsymbol{\varphi}) \int_0^T \|u_\theta(t)\|_{\mathcal{D}} \, |\psi(t)| \mathrm{d}t \leq W_{\mathcal{D}}(\boldsymbol{\varphi}) \, \|u_\theta\|_{L^q(0,T;X_{\mathcal{D},0})} \, \|\psi\|_{L^{q'}(0,T)} \, .$$

As $m \to \infty$, $\|(u_m)_\theta\|_{L^q(0,T;X_{\mathcal{D}_m,0})}$ is bounded and $W_{\mathcal{D}_m}(\varphi) \to 0$ by limit-conformity. We can thus pass to the limit and see that

$$\int_0^T \int_\Omega \Big[\boldsymbol{v}(\boldsymbol{x},t) \cdot (\psi(t)\boldsymbol{\varphi}(\boldsymbol{x})) + u(\boldsymbol{x},t) \mathrm{div}(\psi(t)\boldsymbol{\varphi})(\boldsymbol{x}) \Big] \mathrm{d}\boldsymbol{x}\mathrm{d}t = 0.$$

This relation holds true for linear combinations of functions of the form $(\boldsymbol{x},t) \to \psi(t)\boldsymbol{\varphi}(\boldsymbol{x})$, that is for all tensorial smooth functions. These tensorial functions are dense (*e.g.*, for the $C^1(\overline{\Omega} \times [0,T])^d$ norm) in $C_c^\infty(\mathbb{R}^d \times (0,T))^d$, see [63] or [67, Appendix D]. This shows that, for all $\boldsymbol{\Phi} \in C_c^\infty(\mathbb{R}^d \times (0,T))^d$,

$$\int_0^T \int_\Omega \Big[\boldsymbol{v}(\boldsymbol{x},t) \cdot \boldsymbol{\Phi}(\boldsymbol{x},t) + u(\boldsymbol{x},t) \mathrm{div}(\boldsymbol{\Phi}(\cdot,t))(\boldsymbol{x}) \Big] \mathrm{d}\boldsymbol{x}\mathrm{d}t = 0.$$

Hence, $\boldsymbol{v} = \nabla u$ in the sense of distributions on $\mathbb{R}^d \times (0,T)$, as required.

■

The following result shows that functions depending on time and space can be approximated, along with their gradient, by reconstructed functions and gradients built from space–time-consistent GDs.

Lemma 4.10 (Interpolation of space–time functions). *For $p \in [1,\infty)$, $T > 0$ and $\theta \in [0,1]$, let $((\mathcal{D}_T)_m)_{m\in\mathbb{N}}$ be a sequence of space–time GDs in the sense of Definition 4.1, that is space–time-consistent in the sense of Definition 4.3. Let $\overline{v} \in L^p(0,T;W^{1,p}_\bullet(\Omega))$. Then:*

1. *There exists a sequence $(v_m)_{m\in\mathbb{N}}$ such that $v_m = (v_m^{(n)})_{n=0,\ldots,N_m} \in X^{N_m+1}_{\mathcal{D}_m,\bullet}$ for all $m \in \mathbb{N}$, and, as $m \to \infty$,*

$$\Pi^{(\theta)}_{\mathcal{D}_m} v_m \to \overline{v} \text{ strongly in } L^p(\Omega \times (0,T)), \tag{4.7a}$$

$$\nabla^{(\theta)}_{\mathcal{D}_m} v_m \to \nabla\overline{v} \text{ strongly in } L^p(\Omega \times (0,T))^d. \tag{4.7b}$$

2. *In the case of non-homogeneous Neumann boundary conditions, if the sequence of underlying spatial discretisations is limit–conforming in the sense of Definition 3.14, then the sequence $(v_m)_{m\in\mathbb{N}}$ in Item 1 also satisfies*

$$\mathbb{T}^{(\theta)}_{\mathcal{D}_m} v_m \to \gamma\overline{v} \text{ weakly in } L^p(\partial\Omega \times (0,T)) \text{ as } m \to \infty. \tag{4.8}$$

3. *In the case of non-homogeneous Fourier boundary conditions, the sequence $(v_m)_{m\in\mathbb{N}}$ in Item 1 can be chosen such that*

$$\mathbb{T}^{(\theta)}_{\mathcal{D}_m} v_m \to \gamma\overline{v} \text{ strongly in } L^p(\partial\Omega \times (0,T)) \text{ as } m \to \infty. \tag{4.9}$$

4. *If moreover* $\overline{v} \in C([0,T];L^2(\Omega))$, $\partial_t \overline{v} \in L^2(\Omega \times (0,T))$ *and* $\overline{v}(\cdot,T) = 0$, *then the sequence* $(v_m)_{m\in\mathbb{N}}$ *in Item 1 can be chosen such that, in addition to* (4.7),

$$\forall m \in \mathbb{N}\,,\ v_m^{(N_m-1)} = v_m^{(N_m)} = 0$$
$$\text{(and thus } \Pi_{\mathcal{D}_m}^{(\theta)} v^m = 0 \text{ on } \Omega \times (t^{(N_m-1)}, t^{(N_m)}]), \tag{4.10a}$$
$$\Pi_{\mathcal{D}_m}^{(\theta)} v_m(\cdot,0) \to \overline{v}(\cdot,0) \text{ strongly in } L^2(\Omega) \text{ as } m \to \infty, \tag{4.10b}$$
$$\delta_{\mathcal{D}_m} v_m \to \partial_t \overline{v} \text{ strongly in } L^2(\Omega \times (0,T)) \text{ as } m \to \infty. \tag{4.10c}$$

Proof.
Step 1: proof of Item 1.
Define the set $\mathcal{T}(0,T;W_\bullet^{1,p}(\Omega))$ of space–time tensorial functions the following way: $v \in \mathcal{T}(0,T;W_\bullet^{1,p}(\Omega))$ if there exist $\ell \in \mathbb{N}$, a family $(\varphi_i)_{i=1,\dots,\ell} \subset C^\infty([0,T])$ and a family $(w_i)_{i=1,\dots,\ell} \subset W_\bullet^{1,p}(\Omega)$ such that

$$v(\boldsymbol{x},t) = \sum_{i=1}^{\ell} \varphi_i(t) w_i(\boldsymbol{x}) \text{ for a.e. } \boldsymbol{x} \in \Omega \text{ and all } t \in (0,T). \tag{4.11}$$

By [67, Corollary 1.3.1], $\mathcal{T}(0,T;W_\bullet^{1,p}(\Omega))$ is dense in $L^p(0,T;W_\bullet^{1,p}(\Omega))$ and we can therefore reduce the proof of (4.7) to the case $\overline{v} \in \mathcal{T}(0,T;W_\bullet^{1,p}(\Omega))$ (the proof of this reduction is similar to the proof of Lemma 2.16).
Given the structure (4.11) of functions in $\mathcal{T}(0,T;W_\bullet^{1,p}(\Omega))$, we actually only need to prove the result for $\overline{v}(\boldsymbol{x},t) = \varphi(t)w(\boldsymbol{x})$ with $\varphi \in C^\infty([0,T])$ and $w \in W_\bullet^{1,p}(\Omega)$. Let $v_m \in X_{\mathcal{D}_m,\bullet}^{N_m+1}$ be defined by $v_m^{(n)} = \varphi(t^{(n)}) I_{\mathcal{D}_m} w$ for $n = 0,\dots,N_m$, where

$$I_{\mathcal{D}_m} w = \operatorname*{argmin}_{z \in X_{\mathcal{D}_m,\bullet}} \left(\|\Pi_{\mathcal{D}_m} z - w\|_{L^{\max(p,2)}(\Omega)} + \|\nabla_{\mathcal{D}_m} z - \nabla w\|_{L^p(\Omega)^d} \right). \tag{4.12}$$

Define $\Phi_m : (0,T] \to \mathbb{R}$ as the piecewise constant function equal to $\theta\varphi(t^{(n+1)}) + (1-\theta)\varphi(t^{(n)})$ on $(t^{(n)}, t^{(n+1)}]$ for all $n = 0,\dots,N_m-1$. Then, by definition (4.2) of the space–time reconstruction operator $\Pi_{\mathcal{D}_m}^{(\theta)}$, for all $t \in (0,T)$ and a.e. $\boldsymbol{x} \in \Omega$,

$$\begin{aligned}\overline{v}(\boldsymbol{x},t) - \Pi_{\mathcal{D}_m}^{(\theta)} v_m(\boldsymbol{x},t) &= \varphi(t)w(\boldsymbol{x}) - \Phi_m(t)\Pi_{\mathcal{D}_m}(I_{\mathcal{D}_m} w)(\boldsymbol{x}) \\ &= [\varphi(t) - \Phi_m(t)]w(\boldsymbol{x}) + \Phi_m(t)[w(\boldsymbol{x}) - \Pi_{\mathcal{D}_m}(I_{\mathcal{D}_m} w)(\boldsymbol{x})].\end{aligned} \tag{4.13}$$

Using the definitions of $\widehat{S}_{\mathcal{D}_m}$ and $I_{\mathcal{D}_m}$, we infer that

$$\begin{aligned}&\|\overline{v} - \Pi_{\mathcal{D}_m}^{(\theta)} v_m\|_{L^p(\Omega\times(0,T))} \\ &\le \|\varphi - \Phi_m\|_{L^p(0,T)} \|w\|_{L^p(\Omega)} + \|\Phi_m\|_{L^p(0,T)} \|w - \Pi_{\mathcal{D}_m}(I_{\mathcal{D}_m} w)\|_{L^p(\Omega)} \\ &\le \|\varphi - \Phi_m\|_{L^p(0,T)} \|w\|_{L^p(\Omega)} + \|\Phi_m\|_{L^p(0,T)} \widehat{S}_{\mathcal{D}_m}(w).\end{aligned} \tag{4.14}$$

As $m \to \infty$, the space–time-consistency of $((\mathcal{D}_T)_m)_{m\in\mathbb{N}}$ gives $\widehat{S}_{\mathcal{D}_m}(w) \to 0$ and the smoothness of φ shows that $\Phi_m \to \varphi$ uniformly (and thus in $L^p(0,T)$). Hence, (4.7a) follows from (4.14). The proof of (4.7b) is obtained by the same argument starting from (4.13) and replacing $\overline{v}$ with $\nabla\overline{v}$, w with ∇w, and $\Pi^{(\theta)}_{\mathcal{D}_m}$ with $\nabla^{(\theta)}_{\mathcal{D}_m}$.

Step 2: proof of Items 2 and 3.
In the case of Neumann boundary conditions, applying Lemma 4.8 to $(v_m)_{m\in\mathbb{N}}$ yields (4.8).
In the case of Fourier boundary conditions, the definition (4.12) can be replaced with

$$I_{\mathcal{D}_m} w = \operatorname{argmin}_{z\in X_{\mathcal{D}_m}} \Big(\|\Pi_{\mathcal{D}_m} z - w\|_{L^{\max(p,2)}(\Omega)} + \|\nabla_{\mathcal{D}_m} z - \nabla w\|_{L^p(\Omega)^d} + \|\mathbb{T}_{\mathcal{D}_m} z - \gamma w\|_{L^p(\partial\Omega)} \Big)$$

and the reasoning starting from (4.13) can be done with $(\gamma\overline{v}, \mathbb{T}^{(\theta)}_{\mathcal{D}_m} v_m)$ instead of $(\overline{v}, \Pi^{(\theta)}_{\mathcal{D}_m} v_m)$, since $\|\mathbb{T}_{\mathcal{D}_m}(I_{\mathcal{D}_m} w) - \gamma w\|_{L^p(\partial\Omega)} \le \widehat{S}_{\mathcal{D}_m}(w)$. This shows that (4.9) holds.

Step 3: proof of Item 4.
Assume that $\overline{v} \in L^p(0,T;W^{1,p}_\bullet(\Omega)) \cap C([0,T];L^2(\Omega))$, $\partial_t\overline{v} \in L^2(\Omega\times(0,T))$ and $\overline{v}(\cdot,T) = 0$. By [67, Theorem 2.3.1], we can find a sequence $(\overline{v}_n)_{n\in\mathbb{N}} \subset C^\infty([0,T];W^{1,p}_\bullet(\Omega)\cap L^2(\Omega))$ such that, as $n\to\infty$,

$$\begin{aligned}
&\overline{v}_n \to \overline{v} \text{ in } L^p(0,T;W^{1,p}_\bullet(\Omega)) \cap C([0,T];L^2(\Omega)), \text{ and}\\
&\partial_t\overline{v}_n \to \partial_t\overline{v} \text{ in } L^2(0,T;L^2(\Omega)).
\end{aligned}$$

The proof of [67, Theorem 2.3.1] is based on an even extension of $\overline{v}$ at $t=T$, required to preserve the continuity of the extended function. Since $\overline{v}(\cdot,T)=0$, we can actually use an extension by 0 on $[T,\infty)$ and, by selecting in [67, Theorem 2.3.1] a smoothing kernel with support in $(-T,0)$, we ensure that each $\overline{v}_n$ vanishes on $[T-\epsilon_n,\infty)$ for some $\epsilon_n > 0$.
Having approximated $\overline{v}$ by these $\overline{v}_n$, we just need to prove the result for each $\overline{v}_n$ instead of $\overline{v}$. Let us drop the index n and write $\overline{v}$ for $\overline{v}_n$. We have $\overline{v} \in C^\infty([0,T];W^{1,p}_\bullet(\Omega)\cap L^2(\Omega))$ and $\overline{v}=0$ on $\Omega\times[T-\epsilon,T]$ for some $\epsilon>0$. Let $\tau\in(0,\epsilon/4)$, $\ell_\tau = \lceil T/\tau\rceil$ and take $(\psi_i)_{i=1,\dots,\ell_\tau} \subset C^\infty([0,T])$ a partition of unity on $[0,T]$ subordinate to the open covering $((i\tau-2\tau, i\tau+2\tau)_{i=1,\dots,\ell_\tau})$. Set

$$\begin{aligned}
\overline{v}_\tau(\boldsymbol{x},t) &= \overline{v}(\boldsymbol{x},T) + \sum_{i=1}^{\ell_\tau}\left(\int_T^t \psi_i(s)\mathrm{d}s\right)\partial_t\overline{v}(\boldsymbol{x},i\tau)\\
&= \sum_{i=1}^{\ell_\tau}\left(\int_T^t \psi_i(s)\mathrm{d}s\right)\partial_t\overline{v}(\boldsymbol{x},i\tau).
\end{aligned}$$

Since $\overline{v} = 0$ on $\Omega \times [T-\epsilon, T] \supset \Omega \times [T-4\tau, T]$, the terms corresponding to $i = \ell_\tau - 3, \ldots, \ell_\tau$ in the previous sum vanish (since $i\tau \geq \ell_\tau \tau - 3\tau \geq T - 4\tau$). For $i \leq \ell_\tau - 4$, the support of ψ_i is contained in $[0, T-2\tau]$. This shows that $\overline{v}_\tau(\cdot, t) = 0$ for all $t \in [T-2\tau, T]$.
We write

$$\partial_t \overline{v}_\tau(\boldsymbol{x}, t) = \sum_{i=1}^{\ell_\tau} \psi_i(t) \partial_t \overline{v}(\boldsymbol{x}, i\tau).$$

Since

$$\sum_{i=1}^{\ell_\tau} \psi_i(t) = 1 \text{ for all } t \in (0, T), \tag{4.15}$$

we have $\partial_t \overline{v}(\boldsymbol{x}, t) = \sum_{i=1}^{\ell_\tau} \psi_i(t) \partial_t \overline{v}(\boldsymbol{x}, t)$ and thus

$$\begin{aligned}
&\|\partial_t \overline{v}_\tau(\cdot, t) - \partial_t \overline{v}(\cdot, t)\|_{W^{1,p}_\bullet(\Omega) \cap L^2(\Omega)} \\
&\qquad \leq \sum_{i=1}^{\ell_\tau} \psi_i(t) \, \|\partial_t \overline{v}(\cdot, i\tau) - \partial_t \overline{v}(\cdot, t)\|_{W^{1,p}_\bullet(\Omega) \cap L^2(\Omega)} \, .
\end{aligned}$$

Using the fact that $\psi_i(t) \neq 0$ only if $|t - i\tau| < 2\tau$ (that is, $i\tau \in (t-2\tau, t+2\tau)$), and invoking (4.15), we infer

$$\begin{aligned}
&\|\partial_t \overline{v}_\tau(\cdot, t) - \partial_t \overline{v}(\cdot, t)\|_{W^{1,p}_\bullet(\Omega) \cap L^2(\Omega)} \\
&\qquad \leq \sup_{r \in (t-2\tau, t+2\tau)} \|\partial_t \overline{v}(\cdot, r) - \partial_t \overline{v}(\cdot, t)\|_{W^{1,p}_\bullet(\Omega) \cap L^2(\Omega)} \, .
\end{aligned}$$

By smoothness of $\overline{v}$, this shows that $\partial_t \overline{v}_\tau \to \partial_t \overline{v}$ in $L^\infty(0, T; W^{1,p}_\bullet(\Omega) \cap L^2(\Omega))$ as $\tau \to 0$. Integrating and using $\overline{v}_\tau(\cdot, T) = \overline{v}(\cdot, T) = 0$, we obtain $\overline{v}_\tau \to \overline{v}$ in $C([0, T]; W^{1,p}_\bullet(\Omega) \cap L^2(\Omega))$. Hence, we only need to find approximations in the sense (4.7) and (4.10) for each $\overline{v}_\tau$ instead of $\overline{v}$. Given the structure of $\overline{v}_\tau$, this amounts to finding such approximations when $\overline{v}(\boldsymbol{x}, t) = \varphi(t) w(\boldsymbol{x})$ with $w \in W^{1,p}_\bullet(\Omega) \cap L^2(\Omega)$ and $\varphi \in C^\infty([0, T])$ having support in $[0, T-\nu]$ for some $\nu > 0$.
We set, as before, $v_m^{(n)} = \varphi(t^{(n)}) I_{\mathcal{D}_m} w$ for $n = 0, \ldots, N_m$. The proof of (4.7) is done exactly as in Step 1.
If m is large enough so that $t^{(N_m - 1)} \geq T - \nu$, then $v_m^{(N_m - 1)} = v_m^{(N_m)} = 0$ and (4.10a) is satisfied. We can modify v_m for the remaining m (for example by setting $v_m = 0$) to ensure that this property holds for any m.
By definition (4.3) of $\Pi^{(\theta)}_{\mathcal{D}_m} v_m$ at $t = 0$,

$$\begin{aligned}
u(\boldsymbol{x}, 0) - \Pi^{(\theta)}_{\mathcal{D}_m} v_m(\boldsymbol{x}, 0) &= \varphi(0) w(\boldsymbol{x}) - \varphi(t^{(0)}) \Pi_{\mathcal{D}_m}(I_{\mathcal{D}_m} w)(\boldsymbol{x}) \\
&= \varphi(0)[w(\boldsymbol{x}) - \Pi_{\mathcal{D}_m}(I_{\mathcal{D}_m} w)(\boldsymbol{x})].
\end{aligned}$$

Then, by definition of $I_{\mathcal{D}_m}$ and $\widehat{S}_{\mathcal{D}_m}$,

$$\left\|\overline{v}(0) - \Pi^{(\theta)}_{\mathcal{D}_m} v_m(0)\right\|_{L^2(\Omega)} = |\varphi(0)| \left\|w - \Pi_{\mathcal{D}_m}(I_{\mathcal{D}_m} w)\right\|_{L^2(\Omega)} \leq |\varphi(0)| \widehat{S}_{\mathcal{D}_m}(w)$$

and (4.10b) follows from the space–time-consistency of $((\mathcal{D}_T)_m)_{m\in\mathbb{N}}$.
To establish (4.10c), we define $\Psi_m : (0,T] \to \mathbb{R}$ as the piecewise constant function equal to $\frac{\varphi(t^{(n+1)})-\varphi(t^{(n)})}{\delta t^{(n+\frac{1}{2})}}$ on $(t^{(n)}, t^{(n+1)}]$, for all $n = 0, \ldots, N_m - 1$. Then, by definition (4.4) of $\delta_{\mathcal{D}_m} v_m$, for all $t \in (0,T)$ and a.e. $\boldsymbol{x} \in \Omega$,

$$\partial_t \overline{v}(\boldsymbol{x}, t) - \delta_{\mathcal{D}_m} v_m(\boldsymbol{x}, t) = \varphi'(t) w(\boldsymbol{x}) - \Psi_m(t) \Pi_{\mathcal{D}_m}(I_{\mathcal{D}_m} w)(\boldsymbol{x}).$$

By smoothness of φ we have $\Psi_m \to \varphi'$ uniformly on $[0,T]$ as $m \to \infty$, and the proof of (4.10c) therefore follows by using the same sequence of estimates as in Step 1, starting from (4.13) and replacing $(\overline{v}, \Pi_{\mathcal{D}_m} v_m, \varphi, \Phi_m)$ with $(\partial_t \overline{v}, \delta_{\mathcal{D}_m} v_m, \varphi', \Psi_m)$, and the L^p norms with L^2 norms (note that $\widehat{S}_{\mathcal{D}_m}(w) \geq \|w - \Pi_{\mathcal{D}_m}(I_{\mathcal{D}_m} w)\|_{L^2(\Omega)}$ by definition of $\widehat{S}_{\mathcal{D}}$). ∎

4.2 Compactness results for space-time gradient discretisations

The compactness results established here are consequences of the generic time-DFA results in Appendix C.

4.2.1 Averaged-in-time compactness for space-time GDs

Aubin–Simon theorem

In the context of GDs, the spaces Y_m in Definition C.6 (compactly–continuously embedded sequence) and Theorem C.8 (Aubin–Simon with sequences of spaces and discrete derivative) are often $\Pi_{\mathcal{D}_m}(X_{\mathcal{D}_m,\bullet})$ with the dual norm $\|w\|_{\star,\mathcal{D}_m}$ as defined below.

Definition 4.11 (Dual norm on $\Pi_{\mathcal{D}}(X_{\mathcal{D},\bullet})$). *Let $\mathcal{D}_T$ be a space–time GD in the sense of Definition 4.1. The dual norm $\|\cdot\|_{\star,\mathcal{D}}$ on $\Pi_{\mathcal{D}}(X_{\mathcal{D},\bullet}) \subset L^2(\Omega)$ is defined by:*

$$\begin{aligned} &\forall w \in \Pi_{\mathcal{D}}(X_{\mathcal{D},\bullet}), \\ &\|w\|_{\star,\mathcal{D}} = \sup\left\{\int_\Omega w(\boldsymbol{x}) \Pi_{\mathcal{D}} v(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \,:\, v \in X_{\mathcal{D},\bullet}\,,\ \|v\|_{\mathcal{D}} = 1\right\}. \end{aligned} \tag{4.16}$$

A straightforward consequence of this definition is

$$\forall w \in \Pi_{\mathcal{D}}(X_{\mathcal{D},\bullet})\,,\ \forall v \in X_{\mathcal{D},\bullet}\,,\ \left|\int_\Omega w(\boldsymbol{x}) \Pi_{\mathcal{D}} v(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}\right| \leq \|w\|_{\star,\mathcal{D}} \|v\|_{\mathcal{D}}\,. \tag{4.17}$$

This relation shows that $\|\cdot\|_{\star,\mathcal{D}}$ is a norm (not just a semi-norm). Indeed, if $\|w\|_{\star,\mathcal{D}} = 0$ then $\int_\Omega w(\boldsymbol{x})\Pi_\mathcal{D} v(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = 0$ for all $v \in X_{\mathcal{D},\bullet}$. Taking then v such that $\Pi_\mathcal{D} v = w$ shows that $w = 0$.
The norm $\|\cdot\|_{\star,\mathcal{D}}$ will mostly be used on $\delta_\mathcal{D} v(t)$ for $v \in X_{\mathcal{D},\bullet}^{N+1}$. Recalling the notation (4.4), it is clear that $\delta_\mathcal{D} v(t) \in \Pi_\mathcal{D}(X_{\mathcal{D},\bullet})$ for a.e. $t \in (0,T)$, and thus $\|\delta_\mathcal{D} v(t)\|_{\star,\mathcal{D}}$ is well-defined.

Remark 4.12 (Boundary conditions)
It is also worth noticing that $\|w\|_{\star,\mathcal{D}}$ takes into account the considered boundary conditions, through the norm $\|v\|_\mathcal{D}$ on $X_{\mathcal{D},\bullet}$ (see, *e.g.*, Definitions 2.1 and 3.1).

Remark 4.13 ($\|\cdot\|_{\star,\mathcal{D}}$ is a discrete H^{-1} norm)
Let us consider the case of homogeneous Dirichlet boundary conditions and $p = 2$. Then Definition 2.1 shows that $(X_{\mathcal{D},0}, \|\cdot\|_\mathcal{D})$ is a discrete version of $(H_0^1(\Omega), \|\cdot\|_{H_0^1(\Omega)})$, where $\|\cdot\|_{H_0^1(\Omega)} = \|\nabla\cdot\|_{L^2(\Omega)^d}$ is the standard norm on $H_0^1(\Omega)$.
In the continuous setting, (4.16) therefore reads: for $w \in L^2(\Omega)$,

$$\|w\|_\star := \sup\left\{\int_\Omega w(\boldsymbol{x})v(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \,:\, v \in H_0^1(\Omega)\,,\ \|v\|_{H_0^1(\Omega)} = 1\right\}. \tag{4.18}$$

Identifying w as an element of $H^{-1}(\Omega)$, we have

$$\int_\Omega w(\boldsymbol{x})v(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = \langle w, v\rangle_{H^{-1},H_0^1}.$$

Hence, (4.18) turns out to be the standard dual norm on $H^{-1}(\Omega)$, that is the norm of linear continuous functions $H_0^1(\Omega) \to \mathbb{R}$.
The norm $\|\cdot\|_{\star,\mathcal{D}}$ can thus be considered as a discrete version of the standard dual norm on $H^{-1}(\Omega)$.

The next result is a consequence of the discrete Aubin–Simon theorem (Theorem C.8).

Theorem 4.14 (Aubin–Simon theorem for GDs). *Let $T > 0$, $p \in (1,+\infty)$ and $\theta \in [0,1]$. Assume that $((\mathcal{D}_T)_m)_{m\in\mathbb{N}}$ is a compact sequence of space–time GDs in the sense of Definition 4.6. For any $m \in \mathbb{N}$, let $v_m \in X_{\mathcal{D}_m,\bullet}^{N_m+1}$ be such that there exists $C > 0$ satisfying*

$$\forall m \in \mathbb{N},\ \int_0^T \|(v_m)_\theta(t)\|_{\mathcal{D}_m}^p\,\mathrm{d}t \le C \tag{4.19}$$

and

$$\forall m \in \mathbb{N},\ \int_0^T \|\delta_{\mathcal{D}_m} v_m(t)\|_{\star,\mathcal{D}_m}\,\mathrm{d}t \le C. \tag{4.20}$$

Then the sequence $(\Pi_{\mathcal{D}_m}^{(\theta)} v_m)_{m\in\mathbb{N}}$ is relatively compact in $L^p(\Omega\times(0,T))$.

Proof. To apply Theorem C.8, let $B = L^p(\Omega)$, $X_m = \Pi_{\mathcal{D}_m}(X_{\mathcal{D}_m,\bullet})$, and define the norm on X_m by

$$\|u\|_{X_m} = \min\{\|w\|_{\mathcal{D}_m} \,:\, w \in X_{\mathcal{D}_m,\bullet} \text{ such that } \Pi_{\mathcal{D}_m} w = u\}. \tag{4.21}$$

Set $Y_m = X_m = \Pi_{\mathcal{D}_m}(X_{\mathcal{D}_m,\bullet})$ and $\|\cdot\|_{Y_m} = \|\cdot\|_{\star,\mathcal{D}_m}$.

Let us prove that the sequence $(X_m, Y_m)_{m\in\mathbb{N}}$ is compactly–continuously embedded in B, in the sense of Definition C.6. First, the compactness hypothesis on $(\mathcal{D}_m)_{m\in\mathbb{N}}$ is showing that $(X_m)_{m\in\mathbb{N}}$ is compactly embedded in B, in the sense of Definition C.4 (see, e.g., Remark 2.9 in the case of Dirichlet BCs). Then, by construction, $X_m = Y_m$ for all $m \in \mathbb{N}$. Assume now that $(u_m)_{m\in\mathbb{N}}$ is such that $u_m \in X_m$ for all $m \in \mathbb{N}$, $(\|u_m\|_{X_m})_{m\in\mathbb{N}}$ bounded, $\|u_m\|_{Y_m} \to 0$ as $m \to +\infty$, and $(u_m)_{m\in\mathbb{N}}$ converges in $L^p(\Omega)$. Take $R_m u_m \in X_{\mathcal{D}_m,\bullet}$ a lifting of u_m with minimal norm, *i.e.* $\Pi_{\mathcal{D}_m} R_m u_m = u_m$ and $\|R_m u_m\|_{\mathcal{D}_m} = \|u_m\|_{X_m}$ A use of (4.17) yields

$$\begin{aligned}\int_\Omega u_m(\boldsymbol{x})^2 \mathrm{d}\boldsymbol{x} &= \int_\Omega u_m(\boldsymbol{x}) \Pi_{\mathcal{D}_m} R_m u_m(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \\ &\le \|u_m\|_{\star,\mathcal{D}_m} \|R_m u_m\|_{\mathcal{D}_m} = \|u_m\|_{Y_m} \|u_m\|_{X_m}.\end{aligned}$$

The assumptions on $(u_m)_{m\in\mathbb{N}}$ thus ensure that $\lim_{m\to\infty} \int_\Omega u_m(\boldsymbol{x})^2 \mathrm{d}\boldsymbol{x} = 0$. This shows that, up to a subsequence, $u_m \to 0$ a.e. in Ω, and hence that the limit in $L^p(\Omega)$ of $(u_m)_{m\in\mathbb{N}}$ must be 0. The proof that $(X_m, Y_m)_{m\in\mathbb{N}}$ is compactly–continuously embedded in $B = L^p(\Omega)$ is complete.

The relative compactness of $(\Pi^{(\theta)}_{\mathcal{D}_m} v_m)_{m\in\mathbb{N}}$ in $L^p(0,T;L^p(\Omega))$ follows from Theorem C.8 with $f_m = \Pi^{(\theta)}_{\mathcal{D}_m} v_m$ if we can check the four assumptions stated in this theorem.
The first of these assumptions is obviously satisfied by the definition of $\Pi^{(\theta)}_{\mathcal{D}_m}$ in (4.2).
Since the sequence of underlying spatial discretisations is compact, it is also coercive (see, *e.g.*, Lemma 2.10 for homogeneous Dirichlet boundary conditions). The definition of $C_{\mathcal{D}_m}$ combined with (4.19) and the definition of $\Pi^{(\theta)}_{\mathcal{D}_m}$ then shows that $(\Pi^{(\theta)}_{\mathcal{D}_m} v_m)_{m\in\mathbb{N}}$ is bounded in $L^p(0,T;L^p(\Omega))$. This takes care of the second assumption in Theorem C.8.
The third assumption follows immediately from (4.19) and the fact that

$$\left\|\Pi^{(\theta)}_{\mathcal{D}_m} v_m(t)\right\|_{X_m} = \|\Pi_{\mathcal{D}_m}((v_m)_\theta(t))\|_{X_m} \le \|(v_m)_\theta(t)\|_{\mathcal{D}_m}.$$

To prove the fourth assumption in Theorem C.8, we notice that

$$\left\|\delta_m \Pi^{(\theta)}_{\mathcal{D}_m} v_m(t)\right\|_{Y_m} = \|\delta_{\mathcal{D}_m} v_m(t)\|_{Y_m} = \|\delta_{\mathcal{D}_m} v_m(t)\|_{\star,\mathcal{D}_m}$$

and we use (4.20). ∎

Convergence of a weak–strong product, and identification of non-linear weak limits

Dealing with degenerate parabolic equations often requires fine results to identify non-linear limits of weakly converging sequences. The main result in this section, Theorem 4.17, is one of these fine results. We consider here the particular case $p = 2$ and we restrict ourselves to homogeneous Dirichlet boundary conditions. The adaptation to other boundary conditions is rather simple, but establishing the equivalent of Theorem 4.17 for $p \neq 2$ requires a different approach; see [72, Theorem 5.4] for details.

The inverses of the discrete and continuous Laplace operators with homogeneous Dirichlet boundary conditions now need to be introduced.

Definition 4.15 (Inverse of discrete and continuous Laplace operator). *Let $\mathcal{D} = (X_{\mathcal{D},0}, \Pi_{\mathcal{D}}, \nabla_{\mathcal{D}})$ be a GD in the sense of Definition 2.1, for $p = 2$. We define the operator $\Delta^i_{\mathcal{D}} : X_{\mathcal{D},0} \to X_{\mathcal{D},0}$ such that, for all $v \in X_{\mathcal{D},0}$,*

$$\forall w \in X_{\mathcal{D},0}\,,\ \int_\Omega \nabla_{\mathcal{D}}(\Delta^i_{\mathcal{D}} v)(\boldsymbol{x}) \cdot \nabla_{\mathcal{D}} w(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} = \int_\Omega \Pi_{\mathcal{D}} v(\boldsymbol{x}) \Pi_{\mathcal{D}} w(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}. \quad (4.22)$$

We also define $\Delta^i : L^2(\Omega) \to H^1_0(\Omega)$ such that, for all $v \in L^2(\Omega)$,

$$\forall w \in H^1_0(\Omega)\,,\ \int_\Omega \nabla(\Delta^i v)(\boldsymbol{x}) \cdot \nabla w(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} = \int_\Omega v(\boldsymbol{x}) w(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}. \quad (4.23)$$

Theorem 4.16 (Compactness of Δ^i). *Let $p = 2$, $T > 0$, $\theta \in [0,1]$ and $((\mathcal{D}_T)_m)_{m\in\mathbb{N}} = (\mathcal{D}_m, \mathcal{I}_{\mathcal{D}_m}, (t^{(n)}_m)_{n=0,\dots,N_m})_{m\in\mathbb{N}}$ be a space–time-consistent, limit-conforming and compact sequence of space–time GDs for homogeneous Dirichlet boundary conditions, in the sense of Definitions 4.3 and 4.6. For any $m \in \mathbb{N}$, let $v_m \in X^{N_m+1}_{\mathcal{D}_m,0}$ be such that there exists $C > 0$ and $q \geq 1$ satisfying*

$$\forall m \in \mathbb{N},\ \int_0^T \left\| \Pi^{(\theta)}_{\mathcal{D}_m} v_m(t) \right\|^2_{L^2(\Omega)} \mathrm{d}t \leq C \quad (4.24)$$

and

$$\forall m \in \mathbb{N},\ \int_0^T \|\delta_{\mathcal{D}_m} v_m(t)\|^q_{\star,\mathcal{D}_m} \mathrm{d}t \leq C. \quad (4.25)$$

We also assume that $\Pi^{(\theta)}_{\mathcal{D}_m} v_m$ converges weakly in $L^2(0,T;L^2(\Omega))$ as $m \to \infty$ to some $\overline{v} \in L^2(0,T;L^2(\Omega))$.
Then, as $m \to \infty$,

$$\begin{aligned} &\Pi^{(\theta)}_{\mathcal{D}_m}(\Delta^i_{\mathcal{D}_m} v_m) \to \Delta^i v \text{ in } L^2(0,T;L^2(\Omega)), \text{ and} \\ &\nabla^{(\theta)}_{\mathcal{D}_m}(\Delta^i_{\mathcal{D}_m} v_m) \to \nabla(\Delta^i \overline{v}) \text{ in } L^2(0,T;L^2(\Omega))^d. \end{aligned} \quad (4.26)$$

Moreover, if $q > 1$ then

$$\delta_{\mathcal{D}_m}(\Delta^i_{\mathcal{D}_m} v_m) \to \partial_t(\Delta^i \overline{v}) \text{ weakly in } L^q(0,T;L^2(\Omega)) \text{ as } m \to \infty. \quad (4.27)$$

Proof.
Step 1: we prove, using Proposition C.5 with $p = 2$, that $(\Pi^{(\theta)}_{\mathcal{D}_m}(\Delta^i_{\mathcal{D}_m} v_m))_{m\in\mathbb{N}}$ is relatively compact in $L^2(0,T;L^2(\Omega))$.
Let $u_m = \Delta^i_{\mathcal{D}_m} v_m$. Since the sequence of underlying spatial discretisations is compact, it is coercive (Lemma 2.10). Denote by C_P a coercivity constant of this sequence. Using the definition (4.22) of $\Delta^i_{\mathcal{D}_m}$ with $v = (v_m)_\theta(t)$ and $w = (u_m)_\theta(t)$, the Cauchy–Schwarz inequality in the right-hand side, the definition of $C_P \geq C_{\mathcal{D}_m}$, raising to the power 2 and integrating over $t \in (0,T)$, we see that

$$\int_0^T \|(u_m)_\theta(t)\|^2_{\mathcal{D}_m} \leq C_P^2 \int_0^T \left\|\Pi^{(\theta)}_{\mathcal{D}_m} v_m(t)\right\|^2_{L^2(\Omega)} \leq C_P^2 C. \tag{4.28}$$

Set $B = L^2(\Omega)$ and $X_m = \Pi_{\mathcal{D}_m}(X_{\mathcal{D}_m,0})$, endowed with the norm

$$\|w\|_{X_m} = \inf\{\|z\|_{\mathcal{D}_m} \,:\, \Pi_{\mathcal{D}_m} z = w\}.$$

The compactness of $(\mathcal{D}_m)_{m\in\mathbb{N}}$ ensures that $(X_m)_{m\in\mathbb{N}}$ is compactly embedded in B as per Definition C.4. Estimate (4.28) and the coercivity of $(\mathcal{D}_m)_{m\in\mathbb{N}}$ prove that items 1 and 2 of the hypotheses of Proposition C.5 hold for $f_m = \Pi^{(\theta)}_{\mathcal{D}_m} u_m$. Let us now observe that, for all $z \in X_{\mathcal{D}_m,0}$,

$$\begin{aligned}
&\|\Pi_{\mathcal{D}_m} z\|_{\star,\mathcal{D}_m} \\
&\quad = \sup\left\{\int_\Omega \nabla_{\mathcal{D}_m}(\Delta^i_{\mathcal{D}_m} z)(\boldsymbol{x}) \cdot \nabla_{\mathcal{D}_m} w(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \,:\, w \in X_{\mathcal{D}_m}\,,\ \|w\|_{\mathcal{D}_m} = 1\right\} \\
&\quad = \left\|\Delta^i_{\mathcal{D}_m} z\right\|_{\mathcal{D}_m}.
\end{aligned} \tag{4.29}$$

Therefore, since $\delta_{\mathcal{D}_m} u_m(t) = \frac{\Pi_{\mathcal{D}} u_m^{(n+1)} - \Pi_{\mathcal{D}} u_m^{(n)}}{\delta t^{(n+\frac{1}{2})}}$, Hypothesis (4.25) and the coercivity property (with constant C_P) imply

$$\forall m \in \mathbb{N},\ \int_0^T \|\delta_{\mathcal{D}_m} u_m(t)\|^q_{L^2(\Omega)}\,\mathrm{d}t \leq (C_P)^q C. \tag{4.30}$$

Apply the same computation as in (C.9), followed by (C.10) with Y_m replaced by $L^2(\Omega)$. Using (4.30), this proves that (C.11) holds (still with $L^2(\Omega)$ instead of Y_m). Hence, item 3 of the hypotheses of Proposition C.5 holds with $\eta(h) = h$. Note that, contrary to the proof of Theorem C.8, we do not use an inequality similar to (C.8), which is the discrete equivalent of the Lions lemma.
Therefore, Proposition C.5 provides the existence of $\overline{u} \in L^2(0,T;L^2(\Omega))$ such that, up to a subsequence as $m \to \infty$, $\Pi^{(\theta)}_{\mathcal{D}_m} u_m \to \overline{u}$ in $L^2(0,T;L^2(\Omega))$.

Step 2: we prove that $\overline{u} = \Delta^i \overline{v}$.
By Lemma 4.8, $\overline{u}$ belongs to $L^2(0,T;H^1_0(\Omega))$ and $\nabla^{(\theta)}_{\mathcal{D}_m} u_m \to \nabla\overline{u}$ weakly in $L^2(0,T;L^2(\Omega)^d)$. Let $\overline{w} \in L^2(0,T;H^1_0(\Omega))$ and consider the sequence $(w_m)_{m\in\mathbb{N}}$ given by Lemma 4.10 for $\overline{w}$. Writing (4.22) with $v = (v_m)_\theta(t)$ and

$w = (w_m)_\theta(t)$ and integrating over $t \in (0,T)$, we can pass to the limit to see that $\overline{u}$ satisfies

$$\int_0^T \int_\Omega \nabla\overline{u}(\boldsymbol{x},t)\cdot\nabla\overline{w}(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t = \int_0^T \int_\Omega \overline{v}(\boldsymbol{x},t)\overline{w}(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t. \tag{4.31}$$

This precisely shows that $\overline{u} = \Delta^i\overline{v}$.

Step 3: proof of (4.26).
Write now (4.22) with $v = (v_m)_\theta(t)$ and $w = (u_m)_\theta(t)$, and integrate over $t \in (0,T)$. By the strong convergence of $\Pi^{(\theta)}_{\mathcal{D}_m} u_m$ to $\overline{u}$, we can pass to the limit in the left-hand side and, using (4.31) with $\overline{w} = \overline{u}$, we find

$$\lim_{m\to\infty}\int_0^T \int_\Omega |\nabla^{(\theta)}_{\mathcal{D}_m} u_m(\boldsymbol{x},t)|^2\mathrm{d}\boldsymbol{x}\mathrm{d}t$$
$$= \int_0^T \int_\Omega \overline{v}(\boldsymbol{x})\overline{u}(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\mathrm{d}t = \int_0^T \int_\Omega |\nabla\overline{u}(\boldsymbol{x},t)|^2\mathrm{d}\boldsymbol{x}\mathrm{d}t.$$

This convergence of L^2 norms shows that the convergence of $(\nabla^{(\theta)}_{\mathcal{D}_m} u_m)_{m\in\mathbb{N}}$ to $\nabla\overline{u}$ is actually strong. The proof of (4.26) is thus complete.

Step 4: assuming that $q > 1$, proof of (4.27).
We proved in Step 1 that $(\left\|\delta_{\mathcal{D}_m}(\Delta^i_{\mathcal{D}_m} v_m)\right\|_{L^q(0,T;Y_m)})_{m\in\mathbb{N}}$ is bounded (recall that $\Delta^i_{\mathcal{D}_m} v_m = u_m$). By coercivity of the sequence of GDs, this shows that $\delta_{\mathcal{D}_m}(\Delta^i_{\mathcal{D}_m} v_m)$ is bounded in $L^q(0,T;L^2(\Omega))$ and therefore converges, up to a subsequence, to some V weakly in this space.
Take $\gamma \in C_c^\infty(0,T)$ and $\psi \in C_c^\infty(\Omega)$. Multiply $\delta_{\mathcal{D}_m}(\Delta^i_{\mathcal{D}_m} v_m)(t)$ by $[\nu\gamma(t^{(n)} + (1-\nu)\gamma(t^{(n+1)})]\psi$, where $\nu = 1-\theta$ and n is such that $t \in (t^{(n)}, t^{(n+1)})$, integrate over $(\boldsymbol{x},t) \in \Omega\times(0,T)$ and use the discrete integration-by-part formula (D.17) to transfer the $\delta_{\mathcal{D}_m}$ operator onto $(\gamma(t^{(n)}))_{n=0,\dots,N}$. By smoothness of γ, passing to the limit shows that

$$\int_0^T \int_\Omega V(\boldsymbol{x},t)\gamma(t)\psi(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\mathrm{d}t = -\int_0^T \int_\Omega \overline{u}(\boldsymbol{x},t)\gamma'(t)\psi(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\mathrm{d}t.$$

We infer that $V = \partial_t\overline{u} = \partial_t(\Delta^i\overline{v})$ and the proof is complete. ∎

The next result is characterised as "weak–strong space–time" because it deals with the product of two sequences of functions, one of them being strongly compact in time and weakly in space (estimates on the time derivative), the other one being weakly compact in time and strongly in space (estimate on the spatial derivatives).

Theorem 4.17 (Weak-strong space–time convergence of a product).

Take $T > 0$, $\theta \in [0,1]$, $p = 2$ and a space–time-consistent, limit-conforming and compact sequence $((\mathcal{D}_T)_m)_{m\in\mathbb{N}}$ of space–time GDs for homogeneous Dirichlet boundary conditions, in the sense of Definitions 4.3 and 4.6. For any $m \in \mathbb{N}$, let $\beta_m, \zeta_m \in X^{N_m+1}_{\mathcal{D}_m,0}$ be such that

- *The sequences* $(\int_0^T \|\delta_{\mathcal{D}_m}\beta_m(t)\|_{\star,\mathcal{D}_m})_{m\in\mathbb{N}}$ *and* $(\|\nabla^{(\theta)}_{\mathcal{D}_m}\zeta_m\|_{L^2(0,T;L^2(\Omega)^d)})_{m\in\mathbb{N}}$ *are bounded,*
- *As* $m\to\infty$, $\Pi^{(\theta)}_{\mathcal{D}_m}\beta_m \to \overline{\beta}$ *and* $\Pi^{(\theta)}_{\mathcal{D}_m}\zeta_m \to \overline{\zeta}$ *weakly in* $L^2(\Omega\times(0,T))$.

Then it holds

$$\lim_{m\to\infty}\int_0^T\int_\Omega \Pi^{(\theta)}_{\mathcal{D}_m}\beta_m(\boldsymbol{x},t)\Pi^{(\theta)}_{\mathcal{D}_m}\zeta_m(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t = \int_0^T\int_\Omega \overline{\beta}(\boldsymbol{x},t)\,\overline{\zeta}(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t. \quad (4.32)$$

Proof. The sequence $(\beta_m)_{m\in\mathbb{N}}$ satisfies the hypotheses of Theorem 4.16. Hence, $\nabla^{(\theta)}_{\mathcal{D}_m}(\Delta^i_{\mathcal{D}_m}\beta_m)$ converges strongly to $\nabla(\Delta^i\overline{\beta})$ in $L^2(0,T;L^2(\Omega)^d)$. By definition of $\Delta^i_{\mathcal{D}_m}$, we have

$$\int_0^T\int_\Omega \Pi^{(\theta)}_{\mathcal{D}_m}\beta_m(\boldsymbol{x},t)\Pi^{(\theta)}_{\mathcal{D}_m}\zeta_m(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t = \int_0^T\int_\Omega \nabla^{(\theta)}_{\mathcal{D}_m}(\Delta^i_{\mathcal{D}_m}\beta_m)(\boldsymbol{x},t)\cdot\nabla^{(\theta)}_{\mathcal{D}_m}\zeta_m(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t. \quad (4.33)$$

By assumption on $(\zeta_m)_{m\in\mathbb{N}}$, the sequence $(\|\zeta_m\|_{L^2(0,T;X_{\mathcal{D}_m,0})})_{m\in\mathbb{N}}$ is bounded and thus, by Lemma 4.8, $\nabla^{(\theta)}_{\mathcal{D}_m}\zeta_m \to \nabla\overline{\zeta}$ weakly in $L^2(0,T;L^2(\Omega)^d)$. Passing to the limit in the right-hand side of (4.33), we infer

$$\lim_{m\to\infty}\int_0^T\int_\Omega \Pi^{(\theta)}_{\mathcal{D}_m}\beta_m(\boldsymbol{x},t)\Pi^{(\theta)}_{\mathcal{D}_m}\zeta_m(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t = \int_0^T\int_\Omega \nabla(\Delta^i\overline{\beta})(\boldsymbol{x},t)\cdot\nabla\overline{\zeta}(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t.$$

The definition of Δ^i concludes the proof of (4.32). ∎

4.2.2 Uniform-in-time compactness for space-time GDs

We now consider applications to space-time gradient discretisations (for generic BCs, as described in Section 4.1) of the results in Section C.2. The following theorem is a consequence of Corollary C.12.

Theorem 4.18 ($L^\infty(0,T;L^p(\Omega))$ compactness). *Let* $p\in(1,+\infty)$, $T>0$, $\theta\in[0,1]$, *and* $((\mathcal{D}_T)_m)_{m\in\mathbb{N}}$ *be a space–time-consistent, limit-conforming and compact sequence of space–time GDs in the sense of Definitions 4.3 and 4.6. For each* $m\in\mathbb{N}$, *let* $v_m\in X^{N_m+1}_{\mathcal{D}_m,\bullet}$. *Assume that there exist* $C>0$ *and* $q>1$ *satisfying*

$$\forall m\in\mathbb{N}\,,\ \|(v_m)_\theta\|_{L^\infty(0,T;X_{\mathcal{D}_m,\bullet})}\le C, \quad (4.34)$$

and

$$\forall m \in \mathbb{N}\,,\ \|\delta_{\mathcal{D}_m} v_m\|_{L^q(0,T;L^p(\Omega))} \leq C. \tag{4.35}$$

Then, there exists $u \in C([0,T];L^p(\Omega)) \cap L^\infty(0,T;W^{1,p}_\bullet(\Omega))$ *and a subsequence, again denoted by* $((\mathcal{D}_T)_m, v_m)_{m\in\mathbb{N}}$*, such that*

$$\lim_{m\to\infty} \sup_{t\in[0,T]} \left\| \Pi^{(\theta)}_{\mathcal{D}_m} v_m(t) - u(t) \right\|_{L^p(\Omega)} = 0. \tag{4.36}$$

Moreover, $\partial_t u \in L^q(0,T;L^p(\Omega))$ *and, along the same subsequence,* $\delta_{\mathcal{D}_m} v_m \to \partial_t u$ *weakly in* $L^q(0,T;L^p(\Omega))$*.*

Proof. We apply Corollary C.12 with $B = L^p(\Omega)$, $X_m = \Pi_{\mathcal{D}_m}(X_{\mathcal{D}_m,\bullet})$ endowed with the norm (4.21), and $u_m^{(n)} = \Pi_{\mathcal{D}_m} v_m^{(n)}$.

The compactness hypothesis on $(\mathcal{D}_m)_{m\in\mathbb{N}}$ states that $(X_m)_{m\in\mathbb{N}}$ is compactly embedded in B in the sense of Definition C.4, which yields Hypothesis (h1) in Corollary C.12. Hypothesis (h2) is satisfied owing to (4.34) and

$$\|(u_m)_\theta(t)\|_{X_m} = \|\Pi_{\mathcal{D}_m}[(v_m)_\theta(t)]\|_{X_m} \leq \|(v_m)_\theta(t)\|_{\mathcal{D}_m}\,.$$

Hypothesis (h3) of Corollary C.12 is obtained by (4.35) since $\delta_m u_m = \delta_{\mathcal{D}_m} v_m$. Hypothesis (h4) is included in the definition of space–time-consistency of $((\mathcal{D}_T)_m)_{m\in\mathbb{N}}$ (Definition 4.3).

By Corollary C.12, we obtain $u \in C([0,T];L^p(\Omega))$ such that, up to a subsequence, (4.36) holds. The fact that u belongs to $L^\infty(0,T;W^{1,p}_\bullet(\Omega))$ follows by Lemma 4.8.

It remains to prove the convergence of the discrete time derivative. By (4.35) we can assume, upon extraction of a new subsequence, that $\delta_{\mathcal{D}_m} v_m \to U$ weakly in $L^q(0,T;L^p(\Omega))$. The proof is complete by showing that $U = \partial_t u$ in the sense of distributions on $\Omega \times (0,T)$ (this also proves in particular that no further extraction was necessary). Take $\psi \in C_c^\infty(\Omega \times (0,T))$ and write, by definition (4.4) of $\delta_{\mathcal{D}_m} v_m$,

$$\begin{aligned}
&\int_0^T \int_\Omega \delta_{\mathcal{D}_m} v_m(\boldsymbol{x},t)\psi(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t \\
&= \sum_{n=1}^{N_m-1} \frac{1}{\eth t^{(n+\frac{1}{2})}} \int_{t^{(n)}}^{t^{(n+1)}} \int_\Omega \left(\Pi_{\mathcal{D}_m} v_m^{(n+1)}(\boldsymbol{x}) - \Pi_{\mathcal{D}_m} v_m^{(n)}(\boldsymbol{x}) \right) \psi(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t.
\end{aligned} \tag{4.37}$$

Set $\psi_n(\boldsymbol{x}) = \frac{1}{\eth t^{(n+\frac{1}{2})}} \int_{t^{(n)}}^{t^{(n+1)}} \psi(\boldsymbol{x},t)\mathrm{d}t$ and, for $\nu = 1-\theta$, $\psi_{n+\nu} = \nu\psi_{n+1} + (1-\nu)\psi_n$. Since ψ is smooth, $|\psi_n(\boldsymbol{x}) - \psi_{n+\nu}(\boldsymbol{x})| \leq C_\psi \eth t_{\mathcal{D}_m}$ for some C_ψ not depending on $\boldsymbol{x}$ or n. Using the discrete integration-by-parts formula (D.17), (4.37) yields, for m large enough so that $\psi_0 = \psi_{N_m} = 0$ (which is possible due to ψ vanishing on a neighbourhood of 0 and T),

$$\int_0^T \int_\Omega \delta_{\mathcal{D}_m} v_m(\boldsymbol{x},t)\psi(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t \tag{4.38}$$

$$= \sum_{n=1}^{N_m-1} \int_\Omega \left(\Pi_{\mathcal{D}_m} v_m^{(n+1)}(\boldsymbol{x}) - \Pi_{\mathcal{D}_m} v_m^{(n)}(\boldsymbol{x})\right)\psi_n(\boldsymbol{x})\mathrm{d}\boldsymbol{x}$$

$$= \sum_{n=1}^{N_m-1} \int_\Omega \left(\Pi_{\mathcal{D}_m} v_m^{(n+1)}(\boldsymbol{x}) - \Pi_{\mathcal{D}_m} v_m^{(n)}(\boldsymbol{x})\right)\psi_{n+\nu}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} + R_m$$

$$= -\sum_{n=1}^{N_m-1} \int_\Omega \Pi_{\mathcal{D}_m} v_m^{(n+\theta)}(\boldsymbol{x})\left(\psi_{n+1}(\boldsymbol{x}) - \psi_n(\boldsymbol{x})\right)\mathrm{d}\boldsymbol{x} + R_m$$

$$= -\sum_{n=1}^{N_m-1} \int_{t^{(n)}}^{t^{(n+1)}} \int_\Omega \Pi_{\mathcal{D}_m} v_m^{(n+\theta)}(\boldsymbol{x}) \frac{\psi_{n+1}(\boldsymbol{x}) - \psi_n(\boldsymbol{x})}{\delta t^{(n+\frac{1}{2})}}\mathrm{d}\boldsymbol{x} + R_m \tag{4.39}$$

where, owing to (4.35),

$$|R_m| \le C_\psi \delta t_{\mathcal{D}_m} \|\delta_{\mathcal{D}_m} v_m\|_{L^1(\Omega\times(0,T))} \to 0 \text{ as } m \to \infty.$$

The term (4.38) converges to

$$\int_0^T \int_\Omega U(\boldsymbol{x},t)\psi(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t \tag{4.40}$$

and, owing to the smoothness of ψ and the convergence of $\Pi_{\mathcal{D}_m}^{(\theta)} v_m$, the term (4.39) converges to

$$-\int_0^T \int_\Omega u(\boldsymbol{x},t)\partial_t\psi(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t. \tag{4.41}$$

The proof that $U = \partial_t u$ is complete by equating (4.40) and (4.41). ■

The uniform-in-time weak-in-space compactness result provided by the next theorem is an essential step to proving a uniform-in-time strong-in-space convergence result for gradient scheme approximations of parabolic equations (see, *e.g.*, the proof of Theorem 5.19).

Theorem 4.19 (Uniform-in-time $L^2(\Omega)$-weak compactness). *Let $T > 0$, $\theta \in [0,1]$ and $((\mathcal{D}_T)_m)_{m\in\mathbb{N}}$ be a sequence of space–time-consistent space–time GDs in the sense of Definition 4.3. For each $m \in \mathbb{N}$, let $v_m \in X_{\mathcal{D}_m,\bullet}^{N_m+1}$. Assume that there exists $C > 0$ and $q > 1$ such that, for all $m \in \mathbb{N}$,*

$$\sup_{t\in[0,T]} \left\|\Pi_{\mathcal{D}_m}^{(\theta)} v_m(t)\right\|_{L^2(\Omega)} \le C \quad \text{and} \quad \int_0^T \|\delta_{\mathcal{D}_m} v_m(t)\|_{\star,\mathcal{D}_m}^q \,\mathrm{d}t \le C \tag{4.42}$$

(see Definition 4.11 of $\|\cdot\|_{\star,\mathcal{D}}$).

Then, the sequence $(\Pi^{(\theta)}_{\mathcal{D}_m} v_m)_{m\in\mathbb{N}}$ *is relatively compact weakly in* $L^2(\Omega)$ *uniformly on* $[0,T]$*, that is, it has a subsequence which converges in the sense of Definition C.14 in page 454.*
Moreover, any limit of such a subsequence is continuous $[0,T] \to L^2(\Omega)$ *for the weak topology.*

Proof. Theorem 4.19 is a consequence of the discontinuous Arzelà–Ascoli theorem C.11, with $K = [0,T]$ and E the ball of radius C in $L^2(\Omega)$, endowed with the weak topology. Let $\{\varphi_l : l \in \mathbb{N}\} \subset C_c^\infty(\Omega)$ be a dense set in $L^2(\Omega)$ and endow E with the metric (C.24) from these φ_l. By Proposition C.15, this metric defines the weak $L^2(\Omega)$ topology.
The set E is metric compact and therefore complete, and the functions $\Pi^{(\theta)}_{\mathcal{D}_m} v_m$ have values in E. It remains to estimate $d_E(\Pi^{(\theta)}_{\mathcal{D}_m} v_m(s), \Pi^{(\theta)}_{\mathcal{D}_m} v_m(s'))$. We drop the index m in $\mathcal{D}$ for legibility.
Let $0 \le s \le s' \le T$ and take $n_1, n_2 \in \{0,\dots,N-1\}$ such that $s \in (t^{(n_1)}, t^{(n_1+1)}]$ and $s' \in (t^{(n_2)}, t^{(n_2+1)}]$. If $s = 0$ we let $n_1 = -1$ and $t^{(-1)} = 0$. In a similar way as (C.19), we write

$$\begin{aligned}
&\Pi^{(\theta)}_{\mathcal{D}} v_m(s') - \Pi^{(\theta)}_{\mathcal{D}} v_m(s) \\
&\quad = \theta \sum_{n=n_1+1}^{n_2} \mathfrak{d}t^{(n+\frac12)} \delta^{(n+\frac12)}_{\mathcal{D}} v_m + (1-\theta) \sum_{n=n_1+1}^{n_2} \mathfrak{d}t^{(n-\frac12)} \delta^{(n-\frac12)}_{\mathcal{D}} v_m,
\end{aligned}$$

where $\mathfrak{d}t^{(-\frac12)} = 0$ and $\delta^{(-\frac12)}_{\mathcal{D}} v_m = 0$. Take $I_{\mathcal{D}}\varphi_l \in X_{\mathcal{D},\bullet}$ that realises the minimum defining $S_{\mathcal{D}}(\varphi_l)$, multiply the previous relation by by $\Pi_{\mathcal{D}} I_{\mathcal{D}}\varphi_l$ and integrate over Ω. Estimates (4.17), (4.42) and the Hölder inequality (D.3) (used as in (C.20)) yield

$$\begin{aligned}
&\left| \int_\Omega \left(\Pi^{(\theta)}_{\mathcal{D}} v_m(\boldsymbol{x}, s') - \Pi^{(\theta)}_{\mathcal{D}} v_m(\boldsymbol{x}, s) \right) \Pi_{\mathcal{D}} I_{\mathcal{D}}\varphi_l(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \right| \\
&\le \left| \theta \sum_{n=n_1+1}^{n_2} \mathfrak{d}t^{(n+\frac12)} \int_\Omega \delta^{(n+\frac12)}_{\mathcal{D}} v_m(\boldsymbol{x}) \Pi_{\mathcal{D}} I_{\mathcal{D}}\varphi_l(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \right| \\
&\quad + \left| (1-\theta) \sum_{n=n_1+1}^{n_2} \mathfrak{d}t^{(n-\frac12)} \int_\Omega \delta^{(n-\frac12)}_{\mathcal{D}} v_m(\boldsymbol{x}) \Pi_{\mathcal{D}} I_{\mathcal{D}}\varphi_l(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \right| \\
&\le \theta \left\| I_{\mathcal{D}}\varphi_l \right\|_{\mathcal{D}} \sum_{n=n_1+1}^{n_2} \mathfrak{d}t^{(n+\frac12)} \left\| \delta^{(n+\frac12)}_{\mathcal{D}} v_m \right\|_{\star,\mathcal{D}} \\
&\quad + (1-\theta) \left\| I_{\mathcal{D}}\varphi_l \right\|_{\mathcal{D}} \sum_{n=n_1+1}^{n_2} \mathfrak{d}t^{(n-\frac12)} \left\| \delta^{(n-\frac12)}_{\mathcal{D}} v_m \right\|_{\star,\mathcal{D}} \\
&\le C^{1/q} \Big[\theta (t^{(n_2+1)} - t^{(n_1+1)})^{1/q'} \\
&\qquad\qquad + (1-\theta)(t^{(n_2)} - t^{(n_1)})^{1/q'} \Big] \left\| I_{\mathcal{D}}\varphi_l \right\|_{\mathcal{D}}.
\end{aligned} \tag{4.43}$$

By definition of $I_{\mathcal{D}}$ and of $\|\cdot\|_{\mathcal{D}}$ (depending on the specific boundary conditions), we have

$$\|\Pi_{\mathcal{D}} I_{\mathcal{D}} \varphi_l - \varphi_l\|_{L^2(\Omega)} \leq \widehat{S}_{\mathcal{D}}(\varphi_l)$$

and, using a triangle inequality,

$$\|I_{\mathcal{D}} \varphi_l\|_{\mathcal{D}} \leq \widehat{S}_{\mathcal{D}}(\varphi_l) + D_{\varphi_l} \leq C_{\varphi_l}$$

where D_{φ_l} and C_{φ_l} do not depend on $\mathcal{D}$ (and therefore on m). Since $t^{(n_2+1)} - t^{(n_1+1)} \leq |s'-s| + \delta t$ and $t^{(n_2)} - t^{(n_1)} \leq |s'-s| + \delta t$, the estimate on $\Pi^{(\theta)}_{\mathcal{D}_m} v_m$ in (4.42) gives, owing to (4.43),

$$\begin{aligned}
&\left| \int_\Omega \left(\Pi^{(\theta)}_{\mathcal{D}} v_m(\boldsymbol{x}, s') - \Pi^{(\theta)}_{\mathcal{D}} v_m(\boldsymbol{x}, s) \right) \varphi_l(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \right| \\
&\leq \left| \int_\Omega \left(\Pi^{(\theta)}_{\mathcal{D}} v_m(\boldsymbol{x}, s') - \Pi^{(\theta)}_{\mathcal{D}} v_m(\boldsymbol{x}, s) \right) \Pi_{\mathcal{D}} I_{\mathcal{D}} \varphi_l(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \right| + 2C \widehat{S}_{\mathcal{D}}(\varphi_l) \\
&\leq C^{1/q} C_{\varphi_l} |s'-s|^{1/q'} + C^{1/q} C_{\varphi_l} \delta t^{1/q'} + 2C \widehat{S}_{\mathcal{D}}(\varphi_l). \qquad (4.44)
\end{aligned}$$

Plugged into the definition (C.24) of the distance in E, this yields

$$\begin{aligned}
&d_E\left(\Pi^{(\theta)}_{\mathcal{D}} v_m(s'), \Pi^{(\theta)}_{\mathcal{D}} v_m(s) \right) \\
&\quad \leq \sum_{l \in \mathbb{N}} \frac{\min(1, C^{1/q'} C_{\varphi_l} |s'-s|^{1/q'})}{2^l} \\
&\qquad + \sum_{l \in \mathbb{N}} \frac{\min(1, 2C \widehat{S}_{\mathcal{D}_m}(\varphi_l) + C^{1/q'} C_{\varphi_l} \delta t_m^{1/q'})}{2^l} =: \omega(s, s') + \tau_m.
\end{aligned}$$

Using the dominated convergence theorem for series, we see that $\omega(s, s') \to 0$ as $s - s' \to 0$, and that $\tau_m \to 0$ as $m \to \infty$ (we invoke the space–time-consistency of $((\mathcal{D}_T)_m)_{m \in \mathbb{N}}$ to see that $\lim_{m \to \infty} \widehat{S}_{\mathcal{D}_m}(\varphi_l) \to 0$ for any l). Hence, the assumptions of Theorem C.11 are satisfied and the proof is complete. ∎

5

Parabolic problems

In this chapter, we consider time-dependent problems and their approximation by the gradient discretisation method (GDM).
First, in Section 5.1, we study a quasi-linear problem, which is the transient version of the quasi-linear problem studied in Section 2.1.4. An error estimate for the GDM approximation of the linear version of this problem is first proved, under additional regularity hypotheses. For the complete non-linear problem, the mathematical arguments used in the convergence analysis of the GDM come from Section 4.2. The convergence of the gradient schemes (GS) for this problem is proved under minimal regularity on the solution.
In Section 5.2, we analyse the convergence of the GDM applied to a non-conservative parabolic equation, which includes the regularised level-set equations. For this model, additional regularity on the initial condition must be assumed.
Finally, in Section 5.3, we turn to (non-local) Leray–Lions type parabolic problems with Neumann boundary conditions. These problems arise in particular from image processing models. Using again the results of Section 4.2, several convergence results for the GDM are obtained, including a uniform-in-time strong-in-space convergence result. We stress that such a convergence implies in particular the pointwise-in-time convergence, which is of high practical interest. Indeed, users of numerical techniques are often more interested in approximating a quantity of interest at a given time, rather than averaged over a time span.

5.1 The gradient discretisation method for a quasi-linear parabolic problem

In this whole section, we let $p = 2$.

5.1.1 The continuous problem

We consider the following problem: approximate the solution $\overline{u}$ of

$$\partial_t \overline{u} - \operatorname{div}\left(\Lambda(\boldsymbol{x}, \overline{u})\nabla\overline{u}\right) = f + \operatorname{div}(\boldsymbol{F}), \text{ in } \Omega \times (0,T) \tag{5.1a}$$

with initial condition

$$\overline{u}(\cdot, 0) = u_{\text{ini}}, \text{ on } \Omega, \tag{5.1b}$$

and homogeneous Dirichlet boundary conditions

$$\overline{u} = 0 \text{ on } \partial\Omega \times (0,T). \tag{5.1c}$$

The following hypotheses are assumed:

- Ω is an open bounded connected subset of $\mathbb{R}^d$, $d \in \mathbb{N}^\star$, and $T > 0$, (5.2a)
- $\Lambda : \Omega \times \mathbb{R} \to \mathcal{M}_d(\mathbb{R})$ is a Caratheodory function (*i.e.* $\Lambda(\boldsymbol{x}, s)$ is measurable w.r.t. $\boldsymbol{x}$ and continuous w.r.t. s), and there exists $\underline{\lambda}, \overline{\lambda} > 0$ such that, for a.e. $\boldsymbol{x} \in \Omega$, for all $s \in \mathbb{R}$, $\Lambda(\boldsymbol{x}, s)$ is symmetric with eigenvalues in $[\underline{\lambda}, \overline{\lambda}]$, (5.2b)
- $f \in L^2(\Omega \times (0,T))$, $\boldsymbol{F} \in L^2(\Omega \times (0,T))^d$, (5.2c)
- $u_{\text{ini}} \in L^2(\Omega)$. (5.2d)

Under Hypotheses (5.2), a function $\overline{u}$ is a weak solution of (5.1) if

$$\begin{cases} \overline{u} \in L^2(0,T; H_0^1(\Omega)) \text{ and, for all } \overline{v} \in L^2(0,T; H_0^1(\Omega)) \\ \text{ such that } \partial_t \overline{v} \in L^2(\Omega \times (0,T)) \text{ and } \overline{v}(\cdot, T) = 0, \\ -\displaystyle\int_0^T \int_\Omega \overline{u}(\boldsymbol{x},t)\partial_t \overline{v}(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t - \int_\Omega u_{\text{ini}}(\boldsymbol{x})\overline{v}(\boldsymbol{x},0)\mathrm{d}\boldsymbol{x} \\ \qquad + \displaystyle\int_0^T \int_\Omega \Lambda(\boldsymbol{x}, \overline{u}(\boldsymbol{x},t))\nabla\overline{u}(\boldsymbol{x},t) \cdot \nabla\overline{v}(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t \\ \qquad\quad = \displaystyle\int_0^T \int_\Omega (f(\boldsymbol{x},t)\overline{v}(\boldsymbol{x},t) - \boldsymbol{F}(\boldsymbol{x},t) \cdot \nabla\overline{v}(\boldsymbol{x},t))\mathrm{d}\boldsymbol{x}\mathrm{d}t. \end{cases} \tag{5.3}$$

Taking $\overline{v} \in C_c^\infty(\Omega \times (0,T))$ in this equation shows that (5.1a) then holds in the sense of distributions. Since $\Lambda(\boldsymbol{x}, \overline{u})\nabla\overline{u}$ and $\boldsymbol{F}$ both belong to $L^2(\Omega \times (0,T))^d$, this implies that $\partial_t \overline{u} \in L^2(0,T; H^{-1}(\Omega))$. As a consequence, $\overline{u} \in C([0,T]; L^2(\Omega))$ and, integrating by parts the first term in (5.3) and using the density of $C_c^\infty([0,T]; H_0^1(\Omega))$ in $L^2(0,T; H_0^1(\Omega))$ (see [67, Corollary 1.3.1]), we see that $\overline{u}$ satisfies

$$\begin{cases} \overline{u} \in L^2(0,T;H_0^1(\Omega)) \cap C([0,T];L^2(\Omega))\,,\ \partial_t \overline{u} \in L^2(0,T;H^{-1}(\Omega))\,, \\ \overline{u}(\cdot,0) = u_{\rm ini} \text{ and, for all } w \in L^2(0,T;H_0^1(\Omega)), \\ \displaystyle\int_0^T \langle \partial_t \overline{u}(\cdot,t), w(\cdot,t)\rangle_{H^{-1},H_0^1} {\rm d}t \\ \qquad \displaystyle + \int_0^T \int_\Omega \Lambda(\boldsymbol{x},\overline{u}(\boldsymbol{x},t)) \nabla \overline{u}(\boldsymbol{x},t) \cdot \nabla w(\boldsymbol{x},t) {\rm d}\boldsymbol{x}{\rm d}t \\ \qquad\qquad \displaystyle = \int_0^T \int_\Omega (f(\boldsymbol{x},t) w(\boldsymbol{x},t) - \boldsymbol{F}(\boldsymbol{x},t) \cdot \nabla w(\boldsymbol{x},t)) {\rm d}\boldsymbol{x}{\rm d}t. \end{cases} \tag{5.4}$$

Remark 5.1. The existence of at least one solution $\overline{u}$ to (5.3), and therefore to (5.4), will be a consequence of the convergence analysis of the GDM (see Remark 5.5).
In the linear case, that is $\Lambda(\boldsymbol{x},\overline{u}) = \Lambda(\boldsymbol{x})$, estimates on the continuous solution show that this solution $\overline{u}$ is also unique.

5.1.2 The gradient scheme

Recalling that $p = 2$, let $\mathcal{D}_T = (X_{\mathcal{D},0}, \Pi_{\mathcal{D}}, \nabla_{\mathcal{D}}, \mathcal{I}_{\mathcal{D}}, (t^{(n)})_{n=0,\ldots,N})$ and $\theta \in [\frac{1}{2},1]$ be a space–time GD for homogeneous Dirichlet boundary conditions in the sense of Definition 4.1. Using a θ-scheme for the time stepping, the GDM applied to Problem (5.4) leads to the following GS: find a family $(u^{(n)})_{n=0,\ldots,N} \in X_{\mathcal{D},0}^{N+1}$ such that, recalling the notations (4.2) and (4.4),

$$\begin{cases} u^{(0)} = \mathcal{I}_{\mathcal{D}} u_{\rm ini} \text{ and, for all } n = 0,\ldots,N-1,\ u^{(n+1)} \text{ satisfies} \\ \displaystyle\int_\Omega \delta_{\mathcal{D}}^{(n+\frac{1}{2})} u(\boldsymbol{x}) \Pi_{\mathcal{D}} v(\boldsymbol{x}) {\rm d}\boldsymbol{x} \\ \qquad \displaystyle + \int_\Omega \Lambda(\boldsymbol{x}, \Pi_{\mathcal{D}} u^{(n+\theta)}(\boldsymbol{x})) \nabla_{\mathcal{D}} u^{(n+\theta)}(\boldsymbol{x}) \cdot \nabla_{\mathcal{D}} v(\boldsymbol{x}) {\rm d}\boldsymbol{x} \\ \qquad \displaystyle = \frac{1}{\delta t^{(n+\frac{1}{2})}} \int_{t^{(n)}}^{t^{(n+1)}} \int_\Omega (f(\boldsymbol{x},t) \Pi_{\mathcal{D}} v(\boldsymbol{x}) - \boldsymbol{F}(\boldsymbol{x},t) \cdot \nabla_{\mathcal{D}} v(\boldsymbol{x})) {\rm d}\boldsymbol{x}{\rm d}t, \\ \qquad\qquad \forall v \in X_{\mathcal{D},0}. \end{cases} \tag{5.5}$$

Here, of course, $u^{(n)}$ is expected to provide an approximation of $\overline{u}$ at time t_n.

Remark 5.2 (Practical implementation of the GS (5.5))
For any $n = 0,\ldots,N-1$, taking $u^{(n+\theta)}$ as unknown, and using

$$u^{(n+1)} = \frac{u^{(n+\theta)} - (1-\theta) u^{(n)}}{\theta},$$

the implementation of the GS (5.5) is similar to that of the GS (2.52) for the steady quasi-linear problem.

5.1.3 Error estimate in the linear case

We now consider Problem (5.1) under Hypotheses (5.2) and the following additional hypotheses.

$$\boldsymbol{F} = 0 \quad \text{and} \quad \Lambda(\cdot, s) = \mathrm{Id}. \tag{5.6}$$

The equation considered here is therefore $\partial_t \overline{u} - \Delta \overline{u} = f$, with homogeneous Dirichlet boundary conditions. The following theorem states, under some regularity assumptions, an error estimate based on the notion of space size of a GD as in Definition 2.22 page 29.

Theorem 5.3 (Error estimate, linear case and regular solution). *Under Hypotheses (5.2) and (5.6), let $\mathcal{D}_T$ be a space–time GD for homogeneous Dirichlet boundary conditions, in the sense of Definition 4.1. Let $h_{\mathcal{D}} := h_{\mathcal{D}}(W^{2,\infty}(\Omega) \cap W_0^{1,p}(\Omega); W^{1,\infty}(\Omega)^d) > 0$ be given by Definition 2.22. Assume that the solution $\overline{u}$ to (5.4) is Lipschitz-continuous $[0,T] \to W^{2,\infty}(\Omega)$, and let u be the solution to the GS (5.5) with $\theta = 1$. The error due to the interpolation of the initial condition is denoted by*

$$e_{\mathcal{D}}^{\mathrm{ini}} = \|u_{\mathrm{ini}} - \Pi_{\mathcal{D}} \mathcal{I}_{\mathcal{D}} u_{\mathrm{ini}}\|_{L^2(\Omega)}. \tag{5.7}$$

Let $C_P \geq C_{\mathcal{D}}$; then there exists $C > 0$, depending only on $\overline{u}$, Ω, T and C_P, such that

$$\max_{t \in [0,T]} \left\| \Pi_{\mathcal{D}}^{(1)} u(\cdot, t) - \overline{u}(\cdot, t) \right\|_{L^2(\Omega)} \leq C(\delta t_{\mathcal{D}} + h_{\mathcal{D}} + e_{\mathcal{D}}^{\mathrm{ini}})$$

and

$$\left\| \nabla_{\mathcal{D}}^{(1)} u - \nabla \overline{u} \right\|_{L^2(\Omega \times (0,T))^d} \leq C(\delta t_{\mathcal{D}} + h_{\mathcal{D}} + e_{\mathcal{D}}^{\mathrm{ini}}).$$

Proof. In the following proof, we denote by C_i various quantities depending only on $\overline{u}$, Ω, T and C_P. For the sake of brevity, if $n \in \{0, \ldots, N-1\}$ and $g = f$, $\overline{u}$ or $\partial_t \overline{u}$, we set

$$g^{(n+1)}(\boldsymbol{x}) = \frac{1}{\delta t^{(n+\frac{1}{2})}} \int_{t^{(n)}}^{t^{(n+1)}} g(\boldsymbol{x}, t) \mathrm{d}t.$$

We also let $\overline{u}^{(0)} = \overline{u}(0)$.

Step 1: a linear spatial interpolator.
The interpolator $I_{\mathcal{D}}$ defined by (2.34) enables us to "plug" the exact solution into the scheme, which is an essential process in establishing error estimates. However, this $I_{\mathcal{D}}$ is not necessarily linear, which becomes a problem for parabolic equations. We therefore need a slightly modified version of this interpolator. We define $I_{\mathcal{D}}^{(2)} : H_0^1(\Omega) \to X_{\mathcal{D},0}$ by: for $\varphi \in H_0^1(\Omega)$,

$$I_{\mathcal{D}}^{(2)} \varphi = \operatorname*{argmin}_{w \in X_{\mathcal{D},0}} \left(\|\Pi_{\mathcal{D}} w - \varphi\|_{L^2(\Omega)}^2 + \|\nabla_{\mathcal{D}} w - \nabla \varphi\|_{L^2(\Omega)^d}^2 \right). \tag{5.8}$$

Let $V = \{(\Pi_{\mathcal{D}} w, \nabla_{\mathcal{D}} w) : w \in X_{\mathcal{D},0}\}$ and $\mathcal{P} : L^2(\Omega) \times L^2(\Omega)^d \to V$ be the orthogonal projection. Since $\|\nabla_{\mathcal{D}} \cdot\|_{L^2(\Omega)^d}$ is a norm on $X_{\mathcal{D},0}$, for any $z \in V$ there exists a unique $\mathcal{R}z \in X_{\mathcal{D},0}$ such that $(\Pi_{\mathcal{D}} \mathcal{R}z, \nabla_{\mathcal{D}} \mathcal{R}z) = z$. This defines a linear continuous mapping $\mathcal{R} : V \to X_{\mathcal{D},0}$, and (5.8) shows that $I_{\mathcal{D}}^{(2)} \varphi = \mathcal{R} \circ \mathcal{P}(\varphi, \nabla\varphi)$ for all $\varphi \in H_0^1(\Omega)$. Hence, $I_{\mathcal{D}}^{(2)} \varphi$ is uniquely defined and $I_{\mathcal{D}}^{(2)}$ is linear continuous. The characterisation of the orthogonal projection $\mathcal{P}$ also shows that, for all $\varphi \in H_0^1(\Omega)$ and $w \in X_{\mathcal{D},0}$,

$$\int_\Omega \Pi_{\mathcal{D}} I_{\mathcal{D}}^{(2)} \varphi(\boldsymbol{x}) \Pi_{\mathcal{D}} w(\boldsymbol{x}) + \nabla_{\mathcal{D}} I_{\mathcal{D}}^{(2)} \varphi(\boldsymbol{x}) \cdot \nabla_{\mathcal{D}} w(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} = \int_\Omega \varphi(\boldsymbol{x}) \Pi_{\mathcal{D}} w(\boldsymbol{x}) + \nabla\varphi(\boldsymbol{x}) \cdot \nabla_{\mathcal{D}} w(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}.$$

Let $\varphi \in H_0^1(\Omega)$. Taking $v \in X_{\mathcal{D},0}$ that realises the minimum defining $S_{\mathcal{D}}(\varphi)$ and using the definition of $I_{\mathcal{D}}^{(2)}$ shows that

$$\Big(\Big\| \Pi_{\mathcal{D}} I_{\mathcal{D}}^{(2)} \varphi - \varphi \Big\|_{L^2(\Omega)}^2 + \Big\| \nabla_{\mathcal{D}} I_{\mathcal{D}}^{(2)} \varphi - \nabla\varphi \Big\|_{L^2(\Omega)^d}^2 \Big)^{1/2} \le \Big(\|\Pi_{\mathcal{D}} v - \varphi\|_{L^2(\Omega)}^2 + \|\nabla_{\mathcal{D}} v - \nabla\varphi\|_{L^2(\Omega)^d}^2 \Big)^{1/2} \le \sqrt{2} S_{\mathcal{D}}(\varphi). \quad (5.9)$$

Since $\overline{u} \in C([0,T]; W^{2,\infty}(\Omega))$, (5.9) can be applied to $\varphi = \overline{u}(t^{(n+1)})$. The regularity on $\overline{u}$ shows that $\nabla\overline{u} : [0,T] \to L^2(\Omega)^d$ is Lipschitz-continuous. Hence, using (2.14) and (2.16) in Definition 2.22,

$$\begin{aligned} \Big\| \nabla\overline{u}^{(n+1)} - \nabla_{\mathcal{D}} I_{\mathcal{D}}^{(2)} \overline{u}(t^{(n+1)}) \Big\|_{L^2(\Omega)^d} &\le \Big\| \nabla\overline{u}^{(n+1)} - \nabla\overline{u}(t^{(n+1)}) \Big\|_{L^2(\Omega)^d} + S_{\mathcal{D}}(\overline{u}(t^{(n+1)})) \\ &\le C_1(\delta t_{\mathcal{D}} + h_{\mathcal{D}}). \end{aligned} \quad (5.10)$$

By regularity assumption on $\overline{u}$, the quantity $\|\partial_t \overline{u}^{(n+1)}\|_{W^{2,\infty}(\Omega)}$ is bounded independently of n. Applying (5.9) to $\varphi = \partial_t \overline{u}^{(n+1)} = \frac{\overline{u}(t^{(n+1)}) - \overline{u}(t^{(n)})}{\delta t^{(n+\frac{1}{2})}}$, using the linearity of $I_{\mathcal{D}}^{(2)}$ and invoking (2.14), we obtain

$$\left\| \frac{\Pi_{\mathcal{D}} I_{\mathcal{D}}^{(2)} \overline{u}(t^{(n+1)}) - \Pi_{\mathcal{D}} I_{\mathcal{D}}^{(2)} \overline{u}(t^{(n)})}{\delta t^{(n+\frac{1}{2})}} - \partial_t \overline{u}^{(n+1)} \right\|_{L^2(\Omega)} \le C_2 h_{\mathcal{D}}. \quad (5.11)$$

Step 2: proof of the error estimates.
Since $\nabla\overline{u}^{(n+1)} \in H_{\mathrm{div}}(\Omega)$ we can write, for all $v \in X_{\mathcal{D},0}$,

$$\int_\Omega \Big(\Pi_{\mathcal{D}} v(\boldsymbol{x}) \mathrm{div}(\nabla\overline{u}^{(n+1)})(\boldsymbol{x}) + \nabla\overline{u}^{(n+1)}(\boldsymbol{x}) \cdot \nabla_{\mathcal{D}} v(\boldsymbol{x}) \Big) \mathrm{d}\boldsymbol{x}$$

$$\leq W_{\mathcal{D}}(\nabla \overline{u}^{(n+1)}) \left\| v \right\|_{\mathcal{D}}.$$

Owing to the regularity of $\overline{u}$, the equation $\partial_t \overline{u} - f = \operatorname{div}(\nabla \overline{u})$ is satisfied a.e. in space and time. Averaging over time in $(t^{(n)}, t^{(n+1)})$ gives $\partial_t \overline{u}^{(n+1)} - f^{(n+1)} = \operatorname{div}(\nabla \overline{u}^{(n+1)})$ a.e. in space, and thus

$$\int_\Omega \Big(\Pi_{\mathcal{D}} v(\boldsymbol{x}) \left(\partial_t \overline{u}^{(n+1)}(\boldsymbol{x}) - f^{(n+1)}(\boldsymbol{x}) \right) + \nabla \overline{u}^{(n+1)}(\boldsymbol{x}) \cdot \nabla_{\mathcal{D}} v(\boldsymbol{x}) \Big) \mathrm{d}\boldsymbol{x}$$
$$\leq W_{\mathcal{D}}(\nabla \overline{u}^{(n+1)}) \left\| v \right\|_{\mathcal{D}}.$$

Use the GS (5.5) to replace the term $f^{(n+1)}$ in the left-hand side. Since $\overline{u} \in C([0,T]; W^{2,\infty}(\Omega))$, the quantity $\|\nabla \overline{u}^{(n+1)}\|_{W^{1,\infty}(\Omega)^d}$ is bounded independently on n and thus, using (2.15)–(2.16) in Definition 2.22,

$$\int_\Omega \Pi_{\mathcal{D}} v(\boldsymbol{x}) \left(\partial_t \overline{u}^{(n+1)}(\boldsymbol{x}) - \delta_{\mathcal{D}}^{(n+\frac{1}{2})} u(\boldsymbol{x}) \right) \mathrm{d}\boldsymbol{x}$$
$$+ \int_\Omega \left(\nabla \overline{u}^{(n+1)}(\boldsymbol{x}) - \nabla_{\mathcal{D}} u^{(n+1)} \right) \cdot \nabla_{\mathcal{D}} v(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \leq C_3 h_{\mathcal{D}} \left\| v \right\|_{\mathcal{D}}. \quad (5.12)$$

For $k = 0, \ldots, N$, set $e^{(k)} = I_{\mathcal{D}}^{(2)} \overline{u}(t^{(k)}) - u^{(k)}$. We have

$$\delta_{\mathcal{D}}^{(n+\frac{1}{2})} e = \left[\frac{\Pi_{\mathcal{D}} I_{\mathcal{D}}^{(2)} \overline{u}(t^{(n+1)}) - \Pi_{\mathcal{D}} I_{\mathcal{D}}^{(2)} \overline{u}(t^{(n)})}{\delta t^{(n+\frac{1}{2})}} - \partial_t \overline{u}^{(n+1)} \right]$$
$$+ \left[\partial_t \overline{u}^{(n+1)} - \delta_{\mathcal{D}}^{(n+\frac{1}{2})} u \right].$$

and

$$\nabla_{\mathcal{D}} e^{(n+1)} = \left[\nabla_{\mathcal{D}} I_{\mathcal{D}}^{(2)} \overline{u}(t^{(n+1)}) - \nabla \overline{u}^{(n+1)} \right] + \left[\nabla \overline{u}^{(n+1)} - \nabla_{\mathcal{D}} u^{(n+1)} \right]$$

Then (5.12), (5.11), (5.10) and the definition of $C_{\mathcal{D}}$ give

$$\int_\Omega \Pi_{\mathcal{D}} v(\boldsymbol{x}) \delta_{\mathcal{D}}^{(n+\frac{1}{2})} e(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} + \int_\Omega \nabla_{\mathcal{D}} e^{(n+1)}(\boldsymbol{x}) \cdot \nabla_{\mathcal{D}} v(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}$$
$$\leq C_4 (\delta t_{\mathcal{D}} + h_{\mathcal{D}}) \left\| v \right\|_{\mathcal{D}}.$$

Take $v = \delta t^{(n+\frac{1}{2})} e^{(n+1)}$, and sum over $n = 0, \ldots, m-1$ for some $m \in \{1, \ldots, N\}$. Recalling the definition of $\|\cdot\|_{\mathcal{D}}$,

$$\sum_{n=0}^{m-1} \int_\Omega \Pi_{\mathcal{D}} e^{(n+1)}(\boldsymbol{x}) \left[\Pi_{\mathcal{D}} e^{(n+1)}(\boldsymbol{x}) - \Pi_{\mathcal{D}} e^{(n)}(\boldsymbol{x}) \right] \mathrm{d}\boldsymbol{x}$$
$$+ \sum_{n=0}^{m-1} \delta t^{(n+\frac{1}{2})} \int_\Omega |\nabla_{\mathcal{D}} e^{(n+1)}(\boldsymbol{x})|^2 \mathrm{d}\boldsymbol{x}$$

$$\leq \sum_{n=0}^{m-1} C_4(\delta t_{\mathcal{D}} + h_{\mathcal{D}})\delta t^{(n+\frac{1}{2})} \left(\int_\Omega |\nabla_{\mathcal{D}} e^{(n+1)}(\boldsymbol{x})|^2 \mathrm{d}\boldsymbol{x} \right)^{1/2}. \quad (5.13)$$

We now apply the relation

$$\forall a, b \in \mathbb{R},\ b(b-a) = \frac{1}{2}b^2 - \frac{1}{2}a^2 + \frac{1}{2}(b-a)^2 \geq \frac{1}{2}b^2 - \frac{1}{2}a^2 \quad (5.14)$$

to $a = \Pi_{\mathcal{D}} e^{(n)}(\boldsymbol{x})$ and $b = \Pi_{\mathcal{D}} e^{(n+1)}(\boldsymbol{x})$. Using the Young inequality (D.8) with $p = p' = 2$ in the right-hand side of (5.13), this leads to

$$\begin{aligned}
&\frac{1}{2}\int_\Omega (\Pi_{\mathcal{D}} e^{(m)}(\boldsymbol{x}))^2 \mathrm{d}\boldsymbol{x} + \sum_{n=0}^{m-1} \delta t^{(n+\frac{1}{2})} \int_\Omega |\nabla_{\mathcal{D}} e^{(n+1)}(\boldsymbol{x})|^2 \mathrm{d}\boldsymbol{x} \\
&\quad \leq \frac{1}{2}\int_\Omega (\Pi_{\mathcal{D}} e^{(0)}(\boldsymbol{x}))^2 \mathrm{d}\boldsymbol{x} + \frac{1}{2}\sum_{n=0}^{m-1} \delta t^{(n+\frac{1}{2})} \int_\Omega |\nabla_{\mathcal{D}} e^{(n+1)}(\boldsymbol{x})|^2 \mathrm{d}\boldsymbol{x} \\
&\qquad + \frac{1}{2}\sum_{n=0}^{m-1} C_4^2 (\delta t_{\mathcal{D}} + h_{\mathcal{D}})^2 \delta t^{(n+\frac{1}{2})}. \qquad (5.15)
\end{aligned}$$

Owing to (2.15)–(2.16), (5.7) and Estimate (5.9), since $u^{(0)} = \mathcal{I}_{\mathcal{D}} u_{\mathrm{ini}} = \mathcal{I}_{\mathcal{D}} \overline{u}(0)$,

$$\begin{aligned}
\left\| \Pi_{\mathcal{D}} e^{(0)} \right\|_{L^2(\Omega)} &\leq \left\| \Pi_{\mathcal{D}} I_{\mathcal{D}}^{(2)} \overline{u}(0) - \overline{u}(0) \right\|_{L^2(\Omega)} + \| \overline{u}(0) - \Pi_{\mathcal{D}} \mathcal{I}_{\mathcal{D}} \overline{u}(0) \|_{L^2(\Omega)} \\
&\leq C_5 h_{\mathcal{D}} + e_{\mathcal{D}}^{\mathrm{ini}}.
\end{aligned}$$

Hence, recalling the definition of $\nabla^{(1)}$ and using $\sum_{n=0}^{m-1} \delta t^{(n+\frac{1}{2})} \leq T$, Equation (5.15) yields

$$\begin{aligned}
&\frac{1}{2}\int_\Omega (\Pi_{\mathcal{D}} e^{(m)}(\boldsymbol{x}))^2 \mathrm{d}\boldsymbol{x} + \frac{1}{2}\int_0^{t^{(m)}} \int_\Omega |\nabla_{\mathcal{D}}^{(1)} e(\boldsymbol{x}, t)|^2 \mathrm{d}\boldsymbol{x}\mathrm{d}t \\
&\qquad\qquad \leq C_6 (\delta t_{\mathcal{D}} + h_{\mathcal{D}} + e_{\mathcal{D}}^{\mathrm{ini}})^2. \quad (5.16)
\end{aligned}$$

Using a triangle inequality, (5.9), and the power-of-sums inequality (D.13) with $\alpha = 1/2$, Equation (5.16) leads on one hand to: for all $m = 1, \ldots, N$,

$$\begin{aligned}
\left\| \Pi_{\mathcal{D}} u^{(m)} - \overline{u}(t^{(m)}) \right\|_{L^2(\Omega)} &\leq C_7 (\delta t_{\mathcal{D}} + h_{\mathcal{D}} + e_{\mathcal{D}}^{\mathrm{ini}}) + \sqrt{2} S_{\mathcal{D}}(\overline{u}(t^{(m)})) \\
&\leq C_8 (\delta t_{\mathcal{D}} + h_{\mathcal{D}} + e_{\mathcal{D}}^{\mathrm{ini}}). \qquad (5.17)
\end{aligned}$$

On the other hand, using again (5.9) and a triangle inequality, Equation (5.16) with $m = N-1$ and the power-of-sums inequality (D.12) with $\alpha = 2$ lead to

$$\sum_{n=0}^{N-1} \delta t^{(n+\frac{1}{2})} \left\| \nabla_{\mathcal{D}} u^{(n+1)} - \nabla \overline{u}(t^{(n+1)}) \right\|_{L^2(\Omega)}^2$$

$$\leq 4C_6(\delta t_{\mathcal{D}} + h_{\mathcal{D}} + e_{\mathcal{D}}^{\text{ini}})^2 + 4\sum_{n=0}^{N-1} \delta t^{(n+\frac{1}{2})} S_{\mathcal{D}}(\overline{u}(t^{(n+1)}))^2$$
$$\leq C_9(\delta t_{\mathcal{D}} + h_{\mathcal{D}} + e_{\mathcal{D}}^{\text{ini}})^2. \tag{5.18}$$

The conclusion follows from (5.17), (5.18) and the Lipschitz-continuity of $\overline{u} : [0,T] \to H^1(\Omega)$ to compare $\overline{u}(t)$ (resp. $\nabla\overline{u}(t)$) with $\overline{u}(t^{(n+1)})$ (resp. $\nabla\overline{u}(t^{(n+1)})$) when $t \in (t^{(n)}, t^{(n+1)}]$. ∎

5.1.4 Convergence analysis in the non-linear case

We come back to the generic quasi-linear model (5.1). The convergence result we intend to prove is the following.

Theorem 5.4 (Convergence of the GDM). *Under Assumptions (5.2), let $\theta \in [\frac{1}{2}, 1]$ and $((\mathcal{D}_T)_m)_{m\in\mathbb{N}}$ be a sequence of space–time GDs for homogeneous Dirichlet boundary conditions in the sense of Definition 4.1, which is space–time-consistent, limit-conforming and compact in the sense of Definitions 4.3 and 4.6. For any $m \in \mathbb{N}$, let u_m be a solution to (5.5) with $\mathcal{D}_T = (\mathcal{D}_T)_m$. Then, up to a subsequence as $m \to \infty$,*

$$\sup_{t\in[0,T]} \left\| \Pi_{\mathcal{D}_m}^{(\theta)} u_m(t) - \overline{u}(t) \right\|_{L^2(\Omega)} \to 0 \tag{5.19a}$$

$$\nabla_{\mathcal{D}_m}^{(\theta)} u_m \to \nabla\overline{u} \quad \text{in } L^2(\Omega \times (0,T))^d, \tag{5.19b}$$

where $\overline{u}$ is a solution to (5.3) (and thus also (5.4)).

Remark 5.5. We do not assume the existence of a solution $\overline{u}$ to the continuous problem. The convergence analysis establishes this existence.

The analysis of any GDM for non-linear models starts by establishing *a priori* estimates on the solution to the GS. These estimates are useful to establish the existence of this solution, and to invoke the compactness results of Section 4.2.

Lemma 5.6 ($L^\infty(0,T;L^2(\Omega))$ estimate and discrete $L^2(0,T;H_0^1(\Omega))$ estimate). *Under Assumptions (5.2), let $\theta \in [\frac{1}{2}, 1]$ and $\mathcal{D}_T$ be a space–time GD for homogeneous Dirichlet boundary conditions, in the sense of Definition 4.1. Let u be a solution to the corresponding GS (5.5). Then, for any $k = 0, \ldots, N$,*

$$\begin{aligned}
&\frac{1}{2}\int_\Omega (\Pi_{\mathcal{D}} u^{(k)}(\boldsymbol{x}))^2 \mathrm{d}\boldsymbol{x} \\
&+ \int_0^{t^{(k)}} \int_\Omega \Lambda(\boldsymbol{x}, \Pi_{\mathcal{D}}^{(\theta)} u(\boldsymbol{x},t)) \nabla_{\mathcal{D}}^{(\theta)} u(\boldsymbol{x},t) \cdot \nabla_{\mathcal{D}}^{(\theta)} u(\boldsymbol{x},t) \mathrm{d}\boldsymbol{x}\mathrm{d}t \\
&\quad \leq \frac{1}{2}\int_\Omega (\Pi_{\mathcal{D}} \mathcal{I}_{\mathcal{D}} u_{\text{ini}}(\boldsymbol{x}))^2 \mathrm{d}\boldsymbol{x} \\
&\qquad + \int_0^{t^{(k)}} \int_\Omega (f(\boldsymbol{x},t) \Pi_{\mathcal{D}}^{(\theta)} u(\boldsymbol{x},t) - \boldsymbol{F}(\boldsymbol{x},t) \cdot \nabla_{\mathcal{D}}^{(\theta)} u(\boldsymbol{x},t)) \mathrm{d}\boldsymbol{x}\mathrm{d}t.
\end{aligned} \tag{5.20}$$

Consequently, there exists $C_{10} > 0$ depending only on $C_P \geq C_{\mathcal{D}}$ (see Definition 2.2), $C_{\rm ini} \geq \|\Pi_{\mathcal{D}}\mathcal{I}_{\mathcal{D}}u_{\rm ini}\|_{L^2(\Omega)}$, f, $\boldsymbol{F}$, and $\underline{\lambda}$ such that

$$\sup_{t\in[0,T]} \left\|\Pi_{\mathcal{D}}^{(\theta)}u(t)\right\|_{L^2(\Omega)} \leq C_{10} \quad \text{and} \quad \left\|\nabla_{\mathcal{D}}^{(\theta)}u\right\|_{L^2(\Omega\times(0,T))^d} \leq C_{10}. \tag{5.21}$$

Moreover, there exists at least one solution u to the GS (5.5).

Proof. Relation (5.14) is generalised to the following: for all $a, b \in \mathbb{R}$,

$$\begin{aligned}(a-b)(\theta a + (1-\theta)b) &= (a-b)\left[\left(\theta - \frac{1}{2}\right)a + \left(\frac{1}{2} - \theta\right)b\right] + \frac{1}{2}(a-b)(a+b)\\ &= \left(\theta - \frac{1}{2}\right)(a-b)^2 + \frac{1}{2}(a^2-b^2) \geq \frac{1}{2}(a^2-b^2).\end{aligned}$$

Let $n \in \{0, \ldots, N-1\}$. Applying the above relation to $a = \Pi_{\mathcal{D}}u^{(n+1)}$ and $b = \Pi_{\mathcal{D}}u^{(n)}$ yields

$$\delta t^{(n+\frac{1}{2})}\delta_{\mathcal{D}}^{(n+\frac{1}{2})}u\ \Pi_{\mathcal{D}}u^{(n+\theta)} \geq \frac{1}{2}\left((\Pi_{\mathcal{D}}u^{(n+1)})^2 - (\Pi_{\mathcal{D}}u^{(n)})^2\right). \tag{5.22}$$

Setting $v = \delta t^{(n+\frac{1}{2})}u^{(n+\theta)}$ in (5.5) and summing over $n = 0, \ldots, k-1$ (we assume here that $k \geq 1$, the case $k = 0$ in (5.20) is trivial) therefore leads to

$$\begin{aligned}&\frac{1}{2}\sum_{n=0}^{k-1}\left(\int_\Omega (\Pi_{\mathcal{D}}u^{(n+1)}(\boldsymbol{x}))^2\mathrm{d}\boldsymbol{x} - \int_\Omega (\Pi_{\mathcal{D}}u^{(n)}(\boldsymbol{x}))^2\mathrm{d}\boldsymbol{x}\right)\\ &+\sum_{n=0}^{k-1}\delta t^{(n+\frac{1}{2})}\int_\Omega \Lambda(\boldsymbol{x}, \Pi_{\mathcal{D}}u^{(n+\theta)}(\boldsymbol{x}))\nabla_{\mathcal{D}}u^{(n+\theta)}(\boldsymbol{x})\cdot\nabla_{\mathcal{D}}u^{(n+\theta)}(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\\ &\leq \sum_{n=0}^{k-1}\int_{t^{(n)}}^{t^{(n+1)}}\int_\Omega (f(\boldsymbol{x},t)\Pi_{\mathcal{D}}u^{(n+\theta)}(\boldsymbol{x}) - \boldsymbol{F}(\boldsymbol{x},t)\cdot\nabla_{\mathcal{D}}u^{(n+\theta)}(\boldsymbol{x}))\mathrm{d}\boldsymbol{x}\mathrm{d}t.\end{aligned} \tag{5.23}$$

The first sum is telescopic and reduces to

$$\begin{aligned}&\int_\Omega (\Pi_{\mathcal{D}}u^{(k)}(\boldsymbol{x}))^2\mathrm{d}\boldsymbol{x} - \int_\Omega (\Pi_{\mathcal{D}}u^{(0)}(\boldsymbol{x}))^2\mathrm{d}\boldsymbol{x}\\ &\qquad = \int_\Omega (\Pi_{\mathcal{D}}u^{(k)}(\boldsymbol{x}))^2\mathrm{d}\boldsymbol{x} - \int_\Omega (\Pi_{\mathcal{D}}\mathcal{I}_{\mathcal{D}}u_{\rm ini}(\boldsymbol{x}))^2\mathrm{d}\boldsymbol{x}.\end{aligned}$$

Recalling that $\Pi_{\mathcal{D}}^{(\theta)}u(\boldsymbol{x},t) = \Pi_{\mathcal{D}}u^{(n+\theta)}(\boldsymbol{x})$ and $\nabla_{\mathcal{D}}^{(\theta)}u(\boldsymbol{x},t) = \nabla_{\mathcal{D}}u^{(n+\theta)}(\boldsymbol{x})$ whenever $t \in (t^{(n)}, t^{(n+1)}]$, Equation (5.23) can then be recast as (5.20).
Using the Cauchy–Schwarz inequality (*i.e.* (D.5) with $p = p' = 2$), the Young inequality (D.9) and the definition (2.1) of $C_{\mathcal{D}}$, we write

$$\int_0^{t^{(k)}}\int_\Omega (f(\boldsymbol{x},t)\Pi_{\mathcal{D}}^{(\theta)}u(\boldsymbol{x},t) - \boldsymbol{F}(\boldsymbol{x},t)\cdot\nabla_{\mathcal{D}}^{(\theta)}u(\boldsymbol{x},t))\mathrm{d}\boldsymbol{x}\mathrm{d}t$$

$$
\begin{aligned}
&\leq \|f\|_{L^2(\Omega\times(0,t^{(k)}))} \left\|\Pi_{\mathcal{D}}^{(\theta)} u\right\|_{L^2(\Omega\times(0,t^{(k)}))} \\
&\quad + \|\boldsymbol{F}\|_{L^2(\Omega\times(0,t^{(k)}))^d} \left\|\nabla_{\mathcal{D}}^{(\theta)} u\right\|_{L^2(\Omega\times(0,t^{(k)}))^d} \\
&\leq \frac{C_{\mathcal{D}}^2}{\underline{\lambda}} \|f\|^2_{L^2(\Omega\times(0,t^{(k)}))} + \frac{\underline{\lambda}}{4C_{\mathcal{D}}^2} \left\|\Pi_{\mathcal{D}}^{(\theta)} u\right\|^2_{L^2(\Omega\times(0,t^{(k)}))} \\
&\quad + \frac{1}{\underline{\lambda}} \|\boldsymbol{F}\|^2_{L^2(\Omega\times(0,t^{(k)}))^d} + \frac{\underline{\lambda}}{4} \left\|\nabla_{\mathcal{D}}^{(\theta)} u\right\|^2_{L^2(\Omega\times(0,t^{(k)}))^d} \\
&\leq \frac{C_{\mathcal{D}}^2}{\underline{\lambda}} \|f\|^2_{L^2(\Omega\times(0,t^{(k)}))} + \frac{1}{\underline{\lambda}} \|\boldsymbol{F}\|^2_{L^2(\Omega\times(0,t^{(k)}))^d} \\
&\quad + \frac{\underline{\lambda}}{2} \left\|\nabla_{\mathcal{D}}^{(\theta)} u\right\|^2_{L^2(\Omega\times(0,t^{(k)}))^d}. \qquad (5.24)
\end{aligned}
$$

Plugged into (5.20) and using the coercivity of Λ, this gives

$$
\begin{aligned}
&\frac{1}{2}\int_\Omega (\Pi_{\mathcal{D}} u^{(k)}(\boldsymbol{x}))^2 \mathrm{d}\boldsymbol{x} + \underline{\lambda} \left\|\nabla_{\mathcal{D}}^{(\theta)} u\right\|^2_{L^2(\Omega\times(0,t^{(k)}))^d} \\
&\quad \leq \frac{1}{2} C_{\mathrm{ini}}^2 + \frac{C_{\mathcal{D}}^2}{\underline{\lambda}} \|f\|^2_{L^2(\Omega\times(0,t^{(k)}))} + \frac{1}{\underline{\lambda}} \|\boldsymbol{F}\|^2_{L^2(\Omega\times(0,t^{(k)}))^d} \\
&\qquad + \frac{\underline{\lambda}}{2} \left\|\nabla_{\mathcal{D}}^{(\theta)} u\right\|^2_{L^2(\Omega\times(0,t^{(k)}))^d}.
\end{aligned}
$$

Hence,

$$
\begin{aligned}
&\frac{1}{2} \max_{k=0,\ldots,N} \left\|\Pi_{\mathcal{D}} u^{(k)}\right\|^2_{L^2(\Omega)} + \frac{\underline{\lambda}}{2} \left\|\nabla_{\mathcal{D}}^{(\theta)} u\right\|^2_{L^2(\Omega\times(0,T)^d} \\
&\qquad \leq \frac{1}{2} C_{\mathrm{ini}}^2 + \frac{C_{\mathcal{D}}^2}{\underline{\lambda}} \|f\|^2_{L^2(\Omega\times(0,T))} + \frac{1}{\underline{\lambda}} \|\boldsymbol{F}\|^2_{L^2(\Omega\times(0,T))^d}.
\end{aligned}
$$

The estimates in (5.21) follow from this inequality and from the fact that, by definition (4.2) of $\Pi_{\mathcal{D}}^{(\theta)}$,

$$
\|\Pi_{\mathcal{D}} u^{(n+\theta)}\|_{L^2(\Omega)} \leq \theta \|\Pi_{\mathcal{D}} u^{(n+1)}\|_{L^2(\Omega)} + (1-\theta)\|\Pi_{\mathcal{D}} u^{(n)}\|_{L^2(\Omega)}.
$$

Following the same arguments as in the proof of Theorem 2.35, it is easy to establish by induction that, for each $n = 0, \ldots, N-1$, there is a solution $u^{(n+1)}$ to the equation in (5.5). This shows that this GS has at least one solution u. ∎

Lemma 5.7 (Estimate on the dual norm of the discrete time derivative). *Under Assumptions (5.2), let $\theta \in [\frac{1}{2}, 1]$ and $\mathcal{D}_T$ be a space–time GD for homogeneous Dirichlet boundary conditions, in the sense of Definition 4.1. Let u be a solution to the corresponding GS (5.5). Then there exists C_{11}, depending only on $C_P \geq C_{\mathcal{D}}$, $C_{\mathrm{ini}} \geq \|\Pi_{\mathcal{D}} I_{\mathcal{D}} u_{\mathrm{ini}}\|_{L^2(\Omega)}$, f, $\boldsymbol{F}$, $\underline{\lambda}$ and $\overline{\lambda}$, such that*

$$\int_0^T \|\delta_{\mathcal{D}} u(t)\|_{\star,\mathcal{D}}^2 \,\mathrm{d}t \le C_{11}, \tag{5.25}$$

where the dual norm $\|\cdot\|_{\star,\mathcal{D}}$ *is defined by* (4.16).

Proof. In (5.5), choose $v \in X_{\mathcal{D},0}$ which realises the supremum in the definition (4.16) of $\|\delta_{\mathcal{D}}^{(n+\frac{1}{2})} u\|_{\star,\mathcal{D}}$. Recalling that $\|v\|_{\mathcal{D}} = 1$ and applying the Cauchy–Schwarz inequality as well as the definition (2.1) of $C_{\mathcal{D}}$, we get

$$\begin{aligned}
\left\|\delta_{\mathcal{D}}^{(n+\frac{1}{2})} u\right\|_{\star,\mathcal{D}} &\le \overline{\lambda} \left\|\nabla_{\mathcal{D}} u^{(n+\theta)}\right\|_{L^2(\Omega)} \\
&\quad + \frac{1}{\mathscr{d}t^{(n+\frac{1}{2})}} \int_{t^{(n)}}^{t^{(n+1)}} (C_{\mathcal{D}} \|f(\cdot,t)\|_{L^2(\Omega)} + \|\boldsymbol{F}(\cdot,t)\|_{L^2(\Omega)^d}) \mathrm{d}t \\
&= \frac{1}{\mathscr{d}t^{(n+\frac{1}{2})}} \int_{t^{(n)}}^{t^{(n+1)}} \Big[\overline{\lambda} \left\|\nabla_{\mathcal{D}}^{(\theta)} u(t)\right\|_{L^2(\Omega)} + C_{\mathcal{D}} \|f(\cdot,t)\|_{L^2(\Omega)} \\
&\qquad\qquad + \|\boldsymbol{F}(\cdot,t)\|_{L^2(\Omega)^d}\Big] \mathrm{d}t
\end{aligned}$$

Square, use the Jensen inequality (D.10), multiply by $\mathscr{d}t^{(n+\frac{1}{2})}$ and apply the power-of-sums inequality (D.14). Recalling the definition (4.4) of $\delta_{\mathcal{D}} u$, this yields

$$\begin{aligned}
&\int_{t^{(n)}}^{t^{(n+1)}} \|\delta_{\mathcal{D}} u(t)\|_{\star,\mathcal{D}}^2 \,\mathrm{d}t \le \\
&\quad 3 \int_{t^{(n)}}^{t^{(n+1)}} \left[\overline{\lambda}^2 \left\|\nabla_{\mathcal{D}}^{(\theta)} u(t)\right\|_{\mathcal{D}}^2 + C_P^2 \|f(\cdot,t)\|_{L^2(\Omega)}^2 + \|\boldsymbol{F}(\cdot,t)\|_{L^2(\Omega)^d}^2\right] \mathrm{d}t.
\end{aligned}$$

We conclude the proof of (5.25) by summing over $n = 0, \ldots, N-1$ and by invoking Estimates (5.21). ∎

We are now ready to prove the convergence of the GS (5.5).

Proof of Theorem 5.4.

We note that since $((\mathcal{D}_T)_m)_{m\in\mathbb{N}}$ is compact, it is also coercive (see Lemma 2.10).

Step 1: Application of compactness results.

By Estimates (5.21), Lemma 4.8 gives the existence of some $\overline{u} \in L^2(0,T; H_0^1(\Omega))$ such that, up to a subsequence as $m \to \infty$, $\Pi_{\mathcal{D}_m}^{(\theta)} u_m \to \overline{u}$ weakly in $L^2(\Omega \times (0,T))$ and $\nabla_{\mathcal{D}_m}^{(\theta)} u_m \to \nabla \overline{u}$ weakly in $L^2(\Omega \times (0,T))^d$. Estimate (5.25) and Theorem 4.14 show that, in fact, $\Pi_{\mathcal{D}_m}^{(\theta)} u_m$ converges strongly to $\overline{u}$ in $L^2(\Omega \times (0,T))$.

Step 2: $\overline{u}$ is a solution to (5.3) (and thus also (5.4)).

Let $\overline{v} \in L^2(0,T; H_0^1(\Omega))$ be such that $\partial_t \overline{v} \in L^2(\Omega \times (0,T))$ and $\overline{v}(T,\cdot) = 0$. Let $(v_m)_{m\in\mathbb{N}}$ be given for $\overline{v}$ by Lemma 4.10 (with $1-\theta$ instead of θ).

In the following, we drop the index m in $\mathcal{D}_m$, N_m and v_m for legibility reasons. Introduce $v^{(n+(1-\theta))}$ as test function in (5.5), multiply by $\delta t^{(n+\frac{1}{2})}$, and sum the result on $n = 0, \ldots, N-1$. Recalling the definitions (4.2), this gives $T_1^{(m)} + T_2^{(m)} = T_3^{(m)}$ with

$$T_1^{(m)} = \sum_{n=0}^{N-1} \int_\Omega \left[\Pi_\mathcal{D} u^{(n+1)}(\boldsymbol{x}) - \Pi_\mathcal{D} u^{(n)}(\boldsymbol{x}) \right] \Pi_\mathcal{D} v^{(n+(1-\theta))}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x},$$

$$T_2^{(m)} = \int_0^T \int_\Omega \Lambda(\boldsymbol{x}, \Pi_\mathcal{D}^{(\theta)} u(\boldsymbol{x},t)) \nabla_\mathcal{D}^{(\theta)} u(\boldsymbol{x},t) \cdot \nabla_\mathcal{D}^{(1-\theta)} v(\boldsymbol{x},t) \mathrm{d}\boldsymbol{x}\mathrm{d}t,$$

and

$$T_3^{(m)} = \int_0^T \int_\Omega \left(f(\boldsymbol{x},t) \Pi_\mathcal{D}^{(1-\theta)} v(\boldsymbol{x},t) - \boldsymbol{F}(\boldsymbol{x},t) \cdot \nabla_\mathcal{D}^{(1-\theta)} v(\boldsymbol{x},t) \right) \mathrm{d}\boldsymbol{x}\mathrm{d}t.$$

Applying the discrete integration-by-parts (D.17), with $\nu = 1-\theta$, to $T_1^{(m)}$ and using the fact that $v^{(N)} = 0$, we write

$$\begin{aligned}
T_1^{(m)} &= -\sum_{n=0}^{N-1} \int_\Omega \Pi_\mathcal{D} u^{(n+\theta)}(\boldsymbol{x}) \left[\Pi_\mathcal{D} v^{(n+1)}(\boldsymbol{x}) - \Pi_\mathcal{D} v^{(n)}(\boldsymbol{x}) \right] \mathrm{d}\boldsymbol{x} \\
&\quad - \int_\Omega \Pi_\mathcal{D} u^{(0)}(\boldsymbol{x}) \Pi_\mathcal{D} v^{(0)}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \\
&= -\sum_{n=0}^{N-1} \int_{t^{(n)}}^{t^{(n+1)}} \int_\Omega \Pi_\mathcal{D}^{(\theta)} u(\boldsymbol{x},t) \delta_\mathcal{D} v(\boldsymbol{x},t) \mathrm{d}\boldsymbol{x}\mathrm{d}t \\
&\quad - \int_\Omega \Pi_\mathcal{D} u^{(0)}(\boldsymbol{x}) \Pi_\mathcal{D} v^{(0)}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \\
&= -\int_0^T \int_\Omega \Pi_\mathcal{D}^{(\theta)} u(\boldsymbol{x},t) \delta_\mathcal{D} v(\boldsymbol{x},t) \mathrm{d}\boldsymbol{x}\mathrm{d}t \\
&\quad - \int_\Omega \Pi_\mathcal{D} \mathcal{I}_\mathcal{D} u_{\mathrm{ini}}(\boldsymbol{x}) \Pi_\mathcal{D}^{(1-\theta)} v(\boldsymbol{x},0) \mathrm{d}\boldsymbol{x}.
\end{aligned}$$

Recall that $\Pi_\mathcal{D}^{(\theta)} u \to \overline{u}$ in $L^2(\Omega \times (0,T))$ and, by space–time-consistency of $((\mathcal{D}_T)_m)_{m\in\mathbb{N}}$, that $\Pi_\mathcal{D} \mathcal{I}_\mathcal{D} u_{\mathrm{ini}} \to u_{\mathrm{ini}}$ in $L^2(\Omega)$. The convergence properties of $(v_m)_{m\in\mathbb{N}}$ stated in (4.10c) and (4.10b) (with $1-\theta$ instead of θ) show that

$$\lim_{m\to\infty} T_1^{(m)} = -\int_0^T \int_\Omega \overline{u}(\boldsymbol{x},t) \partial_t \overline{v}(\boldsymbol{x},t) \mathrm{d}\boldsymbol{x}\mathrm{d}t - \int_\Omega u_{\mathrm{ini}}(\boldsymbol{x}) \overline{v}(\boldsymbol{x},0) \mathrm{d}\boldsymbol{x}. \quad (5.26)$$

Since $\Pi_\mathcal{D}^{(\theta)} u \to \overline{u}$ in $L^2(\Omega \times (0,T))$, Lemma D.9 (non-linear strong convergence property) shows that $\Lambda(\cdot, \Pi_\mathcal{D}^{(\theta)} u) \nabla_\mathcal{D}^{(1-\theta)} v$ converges to $\Lambda(\cdot, \overline{u}) \nabla \overline{v}$ in $L^2(\Omega \times (0,T))^d$ as $m \to \infty$. Hence, using the symmetry of Λ and the weak-strong convergence result of Lemma D.8,

$$\lim_{m\to\infty} T_2^{(m)} = \lim_{m\to\infty} \int_0^T \int_\Omega \nabla_{\mathcal{D}}^{(\theta)} u(\boldsymbol{x},t) \cdot \Lambda(\boldsymbol{x}, \Pi_{\mathcal{D}}^{(\theta)} u(\boldsymbol{x},t)) \nabla_{\mathcal{D}}^{(1-\theta)} v(\boldsymbol{x},t) \mathrm{d}\boldsymbol{x}\mathrm{d}t$$
$$= \int_0^T \int_\Omega \Lambda(\boldsymbol{x}, \overline{u}(\boldsymbol{x},t)) \nabla \overline{u}(\boldsymbol{x},t) \cdot \nabla \overline{v}(\boldsymbol{x},t) \mathrm{d}\boldsymbol{x}\mathrm{d}t. \tag{5.27}$$

The convergences of $\Pi_{\mathcal{D}}^{(1-\theta)} v$ and $\nabla_{\mathcal{D}}^{(1-\theta)} v$ readily give

$$\lim_{m\to\infty} T_3^{(m)} = \int_0^T \int_\Omega \left(f(\boldsymbol{x},t)\overline{v}(\boldsymbol{x},t) - \boldsymbol{F}(\boldsymbol{x},t) \cdot \nabla \overline{v}(\boldsymbol{x},t) \right) \mathrm{d}\boldsymbol{x}\mathrm{d}t. \tag{5.28}$$

Using (5.26), (5.27) and (5.28) to pass to the limit $m \to \infty$ in $T_1^{(m)} + T_2^{(m)} = T_3^{(m)}$ shows that $\overline{u}$ satisfies the equation in (5.3).

Step 3: Uniform-in-time convergence of $\Pi_{\mathcal{D}_m}^{(\theta)} u_m$.
Let $s \in [0,T]$ and $(s_m)_{m\geq 1}$ be a sequence in $[0,T]$ that converges to s. Assume first that $s_m > 0$ and let $k(m) \in \{0,\ldots,N_m - 1\}$ be such that $s_m \in (t^{(k(m))}, t^{(k(m)+1)}]$. By convexity of the square function and by Definition (4.2) of $\Pi_{\mathcal{D}}^{(\theta)}$,

$$(\Pi_{\mathcal{D}_m}^{(\theta)} u_m(\boldsymbol{x}, s_m))^2 = \left(\theta \Pi_{\mathcal{D}_m} u_m^{(k(m)+1)} + (1-\theta) \Pi_{\mathcal{D}_m} u_m^{(k(m))} \right)^2$$
$$\leq \theta (\Pi_{\mathcal{D}_m} u_m^{(k(m)+1)})^2 + (1-\theta)(\Pi_{\mathcal{D}_m} u_m^{(k(m))})^2. \tag{5.29}$$

Set $s_m^{(-)} := t^{(k(m))}$ and $s_m^{(+)} := t^{(k(m)+1)}$, which both converge to s as $m \to \infty$. Write (5.20) for $k = k(m)+1$, multiply by θ, write (5.20) with $k = k(m)$, and multiply by $1-\theta$. Summing the two inequalities thus obtained and using (5.29) yields

$$\frac{1}{2} \int_\Omega (\Pi_{\mathcal{D}_m}^{(\theta)} u(\boldsymbol{x}, s_m))^2 \mathrm{d}\boldsymbol{x}$$
$$+ \int_0^{s_m^{(-)}} \int_\Omega \Lambda(\boldsymbol{x}, \Pi_{\mathcal{D}_m}^{(\theta)} u(\boldsymbol{x},t)) \nabla_{\mathcal{D}_m}^{(\theta)} u(\boldsymbol{x},t) \cdot \nabla_{\mathcal{D}_m}^{(\theta)} u(\boldsymbol{x},t) \mathrm{d}\boldsymbol{x}\mathrm{d}t$$
$$+ \left[\theta \int_{s_m^{(-)}}^{s_m^{(+)}} \int_\Omega \Lambda(\boldsymbol{x}, \Pi_{\mathcal{D}_m}^{(\theta)} u(\boldsymbol{x},t)) \nabla_{\mathcal{D}_m}^{(\theta)} u(\boldsymbol{x},t) \cdot \nabla_{\mathcal{D}_m}^{(\theta)} u(\boldsymbol{x},t) \mathrm{d}\boldsymbol{x}\mathrm{d}t \right]$$
$$\leq \frac{1}{2} \int_\Omega (\Pi_{\mathcal{D}_m} \mathcal{I}_{\mathcal{D}_m} u_{\mathrm{ini}}(\boldsymbol{x}))^2 \mathrm{d}\boldsymbol{x}$$
$$+ \int_0^{s_m^{(-)}} \int_\Omega (f(\boldsymbol{x},t) \Pi_{\mathcal{D}_m}^{(\theta)} u(\boldsymbol{x},t) - \boldsymbol{F}(\boldsymbol{x},t) \cdot \nabla_{\mathcal{D}_m}^{(\theta)} u(\boldsymbol{x},t)) \mathrm{d}\boldsymbol{x}\mathrm{d}t$$
$$+ \theta \int_{s_m^{(-)}}^{s_m^{(+)}} \int_\Omega (f(\boldsymbol{x},t) \Pi_{\mathcal{D}_m}^{(\theta)} u(\boldsymbol{x},t) - \boldsymbol{F}(\boldsymbol{x},t) \cdot \nabla_{\mathcal{D}_m}^{(\theta)} u(\boldsymbol{x},t)) \mathrm{d}\boldsymbol{x}\mathrm{d}t. \tag{5.30}$$

Inequality (5.30) also obviously holds if $s_m = 0$ (with, in this case, $s_m^{(+)} = s_m^{(-)} = 0$). Our aim is to take the superior limit of (5.30). We first analyse the behaviour of all the terms, except the first one.

The Cauchy–Schwarz inequality for the semi-definite positive symmetric form

$$W \in L^2(\Omega \times (0,T))^d \to \int_0^{s_m^{(-)}} \int_\Omega \Lambda(\boldsymbol{x}, \Pi_{\mathcal{D}_m}^{(\theta)} u_m(\boldsymbol{x},t)) W(\boldsymbol{x},t) \cdot W(\boldsymbol{x},t) \mathrm{d}\boldsymbol{x}\mathrm{d}t$$

shows that

$$\begin{aligned}
&\left(\int_0^{s_m^{(-)}} \int_\Omega \Lambda(\boldsymbol{x}, \Pi_{\mathcal{D}_m}^{(\theta)} u_m(\boldsymbol{x},t)) \nabla_{\mathcal{D}_m}^{(\theta)} u_m(\boldsymbol{x},t) \cdot \nabla \overline{u}(\boldsymbol{x},t) \mathrm{d}\boldsymbol{x}\mathrm{d}t \right)^2 \\
&\quad \le \left(\int_0^{s_m^{(-)}} \int_\Omega \Lambda(\boldsymbol{x}, \Pi_{\mathcal{D}_m}^{(\theta)} u_m(\boldsymbol{x},t)) \nabla_{\mathcal{D}_m}^{(\theta)} u_m(\boldsymbol{x},t) \cdot \nabla_{\mathcal{D}_m}^{(\theta)} u_m(\boldsymbol{x},t) \mathrm{d}\boldsymbol{x}\mathrm{d}t \right) \\
&\qquad \times \left(\int_0^{s_m^{(-)}} \int_\Omega \Lambda(\boldsymbol{x}, \Pi_{\mathcal{D}_m}^{(\theta)} u_m(\boldsymbol{x},t)) \nabla \overline{u}(\boldsymbol{x},t) \cdot \nabla \overline{u}(\boldsymbol{x},t) \mathrm{d}\boldsymbol{x}\mathrm{d}t \right).
\end{aligned} \tag{5.31}$$

As $m \to \infty$, we have

$$\begin{aligned}
&\Pi_{\mathcal{D}_m}^{(\theta)} u_m \to \overline{u} \text{ strongly in } L^2(\Omega \times (0,T)), \text{ and} \\
&\mathbf{1}_{[0,s_m^{(-)}]} \nabla \overline{u} \to \mathbf{1}_{[0,s]} \nabla \overline{u} \text{ strongly in } L^2(\Omega \times (0,T))^d.
\end{aligned}$$

Here, $\mathbf{1}_{[a,b]}$ is the function of time such that $\mathbf{1}_{[a,b]}(t) = 1$ if $t \in [a,b]$, and $\mathbf{1}_{[a,b]}(t) = 0$ otherwise. The non-linear strong convergence property stated in Lemma D.9 page 465 then shows that, as $m \to \infty$,

$$\mathbf{1}_{[0,s_m^{(-)}]} \Lambda(\cdot, \Pi_{\mathcal{D}_m}^{(\theta)} u_m) \nabla \overline{u} \to \mathbf{1}_{[0,s]} \Lambda(\cdot, \overline{u}) \nabla \overline{u} \text{ strongly in } L^2(\Omega \times (0,T))^d.$$

Owing to Lemma D.8 (weak-strong convergence property) and to the weak convergence in $L^2(\Omega \times (0,T))^d$ of $\nabla_{\mathcal{D}_m}^{(\theta)} u_m$ to $\nabla \overline{u}$, the left-hand side of (5.31) and the second term in the right-hand side of (5.31) pass to the limit. Taking the inferior limit of this inequality and dividing by $\int_0^s \int_\Omega \Lambda(\boldsymbol{x}, \overline{u}) \nabla \overline{u} \cdot \nabla \overline{u}$, we deduce that

$$\begin{aligned}
&\int_0^s \int_\Omega \Lambda(\boldsymbol{x}, \overline{u}(\boldsymbol{x},t)) \nabla \overline{u}(\boldsymbol{x},t) \cdot \nabla \overline{u}(\boldsymbol{x},t) \mathrm{d}\boldsymbol{x}\mathrm{d}t \\
&\le \liminf_{m \to \infty} \int_0^{s_m^{(-)}} \int_\Omega \Lambda(\boldsymbol{x}, \Pi_{\mathcal{D}_m}^{(\theta)} u_m(\boldsymbol{x},t)) \nabla_{\mathcal{D}_m}^{(\theta)} u_m(\boldsymbol{x},t) \cdot \nabla_{\mathcal{D}_m}^{(\theta)} u_m(\boldsymbol{x},t) \mathrm{d}\boldsymbol{x}\mathrm{d}t.
\end{aligned} \tag{5.32}$$

The space–time-consistency of $((\mathcal{D}_T)_m)_{m \in \mathbb{N}}$ (Definition 4.3) gives

$$\int_\Omega (\Pi_{\mathcal{D}_m} \mathcal{I}_{\mathcal{D}_m} u_{\mathrm{ini}}(\boldsymbol{x}))^2 \mathrm{d}\boldsymbol{x} \to \int_\Omega (u_{\mathrm{ini}}(\boldsymbol{x}))^2 \mathrm{d}\boldsymbol{x} \text{ as } m \to \infty. \tag{5.33}$$

Still considering $m \to \infty$, we have $\mathbf{1}_{[0,s_m^{(-)}]} f \to \mathbf{1}_{[0,s]} f$ in $L^2(\Omega \times (0,T))$ and $\mathbf{1}_{[0,s_m^{(-)}]} \boldsymbol{F} \to \mathbf{1}_{[0,s]} \boldsymbol{F}$ in $L^2(\Omega \times (0,T))^d$. The weak convergences of $\Pi_{\mathcal{D}_m}^{(\theta)} u_m$ and $\nabla_{\mathcal{D}_m}^{(\theta)} u_m$ thus give, as $m \to \infty$,

$$\int_0^{s_m^{(-)}} \int_\Omega (f(\boldsymbol{x},t)\Pi_{\mathcal{D}_m}^{(\theta)} u(\boldsymbol{x},t) - \boldsymbol{F}(\boldsymbol{x},t)\cdot\nabla_{\mathcal{D}_m}^{(\theta)} u(\boldsymbol{x},t))\mathrm{d}\boldsymbol{x}\mathrm{d}t$$
$$\to \int_0^s \int_\Omega (f(\boldsymbol{x},t)\overline{u}(\boldsymbol{x},t) - \boldsymbol{F}(\boldsymbol{x},t)\cdot\nabla\overline{u}(\boldsymbol{x},t))\mathrm{d}\boldsymbol{x}\mathrm{d}t. \quad (5.34)$$

Finally, since $\mathbf{1}_{[s_m^{(-)},s_m^{(+)}]} f \to 0$ in $L^2(\Omega\times(0,T))$ and $\mathbf{1}_{[s_m^{(-)},s_m^{(+)}]}\boldsymbol{F} \to 0$ in $L^2(\Omega\times(0,T))^d$,

$$\int_{s_m^{(-)}}^{s_m^{(+)}} \int_\Omega (f(\boldsymbol{x},t)\Pi_{\mathcal{D}_m}^{(\theta)} u(\boldsymbol{x},t) - \boldsymbol{F}(\boldsymbol{x},t)\cdot\nabla_{\mathcal{D}_m}^{(\theta)} u(\boldsymbol{x},t))\mathrm{d}\boldsymbol{x}\mathrm{d}t \to 0. \quad (5.35)$$

We now come back to (5.30), drop the non-negative term in brackets, move the second term from the left-hand side to the right-hand side, and take the superior limit. The convergences (5.32), (5.33), (5.34) and (5.35) yield

$$\begin{aligned}
&\limsup_{m\to\infty} \frac{1}{2}\int_\Omega (\Pi_{\mathcal{D}_m}^{(\theta)} u_m(\boldsymbol{x},s_m))^2\mathrm{d}\boldsymbol{x}\\
&\quad\le \frac{1}{2}\int_\Omega u_{\mathrm{ini}}(\boldsymbol{x})^2\mathrm{d}\boldsymbol{x} + \int_0^s\int_\Omega (f(\boldsymbol{x},t)\overline{u}(\boldsymbol{x},t) - \boldsymbol{F}(\boldsymbol{x},t)\cdot\nabla\overline{u}(\boldsymbol{x},t))\mathrm{d}\boldsymbol{x}\mathrm{d}t\\
&\qquad - \liminf_{m\to\infty}\int_0^{s_m^{(-)}}\int_\Omega \Lambda(\boldsymbol{x},\Pi_{\mathcal{D}_m}^{(\theta)} u_m(\boldsymbol{x},t))\nabla_{\mathcal{D}_m}^{(\theta)} u_m(\boldsymbol{x},t)\cdot\nabla_{\mathcal{D}_m}^{(\theta)} u_m(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t\\
&\quad\le \frac{1}{2}\int_\Omega u_{\mathrm{ini}}(\boldsymbol{x})^2\mathrm{d}\boldsymbol{x} + \int_0^s\int_\Omega (f(\boldsymbol{x},t)\overline{u}(\boldsymbol{x},t) - \boldsymbol{F}(\boldsymbol{x},t)\cdot\nabla\overline{u}(\boldsymbol{x},t))\mathrm{d}\boldsymbol{x}\mathrm{d}t\\
&\qquad - \int_0^s\int_\Omega \Lambda(\boldsymbol{x},\overline{u}(\boldsymbol{x},t))\nabla\overline{u}(\boldsymbol{x},t)\cdot\nabla\overline{u}(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t. \qquad (5.36)
\end{aligned}$$

Since $\overline{u}\in L^2(0,T;H_0^1(\Omega))$ and $\partial_t\overline{u}\in L^2(0,T;H^{-1}(\Omega))$, the following integration by parts is justified (see [67, Section 2.5.2]):

$$\int_0^s \langle\partial_t\overline{u}(t),\overline{u}(t)\rangle_{H^{-1},H_0^1}\mathrm{d}t = \frac{1}{2}\int_\Omega \overline{u}(\boldsymbol{x},s)^2\mathrm{d}\boldsymbol{x} - \frac{1}{2}\int_\Omega \overline{u}(\boldsymbol{x},0)^2\mathrm{d}\boldsymbol{x}.$$

Making $w = \overline{u}\mathbf{1}_{[0,s]}(t)$ in (5.4), we therefore see that

$$\begin{aligned}
&\frac{1}{2}\int_\Omega \overline{u}(\boldsymbol{x},s)^2\mathrm{d}\boldsymbol{x} + \int_0^s\int_\Omega \Lambda(\boldsymbol{x},\overline{u}(\boldsymbol{x},t))\nabla\overline{u}(\boldsymbol{x},t)\cdot\nabla\overline{u}(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t\\
&\quad = \frac{1}{2}\int_\Omega u_{\mathrm{ini}}(\boldsymbol{x})^2\mathrm{d}\boldsymbol{x} + \int_0^s\int_\Omega (f(\boldsymbol{x},t)\overline{u}(\boldsymbol{x},t) - \boldsymbol{F}(\boldsymbol{x},t)\cdot\nabla\overline{u}(\boldsymbol{x},t))\mathrm{d}\boldsymbol{x}\mathrm{d}t. \qquad (5.37)
\end{aligned}$$

Used in (5.36), this relation gives

$$\limsup_{m\to\infty}\int_\Omega (\Pi_{\mathcal{D}_m}^{(\theta)} u_m(\boldsymbol{x},s_m))^2\mathrm{d}\boldsymbol{x} \le \int_\Omega \overline{u}(\boldsymbol{x},s)^2\mathrm{d}\boldsymbol{x}, \quad (5.38)$$

Owing to Theorem 4.19 and to Estimates (5.21) and (5.25), $(\Pi^{(\theta)}_{\mathcal{D}_m} u_m)_{m\in\mathbb{N}}$ converges to $\overline{u}$ weakly in $L^2(\Omega)$ uniformly in $[0,T]$ (in the sense of Definition C.14). Hence, $\Pi^{(\theta)}_{\mathcal{D}_m} u_m(\cdot, s_m) \to \overline{u}(\cdot, s)$ weakly in $L^2(\Omega)$ as $m \to \infty$. Estimate (5.38) and a standard reasoning in Hilbert spaces then show that this convergence is actually strong in $L^2(\Omega)$. By Lemma C.13, we infer that (5.19a) holds.

Step 4: Strong convergence of $\nabla^{(\theta)}_{\mathcal{D}_m} u_m$.
Note that, by (5.19a), $\Pi^{(\theta)}_{\mathcal{D}_m} u_m(\cdot, T) \to \overline{u}(\cdot, T)$ in $L^2(\Omega)$. Write (5.30) with $s_m = T$ (so that $s_m^{(-)} = t^{(N_m-1)}$ and $s_m^{(+)} = T$), move the first term to the right-hand side and take the superior limit. We can pass, as in the previous step, to the limit in all the terms on the right-hand side. Let $g_m : [0,T] \to \mathbb{R}$ be the function such that $g_m = 1$ on $[0, s_m^{(-)}]$ and $g_m = \theta$ on $(s_m^{(-)}, T]$. Using (5.37) with $s = T$, we obtain

$$\begin{aligned}
&\limsup_{m\to\infty} \int_0^T \int_\Omega g_m(t) \Lambda(\boldsymbol{x}, \Pi^{(\theta)}_{\mathcal{D}_m} u_m(\boldsymbol{x},t)) \nabla^{(\theta)}_{\mathcal{D}_m} u_m(\boldsymbol{x},t) \cdot \nabla^{(\theta)}_{\mathcal{D}_m} u_m(\boldsymbol{x},t) \mathrm{d}\boldsymbol{x}\mathrm{d}t \\
&\quad \le \frac{1}{2} \int_\Omega u_{\mathrm{ini}}(\boldsymbol{x})^2 \mathrm{d}\boldsymbol{x} + \int_0^T (f(\boldsymbol{x},t)\overline{u}(\boldsymbol{x},t) - \boldsymbol{F}(\boldsymbol{x},t) \cdot \nabla \overline{u}(\boldsymbol{x},t)) \mathrm{d}\boldsymbol{x}\mathrm{d}t \\
&\qquad - \frac{1}{2} \int_\Omega \overline{u}(\boldsymbol{x},T)^2 \mathrm{d}\boldsymbol{x} \\
&\quad = \int_0^T \int_\Omega \Lambda(\boldsymbol{x}, \overline{u}(\boldsymbol{x},t)) \nabla \overline{u}(\boldsymbol{x},t) \cdot \nabla \overline{u}(\boldsymbol{x},t) \mathrm{d}\boldsymbol{x}\mathrm{d}t.
\end{aligned}$$

Using this estimate, the strong convergence in $L^2(\Omega \times (0,T))$ of $\Pi^{(\theta)}_{\mathcal{D}_m} u_m$ to $\overline{u}$, Lemma D.9, the strong convergence in $L^2(\Omega \times (0,T))^d$ of $g_m \nabla \overline{u}$ to $\nabla \overline{u}$, the weak convergence in $L^2(\Omega \times (0,T))^d$ of $\nabla^{(\theta)}_{\mathcal{D}_m} u_m$ to $\nabla \overline{u}$, and developing the following expression in a similar fashion as (2.60), we infer that

$$\begin{aligned}
\limsup_{m\to\infty} \int_0^T \int_\Omega g_m(t) \Lambda(\boldsymbol{x}, \Pi^{(\theta)}_{\mathcal{D}_m} u_m(\boldsymbol{x},t)) (\nabla^{(\theta)}_{\mathcal{D}_m} u_m(\boldsymbol{x},t) - \nabla \overline{u}(\boldsymbol{x},t)) \\
\cdot (\nabla^{(\theta)}_{\mathcal{D}_m} u_m(\boldsymbol{x},t) - \nabla \overline{u}(\boldsymbol{x},t)) \mathrm{d}\boldsymbol{x}\mathrm{d}t \le 0.
\end{aligned}$$

By coercivity of Λ and since $g_m \ge \theta \ge \frac{1}{2}$, this shows that, as $m \to \infty$,

$$\int_0^T \int_\Omega \left| \nabla^{(\theta)}_{\mathcal{D}_m} u_m(\boldsymbol{x},t) - \nabla \overline{u}(\boldsymbol{x},t) \right|^2 \mathrm{d}\boldsymbol{x}\mathrm{d}t \to 0.$$

This concludes the proof that $\nabla^{(\theta)}_{\mathcal{D}_m} u_m \to \nabla \overline{u}$ strongly in $L^2(\Omega \times (0,T))^d$ as $m \to \infty$. ∎

Remark 5.8 (About the discrete IBP formula (D.17)*)*
The usage in Step 2 of the "ν"-discrete integration by parts formula (D.17) is non-standard. A usual way of proceeding, see, *e.g.*, [76] or the proof of Theorem 5.19, is to analyse in this proof the convergence of $\Pi^{(1)}_{\mathcal{D}_m} u_m$ towards $\overline{u}$. Using this analysis, the test function $v^{(n+1)}$, instead of $v^{(n+(1-\theta))}$, can be used in Step 2, and the more standard discrete integration-by-parts formula (D.15) can then be applied.
Thanks to (D.17), we can however fully analyse in the proof of Theorem 5.4 the convergence of $\Pi^{(\theta)}_{\mathcal{D}_m} u_m$ without having to analyse at the same time $\Pi^{(1)}_{\mathcal{D}_m} u_m$, which is a less natural reconstruction for θ-schemes.

5.2 Non-conservative problems

5.2.1 The continuous problem

We focus in this section on the approximation of some non-linear problems under the following non-conservative form:

$$\begin{aligned}\nu(\boldsymbol{x},t,u(\boldsymbol{x},t),\nabla u(\boldsymbol{x},t))\partial_t u(\boldsymbol{x},t) - \operatorname{div}(\mu(|\nabla u(\boldsymbol{x},t)|)\nabla u(\boldsymbol{x},t)) \\ = f(\boldsymbol{x},t), \text{ for a.e. } (\boldsymbol{x},t) \in \Omega \times (0,T)\end{aligned} \tag{5.39a}$$

with the initial condition

$$u(\boldsymbol{x},0) = u_{\text{ini}}(\boldsymbol{x}), \text{ for a.e. } \boldsymbol{x} \in \Omega, \tag{5.39b}$$

and boundary conditions

$$u(\boldsymbol{x},t) = 0, \text{ for a.e. } (\boldsymbol{x},t) \in \partial\Omega \times (0,T). \tag{5.39c}$$

The hypotheses are as follows:

- Ω is an open bounded connected subset of $\mathbb{R}^d$ ($d \in \mathbb{N}^\star$) and $T > 0$, (5.40a)
- $u_{\text{ini}} \in H^1_0(\Omega)$ (5.40b)
- $f \in L^2(\Omega \times (0,T))$, (5.40c)
- $\nu : \Omega \times (0,T) \times \mathbb{R} \times \mathbb{R}^d \to \mathbb{R}$ is a Caratheodory function and there exists $\nu_{\max} \geq \nu_{\min} > 0$ such that $\nu(\boldsymbol{x},t,s,\xi) \in [\nu_{\min},\nu_{\max}]$ for a.e. $\boldsymbol{x},t$ and for all s,ξ, (5.40d)

(here, Caratheodory means that, for all $(s,\xi) \in \mathbb{R} \times \mathbb{R}^d$, the function $(\boldsymbol{x},t) \to \nu(\boldsymbol{x},t,s,\xi)$ is measurable and, for a.e. $(\boldsymbol{x},t) \in \Omega \times (0,T)$, the function $(s,\xi) \to \nu(\boldsymbol{x},t,s,\xi)$ it is continuous)

- $\mu : \mathbb{R}^+ \to \mathbb{R}$ is Lipschitz-continuous, non-increasing, and there exists $\mu_{\max} \geq \mu_{\min} > 0$ and $\alpha > 0$ such that $\mu(s) \in [\mu_{\min}, \mu_{\max}]$ and $(s\mu(s))' \geq \alpha$ for all $s \in \mathbb{R}^+$. (5.40e)

One specific choice of μ and ν is of particular interest. For given real numbers $0 < a \leq b$, using the functions

$$\mu(s) = \max\left(\frac{1}{\sqrt{s^2+a^2}}, \frac{1}{b}\right), \ \forall s \in \mathbb{R}^+,$$
$$\nu(\boldsymbol{x}, t, z, \xi) = \mu(|\xi|), \ \forall (\boldsymbol{x}, t) \in \Omega \times (0,T), \ z \in \mathbb{R}, \ \forall \xi \in \mathbb{R}^d$$

in (5.39a) lead to the regularised level set equation [89]. These functions satisfy (5.40d)–(5.40e) with $\alpha = a^2/b^3$.

Let us now give the precise mathematical meaning of a solution to Problem (5.39) under Hypotheses (5.40).

Definition 5.9 (Weak solution of (5.39)). *Under Hypotheses (5.40), we say that u is a weak solution of (5.39) if*

1. $u \in L^2(0,T; H_0^1(\Omega))$ *and* $\partial_t u \in L^2(\Omega \times (0,T))$ *(which implies* $u \in C([0,T]; L^2(\Omega))$*),*
2. $u(\cdot, 0) = u_{\text{ini}}$,
3. *the following holds*

$$\int_0^T \int_\Omega \nu(\boldsymbol{x}, t, u(\boldsymbol{x}, t), \nabla u(\boldsymbol{x}, t)) \partial_t u(\boldsymbol{x}, t) v(\boldsymbol{x}, t) \mathrm{d}\boldsymbol{x}\mathrm{d}t$$
$$+ \int_0^T \int_\Omega \mu(|\nabla u(\boldsymbol{x}, t)|) \nabla u(\boldsymbol{x}, t) \cdot \nabla v(\boldsymbol{x}, t) \mathrm{d}\boldsymbol{x}\mathrm{d}t$$
$$= \int_0^T \int_\Omega f(\boldsymbol{x}, t) v(\boldsymbol{x}, t) \mathrm{d}\boldsymbol{x}\mathrm{d}t, \quad \forall v \in L^2(0,T; H_0^1(\Omega)). \quad (5.41)$$

The third item shows that a weak solution to (5.39) satisfies (5.39a) in the sense of distributions. In particular, for such a solution, $\operatorname{div}(\mu(|\nabla u|)\nabla u) \in L^2(\Omega \times (0,T))$.

Our aim is to use the GDM to construct gradient schemes for (5.41), and to prove their convergence to a weak solution of (5.39). As usual for non-linear models, convergence proofs start with *a priori* estimates. Let us formally show the kind of estimates that can be obtained on (5.39).
Defining F by

$$\forall s \in \mathbb{R}_+, \ F(s) = \int_0^s z\mu(z)\mathrm{d}z \in \left[\mu_{\min}\frac{s^2}{2}, \mu_{\max}\frac{s^2}{2}\right], \quad (5.42)$$

any sufficiently regular function u satisfies

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_\Omega F(|\nabla u(\boldsymbol{x},t)|)\mathrm{d}\boldsymbol{x} = \int_\Omega \mu(|\nabla u(\boldsymbol{x},t)|)\nabla u(\boldsymbol{x},t)\cdot\nabla\partial_t u(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t. \quad (5.43)$$

Therefore, assuming that u is solution of (5.39a) with $f = 0$ (for the sake of simplicity of this brief presentation) and taking $v = \partial_t u$ in (5.41), we see that

$$\int_0^T\int_\Omega \nu(u,\nabla u)\partial_t u(\boldsymbol{x},t)^2\mathrm{d}\boldsymbol{x}\mathrm{d}t + \int_\Omega F(|\nabla u(\boldsymbol{x},t)|)\mathrm{d}\boldsymbol{x} = \int_\Omega F(|\nabla u_{\mathrm{ini}}(\boldsymbol{x})|)\mathrm{d}\boldsymbol{x}. \quad (5.44)$$

The discrete equivalent of this essential estimate is established in Lemma 5.10 for the fully-implicit scheme (using that $s \mapsto s\mu(s)$ is strictly increasing), and in Lemma 5.14 for the semi-implicit scheme (using that μ is decreasing). For both schemes, the hypothesis that $s \mapsto s\mu(s)$ is instrumental to proving that the reconstructed gradients converge strongly.

5.2.2 Fully implicit scheme

Let $\mathcal{D}_T = (X_{\mathcal{D},0}, \Pi_{\mathcal{D}}, \nabla_{\mathcal{D}}, \mathcal{I}_{\mathcal{D}}, (t^{(n)})_{n=0,\dots,N})$ be a space–time GD, for homogeneous Dirichlet boundary conditions, in the sense of Definition 4.1 with $p = 2$ and $\theta = 1$. Using a fully implicit time-stepping, the GDM applied to Problem (5.41) leads to the following GS: find a family $u = (u^{(n)})_{n=0,\dots,N} \in X_{\mathcal{D},0}^{N+1}$ such that

$$\begin{cases} u^{(0)} = \mathcal{I}_{\mathcal{D}} u_{\mathrm{ini}} \text{ and, for } n = 0,\dots,N-1,\ u^{(n+1)} \text{ satisfies} \\ \displaystyle\int_{t^{(n)}}^{t^{(n+1)}}\int_\Omega \nu(\boldsymbol{x},t,\Pi_{\mathcal{D}}u^{(n+1)},\nabla_{\mathcal{D}}u^{(n+1)})\delta_{\mathcal{D}}^{(n+\frac{1}{2})}u(\boldsymbol{x})\Pi_{\mathcal{D}}v(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\mathrm{d}t \\ \quad \displaystyle+\delta t^{(n+\frac{1}{2})}\int_\Omega \mu(|\nabla_{\mathcal{D}}u^{(n+1)}(\boldsymbol{x})|)\nabla_{\mathcal{D}}u^{(n+1)}(\boldsymbol{x})\cdot\nabla_{\mathcal{D}}v(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \\ \quad\quad \displaystyle= \int_{t^{(n)}}^{t^{(n+1)}}\int_\Omega f(\boldsymbol{x},t)\Pi_{\mathcal{D}}v(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\mathrm{d}t\,, \quad \forall v \in X_{\mathcal{D},0}. \end{cases} \quad (5.45)$$

We recall the notations (4.2) and (4.4), and that $\theta = 1$ here. The operators $\Pi_{\mathcal{D}}^{(1)}$ and $\nabla_{\mathcal{D}}^{(1)}$ will therefore be our natural space–time function and gradient reconstructions.

Estimates and existence of a solution to the fully implicit scheme

Lemma 5.10 ($L^2(\Omega\times(0,T))$ estimate on $\delta_{\mathcal{D}}u$ and $L^\infty(0,T;X_{\mathcal{D},0})$ estimate on u, fully implicit scheme). *Under Hypotheses (5.40), let $\mathcal{D}_T$ be a space–time GD for homogeneous Dirichlet boundary conditions, in the sense of Definition 4.1. Then, for any solution u to the GS (5.45) and for all $m = 1,\dots,N$,*

$$\nu_{\min} \int_0^{t^{(m)}} \int_\Omega \delta_{\mathcal{D}} u(\boldsymbol{x},t)^2 \mathrm{d}\boldsymbol{x}\mathrm{d}t + \mu_{\min} \left\| \nabla_{\mathcal{D}} u^{(m)} \right\|^2_{L^2(\Omega)^d} \\ \leq \mu_{\max} \left\| \nabla_{\mathcal{D}} \mathcal{I}_{\mathcal{D}} u_{\mathrm{ini}} \right\|^2_{L^2(\Omega)^d} + \frac{1}{\nu_{\min}} \left\| f \right\|^2_{L^2(\Omega\times(0,T))}. \quad (5.46)$$

As a consequence, there exists at least one solution u to the GS (5.45).

Proof. Setting $v = \frac{u^{(n+1)}-u^{(n)}}{\delta t^{(n+\frac{1}{2})}}$ in (5.45) and summing over $n = 0, \ldots, m-1$ leads to

$$\nu_{\min} \int_0^{t^{(m)}} \int_\Omega \delta_{\mathcal{D}} u(\boldsymbol{x},t)^2 \mathrm{d}\boldsymbol{x}\mathrm{d}t \\ + \sum_{n=0}^{m-1} \int_\Omega \mu(|\nabla_{\mathcal{D}} u^{(n+1)}(\boldsymbol{x})|) \nabla_{\mathcal{D}} u^{(n+1)}(\boldsymbol{x}) \cdot \left[\nabla_{\mathcal{D}} u^{(n+1)}(\boldsymbol{x}) - \nabla_{\mathcal{D}} u^{(n)}(\boldsymbol{x}) \right] \mathrm{d}\boldsymbol{x} \\ \leq \int_0^{t^{(m)}} \int_\Omega f(\boldsymbol{x},t) \delta_{\mathcal{D}} u(\boldsymbol{x},t) \mathrm{d}\boldsymbol{x}\mathrm{d}t. \quad (5.47)$$

Hypothesis (5.40e) implies the convexity of F, defined by (5.42), and thus

$$\forall c_1, c_2 \in \mathbb{R}^+ \,,\ F(c_2) - F(c_1) = \int_{c_1}^{c_2} z\mu(z)\mathrm{d}z \leq c_2 \mu(c_2)(c_2 - c_1).$$

This gives in particular

$$F(|\nabla_{\mathcal{D}} u^{(n+1)}(\boldsymbol{x})|) - F(|\nabla_{\mathcal{D}} u^{(n)}(\boldsymbol{x})|) \\ \leq \mu(|\nabla_{\mathcal{D}} u^{(n+1)}(\boldsymbol{x})|) |\nabla_{\mathcal{D}} u^{(n+1)}(\boldsymbol{x})| \left[|\nabla_{\mathcal{D}} u^{(n+1)}(\boldsymbol{x})| - |\nabla_{\mathcal{D}} u^{(n)}(\boldsymbol{x})| \right]. \quad (5.48)$$

The Cauchy–Schwarz inequality implies

$$|\nabla_{\mathcal{D}} u^{(n+1)}(\boldsymbol{x})| \left[|\nabla_{\mathcal{D}} u^{(n+1)}(\boldsymbol{x})| - |\nabla_{\mathcal{D}} u^{(n)}(\boldsymbol{x})| \right] \\ = \nabla_{\mathcal{D}} u^{(n+1)}(\boldsymbol{x}) \cdot \nabla_{\mathcal{D}} u^{(n+1)}(\boldsymbol{x}) - |\nabla_{\mathcal{D}} u^{(n+1)}(\boldsymbol{x})| |\nabla_{\mathcal{D}} u^{(n)}(\boldsymbol{x})| \\ \leq \nabla_{\mathcal{D}} u^{(n+1)}(\boldsymbol{x}) \cdot \left[\nabla_{\mathcal{D}} u^{(n+1)}(\boldsymbol{x}) - \nabla_{\mathcal{D}} u^{(n)}(\boldsymbol{x}) \right]. \quad (5.49)$$

Combining (5.48) and (5.49) and plugging the result into (5.47) yields

$$\nu_{\min} \int_0^{t^{(m)}} \int_\Omega \delta_{\mathcal{D}} u(\boldsymbol{x},t)^2 \mathrm{d}\boldsymbol{x}\mathrm{d}t \\ + \sum_{n=0}^{m-1} \int_\Omega \left[F(|\nabla_{\mathcal{D}} u^{(n+1)}(\boldsymbol{x})|) - F(|\nabla_{\mathcal{D}} u^{(n)}(\boldsymbol{x})|) \right] \mathrm{d}\boldsymbol{x} \\ \leq \int_0^{t^{(m)}} \int_\Omega f(\boldsymbol{x},t) \delta_{\mathcal{D}} u(\boldsymbol{x},t) \mathrm{d}\boldsymbol{x}\mathrm{d}t. \quad (5.50)$$

The sum in the left-hand side is telescopic and reduces to

$$\int_\Omega \left[F(|\nabla_\mathcal{D} u^{(m)}(\boldsymbol{x})|) - F(|\nabla_\mathcal{D} u^{(0)}(\boldsymbol{x})|)\right] \mathrm{d}\boldsymbol{x}.$$

The right-hand side of (5.50) can be estimated by means of the Cauchy–Schwarz inequality and the Young inequality (D.9). Since the range of F is in $[\mu_{\min} s^2/2, \mu_{\max} s^2/2]$, this gives

$$\begin{aligned}
&\nu_{\min} \int_0^{t^{(m)}} \int_\Omega \delta_\mathcal{D} u(\boldsymbol{x},t)^2 \mathrm{d}\boldsymbol{x}\mathrm{d}t + \frac{\mu_{\min}}{2} \int_\Omega |\nabla_\mathcal{D} u^{(m)}(\boldsymbol{x})|^2 \mathrm{d}\boldsymbol{x} \\
&\quad \le \frac{\mu_{\max}}{2} \int_\Omega |\nabla_\mathcal{D} u^{(0)}(\boldsymbol{x})|^2 \mathrm{d}\boldsymbol{x} + \|f\|_{L^2(\Omega\times(0,T))} \|\delta_\mathcal{D} u\|_{L^2(\Omega\times(0,t^{(m)}))} \\
&\quad \le \frac{\mu_{\max}}{2} \int_\Omega |\nabla_\mathcal{D} \mathcal{I}_\mathcal{D} u_{\mathrm{ini}}(\boldsymbol{x})|^2 \mathrm{d}\boldsymbol{x} + \frac{1}{2\nu_{\min}} \|f\|^2_{L^2(\Omega\times(0,T))} \\
&\qquad + \frac{\nu_{\min}}{2} \|\delta_\mathcal{D} u\|^2_{L^2(\Omega\times(0,t^{(m)}))}.
\end{aligned}$$

Moving the last term to the left-hand side yields Estimate (5.46).
To prove the existence of at least one solution to the GS, we create an homotopy between the model (5.39) and a linear PDE. By induction, it suffices to show that, for a given $u^{(n)} \in X_{\mathcal{D},0}$, there exists $u^{(n+1)} \in X_{\mathcal{D},0}$ satisfying the integral relation in (5.45). For $\lambda \in [0,1]$, define μ_λ and ν_λ by

$$\begin{aligned}
\mu_\lambda(s) &= \mu_{\max}(1-\lambda) + \lambda\mu(s)\,, \text{ and} \\
\nu_\lambda(\boldsymbol{x},t,s,\xi) &= \nu_{\min}(1-\lambda) + \lambda\nu(\boldsymbol{x},t,s,\xi).
\end{aligned}$$

Let $(v_i)_{i=1,\dots,M}$ be a basis of $X_{\mathcal{D},0}$ and define $\Phi : X_{\mathcal{D},0} \times [0,1] \to X_{\mathcal{D},0}$ by its components $(\Phi(w,\lambda)_i)_{i=1,\dots,M}$ on $(v_i)_{i=1,\dots,M}$:

$$\begin{aligned}
\Phi(w,\lambda)_i = &\int_{t^{(n)}}^{t^{(n+1)}} \int_\Omega \nu_\lambda(\boldsymbol{x},t,\Pi_\mathcal{D} w(\boldsymbol{x}), \nabla_\mathcal{D} w(\boldsymbol{x})) \\
&\qquad\qquad \times \frac{\Pi_\mathcal{D} w(\boldsymbol{x}) - \Pi_\mathcal{D} u^{(n)}(\boldsymbol{x})}{\delta t^{(n+\frac{1}{2})}} \Pi_\mathcal{D} v_i(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}\mathrm{d}t \\
&+ \delta t^{(n+\frac{1}{2})} \int_\Omega \mu_\lambda(|\nabla_\mathcal{D} w(\boldsymbol{x})|) \nabla_\mathcal{D} w(\boldsymbol{x}) \cdot \nabla_\mathcal{D} v_i(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \\
&- \int_{t^{(n)}}^{t^{(n+1)}} \int_\Omega f(\boldsymbol{x},t) \Pi_\mathcal{D} v_i(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}\mathrm{d}t.
\end{aligned}$$

Then $u^{(n+1)}$ satisfies the integral equation in (5.45) if and only if $\Phi(u^{(n+1)},1) = 0$.
The mapping Φ is clearly continuous. If $\Phi(w,\lambda) = 0$ then, since μ_λ (resp. ν_λ) has its range in $[\mu_{\min}, \mu_{\max}]$ (resp. $[\nu_{\min}, \nu_{\max}]$), similar estimates to the ones established above give a bound on $\|\nabla_\mathcal{D} w\|_{L^2(\Omega)^d} = \|w\|_\mathcal{D}$ that does

not depend on $\lambda \in [0,1]$. Finally, for $\lambda = 0$, $\Phi(\cdot,0)$ is affine and therefore invertible since, by the previous bound, its kernel is bounded (and thus necessarily reduced to a single point). As a consequence, for some R large enough, $\Phi(\cdot,0) = 0$ has a solution in the ball of radius R in $X_{\mathcal{D},0}$.
A topological degree argument (see Theorem D.1 page 462) can therefore be applied and show that $\Phi(\cdot,1) = 0$ has at least one solution, *i.e.* that there exists $u^{(n+1)}$ solution to the integral equation in (5.45). ■

Convergence of the fully implicit scheme

For $u \in X_{\mathcal{D},0}^{N+1}$, define $w_{\mathcal{D}}$ and $G_{\mathcal{D}}$ by, for a.e. $(\boldsymbol{x},t) \in \Omega \times (0,T)$,

$$w_{\mathcal{D}}(\boldsymbol{x},t) = f(\boldsymbol{x},t) - \nu\left(\boldsymbol{x},t,\Pi_{\mathcal{D}}^{(1)}u(\boldsymbol{x},t),\nabla_{\mathcal{D}}^{(1)}u(\boldsymbol{x},t)\right)\delta_{\mathcal{D}}u(\boldsymbol{x},t), \tag{5.51}$$

$$G_{\mathcal{D}}(\boldsymbol{x},t) = \mu\left(|\nabla_{\mathcal{D}}^{(1)}u(\boldsymbol{x},t)|\right)\nabla_{\mathcal{D}}^{(1)}u(\boldsymbol{x},t). \tag{5.52}$$

With these definitions, (5.45) can be recast as

$$\begin{aligned} &u \in X_{\mathcal{D},0}^{N+1} \text{ and, for all } v \in X_{\mathcal{D},0}^{N+1}, \\ &\int_0^T\int_\Omega G_{\mathcal{D}}(\boldsymbol{x},t)\cdot\nabla_{\mathcal{D}}^{(1)}v(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t \\ &\qquad = \int_0^T\int_\Omega w_{\mathcal{D}}(\boldsymbol{x},t)\Pi_{\mathcal{D}}^{(1)}v(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t. \end{aligned} \tag{5.53}$$

The following lemma is an initial step towards establishing the convergence of the fully implicit GS for (5.39).

Lemma 5.11 (A convergence property of the fully implicit scheme). *Under Hypotheses (5.40), let $((\mathcal{D}_T)_m)_{m\in\mathbb{N}}$ be a sequence of space–time GDs for homogeneous Dirichlet boundary conditions (with $p = 2$). Assume that the sequence $((\mathcal{D}_T)_m)_{m\in\mathbb{N}}$ is space–time-consistent, limit-conforming and compact in the sense of Definitions 4.3 and 4.6. Also assume that $(\nabla_{\mathcal{D}_m}\mathcal{I}_{\mathcal{D}_m}u_{\mathrm{ini}})_{m\in\mathbb{N}}$ is bounded in $L^2(\Omega)^d$.*
For any $m \in \mathbb{N}$, take u_m a solution to the GS (5.45) and define $w_{\mathcal{D}_m}$ and $G_{\mathcal{D}_m}$ from u_m by (5.51)–(5.52). Then there exist functions

$$\begin{aligned} &\overline{u} \in L^\infty(0,T;H_0^1(\Omega)) \cap C([0,T];L^2(\Omega)) \text{ with } \partial_t\overline{u} \in L^2(\Omega\times(0,T)) \\ &\qquad \text{and } u(\cdot,0) = u_{\mathrm{ini}}, \\ &\overline{G} \in L^2(\Omega\times(0,T))^d, \text{ and} \\ &\overline{w} \in L^2(\Omega\times(0,T)) \end{aligned}$$

such that, along a subsequence as $m \to \infty$,

- $\sup_{t\in[0,T]}\|\Pi_{\mathcal{D}_m}^{(1)}u_m(t) - \overline{u}(t)\|_{L^2(\Omega)} \to 0$,

- $\nabla^{(1)}_{\mathcal{D}_m} u_m$ *converges weakly in* $L^2(\Omega \times (0,T))^d$ *to* $\nabla \overline{u}$,
- $\delta_{\mathcal{D}_m} u_m$ *converges weakly in* $L^2(\Omega \times (0,T))$ *to* $\partial_t \overline{u}$,
- $G_{\mathcal{D}_m}$ *converges weakly to* $\overline{G}$ *in* $L^2(\Omega \times (0,T))^d$,
- $w_{\mathcal{D}_m}$ *converges weakly to* $\overline{w}$ *in* $L^2(\Omega \times (0,T))$,
- *it holds*

$$\int_0^T \int_\Omega G_{\mathcal{D}_m}(\boldsymbol{x},t) \cdot \nabla^{(1)}_{\mathcal{D}_m} u_m(\boldsymbol{x},t) \mathrm{d}\boldsymbol{x}\mathrm{d}t \longrightarrow \int_0^T \int_\Omega \overline{G}(\boldsymbol{x},t) \cdot \nabla \overline{u}(\boldsymbol{x},t) \mathrm{d}\boldsymbol{x}\mathrm{d}t. \tag{5.54}$$

Remark 5.12. At this stage, we do not identify $\overline{G}$ and $\overline{w}$, respectively, with $\mu(|\nabla \overline{u}|)\nabla \overline{u}$ and $\nu(\overline{u}, \nabla \overline{u})$. This is done later, in the proof of Theorem 5.13, by using (5.54).

Proof. Owing to (5.46), $(G_{\mathcal{D}_m})_{m\in\mathbb{N}}$ is bounded in $L^\infty(0,T;L^2(\Omega)^d)$ and $(w_{\mathcal{D}_m})_{m\in\mathbb{N}}$ is bounded in $L^2(\Omega \times (0,T))$. Hence, there exists $\overline{G} \in L^2(\Omega \times (0,T))^d$ and $\overline{w} \in L^2(\Omega \times (0,T))$ such that, up to a subsequence as $m \to \infty$, $G_{\mathcal{D}_m} \to \overline{G}$ weakly in $L^2(\Omega \times (0,T))^d$ and $w_{\mathcal{D}_m} \to \overline{w}$ weakly in $L^2(\Omega \times (0,T))$. By Lemma 5.10, the sequence $((u_m)_1)_{m\in\mathbb{N}}$ (see notation (4.2) with $\theta = 1$) is bounded in $L^\infty(0,T;X_{\mathcal{D}_m,0})$ and the sequence $(\delta_{\mathcal{D}_m} u_m)_{m\in\mathbb{N}}$ is bounded in $L^2(0,T;L^2(\Omega))$. Theorem 4.18 thus provides $\overline{u} \in L^\infty(0,T;H^1_0(\Omega)) \cap C([0,T];L^2(\Omega))$ such that $\partial_t \overline{u} \in L^2(\Omega \times (0,T))$ and, up to a subsequence as $m \to \infty$, $\sup_{t\in[0,T]} \|\Pi^{(1)}_{\mathcal{D}_m} u_m(t) - \overline{u}(t)\|_{L^2(\Omega)} \to 0$ and $\delta_{\mathcal{D}_m} u_m \to \partial_t \overline{u}$ weakly in $L^2(\Omega \times (0,T))$. The weak convergence of $\nabla^{(1)}_{\mathcal{D}_m} u_m$ to $\nabla \overline{u}$ is a consequence of Lemma 4.8.
The definition (4.3) gives $\Pi^{(1)}_{\mathcal{D}_m} u_m(0) = \Pi_{\mathcal{D}_m} u^{(0)}_m = \Pi_{\mathcal{D}_m} \mathcal{I}_{\mathcal{D}_m} u_{\mathrm{ini}}$. The space–time-consistency of $((\mathcal{D}_T)_m)_{m\in\mathbb{N}}$ then yields $\Pi^{(1)}_{\mathcal{D}_m} u_m(0) \to u_{\mathrm{ini}}$ in $L^2(\Omega)$ as $m \to \infty$. By the uniform convergence of $(\Pi^{(1)}_{\mathcal{D}_m} u_m)_{m\in\mathbb{N}}$ to $\overline{u}$, we infer that $\overline{u}(\cdot,0) = u_{\mathrm{ini}}$.
We now aim to prove (5.54). Since $\overline{u} \in L^2(0,T;H^1_0(\Omega))$ we can take $(v_m)_{m\in\mathbb{N}}$ given by Lemma 4.10 for $\overline{v} = \overline{u}$. Using v_m as a test function in (5.53) with $\mathcal{D}_T = (\mathcal{D}_T)_m$ and passing to the limit yields

$$\int_0^T \int_\Omega \overline{G}(\boldsymbol{x},t) \cdot \nabla \overline{u}(\boldsymbol{x},t) \mathrm{d}\boldsymbol{x}\mathrm{d}t = \int_0^T \int_\Omega \overline{w}(\boldsymbol{x},t) \overline{u}(\boldsymbol{x},t) \mathrm{d}\boldsymbol{x}\mathrm{d}t. \tag{5.55}$$

Putting $v = u_m$ in (5.53), the weak-strong convergence lemma (Lemma D.8 page 464) enables us to pass to the limit in the right-hand side, since $(w_{\mathcal{D}_m})_{m\in\mathbb{N}}$ converges weakly in $L^2(\Omega \times (0,T))$ and $(\Pi^{(1)}_{\mathcal{D}_m} u_m)_{m\in\mathbb{N}}$ converges strongly in $L^2(\Omega \times (0,T))$. Owing to (5.55), this gives

$$\lim_{m\to\infty} \int_0^T \int_\Omega G_{\mathcal{D}_m}(\boldsymbol{x},t) \cdot \nabla^{(1)}_{\mathcal{D}_m} u_m(\boldsymbol{x},t) \mathrm{d}\boldsymbol{x}\mathrm{d}t = \int_0^T \int_\Omega \overline{w}(\boldsymbol{x},t) \overline{u}(\boldsymbol{x},t) \mathrm{d}\boldsymbol{x}\mathrm{d}t$$

$$= \int_0^T \int_\Omega \overline{G}(\boldsymbol{x},t) \cdot \nabla \overline{u}(\boldsymbol{x},t) \mathrm{d}\boldsymbol{x}\mathrm{d}t$$

and the proof of (5.54) is complete. ■

We now state and prove the convergence of the fully implicit GS for (5.39).

Theorem 5.13.
Assume (5.40) and let $((\mathcal{D}_T)_m)_{m\in\mathbb{N}}$ be a sequence of space–time GDs for homogeneous Dirichlet boundary conditions (with $p = 2$). Assume that the sequence $((\mathcal{D}_T)_m)_{m\in\mathbb{N}}$ is space–time-consistent, limit-conforming and compact in the sense of Definitions 4.3 and 4.6.
We also suppose that $(\nabla_{\mathcal{D}_m}\mathcal{I}_{\mathcal{D}_m} u_{\mathrm{ini}})_{m\in\mathbb{N}}$ is bounded in $L^2(\Omega)^d$ and, for any $m \in \mathbb{N}$, we let u_m be a solution to the GS (5.45).
Then there exists a weak solution $\overline{u}$ of (5.39) in the sense of Definition 5.9 such that, up to a subsequence as $m \to \infty$,

- $\sup_{t\in[0,T]} \|\Pi^{(1)}_{\mathcal{D}_m} u_m(t) - \overline{u}(t)\|_{L^2(\Omega)} \to 0$, *and*
- $\nabla^{(1)}_{\mathcal{D}_m} u_m \to \nabla\overline{u}$ *in* $L^2(\Omega \times (0,T))^d$.

Proof.
Let $\overline{u}$, $\overline{G}$ and $\overline{w}$ be given by Lemma 5.11. Then, $\sup_{t\in[0,T]} \|\Pi^{(1)}_{\mathcal{D}_m} u_m(t) - \overline{u}(t)\|_{L^2(\Omega)} \to 0$ along a subsequence (not explicitly indicated below).

Step 1: a strong monotonicity property.
We aim to prove here that, for all $V, W \in L^2(\Omega \times (0,T))^d$,

$$\int_0^T\int_\Omega [\mu(|W|)W - \mu(|V|)V] \cdot [W - V] \,\mathrm{d}\boldsymbol{x}\mathrm{d}t \geq \alpha \left\||W| - |V|\right\|^2_{L^2(\Omega\times(0,T))^d}. \quad (5.56)$$

Use first the Cauchy–Schwarz inequality for the dot product of $\mathbb{R}^d$ to get

$$\int_0^T \int_\Omega \mu(|W|)W \cdot V \mathrm{d}\boldsymbol{x}\mathrm{d}t \leq \int_0^T \int_\Omega \mu(|W|)|W|\,|V| \mathrm{d}\boldsymbol{x}\mathrm{d}t.$$

Writing the same properties with W and V swapped leads to

$$\int_0^T \int_\Omega (\mu(|W|)W - \mu(|V|)V) \cdot (W - V) \,\mathrm{d}\boldsymbol{x}\mathrm{d}t \geq \int_0^T \int_\Omega (\mu(|W|)|W| - \mu(|V|)|V|)\,(|W| - |V|) \,\mathrm{d}\boldsymbol{x}\mathrm{d}t.$$

By Property (5.40e) on μ, (5.56) follows.

Step 2: Proof that $\overline{G} = \mu(|\nabla\overline{u}|)\nabla\overline{u}$.
We use Minty's trick. For $V \in L^2(\Omega \times (0,T))^d$, set

$$T_m(V) = \int_0^T \int_\Omega \left[\mu(|\nabla^{(1)}_{\mathcal{D}_m} u_m|)\nabla^{(1)}_{\mathcal{D}_m} u_m - \mu(|V|)V\right] \cdot \left[\nabla^{(1)}_{\mathcal{D}_m} u_m - V\right] \mathrm{d}\boldsymbol{x}\mathrm{d}t.$$

Recall that $\mu(|\nabla^{(1)}_{\mathcal{D}_m} u_m|)\nabla^{(1)}_{\mathcal{D}_m} u_m = G_{\mathcal{D}_m}$. Together with (5.54), the weak convergences of $G_{\mathcal{D}_m}$ and $\nabla^{(1)}_{\mathcal{D}_m} u_m$ therefore yield

$$\lim_{m\to\infty} T_m(V) = \int_0^T \int_\Omega \left[\overline{G} - \mu(|V|)V\right] \cdot [\nabla\overline{u} - V] \,\mathrm{d}\boldsymbol{x}\mathrm{d}t. \tag{5.57}$$

By (5.56), $T_m(V) \geq 0$ and thus

$$\int_0^T \int_\Omega \left[\overline{G} - \mu(|V|)V\right] \cdot [\nabla\overline{u} - V] \,\mathrm{d}\boldsymbol{x}\mathrm{d}t \geq 0.$$

Take $W \in L^2(\Omega \times (0,T))^d$ and set $V = \nabla\overline{u} + \lambda W$ for $\lambda \in \mathbb{R}$. This gives

$$\lambda \int_0^T \int_\Omega \left[\overline{G} - \mu(|\nabla\overline{u} + \lambda W|)(\nabla\overline{u} + \lambda W)\right] \cdot W \mathrm{d}\boldsymbol{x}\mathrm{d}t \geq 0.$$

Since λ is any real number, this shows that the integral term is equal to zero. The dominated convergence theorem justifies letting $\lambda \to 0$ in this term, which shows that

$$\int_0^T \int_\Omega \left[\overline{G} - \mu(|\nabla\overline{u}|)\nabla\overline{u}\right] \cdot W \mathrm{d}\boldsymbol{x}\mathrm{d}t = 0.$$

Taking $W = \overline{G} - \mu(|\nabla\overline{u}|)\nabla\overline{u}$ yields

$$\overline{G} = \mu(|\nabla\overline{u}|)\nabla\overline{u} \quad \text{a.e. on } \Omega \times (0,T). \tag{5.58}$$

Step 3: strong convergence of $\nabla^{(1)}_{\mathcal{D}_m} u_m$, and proof that $\overline{u}$ is a solution to (5.39).
Making $W = \nabla^{(1)}_{\mathcal{D}_m} u_m$ and $V = \nabla\overline{u}$ in (5.56) gives

$$\left\| |\nabla^{(1)}_{\mathcal{D}_m} u_m| - |\nabla\overline{u}| \right\|^2_{L^2(\Omega\times(0,T))^d} \leq \frac{1}{\alpha} T_m(\nabla\overline{u}).$$

By (5.57), $\lim_{m\to\infty} T_m(\nabla\overline{u}) = 0$ and thus $|\nabla^{(1)}_{\mathcal{D}_m} u_m| \to |\nabla\overline{u}|$ in $L^2(\Omega \times (0,T))$ as $m \to \infty$. This entails the convergences of the L^2 norms of these functions, that is

$$\left\| \nabla^{(1)}_{\mathcal{D}_m} u_m \right\|_{L^2(\Omega\times(0,T))^d} \to \|\nabla\overline{u}\|_{L^2(\Omega\times(0,T))^d} \text{ as } m \to \infty.$$

This latter convergence shows that the weak convergence of $(\nabla^{(1)}_{\mathcal{D}_m} u_m)_{m\in\mathbb{N}}$ to $\nabla\overline{u}$ in $L^2(\Omega \times (0,T))$ is actually strong.

By a form of non-linear strong convergence property similar to Lemma D.9 page 465, the strong convergences of $(\Pi^{(1)}_{\mathcal{D}_m} u_m)_{m\in\mathbb{N}}$ and $(\nabla^{(1)}_{\mathcal{D}_m} u_m)_{m\in\mathbb{N}}$ show that, as $m \to \infty$,

$$\nu(\cdot,\cdot,\Pi^{(1)}_{\mathcal{D}_m} u_m, \nabla^{(1)}_{\mathcal{D}_m} u_m) \to \nu(\cdot,\cdot,\overline{u},\nabla\overline{u}) \text{ strongly in } L^2(\Omega\times(0,T)).$$

The weak convergence of $\delta_{\mathcal{D}_m} u_m$ towards $\partial_t\overline{u}$ and the weak-strong convergence property in Lemma D.8 page 464 then enable us to identify the limit of $w_{\mathcal{D}_m}$ (defined by (5.51)):

$$\overline{w} = f - \nu(\cdot,\cdot,\overline{u},\nabla\overline{u})\partial_t\overline{u} \quad \text{ a.e. in } \Omega\times(0,T). \tag{5.59}$$

Let $v \in L^2(0,T;H^1_0(\Omega))$ and take $(v_m)_{m\in\mathbb{N}}$ provided by Lemma 4.10 for v. Write (5.53) for $\mathcal{D}_T = (\mathcal{D}_T)_m$ and $v = v_m$. Passing to the limit $m \to \infty$ in this relation is justified by the weak convergences of $G_{\mathcal{D}_m}$ and $w_{\mathcal{D}_m}$, and the strong convergences of $\Pi^{(1)}_{\mathcal{D}_m} v_m$ and $\nabla^{(1)}_{\mathcal{D}_m} v_m$. This leads to

$$\int_0^T\int_\Omega \overline{G}(\boldsymbol{x},t)\cdot\nabla v(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t = \int_0^T\int_\Omega \overline{w}(\boldsymbol{x},t)v(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t.$$

Then (5.58) and (5.59) show that $\overline{u}$ satisfies (5.41). Since the regularity properties of $\overline{u}$ required in Definition 5.9 are ascertained in Lemma 5.11, the proof that $\overline{u}$ is a solution to (5.39) is complete. ∎

5.2.3 Semi-implicit scheme

Given a space–time gradient discretisation $\mathcal{D}_T$ and using a semi-implicit time-stepping, the GDM applied to (5.41) gives the following GS: seek a family $u = (u^{(n)})_{n=0,\dots,N} \in X^{N+1}_{\mathcal{D}}$ such that

$$\begin{cases} u^{(0)} = \mathcal{I}_{\mathcal{D}} u_{\mathrm{ini}} \text{ and, for } n = 0,\dots,N-1,\ u^{(n+1)} \text{ satisfies} \\ \displaystyle\int_{t^{(n)}}^{t^{(n+1)}}\int_\Omega \nu(\boldsymbol{x},t,\Pi_{\mathcal{D}}u^{(n)},\nabla_{\mathcal{D}}u^{(n)})\delta^{(n+\frac12)}_{\mathcal{D}} u(\boldsymbol{x})\Pi_{\mathcal{D}}v(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\mathrm{d}t \\ \quad \displaystyle+\delta t^{(n+\frac12)}\int_\Omega \mu(|\nabla_{\mathcal{D}}u^{(n)}|)\nabla_{\mathcal{D}}u^{(n+1)}\cdot\nabla_{\mathcal{D}}v(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \\ \quad\quad \displaystyle= \int_{t^{(n)}}^{t^{(n+1)}}\int_\Omega f(\boldsymbol{x},t)\Pi_{\mathcal{D}}v(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\mathrm{d}t\,, \qquad \forall v\in X_{\mathcal{D},0}. \end{cases} \tag{5.60}$$

Quite naturally, the analysis of this semi-implicit implicit scheme uses both $\Pi^{(1)}_{\mathcal{D}}$, $\nabla^{(1)}_{\mathcal{D}}$ and $\Pi^{(0)}_{\mathcal{D}}$, $\nabla^{(0)}_{\mathcal{D}}$. Recall that the definition of these latter operators (see (4.2)):

$$\Pi^{(0)}_{\mathcal{D}} u(\boldsymbol{x},t) = \Pi_{\mathcal{D}}u^{(n)}(\boldsymbol{x}) \text{ and } \nabla^{(0)}_{\mathcal{D}} u(\boldsymbol{x},t) = \nabla_{\mathcal{D}}u^{(n)}(\boldsymbol{x}),$$
$$\text{for a.e. } (\boldsymbol{x},t)\in\Omega\times(t^{(n)},t^{(n+1)}),\ \forall n = 0,\dots,N-1.$$

Estimates and existence of a solution to the semi-implicit scheme

Lemma 5.14 ($L^2(\Omega\times(0,T))$ estimate on $\delta_{\mathcal{D}}u$ and $L^\infty(0,T;X_{\mathcal{D},0})$ estimate on u, semi-implicit scheme.). *Under Hypotheses (5.40), let $\mathcal{D}_T$ be a space–time GD in the sense of Definition 4.1. Then the GS (5.60) has a unique solution u, and it satisfies, for all $m=1,\dots,N$,*

$$\begin{aligned}\nu_{\min}\int_0^{t^{(m)}}\int_\Omega \delta_{\mathcal{D}}u(\boldsymbol{x},t)^2\mathrm{d}\boldsymbol{x}\mathrm{d}t+\mu_{\min}\left\|\nabla_{\mathcal{D}}u^{(m)}\right\|^2_{L^2(\Omega)^d}\\ +\mu_{\min}\sum_{n=0}^{m-1}\int_\Omega\left|\nabla_{\mathcal{D}}u^{(n+1)}(\boldsymbol{x})-\nabla_{\mathcal{D}}u^{(n)}(\boldsymbol{x})\right|^2\mathrm{d}\boldsymbol{x}\\ \le\mu_{\max}\left\|\nabla_{\mathcal{D}}\mathcal{I}_{\mathcal{D}}u_{\mathrm{ini}}\right\|^2_{L^2(\Omega)^d}+\frac{1}{\nu_{\min}}\left\|f\right\|^2_{L^2(\Omega\times(0,T))}.\end{aligned}\tag{5.61}$$

Proof. First notice that, by Hypothesis (5.40e),

$$\forall\xi,\chi\in\mathbb{R}^d\,,\ \int_{|\xi|}^{|\chi|}z\mu(z)\mathrm{d}z+\frac{1}{2}|\chi-\xi|^2\mu(|\xi|)\le\mu(|\xi|)\chi\cdot(\chi-\xi).\tag{5.62}$$

To prove this property, simply remark by developing $|\chi-\xi|^2$ that it simplifies into

$$\forall\xi,\chi\in\mathbb{R}^d\,,\ \frac{1}{2}\mu(|\xi|)(|\chi|^2-|\xi|^2)-\int_{|\xi|}^{|\chi|}z\mu(z)\mathrm{d}z\ge 0.$$

Set, for $a,b\in\mathbb{R}^+$,

$$\Phi(b)=\frac{1}{2}\mu(a)(b^2-a^2)-\int_a^b z\mu(z)\mathrm{d}z.$$

Then $\Phi'(b)=b(\mu(a)-\mu(b))$, whose sign is that of $b-a$ since μ is non-increasing. Hence $\Phi(b)$ is non-increasing for $b\le a$ and non-decreasing for $b\ge a$. Since $\Phi(a)=0$, this shows that $\Phi(b)\ge 0$ for all $b\in\mathbb{R}^+$ and the proof of (5.62) is complete.
Applying this relation to $\xi=\nabla_{\mathcal{D}}u^{(n)}(\boldsymbol{x})$ and $\chi=\nabla_{\mathcal{D}}u^{(n+1)}(\boldsymbol{x})$ and recalling the definition (5.42) of F leads to

$$\begin{aligned}F(|\nabla_{\mathcal{D}}u^{(n+1)}(\boldsymbol{x})|)-F(|\nabla_{\mathcal{D}}u^{(n)}(\boldsymbol{x})|)\\ +\frac{\mu_{\min}}{2}\left|\nabla_{\mathcal{D}}u^{(n+1)}(\boldsymbol{x})-\nabla_{\mathcal{D}}u^{(n)}(\boldsymbol{x})\right|^2\\ \le\mu(|\nabla_{\mathcal{D}}u^{(n)}(\boldsymbol{x})|)\nabla_{\mathcal{D}}u^{(n+1)}(\boldsymbol{x})\cdot\left[\nabla_{\mathcal{D}}u^{(n+1)}(\boldsymbol{x})-\nabla_{\mathcal{D}}u^{(n)}(\boldsymbol{x})\right].\end{aligned}\tag{5.63}$$

Estimate (5.61) is then established as the proof of Lemma 5.10, by plugging $v=\frac{u^{(n+1)}-u^{(n)}}{\delta t^{(n+\frac{1}{2})}}$ in (5.60), summing over $n=0,\dots,m-1$, and using (5.63) in lieu of (5.48). ∎

Convergence of the semi-implicit scheme

If u is the solution to the GS (5.60), let

$$\widetilde{w}_{\mathcal{D}} = f - \nu(\Pi_{\mathcal{D}}^{(0)} u_{\mathcal{D}}, \nabla_{\mathcal{D}}^{(0)} u_{\mathcal{D}}) \delta_{\mathcal{D}} u_{\mathcal{D}}\,, \tag{5.64}$$

$$\widetilde{G}_{\mathcal{D}} = \mu(|\nabla_{\mathcal{D}}^{(0)} u_{\mathcal{D}}|) \nabla_{\mathcal{D}}^{(1)} u_{\mathcal{D}}\,, \tag{5.65}$$

$$\widehat{G}_{\mathcal{D}} = \mu(|\nabla_{\mathcal{D}}^{(0)} u_{\mathcal{D}}|) \nabla_{\mathcal{D}}^{(0)} u_{\mathcal{D}}. \tag{5.66}$$

Note that the GS (5.60) can be recast as:

$$\begin{aligned} &u \in X_{\mathcal{D},0}^{N+1} \text{ and, for all } v \in X_{\mathcal{D},0}^{N+1}\,, \\ &\int_0^T \int_\Omega \widetilde{G}_{\mathcal{D}}(\boldsymbol{x},t) \cdot \nabla_{\mathcal{D}}^{(1)} v(\boldsymbol{x},t) \mathrm{d}\boldsymbol{x}\mathrm{d}t \\ &\qquad = \int_0^T \int_\Omega \widetilde{w}_{\mathcal{D}}(\boldsymbol{x},t) \Pi_{\mathcal{D}}^{(1)} v(\boldsymbol{x},t) \mathrm{d}\boldsymbol{x}\mathrm{d}t. \end{aligned} \tag{5.67}$$

The following lemma is the equivalent, for the semi-implicit scheme, of Lemma 5.11.

Lemma 5.15 (A convergence property of the semi-implicit scheme). *Under Hypotheses (5.40), let $((\mathcal{D}_T)_m)_{m\in\mathbb{N}}$ be a sequence of space–time GDs for homogeneous Dirichlet boundary conditions (with $p = 2$). Assume that this sequence is space–time-consistent, limit-conforming and compact in the sense of Definitions 4.3 and 4.6. Assume also that $(\nabla_{\mathcal{D}_m} \mathcal{I}_{\mathcal{D}_m} u_{\mathrm{ini}})_{m\in\mathbb{N}}$ is bounded in $L^2(\Omega)^d$.*
For $m \in \mathbb{N}$, let u_m be the solution to the GS (5.60), and define $\widetilde{w}_{\mathcal{D}_m}$, $\widetilde{G}_{\mathcal{D}_m}$ and $\widehat{G}_{\mathcal{D}_m}$ from u_m by, respectively, (5.64), (5.65) and (5.66).
Then there exist functions

$$\begin{aligned} &\overline{u} \in L^\infty(0,T; H_0^1(\Omega)) \cap C([0,T]; L^2(\Omega)) \text{ with } \partial_t \overline{u} \in L^2(\Omega \times (0,T)) \\ &\qquad \text{and } u(\cdot,0) = u_{\mathrm{ini}}\,, \\ &\overline{G} \in L^2(\Omega \times (0,T))^d\,, \text{ and} \\ &\overline{w} \in L^2(\Omega \times (0,T)) \end{aligned}$$

such that, along a subsequence as $m \to \infty$,

- $\sup_{t\in[0,T]} \|\Pi_{\mathcal{D}_m}^{(1)} u_m(t) - \overline{u}(t)\|_{L^2(\Omega)} \to 0$,
- $\nabla_{\mathcal{D}_m}^{(0)} u_m$ *and* $\nabla_{\mathcal{D}_m}^{(1)} u_m$ *converge weakly in* $L^2(\Omega \times (0,T))^d$ *to* $\nabla \overline{u}$,
- $\delta_{\mathcal{D}_m} u_m$ *converges weakly in* $L^2(\Omega \times (0,T))$ *to* $\partial_t \overline{u}$,
- $\widetilde{G}_{\mathcal{D}_m}$ *and* $\widehat{G}_{\mathcal{D}_m}$ *both converge weakly to* $\overline{G}$ *in* $L^2(\Omega \times (0,T))^d$, *and*

$$\int_0^T \int_\Omega (\widetilde{G}_{\mathcal{D}_m}(\boldsymbol{x},t) - \widehat{G}_{\mathcal{D}_m}(\boldsymbol{x},t)) \cdot \nabla_{\mathcal{D}}^{(0)} u_m(\boldsymbol{x},t) \mathrm{d}\boldsymbol{x}\mathrm{d}t \to 0, \tag{5.68}$$

- $\widetilde{w}_{\mathcal{D}_m}$ *converges weakly to* $\overline{w}$ *in* $L^2(\Omega \times (0,T))$,
- *it holds*

$$\int_0^T \int_\Omega \widehat{G}_{\mathcal{D}_m}(\boldsymbol{x},t) \cdot \nabla^{(0)}_{\mathcal{D}_m} u_m(\boldsymbol{x},t) \mathrm{d}\boldsymbol{x}\mathrm{d}t \longrightarrow \int_0^T \int_\Omega \overline{G}(\boldsymbol{x},t) \cdot \nabla \overline{u}(\boldsymbol{x},t) \mathrm{d}\boldsymbol{x}\mathrm{d}t. \tag{5.69}$$

Proof. The proof is similar to the proof of Lemma 5.11. The *a priori* estimate (5.61) provide the existence of $\overline{u}$ such that $(\Pi^{(1)}_{\mathcal{D}_m} u_m)_{m\in\mathbb{N}}$, $(\nabla^{(1)}_{\mathcal{D}_m} u_m)_{m\in\mathbb{N}}$ and $(\delta_{\mathcal{D}_m} u_m)_{m\in\mathbb{N}}$ converge as stated in the lemma. The same estimates show that $(\widetilde{G}_{\mathcal{D}_m})_{m\in\mathbb{N}}$ and $(\widehat{G}_{\mathcal{D}_m})_{m\in\mathbb{N}}$ are bounded in $L^2(\Omega\times(0,T))^d$, and therefore have weak limits in this space (up to a subsequence). Likewise, $(\widetilde{w}_{\mathcal{D}_m})_{m\in\mathbb{N}}$ has a weak limit in $L^2(\Omega \times (0,T))$ up to a subsequence.
Let us now prove that $\nabla^{(0)}_{\mathcal{D}_m} u_m$ converges weakly to $\nabla\overline{u}$, that the weak limits of $(\widetilde{G}_{\mathcal{D}_m})_{m\in\mathbb{N}}$ and $(\widehat{G}_{\mathcal{D}_m})_{m\in\mathbb{N}}$ are the same, and that (5.68) holds. Since

$$\begin{aligned}
&\left\|\nabla^{(1)}_{\mathcal{D}_m} u_m - \nabla^{(0)}_{\mathcal{D}_m} u_m\right\|^2_{L^2(\Omega\times(0,T))^d} \\
&\quad = \sum_{n=0}^{N-1} \eth^{(n+\frac{1}{2})} \int_\Omega \left|\nabla_{\mathcal{D}_m} u_m^{(n+1)}(\boldsymbol{x}) - \nabla_{\mathcal{D}_m} u_m^{(n)}(\boldsymbol{x})\right|^2 \mathrm{d}\boldsymbol{x} \\
&\quad \le \eth_{\mathcal{D}_m} \sum_{n=0}^{N-1} \int_\Omega \left|\nabla_{\mathcal{D}_m} u_m^{(n+1)}(\boldsymbol{x}) - \nabla_{\mathcal{D}_m} u_m^{(n)}(\boldsymbol{x})\right|^2 \mathrm{d}\boldsymbol{x}\,,
\end{aligned} \tag{5.70}$$

the estimate (5.61) shows that

$$\left\|\nabla^{(1)}_{\mathcal{D}_m} u_m - \nabla^{(0)}_{\mathcal{D}_m} u_m\right\|_{L^2(\Omega\times(0,T))^d} \to 0 \quad \text{as } m \to \infty. \tag{5.71}$$

This proves in particular that $\nabla^{(0)}_{\mathcal{D}_m} u_m \to \nabla\overline{u}$ weakly in $L^2(\Omega\times(0,T))^d$. Take now $(\psi_m)_{m\in\mathbb{N}}$ bounded in $L^2(\Omega \times (0,T))^d$ and write

$$\begin{aligned}
&\left|\int_0^T \int_\Omega (\widetilde{G}_{\mathcal{D}_m}(\boldsymbol{x},t) - \widehat{G}_{\mathcal{D}_m}(\boldsymbol{x},t)) \cdot \psi_m(\boldsymbol{x},t) \mathrm{d}\boldsymbol{x}\mathrm{d}t\right| \\
&\le \int_0^T \int_\Omega \mu(|\nabla^{(0)}_{\mathcal{D}_m} u_m(\boldsymbol{x},t)|) \left|\nabla^{(1)}_{\mathcal{D}_m} u_m(\boldsymbol{x},t) - \nabla^{(0)}_{\mathcal{D}_m} u_m(\boldsymbol{x},t)\right| |\psi_m(\boldsymbol{x},t)| \mathrm{d}\boldsymbol{x}\mathrm{d}t \\
&\le \mu_{\max} \left\|\nabla^{(1)}_{\mathcal{D}_m} u_m - \nabla^{(0)}_{\mathcal{D}_m} u_m\right\|_{L^2(\Omega\times(0,T))^d} \|\psi_m\|_{L^2(\Omega\times(0,T))^d}\,.
\end{aligned}$$

Use then (5.71) to infer

$$\int_0^T \int_\Omega (\widetilde{G}_{\mathcal{D}_m}(\boldsymbol{x},t) - \widehat{G}_{\mathcal{D}_m}(\boldsymbol{x},t)) \cdot \psi_m(\boldsymbol{x},t) \mathrm{d}\boldsymbol{x}\mathrm{d}t \to 0 \quad \text{as } m \to \infty. \tag{5.72}$$

Applied to $\psi_m = \psi$ for a fixed ψ, this relation that the weak limits of $(\widetilde{G}_{\mathcal{D}_m})_{m\in\mathbb{N}}$ and $(\widehat{G}_{\mathcal{D}_m})_{m\in\mathbb{N}}$ are the same function $\overline{G}$. The same relation (5.72) with $\psi_m = \nabla^{(0)}_{\mathcal{D}_m} u_m$ provides (5.68).
Let us conclude by proving (5.69). Relation (5.55) is established as in the proof of Lemma 5.11. The GS (5.67) applied to $\mathcal{D} = \mathcal{D}_m$ and $v = u_m$ and the strong convergence of $\Pi^{(1)}_{\mathcal{D}_m} u_m$ then show that

$$\begin{aligned}\int_0^T \int_\Omega \widetilde{G}_{\mathcal{D}_m}(\boldsymbol{x},t)\cdot\nabla^{(1)}_{\mathcal{D}_m} u_m(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t &= \int_0^T \int_\Omega \widetilde{w}_{\mathcal{D}_m}(\boldsymbol{x},t)\Pi^{(1)}_{\mathcal{D}_m} u_m(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t\\ &\to \int_0^T \int_\Omega \overline{w}(\boldsymbol{x},t)\overline{u}(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t\\ &= \int_0^T \int_\Omega \overline{G}(\boldsymbol{x},t)\cdot\nabla\overline{u}(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t.\end{aligned}$$

Since $\nabla^{(1)}_{\mathcal{D}_m} u_m - \nabla^{(0)}_{\mathcal{D}_m} u_m \to 0$ in $L^2(\Omega\times(0,T))^d$ and $(\widetilde{G}_{\mathcal{D}_m})_{m\in\mathbb{N}}$ is bounded in $L^2(\Omega\times(0,T))^d$, this gives

$$\int_0^T \int_\Omega \widetilde{G}_{\mathcal{D}_m}(\boldsymbol{x},t)\cdot\nabla^{(0)}_{\mathcal{D}_m} u_m(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t \to \int_0^T \int_\Omega \overline{G}(\boldsymbol{x},t)\cdot\nabla\overline{u}(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t.$$

We conclude the proof of (5.69) by using (5.68). ∎

The following theorem states the convergence of the semi-implicit scheme. The proof is omitted, as it is identical to the proof of Theorem 5.13, replacing $G_{\mathcal{D}_m}$ by $\widehat{G}_{\mathcal{D}_m}$ and $\nabla^{(1)}_{\mathcal{D}_m} u_m$ by $\nabla^{(0)}_{\mathcal{D}} u_m$ in the definition of $T_m(V)$ in Step 2 (use of Minty trick).

Theorem 5.16.
Assume (5.40) and let $((\mathcal{D}_T)_m)_{m\in\mathbb{N}}$ be a sequence of space–time GDs for homogeneous Dirichlet boundary conditions (with $p = 2$).
Assume that $((\mathcal{D}_T)_m)_{m\in\mathbb{N}}$ is space–time-consistent, limit-conforming and compact in the sense of Definitions 4.3 and 4.6.
We also suppose that $(\nabla_{\mathcal{D}_m}\mathcal{I}_{\mathcal{D}_m} u_{\mathrm{ini}})_{m\in\mathbb{N}}$ is bounded in $L^2(\Omega)^d$ and, for any $m\in\mathbb{N}$, we let u_m be the solution to the GS (5.60).
Then there exists a weak solution $\overline{u}$ of (5.39) in the sense of Definition 5.9 such that, up to a subsequence as $m\to\infty$,

- $\sup_{t\in[0,T]} \|\Pi^{(1)}_{\mathcal{D}_m} u_m(t) - \overline{u}(t)\|_{L^2(\Omega)} \to 0$, *and*
- $\nabla^{(1)}_{\mathcal{D}_m} u_m \to \nabla\overline{u}$ *and* $\nabla^{(0)}_{\mathcal{D}_m} u_m \to \nabla\overline{u}$ *in* $L^2(\Omega\times(0,T))^d$.

5.3 Non-linear time-dependent Leray–Lions type problems

5.3.1 Model

We consider here an evolution problem based on a Leray–Lions type operator, with non-homogeneous Neumann boundary conditions and non-local dependency on the lower order terms. The model reads

$$\begin{aligned} &\partial_t \overline{u} - \operatorname{div}(\boldsymbol{a}(\boldsymbol{x}, \overline{u}, \nabla \overline{u})) = f && \text{in } \Omega \times (0,T), \\ &\overline{u}(\boldsymbol{x}, 0) = u_{\text{ini}}(\boldsymbol{x}) && \text{in } \Omega, \\ &\boldsymbol{a}(\boldsymbol{x}, \overline{u}, \nabla \overline{u}) \cdot \boldsymbol{n} = g && \text{on } \partial\Omega \times (0,T), \end{aligned} \tag{5.73}$$

where $\boldsymbol{a}$ satisfies (2.85a)–(2.85d) and

$$\begin{aligned} &\bullet\ T \in (0, +\infty), \\ &\bullet\ u_{\text{ini}} \in L^2(\Omega), \\ &\bullet\ f \in L^{p'}(\Omega \times (0,T)) \text{ and } g \in L^{p'}(\partial\Omega \times (0,T))\,, \text{ where } p' = \tfrac{p}{p-1}. \end{aligned} \tag{5.74}$$

The non-linear equation (5.73) covers a number of models, including semi-linear ones appearing in image processing [42, 46]. The analysis of the GDM applied to (5.73) with homogeneous Dirichlet boundary conditions is done in [76]. In the quasi-linear case, that is $\boldsymbol{a}(\boldsymbol{x}, \overline{u}, \nabla \overline{u}(\boldsymbol{x})) = G(\boldsymbol{x}, \overline{u}) \nabla \overline{u}(\boldsymbol{x})$ with G a matrix-valued function having a non-local dependency on $\overline{u}$ as in (2.85a), the convergence of the GDM was established in [103].
The precise notion of solution to (5.73) is the following:

$$\begin{cases} \overline{u} \in L^p(0,T; W^{1,p}(\Omega)) \cap C([0,T]; L^2(\Omega))\,,\ \overline{u}(\cdot, 0) = u_{\text{ini}}, \\ \partial_t \overline{u} \in L^{p'}(0,T; (W^{1,p}(\Omega))') \text{ and} \\ \displaystyle\int_0^T \langle \partial_t \overline{u}(\cdot,t), \overline{v}(\cdot,t) \rangle_{(W^{1,p}(\Omega))', W^{1,p}(\Omega)} \mathrm{d}t \\ \displaystyle + \int_0^T \int_\Omega \boldsymbol{a}(\boldsymbol{x}, \overline{u}(\cdot,t), \nabla \overline{u}(\boldsymbol{x},t)) \cdot \nabla \overline{v}(\boldsymbol{x},t) \mathrm{d}\boldsymbol{x} \mathrm{d}t \\ \displaystyle \quad = \int_0^T \int_\Omega f(\boldsymbol{x},t) \overline{v}(\boldsymbol{x},t) \mathrm{d}\boldsymbol{x} \mathrm{d}t + \int_0^T \int_{\partial\Omega} g(\boldsymbol{x},t) \gamma \overline{v}(\boldsymbol{x},t) \mathrm{d}\gamma(\boldsymbol{x}) \mathrm{d}t\,, \\ \quad \forall \overline{v} \in L^p(0;T; W^{1,p}(\Omega)). \end{cases} \tag{5.75}$$

Remark 5.17. The derivative $\partial_t \overline{u}$ is understood in the sense of distributions on $(0,T)$ with values in $L^2(\Omega)$. Stating that it belongs to $L^{p'}(0,T; (W^{1,p}(\Omega))') = (L^p(0,T; W^{1,p}(\Omega)))'$ amounts to asking that the linear form defined by

$$
\begin{aligned}
&C_c^\infty(0,T;L^2(\Omega)) \cap L^p(0,T;W^{1,p}(\Omega)) \to \mathbb{R} \\
&\varphi \mapsto \langle \partial_t \overline{u}, \varphi \rangle_{\mathcal{D}'(0,T;L^2(\Omega)),\mathcal{D}(0,T;L^2(\Omega))} \\
&\qquad := -\int_0^T \langle \overline{u}(\cdot,t), \partial_t \varphi(\cdot,t) \rangle_{L^2(\Omega),L^2(\Omega)} \mathrm{d}t \\
&\qquad = -\int_0^T \int_\Omega \overline{u}(\boldsymbol{x},t) \partial_t \varphi(\boldsymbol{x},t) \mathrm{d}\boldsymbol{x}\mathrm{d}t
\end{aligned}
\tag{5.76}
$$

is continuous for the norm of $L^p(0,T;W^{1,p}(\Omega))$. Since the set of tensorial functions $\mathcal{S} = \{\sum_{i=1}^q \varphi_i(t)\beta_i(\boldsymbol{x}) \,:\, q \in \mathbb{N}\,,\ \varphi_i \in C_c^\infty(0,T)\,,\ \beta_i \in C^\infty(\overline{\Omega})\}$ is dense in $L^p(0,T;W^{1,p}(\Omega))$ (see [67, Corollary 1.3.1]), the derivative $\partial_t \overline{u}$ belongs to $L^{p'}(0,T;(W^{1,p}(\Omega))')$ if and only if (5.76) is continuous on $\mathcal{S}$ for the $L^p(0,T;W^{1,p}(\Omega))$-norm.

Remark 5.18. Using regularisation and integration-by-parts techniques [67, Section 2.5.2], it is possible to see that any solution $\overline{u}$ to (5.75) also satisfies, for any $s \in [0,T]$,

$$
\begin{aligned}
\frac{1}{2}\|\overline{u}(s)\|^2_{L^2(\Omega)} &+ \int_0^s \int_\Omega \boldsymbol{a}(\boldsymbol{x},\overline{u}(\cdot,\tau),\nabla\overline{u}(\boldsymbol{x},\tau)) \cdot \nabla\overline{u}(\boldsymbol{x},\tau)\mathrm{d}\boldsymbol{x}\mathrm{d}\tau \\
&= \frac{1}{2}\|u_{\mathrm{ini}}\|^2_{L^2(\Omega)} + \int_0^s \int_\Omega f(\boldsymbol{x},\tau)\overline{u}(\boldsymbol{x},\tau)\mathrm{d}\boldsymbol{x}\mathrm{d}\tau \\
&\qquad + \int_0^s \int_{\partial\Omega} g(\boldsymbol{x},\tau)\gamma\overline{u}(\boldsymbol{x},\tau)\mathrm{d}\gamma(\boldsymbol{x})\mathrm{d}\tau.
\end{aligned}
\tag{5.77}
$$

With a reasoning similar to the one employed to establish the equivalence of (5.3) and (5.4), we can see that (5.75) is equivalent to:

$$
\begin{cases}
\overline{u} \in L^p(0,T;W^{1,p}(\Omega)) \cap L^\infty(0,T;L^2(\Omega)) \text{ and,} \\
\text{for all } \overline{v} \in C^1([0,T];W^{1,p}(\Omega) \cap L^2(\Omega)) \text{ such that } \overline{v}(\cdot,T) = 0, \\
-\displaystyle\int_0^T \int_\Omega \overline{u}(\boldsymbol{x},t)\partial_t\overline{v}(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t - \int_\Omega u_{\mathrm{ini}}(\boldsymbol{x})\overline{v}(\boldsymbol{x},0)\mathrm{d}\boldsymbol{x} \\
+\displaystyle\int_0^T \int_\Omega \boldsymbol{a}(\boldsymbol{x},\overline{u}(\cdot,t),\nabla\overline{u}(\boldsymbol{x},t)) \cdot \nabla\overline{v}(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t \\
\quad = \displaystyle\int_0^T \int_\Omega f(\boldsymbol{x},t)\overline{v}(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t + \int_0^T \int_{\partial\Omega} g(\boldsymbol{x},t)\gamma\overline{v}(\boldsymbol{x},t)\mathrm{d}\gamma(\boldsymbol{x})\mathrm{d}t.
\end{cases}
\tag{5.78}
$$

To prove this equivalence, we use [67, Section 2.5.2] to see that if $\overline{u} \in L^p(0,T;W^{1,p}(\Omega) \cap L^2(\Omega))$ satisfies $\partial_t\overline{u} \in L^{p'}(0,T;(W^{1,p}(\Omega) \cap L^2(\Omega))')$, then $\overline{u} \in C([0,T];L^2(\Omega))$. We also use the density in $L^p(0,T;W^{1,p}(\Omega))$ of $C^1([0,T];W^{1,p}(\Omega) \cap L^2(\Omega))$, which is for example a consequence of [67, Corollary 1.3.1].

5.3.2 Gradient scheme and main results

Let $\mathcal{D}_T = (X_\mathcal{D}, \Pi_\mathcal{D}, \mathbb{T}_\mathcal{D}, \nabla_\mathcal{D}, \mathcal{I}_\mathcal{D}, (t^{(n)})_{n=0,\dots,N})$ be a space–time GD for non-homogeneous Neumann conditions in the sense of Definition 4.1, and let $\theta \in$

$[\frac{1}{2}, 1]$. The GDM applied to Problem (5.73) yields the following GS: find a family $(u^{(n)})_{n=0,\dots,N} \in X_{\mathcal{D}}^{N+1}$ such that

$$\begin{cases} u^{(0)} = \mathcal{I}_{\mathcal{D}} u_{\text{ini}} \in X_{\mathcal{D}} \text{ and, for all } n = 0, \dots, N-1,\ u^{(n+1)} \text{ satisfies} \\ \displaystyle\int_\Omega \delta_{\mathcal{D}}^{(n+\frac{1}{2})} u(\boldsymbol{x}) \Pi_{\mathcal{D}} v(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \\ \displaystyle + \int_\Omega \boldsymbol{a}\left(\boldsymbol{x}, \Pi_{\mathcal{D}} u^{(n+\theta)}, \nabla_{\mathcal{D}} u^{(n+\theta)}(\boldsymbol{x})\right) \cdot \nabla_{\mathcal{D}} v(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \\ \displaystyle \quad = \frac{1}{\eth^{(n+\frac{1}{2})}} \int_{t^{(n)}}^{t^{(n+1)}} \int_\Omega f(\boldsymbol{x}, t) \Pi_{\mathcal{D}} v(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \mathrm{d}t \\ \displaystyle \qquad + \frac{1}{\eth^{(n+\frac{1}{2})}} \int_{t^{(n)}}^{t^{(n+1)}} \int_{\partial\Omega} g(\boldsymbol{x}, t) \mathbb{T}_{\mathcal{D}} v(\boldsymbol{x}) \mathrm{d}\gamma(\boldsymbol{x}) \mathrm{d}t\,, \quad \forall v \in X_{\mathcal{D}}. \end{cases} \tag{5.79}$$

The choice $\theta \geq \frac{1}{2}$ is required for stability reasons. As explained in Section 4.1, $\theta = 1$ leads to the classical Euler time implicit discretisation, while $\theta = \frac{1}{2}$ corresponds to the Crank–Nicolson time discretisation.
Recalling the notations in (4.2), we now state our initial convergence results for this GS.

Theorem 5.19 (Convergence of the GS for transient Leray–Lions type problem). *Under Assumptions (2.85a)–(2.85d) and (5.74), let $((\mathcal{D}_T)_m)_{m\in\mathbb{N}}$ be a sequence of space–time GDs for non-homogeneous Neumann boundary conditions, in the sense of Definition 4.1. Assume that this sequence is space–time-consistent, limit-conforming and compact in the sense of Definitions 4.3 and 4.6. Let $\theta \in [\frac{1}{2}, 1]$ be given.*
Then, for any $m \in \mathbb{N}$, there exists a solution u_m to the GS (5.79) with $\mathcal{D} = \mathcal{D}_m$ and, along a subsequence as $m \to \infty$,

- *$\Pi_{\mathcal{D}_m}^{(\theta)} u_m$ converges to $\overline{u}$ strongly in $L^p(\Omega \times (0,T))$,*
- *$\Pi_{\mathcal{D}_m}^{(1)} u_m$ converges to $\overline{u}$ weakly in $L^2(\Omega)$ uniformly on $[0,T]$ (see Definition C.14),*
- *$\nabla_{\mathcal{D}_m}^{(\theta)} u_m$ converges to $\nabla\overline{u}$ weakly in $L^p(\Omega \times (0,T))^d$,*

where $\overline{u}$ is a solution to (5.75).

Remark 5.20. As for the stationary problem (see Remark 2.45), the existence of a solution to (5.75) is a by-product of the proof of convergence of the GDM. Moreover, in the case where the solution $\overline{u}$ of (5.75) is unique, the whole sequence $(u_m)_{m\in\mathbb{N}}$ converges to $\overline{u}$ in the senses above.

The convergence of the function reconstructions is actually much better than in the initial result above. It is uniform-in-time and strong in space.

Theorem 5.21 (Uniform-in-time convergence of the GS). *Under the assumptions and notations of Theorem 5.19, and along the same subsequence as in this theorem, we have*

- $\sup_{t\in[0,T]} \|\Pi^{(\theta)}_{\mathcal{D}_m} u_m(t) - \overline{u}(t)\|_{L^2(\Omega)} \to 0$,
- $\sup_{t\in[0,T]} \|\Pi^{(1)}_{\mathcal{D}_m} u_m(t) - \overline{u}(t)\|_{L^2(\Omega)} \to 0$.

If the Leray–Lions type operator $\boldsymbol{a}$ is strictly monotone, then a strong convergence result can also be stated on the gradients.

Theorem 5.22 (Strong convergence of the gradients in the strictly monotone case). *Let us assume the hypotheses of Theorem 5.19, and that $\boldsymbol{a}$ is strictly monotone in the sense of (2.87). Then, with the same notations and along the same subsequence as in Theorem 5.19, $\nabla^{(\theta)}_{\mathcal{D}_m} u_m$ converges strongly to $\nabla\overline{u}$ in $L^p(\Omega\times(0,T))^d$.*

5.3.3 *A priori* estimates

We begin by establishing *a priori* estimates.

Lemma 5.23 ($L^\infty(0,T;L^2(\Omega))$ estimate, discrete $L^p(0,T;W^{1,p}(\Omega))$ estimate, and existence of a solution to the GS). *Under Hypotheses (2.85a)–(2.85d) and (5.74), let $\mathcal{D}_T$ be a space–time GD for non-homogeneous Neumann conditions in the sense of Definition 4.1. Then there exists at least one solution to the GS (5.79), and there exists $C_{12} > 0$, depending only on p, $C_P \geq C_{\mathcal{D}}$, $C_{\rm ini} \geq \|\Pi_{\mathcal{D}}\mathcal{I}_{\mathcal{D}} u_{\rm ini}\|_{L^2(\Omega)}$, f, g and $\underline{a}$ such that, for any solution u to this scheme,*

$$\begin{aligned}&\sup_{t\in[0,T]} \left\|\Pi^{(1)}_{\mathcal{D}} u(t)\right\|_{L^2(\Omega)} \leq C_{12}\,, \quad \sup_{t\in[0,T]} \left\|\Pi^{(\theta)}_{\mathcal{D}} u(t)\right\|_{L^2(\Omega)} \leq C_{12}\\ &\text{and } \left\|\nabla^{(\theta)}_{\mathcal{D}} u\right\|_{L^p(\Omega\times(0,T))^d} \leq C_{12}.\end{aligned} \tag{5.80}$$

Proof. Let us first prove the estimates. Recall (5.22), that is

$$\eth^{(n+\frac{1}{2})}\delta^{(n+\frac{1}{2})}_{\mathcal{D}} u\ \Pi_{\mathcal{D}} u^{(n+\theta)} \geq \frac{1}{2}\left((\Pi_{\mathcal{D}} u^{(n+1)})^2 - (\Pi_{\mathcal{D}} u^{(n)})^2\right),$$

choose $v = \eth^{(n+\frac{1}{2})} u^{(n+\theta)}$ in (5.79), and sum over $n = 0,\ldots,k-1$ for a given $k \in \{1,\ldots,N\}$. This yields

$$\begin{aligned}&\frac{1}{2}\left\|\Pi_{\mathcal{D}} u^{(k)}\right\|^2_{L^2(\Omega)} + \int_0^{t^{(k)}}\int_\Omega \boldsymbol{a}\left(\boldsymbol{x}, \Pi^{(\theta)}_{\mathcal{D}} u(\cdot,t), \nabla^{(\theta)}_{\mathcal{D}} u(\boldsymbol{x},t)\right)\cdot\nabla^{(\theta)}_{\mathcal{D}} u(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t\\ &\quad\leq \frac{1}{2}\left\|\Pi_{\mathcal{D}} u^{(0)}\right\|^2_{L^2(\Omega)} + \int_0^{t^{(k)}}\int_\Omega f(\boldsymbol{x},t)\Pi^{(\theta)}_{\mathcal{D}} u(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t\\ &\qquad+\int_0^{t^{(k)}}\int_{\partial\Omega} g(\boldsymbol{x},t)\mathbb{T}^{(\theta)}_{\mathcal{D}} u(\boldsymbol{x},t)\mathrm{d}\gamma(\boldsymbol{x})\mathrm{d}t.\end{aligned} \tag{5.81}$$

In particular, owing to the coercivity property (2.85b) of $\boldsymbol{a}$, and using Hölder's inequality and Young's inequality (D.9) (the latter with $\varepsilon = \frac{p\underline{a}}{4C_{\mathcal{D}}^p}$),

$$
\begin{aligned}
&\frac{1}{2}\left\|\Pi_{\mathcal{D}}u^{(k)}\right\|^2_{L^2(\Omega)} + \underline{a}\int_0^{t^{(k)}}\left\|\nabla^{(\theta)}_{\mathcal{D}}u(\cdot,t)\right\|^p_{L^p(\Omega)^d}\,\mathrm{d}t \\
&\le \frac{1}{2}\left\|\Pi_{\mathcal{D}}u^{(0)}\right\|^2_{L^2(\Omega)} + \frac{4^{1/(p-1)}C_{\mathcal{D}}^{p'}}{(p\underline{a})^{1/(p-1)}\,p'}\left\|f\right\|^{p'}_{L^{p'}(\Omega\times(0,t^{(k)}))} \\
&\quad + \frac{\underline{a}}{4C_{\mathcal{D}}^p}\left\|\Pi^{(\theta)}_{\mathcal{D}}u\right\|^p_{L^p(\Omega\times(0,t^{(k)}))} + \frac{4^{1/(p-1)}C_{\mathcal{D}}^{p'}}{(p\underline{a})^{1/(p-1)}\,p'}\left\|g\right\|^{p'}_{L^{p'}(\partial\Omega\times(0,t^{(k)}))} \\
&\quad + \frac{\underline{a}}{4C_{\mathcal{D}}^p}\left\|\mathbb{T}^{(\theta)}_{\mathcal{D}}u\right\|^p_{L^p(\partial\Omega\times(0,t^{(k)}))}.
\end{aligned}
$$

Apply the definition (3.9) of $C_{\mathcal{D}}$ and recall that $u^{(0)} = \mathcal{I}_{\mathcal{D}}u_{\mathrm{ini}}$ to deduce

$$
\begin{aligned}
&\frac{1}{2}\left\|\Pi_{\mathcal{D}}u^{(k)}\right\|^2_{L^2(\Omega)} + \frac{\underline{a}}{2}\left\|\nabla^{(\theta)}_{\mathcal{D}}u\right\|^p_{L^p(\Omega\times(0,t^{(k)}))^d} \\
&\qquad \le \frac{1}{2}\left\|\Pi_{\mathcal{D}}\mathcal{I}_{\mathcal{D}}u_{\mathrm{ini}}\right\|^2_{L^2(\Omega)} + \frac{4^{1/(p-1)}C_{\mathcal{D}}^{p'}}{(p\underline{a})^{1/(p-1)}\,p'}\left\|f\right\|^{p'}_{L^{p'}(\Omega\times(0,t^{(k)}))} \\
&\qquad\quad + \frac{4^{1/(p-1)}C_{\mathcal{D}}^{p'}}{(p\underline{a})^{1/(p-1)}\,p'}\left\|g\right\|^{p'}_{L^{p'}(\partial\Omega\times(0,t^{(k)}))}.
\end{aligned}
$$

This establishes the estimates on $\Pi^{(1)}_{\mathcal{D}}u$ and $\nabla^{(\theta)}_{\mathcal{D}}u$. The estimate on $\Pi^{(\theta)}_{\mathcal{D}}u$ follows from the inequality

$$
\left\|\Pi_{\mathcal{D}}u^{(n+\theta)}\right\|_{L^2(\Omega)} \le \theta\left\|\Pi_{\mathcal{D}}u^{(n+1)}\right\|_{L^2(\Omega)} + (1-\theta)\left\|\Pi_{\mathcal{D}}u^{(n)}\right\|_{L^2(\Omega)}.
$$

The existence of at least one solution to (5.79) is done as in the proof of Theorem 2.44, reasoning on $u^{(n+\theta)}$ and using the above estimates. ■

The following estimate will be useful to apply the Aubin–Simon theorem for GD (Theorem 4.14).

Lemma 5.24 (Estimate on the dual norm of the discrete time derivative). *Under Hypotheses (2.85a)–(2.85d) and (5.74), let $\mathcal{D}_T$ be a space–time GD for non-homogeneous Neumann conditions in the sense of Definition 4.1. Let u be a solution to the GS (5.79). Then there exists C_{13}, depending only on p, μ, $\overline{a}$, $\underline{a}$, $C_{\mathrm{ini}} \ge \|\Pi_{\mathcal{D}}\mathcal{I}_{\mathcal{D}}u_{\mathrm{ini}}\|_{L^2(\Omega)}$, f, g, T and $C_P \ge C_{\mathcal{D}}$, such that*

$$
\int_0^T \|\delta_{\mathcal{D}}u(t)\|^{p'}_{\star,\mathcal{D}}\,\mathrm{d}t \le C_{13}, \tag{5.82}
$$

where the dual norm $\|\cdot\|_{\star,\mathcal{D}}$ is given by Definition 4.11.

Proof. Let us take a generic $v \in X_{\mathcal{D}}$ as a test function in (5.79). We have, thanks to Assumption (2.85d) on $\boldsymbol{a}$,

$$\begin{aligned}\int_\Omega \delta_{\mathcal{D}}^{(n+\frac{1}{2})} u(\boldsymbol{x}) \Pi_{\mathcal{D}} v(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \\ &\leq \int_\Omega \left(\overline{a}(\boldsymbol{x}) + \mu |\nabla_{\mathcal{D}} u^{(n+\theta)}(\boldsymbol{x})|^{p-1}\right) |\nabla_{\mathcal{D}} v(\boldsymbol{x})| \mathrm{d}\boldsymbol{x} \\ &\quad + \frac{1}{\eth t^{(n+\frac{1}{2})}} \int_{t^{(n)}}^{t^{(n+1)}} \int_\Omega f(\boldsymbol{x},t) \Pi_{\mathcal{D}} v(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \mathrm{d}t \\ &\quad + \frac{1}{\eth t^{(n+\frac{1}{2})}} \int_{t^{(n)}}^{t^{(n+1)}} \int_{\partial\Omega} g(\boldsymbol{x},t) \mathbb{T}_{\mathcal{D}} v(\boldsymbol{x}) \mathrm{d}\gamma(\boldsymbol{x}) \mathrm{d}t.\end{aligned}$$

This leads, by definition (3.9) of $C_{\mathcal{D}}$, to the existence of $C_{14} > 0$ depending only on p, μ such that

$$\begin{aligned}\int_\Omega \delta_{\mathcal{D}}^{(n+\frac{1}{2})} u(\boldsymbol{x}) \Pi_{\mathcal{D}} v(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \\ &\leq C_{14} \begin{pmatrix} \|\overline{a}\|_{L^{p'}(\Omega)} + \left\|\nabla_{\mathcal{D}} u^{(n+\theta)}\right\|_{L^p(\Omega)^d}^{p-1} \\ + \dfrac{C_{\mathcal{D}}}{\eth t^{(n+\frac{1}{2})}} \displaystyle\int_{t^{(n)}}^{t^{(n+1)}} \|f(\cdot,t)\|_{L^{p'}(\Omega)} \mathrm{d}t \\ + \dfrac{C_{\mathcal{D}}}{\eth t^{(n+\frac{1}{2})}} \displaystyle\int_{t^{(n)}}^{t^{(n+1)}} \|g(\cdot,t)\|_{L^{p'}(\partial\Omega)} \mathrm{d}t \end{pmatrix} \|v\|_{\mathcal{D}}.\end{aligned}$$

Taking the supremum on $v \in X_{\mathcal{D}}$ such that $\|v\|_{\mathcal{D}} = 1$ gives

$$\begin{aligned}\left\|\delta_{\mathcal{D}}^{(n+\frac{1}{2})} u\right\|_{\star,\mathcal{D}} &\leq C_{14} \|\overline{a}\|_{L^{p'}(\Omega)} + C_{14} \left\|\nabla_{\mathcal{D}} u^{(n+\theta)}\right\|_{L^p(\Omega)^d}^{p-1} \\ &+ \frac{C_{14} C_{\mathcal{D}}}{\eth t^{(n+\frac{1}{2})}} \int_{t^{(n)}}^{t^{(n+1)}} \|f(\cdot,t)\|_{L^{p'}(\Omega)} \mathrm{d}t + \frac{C_{14} C_{\mathcal{D}}}{\eth t^{(n+\frac{1}{2})}} \int_{t^{(n)}}^{t^{(n+1)}} \|g(\cdot,t)\|_{L^{p'}(\partial\Omega)} \mathrm{d}t.\end{aligned}$$

The proof is concluded by raising this estimate to the power p', distributing this power to each term on the right-hand side thanks to the power-of-sums inequality (D.14), using Jensen's inequality for the integral terms, multiplying by $\eth t^{(n+\frac{1}{2})}$, summing on n and invoking Lemma 5.23 to estimate $\|\nabla_{\mathcal{D}}^{(\theta)} u\|_{L^p(\Omega\times(0,T))^d}^p$. ∎

5.3.4 Proof of the convergence results

We now prove the convergence of the GDM for the transient Leray–Lions type model (5.73).

Proof of Theorem 5.19.

Step 1: Application of compactness results.
The definition (3.8) of $\|\cdot\|_{\mathcal{D}_m}$ and Estimates (5.80) and (5.82) show that the hypotheses of Lemma 4.8 (regularity of the limit) and Theorem 4.14

(Aubin–Simon for GD) are satisfied by $(u_m)_{m\in\mathbb{N}}$. Therefore, there exists $\overline{u} \in L^p(0,T;W^{1,p}(\Omega))$ such that, up to a subsequence as $m \to \infty$, $\Pi^{(\theta)}_{\mathcal{D}_m} u_m \to \overline{u}$ strongly in $L^p(\Omega \times (0,T))$, $\nabla^{(\theta)}_{\mathcal{D}_m} u_m \to \nabla\overline{u}$ weakly in $L^p(\Omega \times (0,T))^d$, and $\mathbb{T}^{(\theta)}_{\mathcal{D}_m} u_m \to \gamma\overline{u}$ weakly in $L^p(\partial\Omega \times (0,T))$. Moreover, since $(\Pi^{(\theta)}_{\mathcal{D}_m} u_m)_{m\in\mathbb{N}}$ is bounded in $L^\infty(0,T;L^2(\Omega))$, the convergence $\Pi^{(\theta)}_{\mathcal{D}_m} u_m \to \overline{u}$ also holds in $L^\infty(0,T;L^2(\Omega))$ weak-$*$.
Estimates (5.80) and (5.82) show that $(u_m)_{m\in\mathbb{N}}$ satisfies the assumptions of Theorem 4.19 with $\theta = 1$. Hence, up to a subsequence as $m \to \infty$, $(\Pi^{(1)}_{\mathcal{D}_m} u_m)_{m\in\mathbb{N}}$ converges to some $\widetilde{u}$ uniformly-in-time for the weak topology of $L^2(\Omega)$ (as per Definition C.14).
Estimates (5.80) and Assumption (2.85d) show that the functions $\mathcal{A}_{\mathcal{D}_m}(\boldsymbol{x},t) = \boldsymbol{a}(\boldsymbol{x}, \Pi^{(\theta)}_{\mathcal{D}_m} u_m(\cdot,t), \nabla^{(\theta)}_{\mathcal{D}_m} u_m(\boldsymbol{x},t))$ remain bounded in $L^{p'}(\Omega\times(0,T))^d$. Up to a subsequence, $\mathcal{A}_{\mathcal{D}_m}$ therefore converges to some $\boldsymbol{A}$ weakly in $L^{p'}(\Omega \times (0,T))^d$ as $m \to \infty$.

Step 2: Proof that $\overline{u} = \widetilde{u}$.
First notice that the convergence of $(\Pi^{(1)}_{\mathcal{D}_m} u_m)_{m\in\mathbb{N}}$ towards $\widetilde{u}$ also holds for the weak topology of $L^2(\Omega\times(0,T))$ (this is an easy consequence of its convergence uniformly-in-time and weakly in $L^2(\Omega)$).
Take $\varphi \in C^\infty_c(\Omega \times (0,T))$ and let

$$I_{\mathcal{D}_m}\varphi(t) = \operatorname*{argmin}_{w\in X_{\mathcal{D}_m}} \Big(\|\Pi_{\mathcal{D}_m} w - \varphi(t)\|_{L^{\max(p,2)}(\Omega)} + \|\nabla_{\mathcal{D}_m} w - \nabla\varphi(t)\|_{L^p(\Omega)^d} \Big).$$

Since $0 \in X_{\mathcal{D}_m}$, a triangle inequality shows that

$$\|\Pi_{\mathcal{D}_m} I_{\mathcal{D}_m}\varphi(t)\|_{L^{\max(p,2)}(\Omega)} + \|\nabla_{\mathcal{D}_m} I_{\mathcal{D}_m}\varphi(t)\|_{L^p(\Omega)^d} \le 2\,\|\varphi(t)\|_{L^{\max(p,2)}(\Omega)} + 2\,\|\nabla\varphi(t)\|_{L^p(\Omega)^d}\,. \tag{5.83}$$

In particular, by definition of $\|\cdot\|_{\mathcal{D}_m}$ and smoothness of φ, $\|I_{\mathcal{D}_m}\varphi\|_{L^p(0,T;X_{\mathcal{D}_m})}$ remains bounded. Moreover, the space–time-consistency of $((\mathcal{D}_T)_m)_{m\in\mathbb{N}}$ ensures that, for all $t \in (0,T)$, $\Pi_{\mathcal{D}_m} I_{\mathcal{D}_m}\varphi(t) \to \varphi(t)$ in $L^2(\Omega)$ as $m \to \infty$. Combined with the dominated convergence theorem and (5.83), this yields $\Pi_{\mathcal{D}_m} I_{\mathcal{D}_m}\varphi \to \varphi$ in $L^2(\Omega \times (0,T))$ as $m \to \infty$.
For $n \in \{0,\dots,N-1\}$ and $t \in (t^{(n)}, t^{(n+1)}]$,

$$\begin{aligned}\Pi^{(1)}_{\mathcal{D}_m} u_m(t) - \Pi^{(\theta)}_{\mathcal{D}_m} u_m(t) &= \Pi_{\mathcal{D}_m} u_m^{(n+1)} - \Pi_{\mathcal{D}_m} u_m^{(n+\theta)} \\ &= (1-\theta)(\Pi_{\mathcal{D}_m} u_m^{(n+1)} - \Pi_{\mathcal{D}_m} u_m^{(n)}) \\ &= (1-\theta)\delta t^{(n+\frac{1}{2})}\delta_{\mathcal{D}_m} u_m(t)\end{aligned} \tag{5.84}$$

and thus, by (4.17),

$$\left|\int_0^T\int_\Omega\left(\Pi^{(1)}_{\mathcal{D}_m}u_m(\boldsymbol{x},t)-\Pi^{(\theta)}_{\mathcal{D}_m}u_m(\boldsymbol{x},t)\right)\Pi_{\mathcal{D}_m}[I_{\mathcal{D}_m}\varphi(t)](\boldsymbol{x})\mathrm{d}\boldsymbol{x}\mathrm{d}t\right|$$
$$\leq(1-\theta)\delta t_{\mathcal{D}_m}\int_0^T\|\delta_{\mathcal{D}_m}u(t)\|_{\star,\mathcal{D}_m}\|I_{\mathcal{D}_m}\varphi(t)\|_{\mathcal{D}_m}\,\mathrm{d}t.$$

Use Lemma 5.24 and Hölder's inequality to see that the right-hand side of this relation tends to 0 as $m\to\infty$. Since $\Pi^{(1)}_{\mathcal{D}_m}u_m$ and $\Pi^{(\theta)}_{\mathcal{D}_m}u_m$ converge weakly in $L^2(\Omega\times(0,T))$ towards $\widetilde{u}$ and $\overline{u}$, respectively, we deduce

$$\begin{aligned}&\int_0^T\int_\Omega(\widetilde{u}(\boldsymbol{x},t)-\overline{u}(\boldsymbol{x},t))\varphi(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t\\&=\lim_{m\to\infty}\int_0^T\int_\Omega(\Pi^{(1)}_{\mathcal{D}_m}u_m(\boldsymbol{x},t)-\Pi^{(\theta)}_{\mathcal{D}_m}u_m(\boldsymbol{x},t))\Pi_{\mathcal{D}_m}[I_{\mathcal{D}_m}\varphi(t)](\boldsymbol{x})\mathrm{d}\boldsymbol{x}\mathrm{d}t\\&=0.\end{aligned}\tag{5.85}$$

This proves that $\widetilde{u}=\overline{u}$, and thus that $\Pi^{(1)}_{\mathcal{D}_m}u_m\to\overline{u}$ uniformly on $[0,T]$ weakly in $L^2(\Omega)$. In particular, $\Pi^{(1)}_{\mathcal{D}_m}u_m(T)\to\overline{u}(T)$ weakly in $L^2(\Omega)$ and thus

$$\int_\Omega\overline{u}(\boldsymbol{x},T)^2\mathrm{d}\boldsymbol{x}\leq\liminf_{m\to\infty}\int_\Omega\Pi^{(1)}_{\mathcal{D}_m}u_m(\boldsymbol{x},T)^2\mathrm{d}\boldsymbol{x}.\tag{5.86}$$

Step 3: Proof that $\overline{u}$ is a solution to (5.75).
Let $\overline{v}\in C^1([0,T];W^{1,p}(\Omega)\cap L^2(\Omega))$ such that $\overline{v}(\cdot,T)=0$, and let $(v_m)_{m\in\mathbb{N}}$ be given by Lemma 4.10. Properties (4.7), (4.8) and (4.10) therefore hold, with $\theta=0$.
We drop some indices m for legibility. Using $\delta t^{(n+\frac12)}v^{(n)}$ as test function in (5.79) yields $T_1^{(m)}+T_2^{(m)}=T_3^{(m)}+T_4^{(m)}$ with

$$T_1^{(m)}=\sum_{n=0}^{N-1}\int_\Omega\left[\Pi_{\mathcal{D}}u^{(n+1)}(\boldsymbol{x})-\Pi_{\mathcal{D}}u^{(n)}(\boldsymbol{x})\right]\Pi_{\mathcal{D}}v^{(n)}(\boldsymbol{x})\mathrm{d}\boldsymbol{x},$$

$$T_2^{(m)}=\int_0^T\int_\Omega\boldsymbol{a}\left(\boldsymbol{x},\Pi^{(\theta)}_{\mathcal{D}}u(\cdot,t),\nabla^{(\theta)}_{\mathcal{D}}u(\boldsymbol{x},t)\right)\cdot\nabla^{(0)}_{\mathcal{D}}v(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t,$$

$$T_3^{(m)}=\int_0^T\int_\Omega f(\boldsymbol{x},t)\Pi^{(0)}_{\mathcal{D}}v(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t,$$

and

$$T_4^{(m)}=\int_0^T\int_{\partial\Omega}g(\boldsymbol{x},t)\mathbb{T}^{(0)}_{\mathcal{D}}v(\boldsymbol{x},t)\mathrm{d}\gamma(\boldsymbol{x})\mathrm{d}t.$$

Accounting for $v^{(N)}=0$, the discrete integrate-by-parts formula (D.15) gives

$$T_1^{(m)}=-\sum_{n=0}^{N-1}\int_\Omega\Pi_{\mathcal{D}}u^{(n+1)}(\boldsymbol{x})\left[\Pi_{\mathcal{D}}v^{(n+1)}(\boldsymbol{x})-\Pi_{\mathcal{D}}v^{(n)}(\boldsymbol{x})\right]\mathrm{d}\boldsymbol{x}$$

$$
\begin{aligned}
&- \int_\Omega \Pi_{\mathcal{D}} u^{(0)}(\boldsymbol{x}) \Pi_{\mathcal{D}} v^{(0)}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \\
&= - \int_0^T \int_\Omega \Pi_{\mathcal{D}}^{(1)} u(\boldsymbol{x},t) \delta_{\mathcal{D}} v(\boldsymbol{x},t) \mathrm{d}\boldsymbol{x}\mathrm{d}t - \int_\Omega \Pi_{\mathcal{D}} \mathcal{I}_{\mathcal{D}} u_{\mathrm{ini}}(\boldsymbol{x}) \Pi_{\mathcal{D}}^{(0)} v(\boldsymbol{x},0) \mathrm{d}\boldsymbol{x}.
\end{aligned}
$$

The strong convergences (4.10c) and (4.10b) of v_m and the weak convergence in $L^2(\Omega \times (0,T))$ of $\Pi_{\mathcal{D}_m}^{(1)} u_m$ thus ensure that, as $m \to \infty$,

$$
T_1^{(m)} \to - \int_0^T \int_\Omega \overline{u}(\boldsymbol{x},t) \partial_t \overline{v}(\boldsymbol{x},t) \mathrm{d}\boldsymbol{x}\mathrm{d}t - \int_\Omega u_{\mathrm{ini}}(\boldsymbol{x}) \overline{v}(\boldsymbol{x},0) \mathrm{d}\boldsymbol{x}. \tag{5.87}
$$

Owing to the weak convergence of $\mathcal{A}_{\mathcal{D}_m}$ and the strong convergence (4.7b) of $\nabla_{\mathcal{D}_m}^{(0)} v_m$, as $m \to \infty$,

$$
T_2^{(m)} \to \int_0^T \int_\Omega \boldsymbol{A}(\boldsymbol{x},t) \cdot \nabla \overline{v}(\boldsymbol{x},t) \mathrm{d}\boldsymbol{x}\mathrm{d}t. \tag{5.88}
$$

Finally, by (4.7a) and (4.8), as $m \to \infty$,

$$
\begin{aligned}
T_3^{(m)} &\to \int_0^T \int_\Omega f(\boldsymbol{x},t) \overline{v}(\boldsymbol{x},t) \mathrm{d}\boldsymbol{x}\mathrm{d}t, \quad \text{and} \\
T_4^{(m)} &\to \int_0^T \int_{\partial\Omega} g(\boldsymbol{x},t) \gamma \overline{v}(\boldsymbol{x},t) \mathrm{d}\gamma(\boldsymbol{x}) \mathrm{d}t.
\end{aligned} \tag{5.89}
$$

Using (5.87)–(5.89) we can pass to the limit in $T_1^{(m)} + T_2^{(m)} = T_3^{(m)} + T_4^{(m)}$ to see that

$$
\begin{aligned}
&- \int_0^T \int_\Omega \overline{u}(\boldsymbol{x},t) \partial_t \overline{v}(\boldsymbol{x},t) \mathrm{d}\boldsymbol{x}\mathrm{d}t - \int_\Omega u_{\mathrm{ini}}(\boldsymbol{x}) \overline{v}(\boldsymbol{x},0) \mathrm{d}\boldsymbol{x} \\
&+ \int_0^T \int_\Omega \boldsymbol{A}(\boldsymbol{x},t) \cdot \nabla \overline{v}(\boldsymbol{x},t) \mathrm{d}\boldsymbol{x}\mathrm{d}t \\
&\quad = \int_0^T \int_\Omega f(\boldsymbol{x},t) \overline{v}(\boldsymbol{x},t) \mathrm{d}\boldsymbol{x}\mathrm{d}t + \int_0^T \int_{\partial\Omega} g(\boldsymbol{x},t) \gamma \overline{v}(\boldsymbol{x},t) \mathrm{d}\gamma(\boldsymbol{x}) \mathrm{d}t.
\end{aligned}
$$

This holds for all $\overline{v} \in C^1([0,T]; W^{1,p}(\Omega) \cap L^2(\Omega))$ such that $\overline{v}(\cdot,T) = 0$. By a density argument similar to the one used to prove the equivalence of (5.3) and (5.4), or of (5.75) and (5.78), we infer that $\overline{u} \in L^p(0,T; W^{1,p}(\Omega)) \cap C([0,T]; L^2(\Omega))$, $\partial_t \overline{u} \in L^{p'}(0,T; (W^{1,p}(\Omega))')$, $\overline{u}(\cdot,0) = u_{\mathrm{ini}}$ and, for all $\overline{v} \in L^p(0,T; W^{1,p}(\Omega))$,

$$
\begin{aligned}
&\int_0^T \langle \partial_t \overline{u}(\cdot,t), \overline{v}(\cdot,t) \rangle_{(W^{1,p}(\Omega))', W^{1,p}(\Omega)} \mathrm{d}t + \int_0^T \int_\Omega \boldsymbol{A}(\boldsymbol{x},t) \cdot \nabla \overline{v}(\boldsymbol{x},t) \mathrm{d}\boldsymbol{x}\mathrm{d}t \\
&\quad = \int_0^T \int_\Omega f(\boldsymbol{x},t) \overline{v}(\boldsymbol{x},t) \mathrm{d}\boldsymbol{x}\mathrm{d}t + \int_0^T \int_{\partial\Omega} g(\boldsymbol{x},t) \gamma \overline{v}(\boldsymbol{x},t) \mathrm{d}\gamma(\boldsymbol{x}) \mathrm{d}t.
\end{aligned} \tag{5.90}
$$

It remains to prove that

$$\boldsymbol{A}(\boldsymbol{x},t) = \boldsymbol{a}(\boldsymbol{x},\overline{u}(\cdot,t),\nabla\overline{u}(\boldsymbol{x},t)), \text{ for a.e. } (\boldsymbol{x},t) \in \Omega \times (0,T). \tag{5.91}$$

The formula

$$\int_0^T \langle \partial_t \overline{u}(\cdot,t), \overline{u}(\cdot,t)\rangle_{(W^{1,p}(\Omega))',W^{1,p}(\Omega)} \mathrm{d}t = \frac{1}{2}\int_\Omega \overline{u}(\boldsymbol{x},T)^2 \mathrm{d}\boldsymbol{x} - \frac{1}{2}\int_\Omega \overline{u}(\boldsymbol{x},0)^2 \mathrm{d}\boldsymbol{x}$$

is justified by [67, Section 2.5.2] since $\overline{u} \in L^p(0,T;W^{1,p}(\Omega)\cap L^2(\Omega))$ and $\partial_t\overline{u} \in L^{p'}(0,T;(W^{1,p}(\Omega))')$. Writing (5.90) with $\overline{v} = \overline{u}$ thus yields

$$\begin{aligned}\frac{1}{2}\int_\Omega \overline{u}(\boldsymbol{x},T)^2\mathrm{d}\boldsymbol{x} - \frac{1}{2}\int_\Omega u_{\mathrm{ini}}(\boldsymbol{x})^2\mathrm{d}\boldsymbol{x} + \int_0^T\int_\Omega \boldsymbol{A}(\boldsymbol{x},t)\cdot\nabla\overline{u}(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t \\ = \int_0^T\int_\Omega f(\boldsymbol{x},t)\overline{u}(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t + \int_0^T\int_{\partial\Omega} g(\boldsymbol{x},t)\gamma\overline{u}(\boldsymbol{x},t)\mathrm{d}\gamma(\boldsymbol{x})\mathrm{d}t.\end{aligned} \tag{5.92}$$

Relation (5.81) with $k = N$ yields

$$\begin{aligned}&\frac{1}{2}\int_\Omega (\Pi_{\mathcal{D}}^{(1)}u(\boldsymbol{x},T))^2\mathrm{d}\boldsymbol{x} \\ &+ \int_0^T\int_\Omega \boldsymbol{a}\left(\boldsymbol{x},\Pi_{\mathcal{D}}^{(\theta)}u(\cdot,t),\nabla_{\mathcal{D}}^{(\theta)}u(\boldsymbol{x},t)\right)\cdot\nabla_{\mathcal{D}}^{(\theta)}u(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t \\ &\quad\le \frac{1}{2}\int_\Omega (\Pi_{\mathcal{D}}\mathcal{I}_{\mathcal{D}}u_{\mathrm{ini}}(\boldsymbol{x}))^2\mathrm{d}\boldsymbol{x} + \int_0^T\int_\Omega f(\boldsymbol{x},t)\Pi_{\mathcal{D}}^{(\theta)}u(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t \\ &\qquad + \int_0^T\int_{\partial\Omega} g(\boldsymbol{x},t)\mathbb{T}_{\mathcal{D}}^{(\theta)}u(\boldsymbol{x},t)\mathrm{d}\gamma(\boldsymbol{x})\mathrm{d}t.\end{aligned} \tag{5.93}$$

Recall that $\Pi_{\mathcal{D}_m}^{(\theta)}u_m \to \overline{u}$ strongly in $L^p(\Omega\times(0,T))$, that $\mathbb{T}_{\mathcal{D}_m}^{(\theta)}u_m \to \gamma\overline{u}$ weakly in $L^p(\partial\Omega\times(0,T))$ and, by space–time consistency, that $\Pi_{\mathcal{D}_m}\mathcal{I}_{\mathcal{D}_m}u_{\mathrm{ini}} \to u_{\mathrm{ini}}$ in $L^2(\Omega)$. Moving the first term of (5.93) into the right-hand side, taking the superior limit of the resulting inequality and using (5.86) therefore leads to

$$\begin{aligned}&\limsup_{m\to\infty}\int_0^T\int_\Omega \boldsymbol{a}\left(\boldsymbol{x},\Pi_{\mathcal{D}}^{(\theta)}u(\cdot,t),\nabla_{\mathcal{D}}^{(\theta)}u(\boldsymbol{x},t)\right)\cdot\nabla_{\mathcal{D}}^{(\theta)}u(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t \\ &\quad\le -\liminf_{m\to\infty}\frac{1}{2}\int_\Omega (\Pi_{\mathcal{D}}^{(1)}u(\boldsymbol{x},T))^2\mathrm{d}\boldsymbol{x}\mathrm{d}t + \frac{1}{2}\int_\Omega u_{\mathrm{ini}}(\boldsymbol{x})^2\mathrm{d}\boldsymbol{x} \\ &\qquad + \int_0^T\int_\Omega f(\boldsymbol{x},t)\overline{u}(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t + \int_0^T\int_{\partial\Omega} g(\boldsymbol{x},t)\gamma\overline{u}(\boldsymbol{x},t)\mathrm{d}\gamma(\boldsymbol{x})\mathrm{d}t \\ &\quad\le -\frac{1}{2}\int_\Omega \overline{u}(\boldsymbol{x},T)^2\mathrm{d}\boldsymbol{x}\mathrm{d}t + \frac{1}{2}\int_\Omega u_{\mathrm{ini}}(\boldsymbol{x})^2\mathrm{d}\boldsymbol{x}\end{aligned}$$

$$+\int_0^T\int_\Omega f(\boldsymbol{x},t)\overline{u}(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t+\int_0^T\int_{\partial\Omega} g(\boldsymbol{x},t)\gamma\overline{u}(\boldsymbol{x},t)\mathrm{d}\gamma(\boldsymbol{x})\mathrm{d}t.$$

Relation (5.92) then yields

$$\limsup_{m\to\infty}\int_0^T\int_\Omega \boldsymbol{a}\left(\boldsymbol{x},\Pi_{\mathcal{D}_m}^{(\theta)}u_m(\cdot,t),\nabla_{\mathcal{D}_m}^{(\theta)}u_m(\boldsymbol{x},t)\right)\cdot\nabla_{\mathcal{D}_m}^{(\theta)}u_m(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t$$
$$\leq\int_0^T\int_\Omega \boldsymbol{A}(\boldsymbol{x},t)\cdot\nabla\overline{u}(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t. \quad (5.94)$$

It is now possible to apply Minty's trick. Consider, for $\boldsymbol{G}\in L^p(0,T;L^p(\Omega))^d$, the quantity

$$\int_0^T\int_\Omega\left[\boldsymbol{a}\left(\boldsymbol{x},\Pi_{\mathcal{D}}^{(\theta)}u(\cdot,t),\nabla_{\mathcal{D}}^{(\theta)}u(\boldsymbol{x},t)\right)-\boldsymbol{a}\left(\boldsymbol{x},\Pi_{\mathcal{D}}^{(\theta)}u(\cdot,t),\boldsymbol{G}(\boldsymbol{x},t)\right)\right]$$
$$\cdot\left[\nabla_{\mathcal{D}}^{(\theta)}u(\boldsymbol{x},t)-\boldsymbol{G}(\boldsymbol{x},t)\right]\mathrm{d}\boldsymbol{x}\mathrm{d}t\geq 0. \quad (5.95)$$

Since $\Pi_{\mathcal{D}_m}^{(\theta)}u_m\to\overline{u}$ strongly in $L^p(0,T;L^p(\Omega))$, up to a subsequence we can assume that $\Pi_{\mathcal{D}_m}^{(\theta)}u_m(t)\to\overline{u}(t)$ strongly in $L^p(\Omega)$ for a.e. $t\in(0,T)$. Assumptions (2.85a) and (2.85d) and the dominated convergence theorem then show that $\boldsymbol{a}(\cdot,\Pi_{\mathcal{D}}^{(\theta)}u,\boldsymbol{G})\to\boldsymbol{a}(\cdot,\overline{u},\boldsymbol{G})$ strongly in $L^{p'}(\Omega\times(0,T))^d$. Developing (5.95), all the terms except one pass to the limit by "weak-strong" convergence (cf. Lemma D.8). For the only "weak-weak" limit, apply (5.94) and, taking the superior limit as $m\to\infty$, write

$$\int_0^T\int_\Omega\left[\boldsymbol{A}(\boldsymbol{x},t)-\boldsymbol{a}\left(\boldsymbol{x},\overline{u}(\cdot,t),\boldsymbol{G}(\boldsymbol{x},t)\right)\right]\cdot\left[\nabla\overline{u}(\boldsymbol{x},t)-\boldsymbol{G}(\boldsymbol{x},t)\right]\mathrm{d}\boldsymbol{x}\mathrm{d}t\geq 0.$$

In a similar way as in Step 2 of the proof of Theorem 2.44, take then $\boldsymbol{G}=\nabla\overline{u}+\alpha\boldsymbol{\varphi}$ for $\alpha\in\mathbb{R}$ and $\boldsymbol{\varphi}\in L^p(0,T;L^p(\Omega))^d$, divide by α and let $\alpha\to 0$. This gives

$$\int_0^T\int_\Omega\left[\boldsymbol{A}(\boldsymbol{x},t)-\boldsymbol{a}\left(\boldsymbol{x},\overline{u}(\cdot,t),\nabla\overline{u}(\boldsymbol{x},t)\right)\right]\cdot\boldsymbol{\varphi}(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}=0,$$

which shows that (5.91) holds. The proof that $\overline{u}$ is a weak solution to (5.75) is therefore complete. ∎

Proof of Theorem 5.21.
Step 1: a preliminary result.
Take $(s_m)_{m\in\mathbb{N}}\subset[0,T]$ that converges to some $s\in[0,T]$. Since $\Pi_{\mathcal{D}_m}^{(\theta)}u_m\to\overline{u}$ strongly in $L^p(\Omega\times(0,T))$, as in Step 3 of the proof of Theorem 5.19, Assumptions (2.85a) and (2.85d) show that $(\mathbf{1}_{[0,s_m]}\boldsymbol{a}(\boldsymbol{x},\Pi_{\mathcal{D}_m}^{(\theta)}u_m,\nabla\overline{u}))_{m\in\mathbb{N}}$ converges strongly in $L^{p'}(\Omega\times(0,T))^d$. The weak convergence of $(\nabla_{\mathcal{D}_m}^{(\theta)}u_m)_{m\in\mathbb{N}}$ to $\nabla\overline{u}$ in $L^p(\Omega\times(0,T))^d$ then yields

$$\int_0^{s_m}\int_\Omega \boldsymbol{a}\left(\boldsymbol{x}, \Pi^{(\theta)}_{\mathcal{D}_m} u_m(\cdot,t), \nabla\overline{u}(\boldsymbol{x},t)\right) \cdot \left[\nabla^{(\theta)}_{\mathcal{D}_m} u_m(\boldsymbol{x},t) - \nabla\overline{u}(\boldsymbol{x},t)\right]\mathrm{d}\boldsymbol{x}\mathrm{d}t \to 0. \quad (5.96)$$

Write (5.95) with s_m instead of T and $\nabla\overline{u}$ instead of $\mathbf{G}$, and develop the terms. Using (5.96), the weak convergence $\boldsymbol{a}(\cdot, \Pi^{(\theta)}_{\mathcal{D}_m} u_m, \nabla^{(\theta)}_{\mathcal{D}_m} u_m) \to \boldsymbol{a}(\cdot, \overline{u}, \nabla\overline{u})$ in $L^{p'}(\Omega\times(0,T))$ (see (5.91)), and the strong convergence $\mathbf{1}_{[0,s_m]}\nabla\overline{u} \to \mathbf{1}_{[0,s]}\nabla\overline{u}$ in $L^p(\Omega\times(0,T))$, we obtain

$$\liminf_{m\to\infty}\int_0^{s_m}\int_\Omega \boldsymbol{a}\left(\boldsymbol{x}, \Pi^{(\theta)}_{\mathcal{D}_m} u_m(\cdot,t), \nabla^{(\theta)}_{\mathcal{D}_m} u_m(\boldsymbol{x},t)\right)\cdot\nabla^{(\theta)}_{\mathcal{D}_m} u_m(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t \geq \int_0^s\int_\Omega \boldsymbol{a}\left(\boldsymbol{x}, \overline{u}(\cdot,t), \nabla\overline{u}(\boldsymbol{x},t)\right)\cdot\nabla\overline{u}(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t. \quad (5.97)$$

Step 2: proof of the uniform-in-time strong in $L^2(\Omega)$ convergences.
Let $s\in[0,T]$ and $k(s)$ such that $s\in(t^{(k(s))}, t^{(k(s)+1)}]$. Apply (5.81) to $k = k(s)+1$ to write

$$\begin{aligned}\frac{1}{2}&\left\|\Pi^{(1)}_{\mathcal{D}_m} u_m(s)\right\|^2_{L^2(\Omega)} \\ &+ \int_0^s\int_\Omega \boldsymbol{a}(\boldsymbol{x}, \Pi^{(\theta)}_{\mathcal{D}_m} u_m(\cdot,t), \nabla^{(\theta)}_{\mathcal{D}_m} u_m(\boldsymbol{x},t))\cdot\nabla^{(\theta)}_{\mathcal{D}_m} u_m(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t \\ &\leq \frac{1}{2}\left\|\Pi_{\mathcal{D}_m}\mathcal{I}_{\mathcal{D}_m} u_{\mathrm{ini}}\right\|^2_{L^2(\Omega)} + \int_0^s\int_\Omega f(\boldsymbol{x},t)\Pi_{\mathcal{D}_m} u_m(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t \\ &\quad + \int_0^s\int_{\partial\Omega} g(\boldsymbol{x},t)\mathbb{T}_{\mathcal{D}_m} u_m(\boldsymbol{x},t)\mathrm{d}\gamma(\boldsymbol{x})\mathrm{d}t + \rho(\delta t_{\mathcal{D}_m}) \quad (5.98)\end{aligned}$$

where $\rho(\delta t_{\mathcal{D}_m}) \to 0$ as $\delta t_{\mathcal{D}_m} \to 0$ (all time integrals should be up to $t^{(k(s)+1)}$, but we used the non-negativity of the integrand involving $\boldsymbol{a}$ to limit its integral to s, and ρ is the quantity that includes the remaining parts of the integrals in the right-hand side, estimated using to (5.80)).
The proof of the uniform convergence of $(\Pi^{(1)}_{\mathcal{D}_m} u_m)_{m\in\mathbb{N}}$ is done by invoking Lemma C.13. As in Step 1, take $(s_m)_{m\in\mathbb{N}}\subset[0,T]$ that converges to some $s\in[0,T]$. We want to show that $\Pi^{(1)}_{\mathcal{D}_m} u_m(s_m) \to \overline{u}(s)$ in $L^2(\Omega)$. Apply (5.98) with $s = s_m$, move the second term to the right-hand side, and take the superior limit as $m\to\infty$. Relation (5.97) and the strong (resp. weak) convergence of $\Pi^{(\theta)}_{\mathcal{D}_m} u_m$ (resp. $\mathbb{T}^{(\theta)}_{\mathcal{D}_m} u_m$) enable us to pass to the limit in all the terms except the first one. Owing to (5.77), this gives

$$\begin{aligned}\limsup_{m\to\infty}\frac{1}{2}&\left\|\Pi^{(1)}_{\mathcal{D}_m} u_m(s_m)\right\|^2_{L^2(\Omega)} \\ &\leq -\int_0^s\int_\Omega \boldsymbol{a}(\boldsymbol{x}, \overline{u}(\cdot,t), \nabla\overline{u}(\boldsymbol{x},t))\cdot\nabla\overline{u}(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t\end{aligned}$$

$$+\frac{1}{2}\left\|u_{\mathrm{ini}}\right\|_{L^2(\Omega)}^2+\int_0^s\int_\Omega f(\boldsymbol{x},t)\overline{u}(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t$$
$$+\int_0^s\int_{\partial\Omega} g(\boldsymbol{x},t)\gamma\overline{u}(\boldsymbol{x},t)\mathrm{d}\gamma(\boldsymbol{x})\mathrm{d}t=\frac{1}{2}\left\|\overline{u}(s)\right\|_{L^2(\Omega)}^2. \tag{5.99}$$

The uniform-in-time weak $L^2(\Omega)$ convergence of $(\Pi_{\mathcal{D}_m}^{(1)}u_m)_{m\in\mathbb{N}}$ towards $\overline{u}$ and Lemma C.13 show that $\Pi_{\mathcal{D}_m}^{(1)}u_m(s_m)\to\overline{u}(s)$ in $L^2(\Omega)$ weak. Owing to (5.99), this convergence is actually strong in $L^2(\Omega)$. Invoke again Lemma C.13 to conclude that $\sup_{t\in[0,T]}\|\Pi_{\mathcal{D}_m}^{(1)}u_m(t)-\overline{u}(t)\|_{L^2(\Omega)}\to 0$.

The strong convergence of $(\Pi_{\mathcal{D}_m}^{(\theta)}u_m)_{m\in\mathbb{N}}$ in the same sense follows immediately from the definition of these functions, the strong convergence of $(\Pi_{\mathcal{D}_m}^{(1)}u_m)_{m\in\mathbb{N}}$, and the continuity of $\overline{u}:[0,T]\to L^2(\Omega)$. Indeed, $\Pi_{\mathcal{D}_m}^{(\theta)}u_m(\cdot,t)$ is a convex combination of values of $\Pi_{\mathcal{D}_m}^{(1)}u_m$ at two times within distance $\delta t_{\mathcal{D}_m}$ of t. ∎

Proof of Theorem 5.22.
Using (5.94) and (5.96) with $s_m=T$,

$$\limsup_{m\to\infty}\int_0^T\int_\Omega\left[\boldsymbol{a}\left(\boldsymbol{x},\Pi_{\mathcal{D}_m}^{(\theta)}u_m,\nabla_{\mathcal{D}_m}^{(\theta)}u_m\right)-\boldsymbol{a}\left(\boldsymbol{x},\Pi_{\mathcal{D}_m}^{(\theta)}u_m,\nabla\overline{u}\right)\right]$$
$$\cdot\left[\nabla_{\mathcal{D}_m}^{(\theta)}u_m-\nabla\overline{u}\right]\mathrm{d}\boldsymbol{x}\mathrm{d}t\le 0.$$

This relation and the strict monotonicity of $\boldsymbol{a}$ enable us to conclude, as in Step 3 of the proof of Theorem 2.44, that $\nabla_{\mathcal{D}_m}^{(\theta)}u_m\to\nabla\overline{u}$ a.e. on $\Omega\times(0,T)$. From (5.94) and (5.91) we also infer

$$\limsup_{m\to\infty}\int_0^T\int_\Omega\boldsymbol{a}\left(\boldsymbol{x},\Pi_{\mathcal{D}_m}^{(\theta)}u_m,\nabla_{\mathcal{D}_m}^{(\theta)}u_m\right)\cdot\nabla_{\mathcal{D}_m}^{(\theta)}u_m\mathrm{d}\boldsymbol{x}\mathrm{d}t$$
$$\le\int_0^T\int_\Omega\boldsymbol{a}\left(\boldsymbol{x},\overline{u},\nabla\overline{u}\right)\cdot\nabla\overline{u}\mathrm{d}\boldsymbol{x}\mathrm{d}t.$$

Together with (5.97) (with $s_m=T$), this proves that this relation holds with a limit instead of a superior limit, and an equality instead of an inequality. The same technique as in Step 3 of the proof of Theorem 2.44 then yields the strong convergence of $\nabla_{\mathcal{D}_m}u_m$ to $\nabla\overline{u}$ in $L^p(\Omega\times(0,T))^d$. ∎

Remark 5.25 (Navier–Stokes equations)
The GDM has been adapted to the Stokes and Navier–Stokes equations in [74, 90], with similar convergence results as for the transient Leray–Lions type model.

6

Degenerate parabolic problems

In this chapter, we study the following generic non-linear parabolic model

$$\begin{aligned} &\partial_t \beta(\overline{u}) - \operatorname{div}\left(\Lambda(\boldsymbol{x})\nabla\zeta(\overline{u})\right) = f && \text{in } \Omega \times (0,T), \\ &\beta(\overline{u})(\boldsymbol{x},0) = \beta(u_{\text{ini}})(\boldsymbol{x}) && \text{in } \Omega, \\ &\zeta(\overline{u}) = 0 && \text{on } \partial\Omega \times (0,T), \end{aligned} \tag{6.1}$$

where β and ζ are non-decreasing. This model arises in various frameworks (see next section for precise hypotheses on the data). This model includes

1. Richards' model, setting $\zeta(s) = s$, which describes the flow of water in a heterogeneous anisotropic underground medium,
2. Stefan's model [23], setting $\beta(s) = s$, which arises in the study of a simplified heat diffusion in a melting medium.

The purpose of this chapter is to study the convergence of gradient schemes for (6.1). Although Richards' and Stefan's models are formally equivalent when β and ζ are strictly increasing (consider $\beta = \zeta^{-1}$ to pass from one model to the other), they change nature when these functions are allowed to have plateaux. Stefan's model can degenerate to an ODE (if ζ is constant on the range of the solution), and Richards' model can become a non-transient elliptic equation (if β is constant on this range). The techniques developed in this chapter also apply to the following more general non-linear PDE, which mixes (6.1) and Leray–Lions type operators as in Section 5.3:

$$\partial_t \beta(\overline{u}) - \operatorname{div} a\left(\boldsymbol{x}, \nu(u), \nabla\zeta(\overline{u})\right) = f, \tag{6.2}$$

where $\nu' = \beta'\zeta'$ [72].

The chapter is organised as follows. Section 6.1 is devoted to the assumptions and the notion of weak solution for the problem (6.1), which is reformulated using the notion of maximal monotone graph. Section 6.2 presents the gradient schemes (GSs) obtained by applying the gradient discretisation method (GDM) to the generic model (6.1). Based on estimates proved in Section 6.3,

Section 6.4 contains the convergence proof of these GSs. Section 6.5 is focused on a uniform-in-time convergence result. A uniqueness result, based on the existence of a solution to the adjoint problem, is given in Section 6.7.

6.1 The continuous problem

6.1.1 Hypotheses and notion of solution

We consider the evolution problem (6.1) under the following hypotheses.

- Ω is an open bounded connected subset of $\mathbb{R}^d$ ($d \in \mathbb{N}^\star$) and $T > 0$, (6.3a)
- $\zeta : \mathbb{R} \to \mathbb{R}$ is non-decreasing, Lipschitz continuous with Lipschitz constant $L_\zeta > 0$, $\zeta(0) = 0$ and, for some $M_0, M_1 > 0$, $|\zeta(s)| \geq M_0|s| - M_1$ for all $s \in \mathbb{R}$, (6.3b)
- $\beta : \mathbb{R} \to \mathbb{R}$ is non-decreasing, Lipschitz continuous with Lipschitz constant $L_\beta > 0$, and $\beta(0) = 0$, (6.3c)
- $\beta + \zeta$ is strictly increasing, (6.3d)
- $\Lambda : \Omega \to \mathcal{M}_d(\mathbb{R})$ is measurable and there exists $\overline{\lambda} \geq \underline{\lambda} > 0$ such that, for a.e. $\boldsymbol{x} \in \Omega$, $\Lambda(\boldsymbol{x})$ is symmetric with eigenvalues in $[\underline{\lambda}, \overline{\lambda}]$. (6.3e)
- $u_{\text{ini}} \in L^2(\Omega)$, $f \in L^2(\Omega \times (0,T))$. (6.3f)

Remark 6.1 (Common plateaux of ζ and β)
Hypothesis (6.3d) does not restrict the generality of the model. Indeed, if we only assume (6.3b)–(6.3c), and if there exist $s_1 < s_2$ such that $(\beta+\zeta)(s_1) = (\beta+\zeta)(s_2)$, then $[s_1, s_2]$ is a common plateau of β and ζ. Denoting by $\widetilde{\beta}$, $\widetilde{\zeta}$ and $\widetilde{\nu}$ the functions obtained from β and ζ by removing this common plateau (by a contraction of the s-ordinate), we see that u is a solution to (6.1) if and only if u is a solution of the same problem with β and ζ replaced with $\widetilde{\beta}$ and $\widetilde{\zeta}$.

The precise notion of solution to (6.1) that we consider is the following:

$$\begin{cases} \zeta(\overline{u}) \in L^2(0,T;H_0^1(\Omega)), \\ -\displaystyle\int_0^T \int_\Omega \beta(\overline{u})(\boldsymbol{x},t)\partial_t\overline{v}(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t - \int_\Omega \beta(u_{\text{ini}}(\boldsymbol{x}))\overline{v}(\boldsymbol{x},0)\mathrm{d}\boldsymbol{x} \\ \qquad + \displaystyle\int_0^T \int_\Omega \Lambda(\boldsymbol{x})\nabla\zeta(\overline{u})(\boldsymbol{x},t)\cdot\nabla\overline{v}(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t \\ \qquad = \displaystyle\int_0^T \int_\Omega f(\boldsymbol{x},t)\overline{v}(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t, \\ \qquad\qquad \forall \overline{v} \in L^2(0;T;H_0^1(\Omega)) \text{ such that } \partial_t\overline{v} \in L^2((0,T)\times\Omega) \\ \qquad\qquad \text{and } \overline{v}(\cdot,T) = 0. \end{cases} \tag{6.4}$$

Remark 6.2 (All the terms in (6.4) make sense). If $\overline{v}$ and $\partial_t \overline{v}$ belong to $L^2(0,T;L^2(\Omega))$, then $\overline{v} \in C([0,T];L^2(\Omega))$ (see [67]), and we can therefore impose the pointwise-in-time value of $\overline{v}(\cdot,T)$. Moreover, Assumptions (6.3b) and (6.3c) ensure that, if $\zeta(\overline{u}) \in L^2((0,T)\times\Omega)$, then $\overline{u}$ and $\beta(\overline{u})$ also belong to $L^2((0,T)\times\Omega)$. Hence, all the terms in (6.4) are well-defined.

The existence of a solution to this problem follows from the proof of convergence of the GS (see Remark 6.13). The uniqueness of this solution is proved in Section 6.7.

Theorem 6.3 (Existence and uniqueness of the weak solution). *Under Hypotheses (6.3), there exists a unique solution to (6.4).*

Remark 6.4. We will see in Corollary 6.15 that the solution to (6.4) enjoys additional regularity properties, and that (6.4) can be recast in a stronger form.

6.1.2 A maximal monotone operator viewpoint

Following [78], we show here that (6.1) can be recast in a maximal monotone operator framework. Let us first recall a few definitions.

- A multi-valued operator $\mathcal{T}$ with domain $\mathbb{R}$ is a function from $\mathbb{R}$ to the set $\mathcal{P}(\mathbb{R})$ of all subsets of $\mathbb{R}$.
- The graph of $\mathcal{T}$ is defined by $\mathrm{Gr}(\mathcal{T}) = \{(x,y) : x \in \mathbb{R}, y \in \mathcal{T}(x)\}$.
- The operator $\mathcal{T}$ is said to be monotone if, for any (x,y) and (x',y') in $\mathrm{Gr}(\mathcal{T})$, $(x'-x)(y'-y) \geq 0$.
- The operator $\mathcal{T}$ is said to be maximal monotone if it is monotone and, for any $(x,y) \in \mathbb{R}^2$ such that

$$\text{for all } (x',y') \in \mathrm{Gr}(\mathcal{T}), \text{ there holds } (x'-x)(y'-y) \geq 0, \tag{6.5}$$

  we have $(x,y) \in \mathrm{Gr}(\mathcal{T})$. Then the resolvent $\mathcal{R}(\mathcal{T}) = (\mathrm{Id} + \mathcal{T})^{-1}$ of the maximal monotone operator $\mathcal{T}$ is a single-valued function $\mathbb{R} \to \mathbb{R}$, is non-decreasing and Lipschitz continuous with Lipschitz constant 1.

We refer to [33] for more definitions and properties of maximal monotone operators.

Lemma 6.5 (Caracterisation of sublinear maximal monotone operators). *Let $\mathcal{T}$ be a multi-valued operator with domain $\mathbb{R}$. The following properties are equivalent:*

1. *$\mathcal{T}$ is a maximal monotone operator, $0 \in \mathcal{T}(0)$ and $\mathcal{T}$ is sublinear in the sense that there exist $T_1, T_2 \geq 0$ such that, for all $x \in \mathbb{R}$ and all $y \in \mathcal{T}(x)$, $|y| \leq T_1|x| + T_2$;*
2. *There exist ζ and β satisfying (6.3b) and (6.3c) such that the graph of $\mathcal{T}$ is given by $\mathrm{Gr}(\mathcal{T}) = \{(\zeta(s), \beta(s)), s \in \mathbb{R}\}$.*

Proof. *(2)⇒(1).* Clearly $0 = (\zeta(0), \beta(0)) \in \mathcal{T}(0)$. The monotonicity of $\mathcal{T}$ follows from the fact that ζ and β are non-decreasing. We now have to prove that $\mathcal{T}$ is maximal monotone. Let $(x, y) \in \mathbb{R}^2$ be such that (6.5) holds, which implies that, for all $s \in \mathbb{R}$, $(\zeta(s) - x)(\beta(s) - y) \geq 0$. By (6.3b) and (6.3c), the mapping $\beta + \zeta : \mathbb{R} \to \mathbb{R}$ is onto, so there exists $s \in \mathbb{R}$ such that

$$\beta(s) + \zeta(s) = x + y. \tag{6.6}$$

Then $\zeta(s) - x = y - \beta(s)$ and therefore $-(\beta(s) - y)^2 = (\zeta(s) - x)(\beta(s) - y) \geq 0$. This implies $\beta(s) = y$ and, combined with (6.6), $\zeta(s) = x$. Hence $(x, y) \in \mathrm{Gr}(\mathcal{T})$. The sub-linearity of $\mathcal{T}$ follows from $|\beta(s)| \leq L_\beta |s| \leq L_\beta(|\zeta(s)| + M_2)/M_1$.
(1)⇒(2). Set $\zeta = \mathcal{R}(\mathcal{T})$ and $\beta = \mathrm{Id} - \zeta$. These functions are non-decreasing and Lipschitz continuous with constant 1. By definition of the resolvent,

$$(x, y) \in \mathrm{Gr}(\mathcal{T}) \Leftrightarrow (x, x+y) \in \mathrm{Gr}(\mathrm{Id}+\mathcal{T}) \Leftrightarrow (x+y, x) \in \mathrm{Gr}(\zeta) \Leftrightarrow x = \zeta(x+y).$$

Since $\beta = \mathrm{Id} - \zeta$, setting $s = x + y$ shows that $(x, y) \in \mathrm{Gr}(\mathcal{T})$ is equivalent to $(x, y) = (\zeta(s), \beta(s))$. Since $0 \in T(0)$ this gives $\beta(0) = \zeta(0) = 0$. Finally, the existence of M_1 and M_2 in (6.3b) follows from the sublinearity of $\mathcal{T}$. If $(x, y) \in \mathrm{Gr}(\mathcal{T})$ then $|y| \leq T_1|x| + T_2$ and $x = \zeta(x + y)$, which gives $|x + y| \leq ((1 + T_1)|\zeta(x + y)| + T_2)$. ∎

Using this lemma, we recast (6.1) as

$$\begin{cases} \partial_t \mathcal{T}(z) - \mathrm{div}\,(\Lambda(\boldsymbol{x})\nabla z) = f & \text{in } \Omega \times (0, T), \\ \mathcal{T}(z)(\cdot, 0) = b_{\mathrm{ini}} & \text{in } \Omega, \\ z = 0 & \text{on } \partial\Omega \times (0, T) \end{cases} \tag{6.7}$$

where $b_{\mathrm{ini}} = \beta(u_{\mathrm{ini}}) \in L^2(\Omega)$. Hypotheses (6.3c) and (6.3b) are translated into:

$$\begin{aligned} &\mathcal{T} : \mathbb{R} \to \mathcal{P}(\mathbb{R}) \text{ is a maximal monotone operator, } 0 \in \mathcal{T}(0) \\ &\text{and } \mathcal{T} \text{ is sublinear: } \exists T_1, T_2 \geq 0 \text{ such that, for all } x \in \mathbb{R} \\ &\text{all } y \in \mathcal{T}(x),\ |y| \leq T_1|x| + T_2. \end{aligned} \tag{6.8}$$

Definition 6.6. *Let us assume* (6.3a), (6.3e), (6.3f) *and* (6.8). *Let* $z_{\mathrm{ini}} \in L^2(\Omega)$ *and* $b_{\mathrm{ini}} : \Omega \to \mathbb{R}$ *such that, for a.e.* $\boldsymbol{x} \in \Omega$, $b_{\mathrm{ini}}(\boldsymbol{x}) \in \mathcal{T}(z_{\mathrm{ini}}(\boldsymbol{x}))$. *A* solution *to* (6.7) *is a pair of functions* (z, b) *satisfying*

$$\begin{cases} z \in L^2(0, T; H_0^1(\Omega)), \\ b(\boldsymbol{x}, t) \in \mathcal{T}(z(\boldsymbol{x}, t)) \text{ for a.e. } (\boldsymbol{x}, t) \in \Omega \times (0, T), \\ -\displaystyle\int_0^T \int_\Omega b(\boldsymbol{x}, t)\partial_t \overline{v}(\boldsymbol{x}, t)\mathrm{d}\boldsymbol{x}\mathrm{d}t - \int_\Omega b_{\mathrm{ini}}(\boldsymbol{x})\overline{v}(\boldsymbol{x}, 0)\mathrm{d}\boldsymbol{x} \\ \qquad + \displaystyle\int_0^T \int_\Omega \Lambda(\boldsymbol{x})\nabla z(\boldsymbol{x}, t) \cdot \nabla \overline{v}(\boldsymbol{x}, t)\mathrm{d}x\mathrm{d}t \\ \qquad = \displaystyle\int_0^T \int_\Omega f(\boldsymbol{x}, t)\overline{v}(\boldsymbol{x}, t)\mathrm{d}\boldsymbol{x}\mathrm{d}t \\ \qquad\quad \forall \overline{v} \in L^2(0; T; H_0^1(\Omega)) \text{ such that } \partial_t \overline{v} \in L^2((0, T) \times \Omega) \\ \qquad\quad \text{and } \overline{v}(\cdot, T) = 0. \end{cases} \tag{6.9}$$

Remark 6.7. The sublinearity of $\mathcal{T}$ ensures that $b \in L^2(0,T;L^2(\Omega))$ and $b_{\rm ini} \in L^2(\Omega)$, since $z \in L^2(0,T;L^2(\Omega))$ and $z_{\rm ini} \in L^2(\Omega)$.

Given $(z_{\rm ini}, b_{\rm ini})$ as in Definition 6.6 and fixing $\zeta = \mathcal{R}(\mathcal{T})$ and $\beta = {\rm Id} - \zeta$ (as in the proof of Item 2 of Lemma 6.5), we can find a measurable $u_{\rm ini}$ such that $z_{\rm ini} = \zeta(u_{\rm ini})$ and $b_{\rm ini} = \beta(u_{\rm ini})$. the estimate $|z_{\rm ini}| \geq M_0|u_{\rm ini}| - M_1$ ensures that $u_{\rm ini} \in L^2(\Omega)$. These ζ, β and $u_{\rm ini}$ being fixed, the existence and uniqueness of the solution to (6.4) (Theorem 6.3) gives the existence and uniqueness of the solution to (6.9). This solution satisfies that $z = \zeta(\overline{u})$ and $b = \beta(\overline{u})$, where $\overline{u}$ is the unique solution to (6.4).

6.2 Gradient scheme

Let $p = 2$ and $\mathcal{D}_T = (\mathcal{D}, \mathcal{I}_\mathcal{D}, (t^{(n)})_{n=0,\ldots,N})$ be a space–time gradient discretisation for homogeneous Dirichlet boundary conditions, in the sense of Definition 4.1. Assume that $\mathcal{D}$ has the piecewise constant reconstruction property in the sense of Definition 2.12. We take $\theta = 1$ in (4.2), which means that an implicit time-stepping is considered. We recall the corresponding notations $\Pi_\mathcal{D}^{(1)}$ and $\nabla_\mathcal{D}^{(1)}$.

Formally integrating (6.4) by parts in time, we obtain a new formulation of (6.1) (see (6.26)). The GDM applied to (6.4) leads to a GS which merely consists in using, in this new formulation, the discrete space and mappings of the GD. The GS is therefore: seek a family $(u^{(n)})_{n=0,\ldots,N} \subset X_{\mathcal{D},0}$ such that

$$\left\{\begin{array}{l} u^{(0)} = \mathcal{I}_\mathcal{D} u_{\rm ini} \text{ and, for all } v = (v^{(n)})_{n=1,\ldots,N} \subset X_{\mathcal{D},0}, \\ \displaystyle\int_0^T \int_\Omega \Big[\delta_\mathcal{D}\beta(u)(\boldsymbol{x},t)\Pi_\mathcal{D}^{(1)} v(\boldsymbol{x},t) + \Lambda(\boldsymbol{x})\nabla_\mathcal{D}^{(1)}\zeta(u)(\boldsymbol{x},t)\cdot\nabla_\mathcal{D}^{(1)} v(\boldsymbol{x},t)\Big]{\rm d}\boldsymbol{x}{\rm d}t \\ \qquad\displaystyle = \int_0^T \int_\Omega f(\boldsymbol{x},t)\Pi_\mathcal{D}^{(1)} v(\boldsymbol{x},t){\rm d}\boldsymbol{x}{\rm d}t. \end{array}\right. \tag{6.10}$$

We recall the definition, in Remark 2.13, of $\zeta(u)$ and $\beta(u)$, which is coherent with $\Pi_\mathcal{D}$ since this reconstruction is piecewise constant.

6.3 Estimates on the approximate solution

As it is usual in the study of numerical methods for PDE with strong nonlinearities or without regularity assumptions on the data, everything starts with *a priori* estimates.

Lemma 6.8 ($L^\infty(0,T;L^2(\Omega))$ estimate and discrete $L^2(0,T;H_0^1(\Omega))$ estimate). *Under Assumptions (6.3), let $\mathcal{D}_T$ be a space–time GD for homogeneous Dirichlet boundary conditions, in the sense of Definition 4.1. Assume*

that the underlying spatial discretisation has a piecewise constant reconstruction in the sense of Definition 2.12, and that u is a solution to the corresponding GS (6.10). Let $\eta : \mathbb{R} \to \mathbb{R}$ be defined by

$$\forall s \in \mathbb{R}\,, \quad \eta(s) = \int_0^s \zeta(q)\beta'(q)\mathrm{d}q. \tag{6.11}$$

Let $T_0 \in (0, T]$ and denote by $k = 1, \ldots, N$ the index such that $T_0 \in (t^{(k-1)}, t^{(k)}]$. Then

$$\begin{aligned}\int_\Omega \Pi_{\mathcal{D}}^{(1)}\eta(u)(\boldsymbol{x}, T_0)\mathrm{d}\boldsymbol{x} + \int_0^{T_0}\int_\Omega \Lambda(\boldsymbol{x})\nabla_{\mathcal{D}}^{(1)}\zeta(u)(\boldsymbol{x},t)\cdot\nabla_{\mathcal{D}}^{(1)}\zeta(u)(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t \\ \leq \int_\Omega \Pi_{\mathcal{D}}\eta(\mathcal{I}_{\mathcal{D}}u_{\mathrm{ini}})(\boldsymbol{x})\mathrm{d}\boldsymbol{x} + \int_0^{t^{(k)}}\int_\Omega f(\boldsymbol{x},t)\Pi_{\mathcal{D}}^{(1)}\zeta(u)(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t.\end{aligned} \tag{6.12}$$

Consequently, there exists $C_1 > 0$, depending only on L_β, L_ζ, $C_P \geq C_{\mathcal{D}}$ (see Definition 2.2), $C_{\mathrm{ini}} \geq \|\Pi_{\mathcal{D}}\mathcal{I}_{\mathcal{D}}u_{\mathrm{ini}}\|_{L^2(\Omega)}$, f and $\underline{\lambda}$ such that

$$\begin{aligned}&\sup_{t\in[0,T]}\left\|\Pi_{\mathcal{D}}^{(1)}\eta(u)(t)\right\|_{L^1(\Omega)} \leq C_1\,,\ \left\|\nabla_{\mathcal{D}}^{(1)}\zeta(u)\right\|_{L^2(\Omega\times(0,T))^d} \leq C_1 \\ &\textit{and}\ \sup_{t\in[0,T]}\left\|\Pi_{\mathcal{D}}^{(1)}\beta(u)(t)\right\|_{L^2(\Omega)} \leq C_1.\end{aligned} \tag{6.13}$$

Proof. Let us first remark that, for all $a, b \in \mathbb{R}$, an integration by parts gives

$$\eta(b) - \eta(a) = \int_a^b \zeta(q)\beta'(q)\mathrm{d}q = \zeta(b)(\beta(b) - \beta(a)) - \int_a^b \zeta'(q)(\beta(q) - \beta(a))\mathrm{d}q.$$

Since $\int_a^b \zeta'(q)(\beta(q) - \beta(a))\mathrm{d}q \geq 0$ (as ζ and β are non-decreasing), we get

$$\eta(b) - \eta(a) \leq \zeta(b)(\beta(b) - \beta(a)). \tag{6.14}$$

Using Remark 2.13 (consequence of the definition 2.12 of piecewise constant reconstruction) and (6.14), we infer that for any $n = 0, \ldots, N-1$, any $t \in (t^{(n)}, t^{(n+1)}]$,

$$\begin{aligned}\delta_{\mathcal{D}}\beta(u)(t)\Pi_{\mathcal{D}}\zeta(u^{(n+1)}) &= \frac{1}{\eth^{(n+\frac{1}{2})}}\left(\beta(\Pi_{\mathcal{D}}u^{(n+1)}) - \beta(\Pi_{\mathcal{D}}u^{(n)})\right)\zeta(\Pi_{\mathcal{D}}u^{(n+1)}) \\ &\geq \frac{1}{\eth^{(n+\frac{1}{2})}}\left(\eta(\Pi_{\mathcal{D}}u^{(n+1)}) - \eta(\Pi_{\mathcal{D}}u^{(n)})\right).\end{aligned}$$

Hence, taking $v = (\zeta(u^{(0)}), \zeta(u^{(1)}), \ldots, \zeta(u^{(k)}), 0, \ldots, 0) \subset X_{\mathcal{D},0}$ in (6.10), we find

$$\int_\Omega \eta(\Pi_{\mathcal{D}}^{(1)}u(\boldsymbol{x}, t^{(k)}))\mathrm{d}\boldsymbol{x} + \int_0^{t^{(k)}}\int_\Omega \Lambda(\boldsymbol{x})\nabla_{\mathcal{D}}^{(1)}\zeta(u)(\boldsymbol{x},t)\cdot\nabla_{\mathcal{D}}^{(1)}\zeta(u)(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t$$

$$\leq \int_\Omega \eta(\Pi_\mathcal{D} u^{(0)}(\boldsymbol{x}))\mathrm{d}\boldsymbol{x} + \int_0^{t^{(k)}} \int_\Omega f(\boldsymbol{x},t)\Pi_\mathcal{D}^{(1)}\zeta(u)(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t. \quad (6.15)$$

Equation (6.12) is a straightforward consequence of this estimate, of the relation $\Pi_\mathcal{D}^{(1)}u(\cdot,T_0) = \Pi_\mathcal{D}^{(1)}u(\cdot,t^{(k)})$ (see (4.2)) and of the fact that the integrand involving Λ is non-negative on $[T_0, t^{(k)}]$.
Using the Young inequality (D.9), we write

$$\int_0^{t^{(k)}} \int_\Omega f(\boldsymbol{x},t)\Pi_\mathcal{D}^{(1)}\zeta(u)(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t$$
$$\leq \frac{C_\mathcal{D}^2}{2\underline{\lambda}}\|f\|^2_{L^2(\Omega\times(0,t^{(k)}))} + \frac{\underline{\lambda}}{2C_\mathcal{D}^2}\|\Pi_\mathcal{D}^{(1)}\zeta(u)\|^2_{L^2(\Omega\times(0,t^{(k)}))}. \quad (6.16)$$

We also notice that

$$0 \leq \eta(s) \leq L_\beta L_\zeta \int_0^s q\mathrm{d}q = L_\beta L_\zeta \frac{s^2}{2}, \quad (6.17)$$

so that

$$\left\|\eta(\Pi_\mathcal{D}^{(1)}u(\cdot,T_0))\right\|_{L^1(\Omega)} = \int_\Omega \eta(\Pi_\mathcal{D}^{(1)}u(x,T_0))\mathrm{d}\boldsymbol{x}$$

and

$$\left\|\eta(\Pi_\mathcal{D} u^{(0)})\right\|_{L^1(\Omega)} = \|\eta(\Pi_\mathcal{D}\mathcal{I}_\mathcal{D} u_{\mathrm{ini}})\|_{L^1(\Omega)} \leq \frac{L_\beta L_\zeta}{2}\|\Pi_\mathcal{D}\mathcal{I}_\mathcal{D} u_{\mathrm{ini}}\|^2_{L^2(\Omega)}.$$

The first two estimates in (6.13) therefore follow from (6.15), (6.16), Assumption (6.3e) on Λ, and the definition (2.1) of $C_\mathcal{D}$.
Let us now prove that the uniform-in-time $L^1(\Omega)$ estimate on $\Pi_\mathcal{D}^{(1)}\eta(u)$ implies the uniform-in-time $L^2(\Omega)$ estimate on $\Pi_\mathcal{D}^{(1)}\beta(u) = \beta(\Pi_\mathcal{D}^{(1)}u)$. Owing to (6.3b), for all $s \geq 0$ there holds $\zeta(s) \geq M_0 s - M_1 \geq \frac{M_0}{L_\beta}\beta(s) - M_1$. Hence, using the Young inequality,

$$\begin{aligned}\eta(s) = \int_0^s \zeta(q)\beta'(q)\mathrm{d}q &\geq \frac{M_0}{L_\beta}\int_0^s \beta(q)\beta'(q)\mathrm{d}q - M_1\int_0^s \beta'(q)\mathrm{d}q \\ &= \frac{M_0}{2L_\beta}\beta(s)^2 - M_1\beta(s) \\ &\geq \frac{M_0}{2L_\beta}\beta(s)^2 - \frac{M_0}{4L_\beta}\beta(s)^2 - \frac{L_\beta M_1^2}{M_0}.\end{aligned}$$

For $s \leq 0$, we use $-\zeta(s) \geq -M_0 s - M_1 \geq -\frac{M_0}{L_\beta}\beta(s) - M_1$ to infer the same estimate. Therefore,

$$\forall s \in \mathbb{R}, \quad \frac{M_0}{4L_\beta}\beta(s)^2 - \frac{L_\beta M_1^2}{M_0} \leq \eta(s). \quad (6.18)$$

Making $s = \Pi_{\mathcal{D}}^{(1)} u$ in this inequality and using the uniform-in-time $L^1(\Omega)$ estimate on $\eta(\Pi_{\mathcal{D}}^{(1)} u)$, we deduce the uniform-in-time $L^2(\Omega)$ estimate on $\beta(\Pi_{\mathcal{D}}^{(1)} u)$ stated in (6.13). ∎

Corollary 6.9 (Existence of a solution to the GS). *Under Assumptions (6.3), let $\mathcal{D}_T$ be a space–time GD for homogeneous Dirichlet boundary conditions, in the sense of Definition 4.1. Assume that the underlying spatial discretisation has a piecewise constant reconstruction in the sense of Definition 2.12. Then there exists at least a solution to the GS (6.10).*

Proof. For $\rho \in [0,1]$ we let $\beta_\rho(u) = \rho u + (1-\rho)\beta(u)$ and $\zeta_\rho(u) = \rho u + (1-\rho)\zeta(u)$. It is clear that β_ρ and ζ_ρ satisfy the same assumptions as β and ζ for some L_β and M_0, M_1 not depending on ρ. We can therefore apply Lemma 6.8 to see that there exists C_2, not depending on ρ, such that any solution u_ρ to (6.10) with $\beta = \beta_\rho$ and $\zeta = \zeta_\rho$ satisfies

$$\left\| \nabla_{\mathcal{D}}^{(1)} \zeta_\rho(u_\rho) \right\|_{L^2((0,T)\times\Omega)^d} \leq C_2.$$

Since $\|\nabla_{\mathcal{D}} \cdot\|_{L^2(\Omega)^d}$ is a norm on $X_{\mathcal{D},0}$, this shows that $(\zeta_\rho(u_\rho))_{\rho\in[0,1]}$ remains bounded in this finite dimensional space. In particular, for all $i \in I$, $(\zeta_\rho(u_\rho)_i)_{\rho\in[0,1]}$ is bounded. Using Assumption (6.3b) for ζ_ρ with constants not depending on ρ, we deduce that $((u_\rho)_i)_{\rho\in[0,1]}$ remains bounded for any $i \in I$, and thus that $(u_\rho)_{\rho\in[0,1]}$ is bounded in $X_{\mathcal{D},0}$.
If $\rho = 0$ then (6.10) is a square linear system. Any solution to this system being bounded in $X_{\mathcal{D},0}$, this shows that the underlying linear system is invertible. A topological degree argument (see Theorem D.1) combined with the uniform bound on $(u_\rho)_{\rho\in[0,1]}$ then shows that the scheme corresponding to $\rho = 1$, that is (6.10), possesses at least one solution. ∎

Lemma 6.10 (Uniqueness of the solution to the GS). *Under Assumptions (6.3), let $\mathcal{D}_T$ be a space–time GD for homogeneous Dirichlet boundary conditions, in the sense of Definition 4.1. Assume that the underlying spatial discretisation has a piecewise constant reconstruction in the sense of Definition 2.12. Let $u, \widetilde{u}$ be solutions to the GS (6.10). Then, for all $n = 0, \ldots, N$, $\Pi_{\mathcal{D}} u^{(n)} = \Pi_{\mathcal{D}} \widetilde{u}^{(n)}$ in $L^2(\Omega)$, and $\zeta(u^{(n)}) = \zeta(\widetilde{u}^{(n)})$ in $X_{\mathcal{D},0}$.*

Proof. The proof is done by induction on n. The result is clearly true for $n = 0$, since $u^{(0)} = \widetilde{u}^{(0)} = \mathcal{I}_{\mathcal{D}} u_{\mathrm{ini}}$. Let us now assume that, for some $n \leq N-1$, $\Pi_{\mathcal{D}} u^{(n)}(\boldsymbol{x}) = \Pi_{\mathcal{D}} \widetilde{u}^{(n)}(\boldsymbol{x})$ for a.e. $\boldsymbol{x} \in \Omega$. Subtracting the equation corresponding to $\widetilde{u}^{(n+1)}$ to the equation corresponding to $u^{(n+1)}$, we get

$$\int_\Omega \Big[\frac{\Pi_{\mathcal{D}}(\beta(u^{(n+1)}) - \beta(\widetilde{u}^{(n+1)}))(\boldsymbol{x})}{\delta t^{(n+\frac{1}{2})}} \Pi_{\mathcal{D}} v(\boldsymbol{x}) \\ + \nabla_{\mathcal{D}}(\zeta(u^{(n+1)}) - \zeta(\widetilde{u}^{(n+1)}))(\boldsymbol{x}) \cdot \nabla_{\mathcal{D}} v(\boldsymbol{x}) \Big] \mathrm{d}\boldsymbol{x} = 0, \quad \forall v \in X_{\mathcal{D},0}. \quad (6.19)$$

Using (6.3b)–(6.3c) we have

$$\Pi_{\mathcal{D}}\left[\beta(u^{(n+1)})-\beta(\widetilde{u}^{(n+1)})\right]\times\Pi_{\mathcal{D}}\left[\zeta(u^{(n+1)})-\zeta(\widetilde{u}^{(n+1)})\right]=$$
$$\left[\beta(\Pi_{\mathcal{D}}u^{(n+1)})-\beta(\Pi_{\mathcal{D}}\widetilde{u}^{(n+1)})\right]\left[\zeta(\Pi_{\mathcal{D}}u^{(n+1)})-\zeta(\Pi_{\mathcal{D}}\widetilde{u}^{(n+1)}))\right]\geq 0.$$

Hence, making $v=\zeta(u^{(n+1)})-\zeta(\widetilde{u}^{(n+1)})$ in (6.19),

$$\int_\Omega|\nabla_{\mathcal{D}}(\zeta(u^{(n+1)})-\zeta(\widetilde{u}^{(n+1)}))(\boldsymbol{x})|^2\mathrm{d}\boldsymbol{x}=0.$$

Since $\|\nabla\cdot\|_{L^2(\Omega)}$ is a norm on $X_{\mathcal{D},0}$, this shows that $\zeta(u^{(n+1)})=\zeta(\widetilde{u}^{(n+1)})$. We then get, from (6.19), that

$$\int_\Omega\left[\Pi_{\mathcal{D}}(\beta(u^{(n+1)})-\beta(\widetilde{u}^{(n+1)}))(\boldsymbol{x})\right]\Pi_{\mathcal{D}}v(\boldsymbol{x})\mathrm{d}\boldsymbol{x}=0,\quad\forall v\in X_{\mathcal{D},0}.$$

Letting $v=\beta(u^{(n+1)})-\beta(\widetilde{u}^{(n+1)})$ gives $\Pi_{\mathcal{D}}\beta(u^{(n+1)})=\Pi_{\mathcal{D}}\beta(\widetilde{u}^{(n+1)})$ a.e. on Ω. Since $\Pi_{\mathcal{D}}\zeta(u^{(n+1)})=\Pi_{\mathcal{D}}\zeta(\widetilde{u}^{(n+1)})$ a.e. on Ω, Assumption (6.3d) and the fact that $\Pi_{\mathcal{D}}(\beta(w)+\zeta(w))=\beta(\Pi_{\mathcal{D}}w)+\zeta(\Pi_{\mathcal{D}}w)$ for all $w\in X_{\mathcal{D},0}$ imply $\Pi_{\mathcal{D}}u^{(n+1)}=\Pi_{\mathcal{D}}\widetilde{u}^{(n+1)}$ a.e. on Ω. ■

Lemma 6.11 (Estimate on the dual norm of the discrete time derivative). *Under Assumptions (6.3), let $\mathcal{D}_T$ be a space–time GD for homogeneous Dirichlet boundary conditions, in the sense of Definition 4.1. Assume that the underlying spatial discretisation has a piecewise constant reconstruction in the sense of Definition 2.12. Let u be a solution to Scheme (6.10). Then there exists C_3, depending only on L_β, L_ζ, $C_P\geq C_{\mathcal{D}}$, $C_{\mathrm{ini}}\geq\|\Pi_{\mathcal{D}}I_{\mathcal{D}}u_{\mathrm{ini}}\|_{L^2(\Omega)}$, f, $\underline{\lambda}$, $\overline{\lambda}$ and T, such that*

$$\int_0^T\|\delta_{\mathcal{D}}\beta(u)(t)\|_{\star,\mathcal{D}}^2\,\mathrm{d}t\leq C_3,\tag{6.20}$$

where the dual norm $\|\cdot\|_{\star,\mathcal{D}}$ is given by Definition 4.11.

Proof. Let us take a generic $v=(v^{(n)})_{n=1,\dots,N}\subset X_{\mathcal{D},0}$ as test function in (6.10). We have

$$\int_0^T\int_\Omega\delta_{\mathcal{D}}\beta(u)(\boldsymbol{x},t)\Pi_{\mathcal{D}}^{(1)}v(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t\leq$$
$$\overline{\lambda}\int_0^T\int_\Omega|\nabla_{\mathcal{D}}^{(1)}\zeta(u)(\boldsymbol{x},t)|\,|\nabla_{\mathcal{D}}^{(1)}v(\boldsymbol{x},t)|\mathrm{d}\boldsymbol{x}\mathrm{d}t+\int_0^T\int_\Omega f(\boldsymbol{x},t)\Pi_{\mathcal{D}}^{(1)}v(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t.$$

Using the Cauchy–Schwarz inequality, the definition 2.2 of $C_{\mathcal{D}}$, and Estimates (6.13), this gives $C_4>0$ depending only on L_β, C_P, C_{ini}, f, $\underline{\lambda}$ and $\overline{\lambda}$ such that

$$\int_0^T \int_\Omega \delta_{\mathcal{D}}\beta(u)(\boldsymbol{x},t)\Pi_{\mathcal{D}}^{(1)}v(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t \leq C_4\|\nabla_{\mathcal{D}}^{(1)}v\|_{L^2(0,T;L^2(\Omega))^d}.$$

The proof of (6.20) is completed by selecting

$$v = \left(\left\|\delta_{\mathcal{D}}^{(n+\frac{1}{2})}\beta(u)\right\|_{\star,\mathcal{D}} z^{(n)}\right)_{n=0,\ldots,N}$$

with $(z^{(n)})_{n=0,\ldots,N} \subset X_{\mathcal{D},0}$ such that, for any $n = 0,\ldots,N-1$, $z^{(n+1)}$ realises the supremum in (4.16) with $w = \delta_{\mathcal{D}}^{(n+\frac{1}{2})}\beta(u)$. ∎

6.4 A first convergence theorem

The following theorem states initial convergence properties of the GS for (6.1).

Theorem 6.12 (Convergence of the GS). *Under Assumptions (6.3), let $((\mathcal{D}_T)_m)_{m\in\mathbb{N}}$ be a space–time-consistent, limit-conforming and compact sequence of space–time GDs, for homogeneous Dirichlet boundary conditions, in the sense of Definitions 4.3 and 4.6. We assume that the sequence of underlying spatial discretisations has the piecewise constant reconstruction property (Definition 2.12). Let $\nu : \mathbb{R} \to \mathbb{R}$ be defined by*

$$\forall s \in \mathbb{R}, \quad \nu(s) = \int_0^s \zeta'(q)\beta'(q)\mathrm{d}q. \tag{6.21}$$

For any $m \in \mathbb{N}$, let u_m be a solution to (6.10) with $\mathcal{D} = \mathcal{D}_m$. Then, as $m \to \infty$,

$$\begin{aligned}
&\Pi_{\mathcal{D}_m}^{(1)}\beta(u_m) \to \beta(\overline{u}) && \text{weakly in } L^2(\Omega) \text{ uniformly on } [0,T] \\
& && \text{(see Definition C.14)}, \\
&\Pi_{\mathcal{D}_m}^{(1)}\zeta(u_m) \to \zeta(\overline{u}) && \text{weakly in } L^2(\Omega\times(0,T)), \\
&\Pi_{\mathcal{D}_m}^{(1)}\nu(u_m) \to \nu(\overline{u}) && \text{in } L^2(\Omega\times(0,T)), \\
&\nabla_{\mathcal{D}_m}^{(1)}\zeta(u_m) \to \nabla\zeta(\overline{u}) && \text{weakly in } L^2(\Omega\times(0,T))^d,
\end{aligned} \tag{6.22}$$

where $\overline{u}$ is the unique solution to (6.4).

Remark 6.13. We do not assume the existence of a solution $\overline{u}$ to the continuous problem, the convergence analysis establishes this existence, which proves part of Theorem 6.3.

Proof.
Note that, since $((\mathcal{D}_T)_m)_{m\in\mathbb{N}}$ is compact, it is also coercive (Lemma 2.10).

Step 1: Application of compactness results.

Thanks to Theorem 4.19 and Estimates (6.13) and (6.20), we first extract a subsequence, without changing the notations, such that $(\Pi^{(1)}_{\mathcal{D}_m}\beta(u_m))_{m\in\mathbb{N}}$ converges weakly in $L^2(\Omega)$ uniformly in $[0,T]$ (in the sense of Definition C.14) to some function $\overline{\beta} \in C([0,T];L^2(\Omega)\text{-w})$. By space–time-consistency of $((\mathcal{D}_T)_m)_{m\in\mathbb{N}}$, $\Pi^{(1)}_{\mathcal{D}_m}\beta(u_m)(\cdot,0) = \Pi_{\mathcal{D}_m}\beta(\mathcal{I}_{\mathcal{D}_m}u_{\rm ini}) = \beta(\Pi_{\mathcal{D}_m}\mathcal{I}_{\mathcal{D}_m}u_{\rm ini}) \to \beta(u_{\rm ini})$ in $L^2(\Omega)$. Hence, the uniform-in-time weak $L^2(\Omega)$ convergence of $(\Pi^{(1)}_{\mathcal{D}_m}\beta(u_m))_{m\in\mathbb{N}}$ shows that $\overline{\beta}(\cdot,0) = \beta(u_{\rm ini})$ in $L^2(\Omega)$. Using again Estimates (6.13) and applying Lemma 4.8, we extract another subsequence such that, for some $\overline{\zeta} \in L^2(0,T;H^1_0(\Omega))$, $\Pi^{(1)}_{\mathcal{D}_m}\zeta(u_m) \to \overline{\zeta}$ weakly in $L^2(\Omega\times(0,T))$ and $\nabla^{(1)}_{\mathcal{D}_m}\zeta(u_m) \to \nabla\overline{\zeta}$ weakly in $L^2(\Omega\times(0,T))^d$.
Estimates (6.13) and (6.20) also show that $\beta_m = \beta(u_m)$ and $\zeta_m = \zeta(u_m)$ satisfy the assumptions of Theorem 4.17 (weak-strong time-space convergence of a product theorem). Hence,

$$\lim_{m\to\infty}\int_0^T\int_\Omega \beta\left(\Pi^{(1)}_{\mathcal{D}_m}u_m(\boldsymbol{x},t)\right)\zeta\left(\Pi^{(1)}_{\mathcal{D}_m}u_m(\boldsymbol{x},t)\right)\mathrm{d}\boldsymbol{x}\mathrm{d}t = \int_0^T\int_\Omega \overline{\beta}(\boldsymbol{x},t)\overline{\zeta}(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t. \quad (6.23)$$

Assumptions (6.3b)–(6.3d) allow us to apply Lemma D.10 to $w_m = \Pi^{(1)}_{\mathcal{D}_m}u_m$. This gives the existence of a measurable function $\overline{u}$ such that $\overline{\beta} = \beta(\overline{u})$ and $\overline{\zeta} = \zeta(\overline{u})$ a.e. on $\Omega\times(0,T)$. Since $\overline{\zeta} \in L^2(\Omega\times(0,T))$, the growth assumption (6.3b) on ζ ensures that $\overline{u} \in L^2(\Omega\times(0,T))$.
Since $0 \le \zeta'(q)\beta'(q) \le \sqrt{L_\zeta L_\beta}\sqrt{\zeta'(q)\beta'(q)}$, the following inequality holds for all $a,b\in\mathbb{R}$:

$$\begin{aligned}(\nu(a)-\nu(b))^2 &= \left(\int_a^b \zeta'(q)\beta'(q)\mathrm{d}q\right)^2\\ &\le \left(\sqrt{L_\zeta L_\beta}\int_a^b\sqrt{\beta'(q)\zeta'(q)}\mathrm{d}q\right)^2\\ &\le L_\zeta L_\beta\left(\int_a^b\beta'(q)\mathrm{d}q\right)\left(\int_a^b\zeta'(q)\mathrm{d}q\right)\\ &= L_\zeta L_\beta[\beta(b)-\beta(a)][\zeta(b)-\zeta(a)].\end{aligned}$$

It can therefore be deduced that

$$\begin{aligned}&\int_0^T\int_\Omega\left[\nu(\Pi^{(1)}_{\mathcal{D}_m}u_m(\boldsymbol{x},t)) - \nu(\overline{u}(\boldsymbol{x},t))\right]^2\mathrm{d}\boldsymbol{x}\mathrm{d}t\\ &\qquad\le L_\zeta L_\beta\int_0^T\int_\Omega\left[\beta(\Pi^{(1)}_{\mathcal{D}_m}u_m(\boldsymbol{x},t)) - \beta(\overline{u}(\boldsymbol{x},t))\right]\end{aligned}$$

$$\times \left[\zeta(\Pi^{(1)}_{\mathcal{D}_m} u_m(\boldsymbol{x},t)) - \zeta(\overline{u}(\boldsymbol{x},t))\right] \mathrm{d}\boldsymbol{x}\mathrm{d}t. \quad (6.24)$$

Developing the right-hand side of this inequality, using (6.23) and the weak convergences $\beta(\Pi^{(1)}_{\mathcal{D}_m} u_m) \to \overline{\beta} = \beta(\overline{u})$ and $\zeta(\Pi^{(1)}_{\mathcal{D}_m} u_m) \to \overline{\zeta} = \zeta(\overline{u})$, we see that this right-hand side goes to 0 as $m \to \infty$. Hence, taking the superior limit as $m \to \infty$ in (6.24) shows that $\nu(\Pi^{(1)}_{\mathcal{D}_m} u_m) \to \nu(\overline{u})$ in $L^2(\Omega \times (0,T))$.

Step 2: $\overline{u}$ is a solution to (6.4).
We drop some indices m for legibility. Let $\overline{v} \in L^2(0,T;H^1_0(\Omega))$ such that $\partial_t \overline{v} \in L^2(\Omega \times (0,T))$ and $\overline{v}(\cdot,T) = 0$. Let $(v_m)_{m\in\mathbb{N}}$ be given by Lemma 4.10 (for $\theta = 0$) and introduce $(0, v^{(0)}, \ldots, v^{(N-1)})$ as test function in (6.10). This gives $T_1^{(m)} + T_2^{(m)} = T_3^{(m)}$ with

$$T_1^{(m)} = \sum_{n=0}^{N-1} \int_\Omega \left[\Pi_{\mathcal{D}}\beta(u^{(n+1)})(\boldsymbol{x}) - \Pi_{\mathcal{D}}\beta(u^{(n)})(\boldsymbol{x})\right] \Pi_{\mathcal{D}} v^{(n)}(\boldsymbol{x})\mathrm{d}\boldsymbol{x},$$

$$T_2^{(m)} = \int_0^T \int_\Omega \Lambda(\boldsymbol{x}) \nabla^{(1)}_{\mathcal{D}} \zeta(u)(\boldsymbol{x},t) \cdot \nabla^{(0)}_{\mathcal{D}} v(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t,$$

and

$$T_3^{(m)} = \int_0^T \int_\Omega f(\boldsymbol{x},t) \Pi^{(0)}_{\mathcal{D}} v(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t.$$

Use the discrete integration-by-parts formula (D.15) in $T_1^{(m)}$:

$$\begin{aligned} T_1^{(m)} = & -\int_0^T \int_\Omega \Pi^{(1)}_{\mathcal{D}} \beta(u)(\boldsymbol{x},t) \delta_{\mathcal{D}} v(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t \\ & - \int_\Omega \beta(\Pi_{\mathcal{D}} \mathcal{I}_{\mathcal{D}} u_{\mathrm{ini}})(\boldsymbol{x}) \Pi_{\mathcal{D}} v^{(0)}\mathrm{d}\boldsymbol{x}. \end{aligned}$$

Hence, by the convergence properties (4.10c) and (4.10b) of $(v_m)_{m\in\mathbb{N}}$, as $m \to \infty$ we have

$$T_1^{(m)} \to -\int_0^T \int_\Omega \beta(\overline{u})(\boldsymbol{x},t) \partial_t \overline{v}(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t - \int_\Omega \beta(u_{\mathrm{ini}})(\boldsymbol{x}) \overline{v}(\boldsymbol{x},0)\mathrm{d}\boldsymbol{x}. \quad (6.25)$$

Using the convergence (4.7b) and (4.7a) of $(v_m)_{m\in\mathbb{N}}$, we have, as $m \to \infty$,

$$\begin{aligned} T_2^{(m)} &\to \int_0^T \int_\Omega \Lambda(\boldsymbol{x}) \nabla \zeta(\overline{u})(\boldsymbol{x},t) \cdot \nabla \overline{v}(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t, \\ T_3^{(m)} &\to \int_0^T \int_\Omega f(\boldsymbol{x},t) \overline{v}(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t. \end{aligned}$$

Plugged alongside (6.25) in $T_1^{(m)} + T_2^{(m)} = T_3^{(m)}$, these convergences show that $\overline{u}$ satisfies (6.4). ∎

Remark 6.14 (Convergence of $\Pi^{(1)}_{\mathcal{D}} u_m$?)
We do not prove here that $\overline{u}$ is a weak limit of $\Pi^{(1)}_{\mathcal{D}_m} u_m$. Such a limit is not stated in (6.22) and can actually be considered as irrelevant for the model (6.1) since, in this model, the quantities of interest (physically relevant when this PDE models a natural phenomenon) are $\beta(\overline{u})$ and $\zeta(\overline{u})$.

As a corollary to this convergence analysis and to the uniqueness of the solution (Theorem 6.3), an equivalent form of (6.4) can be stated.

Corollary 6.15 (Equivalent from of (6.4)). *Under Hypotheses (6.3), Problem (6.4) is equivalent to*

$$\begin{cases} \overline{u} \in L^2(0,T;L^2(\Omega))\,,\ \zeta(\overline{u}) \in L^2(0,T;H^1_0(\Omega))\,, \\ \beta(\overline{u}) \in C([0,T],L^2(\Omega)\text{-w})\,,\ \partial_t\beta(\overline{u}) \in L^2(0,T;H^{-1}(\Omega))\,, \\ \beta(\overline{u})(\cdot,0) = \beta(u_{\rm ini}) \ \textit{in}\ L^2(\Omega), \\ \displaystyle\int_0^T \langle \partial_t\beta(\overline{u})(\cdot,t), \overline{v}(\cdot,t)\rangle_{H^{-1},H^1_0}\mathrm{d}t \\ \qquad \displaystyle + \int_0^T\int_\Omega \Lambda(\boldsymbol{x})\nabla\zeta(\overline{u})(\boldsymbol{x},t)\cdot\nabla\overline{v}(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t \\ \qquad \displaystyle = \int_0^T\int_\Omega f(\boldsymbol{x},t)\overline{v}(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t, \quad \forall \overline{v} \in L^2(0;T;H^1_0(\Omega)), \end{cases} \tag{6.26}$$

where $C([0,T];L^2(\Omega)\text{-w})$ denotes the space of continuous functions $[0,T] \mapsto L^2(\Omega)$ for the weak-$*$ topology of $L^2(\Omega)$.

Proof. Let us prove that (6.4) implies (6.26). There is a unique solution to (6.4) (Theorem 6.3), so it must be the $\overline{u}$ constructed in the proof of Theorem 6.12. We saw in Step 1 of this proof that $\beta(\overline{u}) = \overline{\beta} \in C([0,T];L^2(\Omega)\text{-w})$ and that $\beta(\overline{u})(0,\cdot) = \overline{\beta}(0,\cdot) = \beta(u_{\rm ini})$. Using $C^\infty_c((0,T)\times\Omega)$ test functions in (6.4), we see that

$$\partial_t\beta(\overline{u}) = \mathrm{div}(\Lambda\nabla\zeta(\overline{u})) + f$$

in the sense of distributions. Since $\nabla\zeta(\overline{u}) \in L^2(0,T;L^2(\Omega))$, this shows that $\partial_t\beta(\overline{u}) \in L^2(0,T;H^{-1}(\Omega))$. Let $\overline{v} \in C^\infty_c((0,T)\times\Omega)$. By definition of the distribution derivative,

$$-\int_0^T\int_\Omega \beta(\overline{u})(\boldsymbol{x},t)\partial_t\overline{v}(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t = \int_0^T \langle \partial_t\beta(\overline{u})(t), \overline{v}(t)\rangle_{H^{-1},H^1_0}\mathrm{d}t$$

and thus (6.4) shows that the equation in (6.26) is satisfied for such smooth compactly supported $\overline{v}$. Since these functions are dense in $L^2(0,T;H^1_0(\Omega))$ (see [67]), we infer that (6.26) is fully satisfied.

Let us now assume that $\overline{u}$ satisfies (6.26). Then it clearly has all the regularity properties expected in (6.4). To prove that it also satisfies the equation in this latter problem, we start by taking $\overline{v} \in C^\infty_c((-\infty,T)\times\Omega)$. By smoothness

of this function and regularity assumptions on $\beta(\overline{u})$ an integration-by-parts gives

$$\int_0^T \langle \partial_t \beta(\overline{u})(t), \overline{v}(t)\rangle_{H^{-1},H_0^1} \mathrm{d}t = -\int_0^T \beta(\overline{u})(\boldsymbol{x},t)\partial_t \overline{v}(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t \\ - \int_\Omega \beta(\overline{u})(\boldsymbol{x},0)\overline{v}(\boldsymbol{x},0)\mathrm{d}\boldsymbol{x}.$$

Since $\beta(\overline{u})(\boldsymbol{x},0) = \beta(u_{\mathrm{ini}})$, (6.26) proves that (6.4) is satisfied for such $\overline{v}$. As discussed at the end of the proof of Theorem 6.12, this shows that (6.26) is satisfied for all required test functions. ■

Remark 6.16 (The continuity property of $\beta(\overline{u})$)
The continuity property of $\beta(\overline{u}) : [0,T] \to L^2(\Omega)$-w is rather natural. Indeed, the PDE in the sense of distributions shows that $T_\varphi : t \mapsto \langle \beta(\overline{u})(t), \varphi\rangle_{L^2}$ belongs to $W^{1,1}(0,T)$, and is therefore continuous, for any $\varphi \in C_c^\infty(\Omega)$. The density of $\varphi \in C_c^\infty(\Omega)$ in $L^2(\Omega)$ combined with the fact that $\beta(\overline{u}) \in L^\infty(0,T;L^2(\Omega))$, proves the continuity of T_φ for any $\varphi \in L^2(\Omega)$, that is to say the continuity of $\beta(\overline{u}) : [0,T] \to L^2(\Omega)$-w.
This notion of $\beta(\overline{u})$ as a function continuous in time is nevertheless a subtle one. It is to be understood in the sense that the function $(\boldsymbol{x},t) \mapsto \beta(\overline{u}(\boldsymbol{x},t))$ has an a.e. representative which is continuous $[0,T] \mapsto L^2(\Omega)$-w. In other words, there is a function $Z \in C([0,T];L^2(\Omega)\text{-w})$ such that $Z(t)(\boldsymbol{x}) = \beta(\overline{u}(\boldsymbol{x},t))$ for a.e. $(\boldsymbol{x},t) \in \Omega \times (0,T)$. We must however make sure, when dealing with pointwise values in time to separate Z from $\beta(\overline{u}(\cdot,\cdot))$ as $\beta(\overline{u}(\cdot,t_1))$ may not make sense for a particular $t_1 \in [0,T]$.
That being said, in order to adopt a simple notation, in the following we denote by $\beta(\overline{u})(\cdot,\cdot)$ the function Z, and by $\beta(\overline{u}(\cdot,\cdot))$ the a.e.-defined composition of β and $\overline{u}$. Hence, it will make sense to talk about $\beta(\overline{u})(\cdot,t)$ for a particular $t_1 \in [0,T]$, and we will only write $\beta(\overline{u})(\boldsymbol{x},t) = \beta(\overline{u}(\boldsymbol{x},t))$ for a.e. $(\boldsymbol{x},t) \in \Omega \times (0,T)$.

6.5 Uniform-in-time, strong L^2 convergence results

We denote by R_β the range of β and define the pseudo-inverse function $\beta^i : R_\beta \to \mathbb{R}$ of β by

$$\forall s \in R_\beta\,,\ \beta^i(s) = \begin{cases} \inf\{t \in \mathbb{R} \,|\, \beta(t) = s\} & \text{if } s \geq 0, \\ \sup\{t \in \mathbb{R} \,|\, \beta(t) = s\} & \text{if } s < 0, \end{cases} \\ = \text{closest } t \text{ to } 0 \text{ such that } \beta(t) = s. \tag{6.27}$$

See Figure 6.1 for an illustration of β^i.
Since $\beta(0) = 0$, it holds $\beta^i \geq 0$ on $R_\beta \cap \mathbb{R}^+$ and $\beta^i \leq 0$ on $R_\beta \cap \mathbb{R}^-$. The function $B : R_\beta \to [0,\infty]$ is defined by

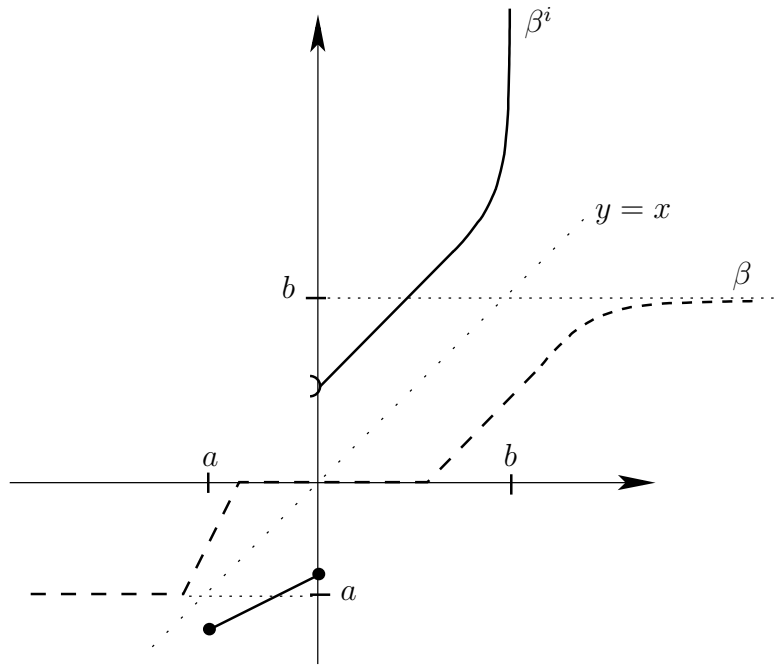


Fig. 6.1. An example of β (dashed line) and its pseudo-inverse function β^i (continuous line). Here, the range of β is $[a, b)$.

$$B(z) = \int_0^z \zeta(\beta^i(s))\, ds. \tag{6.28}$$

The function β^i is non-decreasing, and thus $B(z)$ is always well-defined in $[0, \infty)$. The signs of β^i and ζ also ensure that that B is non-decreasing on $R_\beta \cap \mathbb{R}^+$ and non-increasing on $R_\beta \cap R^-$. B can therefore be extended to the closure $\overline{R_\beta}$ of R_β, by defining $B(a) = \lim_{z \to a} B(z) \in [0, +\infty]$ at any endpoint a of R_β that does not belong to R_β. Lemma 6.21 in Section 6.6 states a few useful properties of B.

Remark 6.17 (Range of $\beta(\overline{u})$)
The a.e. equality $\beta(\overline{u})(\boldsymbol{x}, t) = \beta(\overline{u}(\boldsymbol{x}, t))$ (see Remark 6.16) ensures that $\beta(\overline{u})(\cdot, \cdot)$ takes its values in $\overline{R_\beta}$.

The following theorem shows that the solutions to GSs for (6.1) actually enjoy stronger convergence results than those established in Theorem 6.12.

Theorem 6.18 (Uniform-in-time convergence of the GS). *Under the assumptions of Theorem 6.12, the solution u_m to the GS (6.10) with $\mathcal{D}_T = (\mathcal{D}_T)_m$ satisfies the following convergence results, as $m \to \infty$:*

$$\begin{aligned}
&\sup_{t\in[0,T]}\left\|\Pi^{(1)}_{\mathcal{D}_m}\nu(u_m)(t)-\nu(\overline{u})(t)\right\|_{L^2(\Omega)}\to 0,\\
&\Pi^{(1)}_{\mathcal{D}_m}\zeta(u_m)\to\zeta(\overline{u})\ \textit{in}\ L^2(\Omega\times(0,T)),\\
&\nabla^{(1)}_{\mathcal{D}_m}\zeta(u_m)\to\nabla\zeta(\overline{u})\ \textit{in}\ L^2(\Omega\times(0,T))^d,
\end{aligned}\tag{6.29}$$

where $\overline{u}$ is the unique solution to (6.4)*, and ν is defined by* (6.21)*.*

Remark 6.19. For the Stefan problem, $\beta=\mathrm{Id}$ and thus $\nu(\overline{u})=\zeta(\overline{u})$ is the temperature of the melting material. For the Richards model, $\zeta=\mathrm{Id}$ and thus $\nu(\overline{u})=\beta(\overline{u})$ is the water saturation. Hence, in both cases, $\nu(\overline{u})$ is the quantity of interest to approximate.

Proof.
By (6.39) in Lemma 6.21, $\eta=B\circ\beta$. The energy estimate (6.12) can thus be written

$$\begin{aligned}
&\int_\Omega B(\beta(\Pi^{(1)}_{\mathcal{D}_m}u_m))(\boldsymbol{x},T_0)\mathrm{d}\boldsymbol{x}\\
&\qquad+\int_0^{T_0}\int_\Omega \Lambda(\boldsymbol{x})\nabla^{(1)}_{\mathcal{D}_m}\zeta(u_m)(\boldsymbol{x},t)\cdot\nabla^{(1)}_{\mathcal{D}_m}\zeta(u_m)(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t\\
&\qquad\qquad\le\int_\Omega B(\beta(\Pi_{\mathcal{D}_m}\mathcal{I}_{\mathcal{D}_m}u_{\mathrm{ini}}))(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\\
&\qquad\qquad\qquad+\int_0^{t^{(k)}}\int_\Omega f(\boldsymbol{x},t)\Pi^{(1)}_{\mathcal{D}_m}\zeta(u_m)(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t.
\end{aligned}\tag{6.30}$$

Here, we recall that $t^{(k)}$ is the time such that $T_0\in(t^{(k-1)},t^{(k)}]$.

Step 1 Uniform-in-time convergence of $\Pi^{(1)}_{\mathcal{D}_m}\nu(u_m)$.
Let us take $T_0\in[0,T]$ and $(T_m)_{m\ge1}$ a sequence in $[0,T]$ which converges to T_0. The Cauchy–Schwarz inequality for the semi-definite positive symmetric form

$$W\in L^2((0,T)\times\Omega)^d\to\int_0^{T_m}\int_\Omega\Lambda(\boldsymbol{x})W(t,\boldsymbol{x})\cdot W(t,\boldsymbol{x})\mathrm{d}\boldsymbol{x}\mathrm{d}t$$

shows that

$$\begin{aligned}
&\left(\int_0^{T_m}\int_\Omega\Lambda(\boldsymbol{x})\nabla^{(1)}_{\mathcal{D}_m}\zeta(u_m)(t,\boldsymbol{x})\cdot\nabla\zeta(\overline{u})(t,\boldsymbol{x})\mathrm{d}\boldsymbol{x}\mathrm{d}t\right)^2\\
&\qquad\le\left(\int_0^{T_m}\int_\Omega\Lambda(\boldsymbol{x})\nabla^{(1)}_{\mathcal{D}_m}\zeta(u_m)(t,\boldsymbol{x})\cdot\nabla^{(1)}_{\mathcal{D}_m}\zeta(u_m)(t,\boldsymbol{x})\mathrm{d}\boldsymbol{x}\mathrm{d}t\right)\\
&\qquad\qquad\times\left(\int_0^{T_m}\int_\Omega\Lambda(\boldsymbol{x})\nabla\zeta(\overline{u})(t,\boldsymbol{x})\cdot\nabla\zeta(\overline{u})(t,\boldsymbol{x})\mathrm{d}\boldsymbol{x}\mathrm{d}t\right)
\end{aligned}$$

By weak convergence in $L^2((0,T)\times\Omega)^d$ of $\nabla^{(1)}_{\mathcal{D}_m}\zeta(u_m)$ to $\nabla\zeta(\overline{u})$ and strong convergence in the same space of $\mathbf{1}_{[0,T_m]}\nabla\zeta(\overline{u})$ to $\nabla\zeta(\overline{u})$, we can pass to the limit in the left-hand side (by weak-strong convergence, see Lemma D.8 page 464) and in the second term in the right-hand side. Hence, taking the inferior limit of this inequality and dividing by $\int_0^{T_0}\int_\Omega \Lambda\nabla\zeta(\overline{u})\cdot\nabla\zeta(\overline{u})$, we deduce that

$$\liminf_{m\to\infty}\int_0^{T_m}\int_\Omega \Lambda(\boldsymbol{x})\nabla^{(1)}_{\mathcal{D}_m}\zeta(u_m)(\boldsymbol{x},t)\cdot\nabla^{(1)}_{\mathcal{D}_m}\zeta(u_m)(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t \\ \geq \int_0^{T_0}\int_\Omega \Lambda(\boldsymbol{x})\nabla\zeta(\overline{u})(\boldsymbol{x},t)\cdot\nabla\zeta(\overline{u})(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t. \quad (6.31)$$

By space–time-consistency of $((\mathcal{D}_T)_m)_{m\in\mathbb{N}}$ (Definition 4.3) and quadratic growth (6.17) of $\eta = B\circ\beta$, it holds

$$B(\beta(\Pi_{\mathcal{D}_m}\mathcal{I}_{\mathcal{D}_m}u_{\mathrm{ini}})) \to B(\beta(u_{\mathrm{ini}}) \text{ in } L^1(\Omega) \text{ as } m\to\infty. \quad (6.32)$$

We then write (6.30) with T_m instead of T_0. The time $t^{(k(m))}$ such $T_m \in (t^{(k(m)-1)}, t^{(k(m))}]$ satisfies $t^{(k(m))}\to T_0$ as $m\to\infty$. Hence, using (6.31) and the weak convergence of $\Pi^{(1)}_{\mathcal{D}_m}\zeta(u_m)$ to $\zeta(\overline{u})$,

$$\begin{aligned}
&\limsup_{m\to\infty}\int_\Omega B(\beta(\Pi^{(1)}_{\mathcal{D}_m}u_m(\boldsymbol{x},T_m)))\mathrm{d}\boldsymbol{x} \\
&\quad\leq \limsup_{m\to\infty}\Bigg(\int_\Omega B(\beta(\Pi_{\mathcal{D}_m}\mathcal{I}_{\mathcal{D}_m}u_{\mathrm{ini}}))(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \\
&\qquad + \int_0^{t^{(k(m))}}\int_\Omega f(\boldsymbol{x},t)\Pi^{(1)}_{\mathcal{D}_m}\zeta(u_m)(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t \\
&\qquad - \int_0^{T_m}\int_\Omega \Lambda(\boldsymbol{x})\nabla^{(1)}_{\mathcal{D}_m}\zeta(u_m)(\boldsymbol{x},t)\cdot\nabla^{(1)}_{\mathcal{D}_m}\zeta(u_m)(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t\Bigg) \\
&\quad\leq \int_\Omega B(\beta(u_{\mathrm{ini}}))(\boldsymbol{x})\mathrm{d}\boldsymbol{x} + \int_0^{T_0}\int_\Omega f(\boldsymbol{x},t)\zeta(\overline{u})(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t \\
&\qquad - \liminf_{m\to\infty}\int_0^{T_m}\int_\Omega \Lambda(\boldsymbol{x})\nabla_{\mathcal{D}_m}\zeta(u_m)(\boldsymbol{x},t)\cdot\nabla_{\mathcal{D}_m}\zeta(u_m)(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t \\
&\quad\leq \int_\Omega B(\beta(u_{\mathrm{ini}}))(\boldsymbol{x})\mathrm{d}\boldsymbol{x} + \int_0^{T_0}\int_\Omega f(\boldsymbol{x},t)\zeta(\overline{u})(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t \\
&\qquad - \int_0^{T_0}\int_\Omega \Lambda(\boldsymbol{x})\nabla\zeta(\overline{u})(\boldsymbol{x},t)\cdot\nabla\zeta(\overline{u})(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t.
\end{aligned}$$

Corollary 6.24 therefore gives

$$\limsup_{m\to\infty}\int_\Omega B(\beta(\Pi^{(1)}_{\mathcal{D}_m}u_m(\boldsymbol{x},T_m)))\mathrm{d}\boldsymbol{x} \leq \int_\Omega B(\beta(\overline{u})(\boldsymbol{x},T_0))\mathrm{d}\boldsymbol{x}. \quad (6.33)$$

By Lemma C.13, the uniform-in-time weak L^2 convergence of $\beta(\Pi^{(1)}_{\mathcal{D}_m} u_m)$ to $\beta(\overline{u})$ and the continuity of $\beta(\overline{u}) : [0,T] \to L^2(\Omega)$-w (see Corollary 6.15), we have

$$\beta(\Pi^{(1)}_{\mathcal{D}_m} u_m)(T_m) \to \beta(\overline{u})(T_0) \text{ weakly in } L^2(\Omega) \text{ as } m \to \infty. \tag{6.34}$$

Therefore, for any $(s_m)_{m\in\mathbb{N}}$ converging to T_0,

$$\frac{\beta(\Pi^{(1)}_{\mathcal{D}_m} u_m(T_m)) + \beta(\overline{u})(s_m)}{2} \to \beta(\overline{u})(T_0) \text{ weakly in } L^2(\Omega) \text{ as } m \to \infty.$$

Lemma D.11 then gives, by convexity of B,

$$\begin{aligned}\int_\Omega & B(\beta(\overline{u})(\boldsymbol{x}, T_0))\mathrm{d}\boldsymbol{x} \\ & \leq \liminf_{m\to\infty} \int_\Omega B\left(\frac{\beta(\Pi^{(1)}_{\mathcal{D}_m} u_m(\boldsymbol{x}, T_m)) + \beta(\overline{u})(\boldsymbol{x}, s_m)}{2}\right)\mathrm{d}\boldsymbol{x}.\end{aligned} \tag{6.35}$$

Property (6.41) of B and the two inequalities (6.33) and (6.35) allow us to conclude the proof. Let $\mathcal{T}$ be the set of $\tau \in [0,T]$ such that $\beta(\overline{u}(\cdot,\tau)) = \beta(\overline{u})(\cdot,\tau)$ and $\nu(\overline{u}(\cdot,\tau)) = \nu(\overline{u})(\cdot,\tau)$ a.e. on Ω (see Remarks 6.16 and 6.25), and let $(s_m)_{m\in\mathbb{N}}$ be a sequence in $\mathcal{T}$ which converges to T_0. Since $\nu(\overline{u}) \in C([0,T]; L^2(\Omega))$ by Corollary 6.24, we have

$$\nu(\overline{u}(\cdot, s_m)) \to \nu(\overline{u})(\cdot, T_0) \text{ in } L^2(\Omega) \text{ as } m \to \infty. \tag{6.36}$$

Inequality (6.41) gives

$$\begin{aligned}\left\|\nu(\Pi^{(1)}_{\mathcal{D}_m} u_m(\cdot, T_m)) - \nu(\overline{u})(\cdot, T_0)\right\|^2_{L^2(\Omega)} & \\ \leq 2 \left\|\nu(\Pi^{(1)}_{\mathcal{D}_m} u_m(\cdot, T_m)) - \nu(\overline{u}(\cdot, s_m))\right\|^2_{L^2(\Omega)} & \\ + 2 \left\|\nu(\overline{u}(\cdot, s_m)) - \nu(\overline{u})(\cdot, T_0))\right\|^2_{L^2(\Omega)} & \\ \leq 8 L_\beta L_\zeta \int_\Omega \left[B(\beta(\Pi^{(1)}_{\mathcal{D}_m} u_m(\boldsymbol{x}, T_m))) + B(\beta(\overline{u}(\boldsymbol{x}, s_m)))\right]\mathrm{d}\boldsymbol{x} & \\ - 16 L_\beta L_\zeta \int_\Omega B\left(\frac{\beta(\Pi^{(1)}_{\mathcal{D}_m} u_m(\boldsymbol{x}, T_m)) + \beta(\overline{u}(\boldsymbol{x}, s_m))}{2}\right)\mathrm{d}\boldsymbol{x} & \\ + 2 \left\|\nu(\overline{u}(\cdot, s_m)) - \nu(\overline{u})(\cdot, T_0))\right\|^2_{L^2(\Omega)}. &\end{aligned}$$

We then take the lim sup as $m \to \infty$ of this expression. Thanks to (6.33) and to the boundedness of $\mathcal{B} : t \in [0,T] \mapsto \int_\Omega B(\beta(\overline{u})(\boldsymbol{x}, t))\mathrm{d}\boldsymbol{x} \in [0,\infty)$ (see Corollary 6.24), the first term in the right-hand side has a finite lim sup. We can therefore split the lim sup of this right-hand side without risking writing

$\infty - \infty$ and we get, thanks to (6.33), (6.35), (6.36) and to the continuity of B (Corollary 6.24),

$$\limsup_{m\to\infty} \left\| \nu(\Pi^{(1)}_{\mathcal{D}_m} u_m(\cdot, T_m)) - \nu(\overline{u})(\cdot, T_0) \right\|^2_{L^2(\Omega)} \leq 0.$$

Thus, $\nu(\Pi^{(1)}_{\mathcal{D}_m} u_m(\cdot, T_m)) \to \nu(\overline{u})(T_0)$ strongly in $L^2(\Omega)$. By Lemma C.13, this concludes the proof that $\sup_{t\in[0,T]} \|\nu(\Pi^{(1)}_{\mathcal{D}_m} u_m)(t) - \nu(\overline{u})(t)\|_{L^2(\Omega)} \to 0$.

Step 2: Strong convergence of $\nabla^{(1)}_{\mathcal{D}_m} \zeta(u_m)$.
Since B is convex, the convergence property (6.34) (with $T_m = T_0 = T$) and Lemma D.11 give

$$\int_\Omega B(\beta(\overline{u})(\boldsymbol{x}, T))\mathrm{d}\boldsymbol{x} \leq \liminf_{m\to\infty} \int_\Omega B(\beta(\Pi^{(1)}_{\mathcal{D}_m} u_m)(\boldsymbol{x}, T))\mathrm{d}\boldsymbol{x}.$$

Writing (6.30) with $T_0 = T$, taking the lim sup as $m \to \infty$, using (6.32) and the continuous integration-by-part formula (6.50), we therefore find

$$\limsup_{m\to\infty} \int_0^T \int_\Omega \Lambda(\boldsymbol{x}) \nabla^{(1)}_{\mathcal{D}_m} \zeta(u_m)(\boldsymbol{x}, t) \cdot \nabla^{(1)}_{\mathcal{D}_m} \zeta(u_m)(\boldsymbol{x}, t)\mathrm{d}\boldsymbol{x}\mathrm{d}t \\ \leq \int_0^{T_0} \int_\Omega \Lambda(\boldsymbol{x}) \nabla\zeta(\overline{u})(\boldsymbol{x}, t) \cdot \nabla\zeta(\overline{u})(\boldsymbol{x}, t)\mathrm{d}\boldsymbol{x}\mathrm{d}t.$$

Combined with (6.31) with $T_m = T_0 = T$, this shows that

$$\lim_{m\to\infty} \int_0^T \int_\Omega \Lambda(\boldsymbol{x}) \nabla^{(1)}_{\mathcal{D}_m} \zeta(u_m)(\boldsymbol{x}, t) \cdot \nabla^{(1)}_{\mathcal{D}_m} \zeta(u_m)(\boldsymbol{x}, t)\mathrm{d}\boldsymbol{x}\mathrm{d}t \\ = \int_0^{T_0} \int_\Omega \Lambda(\boldsymbol{x}) \nabla\zeta(\overline{u})(\boldsymbol{x}, t) \cdot \nabla\zeta(\overline{u})(\boldsymbol{x}, t)\mathrm{d}\boldsymbol{x}\mathrm{d}t. \quad (6.37)$$

Developing all the terms and using the weak convergence of $\nabla^{(1)}_{\mathcal{D}_m} \zeta(u_m)$ to $\nabla\zeta(\overline{u})$, we deduce

$$\lim_{m\to\infty} \int_0^T \int_\Omega \Lambda(\boldsymbol{x}) \left[\nabla^{(1)}_{\mathcal{D}_m} \zeta(u_m)(\boldsymbol{x}, t) - \nabla\zeta(\overline{u})(\boldsymbol{x}, t)\right] \\ \cdot \left[\nabla^{(1)}_{\mathcal{D}_m} \zeta(u_m)(\boldsymbol{x}, t) - \nabla\zeta(\overline{u})(\boldsymbol{x}, t)\right] \mathrm{d}\boldsymbol{x}\mathrm{d}t = 0.$$

The coercivity of Λ therefore implies $\nabla^{(1)}_{\mathcal{D}_m} \zeta(u_m) \to \nabla\zeta(\overline{u})$ strongly in $L^2(\Omega \times (0, T))^d$ as $m \to \infty$.

Step 3: Strong convergence of $\Pi^{(1)}_{\mathcal{D}_m} \zeta(u_m)$.
Apply Lemma 4.10 to $\overline{v} = \zeta(\overline{u})$. This gives $(v_m)_{m\in\mathbb{N}}$ such that $\Pi^{(1)}_{\mathcal{D}_m} v_m \to \zeta(\overline{u})$ in $L^2(\Omega \times (0, T))$ and $\nabla^{(1)}_{\mathcal{D}_m} v_m \to \nabla\zeta(\overline{u})$ in $L^2(\Omega \times (0, T))^d$. The coercivity definition 2.2 gives

$$\left\| \Pi^{(1)}_{\mathcal{D}_m} \zeta(u_m) - \Pi^{(1)}_{\mathcal{D}_m} v_m \right\|_{L^2(\Omega\times(0,T))} \leq C_P \left\| \nabla^{(1)}_{\mathcal{D}_m} \zeta(u_m) - \nabla^{(1)}_{\mathcal{D}_m} v_m \right\|_{L^2(\Omega\times(0,T))^d}.$$

By strong convergence of $\nabla^{(1)}_{\mathcal{D}_m} \zeta(u_m)$, letting $m \to \infty$ in this estimate proves that $\Pi^{(1)}_{\mathcal{D}_m} \zeta(u_m) \to \zeta(\overline{u})$ in $L^2(\Omega \times (0,T))$ follows. ■

Remark 6.20 (Convergence of $B(\beta(\Pi^{(1)}_{\mathcal{D}_m} u_m(T_m)))$)
Let $T_m \to T_0$. The convergence property (6.34), the convexity of B and Lemma D.11 show that

$$\int_\Omega B(\beta(\overline{u})(\boldsymbol{x}, T_0))\mathrm{d}\boldsymbol{x} \leq \liminf_{m\to\infty} \int_\Omega B(\beta(\Pi^{(1)}_{\mathcal{D}_m} u_m)(\boldsymbol{x}, T_m))\mathrm{d}\boldsymbol{x}.$$

Combined with (6.33), this gives

$$\lim_{m\to\infty} \int_\Omega B(\beta(\Pi^{(1)}_{\mathcal{D}_m} u_m(\boldsymbol{x}, T_m)))\mathrm{d}\boldsymbol{x} = \int_\Omega B(\beta(\overline{u})(\boldsymbol{x}, T_0))\mathrm{d}\boldsymbol{x}. \tag{6.38}$$

6.6 Auxiliary results

We state here a family of technical lemmas, starting with a few properties on ν and B.

Lemma 6.21. *Under Assumptions (6.3b) and (6.3c), let ν be defined by (6.21), B be defined by (6.28), and η be defined by (6.11). Then the function B is convex lower semi-continuous on $\overline{R_\beta}$, the function $B \circ \beta : \mathbb{R} \to [0, \infty)$ is continuous,*

$$\forall s \in \mathbb{R}, \quad \eta(s) = B(\beta(s)) = \int_0^s \zeta(q)\beta'(q)\mathrm{d}q, \tag{6.39}$$

$$\forall a \in \mathbb{R}, \forall r \in \overline{R_\beta}, \quad B(r) - B(\beta(a)) \geq \zeta(a)(r - \beta(a)), \tag{6.40}$$

and

$$\forall s, s' \in \mathbb{R}, (\nu(s) - \nu(s'))^2 \leq 4L_\beta L_\zeta \left[B(\beta(s)) + B(\beta(s')) - 2B\left(\frac{\beta(s) + \beta(s')}{2}\right)\right]. \tag{6.41}$$

Proof.
Let us first notice that, since $\beta \geq 0$ on $\mathbb{R}^+$ and $\beta \leq 0$ on $\mathbb{R}^-$, $\beta^i(s)$ is a real number for all $s \in R_\beta$. Moreover, since β is non-decreasing, β^i is also non-decreasing on R_β and therefore locally bounded on R_β. Hence, B is well defined and locally Lipschitz-continuous, with an a.e. derivative $B' = \zeta(\beta^i)$. B' is therefore non-decreasing and B is convex. Since B is continuous on R_β and extended by its (possibly infinite) limit at the endpoints of this interval, B is lower semi-continuous on $\overline{R_\beta}$.
To prove (6.39), we denote by $P \subset R_\beta$ the countable set of plateaux values of β, *i.e.* the numbers $y \in \mathbb{R}$ such that $\beta^{-1}(\{y\})$ is not reduced to a singleton (P is countable since $(\beta^{-1}(\{y\}))_{y\in P}$ is a family of disjoint intervals of positive lengths, and can therefore be counted by associating to each interval a rational number in it). If $s \notin \beta^{-1}(P)$ then $\beta^{-1}(\{\beta(s)\})$ is the singleton $\{s\}$ and therefore $\beta^i(\beta(s)) = s$. Moreover, β^i is continuous at $\beta(s)$ and thus B is differentiable at $\beta(s)$. Since β is differentiable a.e., we deduce that, for a.e. $s \notin \beta^{-1}(P)$, $(B(\beta))'(s) = B'(\beta(s))\beta'(s) = \zeta(\beta^i(\beta(s)))\beta'(s) = \zeta(s)\beta'(s)$. The set $\beta^{-1}(P)$ is a union of intervals on which β, and thus $B(\beta)$, are locally constant; hence, for a.e. s in this set, $(B(\beta))'(s) = 0$ and $\zeta(s)\beta'(s) = 0$. As a consequence, the locally Lipschitz-continuous functions $B(\beta)$ and $s \to \int_0^s \zeta(q)\beta'(q)\mathrm{d}q$ have identical derivatives a.e. on $\mathbb{R}$. Since they have the same value at $s = 0$, they are thus equal on $\mathbb{R}$ and the proof of (6.39) is complete. The continuity of $B \circ \beta$ follows from this relation.
We now prove (6.40), which states that $\zeta(a)$ belongs to the convex subdifferential of B at $\beta(a)$. We first start with the case $r \in R_\beta$, that is $r = \beta(b)$ for some $b \in \mathbb{R}$. If β^i is continuous at $\beta(a)$ then B is differentiable at $\beta(a)$, with $B'(\beta(a)) = \zeta(\beta^i(\beta(a))) = \zeta(a)$, and (6.40) is an obvious consequence of the convexity of B. Otherwise, a plain reasoning also does the job as

$$\begin{aligned}
B(r) - B(\beta(a)) &= B(\beta(b)) - B(\beta(a)) \\
&= \int_a^b \zeta(q)\beta'(q)\mathrm{d}q \\
&= \int_a^b (\zeta(q) - \zeta(a))\beta'(q)\mathrm{d}q + \zeta(a)(\beta(b) - \beta(a)) \\
&\geq \zeta(a)(r - \beta(a)).
\end{aligned}$$

Here, the inequality comes from the fact that $\beta' \geq 0$ and that $\zeta(q) - \zeta(a)$ has the same sign as $b - a$ if q is between a and b. The general case $r \in \overline{R_\beta}$ is obtained by passing to the limit on b_n such that $\beta(b_n) \to r$, and by using the fact that B has limits (possibly $+\infty$) at the endpoints of R_β.
Let us now take $s, s' \in \mathbb{R}$, and let $\bar{s} \in \mathbb{R}$ be such that $\beta(\bar{s}) = \frac{\beta(s)+\beta(s')}{2}$. We have

$$\int_{\bar{s}}^{s} \beta'(q)\mathrm{d}q + \int_{\bar{s}}^{s'} \beta'(q)\mathrm{d}q = \beta(s) + \beta(s') - 2\beta(\bar{s}) = 0.$$

Hence, using (6.39),

$$
\begin{aligned}
&B(\beta(s))+B(\beta(s'))-2B(\beta(\bar{s}))\\
&\quad= \int_0^s \zeta(q)\beta'(q)\mathrm{d}q + \int_0^s \zeta(q)\beta'(q)\mathrm{d}q - 2\int_0^{\bar{s}} \zeta(q)\beta'(q)\mathrm{d}q\\
&\quad= \int_{\bar{s}}^s \zeta(q)\beta'(q)\mathrm{d}q + \int_{\bar{s}}^{s'} \zeta(q)\beta'(q)\mathrm{d}q\\
&\quad= \int_{\bar{s}}^s (\zeta(q)-\zeta(\bar{s}))\beta'(q)\mathrm{d}q + \int_{\bar{s}}^{s'} (\zeta(q)-\zeta(\bar{s}))\beta'(q)\mathrm{d}q. \qquad (6.42)
\end{aligned}
$$

We then use $|\zeta(q)-\zeta(\bar{s})| \geq \frac{1}{L_\beta}|\nu(q)-\nu(\bar{s})|$ and $\beta'(q) \geq \beta'(q)\frac{\zeta'(q)}{L_\zeta} = \frac{\nu'(q)}{L_\zeta}$ to write

$$
\begin{aligned}
\int_{\bar{s}}^s (\zeta(q)-\zeta(\bar{s}))\beta'(q)\mathrm{d}q &\geq \frac{1}{L_\beta L_\zeta}\int_{\bar{s}}^s \nu'(q)(\nu(q)-\nu(\bar{s}))\mathrm{d}q\\
&= \frac{1}{2L_\beta L_\zeta}(\nu(s)-\nu(\bar{s}))^2.
\end{aligned}
$$

The same relation holds with s replaced by s'. Owing to

$$
(\nu(s)-\nu(s'))^2 \leq 2(\nu(s)-\nu(\bar{s}))^2 + 2(\nu(s')-\nu(\bar{s}))^2,
$$

the inequality (6.41) follows from (6.42). ∎

The following property states an expected integration-by-parts result, which can be formally obtained by writing $(\partial_t\beta(v))\zeta(v) = \beta'(v)\zeta(v)\partial_t v = \partial_t B(\beta(v))$ (owing to (6.39)). The rigorous proof of this result is however a bit technical, due to the lack of regularity on $\overline{u}$ and to the non-linearities involved.

Lemma 6.22. *Let us assume* (6.3b) *and* (6.3c). *Let* $v : \Omega \times (0,T) \to \mathbb{R}$ *be measurable such that*

$$
\begin{aligned}
\zeta(v) \in L^2(0,T;H_0^1(\Omega))\,, \quad B(\beta(v)) \in L^\infty(0,T;L^1(\Omega))\,,\\
\beta(v) \in C([0,T];L^2(\Omega)\text{-w})\,, \quad \partial_t\beta(v) \in L^2(0,T;H^{-1}(\Omega)).
\end{aligned}
$$

Then $t \in [0,T] \to \int_\Omega B(\beta(v)(\boldsymbol{x},t))\mathrm{d}\boldsymbol{x} \in [0,\infty)$ *is continuous and, for all* $t_1, t_2 \in [0,T]$,

$$
\begin{aligned}
&\int_{t_1}^{t_2} \langle \partial_t\beta(v)(t), \zeta(v(t))\rangle_{H^{-1},H_0^1}\mathrm{d}t\\
&\qquad = \int_\Omega B(\beta(v)(\boldsymbol{x},t_2))\mathrm{d}\boldsymbol{x} - \int_\Omega B(\beta(v)(\boldsymbol{x},t_1))\mathrm{d}\boldsymbol{x}. \qquad (6.43)
\end{aligned}
$$

Remark 6.23 (Continuity of $\beta(v)$)
Since $\eta = B \circ \beta$ satisfies (6.18), the condition $B(\beta(v)) \in L^\infty(0,T;L^1(\Omega))$ ensures that $\beta(v) \in L^\infty(0,T;L^2(\Omega))$. Combined with the condition $\partial_t\beta(v) \in$

$L^2(0,T;H^{-1}(\Omega))$, this shows that $\beta(v) \in C([0,T];L^2(\Omega)\text{-w})$. Hence, this continuity property on $\beta(v)$ is actually a consequence of the other assumptions on v.
We also point out that, as in Remark 6.16, it is important to keep in mind the separation between $\beta(v(\cdot,\cdot))$ and its continuous representative $\beta(v)(\cdot,\cdot)$.

Proof.
We obviously only need to make the proof when $t_1 < t_2$.
Step 1: truncation, extension and approximation of $\beta(v)$.
We define $\overline{\beta(v)} : \mathbb{R} \to L^2(\Omega)$ by setting

$$\overline{\beta(v)}(t) = \begin{cases} \beta(v)(t) & \text{if } t \in [t_1,t_2], \\ \beta(v)(t_1) & \text{if } t \le t_1, \\ \beta(v)(t_2) & \text{if } t \ge t_2. \end{cases}$$

By continuity property of $\beta(v)$, we have $\overline{\beta(v)} \in C(\mathbb{R};L^2(\Omega)\text{-w})$ and $\partial_t\overline{\beta(v)} = \mathbf{1}_{(t_1,t_2)}\partial_t\beta(v) \in L^2(\mathbb{R};H^{-1}(\Omega))$ (no Dirac masses have been introduced at $t=t_1$ or $t=t_2$). This regularity of $\partial_t\overline{\beta(v)}$ ensures that the function

$$t \in \mathbb{R} \to D_h\overline{\beta(v)} := \frac{1}{h}\int_t^{t+h} \partial_t\overline{\beta(v)}(s)\mathrm{d}s \tag{6.44}$$

$$= \frac{\overline{\beta(v)}(t+h) - \overline{\beta(v)}(t)}{h} \in H^{-1}(\Omega) \tag{6.45}$$

tend to $\partial_t\overline{\beta(v)}$ in $L^2(\mathbb{R};H^{-1}(\Omega))$ as $h \to 0$.

Step 2: we prove that $\|B(\overline{\beta(v)}(t)\|_{L^1(\Omega)} \le \|B(\beta(v))\|_{L^\infty(0,T;L^1(\Omega))}$ for *all* $t \in \mathbb{R}$ (not only for a.e. t).
Let $t \in [t_1,t_2]$. Since $\beta(v)(\cdot,\cdot) = \beta(v(\cdot,\cdot))$ a.e. on $\Omega \times (t_1,t_2)$, there exists a sequence $t_n \to t$ such that, for all n, $\beta(v)(\cdot,t_n) = \beta(v(\cdot,t_n))$ in $L^2(\Omega)$ and $\|B(\beta(v)(\cdot,t_n))\|_{L^1(\Omega)} \le \|B(\beta(v))\|_{L^\infty(0,T;L^1(\Omega))}$. Using the continuity of $\beta(v)$ with values in $L^2(\Omega)$-w, we have $\beta(v)(\cdot,t_n) \to \beta(v)(\cdot,t)$ weakly in $L^2(\Omega)$. We then use the convexity of B and Lemma D.11 to write, thanks to our choice of t_n,

$$\int_\Omega B(\beta(v)(\boldsymbol{x},t))\mathrm{d}\boldsymbol{x} \le \liminf_{n\to\infty}\int_\Omega B(\beta(v)(\boldsymbol{x},t_n))\mathrm{d}\boldsymbol{x} \le \|B(\beta(v))\|_{L^\infty(0,T;L^1(\Omega))}.$$

The estimate on $B(\overline{\beta(v)})(t)$ is thus complete for $t \in [t_1,t_2]$. The result for $t \le t_1$ or $t \ge t_2$ is obvious since $\overline{\beta(v)}(t)$ is then either $\beta(v)(t_1)$ or $\beta(v)(t_2)$.

Step 3: We prove that for all $\tau \in \mathbb{R}$ and a.e. $t \in (t_1,t_2)$,

$$\langle\overline{\beta(v)}(\tau) - \beta(v)(t), \zeta(v(\cdot,t))\rangle_{H^{-1},H_0^1} \le \int_\Omega B(\overline{\beta(v)}(\boldsymbol{x},\tau)) - B(\beta(v)(\boldsymbol{x},t))\mathrm{d}\boldsymbol{x}. \tag{6.46}$$

Let t such that $\beta(v)(\cdot,t) = \beta(v(\cdot,t))$ a.e. on Ω. Almost every t satisfies this property. By Remark 6.17, for a.e. $\boldsymbol{x} \in \Omega$ we have $\overline{\beta(v)}(\boldsymbol{x},\tau) \in \overline{R_\beta}$ and we can therefore write, by (6.40) with $r = \overline{\beta(v)}(\boldsymbol{x},\tau)$ and $a = v(\boldsymbol{x},t)$,

$$B(\overline{\beta(v)}(\boldsymbol{x},\tau)) - B(\beta(v(\boldsymbol{x},t))) \geq \zeta(v(\boldsymbol{x},t))(\overline{\beta(v)}(\boldsymbol{x},\tau) - \beta(v(\boldsymbol{x},t))).$$

Integrating this relation over $\boldsymbol{x} \in \Omega$, Property (6.46) follows since the H^{-1}–H^1_0 duality product in (6.46) can be replaced with an L^2 inner product, as all terms in this product belong to $L^2(\Omega)$.

Step 4: proof of (6.43)
By convergence of $D_h\overline{\beta(v)}$ to $\partial_t\overline{\beta(v)}$ in $L^2(0,T;H^{-1}(\Omega))$ and since $\mathbf{1}_{(t_1,t_2)}\zeta(v) \in L^2(\mathbb{R};H^1_0(\Omega))$, we have

$$\begin{aligned}
&\int_{t_1}^{t_2} \langle \partial_t\beta(v)(t), \zeta(v(t))\rangle_{H^{-1},H^1_0}\mathrm{d}t \\
&\qquad = \int_{\mathbb{R}} \langle \partial_t\overline{\beta(v)}(t), \mathbf{1}_{(t_1,t_2)}(t)\zeta(v(\cdot,t))\rangle_{H^{-1},H^1_0}\mathrm{d}t \\
&\qquad = \lim_{h\to 0}\int_{\mathbb{R}} \langle D_h\overline{\beta(v)}(t), \mathbf{1}_{(t_1,t_2)}(t)\zeta(v(\cdot,t))\rangle_{H^{-1},H^1_0}\mathrm{d}t \\
&\qquad = \lim_{h\to 0}\frac{1}{h}\int_{t_1}^{t_2} \langle \overline{\beta(v)}(s+h) - \overline{\beta(v)}(t), \zeta(v(\cdot,t)\rangle_{H^{-1},H^1_0}\mathrm{d}t. \qquad (6.47)
\end{aligned}$$

We then use (6.46) for a.e. $t \in (t_1,t_2)$ to obtain, for h small enough such that $t_1 + h < t_2$,

$$\begin{aligned}
&\frac{1}{h}\int_{t_1}^{t_2} \langle \overline{\beta(v)}(t+h) - \overline{\beta(v)}(t), \zeta(v(\cdot,t))\rangle_{H^{-1},H^1_0}\mathrm{d}t \\
&\qquad \leq \frac{1}{h}\int_{t_1}^{t_2}\int_\Omega B(\overline{\beta(v)}(\boldsymbol{x},t+h)) - B(\overline{\beta(v)}(\boldsymbol{x},t))\mathrm{d}\boldsymbol{x}\mathrm{d}t \\
&\qquad = \frac{1}{h}\int_{t_2}^{t_2+h}\int_\Omega B(\overline{\beta(v)}(\boldsymbol{x},t))\mathrm{d}\boldsymbol{x}\mathrm{d}t - \frac{1}{h}\int_{t_1}^{t_1+h}\int_\Omega B(\overline{\beta(v)}(\boldsymbol{x},t))\mathrm{d}\boldsymbol{x}\mathrm{d}t \\
&\qquad = \int_\Omega B(\beta(v)(\boldsymbol{x},t_2))\mathrm{d}\boldsymbol{x} - \frac{1}{h}\int_{t_1}^{t_1+h}\int_\Omega B(\beta(v)(\boldsymbol{x},t))\mathrm{d}\boldsymbol{x}\mathrm{d}t. \qquad (6.48)
\end{aligned}$$

In the last line, we used $\overline{\beta(v)}(t) = \beta(v)(t_2)$ for all $t \geq t_2$. We then take the superior limit of (6.48), and use the fact that $B(\beta(v)(\cdot,t_2))$ is integrable (Step 2) to take its integral out of the lim sup. Coming back to (6.47) we obtain

$$\begin{aligned}
&\int_{t_1}^{t_2} \langle \partial_t\beta(v)(t), \zeta(v(t))\rangle_{H^{-1},H^1_0}\mathrm{d}t \\
&\qquad \leq \int_\Omega B(\beta(v)(\boldsymbol{x},t_2))\mathrm{d}\boldsymbol{x} - \liminf_{h\to 0}\frac{1}{h}\int_{t_1}^{t_1+h}\int_\Omega B(\beta(v)(\boldsymbol{x},t))\mathrm{d}\boldsymbol{x}\mathrm{d}t. \qquad (6.49)
\end{aligned}$$

But since $\beta(v) \in C([0,T]; L^2(\Omega)\text{-w})$, as $h \to 0$ we have $\frac{1}{h}\int_{t_1}^{t_1+h} \beta(v)(t)\mathrm{d}t \to \beta(v)(t_1)$ weakly in $L^2(\Omega)$. Hence, the convexity of B, Lemma D.11 and Jensen's inequality give

$$\int_\Omega B(\beta(v)(\boldsymbol{x},t_1))\mathrm{d}\boldsymbol{x} \le \liminf_{h\to 0} \int_\Omega B\left(\frac{1}{h}\int_{t_1}^{t_1+h} \beta(v)(\boldsymbol{x},t)\mathrm{d}t\right)\mathrm{d}\boldsymbol{x}$$
$$\le \liminf_{h\to 0} \int_\Omega \frac{1}{h}\int_{t_1}^{t_1+h} B(\beta(v)(\boldsymbol{x},t))\mathrm{d}t\mathrm{d}\boldsymbol{x}.$$

Plugged into (6.49), this inequality shows that (6.43) holds with $\le$ instead of $=$. The reverse inequality is obtained by reversing time. We consider $\widetilde{v}(t) = v(t_1+t_2-t)$. Then $\zeta(\widetilde{v})$, $B(\beta(\widetilde{v}))$ and $\beta(\widetilde{v})$ have the same properties as $\zeta(v)$, $B(\beta(v))$ and $\beta(v)$, and $\beta(\widetilde{v})$ takes values $\beta(v)(t_1)$ at $t = t_2$ and $\beta(v)(t_2)$ at $t = t_1$. Applying (6.43) with "$\le$" instead of "$=$" to $\widetilde{v}$ and using the fact that $\partial_t\beta(\widetilde{v})(t) = -\partial_t\beta(v)(t_1+t_2-t)$, we obtain (6.43) with "$\ge$" instead of "$=$" and the proof of (6.43) is complete.
The continuity of $t \in [0,T] \mapsto \int_\Omega B(\beta(v)(\boldsymbol{x},t))\mathrm{d}\boldsymbol{x}$ is straightforward from (6.43), since the left-hand side of this relation is continuous with respect to t_1 and t_2. ∎

The following corollary states continuity properties and an essential formula on the solution to (6.4).

Corollary 6.24. *Under Assumption (6.3), if $\overline{u}$ is a solution of (6.4) then:*

1. *The function $t \in [0,T] \mapsto \int_\Omega B(\beta(\overline{u})(\boldsymbol{x},t))\mathrm{d}\boldsymbol{x} \in [0,\infty)$ is continuous (and thus bounded);*
2. *For any $T_0 \in [0,T]$,*

$$\int_\Omega B(\beta(\overline{u})(\boldsymbol{x},T_0))\mathrm{d}\boldsymbol{x} + \int_0^{T_0}\int_\Omega \Lambda(\boldsymbol{x})\nabla\zeta(\overline{u})(\boldsymbol{x},t)\cdot\nabla\zeta(\overline{u})(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t$$
$$= \int_\Omega B(\beta(u_{\mathrm{ini}}(\boldsymbol{x})))\mathrm{d}\boldsymbol{x} + \int_0^{T_0}\int_\Omega f(\boldsymbol{x},t)\zeta(\overline{u})(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t; \quad (6.50)$$

3. *$\nu(\overline{u})$ is continuous $[0,T] \to L^2(\Omega)$.*

Remark 6.25 (Continuity of $\nu(\overline{u})$)
The continuity of $\nu(\overline{u})$ has to be understood in the same sense as the continuity of $\beta(\overline{u})$ (see Remark 6.16), that is, $\nu(\overline{u})$ is a.e. on $\Omega \times (0,T)$ equal to a continuous function $[0,T] \to L^2(\Omega)$. We use in particular a similar notation $\nu(\overline{u})(\cdot,\cdot)$ for the continuous representative of $\nu(\overline{u}(\cdot,\cdot))$ as we did for the continuous representative of $\beta(\overline{u})$.

Proof.
We first notice that Corollary 6.15 was established using Theorems 6.3 and 6.12, which do not make use of Corollary 6.24. Hence, we invoke Corollary

6.15, which tells us that $\overline{u}$ is also a solution to (6.26). The continuity of $t \in [0,T] \mapsto \int_\Omega B(\beta(\overline{u})(\boldsymbol{x},t))\mathrm{d}\boldsymbol{x} \in [0,\infty)$ and Formula (6.50) therefore follow from Lemma 6.22 applied to $v = \overline{u}$, by using $\overline{v} = \zeta(\overline{u})\mathbf{1}_{[0,T_0]}$ in (6.26).

Let us prove the strong continuity of $\nu(\overline{u}) : [0,T] \mapsto L^2(\Omega)$. Let $\mathcal{T}$ be the set of $\tau \in [0,T]$ such that $\beta(\overline{u}(\cdot,\tau)) = \beta(\overline{u})(\cdot,\tau)$ a.e. on Ω. The set $[0,T]\backslash\mathcal{T}$ has zero measure. Let $(s_l)_{l\in\mathbb{N}}$ and $(t_k)_{k\in\mathbb{N}}$ be two sequences in $\mathcal{T}$ that converge to the same value s. Owing to (6.41),

$$\begin{aligned}
&\int_\Omega [\nu(\overline{u}(\boldsymbol{x},s_l)) - \nu(\overline{u}(\boldsymbol{x},t_k))]^2 \mathrm{d}\boldsymbol{x} \\
&\quad \le 4L_\beta L_\zeta \left(\int_\Omega B(\beta(\overline{u}(\boldsymbol{x},s_l)))\mathrm{d}\boldsymbol{x} + \int_\Omega B(\beta(\overline{u}(\boldsymbol{x},t_k)))\mathrm{d}\boldsymbol{x} \right) \\
&\qquad - 8L_\beta L_\zeta \int_\Omega B\left(\frac{\beta(\overline{u}(\boldsymbol{x},s_l)) + \beta(\overline{u}(\boldsymbol{x},t_k))}{2} \right) \mathrm{d}\boldsymbol{x} \\
&\quad = 4L_\beta L_\zeta \left(\int_\Omega B(\beta(\overline{u})(\boldsymbol{x},s_l))\mathrm{d}\boldsymbol{x} + \int_\Omega B[\beta(\overline{u})(\boldsymbol{x},t_k))\mathrm{d}\boldsymbol{x} \right) \\
&\qquad - 8L_\beta L_\zeta \int_\Omega B\left(\frac{\beta(\overline{u})(\boldsymbol{x},s_l) + \beta(\overline{u})(\boldsymbol{x},t_k)}{2} \right) \mathrm{d}\boldsymbol{x}.
\end{aligned} \tag{6.51}$$

Since $\frac{\beta(\overline{u})(\cdot,s_l)+\beta(\overline{u})(\cdot,t_k)}{2} \to \beta(\overline{u})(\cdot,s)$ weakly in $L^2(\Omega)$ as $l,k \to \infty$, Lemma D.11 and the convexity of B (Lemma 6.21) give

$$\int_\Omega B\left(\beta(\overline{u})(\boldsymbol{x},s)\right)\mathrm{d}\boldsymbol{x} \le \liminf_{l,k\to\infty} \int_\Omega B\left(\frac{\beta(\overline{u})(\boldsymbol{x},s_l) + \beta(\overline{u})(\boldsymbol{x},t_k)}{2} \right) \mathrm{d}\boldsymbol{x}.$$

Taking the superior limit as $l,k \to \infty$ of (6.51) and using the continuity of $t \mapsto \int_\Omega B(\beta(\overline{u})(\boldsymbol{x},t))\mathrm{d}\boldsymbol{x}$ thus shows that

$$\|\nu(\overline{u}(\cdot,s_l)) - \nu(\overline{u}(\cdot,t_k))\|_{L^2(\Omega)} \to 0 \quad \text{as } l,k \to \infty. \tag{6.52}$$

The existence of an a.e. representative of $\nu(\overline{u}(\cdot,\cdot))$ that is continuous $[0,T] \mapsto L^2(\Omega)$ is a direct consequence of this convergence.
Let $s \in [0,T]$ and $(s_l)_{l\in\mathbb{N}} \subset \mathcal{T}$ that converges to s. Applied with $t_k = s_k$, (6.52) shows that $(\nu(\overline{u}(\cdot,s_l)))_{l\in\mathbb{N}}$ is a Cauchy sequence in $L^2(\Omega)$, and therefore that $\lim_{l\to\infty} \nu(\overline{u}(\cdot,s_l))$ exists in $L^2(\Omega)$. Relation (6.52) also shows that this limit, that we can call $\nu(\overline{u})(\cdot,s)$, does not depend on the Cauchy sequence in $\mathcal{T}$ which converges to s. With $t_k = s$, we also see that whenever $s \in \mathcal{T}$ we have $\nu(\overline{u}(\cdot,s)) = \nu(\overline{u})(\cdot,s)$ a.e. on Ω, and $\nu(\overline{u})(\cdot,\cdot)$ is therefore equal to $\nu(\overline{u}(\cdot,\cdot))$ a.e. on $\Omega \times (0,T)$.
It remains to establish that $\nu(\overline{u})$ thus defined is continuous $[0,T] \mapsto L^2(\Omega)$. For any $(\tau_r)_{r\in\mathbb{N}} \subset [0,T]$ which converges to $\tau \in [0,T]$, we can pick $s_r \in \mathcal{T} \cap (\tau_r - \frac{1}{r}, \tau_r + \frac{1}{r})$ and $t_r \in \mathcal{T} \cap (\tau - \frac{1}{r}, \tau + \frac{1}{r})$ such that

$$\|\nu(\overline{u})(\cdot,\tau_r) - \nu(\overline{u}(\cdot,s_r))\|_{L^2(\Omega)} \le \frac{1}{r},$$

$$\|\nu(\overline{u})(\cdot,\tau)-\nu(\overline{u}(\cdot,t_r))\|_{L^2(\Omega)}\leq\frac{1}{r}.$$

We therefore have

$$\sup_{t\in[0,T]}\|\nu(\overline{u})(\cdot,\tau_r)-\nu(\overline{u})(\cdot,\tau)\|_{L^2(\Omega)}$$
$$\leq\frac{2}{r}+\sup_{t\in[0,T]}\|\nu(\overline{u}(\cdot,s_r))-\nu(\overline{u}(\cdot,t_r))\|_{L^2(\Omega)}.$$

By (6.52) with $l=k=r$, this proves that $\nu(\overline{u})(\cdot,\tau_r)\to\nu(\overline{u})(\cdot,\tau)$ in $L^2(\Omega)$ as $r\to\infty$. ■

6.7 Proof of the uniqueness of the solution to the model

We give here a proof of the uniqueness of the solution to (6.4) (and thus also to the solution to (6.7)). The uniqueness of *entropy* solutions to $\partial_t\beta(u)-\Delta\zeta(u)=f$ (with an additional convective term, and a merely integrable f) has been established in [41], using the doubling variable technique. Although this proof could be extended to our framework, we rather provide here a much shorter proof, following the idea due to J. Hadamard [113]. This idea consists in using the solution to an approximate dual problem. It was successfully applied to the one-dimensional Stefan problem in [24], and subsequently generalised to the higher dimensional case in [112].
The proof provided here was originally developed in [78] and applies the approximate duality technique to the doubly degenerate model (6.1), which contains both Richards' and Stefan's models as particular case.

Proof of uniqueness of the solution to (6.4).
Set $u_d=\beta(u_1)+\zeta(u_1)-\beta(u_2)-\zeta(u_2)$, and for all $(\boldsymbol{x},t)\in\Omega\times[0,T]$, define

$$q(\boldsymbol{x},t)=\begin{cases}\frac{\zeta(u_1(\boldsymbol{x},t))-\zeta(u_2(\boldsymbol{x},t))}{u_d(\boldsymbol{x},t)} & \text{if } u_d(\boldsymbol{x},t)\neq 0,\\ 0 & \text{otherwise.}\end{cases}$$

Take $\psi\in L^2(0,T;H_0^1(\Omega))$ with $\partial_t\psi\in L^2(\Omega\times(0,T))$, $\psi(\cdot,T)=0$ and $\mathrm{div}(\Lambda\nabla\psi)\in L^2(\Omega\times(0,T))$. Subtract the two equations (6.4) satisfied by u_1 and u_2, and use ψ as a test function. The assumed regularity $\mathrm{div}(\Lambda\nabla\psi)\in L^2(\Omega\times(0,T))$ enables us to integrate by parts the term involving $\Lambda\nabla\overline{u}\cdot\nabla\psi$, and we obtain

$$\int_0^T\int_\Omega u_d(\boldsymbol{x},t)\Big((1-q(\boldsymbol{x},t))\partial_t\psi(\boldsymbol{x},t)+q(\boldsymbol{x},t)\mathrm{div}(\Lambda\nabla\psi)(\boldsymbol{x},t)\Big)\mathrm{d}\boldsymbol{x}\mathrm{d}t=0. \tag{6.53}$$

For $\varepsilon\in(0,1/2)$ set $q_\varepsilon=(1-2\varepsilon)q+\varepsilon$. Since $0\leq q\leq 1$ we have $\varepsilon\leq q_\varepsilon\leq 1-\varepsilon$, and

$$\frac{(q_\varepsilon - q)^2}{q_\varepsilon} \le \varepsilon \quad \text{and} \quad \frac{(q_\varepsilon - q)^2}{1 - q_\varepsilon} \le \varepsilon. \tag{6.54}$$

Let ψ_ε be given by Lemma 6.26 below, with $g = q_\varepsilon$ and some $w \in C_c^\infty(\Omega \times (0,T))$. Making $\psi = \psi_\varepsilon$ in (6.53) and using (6.57),

$$\begin{aligned}
&\left| \int_0^T \int_\Omega u_d(\boldsymbol{x},t) w(\boldsymbol{x},t) \mathrm{d}\boldsymbol{x}\mathrm{d}t \right| \\
&\le \left| \int_0^T \int_\Omega u_d(\boldsymbol{x},t)(q_\varepsilon(\boldsymbol{x},t) - q(\boldsymbol{x},t))(\mathrm{div}(\Lambda \nabla \psi_\varepsilon)(\boldsymbol{x},t) - \partial_t \psi_\varepsilon(\boldsymbol{x},t)) \mathrm{d}\boldsymbol{x}\mathrm{d}t \right|.
\end{aligned} \tag{6.55}$$

The Cauchy–Schwarz inequality, (6.58) and (6.54) imply

$$\begin{aligned}
&\left[\int_0^T \int_\Omega u_d(x,t)(q_\varepsilon(x,t) - q(x,t))(\mathrm{div}(\Lambda \nabla \psi_\varepsilon)(x,t) - \partial_t \psi_\varepsilon(x,t)) \mathrm{d}x\mathrm{d}t \right]^2 \\
&\quad \le 2 \left(\int_0^T \int_\Omega u_d(\boldsymbol{x},t)^2 \frac{(q(\boldsymbol{x},t) - q_\varepsilon(\boldsymbol{x},t))^2}{q_\varepsilon(\boldsymbol{x},t)} \mathrm{d}\boldsymbol{x}\mathrm{d}t \right) \\
&\qquad \times \left(\int_0^T \int_\Omega q_\varepsilon(\boldsymbol{x},t) \Big(\mathrm{div}(\Lambda \nabla \psi_\varepsilon)(\boldsymbol{x},t) \Big)^2 \mathrm{d}\boldsymbol{x}\mathrm{d}t \right) \\
&\quad + 2 \left(\int_0^T \int_\Omega u_d(\boldsymbol{x},t)^2 \frac{(q(\boldsymbol{x},t) - q_\varepsilon(x,t))^2}{1 - q_\varepsilon(\boldsymbol{x},t)} \mathrm{d}\boldsymbol{x}\mathrm{d}t \right) \\
&\qquad \times \left(\int_0^T \int_\Omega (1 - q_\varepsilon(\boldsymbol{x},t)) \Big(\partial_t \psi_\varepsilon(\boldsymbol{x},t) \Big)^2 \mathrm{d}\boldsymbol{x}\mathrm{d}t \right) \\
&\quad \le 2\varepsilon C_0 \left\| u_d \right\|_{L^2(\Omega \times (0,T))} \\
&\qquad \times \left(\left\| \nabla w \right\|^2_{L^2(\Omega \times (0,T))^d} + \left\| w \right\|^2_{L^2(\Omega \times (0,T))} + \left\| \partial_t w \right\|^2_{L^2(\Omega \times (0,T))} \right).
\end{aligned} \tag{6.56}$$

Letting $\varepsilon \to 0$ and using (6.55) gives

$$\int_0^T \int_\Omega u_d(\boldsymbol{x},t) w(\boldsymbol{x},t) \mathrm{d}\boldsymbol{x}\mathrm{d}t = 0.$$

Since this holds for any function $w \in C_c^\infty(\Omega \times (0,T))$, we deduce that $u_d = 0$ a.e. on $\Omega \times (0,T)$. Hence $\beta(u_1) + \zeta(u_1) = \beta(u_2) + \zeta(u_2)$, and the proof is complete since $\beta + \zeta$ is one-to-one. ∎

The following lemma ensures the existence of the function ψ, used in the proof above.

Lemma 6.26. *Let $T > 0$, and let Ω be a bounded open subset of $\mathbb{R}^d$ ($d \in \mathbb{N}$). Assume Hypothesis (6.3e). Let $w \in C_c^\infty(\Omega \times (0,T))$ and $g \in L^\infty(\Omega \times (0,T))$ such that $g(\boldsymbol{x},t) \in [g_{\min}, 1 - g_{\min}]$ for a.e. $(\boldsymbol{x},t) \in \Omega \times (0,T)$, where $g_{\min}$ is a fixed number in $(0, \frac{1}{2})$. Then there exists a function ψ such that:*

1. *$\psi \in L^\infty(0,T;H_0^1(\Omega))$, $\partial_t\psi \in L^2(\Omega\times(0,T))$, $\mathrm{div}(\Lambda\nabla\psi) \in L^2(\Omega\times(0,T))$ (this implies $\psi \in C([0,T];L^2(\Omega))$);*
2. *$\psi(\cdot,T) = 0$;*
3. *For a.e. $(\boldsymbol{x},t) \in \Omega\times(0,T)$,*

$$(1-g(\boldsymbol{x},t))\partial_t\psi(\boldsymbol{x},t) + g(\boldsymbol{x},t)\mathrm{div}(\Lambda\nabla\psi)(\boldsymbol{x},t) = w(\boldsymbol{x},t); \tag{6.57}$$

4. *There exists $C_0 > 0$, depending only on T, $\mathrm{diam}(\Omega)$, $\underline{\lambda}$ and $\overline{\lambda}$ (and not on $g_{\min}$), such that*

$$\int_0^T\int_\Omega \left((1-g(\boldsymbol{x},t))\Big(\partial_t\psi(\boldsymbol{x},t)\Big)^2 + g(\boldsymbol{x},t)\Big(\mathrm{div}(\Lambda\nabla\psi)(\boldsymbol{x},t)\Big)^2\right)\mathrm{d}\boldsymbol{x}\mathrm{d}t$$
$$\leq C_0\left(\|\nabla w\|^2_{L^2(\Omega\times(0,T))^d} + \|w\|^2_{L^2(\Omega\times(0,T))} + \|\partial_t w\|^2_{L^2(\Omega\times(0,T))}\right). \tag{6.58}$$

Proof.

Step 1: existence of ψ satisfying 1, 2 and 3.
After dividing through by g, observe that (6.57) is equivalent to

$$\text{For a.e. } (\boldsymbol{x},t) \in \Omega\times(0,T)\,,$$
$$\Phi(\boldsymbol{x},t)\partial_t\psi(\boldsymbol{x},t) + \mathrm{div}(\Lambda(\boldsymbol{x})\nabla\psi(\boldsymbol{x},t)) = f(\boldsymbol{x},t), \tag{6.59}$$

where $f \in L^\infty(\Omega\times(0,T))$, $\Phi \in L^\infty(\Omega\times(0,T))$ and, for some fixed numbers $\varphi^* \geq \varphi_* > 0$, $\varphi_* \leq \Phi(\boldsymbol{x},t) \leq \varphi^*$ for a.e. $(\boldsymbol{x},t) \in \Omega\times(0,T)$. The parabolic equation (6.59) is slightly non-standard because of the time-dependent coefficient Φ in front of $\partial_t\psi$. However, as we shall now see, a standard Galerkin approximation provides the existence of a solution to this equation.
Let $(V_k)_{k\in\mathbb{N}}$ be a non-decreasing family of finite-dimensional subspaces of $H_0^1(\Omega)$ such that $\overline{\bigcup_{k\in\mathbb{N}} V_k} = H_0^1(\Omega)$. We look for $\psi_k : [0,T] \to V_k$ solution to the following Galerkin approximation of (6.59), with final condition:

$$\begin{aligned}&\psi_k(T) = 0 \text{ and } \forall t \in [0,T]\,, \forall v \in V_k\,,\\ &(\Phi(\cdot,t)\psi_k'(t),v)_{L^2} - (\Lambda\nabla\psi_k(t),\nabla v)_{(L^2)^d} = (f(\cdot,t),v)_{L^2}.\end{aligned} \tag{6.60}$$

Here, $(\cdot,\cdot)_{L^2}$ is the $L^2(\Omega)$ inner product. Choosing an orthonormal (for this inner product) basis $(\mathbf{e}_i)_{i=1,\ldots,N_k}$ of V_k and writing $\psi_k(t) = \sum_{i=1}^{N_k}\theta_i(t)\mathbf{e}_i$, (6.60) can be re-cast as

$$\Theta(T) = 0 \text{ and, for all } t \in [0,T],\ M(t)\Theta'(t) - S(t)\Theta(t) = F(t) \tag{6.61}$$

where $\Theta(t) = (\theta_i(t))_{i=1,\ldots,N_k}$, $M(t)$ and $S(t)$ are the symmetric matrices with respective entries $M_{i,j}(t) = (\Phi(\cdot,t)\mathbf{e}_i,\mathbf{e}_j)_{L^2}$ and $S_{i,j}(t) = (\Lambda\nabla\mathbf{e}_i,\nabla\mathbf{e}_j)_{(L^2)^d}$, and $F(t) = ((f(\cdot,t),\mathbf{e}_j))_{j=1,\ldots,N_k}$. Since $\Phi \geq \varphi^*$ and $(\mathbf{e}_i)_{i=1,\ldots,N_k}$ is orthonormal for $(\cdot,\cdot)_{L^2}$, it holds $M(t) \geq \varphi_*\mathrm{Id}$. $M(t)^{-1}$ is therefore well defined and measurable bounded over $[0,T]$. Hence, the initial value problem (6.61) can be

put in standard form, with bounded measurable coefficients, and it therefore has a unique solution Θ such that Θ' is bounded.
There exists thus a unique solution ψ_k to (6.60), with $\psi_k \in W^{1,\infty}(0,T;V_k) \subset W^{1,\infty}(0,T;H_0^1(\Omega))$. Let us now prove some *a priori* estimates on ψ_k. We make, for a.e. $t \in (0,T)$, $w = \psi_k'(t)$ in (6.60) and we integrate over $t \in (\tau,T)$, for some $\tau \in (0,T)$. Since Λ is symmetric and does not depend on t,

$$(\Lambda\nabla\psi_k(t), \nabla\psi_k'(t))_{(L^2)^d} = \frac{1}{2}\frac{d}{dt}(\Lambda\nabla\psi_k(t), \nabla\psi_k(t))_{(L^2)^d}$$

and we therefore obtain, using $\psi_k(\cdot,T) = 0$ and the Young inequality (D.9),

$$\begin{aligned}\int_\tau^T \int_\Omega \Phi(\boldsymbol{x},t)|\partial_t\psi_k(\boldsymbol{x},t)|^2 \mathrm{d}\boldsymbol{x}\mathrm{d}t + \frac{1}{2}\int_\Omega \Lambda(\boldsymbol{x})\nabla\psi_k(\boldsymbol{x},\tau)\cdot\nabla\psi_k(\boldsymbol{x},\tau)\mathrm{d}\boldsymbol{x} \\ \leq \|f\|_{L^2(\Omega\times(0,T))}\|\partial_t\psi_k\|_{L^2(\Omega\times(\tau,T)} \\ \leq \frac{1}{2\varphi_*}\|f\|_{L^2(\Omega\times(0,T))} + \frac{\varphi_*}{2}\|\partial_t\psi_k\|^2_{L^2(\Omega\times(\tau,T)}\,.\end{aligned}$$

This estimate holds for any $\tau \in (0,T)$. Given that Λ is uniformly coercive and that $\Phi \geq \varphi_*$, we deduce that $(\psi_k)_{k\in\mathbb{N}}$ is bounded in $L^\infty(0,T;H_0^1(\Omega))$ and that $(\partial_t\psi_k)_{k\in\mathbb{N}}$ is bounded in $L^2(\Omega\times(0,T))$. Hence, there exists $\psi \in L^\infty(0,T;H_0^1(\Omega))$ such that $\partial_t\psi \in L^2(\Omega\times(0,T))$ and, up to a subsequence as $k\to\infty$, $\psi_k \to \psi$ weakly-$*$ in $L^\infty(0,T;H_0^1(\Omega))$ and $\partial_t\psi_k \to \partial_t\psi$ weakly in $L^2(\Omega\times(0,T))$. Using the Aubin–Simon theorem, we also see that the convergence of $(\psi_k)_{k\in\mathbb{N}}$ holds in $C([0,T];L^2(\Omega))$, which ensures that $\psi(\cdot,T) = 0$.
We then take $\theta \in C_c^\infty(0,T)$ and $v \in V_\ell$ for some $\ell \in \mathbb{N}$, and apply (6.60) for $k \geq \ell$ to $\theta(t)v$ instead of v. Integrating the resulting equation over $t \in (0,T)$, we can take the limit and see that ψ satisfies, with $\rho(\boldsymbol{x},t) = \theta(t)v(\boldsymbol{x})$,

$$\begin{aligned}\int_0^T\int_\Omega \Phi(\boldsymbol{x},t)\partial_t\psi(\boldsymbol{x},t)\rho(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t - \int_0^T\int_\Omega \Lambda(\boldsymbol{x})\nabla\psi(\boldsymbol{x},t)\cdot\nabla\rho(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t \\ = \int_0^T\int_\Omega f(\boldsymbol{x},t)\rho(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t. \quad (6.62)\end{aligned}$$

Any function ρ in $L^2(0,T;H_0^1(\Omega))$ can be approximated in this space by finite sums of functions $(\boldsymbol{x},t) \to \theta(t)v(\boldsymbol{x})$, with $\theta \in C_c^\infty(0,T)$ and $v \in \cup_{\ell\in\mathbb{N}}V_\ell$ (see [67]). Hence, (6.62) also holds for any $\rho \in L^2(0,T;H_0^1(\Omega))$. Considering smooth compactly supported functions ρ, (6.62) shows that $\mathrm{div}(\Lambda\nabla\psi) = f - \Phi\partial_t\psi$ in the sense of distributions. This proves that $\mathrm{div}(\Lambda\nabla\psi) \in L^2(\Omega\times(0,T))$ and thus, by (6.62), that (6.59) is satisfied.
Note that Lemma 6.27 below provides an additional regularity property and an integration-by-part formula on ψ.

Step 2: proof of (6.58).
Taking $s,\tau \in [0,T]$, we have

$$\int_s^\tau \int_\Omega w(\boldsymbol{x},t)\mathrm{div}(\Lambda\nabla\psi)(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t = -\int_s^\tau \int_\Omega \Lambda(\boldsymbol{x})\nabla w(\boldsymbol{x},t)\cdot\nabla\psi(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t,$$

and

$$\int_s^\tau \int_\Omega w(\boldsymbol{x},t)\partial_t\psi(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t = \int_\Omega (w(\boldsymbol{x},\tau)\psi(\boldsymbol{x},\tau) - w(\boldsymbol{x},s)\psi(\boldsymbol{x},s))\mathrm{d}\boldsymbol{x} - \int_s^\tau \int_\Omega \psi(\boldsymbol{x},t)\partial_t w(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t.$$

Multiplying (6.57) by $\partial_t\psi(\boldsymbol{x},t) + \mathrm{div}(\Lambda\nabla\psi)(\boldsymbol{x},t)$, integrating over $\Omega\times(s,T)$ for $s\in[0,T]$, using (6.66) in Lemma 6.27, and recalling that $\psi(\cdot,T)=0$, we obtain

$$\begin{aligned}
&\frac{1}{2}\int_\Omega \Lambda(\boldsymbol{x})\nabla\psi(\boldsymbol{x},s)\cdot\nabla\psi(\boldsymbol{x},s)\mathrm{d}\boldsymbol{x}\\
&\quad+\int_s^T\int_\Omega \left((1-g(\boldsymbol{x},t))\Big(\partial_t\psi(\boldsymbol{x},t)\Big)^2 + g(\boldsymbol{x},t)\Big(\mathrm{div}(\Lambda\nabla\psi)(\boldsymbol{x},t)\Big)^2\right)\mathrm{d}\boldsymbol{x}\mathrm{d}t\\
&\quad= -\int_s^T\int_\Omega \Lambda(\boldsymbol{x})\nabla w(\boldsymbol{x},t)\cdot\nabla\psi(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t - \int_\Omega w(\boldsymbol{x},s)\psi(\boldsymbol{x},s)\mathrm{d}\boldsymbol{x}\\
&\qquad-\int_s^T\int_\Omega \psi(\boldsymbol{x},t)\partial_t w(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t.
\end{aligned}\tag{6.63}$$

Integrating (6.63) with respect to $s\in(0,T)$ leads to

$$\begin{aligned}
&\frac{1}{2}\int_0^T\int_\Omega \Lambda(\boldsymbol{x})\nabla\psi(\boldsymbol{x},s)\cdot\nabla\psi(\boldsymbol{x},s)\mathrm{d}\boldsymbol{x}\mathrm{d}s\\
&\le T\int_0^T\int_\Omega |\Lambda(\boldsymbol{x})\nabla w(\boldsymbol{x},t)\cdot\nabla\psi(\boldsymbol{x},t)|\mathrm{d}\boldsymbol{x}\mathrm{d}t + \int_0^T\int_\Omega |w(\boldsymbol{x},s)\psi(\boldsymbol{x},s)|\mathrm{d}\boldsymbol{x}\mathrm{d}s\\
&\quad+T\int_0^T\int_\Omega |\psi(\boldsymbol{x},t)\partial_t w(\boldsymbol{x},t)|\mathrm{d}\boldsymbol{x}\mathrm{d}t.
\end{aligned}\tag{6.64}$$

Apply the Cauchy–Schwarz and Poincaré inequalities to obtain

$$\begin{aligned}
\frac{\lambda}{2}\|\nabla\psi\|_{L^2(\Omega\times(0,T))^d} &\le T\overline{\lambda}\|\nabla w\|_{L^2(\Omega\times(0,T))}\\
&\quad+\mathrm{diam}(\Omega)\left(\|w\|_{L^2(\Omega\times(0,T))} + T\|\partial_t w\|_{L^2(\Omega\times(0,T))}\right).
\end{aligned}\tag{6.65}$$

Letting $s=0$ in (6.63), recalling that $w(\cdot,0)=0$, and using (6.65) gives

$$\begin{aligned}
&\int_0^T\int_\Omega \left((1-g(\boldsymbol{x},t))\Big(\partial_t\psi(\boldsymbol{x},t)\Big)^2 + g(\boldsymbol{x},t)\Big(\mathrm{div}(\Lambda\nabla\psi)(\boldsymbol{x},t)\Big)^2\right)\mathrm{d}\boldsymbol{x}\mathrm{d}t\\
&\le\left(\overline{\lambda}\|\nabla w\|_{L^2(\Omega\times(0,T))} + \mathrm{diam}(\Omega)\|\partial_t w\|_{L^2(\Omega\times(0,T))}\right)\|\nabla\psi\|_{L^2(\Omega\times(0,T))^d}.
\end{aligned}$$

Combined with (6.65), this shows that (6.58) holds. ∎

Lemma 6.27. *Assume that Ω, T and Λ satisfy (6.3a) and (6.3e). Let $\psi \in L^\infty(0,T;H_0^1(\Omega))$ such that $\partial_t\psi$ and $\mathrm{div}(\Lambda\nabla\psi)$ belong to $L^2(\Omega\times(0,T))$. Then $\psi \in C([0,T];H_0^1(\Omega))$ and, for all $s,\tau\in[0,T]$,*

$$\int_s^\tau \int_\Omega \partial_t\psi(\boldsymbol{x},t)\mathrm{div}(\Lambda\nabla\psi)(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t = -\frac{1}{2}\int_\Omega \Lambda(\boldsymbol{x})\nabla\psi(\boldsymbol{x},\tau)\cdot\nabla\psi(\boldsymbol{x},\tau)\mathrm{d}\boldsymbol{x} + \frac{1}{2}\int_\Omega \Lambda(\boldsymbol{x})\nabla\psi(\boldsymbol{x},s)\cdot\nabla\psi(\boldsymbol{x},s)\mathrm{d}\boldsymbol{x}. \quad (6.66)$$

Proof.
Step 1: $\psi \in C([0,T];L^2(\Omega))$ and $\psi : [0,T] \to H_0^1(\Omega)$ is continuous for the weak topology of $H_0^1(\Omega)$.
Since $\psi \in L^\infty(0,T;H_0^1(\Omega)) \subset L^2(0,T;L^2(\Omega))$ and $\partial_t\psi \in L^2(0,T;L^2(\Omega))$, we have $\psi \in H^1(0,T;L^2(\Omega)) \subset C([0,T];L^2(\Omega))$.
Let $M = \|\psi\|_{L^\infty(0,T;H_0^1(\Omega))}$ and let $t\in[0,T]$. There exists $(t_n)_{n\in\mathbb{N}}$ converging to t such that $\|\psi(t_n)\|_{H_0^1(\Omega)} \le M$. Since ψ is continuous with values in $L^2(\Omega)$, we have $\psi(t_n) \to \psi(t)$ in $L^2(\Omega)$. Given the bound on $\|\psi(t_n)\|_{H_0^1(\Omega)}$, this convergence also holds in $H_0^1(\Omega)$, and $\|\psi(t)\|_{H_0^1(\Omega)} \le M$. In other words, M is not just an essential bound of $\|\psi(\cdot)\|_{H_0^1(\Omega)}$, but actually a pointwise bound.
Let us now prove the weak continuity of ψ. Let $t\in[0,T]$ and $t_n \to t$. If $\gamma \in C_c^\infty(\Omega)$ we have

$$(\psi(t_n),\gamma)_{H_0^1} = \int_\Omega \nabla\psi(\boldsymbol{x},t_n)\cdot\nabla\gamma(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\mathrm{d}t = -\int_\Omega \psi(\boldsymbol{x},t_n)\Delta\gamma(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\mathrm{d}t$$

and thus, as $n\to\infty$, since $\psi \in C([0,T];L^2(\Omega))$,

$$\begin{aligned}(\psi(t_n),\gamma)_{H_0^1} \to & -\int_\Omega \psi(\boldsymbol{x},t)\Delta\gamma(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\mathrm{d}t \\ &= \int_\Omega \nabla\psi(\boldsymbol{x},t)\cdot\nabla\gamma(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\mathrm{d}t = (\psi(t),\gamma)_{H_0^1}. \end{aligned} \quad (6.67)$$

If $\gamma \in H_0^1(\Omega)$ then we take $\gamma_\varepsilon \in C_c^\infty(\Omega)$ such that $\|\gamma-\gamma_\varepsilon\|_{H_0^1(\Omega)} \le \varepsilon$ and we classically write

$$\begin{aligned}&\Big|(\psi(t_n),\gamma)_{H_0^1} - (\psi(t),\gamma)_{H_0^1}\Big| \\ &\quad\le \Big|(\psi(t_n),\gamma)_{H_0^1} - (\psi(t_n),\gamma_\varepsilon)_{H_0^1}\Big| + \Big|(\psi(t_n),\gamma_\varepsilon)_{H_0^1} - (\psi(t),\gamma_\varepsilon)_{H_0^1}\Big| \\ &\qquad + \Big|(\psi(t),\gamma_\varepsilon)_{H_0^1} - (\psi(t),\gamma)_{H_0^1}\Big| \\ &\quad\le M\varepsilon + \Big|(\psi(t_n),\gamma_\varepsilon)_{H_0^1} - (\psi(t),\gamma_\varepsilon)_{H_0^1}\Big| + M\varepsilon.\end{aligned}$$

Taking the superior limit as $n\to\infty$ (using (6.67) with γ_ε instead of γ), and then the limit as $\varepsilon\to 0$, we deduce that that $(\psi(t_n),\gamma)_{H_0^1} \to (\psi(t),\gamma)_{H_0^1}$ as $n\to\infty$. This concludes the proof of the continuity of $\psi : [0,T] \to H_0^1(\Omega)$-w.

Step 2: proof of (6.66).
We only have to consider the case $s < \tau$. We truncate ψ to $[s, \tau]$ and extend it by its constant values at the endpoints of this interval, which consists in defining $\overline{\psi}$ on $\mathbb{R}$ by

$$\overline{\psi}(t) = \begin{cases} \psi(s) & \text{if } t \leq s, \\ \psi(t) & \text{if } t \in (s, \tau), \\ \psi(\tau) & \text{if } t \geq \tau. \end{cases}$$

Since $\psi \in C([0,T]; L^2(\Omega)) \cap C([0,T]; H_0^1(\Omega)\text{-w})$, this definition makes sense and we have $\overline{\psi} \in C(\mathbb{R}; L^2(\Omega)) \cap C(\mathbb{R}; H_0^1(\Omega)\text{-w})$. By these continuity properties, we have $\partial_t \overline{\psi} = \mathbf{1}_{(s,\tau)} \partial_t \psi$ since no Dirac masses are introduced at s or τ. We also have, on (s, τ), $\operatorname{div}(\Lambda \nabla \overline{\psi}) = \operatorname{div}(\Lambda \nabla \psi) \in L^2(\Omega \times (s, \tau))$. However, because we cannot ensure that $\operatorname{div}(\Lambda \nabla \psi(\tau))$ and $\operatorname{div}(\Lambda \nabla \psi(s))$ belongs to $L^2(\Omega)$, we cannot say that $\operatorname{div}(\Lambda \nabla \overline{\psi}) \in L^2(\Omega \times \mathbb{R})$. We only have $\operatorname{div}(\Lambda \nabla \overline{\psi}) \in C(\mathbb{R}; H^{-1}(\Omega)\text{-w})$, owing to $\overline{\psi} \in C(\mathbb{R}; H_0^1(\Omega)\text{-w})$.

Let $(\rho_n)_{n \in \mathbb{N}}$ be a smoothing kernel in time, such that $\operatorname{supp}(\rho_n) \subset (-(\tau - s), 0)$. We set $\overline{\psi}_n(\boldsymbol{x}, t) = (\overline{\psi}(\boldsymbol{x}, \cdot) * \rho_n)(t)$. Then $\overline{\psi}_n \in C^\infty(\mathbb{R}; H_0^1(\Omega))$ and we can write, since Λ is symmetric and does not depend on time,

$$\begin{aligned} &\int_s^\tau \langle \partial_t \overline{\psi}_n(t), \operatorname{div}(\Lambda \nabla \overline{\psi}_n)(t) \rangle_{H_0^1, H^{-1}} \mathrm{d}t \\ &\qquad = -\int_s^\tau \int_\Omega \partial_t \nabla \overline{\psi}_n(\boldsymbol{x}, t) \cdot \Lambda(\boldsymbol{x}) \nabla \overline{\psi}_n(\boldsymbol{x}, t) \mathrm{d}\boldsymbol{x} \mathrm{d}t \\ &\qquad = -\frac{1}{2} \int_s^\tau \frac{d}{dt} \int_\Omega \Lambda(\boldsymbol{x}) \nabla \overline{\psi}_n(\boldsymbol{x}, t) \cdot \nabla \overline{\psi}_n(\boldsymbol{x}, t) \mathrm{d}\boldsymbol{x} \mathrm{d}t \\ &\qquad = -\frac{1}{2} \int_\Omega \Lambda(\boldsymbol{x}) \nabla \overline{\psi}_n(\boldsymbol{x}, \tau) \cdot \nabla \overline{\psi}_n(\boldsymbol{x}, \tau) \mathrm{d}\boldsymbol{x} \\ &\qquad\quad + \frac{1}{2} \int_\Omega \Lambda(\boldsymbol{x}) \nabla \overline{\psi}_n(\boldsymbol{x}, s) \cdot \nabla \overline{\psi}_n(\boldsymbol{x}, s) \mathrm{d}\boldsymbol{x}. \end{aligned} \tag{6.68}$$

We aim at passing to the limit $n \to \infty$ in this relation. By choice of $\operatorname{supp}(\rho_n)$ and by definition of $\overline{\psi}$,

$$\begin{aligned} \overline{\psi}_n(\boldsymbol{x}, \tau) &= \int_{\mathbb{R}} \overline{\psi}(\boldsymbol{x}, q) \rho_n(\tau - q) \mathrm{d}q \\ &= \int_\tau^\infty \overline{\psi}(\boldsymbol{x}, q) \rho_n(\tau - q) \mathrm{d}s = \psi(\boldsymbol{x}, \tau) \int_\tau^\infty \rho_n(\tau - q) \mathrm{d}q = \psi(\boldsymbol{x}, \tau). \end{aligned}$$

Hence, for all $n \in \mathbb{N}$,

$$\begin{aligned} &\frac{1}{2} \int_\Omega \Lambda(\boldsymbol{x}) \nabla \overline{\psi}_n(\boldsymbol{x}, \tau) \cdot \nabla \overline{\psi}_n(\boldsymbol{x}, \tau) \mathrm{d}\boldsymbol{x} \\ &\qquad\qquad = \frac{1}{2} \int_\Omega \Lambda(\boldsymbol{x}) \nabla \psi(\boldsymbol{x}, \tau) \cdot \nabla \psi(\boldsymbol{x}, \tau) \mathrm{d}\boldsymbol{x}. \end{aligned} \tag{6.69}$$

Since $\overline{\psi} \in C(\mathbb{R}; H_0^1(\Omega)\text{-w})$, as $n \to \infty$ we have $\overline{\psi}_n(s) \to \overline{\psi}(s) = \psi(s)$ weakly in $H_0^1(\Omega)$. The bilinear form $(\Lambda\nabla\cdot, \nabla\cdot)_{(L^2)^d}$ being a Hilbert norm in $H_0^1(\Omega)$, we infer that

$$\liminf_{n\to\infty} \frac{1}{2}\int_\Omega \Lambda(\boldsymbol{x})\nabla\overline{\psi}_n(\boldsymbol{x},s)\cdot\nabla\overline{\psi}_n(\boldsymbol{x},s)\mathrm{d}\boldsymbol{x} \geq \frac{1}{2}\int_\Omega \Lambda(\boldsymbol{x})\nabla\psi(\boldsymbol{x},s)\cdot\nabla\psi(\boldsymbol{x},s)\mathrm{d}\boldsymbol{x}. \quad (6.70)$$

Dealing with the left-hand side of (6.68) is a bit more challenging, due to the lack of regularity of $\mathrm{div}(\Lambda\nabla\overline{\psi})$ outside (s,τ). By definition of $\overline{\psi}$, we have

$$\mathrm{div}(\Lambda\nabla\overline{\psi}(t)) = \mathbf{1}_{(-\infty,s]}(t)\mathrm{div}(\Lambda\nabla\psi(s)) + \mathbf{1}_{(s,\tau)}(t)(\mathrm{div}(\Lambda\nabla\psi(t)) + \mathbf{1}_{[\tau,+\infty)}(t)(\mathrm{div}(\Lambda\nabla\psi(\tau)).$$

The choice of the support of ρ_n ensures that, whenever $t > s$, $\mathbf{1}_{(-\infty,s]} * \rho_n(t) = 0$. Hence, for $t \in (s,\tau)$,

$$\mathrm{div}(\Lambda\nabla\overline{\psi}_n(t)) = \left[\mathrm{div}(\Lambda\nabla\psi(\cdot))\mathbf{1}_{(s,\tau)}\right] * \rho_n(t) + (\mathbf{1}_{[\tau,+\infty)} * \rho_n)(t)(\mathrm{div}(\Lambda\nabla\psi(\tau)).$$

Since $\mathrm{div}(\Lambda\nabla\psi)\mathbf{1}_{(s,\tau)} \in L^2(\Omega\times\mathbb{R})$, the left-hand side of (6.68) can therefore be re-cast as

$$\begin{aligned}
&\int_s^\tau \langle\partial_t\overline{\psi}_n(t), \mathrm{div}(\Lambda\nabla\overline{\psi}_n)(t)\rangle_{H_0^1,H^{-1}}\mathrm{d}t \\
&\quad= \int_s^\tau\int_\Omega \partial_t\overline{\psi}_n(\boldsymbol{x},t)\left[\mathrm{div}(\Lambda\nabla\psi)(\boldsymbol{x},\cdot)\mathbf{1}_{(s,\tau)}\right] * \rho_n(t)\mathrm{d}\boldsymbol{x}\mathrm{d}t \\
&\qquad+ \int_s^\tau \langle\partial_t\overline{\psi}_n(t), \mathrm{div}(\Lambda\nabla\psi(\tau))\rangle_{H_0^1,H^{-1}}(\mathbf{1}_{[\tau,\infty)} * \rho_n)(t)\mathrm{d}t \\
&\quad= \int_s^\tau\int_\Omega \partial_t\overline{\psi}_n(\boldsymbol{x},t)\left[\mathrm{div}(\Lambda\nabla\psi)(\boldsymbol{x},\cdot)\mathbf{1}_{(s,\tau)}\right] * \rho_n(t)\mathrm{d}\boldsymbol{x}\mathrm{d}t + T_n, \qquad (6.71)
\end{aligned}$$

where $T_n = \int_s^\tau F_n'(t)(\mathbf{1}_{[\tau,\infty)} * \rho_n)(t)\mathrm{d}t$ with

$$F_n(t) = F * \rho_n(t)\,, \quad F(t) = \langle\overline{\psi}(t), \mathrm{div}(\Lambda\nabla\psi(\tau)\rangle_{H_0^1,H^{-1}}.$$

Integrating-by-parts, we have

$$\begin{aligned}
T_n &= F_n(\tau)(\mathbf{1}_{[\tau,\infty)} * \rho_n)(\tau) - F_n(s)(\mathbf{1}_{[\tau,\infty)} * \rho_n)(s) \\
&\quad - \int_s^\tau F_n(t)(\mathbf{1}_{[\tau,\infty)} * \rho_n)'(t)\mathrm{d}t.
\end{aligned}$$

The choice of support of ρ_n ensures that $(\mathbf{1}_{[\tau,\infty)} * \rho_n)(s) = 0$ and that $(\mathbf{1}_{[\tau,\infty)} * \rho_n)(\tau) = 1$. We also notice that $(\mathbf{1}_{[\tau,\infty)} * \rho_n)' = \delta_\tau * \rho_n$ has support in (s,τ) and converges weakly in the sense of measures toward the Dirac mass δ_τ. Since

$\overline{\psi} \in C(\mathbb{R}; H_0^1(\Omega)\text{-w})$, we have $F \in C(\mathbb{R})$ and thus $F_n \to F$ locally uniformly on $\mathbb{R}$. Hence, as $n \to \infty$,

$$T_n = F_n(\tau) - \int_s^\tau F_n(t)(\mathbf{1}_{[\tau,\infty)} * \rho_n)'(t)\mathrm{d}t \to F(\tau) - F(\tau) = 0. \tag{6.72}$$

The functions $\partial_t \overline{\psi}$ and $\mathrm{div}(\Lambda\nabla\psi)(\boldsymbol{x},\cdot)\mathbf{1}_{(s,\tau)}$ belong to $L^2(\Omega \times \mathbb{R})$, so

$$\partial_t \overline{\psi}_n = (\partial_t \overline{\psi}) * \rho_n \to \partial_t \overline{\psi} \quad \text{ in } L^2(\Omega \times \mathbb{R})$$

and

$$\left[\mathrm{div}(\Lambda\nabla\psi)(\boldsymbol{x},\cdot)\mathbf{1}_{(s,\tau)}\right] * \rho_n \to \mathrm{div}(\Lambda\nabla\psi)(\boldsymbol{x},\cdot)\mathbf{1}_{(s,\tau)} \quad \text{ in } L^2(\Omega \times \mathbb{R}).$$

Using (6.72), we can therefore pass to the limit in (6.71) and we see, since $\partial_t \overline{\psi} = \partial_t \psi$ on $\Omega \times (s,\tau)$,

$$\lim_{n\to\infty} \int_s^\tau \langle \partial_t \overline{\psi}_n(t), \mathrm{div}(\Lambda\nabla\overline{\psi}_n)(t)\rangle_{H_0^1, H^{-1}} \mathrm{d}t$$
$$= \int_s^\tau \int_\Omega \partial_t \psi(\boldsymbol{x},t)\mathrm{div}(\Lambda\nabla\psi)(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t.$$

Combined with (6.68), (6.69) and (6.70), this gives (6.66) with "$\geq$" instead of "$=$". The converse inequality is obtained by re-doing the previous reasoning with smoothing kernels ρ_n having support in $(0, \tau - s)$, or by reversing the time as at the end of the proof of Lemma 6.22.

Step 3: proof that $\psi : [0,T] \to H_0^1(\Omega)$ is continuous for the strong topology of $H_0^1(\Omega)$.
Since the left-hand side of (6.66) is continuous with respect to s, the mapping $s \to (\Lambda\nabla\psi(s), \nabla\psi(s))_{(L^2)^d}$ is continuous. Assume that $s_n \to s$ in $[0,T]$. Owing to $\psi \in C([0,T]; H_0^1(\Omega)\text{-w})$ we have $\psi(s_n) \to \psi(s)$ weakly in $H_0^1(\Omega)$. Moreover, $(\Lambda\nabla\psi(s_n), \nabla\psi(s_n))_{(L^2)^d} \to (\Lambda\nabla\psi(s), \nabla\psi(s))_{(L^2)^d}$. Since $(\Lambda\nabla\cdot, \nabla\cdot)_{(L^2)^d}$ is a Hilbert norm on $H_0^1(\Omega)$, we conclude that $\psi(s_n) \to \psi(s)$ strongly in $H_0^1(\Omega)$.

■

Part III

Examples of gradient discretisation methods

In this part several classical and popular numerical methods are shown to fit in the gradient discretisation method (GDM).
Chapter 7 introduces the concept of polytopal meshes, on which most schemes considered in the subsequent chapters are built, and presents discrete tools to analyse the properties of various gradient discretisations (GDs). The notions of "control by polytopal toolboxes", of "local linearly exact (LLE) GDs", of "mass lumping", and of "barycentric condensation" provide very easy and short proofs of the consistency, coercivity, limit-conformity and compactness of the considered GDs.
Chapters 8 to 14 are devoted to specific well-known classes of methods, namely:

- Methods based on Galerkin and polynomial interpolations: conforming Galerkin methods, non-conforming finite element methods and derived methods, mixed finite element schemes, discontinuous Galerkin methods;
- Methods derived from the finite volume framework: the multi-point flux approximation (MPFA)-O scheme, hybrid mimetic mixed schemes and nodal mimetic finite difference methods, the cell–certex–face/edge discrete duality finite volume "CeVeFE-DDFV" method .

For each of these methods, a gradient discretisation is constructed such that the corresponding gradient scheme (GS) (2.23) for the standard linear diffusion model (2.20) corresponds to the considered numerical method applied to this model.
The properties (defined in Part I) of each of these GDs thus constructed are then analysed. Once these known numerical methods are recast as GDMs through the choice of appropriate GDs, the analysis developed for various models in Parts I and II directly applies to these methods. A by-product is the convergence of say, the non-conforming $\mathbb{P}_1$, HMM and nMFD schemes, for the Leray–Lions, Stefan and Richards models.
It is worth mentioning that some ideas underlying the methods in Chapters 8 to 14 have been adapted to other models than those covered in Parts I and II, including for instance advection terms: see e.g. [22, 37, 39, 40, 61, 123] and references therein. Although not necessarily amenable to a direct GDM analysis, such adaptations could still benefit from some of the tools developed in this monograph. The GDM itself can be coupled with specific treatments of non diffusive terms; this is the case in [47], which introduces and analyses a scheme coupling the GDM for the diffusion terms and the Eulerian–Lagrangian Localized Adjoint Method (ELLAM) for the convection term.

In all the following chapters in this part, $p \in (1, \infty)$ is an index referring to the regularity of the solution of the considered problem and Ω is an open bounded connected subset of $\mathbb{R}^d$ ($d \in \mathbb{N}^\star$) with Lipschitz-continuous boundary $\partial\Omega$.

7

Analysis tools for gradient discretisations

Polytopal meshes are often used in real applications, for instance in the numerical simulation of complex flows in porous media. The design of numerical methods on such meshes is at the origin of the GDM. Most of the examples of GDs reviewed in this part of the book are based on such meshes. These meshes can be very generic, non-conforming and with possibly non-convex cells.

The convergence analysis of one of the "historical" gradient scheme built upon such polytopal meshes, namely the SUSHI scheme, relies upon a number of discrete functional analysis results which are described in detail in [96, Appendix], see also chapter B in the Appendix below. These results rely on some "polytopal tools", detailed in Section 7.1:

- the polynomial mesh itself (see Definition 7.2 below),
- the set of discrete unknowns on the cells and faces of the mesh,
- piecewise constant reconstruction operators on the cells and faces using these unknowns,
- the (non stabilised) discrete gradient based upon the geometrical property (B.2) and defined by (7.7e) below,
- a discrete $W^{1,p}$ norm defined by (7.7f) below,
- some parameters referring to the regularity of the mesh.

In Section 7.2, these tools are organised in "polytopal toolboxes", consisting of some of the above tools, chosen according to the type of boundary condition. Hence each polytopal toolbox is associated to a given boundary condition. The control of a GD by a polytopal toolbox consists in mapping the discrete unknowns of a GD onto cell- and face-unknowns on a polytopal mesh; if the mapping satisfies three core properties, this control is shown to give the *coercivity*, *compactness* and *limit-conformity* of numerous mesh-based GDs, thanks to the discrete functional analysis results of Appendix B. Precise estimates on the coercivity measure $C_{\mathcal{D}}$ and the limit-conformity $W_{\mathcal{D}}$ are also established.

The notion of local linearly exact (LLE) gradient discretisation is then introduced and analysed in Section 7.3. The linear exactness ensures the *consistency* of sequences of GDs. Section 7.4 extends this analysis to provide explicit estimates on $S_{\mathcal{D}}(\varphi)$, when $\varphi \in W^{2,p}$ with $p > d/2$.
The notion of LLE GD is also used to describe a generic elimination process of unknowns in gradient schemes, replacing them with barycentric combinations of other unknowns. This process, called "barycentric condensation" is standard in the construction of numerous schemes (*e.g.*, SUSHI [96], VAG [99, 101]), but was always performed on a case-by-case basis. In the context of LLE GDs, such barycentric condensations are described without referring to the specificities of each GD, and are shown to preserve the LLE property – and thus the GD-consistency.
A general way to mass-lump any GD is finally presented. Mass-lumping hides various processes which are not always well defined nor justified, and whose purpose is to modify a scheme so as to obtain piecewise constant approximations. In the GDM framework, a rigorous way for performing mass-lumping is set up so that, under a single easily-checked assumption, the mass-lumped GDs enjoy the same properties as the initial GDs.

As a conclusion to this chapter, Section 7.5 presents tracks for further research: a generalised notion of discrete unknowns (Section 7.5.1), and non-linearly exact barycentric combinations (Section 7.5.2) – such combinations may arise in the case of heterogeneous diffusion problems.

Example 7.1 (Illustration of the notions)
Boxes such as the present one one provide illustrative examples of the concepts introduced in this chapter (control by a polytopal toolbox, LLE GDs, etc.). These examples are all based on the non-conforming $\mathbb{P}_1$ finite element method, covered in detail in Chapter 9.

7.1 Polytopal tools

7.1.1 Polytopal meshes

We recall that a 0-polytope is a vertex, a 1-polytope is a segment or an edge, a 2-polytope is a polygon, a 3-polytope is a polyhedron. In order to give a precise definition of a polytope, we first define the k-simplices of $\mathbb{R}^d$ for $k = 0, \ldots, d$. For any family $(\boldsymbol{x}_i)_{i=1,\ldots,k+1}$ of points of $\mathbb{R}^d$ such that the family of vectors $(\boldsymbol{x}_i - \boldsymbol{x}_{k+1})_{i=1,\ldots,k}$ is linearly independent, the k-simplex denoted by $\mathcal{S}((\boldsymbol{x}_i)_{i=1,\ldots,k+1})$ is defined by the convex hull of the points $(\boldsymbol{x}_i)_{i=1,\ldots,k+1}$, that is

$$\mathcal{S}((\boldsymbol{x}_i)_{i=1,\dots,k+1}) = \left\{ \sum_{i=1}^{k+1} \alpha_i \boldsymbol{x}_i \,:\, \alpha_i \geq 0, i = 1, \dots, k+1, \sum_{i=1}^{k+1} \alpha_i = 1 \right\}. \quad (7.1)$$

An open d-polytope Ω (also called poytopal subset of $\mathbb{R}^d$) is defined as the interior of the union of a finite number of d-simplices $(S_j)_{j=1,\dots,M}$, such that the intersection $S_m \cap S_n$ of two different simplices S_m and S_n of the family is either empty or equal to a d'-simplex with $d' < d$. In particular, we have

$$\overline{\Omega} = \bigcup_{j=1,\dots,M} S_j$$

and Ω is the interior of $\overline{\Omega}$. The boundary of Ω is then the union of the faces of the simplices $(S_j)_{j=1,\dots,M}$ which are not common to two different simplices. $\partial\Omega$ is therefore the union of $d-1$-simplices.
In this section, we work with the following conditions and notations:

$$\begin{aligned} &d \in \mathbb{N} \setminus \{0\} \text{ denotes the space dimension,} \\ &\Omega \text{ is a } d\text{-polytopal bounded connected open subset of } \mathbb{R}^d, \\ &\text{with boundary } \partial\Omega. \end{aligned} \quad (7.2)$$

Definition 7.2 (Polytopal mesh). *Let $\Omega \subset \mathbb{R}^d$ satisfy Assumption (7.2); a polytopal mesh of Ω is a quadruplet $\mathfrak{T} = (\mathcal{M}, \mathcal{F}, \mathcal{P}, \mathcal{V})$, where:*

1. *$\mathcal{M}$ is a finite family of non-empty connected polytopal open disjoint subsets of Ω (the "cells") such that $\overline{\Omega} = \cup_{K \in \mathcal{M}} \overline{K}$. For any $K \in \mathcal{M}$, let $\partial K = \overline{K} \setminus K$ be the boundary of K, $|K| > 0$ is the measure of K and h_K denote the diameter of K, that is the maximum distance between two points of $\overline{K}$.*
2. *$\mathcal{F} = \mathcal{F}_{\text{int}} \cup \mathcal{F}_{\text{ext}}$ is a finite family of disjoint subsets of $\overline{\Omega}$ (the "faces" of the mesh – "edges" in 2D), such that any $\sigma \in \mathcal{F}_{\text{int}}$ is contained in Ω and any $\sigma \in \mathcal{F}_{\text{ext}}$ is contained in $\partial\Omega$. Each $\sigma \in \mathcal{F}$ is assumed to be a non-empty open subset of a hyperplane of $\mathbb{R}^d$, with a strictly positive $(d-1)$-dimensional measure $|\sigma|$, and a relative interior $\overline{\sigma} \setminus \sigma$ of zero $(d-1)$-dimensional measure. Furthermore, for all $K \in \mathcal{M}$, there exists a subset $\mathcal{F}_K$ of $\mathcal{F}$ such that $\partial K = \cup_{\sigma \in \mathcal{F}_K} \overline{\sigma}$. We set $\mathcal{M}_\sigma = \{K \in \mathcal{M} \,:\, \sigma \in \mathcal{F}_K\}$ and assume that, for all $\sigma \in \mathcal{F}$, either $\mathcal{M}_\sigma$ has exactly one element and then $\sigma \in \mathcal{F}_{\text{ext}}$, or $\mathcal{M}_\sigma$ has exactly two elements and then $\sigma \in \mathcal{F}_{\text{int}}$. The centre of mass of σ is denoted by $\overline{\boldsymbol{x}}_\sigma$ and, for $K \in \mathcal{M}$ and $\sigma \in \mathcal{F}_K$, $\boldsymbol{n}_{K,\sigma}$ is the (constant) unit vector normal to σ outward to K.*
   For all $K \in \mathcal{M}$, $\mathcal{N}_K$ is the set of the neighbours of K:
   $$\mathcal{N}_K = \{L \in \mathcal{M} \setminus \{K\} \,:\, \exists \sigma \in \mathcal{F}_{\text{int}},\ \mathcal{M}_\sigma = \{K, L\}\}. \quad (7.3)$$
3. *$\mathcal{P}$ is a family of points of Ω indexed by $\mathcal{M}$ and $\mathcal{F}$, denoted by $\mathcal{P} = ((\boldsymbol{x}_K)_{K \in \mathcal{M}}, (\boldsymbol{x}_\sigma)_{\sigma \in \mathcal{F}})$, such that for all $K \in \mathcal{M}$, $\boldsymbol{x}_K \in K$ and for all*

$\sigma \in \mathcal{F}$, $\boldsymbol{x}_\sigma \in \sigma$. We then denote by $d_{K,\sigma}$ the signed orthogonal distance between $\boldsymbol{x}_K$ and $\sigma \in \mathcal{F}_K$ (see Figure 7.1), that is:

$$d_{K,\sigma} = (\boldsymbol{x} - \boldsymbol{x}_K) \cdot \boldsymbol{n}_{K,\sigma}, \text{ for all } \boldsymbol{x} \in \sigma. \tag{7.4}$$

(Note that $(\boldsymbol{x} - \boldsymbol{x}_K) \cdot \boldsymbol{n}_{K,\sigma}$ is constant for $\boldsymbol{x} \in \sigma$.) We then assume that each cell $K \in \mathcal{M}$ is strictly star-shaped with respect to $\boldsymbol{x}_K$, that is $d_{K,\sigma} > 0$ for all $\sigma \in \mathcal{F}_K$. This implies that for all $\boldsymbol{x} \in K$, the line segment $[\boldsymbol{x}_K, \boldsymbol{x}]$ is included in K.
For all $K \in \mathcal{M}$ and $\sigma \in \mathcal{F}_K$, we denote by $D_{K,\sigma}$ the cone with vertex $\boldsymbol{x}_K$ and basis σ, that is

$$D_{K,\sigma} = \{t\boldsymbol{x}_K + (1-t)\boldsymbol{y} \,:\, t \in (0,1),\ \boldsymbol{y} \in \sigma\}. \tag{7.5}$$

We denote, for all $\sigma \in \mathcal{F}$, $D_\sigma = \bigcup_{K \in \mathcal{M}_\sigma} D_{K,\sigma}$ (this set is called the "diamond" associated to the face σ, and for obvious reasons $D_{K,\sigma}$ is also referred to as an "half-diamond").

4. *$\mathcal{V}$ is a set of points (the vertices of the mesh). For $K \in \mathcal{M}$, the set of vertices of K,* i.e. *the vertices contained in $\overline{K}$, is denoted by $\mathcal{V}_K$. Similarly, the set of vertices of $\sigma \in \mathcal{F}$ is $\mathcal{V}_\sigma$.*

The size of the polytopal mesh is defined by:

$$h_{\mathcal{M}} = \sup\{h_K \,:\, K \in \mathcal{M}\}. \tag{7.6}$$

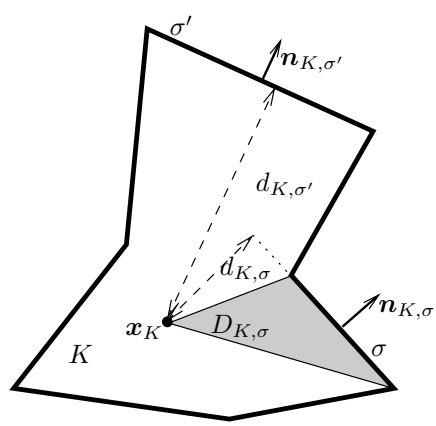


Fig. 7.1. A cell K of a polytopal mesh

Remark 7.3. Definition 7.2 covers a large variety of meshes. In particular, the cells are not assumed to be convex, and the common boundary of two neighbouring cells can include more than one face.

A classical geometry that can be handled through this definition is that of "generalised hexahedron" (see Figure 7.2). This 3D-cell is made of 8 vertices, but the corresponding "physical faces" are not necessarily planar. In that case, by cutting each of these faces in two triangles, we recover a polyhedron with planar faces. This polyhedron has 8 vertices, up to 12 triangular faces (in the case where all physical faces are non-planar and need to be cut in half), but only 6 neighbouring cells.

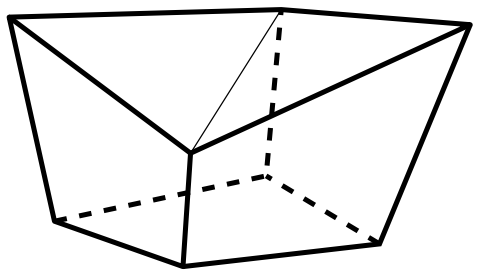


Fig. 7.2. A "generalised hexahedron" with one (non-planar) physical face cut in two (planar and triangular) faces.

A number of finite element methods require the notion of simplicial mesh.

Definition 7.4 (Conforming simplicial mesh). *A conforming simplicial mesh of Ω is a polytopal mesh $\mathfrak{T} = (\mathcal{M}, \mathcal{F}, \mathcal{P}, \mathcal{V})$ in the sense of Definition 7.2, such that for each $K \in \mathcal{M}$ we have* $\mathrm{Card}(\mathcal{F}_K) = d + 1$. *Most often, for these polytopal meshes, $\mathcal{P}$ will be the set of centres of mass of the cells.*

In a conforming simplicial mesh, each cell is therefore a d-simplex (triangle if $d = 2$, tetrahedron if $d = 3$), and there are no hanging nodes, *i.e.* the vertices of the mesh are exactly the "physical" vertices of the cells.

7.1.2 Operators, norm and regularity factors associated with a polytopal mesh

Under Hypothesis (7.2), if $\mathfrak{T} = (\mathcal{M}, \mathcal{F}, \mathcal{P}, \mathcal{V})$ is a polytopal mesh of Ω in the sense of Definition 7.2, we define the space of cell and face unknowns by

$$X_{\mathfrak{T}} = \{v = ((v_K)_{K\in\mathcal{M}}, (v_\sigma)_{\sigma\in\mathcal{F}}) \,:\, v_K \in \mathbb{R}, v_\sigma \in \mathbb{R}\}, \tag{7.7a}$$

and the subspace of vectors with a zero value on the boundary by

$$X_{\mathfrak{T},0} = \{v \in X_{\mathfrak{T}} \,:\, v_\sigma = 0 \text{ for all } \sigma \in \mathcal{F}_{\mathrm{ext}}\}. \tag{7.7b}$$

The function reconstruction $\Pi_{\mathfrak{T}} : X_{\mathfrak{T}} \to L^\infty(\Omega)$, trace reconstruction $\mathbb{T}_{\mathfrak{T}} : X_{\mathfrak{T}} \to L^\infty(\partial\Omega)$ and gradient reconstruction $\overline{\nabla}_{\mathfrak{T}} : X_{\mathfrak{T}} \to L^\infty(\Omega)^d$ are defined by

$$\forall v \in X_{\mathfrak{T}},\ \forall K \in \mathcal{M},\ \text{for a.e. } \boldsymbol{x} \in K,\ \Pi_{\mathfrak{T}} v(\boldsymbol{x}) = v_K, \tag{7.7c}$$

$$\forall v \in X_{\mathfrak{T}},\ \forall \sigma \in \mathcal{F}_{\text{ext}},\ \text{for a.e. } \boldsymbol{x} \in \sigma,\ \mathbb{T}_{\mathfrak{T}} v(\boldsymbol{x}) = v_\sigma, \tag{7.7d}$$

$$\begin{aligned} &\forall v \in X_{\mathfrak{T}},\ \forall K \in \mathcal{M},\ \text{for a.e. } \boldsymbol{x} \in K, \\ &\qquad \overline{\nabla}_{\mathfrak{T}} v(\boldsymbol{x}) = \overline{\nabla}_K v := \frac{1}{|K|} \sum_{\sigma \in \mathcal{F}_K} |\sigma| (v_\sigma - v_K) \boldsymbol{n}_{K,\sigma} \\ &\qquad\qquad\qquad\qquad = \frac{1}{|K|} \sum_{\sigma \in \mathcal{F}_K} |\sigma| v_\sigma \boldsymbol{n}_{K,\sigma}. \end{aligned} \tag{7.7e}$$

The gradient defined by (7.7e) is consistent thanks to the geometrical property (B.2), see lemma B.10; it is however not stable since it only involves the face unknowns $(u_\sigma)_{\sigma \in \mathcal{F}}$ and not the cell unknowns $(u_K)_{K \in \mathfrak{T}}$ (see also Remark 7.8). The last equality in (7.7e) is a consequence of Stokes' formula, which ensures that $\sum_{\sigma \in \mathcal{F}_K} |\sigma| \boldsymbol{n}_{K,\sigma} = 0$ (see the proof of Lemma B.3). Finally, for $p \in [1, +\infty)$ a discrete $W^{1,p}$ semi-norm on $X_{\mathfrak{T}}$ is defined by

$$\forall v \in X_{\mathfrak{T}},\ |v|^p_{\mathfrak{T},p} = \sum_{K \in \mathcal{M}} \sum_{\sigma \in \mathcal{F}_K} |\sigma| d_{K,\sigma} \left| \frac{v_\sigma - v_K}{d_{K,\sigma}} \right|^p. \tag{7.7f}$$

The semi-norm $|\cdot|_{\mathfrak{T},p}$ is in fact a norm when restricted to $X_{\mathfrak{T},0}$.

Remark 7.5 (Cell-centred schemes)
For cell-centred schemes, whose unknowns are $v = (v_K)_{K \in \mathcal{M}}$, a more natural norm than (7.7f) is

$$|v|^p_{\mathfrak{T},p,c} = \sum_{\sigma \in \mathcal{F}} |\sigma| d_{K,L} \left| \frac{v_L - v_K}{d_{K,L}} \right|^p$$

where, in this sum, K and L are the cells around σ and $d_{K,L} = d_{K,\sigma} + d_{L,\sigma}$ if σ is an interior face, or $v_L = 0$ and $d_{K,L} = d_{K,\sigma}$ if $\sigma \in \mathcal{F}_K \cap \mathcal{F}_{\text{ext}}$. Switching from cell unknowns to cell and face unknowns as in (7.7a) and (7.7f) is quite easy. It suffices to extend $u = (u_K)_{K \in \mathcal{M}}$ into $\widetilde{v} = ((v_K)_{K \in \mathcal{M}}, (v_\sigma)_{\sigma \in \mathcal{F}})$ with $v_\sigma = \frac{v_K + v_L}{2}$ if σ is an interior face and K, L are the cells around σ, or $v_\sigma = 0$ if σ is a boundary face. Then, the norms $v \mapsto |\widetilde{v}|_{\mathfrak{T},p}$ and $v \mapsto |v|_{\mathfrak{T},p,c}$ are equivalent, with constants involving $\eta_{\mathfrak{T}}$ given in (7.9) below, and all the results presented in this section can therefore be applied provided that $\eta_{\mathfrak{T}}$ is bounded independently of the mesh size. Note that the converse (adding cell unknowns to a method with face unknowns only, in order to use the results of this section) is also easy – see the analysis of non-conforming finite elements in Chapter 9.

Finally, two numbers are introduced to measure the regularity properties of a polytopal mesh, namely:

$$\theta_{\mathfrak{T}} = \max_{K \in \mathcal{M}} \left(\max_{\sigma \in \mathcal{F}_K} \frac{h_K}{d_{K,\sigma}} + \text{Card}(\mathcal{F}_K) \right), \tag{7.8}$$

$$\eta_{\mathfrak{T}} = \max_{\sigma\in\mathcal{F}_{\rm int}\,,\ \mathcal{M}_\sigma=\{K,L\}} \left(\frac{d_{K,\sigma}}{d_{L,\sigma}} + \frac{d_{L,\sigma}}{d_{K,\sigma}}\right). \tag{7.9}$$

A number of results involving sequences $(\mathfrak{T}_m)_{m\in\mathbb{N}}$ of polytopal meshes will require one or the other, or both, of these corresponding regularity factors to be bounded along the sequence of meshes.
For simplicial meshes, only the following simpler regularity factor is needed:

$$\kappa_{\mathfrak{T}} = \max_{K\in\mathcal{M}} \frac{h_K}{\rho_K}, \tag{7.10}$$

where, for $K \in \mathcal{M}$, ρ_K is the radius of the largest ball included in K and centred at the centre of mass $\overline{\boldsymbol{x}}_K$ of K. It is proved in Lemma B.4 page 409 that, for simplicial meshes, $\kappa_{\mathfrak{T}}$ controls $\theta_{\mathfrak{T}}$ and $\eta_{\mathfrak{T}}$.

7.2 Control of a GD by a polytopal toolbox

Example 7.6 (GD for the non-conforming $\mathbb{P}_1$ finite elements)
The non-conforming $\mathbb{P}_1$ finite element is used to illustrate notions introduced below; the corresponding gradient discretisation is briefly recalled: a simplicial mesh $\mathfrak{T}$ (see Definition 7.4) is used; the set of discrete unknowns, located on the faces of the mesh, is denoted by

$$X_{\mathcal{D},0} = \{v = (v_\sigma)_{\sigma\in\mathcal{F}_K} \,:\, v_\sigma \in \mathbb{R} \text{ for all } \sigma \in \mathcal{F}_{\rm int}\,, \quad v_\sigma = 0 \text{ for all } \sigma \in \mathcal{F}_{\rm ext}\}. \tag{7.11}$$

The function reconstruction $\Pi_{\mathcal{D}} : X_{\mathcal{D},0} \to L^p(\Omega)$ is defined by: for $v \in X_{\mathcal{D},0}$, $\Pi_{\mathcal{D}} v$ is the function on Ω that is linear on each $K \in \mathcal{M}$, continuous at the face centres $(\overline{\boldsymbol{x}}_\sigma)_{\sigma\in\mathcal{F}}$, and takes the values $(v_\sigma)_{\sigma\in\mathcal{F}}$ at these centres. The gradient reconstruction $\nabla_{\mathcal{D}} : X_{\mathcal{D},0} \to L^p(\Omega)^d$ is the "broken" gradient: $\nabla_{\mathcal{D}} v$ is constant equal to $\nabla[(\Pi_{\mathcal{D}} v)_{|K}]$ in each $K \in \mathcal{M}$.

7.2.1 Dirichlet boundary conditions

Definition 7.7 (Polytopal toolbox for homogeneous Dirichlet BCs). *Let Ω satisfy Assumption (7.2), and let $\mathfrak{T}$ be a polytopal mesh in the sense of Definition 7.2. The quadruplet $(X_{\mathfrak{T},0}, \Pi_{\mathfrak{T}}, \overline{\nabla}_{\mathfrak{T}}, |\cdot|_{\mathfrak{T},p})$ is a polytopal toolbox for Dirichlet boundary conditions if:*

1. *The set $X_{\mathfrak{T},0}$ is defined by (7.7b):*

$$X_{\mathfrak{T},0} = \{v \in X_{\mathfrak{T}} \,:\, v_\sigma = 0 \text{ for all } \sigma \in \mathcal{F}_{\rm ext}\}.$$

2. *The function reconstruction $\Pi_{\mathfrak{T}} : X_{\mathfrak{T},0} \to L^\infty(\Omega)$ is defined by (7.7c):*

$$\forall v \in X_{\mathfrak{T},0},\ \forall K \in \mathcal{M},\ \text{for a.e. } \boldsymbol{x} \in K,\ \Pi_{\mathfrak{T}} v(\boldsymbol{x}) = v_K.$$

3. The gradient reconstruction $\overline{\nabla}_{\mathfrak{T}} : X_{\mathfrak{T},0} \to L^\infty(\Omega)^d$ *is defined by* (7.7e)*:*

$$\forall v \in X_{\mathfrak{T},0},\ \forall K \in \mathcal{M},\ \textit{for a.e. } \boldsymbol{x} \in K,$$
$$\overline{\nabla}_{\mathfrak{T}} v(\boldsymbol{x}) = \frac{1}{|K|} \sum_{\sigma \in \mathcal{F}_K} |\sigma|(v_\sigma - v_K)\boldsymbol{n}_{K,\sigma} = \frac{1}{|K|} \sum_{\sigma \in \mathcal{F}_K} |\sigma| v_\sigma \boldsymbol{n}_{K,\sigma}.$$

4. The space $X_{\mathfrak{T},0}$ *is endowed with the norm* (7.7f)*:*

$$\forall v \in X_{\mathfrak{T},0},\ |v|^p_{\mathfrak{T},p} = \sum_{K \in \mathcal{M}} \sum_{\sigma \in \mathcal{F}_K} |\sigma| d_{K,\sigma} \left| \frac{v_\sigma - v_K}{d_{K,\sigma}} \right|^p.$$

Remark 7.8. Note that $(X_{\mathfrak{T},0}, \Pi_{\mathfrak{T}}, \overline{\nabla}_{\mathfrak{T}})$ is not a GD since $\left\|\overline{\nabla}_{\mathfrak{T}} \cdot\right\|_{L^p(\Omega)^d}$ is not a norm on $X_{\mathfrak{T},0}$: if $v \in X_{\mathfrak{T},0}$ has zero values at all the faces but not in the cells, $\overline{\nabla}_{\mathfrak{T}} v = 0$ but $v \neq 0$ in $X_{\mathfrak{T},0}$. The original SUSHI scheme [96] combines the gradient $\overline{\nabla}_{\mathfrak{T}} v$ with a stabilisation on half diamond cells (see Chapter 13).

Often, $\mathfrak{T}$ refers to both the polytopal mesh and to the polytopal toolbox $(X_{\mathfrak{T},0}, \Pi_{\mathfrak{T}}, \overline{\nabla}_{\mathfrak{T}}, |\ |_{\mathfrak{T},p})$. There is an abuse of notation here, since the polytopal mesh does not depend on the considered boundary conditions (Dirichlet, Neumann, etc.), whereas the polytopal toolbox does depend on these conditions as seen in Section 7.2.2. However, the context will always make clear which boundary conditions are considered, and thus which kind of polytopal toolbox should be used.

Definition 7.9 (Control of a GD, hom. Dirichlet BCs). *Let* Ω *satisfy Assumption* (7.2)*, let* $\mathcal{D}$ *be a GD in the sense of Definition 2.1, and let* $\mathfrak{T}$ *be a polytopal toolbox in the sense of Definition 7.7. A control of* $\mathcal{D}$ *by* $\mathfrak{T}$ *is a linear mapping* $\boldsymbol{\Phi} : X_{\mathcal{D},0} \to X_{\mathfrak{T},0}$*. We then define*

$$\|\boldsymbol{\Phi}\|_{\mathcal{D},\mathfrak{T}} = \max_{v \in X_{\mathcal{D},0} \setminus \{0\}} \frac{|\boldsymbol{\Phi}(v)|_{\mathfrak{T},p}}{\|v\|_{\mathcal{D}}}, \tag{7.12}$$
$$\omega^{\Pi}(\mathcal{D}, \mathfrak{T}, \boldsymbol{\Phi}) = \max_{v \in X_{\mathcal{D},0} \setminus \{0\}} \frac{\|\Pi_{\mathcal{D}} v - \Pi_{\mathfrak{T}} \boldsymbol{\Phi}(v)\|_{L^p(\Omega)}}{\|v\|_{\mathcal{D}}},$$
$$\omega^{\nabla}(\mathcal{D}, \mathfrak{T}, \boldsymbol{\Phi}) = \max_{v \in X_{\mathcal{D},0} \setminus \{0\}} \frac{1}{\|v\|_{\mathcal{D}}} \left(\sum_{K \in \mathcal{M}} |K|^{1-p} \left| \int_K \left[\nabla_{\mathcal{D}} v(\boldsymbol{x}) - \overline{\nabla}_{\mathfrak{T}} \boldsymbol{\Phi}(v)(\boldsymbol{x}) \right] \mathrm{d}\boldsymbol{x} \right|^p \right)^{\frac{1}{p}}.$$

Example 7.10 (Control of the non-conforming $\mathbb{P}_1$ GD)
Finding a control of a given gradient discretisation $\mathcal{D}$ by a polytopal toolbox $\mathfrak{T}$ consists in computing – often in the most natural way – face and

cell values (which define an element of $X_{\mathfrak{T},0}$) from the discrete unknowns of $\mathcal{D}$.
Let us consider the case of the non-conforming $\mathbb{P}_1$ gradient discretisation. Recalling the definition (7.11) of $X_{\mathcal{D},0}$, there is not need to actually *compute* face unknowns since they are already in $X_{\mathcal{D},0}$. Cell unknowns are computed by creating equally weighted averages of the $d+1$ face unknowns in each cell.
This leads us to defining the following control $\boldsymbol{\Phi} : X_{\mathcal{D},0} \to X_{\mathfrak{T},0}$: for $v = (v_\sigma)_{\sigma\in\mathcal{F}_K} \in X_{\mathcal{D},0}$, the element $\Phi(v) = \widehat{v} = ((\widehat{v}_K)_{K\in\mathcal{M}}, (\widehat{v}_\sigma)_{\sigma\in\mathcal{F}})$ of $X_{\mathfrak{T},0}$ is given by

$$\forall \sigma \in \mathcal{F}\,,\ \widehat{v}_\sigma = v_\sigma \quad \text{and} \quad \forall K \in \mathcal{M}\,,\ \widehat{v}_K = \frac{1}{d+1}\sum_{\sigma\in\mathcal{F}_K} v_\sigma.$$

We prove in Lemma 9.3 that, for this control, $\|\boldsymbol{\Phi}\|_{\mathcal{D},\mathfrak{T}} \le \kappa_{\mathfrak{T}} d^{1/p}$, $\omega^{\Pi}(\mathcal{D},\mathfrak{T},\boldsymbol{\Phi}) \le h_{\mathcal{M}}$ and $\omega^{\nabla}(\mathcal{D},\mathfrak{T},\boldsymbol{\Phi}) = 0$. Example 7.13 shows how such bounds are used.

Theorem 7.11 (Estimates for a controlled GD, hom. Dirichlet BCs). *Let Ω satisfy Assumption (7.2), let $\mathcal{D}$ be a GD in the sense of Definition 2.1, let $\mathfrak{T}$ be a polytopal toolbox in the sense of Definition 7.7, and let $\boldsymbol{\Phi}$ be a control of $\mathcal{D}$ by $\mathfrak{T}$ in the sense of Definition 7.9. We take $\varrho \ge \theta_{\mathfrak{T}} + \eta_{\mathfrak{T}}$ (see (7.8) and (7.9)).*
Then, there exists C_1 depending only on Ω, p and ϱ such that

$$C_{\mathcal{D}} \le \omega^{\Pi}(\mathcal{D},\mathfrak{T},\boldsymbol{\Phi}) + C_1 \|\boldsymbol{\Phi}\|_{\mathcal{D},\mathfrak{T}} \tag{7.13}$$

and, for all $\boldsymbol{\varphi} \in W^{1,p'}(\Omega)^d$,

$$W_{\mathcal{D}}(\boldsymbol{\varphi}) \le \|\boldsymbol{\varphi}\|_{W^{1,p'}(\Omega)^d} \Big[C_1 h_{\mathcal{M}}(1 + \|\boldsymbol{\Phi}\|_{\mathcal{D},\mathfrak{T}}) + \omega^{\Pi}(\mathcal{D},\mathfrak{T},\boldsymbol{\Phi}) + \omega^{\nabla}(\mathcal{D},\mathfrak{T},\boldsymbol{\Phi}) \Big]. \tag{7.14}$$

Here, $C_{\mathcal{D}}$ and $W_{\mathcal{D}}$ are the coercivity constant and limit-conformity measure defined by (2.1) and (2.6).

Proof. Using the triangle inequality, Lemma B.15 and the Hölder's inequality (D.7) we observe that, for any $v \in X_{\mathcal{D},0}$,

$$\begin{aligned}\|\Pi_{\mathcal{D}} v\|_{L^p(\Omega)} &\le \omega^{\Pi}(\mathcal{D},\mathfrak{T},\boldsymbol{\Phi})\, \|v\|_{\mathcal{D}} + \|\Pi_{\mathfrak{T}}\boldsymbol{\Phi}(v)\|_{L^p(\Omega)} \\ &\le \omega^{\Pi}(\mathcal{D},\mathfrak{T},\boldsymbol{\Phi})\, \|v\|_{\mathcal{D}} + C_8 |\Omega|^{\frac{1}{p}-\frac{1}{q}}\, |\boldsymbol{\Phi}(v)|_{\mathfrak{T},p}\,,\end{aligned}$$

with q and C_8 given in Lemma B.15. The proof of Estimate (7.13) is concluded by dividing by $\|v\|_{\mathcal{D}}$ and using the definition (7.12) of $\|\boldsymbol{\Phi}\|_{\mathcal{D},\mathfrak{T}}$.

We turn to (7.14). Let $\boldsymbol{\varphi} \in W^{1,p'}(\Omega)^d$ and use the triangle inequality, the definition of $\omega^{\Pi}(\mathcal{D}, \mathfrak{T}, \boldsymbol{\Phi})$ and (B.31) (notice that $\mathbb{T}_{\mathcal{D}}\boldsymbol{\Phi}(v) = 0$ since $\boldsymbol{\Phi}(v) \in X_{\mathfrak{T},0}$) to obtain

$$\begin{aligned}
&\left| \int_\Omega \Big(\nabla_{\mathcal{D}} v(\boldsymbol{x}) \cdot \boldsymbol{\varphi}(\boldsymbol{x}) + \Pi_{\mathcal{D}} v(\boldsymbol{x}) \mathrm{div} \boldsymbol{\varphi}(\boldsymbol{x}) \Big) \mathrm{d}\boldsymbol{x} \right| \\
&\quad \le \left| \int_\Omega [\nabla_{\mathcal{D}} v(\boldsymbol{x}) - \overline{\nabla}_{\mathfrak{T}} \boldsymbol{\Phi}(v)(\boldsymbol{x})] \cdot \boldsymbol{\varphi}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \right| + \|\mathrm{div}\boldsymbol{\varphi}\|_{L^{p'}(\Omega)} \, \omega^{\Pi}(\mathcal{D}, \mathfrak{T}, \boldsymbol{\Phi}) \, \|v\|_{\mathcal{D}} \\
&\qquad + \left| \int_\Omega \left(\overline{\nabla}_{\mathfrak{T}} \boldsymbol{\Phi}(v)(\boldsymbol{x}) \cdot \boldsymbol{\varphi}(\boldsymbol{x}) + \Pi_{\mathfrak{T}} \boldsymbol{\Phi}(v)(\boldsymbol{x}) \mathrm{div} \boldsymbol{\varphi}(\boldsymbol{x}) \right) \mathrm{d}\boldsymbol{x} \right| \\
&\quad \le \left| \int_\Omega [\nabla_{\mathcal{D}} v(\boldsymbol{x}) - \overline{\nabla}_{\mathfrak{T}} \boldsymbol{\Phi}(v)(\boldsymbol{x})] \cdot \boldsymbol{\varphi}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \right| + \|\mathrm{div}\boldsymbol{\varphi}\|_{L^{p'}(\Omega)} \, \omega^{\Pi}(\mathcal{D}, \mathfrak{T}, \boldsymbol{\Phi}) \, \|v\|_{\mathcal{D}} \\
&\qquad + C_5 \, \| \, |\nabla \boldsymbol{\varphi}| \, \|_{L^{p'}(\Omega)^d} \, |\boldsymbol{\Phi}(v)|_{\mathfrak{T},p} \, h_{\mathcal{M}}. \qquad (7.15)
\end{aligned}$$

Let $\boldsymbol{\varphi}_K = \frac{1}{|K|} \int_K \boldsymbol{\varphi}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}$. Assuming that $p > 1$ (so that $p' < \infty$) and applying (B.12) in Lemma B.6 to p' instead of p, we find C_2 depending only on d, p and ϱ such that $\|\boldsymbol{\varphi} - \boldsymbol{\varphi}_K\|_{L^{p'}(K)} \le C_2 h_K \, \| \, |\nabla \boldsymbol{\varphi}| \, \|_{L^{p'}(K)}$. Hence, using Hölder's inequality,

$$\begin{aligned}
&\left| \int_\Omega [\nabla_{\mathcal{D}} v(\boldsymbol{x}) - \overline{\nabla}_{\mathfrak{T}} \boldsymbol{\Phi}(v)(\boldsymbol{x})] \cdot \boldsymbol{\varphi}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \right| \\
&\quad = \left| \sum_{K \in \mathcal{M}} \int_K [\nabla_{\mathcal{D}} v(\boldsymbol{x}) - \overline{\nabla}_{\mathfrak{T}} \boldsymbol{\Phi}(v)(\boldsymbol{x})] \cdot \boldsymbol{\varphi}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \right| \\
&\quad = \left| \sum_{K \in \mathcal{M}} \Big(\int_K \nabla_{\mathcal{D}} v(\boldsymbol{x}) \cdot [\boldsymbol{\varphi}(\boldsymbol{x}) - \boldsymbol{\varphi}_K] \mathrm{d}\boldsymbol{x} \right. \\
&\qquad\qquad \left. + \boldsymbol{\varphi}_K \cdot \int_K [\nabla_{\mathcal{D}} v(\boldsymbol{x}) - \overline{\nabla}_{\mathfrak{T}} \boldsymbol{\Phi}(v)(\boldsymbol{x})] \mathrm{d}\boldsymbol{x} \Big) \right| \\
&\quad \le C_2 h_{\mathcal{M}} \, \| \, |\nabla \boldsymbol{\varphi}| \, \|_{L^{p'}(\Omega)} \, \|\nabla_{\mathcal{D}} v\|_{L^p(\Omega)^d} \\
&\qquad\qquad + \sum_{K \in \mathcal{M}} |\boldsymbol{\varphi}_K| \left| \int_K [\nabla_{\mathcal{D}} v(\boldsymbol{x}) - \overline{\nabla}_{\mathfrak{T}} \boldsymbol{\Phi}(v)(\boldsymbol{x})] \mathrm{d}\boldsymbol{x} \right|
\end{aligned}$$

By Hölder's inequality $|\boldsymbol{\varphi}_K| \le |K|^{-1} |K|^{1-\frac{1}{p'}} \, \|\boldsymbol{\varphi}\|_{L^{p'}(K)^d} = |K|^{\frac{1}{p}-1} \, \|\boldsymbol{\varphi}\|_{L^{p'}(K)^d}$ and thus

$$\begin{aligned}
&\left| \int_\Omega [\nabla_{\mathcal{D}} v(\boldsymbol{x}) - \overline{\nabla}_{\mathfrak{T}} \boldsymbol{\Phi}(v)(\boldsymbol{x})] \cdot \boldsymbol{\varphi}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \right| \\
&\quad \le C_2 h_{\mathcal{M}} \, \| \, |\nabla \boldsymbol{\varphi}| \, \|_{L^{p'}(\Omega)} \, \|\nabla_{\mathcal{D}} v\|_{L^p(\Omega)^d} \\
&\qquad + \|\boldsymbol{\varphi}\|_{L^{p'}(\Omega)^d} \left(\sum_{K \in \mathcal{M}} |K|^{1-p} \left| \int_K \left[\nabla_{\mathcal{D}} v(\boldsymbol{x}) - \overline{\nabla}_{\mathfrak{T}} \boldsymbol{\Phi}(v)(\boldsymbol{x}) \right] \mathrm{d}\boldsymbol{x} \right|^p \right)^{1/p}
\end{aligned}$$

$$\leq \left(C_2 h_{\mathcal{M}} + \omega^{\nabla}(\mathcal{D}, \mathfrak{T}, \boldsymbol{\Phi})\right) \|\boldsymbol{\varphi}\|_{W^{1,p'}(\Omega)^d} \|\nabla_{\mathcal{D}} v\|_{L^p(\Omega)^d}. \tag{7.16}$$

Plugged into (7.15) and using the definition (7.12) of $\|\boldsymbol{\Phi}\|_{\mathcal{D},\mathfrak{T}}$ this gives (7.14). In the case $p = 1$ (and thus $p' = +\infty$), extend $\boldsymbol{\varphi}$ into a Lipschitz-continuous function over $\mathbb{R}^d$, with a Lipschitz constant bounded by $C_3 \| |\nabla \boldsymbol{\varphi}| \|_{L^\infty(\Omega)}$ for some C_3 depending only on d. The previous calculations can then be done by using the estimate $|\boldsymbol{\varphi}(\boldsymbol{x}) - \boldsymbol{\varphi}_K| \leq C_3 h_K \| |\nabla \boldsymbol{\varphi}| \|_{L^\infty(\Omega)}$ for any $\boldsymbol{x} \in K$.

■

An immediate consequence of Theorem 7.11 is the following corollary.

Corollary 7.12 (Properties of controlled GDs, hom. Dirichlet BCs). *Let Ω satisfy Assumption (7.2), let $(\mathcal{D}_m)_{m\in\mathbb{N}}$ be a sequence of GDs in the sense of Definition 2.1, and let $(\mathfrak{T}_m)_{m\in\mathbb{N}}$ be a sequence of polytopal toolboxes in the sense of Definition 7.7. We assume that $h_{\mathcal{M}_m} \to 0$ as $m \to \infty$ and that $\sup_{m\in\mathbb{N}}(\theta_{\mathfrak{T}_m} + \eta_{\mathfrak{T}_m}) < +\infty$ (see (7.8) and (7.9)).*
For all $m \in \mathbb{N}$ we take a control $\boldsymbol{\Phi}_m$ of $\mathcal{D}_m$ by $\mathfrak{T}_m$ in the sense of Definition 7.9, and we assume that

$$\sup_{m\in\mathbb{N}} \|\boldsymbol{\Phi}_m\|_{\mathcal{D}_m,\mathfrak{T}_m} < +\infty,$$
$$\lim_{m\to\infty} \omega^{\Pi}(\mathcal{D}_m, \mathfrak{T}_m, \boldsymbol{\Phi}_m) = 0, \text{ and}$$
$$\lim_{m\to\infty} \omega^{\nabla}(\mathcal{D}_m, \mathfrak{T}_m, \boldsymbol{\Phi}_m) = 0.$$

Then $(\mathcal{D}_m)_{m\in\mathbb{N}}$ is coercive in the sense of Definition 2.2, limit-conforming in the sense of Definition 2.5, and compact in the sense of Definition 2.8.

Example 7.13 (Properties of the non-conforming $\mathbb{P}_1$ GD)
Using the control $\boldsymbol{\Phi}$ and the estimates on $\|\boldsymbol{\Phi}\|_{\mathcal{D},\mathfrak{T}}$, $\omega^{\Pi}(\mathcal{D}, \mathfrak{T}, \boldsymbol{\Phi})$ and $\omega^{\nabla}(\mathcal{D}, \mathfrak{T}, \boldsymbol{\Phi})$, from Example 7.10, the above corollary establishes the coercivity, limit-conformity and compactness of the gradient discretisations built on non-conforming $\mathbb{P}_1$ finite elements.

Proof. The coercivity and limit-conformity are trivial since (7.13) and (7.14) ensure that $\sup_{m\in\mathbb{N}} C_{\mathcal{D}_m} < +\infty$ and that $W_{\mathcal{D}_m}(\boldsymbol{\varphi}) \to 0$ as $m \to \infty$, for all $\boldsymbol{\varphi} \in W^{1,p'}(\Omega)^d$ (we use Lemma 2.17 and the fact that $W^{1,p'}(\Omega)^d$ is dense in $W^{p'}_{\mathrm{div}}(\Omega)$ – see Remark 2.18).
It remains to prove the compactness. If $u_m \in X_{\mathcal{D}_m,0}$ is such that $(\|u_m\|_{\mathcal{D}_m})_{m\in\mathbb{N}}$ is bounded, then the bound on $\|\boldsymbol{\Phi}_m\|_{\mathcal{D}_m,\mathfrak{T}_m}$ ensures that $(|\boldsymbol{\Phi}_m(u_m)|_{\mathfrak{T}_m,p})_{m\in\mathbb{N}}$ is bounded. By Lemma B.19, we infer that up to a subsequence $\Pi_{\mathfrak{T}_m}\boldsymbol{\Phi}_m(u_m)$ converges to some u in $L^p(\Omega)$ as $m \to \infty$. Since

$$\|\Pi_{\mathcal{D}_m} u_m - \Pi_{\mathfrak{T}_m}\boldsymbol{\Phi}_m(u_m)\|_{L^p(\Omega)} \leq \omega^{\Pi}(\mathcal{D}_m, \mathfrak{T}_m, \boldsymbol{\Phi}_m) \|u_m\|_{\mathcal{D}_m} \to 0$$

as $m \to \infty$, we deduce that $\Pi_{\mathcal{D}_m} u_m \to u$ in $L^p(\Omega)$ and the proof is complete.

■

Remark 7.14 (Non-homogeneous Dirichlet boundary conditions). The definitions of coercivity, limit-conformity and compactness of GDs for non-homogeneous Dirichlet boundary conditions are identical to the definitions for homogeneous Dirichlet conditions. Hence, all results in this section (in particular, Corollary 7.12) can be used in the context of non-homogeneous Dirichlet boundary conditions.

7.2.2 Neumann and Fourier boundary conditions

We define here the notions of polytopal toolboxes and control by polytopal toolboxes for non-homogeneous Neumann boundary conditions, in a similar way as what we did in Section 7.2.1 for Dirichlet boundary conditions. In Remarks 7.19 and 7.20 we indicate the minor modifications that need to be made to the following definitions and results for homogeneous Neumann and Fourier boundary conditions.

Definition 7.15 (Polytopal toolbox for Neumann BCs). *Let Ω satisfy Assumption (7.2), and let $\mathfrak{T}$ be a polytopal mesh in the sense of Definition 7.2. The family $(X_{\mathfrak{T}}, \Pi_{\mathfrak{T}}, \mathbb{T}_{\mathfrak{T}}, \overline{\nabla}_{\mathfrak{T}}, \| \ \|_{\mathfrak{T},p})$ is a polytopal toolbox for Neumann boundary conditions if:*

1. *The set $X_{\mathfrak{T}}$ is defined by (7.7a):*

$$X_{\mathfrak{T}} = \{v = ((v_K)_{K\in\mathcal{M}}, (v_\sigma)_{\sigma\in\mathcal{F}}) \,:\, v_K \in \mathbb{R}, v_\sigma \in \mathbb{R}\}.$$

2. *The function reconstruction $\Pi_{\mathfrak{T}} : X_{\mathfrak{T}} \to L^\infty(\Omega)$ is defined by (7.7c):*

$$\forall v \in X_{\mathfrak{T}},\ \forall K \in \mathcal{M},\ \textit{for a.e. } \boldsymbol{x} \in K,\ \Pi_{\mathfrak{T}} v(\boldsymbol{x}) = v_K.$$

3. *The trace reconstruction $\mathbb{T}_{\mathfrak{T}} : X_{\mathfrak{T}} \to L^\infty(\partial\Omega)$ is defined by (7.7d):*

$$\forall v \in X_{\mathfrak{T}},\ \forall \sigma \in \mathcal{F}_{\text{ext}},\ \textit{for a.e. } \boldsymbol{x} \in \sigma,\ \mathbb{T}_{\mathfrak{T}} v(\boldsymbol{x}) = v_\sigma.$$

4. *The gradient reconstruction $\overline{\nabla}_{\mathfrak{T}} : X_{\mathfrak{T}} \to L^\infty(\Omega)^d$ is defined by (7.7e):*

$$\begin{aligned}&\forall v \in X_{\mathfrak{T}},\ \forall K \in \mathcal{M},\ \textit{for a.e. } \boldsymbol{x} \in K,\\ &\quad \overline{\nabla}_{\mathfrak{T}} v(\boldsymbol{x}) = \frac{1}{|K|}\sum_{\sigma\in\mathcal{F}_K} |\sigma|(v_\sigma - v_K)\boldsymbol{n}_{K,\sigma} = \frac{1}{|K|}\sum_{\sigma\in\mathcal{F}_K} |\sigma| v_\sigma \boldsymbol{n}_{K,\sigma}.\end{aligned}$$

5. *Recalling the definition (7.7f) of the semi-norm $|\ |_{\mathfrak{T},p}$, the space $X_{\mathfrak{T}}$ is endowed with the norm*

$$\|v\|_{\mathfrak{T},p}^p = |v|_{\mathfrak{T},p}^p + \left|\int_\Omega \Pi_{\mathfrak{T}} v(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\right|^p. \tag{7.17}$$

As mentioned in Section 7.2.1 on Dirichlet boundary conditions, we will often use $\mathfrak{T}$ to denote both the polytopal mesh and the polytopal toolbox $(X_{\mathfrak{T}}, \Pi_{\mathfrak{T}}, \mathbb{T}_{\mathfrak{T}}, \overline{\nabla}_{\mathfrak{T}}, \| \ \|_{\mathfrak{T},p})$.

Definition 7.16 (Control of a GD by a polytopal toolbox – Neumann BCs). *Let Ω satisfy Assumption (7.2), let $\mathcal{D}$ be a GD in the sense of Definition 3.11, and let $\mathfrak{T}$ be a polytopal toolbox in the sense of Definition 7.15. A control of $\mathcal{D}$ by $\mathfrak{T}$ is a linear mapping $\boldsymbol{\Phi} : X_{\mathcal{D}} \to X_{\mathfrak{T}}$. We then define*

$$\|\boldsymbol{\Phi}\|_{\mathcal{D},\mathfrak{T}} = \max_{v\in X_{\mathcal{D}}\setminus\{0\}} \frac{\|\boldsymbol{\Phi}(v)\|_{\mathfrak{T},p}}{\|v\|_{\mathcal{D}}},$$

$$\omega^{\Pi}(\mathcal{D},\mathfrak{T},\boldsymbol{\Phi}) = \max_{v\in X_{\mathcal{D}}\setminus\{0\}} \frac{\|\Pi_{\mathcal{D}} v - \Pi_{\mathfrak{T}}\boldsymbol{\Phi}(v)\|_{L^p(\Omega)}}{\|v\|_{\mathcal{D}}},$$

$$\omega^{\mathbb{T}}(\mathcal{D},\mathfrak{T},\boldsymbol{\Phi}) = \max_{v\in X_{\mathcal{D}}\setminus\{0\}} \frac{\|\mathbb{T}_{\mathcal{D}} v - \mathbb{T}_{\mathfrak{T}}\boldsymbol{\Phi}(v)\|_{L^p(\partial\Omega)}}{\|v\|_{\mathcal{D}}},$$

$$\omega^{\nabla}(\mathcal{D},\mathfrak{T},\boldsymbol{\Phi}) = \max_{v\in X_{\mathcal{D}}\setminus\{0\}} \frac{\left(\sum_{K\in\mathcal{M}} |K|^{1-p} \left|\int_K \left[\nabla_{\mathcal{D}} v(\boldsymbol{x}) - \overline{\nabla}_{\mathfrak{T}}\boldsymbol{\Phi}(v)(\boldsymbol{x})\right] \mathrm{d}\boldsymbol{x}\right|^p\right)^{\frac{1}{p}}}{\|v\|_{\mathcal{D}}}.$$

Theorem 7.17 (Estimates for a GD controlled by polytopal toolboxes – Neumann BCs). *Let Ω satisfy Assumption (7.2), let $\mathcal{D}$ be a GD in the sense of Definition 3.11, and let $\mathfrak{T}$ be a polytopal toolbox in the sense of Definition 7.15. We take $\boldsymbol{\Phi}$ a control of $\mathcal{D}$ by $\mathfrak{T}$ in the sense of Definition 7.16, and $\varrho \geq \theta_{\mathfrak{T}} + \eta_{\mathfrak{T}}$ (see (7.8) and (7.9)).*
Then, there exists C_4 depending only on Ω, p and ϱ such that

$$C_{\mathcal{D}} \leq \max\left(\omega^{\Pi}(\mathcal{D},\mathfrak{T},\boldsymbol{\Phi}), \omega^{\mathbb{T}}(\mathcal{D},\mathfrak{T},\boldsymbol{\Phi})\right) + C_4 \|\boldsymbol{\Phi}\|_{\mathcal{D},\mathfrak{T}} \tag{7.18}$$

and, for all $\boldsymbol{\varphi} \in W^{1,p'}(\Omega)^d$,

$$W_{\mathcal{D}}(\boldsymbol{\varphi}) \leq \|\boldsymbol{\varphi}\|_{W^{1,p'}(\Omega)^d} \Big[C_4 h_{\mathcal{M}}(1 + \|\boldsymbol{\Phi}\|_{\mathcal{D},\mathfrak{T}}) + \omega^{\Pi}(\mathcal{D},\mathfrak{T},\boldsymbol{\Phi}) + \omega^{\nabla}(\mathcal{D},\mathfrak{T},\boldsymbol{\Phi}) + \omega^{\mathbb{T}}(\mathcal{D},\mathfrak{T},\boldsymbol{\Phi})\Big]. \tag{7.19}$$

Here, $C_{\mathcal{D}}$ and $W_{\mathcal{D}}$ are the coercivity constant and limit-conformity measure defined by (3.9) and (3.11).

Proof. The proof is similar to the case of Dirichlet boundary conditions (Theorem 7.11). By Lemmas B.21 and B.25 we have $\|\mathbb{T}_{\mathfrak{T}}\boldsymbol{\Phi}(v)\|_{L^p(\partial\Omega)} \leq C_5 \|\boldsymbol{\Phi}(v)\|_{\mathfrak{T},p}$ for some C_5 depending only on Ω, p and ϱ. Hence, using the triangle inequality,

$$\begin{aligned}\|\mathbb{T}_{\mathcal{D}} v\|_{L^p(\partial\Omega)} &\leq \omega^{\mathbb{T}}(\mathcal{D},\mathfrak{T},\boldsymbol{\Phi}) \|v\|_{\mathcal{D}} + \|\mathbb{T}_{\mathfrak{T}}\boldsymbol{\Phi}(v)\|_{L^p(\partial\Omega)} \\ &\leq \omega^{\mathbb{T}}(\mathcal{D},\mathfrak{T},\boldsymbol{\Phi}) \|v\|_{\mathcal{D}} + C_5 \|\boldsymbol{\Phi}(v)\|_{\mathfrak{T},p} \\ &\leq \omega^{\mathbb{T}}(\mathcal{D},\mathfrak{T},\boldsymbol{\Phi}) \|v\|_{\mathcal{D}} + C_5 \|\boldsymbol{\Phi}\|_{\mathcal{D},\mathfrak{T}} \|v\|_{\mathcal{D}}.\end{aligned}$$

The proof of (7.18) is concluded by reproducing the same steps starting from $\|\Pi_{\mathcal{D}} v\|_{L^p(\partial\Omega)}$ and using Lemma B.25 to control $\|\Pi_{\mathfrak{T}}\boldsymbol{\Phi}(v)\|_{L^p(\Omega)}$ by $\|\boldsymbol{\Phi}(v)\|_{\mathfrak{T},p}$.

We turn to (7.19). Let $\boldsymbol{\varphi} \in W^{1,p'}(\Omega)^d$ and use the triangle inequality, the definition of $\omega^{\Pi}(\mathcal{D},\mathfrak{T},\boldsymbol{\Phi})$ and $\omega^{\mathbb{T}}(\mathcal{D},\mathfrak{T},\boldsymbol{\Phi})$ and (B.31) to obtain

$$\begin{aligned}
&\left|\int_\Omega \Big(\nabla_{\mathcal{D}} v(\boldsymbol{x})\cdot\boldsymbol{\varphi}(\boldsymbol{x}) + \Pi_{\mathcal{D}} v(\boldsymbol{x})\mathrm{div}\boldsymbol{\varphi}(\boldsymbol{x})\Big)\mathrm{d}\boldsymbol{x} - \int_{\partial\Omega} \mathbb{T}_{\mathcal{D}} v(\boldsymbol{x})\gamma_{\boldsymbol{n}}(\boldsymbol{\varphi})(\boldsymbol{x})\mathrm{d}\gamma(\boldsymbol{x})\right| \\
&\quad\le \left|\int_\Omega [\nabla_{\mathcal{D}} v(\boldsymbol{x}) - \overline{\nabla}_{\mathfrak{T}}\boldsymbol{\Phi}(v)(\boldsymbol{x})]\cdot\boldsymbol{\varphi}(\boldsymbol{x})\right| + \|\mathrm{div}\boldsymbol{\varphi}\|_{L^{p'}(\Omega)}\,\omega^{\Pi}(\mathcal{D},\mathfrak{T},\boldsymbol{\Phi})\,\|v\|_{\mathcal{D}} \\
&\qquad + \|\gamma_{\boldsymbol{n}}(\boldsymbol{\varphi})\|_{L^{p'}(\partial\Omega)}\,\omega^{\mathbb{T}}(\mathcal{D},\mathfrak{T},\boldsymbol{\Phi})\,\|v\|_{\mathcal{D}} \\
&\qquad + \left|\int_\Omega \big(\overline{\nabla}_{\mathfrak{T}}\boldsymbol{\Phi}(v)(\boldsymbol{x})\cdot\boldsymbol{\varphi}(\boldsymbol{x}) + \Pi_{\mathfrak{T}}\boldsymbol{\Phi}(v)(\boldsymbol{x})\mathrm{div}\boldsymbol{\varphi}(\boldsymbol{x})\big)\,\mathrm{d}\boldsymbol{x}\right. \\
&\qquad \left. - \int_{\partial\Omega} \mathbb{T}_{\mathfrak{T}} v(\boldsymbol{x})\gamma_{\boldsymbol{n}}(\boldsymbol{\varphi})(\boldsymbol{x})\mathrm{d}\gamma(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\right| \\
&\quad\le \left|\int_\Omega [\nabla_{\mathcal{D}} v(\boldsymbol{x}) - \overline{\nabla}_{\mathfrak{T}}\boldsymbol{\Phi}(v)(\boldsymbol{x})]\cdot\boldsymbol{\varphi}(\boldsymbol{x})\right| + \|\mathrm{div}\boldsymbol{\varphi}\|_{L^{p'}(\Omega)}\,\omega^{\Pi}(\mathcal{D},\mathfrak{T},\boldsymbol{\Phi})\,\|v\|_{\mathcal{D}} \\
&\qquad + \|\gamma_{\boldsymbol{n}}(\boldsymbol{\varphi})\|_{L^{p'}(\partial\Omega)}\,\omega^{\mathbb{T}}(\mathcal{D},\mathfrak{T},\boldsymbol{\Phi})\,\|v\|_{\mathcal{D}} + C_6\,\|\,|\nabla\boldsymbol{\varphi}|\,\|_{L^{p'}(\Omega)^d}\,|\boldsymbol{\Phi}(v)|_{\mathfrak{T},p}\,h_{\mathcal{M}},
\end{aligned}$$

where C_6 depends only on d, p and ϱ. The first term in the right-hand side can be bounded above by using (7.16). Invoking the definition of $\|\boldsymbol{\Phi}\|_{\mathcal{D},\mathfrak{T}}$ then concludes the proof. ∎

Corollary 7.18 (Properties of GDs controlled by polytopal toolboxes – Neumann BCs). *Let $p > 1$, Ω satisfy Assumption (7.2), $(\mathcal{D}_m)_{m\in\mathbb{N}}$ be a sequence of GDs in the sense of Definition 3.11, and $(\mathfrak{T}_m)_{m\in\mathbb{N}}$ be a sequence of polytopal toolboxes in the sense of Definition 7.15. Assume that $h_{\mathcal{M}_m} \to 0$ as $m \to \infty$ and that $\sup_{m\in\mathbb{N}}(\theta_{\mathfrak{T}_m} + \eta_{\mathfrak{T}_m}) < +\infty$ (see (7.8) and (7.9)).*
For all $m \in \mathbb{N}$ we take a control $\boldsymbol{\Phi}_m$ of $\mathcal{D}_m$ by $\mathfrak{T}_m$ in the sense of Definition 7.16, and we assume that

$$\begin{aligned}
&\sup_{m\in\mathbb{N}} \|\boldsymbol{\Phi}_m\|_{\mathcal{D}_m,\mathfrak{T}_m} < +\infty, \\
&\lim_{m\to\infty} \omega^{\Pi}(\mathcal{D}_m,\mathfrak{T}_m,\boldsymbol{\Phi}_m) = 0, \\
&\lim_{m\to\infty} \omega^{\mathbb{T}}(\mathcal{D}_m,\mathfrak{T}_m,\boldsymbol{\Phi}_m) = 0, \ \textit{and} \\
&\lim_{m\to\infty} \omega^{\nabla}(\mathcal{D}_m,\mathfrak{T}_m,\boldsymbol{\Phi}_m) = 0.
\end{aligned}$$

Then $(\mathcal{D}_m)_{m\in\mathbb{N}}$ is coercive in the sense of Definition 3.13, limit-conforming in the sense of Definition 3.14, and compact in the sense of Definition 3.15.

Proof. The coercivity and limit-conformity follow from Estimates (7.18) and (7.19), from Lemma 3.19, and from the fact that $W^{1,p'}(\Omega)^d$ is dense in $W^{p'}_{\mathrm{div},\partial}(\Omega)$ (see Lemma 3.27).

To establish the compactness, we notice that if $v_m \in X_{\mathcal{D}_m}$ is such that $(\|v_m\|_{\mathcal{D}_m})_{m\in\mathbb{N}}$ is bounded, then so is $(\|\boldsymbol{\Phi}_m(v_m)\|_{\mathfrak{T}_m,p})_{m\in\mathbb{N}}$ since

$$\|\boldsymbol{\Phi}_m(v_m)\|_{\mathfrak{T}_m} \leq \|\boldsymbol{\Phi}_m\|_{\mathcal{D}_m,\mathfrak{T}_m} \|v_m\|_{\mathcal{D}_m} .$$

Hence, by Lemma B.27 and the definition (7.17) of $\|\boldsymbol{\Phi}_m(v_m)\|_{\mathfrak{T}_m,p}$, some subsequences of $(\Pi_{\mathfrak{T}_m}\boldsymbol{\Phi}_m(v_m))_{m\in\mathbb{N}}$ and $(\mathbb{T}_{\mathfrak{T}_m}\boldsymbol{\Phi}_m(v_m))_{m\in\mathbb{N}}$ converge strongly in $L^p(\Omega)$ and $L^p(\partial\Omega)$, respectively. The convergences of $\omega^{\Pi}(\mathcal{D}_m,\mathfrak{T}_m,\boldsymbol{\Phi}_m)$ and $\omega^{\mathbb{T}}(\mathcal{D}_m,\mathfrak{T}_m,\boldsymbol{\Phi}_m)$ towards 0 then ensure that, along the same subsequences, $(\Pi_{\mathcal{D}_m}v_m)_{m\in\mathbb{N}}$ converges in $L^p(\Omega)$ and $(\mathbb{T}_{\mathcal{D}_m}v_m)_{m\in\mathbb{N}}$ converges in $L^p(\partial\Omega)$, which completes the proof. ■

Remark 7.19 (Homogeneous Neumann boundary conditions). Homogeneous Neumann conditions are a particular case of non-homogeneous Neumann conditions, so all previous results also apply. However, if one is solely interested in homogeneous Neumann conditions, some simplifications can be made. Precisely, there is no need to include $\mathbb{T}_{\mathfrak{T}}$ in Definition 7.15, $\omega^{\mathbb{T}}$ in Definition 7.16 and Corollary 7.18, and Theorem 7.17 holds with $\omega^{\mathbb{T}}$ replaced with 0.

Remark 7.20 (Fourier boundary conditions). The only differences between GDs for non-homogeneous Neumann conditions and Fourier conditions are the definition of the norm $\|\ \|_{\mathcal{D}}$, and the definition of the GD-consistency. Since GD-consistency is not a notion covered by polytopal toolboxes, all previous results in this section apply to Fourier boundary conditions provided that the norm (7.17) is replaced with the norm defined by

$$\|v\|_{\mathfrak{T},p}^p = |v|_{\mathfrak{T},p}^p + \|\mathbb{T}_{\mathfrak{T}}v\|_{L^p(\partial\Omega)}^p .$$

Estimating $C_{\mathcal{D}}$ in Theorem 7.17 in the case of Fourier boundary conditions is straightforward thanks to Lemma B.22.

7.2.3 Mixed boundary conditions

Polytopal toolboxes for mixed boundary conditions are now going to be defined; the associated results are presented without proofs; they can be established quite similarly to the case of Dirichlet and Neumann boundary conditions, using Lemma B.32 and B.33.

Definition 7.21 (Polytopal toolbox for mixed BCs). *Under Assumptions (7.2) and (3.60), let $\mathfrak{T}$ be a polytopal mesh in the sense of Definition 7.2. The family $(X_{\mathfrak{T},\Omega,\Gamma_n}, \Pi_{\mathfrak{T}}, \mathbb{T}_{\mathfrak{T},\Gamma_n}, \overline{\nabla}_{\mathfrak{T}}, |\ |_{\mathfrak{T},p})$ is a polytopal toolbox for mixed boundary conditions if:*

1. *The set $X_{\mathfrak{T},\Omega,\Gamma_n}$ is defined by (B.77).*
2. *The function reconstruction $\Pi_{\mathfrak{T}} : X_{\mathfrak{T}} \to L^\infty(\Omega)$ is defined by (7.7c).*
3. *The trace reconstruction $\mathbb{T}_{\mathfrak{T},\Gamma_n} : X_{\mathfrak{T}} \to L^\infty(\Gamma_n)$ is the restriction to Γ_n of the discrete trace (7.7d).*

4. *The gradient reconstruction* $\overline{\nabla}_{\mathfrak{T}} : X_{\mathfrak{T}} \to L^\infty(\Omega)^d$ *is defined by* (7.7e).
5. *The space* $X_{\mathfrak{T},\Omega,\Gamma_n}$ *is endowed with the norm* $|\ |_{\mathfrak{T},p}$ *defined by* (7.7f).

Definition 7.22 (Control of a GD by a polytopal toolbox – mixed BCs). *Under Assumptions* (7.2) *and* (3.60)*, let* $\mathcal{D}$ *be a GD in the sense of Definition 3.42, and let* $\mathfrak{T}$ *be a polytopal toolbox in the sense of Definition 7.21. A control of* $\mathcal{D}$ *by* $\mathfrak{T}$ *is a linear mapping* $\boldsymbol{\Phi} : X_{\mathcal{D},\Omega,\Gamma_n} \to X_{\mathfrak{T},\Omega,\Gamma_n}$*. We then define*

$$\|\boldsymbol{\Phi}\|_{\mathcal{D},\mathfrak{T}} = \max_{v\in X_{\mathcal{D},\Omega,\Gamma_n}\setminus\{0\}} \frac{|\boldsymbol{\Phi}(v)|_{\mathfrak{T},p}}{\|v\|_{\mathcal{D}}},$$

$$\omega^{\Pi}(\mathcal{D},\mathfrak{T},\boldsymbol{\Phi}) = \max_{v\in X_{\mathcal{D},\Omega,\Gamma_n}\setminus\{0\}} \frac{\|\Pi_{\mathcal{D}} v - \Pi_{\mathfrak{T}}\boldsymbol{\Phi}(v)\|_{L^p(\Omega)}}{\|v\|_{\mathcal{D}}},$$

$$\omega^{\mathbb{T}}(\mathcal{D},\mathfrak{T},\boldsymbol{\Phi}) = \max_{v\in X_{\mathcal{D},\Omega,\Gamma_n}\setminus\{0\}} \frac{\|\mathbb{T}_{\mathcal{D},\Gamma_n} v - \mathbb{T}_{\mathfrak{T},\Gamma_n}\boldsymbol{\Phi}(v)\|_{L^p(\Gamma_n)}}{\|v\|_{\mathcal{D}}},$$

$$\omega^{\nabla}(\mathcal{D},\mathfrak{T},\boldsymbol{\Phi}) = \max_{\substack{v\,\in\, X_{\mathcal{D},\Omega,\Gamma_n}\\ v\neq 0}} \frac{\left[\displaystyle\sum_{K\in\mathcal{M}} |K|^{1-p} \left|\int_K \left[\nabla_{\mathcal{D}} v(\boldsymbol{x}) - \overline{\nabla}_{\mathfrak{T}}\boldsymbol{\Phi}(v)(\boldsymbol{x})\right] \mathrm{d}\boldsymbol{x}\right|^p\right]^{\frac{1}{p}}}{\|v\|_{\mathcal{D}}}.$$

Theorem 7.23 (Estimates for a GD controlled by polytopal toolboxes – mixed BCs). *Under Assumptions* (7.2) *and* (3.60)*, let* $\mathcal{D}$ *be a GD in the sense of Definition 3.42, and let* $\mathfrak{T}$ *be a polytopal toolbox in the sense of Definition 7.21. We take* $\boldsymbol{\Phi}$ *a control of* $\mathcal{D}$ *by* $\mathfrak{T}$ *in the sense of Definition 7.22, and* $\varrho \geq \theta_{\mathfrak{T}} + \eta_{\mathfrak{T}}$ *(see* (7.8) *and* (7.9)*).*
Then, there exists C_7 *depending only on* Ω*,* p *and* ϱ *such that*

$$C_{\mathcal{D}} \leq \max\left(\omega^{\Pi}(\mathcal{D},\mathfrak{T},\boldsymbol{\Phi}), \omega^{\mathbb{T}}(\mathcal{D},\mathfrak{T},\boldsymbol{\Phi})\right) + C_4 \|\boldsymbol{\Phi}\|_{\mathcal{D},\mathfrak{T}}$$

and, for all $\boldsymbol{\varphi} \in W^{1,p'}(\Omega)^d$*,*

$$W_{\mathcal{D}}(\boldsymbol{\varphi}) \leq \|\boldsymbol{\varphi}\|_{W^{1,p'}(\Omega)^d} \Big[C_4 h_{\mathcal{M}}(1 + \|\boldsymbol{\Phi}\|_{\mathcal{D},\mathfrak{T}}) + \omega^{\Pi}(\mathcal{D},\mathfrak{T},\boldsymbol{\Phi}) + \omega^{\nabla}(\mathcal{D},\mathfrak{T},\boldsymbol{\Phi}) + \omega^{\mathbb{T}}(\mathcal{D},\mathfrak{T},\boldsymbol{\Phi})\Big].$$

Here, $C_{\mathcal{D}}$ *and* $W_{\mathcal{D}}$ *are the coercivity constant and limit-conformity measure defined by* (3.61) *and* (3.65)*.*

Corollary 7.24 (Properties of GDs controlled by polytopal toolboxes – mixed BCs). *Under Assumptions* (7.2) *and* (3.60)*, let* $(\mathcal{D}_m)_{m\in\mathbb{N}}$ *be a sequence of GDs in the sense of Definition 3.42, and let* $(\mathfrak{T}_m)_{m\in\mathbb{N}}$ *be a sequence of polytopal toolboxes in the sense of Definition 7.21. We assume that* $h_{\mathcal{M}_m} \to 0$ *as* $m \to \infty$ *and that* $\sup_{m\in\mathbb{N}}(\theta_{\mathfrak{T}_m} + \eta_{\mathfrak{T}_m}) < +\infty$ *(see* (7.8) *and* (7.9)*).*
For all $m \in \mathbb{N}$ *we take a control* $\boldsymbol{\Phi}_m$ *of* $\mathcal{D}_m$ *by* $\mathfrak{T}_m$ *in the sense of Definition 7.22, and we assume that*

$$\sup_{m\in\mathbb{N}} \|\boldsymbol{\Phi}_m\|_{\mathcal{D}_m,\mathfrak{T}_m} < +\infty,$$
$$\lim_{m\to\infty} \omega^{\Pi}(\mathcal{D}_m,\mathfrak{T}_m,\boldsymbol{\Phi}_m) = 0,$$
$$\lim_{m\to\infty} \omega^{\mathbb{T}}(\mathcal{D}_m,\mathfrak{T}_m,\boldsymbol{\Phi}_m) = 0, \text{ and}$$
$$\lim_{m\to\infty} \omega^{\nabla}(\mathcal{D}_m,\mathfrak{T}_m,\boldsymbol{\Phi}_m) = 0.$$

Then $(\mathcal{D}_m)_{m\in\mathbb{N}}$ is coercive in the sense of Definition 3.43, limit-conforming in the sense of Definition 3.45, and compact in the sense of Definition 3.47.

The only (slightly) non-trivial adaptation of the preceding proofs to establish this corollary is the density of smooth functions in $W^{p'}_{\mathrm{div},\Gamma_n}(\Omega)$, endowed with the norm $\|\boldsymbol{\varphi}\|_{L^{p'}(\Omega)^d}+\|\mathrm{div}\boldsymbol{\varphi}\|_{L^{p'}(\Omega)}+\|\gamma_{\mathbf{n}}(\boldsymbol{\varphi})\|_{L^{p'}(\Gamma_n)}$. This density is actually established in a similar way as in Lemma 3.27, where in Item 2 we take ψ_ϵ such that

$$\|\gamma_{\mathbf{n}}(\boldsymbol{\varphi})-\psi_\epsilon\|_{(W^{1-\frac{1}{p},p}(\partial\Omega))'} + \|\gamma_{\mathbf{n}}(\boldsymbol{\varphi})-\psi_\epsilon\|_{L^{p'}(\Gamma_n)} \le \varepsilon.$$

7.3 Local linearly exact GDs

7.3.1 $\mathbb{P}_0$-exact and $\mathbb{P}_1$-exact reconstructions

Most numerical methods for diffusion equations are based, either explicitly or implicitly, on reconstructions of functions – or gradients – from discrete unknowns. These reconstructions are designed to match certain simple functions (*e.g.*, constant, or affine) – or their gradients – when the discrete unknowns interpolate these functions at certain points. We give here a precise meaning to these notions, and state some of their approximation properties of the corresponding reconstructions.

Definition 7.25 ($\mathbb{P}_0$-exact function reconstruction). *Let I be a finite set, K be a bounded subset of $\mathbb{R}^d$ with non-zero measure, and $p\in[1,+\infty]$.*
A $\mathbb{P}_0$-exact function reconstruction on K is a family $\pi_K=(\pi_K^i)_{i\in I}$ of functions in $L^p(K)$ such that

$$\text{for a.e. } \boldsymbol{x}\in K\,,\ \sum_{i\in I}\pi_K^i(\boldsymbol{x})=1. \tag{7.20}$$

The norm of π_K is defined by (setting $|K|^{-\frac{1}{p}}=1$ if $p=+\infty$)

$$\|\pi_K\|_p = |K|^{-\frac{1}{p}}\left\|\sum_{i\in I}|\pi_K^i|\right\|_{L^p(K)}. \tag{7.21}$$

If $v=(v_i)_{i\in I}$ is a family of real numbers, $\pi_K v$ denotes the function in $L^p(K)$ given by:

$$\text{for a.e. } \boldsymbol{x}\in K\,,\ (\pi_K v)(\boldsymbol{x})=\sum_{i\in I} v_i\pi_K^i(\boldsymbol{x}).$$

Property (7.20) shows that, if $v = (v_i)_{i\in I}$ is such that there exists $c \in \mathbb{R}$ with $v_i = c$ for all $i \in I$, then $\pi_K v = c$ a.e. on K. The reconstruction π_K is therefore exact on interpolants of constant functions.

Example 7.26 (Elementary basis functions for non-conforming $\mathbb{P}_1$ finite element)
Let K be a simplex. For each $\sigma \in \mathcal{F}_K$, let π_K^σ be the affine function in K that has value 1 at $\overline{\boldsymbol{x}}_\sigma$ and 0 at $\overline{\boldsymbol{x}}_{\sigma'}$ for all face $\sigma' \neq \sigma$ of K. Then $\sum_{\sigma\in\mathcal{F}_K} \pi_K^\sigma = 1$ on K, that is, $\pi_K = (\pi_K^\sigma)_{\sigma\in\mathcal{F}_K}$ is a $\mathbb{P}_0$-exact function reconstruction on K.

Since regular functions are locally close to constant functions, it is expected that $\mathbb{P}_0$-exact function reconstructions enjoy some approximation properties when computed on interpolants of regular functions.

Lemma 7.27 (Interpolation estimate for $\mathbb{P}_0$-exact function reconstruction). *Let I be a finite set, K be a bounded subset of $\mathbb{R}^d$ with non-zero measure, $p \in [1, +\infty]$, $\pi_K = (\pi_K^i)_{i\in I}$ be a $\mathbb{P}_0$-exact function reconstruction on K, and $(\boldsymbol{x}_i)_{i\in I}$ be points in $\mathbb{R}^d$.*
Then, if $\varphi \in W^{1,\infty}(\mathbb{R}^d)$ and $v = (\varphi(\boldsymbol{x}_i))_{i\in I}$,

$$\begin{aligned}&\|\pi_K v - \varphi\|_{L^p(K)}\\ &\qquad\le \left(1 + \max_{i\in I} \frac{\operatorname{dist}(\boldsymbol{x}_i, K)}{\operatorname{diam}(K)}\right) \|\pi_K\|_p \, |K|^{\frac{1}{p}} \operatorname{diam}(K) \, \||\nabla\varphi|\|_{L^\infty(\mathbb{R}^d)} .\end{aligned}$$

Proof. For a.e. $\boldsymbol{x} \in K$, using (7.20) yields

$$\varphi(\boldsymbol{x}) = \varphi(\boldsymbol{x}) \sum_{i\in I} \pi_K^i(\boldsymbol{x}) = \sum_{i\in I} \pi_K^i(\boldsymbol{x})\varphi(\boldsymbol{x}).$$

Moreover, for any $i \in I$ and $\boldsymbol{x} \in K$,

$$\begin{aligned}|\varphi(\boldsymbol{x}_i) - \varphi(\boldsymbol{x})| &\le |\boldsymbol{x}_i - \boldsymbol{x}| \, \||\nabla\varphi|\|_{L^\infty(\mathbb{R}^d)}\\ &\le (\operatorname{diam}(K) + \operatorname{dist}(\boldsymbol{x}_i, K)) \, \||\nabla\varphi|\|_{L^\infty(\mathbb{R}^d)} .\end{aligned}$$

Hence, for a.e. $\boldsymbol{x} \in K$,

$$\begin{aligned}|\pi_K v(\boldsymbol{x}) - \varphi(\boldsymbol{x})| &= \left|\sum_{i\in I} \pi_K^i(\boldsymbol{x})(v_i - \varphi(\boldsymbol{x}))\right| \qquad (7.22)\\ &\le \max_{i\in I} |\varphi(\boldsymbol{x}_i) - \varphi(\boldsymbol{x})| \sum_{i\in I} |\pi_K^i(\boldsymbol{x})|\\ &\le \left(1 + \max_{i\in I} \frac{\operatorname{dist}(\boldsymbol{x}_i, K)}{\operatorname{diam}(K)}\right) \operatorname{diam}(K) \, \||\nabla\varphi|\|_{L^\infty(\mathbb{R}^d)} \sum_{i\in I} |\pi_K^i(\boldsymbol{x})|.\end{aligned}$$

The proof is complete by taking the $L^p(K)$ norm over $\boldsymbol{x}$ and by using the definition of $\|\pi_K\|_p$. ∎

We now turn to the notion of gradient reconstructions that are exact on interpolants of affine functions.

Definition 7.28 ($\mathbb{P}_1$-exact gradient reconstruction). *Let I be a finite set, K be a bounded subset of $\mathbb{R}^d$ with non-zero measure, $p \in [1, +\infty]$, and $S = (\boldsymbol{x}_i)_{i\in I}$ be a family of points in $\mathbb{R}^d$.*
A $\mathbb{P}_1$-exact gradient reconstruction on K upon S is a family $\mathcal{G}_K = (\mathcal{G}_K^i)_{i\in I}$ of functions in $L^p(K)^d$ satisfying the following property:

$$\text{for any affine } A : \mathbb{R}^d \to \mathbb{R} \text{ and a.e. } \boldsymbol{x} \in K\,,\ \sum_{i\in I} A(\boldsymbol{x}_i)\mathcal{G}_K^i(\boldsymbol{x}) = \nabla A. \quad (7.23)$$

The norm of $\mathcal{G}_K$ is defined by (setting $|K|^{-\frac{1}{p}} = 1$ if $p = +\infty$)

$$\|\mathcal{G}_K\|_p = \operatorname{diam}(K)|K|^{-\frac{1}{p}} \left\| \sum_{i\in I} |\mathcal{G}_K^i| \right\|_{L^p(K)}. \quad (7.24)$$

If $v = (v_i)_{i\in I}$ is a family of real numbers, $\mathcal{G}_K v$ denotes the function in $L^p(K)^d$ given by:

$$\text{for a.e. } \boldsymbol{x} \in K\,,\ (\mathcal{G}_K v)(\boldsymbol{x}) = \sum_{i\in I} v_i \mathcal{G}_K^i(\boldsymbol{x}).$$

We notice from (7.23) that

$$\text{for all affine function } A, \text{ if } v = (A(\boldsymbol{x}_i))_{i\in I} \text{ then } \mathcal{G}_K v = \nabla A. \quad (7.25)$$

This is the $\mathbb{P}_1$-exactness of the gradient reconstruction $\mathcal{G}_K$.

Example 7.29 ($\mathcal{G}_K$ for non-conforming $\mathbb{P}_1$ finite element)
Let K be a simplex. Recalling the definition of $\pi_K = (\pi_K^\sigma)_{\sigma\in\mathcal{F}_K}$ in Remark 7.26, we let $\mathcal{G}_K^\sigma = \nabla\pi_K^\sigma \in L^p(K)^d$. As proved in Lemma 9.12, the family $\mathcal{G}_K = (\mathcal{G}_K^\sigma)_{\sigma\in\mathcal{F}_K}$ is a $\mathbb{P}_1$-exact gradient reconstruction on K upon $S = (\overline{\boldsymbol{x}}_\sigma)_{\sigma\in\mathcal{F}_K}$. The fact that $\mathcal{G}_K^i$ is the gradient of π_K^i, which also holds for conforming finite elements, is a very specific property. It is not satisfied by a number of other schemes such as mixed finite elements, hybrid mimetic mixed methods, etc. (see Chapters 10, 12, 13 and 14), nor by mass-lumped schemes (see Example 7.46). Hence there are a number of methods for which a full description cannot be given only by the elementary functions π_K^i, but also requires a separate definition of the local gradients $\mathcal{G}_K^i$.

Remark 7.30 (Why the factor $\mathrm{diam}(K)|K|^{-1/p}$ *in* (7.24)*?)*
The scaling $\mathrm{diam}(K)|K|^{-1/p}$ in $\|\mathcal{G}_K\|_p$ is chosen to ensure that this norm provides an intrinsic measure of $\mathcal{G}_K$, that does not blow up or vanish as the size of K goes to zero.
To understand this, take $\boldsymbol{x}_K \in K$ and consider $A(\boldsymbol{x}) = \boldsymbol{\xi} \cdot (\boldsymbol{x} - \boldsymbol{x}_K)$ for some $\boldsymbol{\xi} \in \mathbb{R}^d$. Then (7.23) gives

$$\boldsymbol{\xi} = \sum_{i \in I} \boldsymbol{\xi} \cdot (\boldsymbol{x}_i - \boldsymbol{x}_K) \mathcal{G}_K^i(\boldsymbol{x}). \tag{7.26}$$

The gradient reconstruction over K is expected to be based on points $\boldsymbol{x}_i$ not far from K; in practice, this means that each $\boldsymbol{x}_i$ is expected to remain within distance $\mathcal{O}(\mathrm{diam}(K))$ of K. Hence, assuming that $|\boldsymbol{\xi}| = 1$ and taking the $L^p(K)^d$ norm of (7.26) yields

$$|K|^{\frac{1}{p}} \leq C\mathrm{diam}(K) \left\| \sum_{i \in I} |\mathcal{G}_K^i| \right\|_{L^p(K)}.$$

This shows that $\mathrm{diam}(K)|K|^{-1/p} \left\| \sum_{i \in I} |\mathcal{G}_K^i| \right\|_{L^p(K)} =: \|\mathcal{G}_K\|_p$ remains bounded away from 0, independently of size of K. Any other scaling in $\|\mathcal{G}_K\|_p$ would lead to a quantity that could either go to 0 or to $+\infty$ as the size of K goes to 0.
A similar reasoning can be done to understand the need for the scaling factor $|K|^{-1/p}$ in (7.21).

In a similar way as for $\mathbb{P}_0$-exact function reconstructions above, the fact that any smooth function is locally close to an affine function ensures that $\mathbb{P}_1$-exact gradient reconstructions enjoy approximation properties.

Lemma 7.31 (Interpolation estimate for $\mathbb{P}_1$-exact gradient reconstructions). *Let I be a finite set, K be a bounded subset of $\mathbb{R}^d$ with non-zero measure, $p \in [1, +\infty]$, $S = (\boldsymbol{x}_i)_{i \in I} \subset \mathbb{R}^d$, and $\mathcal{G}_K = (\mathcal{G}_K^i)_{i \in I}$ be a $\mathbb{P}_1$-exact gradient reconstruction on K upon S.*
Then, if $\varphi \in W^{2,\infty}(\mathbb{R}^d)$ and $v = (\varphi(\boldsymbol{x}_i))_{i \in I}$,

$$\begin{aligned} \|\mathcal{G}_K v - \nabla\varphi\|_{L^p(K)^d} \leq & \left(1 + \frac{1}{2}\|\mathcal{G}_K\|_p \left[1 + \max_{i \in I} \frac{\mathrm{dist}(\boldsymbol{x}_i, K)}{\mathrm{diam}(K)}\right]^2\right) \\ & \times |K|^{\frac{1}{p}} \mathrm{diam}(K) \left\| |D^2\varphi| \right\|_{L^\infty(\mathbb{R}^d)}. \end{aligned} \tag{7.27}$$

Proof. Let us first assume that $\varphi \in C_b^2(\mathbb{R}^d)$. Take $\boldsymbol{x}_K \in K$ and let $A(\boldsymbol{x}) = \varphi(\boldsymbol{x}_K) + \nabla\varphi(\boldsymbol{x}_K) \cdot (\boldsymbol{x} - \boldsymbol{x}_K)$ be the first order Taylor expansion of φ around $\boldsymbol{x}_K$. If $\xi = (A(\boldsymbol{x}_i))_{i \in I}$, by $\mathbb{P}_1$-exactness (7.25) of $\mathcal{G}_K$ we have $\mathcal{G}_K \xi = \nabla A = \nabla\varphi(\boldsymbol{x}_K)$ on K. Hence, since $\nabla\varphi$ is Lipschitz-continuous with a Lipschitz constant bounded above by $\left\| |D^2\varphi| \right\|_{L^\infty(\mathbb{R}^d)}$, we write

$$\begin{aligned} \|\mathcal{G}_K \xi - \nabla\varphi\|_{L^p(K)^d} &= \|\nabla\varphi(\boldsymbol{x}_K) - \nabla\varphi\|_{L^p(K)^d} \\ &\leq |K|^{\frac{1}{p}} \|\nabla\varphi(\boldsymbol{x}_K) - \nabla\varphi\|_{L^\infty(K)^d} \\ &\leq |K|^{\frac{1}{p}} \mathrm{diam}(K) \left\| |D^2\varphi| \right\|_{L^\infty(\mathbb{R}^d)}. \end{aligned} \tag{7.28}$$

For any $i \in I$ we have, by Taylor's expansion,

$$\begin{aligned} v_i - \xi_i &= \varphi(\boldsymbol{x}_i) - A(\boldsymbol{x}_i) \\ &= \varphi(\boldsymbol{x}_i) - \varphi(\boldsymbol{x}_K) - \nabla\varphi(\boldsymbol{x}_K)\cdot(\boldsymbol{x}_i - \boldsymbol{x}_K) \\ &= \int_0^1 (1-s) D^2\varphi(\boldsymbol{x}_K + s(\boldsymbol{x}_i - \boldsymbol{x}_K))(\boldsymbol{x}_i - \boldsymbol{x}_K)\cdot(\boldsymbol{x}_i - \boldsymbol{x}_K)\mathrm{d}s. \end{aligned} \tag{7.29}$$

Using $|\boldsymbol{x}_i - \boldsymbol{x}_K| \le \operatorname{diam}(K) + \operatorname{dist}(\boldsymbol{x}_i, K)$ yields

$$|v_i - \xi_i| \le \frac{1}{2}[\operatorname{diam}(K) + \operatorname{dist}(\boldsymbol{x}_i, K)]^2 \left\| |D^2\varphi| \right\|_{L^\infty(\mathbb{R}^d)}. \tag{7.30}$$

Hence, for a.e. $\boldsymbol{x} \in K$,

$$\begin{aligned} |\mathcal{G}_K v(\boldsymbol{x}) - \mathcal{G}_K \xi(\boldsymbol{x})| &= \left| \sum_{i\in I} (v_i - \xi_i)\mathcal{G}_K^i(\boldsymbol{x}) \right| \\ &\le \frac{1}{2}\left[\operatorname{diam}(K) + \max_{i\in I}\operatorname{dist}(\boldsymbol{x}_i, K)\right]^2 \left\| |D^2\varphi| \right\|_{L^\infty(\mathbb{R}^d)} \sum_{i\in I} |\mathcal{G}_K^i(\boldsymbol{x})|. \end{aligned}$$

Taking the $L^p(K)$ norm over $\boldsymbol{x}$ and recalling the definition of $\|\mathcal{G}_K\|_p$ leads to

$$\begin{aligned} &\|\mathcal{G}_K v - \mathcal{G}\xi\|_{L^p(K)^d} \\ &\quad\le \frac{|K|^{\frac{1}{p}}\|\mathcal{G}_K\|_p}{\operatorname{diam}(K)}\frac{1}{2}\left[\operatorname{diam}(K) + \max_{i\in I}\operatorname{dist}(\boldsymbol{x}_i, K)\right]^2 \left\| |D^2\varphi| \right\|_{L^\infty(\mathbb{R}^d)} \\ &\quad= \frac{1}{2}\|\mathcal{G}_K\|_p \left[1 + \max_{i\in I}\frac{\operatorname{dist}(\boldsymbol{x}_i, K)}{\operatorname{diam}(K)}\right]^2 |K|^{\frac{1}{p}}\operatorname{diam}(K) \left\| |D^2\varphi| \right\|_{L^\infty(\mathbb{R}^d)}. \end{aligned}$$

Combined with (7.28) and a triangle inequality, this completes the proof of the lemma if $\varphi \in C_b^2(\mathbb{R}^d)$.
If $\varphi \in W^{2,\infty}(\mathbb{R}^d)$, use convolution to construct a sequence $(\varphi_n)_{n\in\mathbb{N}} \subset C_b^2(\mathbb{R}^d)$ such that $\varphi_n \to \varphi$ and $\nabla\varphi_n \to \nabla\varphi$ uniformly on compact sets as $n \to \infty$, and $\left\| |D^2\varphi_n| \right\|_{L^\infty(\mathbb{R}^d)} \le \left\| |D^2\varphi| \right\|_{L^\infty(\mathbb{R}^d)}$ for all $n \in \mathbb{N}$. Apply then (7.27) to φ_n and pass to the limit $n \to \infty$ to prove the same estimate for φ. ∎

Remark 7.32. For all functions and gradient reconstructions considered in Chapters 8–14, the functions π_K^i (resp. $\mathcal{G}_K^i$) have values in $L^\infty(K)$ (resp. $L^\infty(K)^d$). In that case, by Hölder's inequality,

$$\|\pi_K\|_p \le \|\pi_K\|_\infty = \operatorname*{esssup}_{\boldsymbol{x}\in K} \sum_{i\in I} |\pi_K^i(\boldsymbol{x})|$$

(where esssup is the essential supremum) and

$$\|\mathcal{G}_K\|_p \le \|\mathcal{G}_K\|_\infty \le \operatorname{diam}(K) \sum_{i\in I} \left\|\mathcal{G}_K^i\right\|_{L^\infty(K)^d}.$$

These estimates will be used, when analysing specific GDs in Chapters 8–14, to obtain upper bounds on $\|\pi_K\|_p$ and $\|\mathcal{G}_K\|_p$.

In a number of cases, estimating $\|\pi_K\|_\infty$ (and thus $\|\pi_K\|_p$) is trivial. For example, if for a.e. $\boldsymbol{x} \in K$ the value $\pi_K v(\boldsymbol{x})$ is computed as a convex combination of the real numbers $(v_i)_{i\in I}$, then $\pi_K^i \geq 0$ for all $i \in I$ and $\sum_{i\in I} |\pi_K^i(\boldsymbol{x})| = \sum_{i\in I} \pi_K^i(\boldsymbol{x}) = 1$. This is for instance the case, *e.g.*, if $\pi_K v$ is linear on K, $v_i = \pi_K v(\boldsymbol{x}_i)$ and $(\boldsymbol{x}_i)_{i\in I}$ are extremal points of K (this situation appears in the conforming linear $\mathbb{P}_1$ finite element method).
Another example is the case where for a.e. $\boldsymbol{x} \in K$ there is exactly one $i \in I$ such that $\pi_K^i(\boldsymbol{x}) = 1$, and $\pi_j(\boldsymbol{x}) = 0$ for all other $j \in I$. Then $\sum_{i\in I} |\pi_K^i(\boldsymbol{x})| = 1$ a.e. on K (and $\pi_K v$ is piecewise constant on K). This situation occurs in the case of the mass-lumped $\mathbb{P}_1$ finite element method, see Section 8.4.

7.3.2 Definition and consistency of local linearly exact GDs for Dirichlet boundary conditions

The previous concepts of $\mathbb{P}_0/\mathbb{P}_1$-exact function/gradient reconstructions are useful to establish the GD-consistency, through the following notion of local linearly exact gradient discretisation (LLE GD). This notion applies to the vast majority of GDs analysed in the Chapters 8–14. LLE GDs are the gradient discretisations whose function reconstructions are locally $\mathbb{P}_0$-exact and whose gradient reconstructions are locally $\mathbb{P}_1$-exact, both reconstructions being computed locally. To measure this locality, a regularity parameter $\mathrm{reg}_{\mathrm{LLE}}$ is introduced; the boundedness of this parameter imposes that, at any given point $\boldsymbol{x}$, the reconstructed functions and gradients are computed by discrete unknowns (or zero values) located not far from $\boldsymbol{x}$.

Definition 7.33 (Local linearly exact gradient discretisation (LLE GD)). *A gradient discretisation $\mathcal{D} = (X_{\mathcal{D},0}, \Pi_{\mathcal{D}}, \nabla_{\mathcal{D}})$ in the sense of Definition 2.1 is a local linearly exact gradient discretisation (LLE GD) if:*

1. *There exists a finite set I of geometrical entities attached to the discrete unknowns, where I is partitioned into I_Ω (interior geometrical entities attached to the discrete unknowns) and I_∂ (boundary geometrical entities attached to the discrete unknowns), such that*

$$X_{\mathcal{D},0} = \{v = (v_i)_{i\in I} \,:\, v_i \in \mathbb{R} \text{ for all } i \in I\,,\ v_i = 0 \text{ if } i \in I_\partial\}. \tag{7.31}$$

2. *There exists a family of approximation points $S = (\boldsymbol{x}_i)_{i\in I} \subset \mathbb{R}^d$, a mesh $\mathcal{M}$ of Ω and, for each $K \in \mathcal{M}$, a subset $I_K \subset I$ and*
   a) *a $\mathbb{P}_0$-exact function reconstruction $\pi_K = (\pi_K^i)_{i\in I_K}$ on K (see Definition 7.25) such that*

$$\begin{aligned} &\forall v \in X_{\mathcal{D},0}\,, \quad \text{for a.e. } x \in K\,, \\ &\Pi_{\mathcal{D}} v(\boldsymbol{x}) = \pi_K[(v_i)_{i\in I_K}](\boldsymbol{x}) = \sum_{i\in I_K} v_i \pi_K^i(\boldsymbol{x}), \end{aligned} \tag{7.32}$$

b) a $\mathbb{P}_1$-exact gradient reconstruction $\mathcal{G}_K = (\mathcal{G}_K^i)_{i\in I_K}$ on K upon $(\boldsymbol{x}_i)_{i\in I_K}$ (see Definition 7.28) such that

$$\begin{aligned} &\forall v \in X_{\mathcal{D},0}\,, \quad \textit{for a.e. } x \in K\,, \\ &\nabla_{\mathcal{D}} v(\boldsymbol{x}) = \mathcal{G}_K[(v_i)_{i\in I_K}](\boldsymbol{x}) = \sum_{i\in I_K} v_i \mathcal{G}_K^i(\boldsymbol{x}). \end{aligned} \tag{7.33}$$

Here, the mesh $\mathcal{M}$ is merely a finite family of open disjoint subsets of Ω such that $\bigcup_{K\in\mathcal{M}} \overline{K} = \overline{\Omega}$. Its size is $h_{\mathcal{M}} = \max_{K\in\mathcal{M}} \operatorname{diam}(K)$, and the LLE regularity of $\mathcal{D}$ is defined by

$$\operatorname{reg}_{\text{LLE}}(\mathcal{D}) = \max_{K\in\mathcal{M}} \left(\|\pi_K\|_p + \|\mathcal{G}_K\|_p + \max_{i\in I_K} \frac{\operatorname{dist}(\boldsymbol{x}_i, K)}{\operatorname{diam}(K)} \right). \tag{7.34}$$

Example 7.34 (LLE GD interpretation of the non-conforming $\mathbb{P}_1$ gradient discretisation)
Example 7.6 defines the $\mathbb{P}_1$ gradient discretisation $\mathcal{D}$ in a global way. For both the analysis and the practical implementation, a definition starting from elementary basis functions is necessary. This gradient discretisation is an LLE GD for which the set of geometrical entities attached to the discrete unknowns are the faces of the mesh (that is, $I = \mathcal{F}$), the elementary basis functions are the π_K^σ described in Remark 7.26, and the local gradients $\mathcal{G}_K^\sigma$ are given in Remark 7.29.
It is proved in Lemma 9.12 that, under standard regularity assumptions on $\mathfrak{T}$, $\operatorname{reg}_{\text{LLE}}(\mathcal{D})$ is bounded. Used in Proposition 7.37 below, this bound yields the consistency of non-conforming $\mathbb{P}_1$ gradient discretisations.

Remark 7.35 (Geometrical entities attached to the discrete unknowns)
For most methods considered in Part III, the space $X_{\mathcal{D},0}$ is defined by (7.31) with I a certain set of elements associated with a mesh of Ω. For example, each $i \in I$ can correspond to a cell, a face or a vertex of the mesh. These mesh elements are what we call "geometrical entities attached to the discrete unknowns".
The discrete unknowns themselves are just the components v_i of a vector $v \in X_{\mathcal{D},0}$.

Definition 7.33 calls for a few comments. First, it is not required that the mesh $\mathcal{M}$ satisfies Item 1 in Definition 7.2 of a polytopal mesh. Nevertheless, in all the examples of LLE GDs encountered in Chapters 8–14, the mesh $\mathcal{M}$ in Definition 7.33 is indeed the set of cells of some polytopal mesh $\mathfrak{T} = (\mathcal{M}, \mathcal{F}, \mathcal{P}, \mathcal{V})$. Note that the choice of $h_{\mathcal{M}}$ in Definition 7.33 is the same as (7.6) in Definition 7.2.

The set S in Definition 7.33 might be such that there exist $i, j \in I$ with $i \neq j$ and $\boldsymbol{x}_i = \boldsymbol{x}_j$. This means that two different discrete unknowns may be

located at the same point $\boldsymbol{x}_i = \boldsymbol{x}_j$. This happens for instance in the case of the MPFA-O scheme, see Remark 12.1, page339.

Finally, the function reconstruction $\Pi_{\mathcal{D}}$ of an LLE GD is not necessarily locally $\mathbb{P}_1$-exact; only the local $\mathbb{P}_0$-exactness is required. This enables us to consider gradient discretisations with piecewise constant reconstructions (see Definition 2.12), and in particular mass-lumped GDs (see Section 7.3.5 below).

Remark 7.36 (Generalisation of $\mathrm{reg}_{\mathrm{LLE}}$*)*
The term $\mathrm{diam}(K)$ in $\mathrm{reg}_{\mathrm{LLE}}(\mathcal{D})$ could be replaced by any quantity $\omega_K > 0$: the requirement to prove Proposition 7.37 below is that $\max_{K\in\mathcal{M}_m} \omega_K \to 0$ as $m \to \infty$.

Proposition 7.37 (LLE GDs are consistent). *Let $(\mathcal{D}_m)_{m\in\mathbb{N}}$ be a sequence of LLE GDs (in the sense of Definition 7.33), with associated meshes $(\mathcal{M}_m)_{m\in\mathbb{N}}$. If $(\mathrm{reg}_{\mathrm{LLE}}(\mathcal{D}_m))_{m\in\mathbb{N}}$ is bounded and $h_{\mathcal{M}_m} \to 0$ as $m \to \infty$, then $(\mathcal{D}_m)_{m\in\mathbb{N}}$ is GD-consistent in the sense of Definition 2.4,* i.e. $S_{\mathcal{D}_m}(\varphi) \to 0$ *as $m \to +\infty$, for any $\varphi \in W_0^{1,p}(\Omega)$.*

Proof. Lemma 2.16 states that the convergence of $S_{\mathcal{D}_m}(\varphi)$ to zero only needs to be proved for functions in a dense subspace of $W_0^{1,p}(\Omega)$. Having in mind to use Lemmas 7.27 and 7.31, we take $W_0^{1,p}(\Omega) \cap W^{2,\infty}(\mathbb{R}^d)$ as the dense subspace in $W_0^{1,p}(\Omega)$ (the space $C_c^\infty(\Omega)$ would also be adequate).
Let $\varphi \in W_0^{1,p}(\Omega) \cap W^{2,\infty}(\mathbb{R}^d)$ and let $v^m = (\varphi(\boldsymbol{x}_i^m))_{i\in I^m} \in X_{\mathcal{D}_m,0}$, where $S_m = (\boldsymbol{x}_i^m)_{i\in I^m}$ is the family of approximation points of $\mathcal{D}_m$. Let $K \in \mathcal{M}_m$ and denote by $\pi_{K,m}$ the $\mathbb{P}_0$-exact function reconstruction associated to K for $\mathcal{D}_m$. The definition (7.32) of $\Pi_{\mathcal{D}_m}$, Lemma 7.27 and the definition of $\mathrm{reg}_{\mathrm{LLE}}(\mathcal{D}_m)$ give

$$\begin{aligned}
\|\Pi_{\mathcal{D}_m} v^m - \varphi\|_{L^p(K)} &= \left\|\pi_{K,m}[(v_i^m)_{i\in I_K^m}] - \varphi\right\|_{L^p(K)} \\
&\le \left(1 + \max_{i\in I_K^m} \frac{\mathrm{dist}(\boldsymbol{x}_i, K)}{\mathrm{diam}(K)}\right) \|\pi_{K,m}\|_p |K|^{\frac{1}{p}} \mathrm{diam}(K) \|\varphi\|_{W^{1,\infty}(\mathbb{R}^d)} \\
&\le (1 + \mathrm{reg}_{\mathrm{LLE}}(\mathcal{D}_m)) \mathrm{reg}_{\mathrm{LLE}}(\mathcal{D}_m) |K|^{\frac{1}{p}} h_{\mathcal{M}_m} \|\varphi\|_{W^{1,\infty}(\mathbb{R}^d)} .
\end{aligned}$$

Raise to the power p, sum over $K \in \mathcal{M}_m$ and take the power $1/p$ to obtain

$$\begin{aligned}
&\|\Pi_{\mathcal{D}_m} v^m - \varphi\|_{L^p(\Omega)} \\
&\qquad \le (1 + \mathrm{reg}_{\mathrm{LLE}}(\mathcal{D}_m)) \mathrm{reg}_{\mathrm{LLE}}(\mathcal{D}_m) |\Omega|^{\frac{1}{p}} h_{\mathcal{M}_m} \|\varphi\|_{W^{1,\infty}(\mathbb{R}^d)} . \qquad (7.35)
\end{aligned}$$

Let us now turn to the gradients. For $K \in \mathcal{M}_m$, let $\mathcal{G}_{K,m}$ be the $\mathbb{P}_1$-exact gradient reconstruction associated to K for $\mathcal{D}_m$. Owing to the definition (7.33) of $\nabla_{\mathcal{D}_m}$, to Lemma 7.31 and to the definition of $\mathrm{reg}_{\mathrm{LLE}}(\mathcal{D}_m)$,

$$\|\nabla_{\mathcal{D}_m} v^m - \nabla\varphi\|_{L^p(K)^d} = \left\|\mathcal{G}_{K,m}[(v_i^m)_{i\in I_K^m}] - \nabla\varphi\right\|_{L^p(K)^d}$$

$$\leq \left(1 + \frac{1}{2}\,\mathrm{reg}_{\mathrm{LLE}}(\mathcal{D}_m)[1 + \mathrm{reg}_{\mathrm{LLE}}(\mathcal{D}_m)]^2\right) |K|^{\frac{1}{p}} \mathrm{diam}(K)\, \|\varphi\|_{W^{2,\infty}(\mathbb{R}^d)}\,.$$

Again, raising to the power p, sum over $K \in \mathcal{M}_m$ and take the power $1/p$ to obtain

$$\begin{aligned}
&\|\nabla_{\mathcal{D}_m} v^m - \nabla\varphi\|_{L^p(\Omega)^d} \\
&\quad \leq \left(1 + \frac{1}{2}\,\mathrm{reg}_{\mathrm{LLE}}(\mathcal{D}_m)[1 + \mathrm{reg}_{\mathrm{LLE}}(\mathcal{D}_m)]^2\right) |\Omega|^{\frac{1}{p}} h_{\mathcal{M}_m}\, \|\varphi\|_{W^{2,\infty}(\mathbb{R}^d)}\,. \qquad (7.36)
\end{aligned}$$

Since $(\mathrm{reg}_{\mathrm{LLE}}(\mathcal{D}_m))_{m\in\mathbb{N}}$ is bounded and $S_{\mathcal{D}_m}(\varphi) \leq \|\Pi_{\mathcal{D}_m} v^m - \varphi\|_{L^p(\Omega)} + \|\nabla_{\mathcal{D}_m} v^m - \nabla\varphi\|_{L^p(\Omega)^d}$, we infer from (7.35) and (7.36) the existence of C_8 not depending on m or φ such that

$$S_{\mathcal{D}_m}(\varphi) \leq C_8 h_{\mathcal{M}_m}\, \|\varphi\|_{W^{2,\infty}(\mathbb{R}^d)}\,. \qquad (7.37)$$

Thus, $S_{\mathcal{D}_m}(\varphi) \to 0$ as $m \to \infty$ and the proof is complete. ∎

As demonstrated by (7.37), the notion of LLE GD provides order 1 approximation properties, which is expected when the local reconstructions are only linearly exact. A number of numerical schemes are based on local reconstructions which enjoy a stronger form of exactness: they are exact on interpolants of polynomial of higher order than 1. In this case, a better estimate than (7.37) is expected, as confirmed by the next proposition. As expected, $\mathbb{P}_\ell$ denotes the space of polynomials $\mathbb{R}^d \to \mathbb{R}$ of degree ℓ or less.

Proposition 7.38 (Estimate on $S_{\mathcal{D}}$ for higher order LLE GDs). *Let $k \in \mathbb{N} \setminus \{0\}$ and $\mathcal{D}$ be an LLE GD (in the sense of Definition 7.33), that is* of order k *in the following sense: for any cell K of the underlying mesh $\mathcal{M}$ of $\mathcal{D}$,*

1. *The family $\pi_K = (\pi_K^i)_{i\in I_K}$ is a $\mathbb{P}_{k-1}$-exact function reconstruction on K upon $(\boldsymbol{x}_i)_{i\in I_K}$, that is*

$$\forall q \in \mathbb{P}_{k-1}\,, \quad \textit{for a.e. } \boldsymbol{x} \in K\,, \ \sum_{i\in I_K} q(\boldsymbol{x}_i)\pi_K^i(\boldsymbol{x}) = q(\boldsymbol{x}). \qquad (7.38)$$

2. *The family $\mathcal{G}_K = (\mathcal{G}_K^i)_{i\in I_K}$ is a $\mathbb{P}_k$-exact gradient reconstruction on K upon $(\boldsymbol{x}_i)_{i\in I_K}$, that is*

$$\forall q \in \mathbb{P}_{k}\,, \quad \textit{for a.e. } \boldsymbol{x} \in K\,, \ \sum_{i\in I_K} q(\boldsymbol{x}_i)\mathcal{G}_K^i(\boldsymbol{x}) = \nabla q(\boldsymbol{x}). \qquad (7.39)$$

Let $\theta \geq \mathrm{reg}_{\mathrm{LLE}}(\mathcal{D})$. Then there exists C_9, depending only on θ, k and Ω, such that

$$\forall \varphi \in W^{k+1,\infty}(\Omega) \cap W_0^{1,p}(\Omega)\,,\ S_{\mathcal{D}}(\varphi) \leq C_9 h_{\mathcal{M}}^k\, \|\varphi\|_{W^{k+1,\infty}(\Omega)}\,. \qquad (7.40)$$

Proof. Let $v = (\varphi(\boldsymbol{x}_i))_{i\in I} \in X_{\mathcal{D},0}$ be the same interpolant as in the proof of Proposition 7.37, with $\mathcal{D}_m = \mathcal{D}$. Estimate (7.40) follows if we can prove that

$$\|\Pi_{\mathcal{D}} v - \varphi\|_{L^p(\Omega)} + \|\nabla_{\mathcal{D}} v - \nabla\varphi\|_{L^p(\Omega)^d} \le C_9 h_{\mathcal{M}}^k \|\varphi\|_{W^{k+1,\infty}(\Omega)}. \tag{7.41}$$

By standard extension and approximation techniques in Sobolev spaces, φ can be approximated in $C^k(\overline{\Omega})$ by functions in $C^{k+1}(\mathbb{R}^d)$, in such a way that their $W^{k+1,\infty}(\Omega)$-norms remain bounded by $C_{10}\|\varphi\|_{W^{k+1,\infty}(\Omega)}$ with C_{10} depending only on Ω and k. Proving (7.41) for $\varphi \in C^{k+1}(\overline{\Omega})$ is therefore sufficient.
If $K \in \mathcal{M}$, a Taylor expansion of order $\ell \le k$ about a given point in K gives $P_{\ell,K} \in \mathbb{P}_\ell$ such that, for all $\boldsymbol{y} \in \mathbb{R}^d$,

$$|P_{\ell,i}(\boldsymbol{y}) - \varphi(\boldsymbol{y})| \le C_{11}\left(\mathrm{dist}(\boldsymbol{y},K) + \mathrm{diam}(K)\right)^{\ell+1} \|\varphi\|_{W^{k+1,\infty}(\Omega)}, \tag{7.42}$$

$$|\nabla P_{\ell,i}(\boldsymbol{y}) - \nabla\varphi(\boldsymbol{y})| \le C_{11}\left(\mathrm{dist}(\boldsymbol{y},K) + \mathrm{diam}(K)\right)^{\ell} \|\varphi\|_{W^{k+1,\infty}(\Omega)}, \tag{7.43}$$

with C_{11} depending only on k and Ω. For any $\boldsymbol{x} \in K$, the $\mathbb{P}_{k-1}$-exactness (7.38) of π_K gives $\pi_K[(P_{k-1,K}(\boldsymbol{x}_i))_{i\in I_K}](\boldsymbol{x}) = P_{k-1,K}(\boldsymbol{x})$ and therefore

$$\pi_K v(\boldsymbol{x}) - \varphi(\boldsymbol{x}) = \pi_K v(\boldsymbol{x}) - \pi_K[(P_{k-1,K}(\boldsymbol{x}_i))_{i\in I_K}](\boldsymbol{x}) + P_{k-1,K}(\boldsymbol{x}) - \varphi(\boldsymbol{x}).$$

Using the bound (7.42) (with $\boldsymbol{y} = \boldsymbol{x}$ and $\boldsymbol{y} = \boldsymbol{x}_i$), and the fact that $\theta \ge \mathrm{reg}_{\mathrm{LLE}}(\mathcal{D}) \ge \mathrm{dist}(\boldsymbol{x}_i, K)/\mathrm{diam}(K)$ for all $i \in I_K$, we infer

$$\begin{aligned}
|\pi_K v(\boldsymbol{x}) - \varphi(\boldsymbol{x})| &\le \left|\sum_{i\in I_K} \pi_K^i(\boldsymbol{x})(v_i - P_{k-1,K}(\boldsymbol{x}_i))\right| + |P_{k-1,K}(\boldsymbol{x}) - \varphi(\boldsymbol{x})| \\
&\le C_{11}(\theta+1)^k \mathrm{diam}(K)^k \|\varphi\|_{W^{k+1,\infty}(\Omega)} \sum_{i\in I_K} |\pi_K^i(\boldsymbol{x})| \\
&\quad + C_{11}\mathrm{diam}(K)^k \|\varphi\|_{W^{k+1,\infty}(\Omega)}.
\end{aligned}$$

Taking the $L^p(K)$ norm of this inequality and recalling the definition (7.21) of $\|\pi_K\|_p$ yields C_{12}, depending only on θ, k and Ω, such that

$$\|\Pi_{\mathcal{D}} v - \varphi\|_{L^p(K)} \le C_{12} h_{\mathcal{M}}^k |K|^{\frac{1}{p}} \|\varphi\|_{W^{k+1,\infty}(\Omega)}. \tag{7.44}$$

Similarly, by (7.39), $\mathcal{G}_K[(P_{k,K}(\boldsymbol{x}_i))_{i\in I_K}](\boldsymbol{x}) = \nabla P_{k,K}(\boldsymbol{x})$ and thus

$$\mathcal{G}_K v(\boldsymbol{x}) - \nabla\varphi(\boldsymbol{x}) = \mathcal{G}_K v(\boldsymbol{x}) - \mathcal{G}_K[(P_{k,K}(\boldsymbol{x}_i))_{i\in I_K}](\boldsymbol{x}) + \nabla P_{k,K}(\boldsymbol{x}) - \nabla\varphi(\boldsymbol{x}).$$

Using (7.42) and (7.43) with $\ell = k$, this shows that

$$\begin{aligned}
&|\mathcal{G}_K v(\boldsymbol{x}) - \nabla\varphi(\boldsymbol{x})| \\
&\quad\le \left|\sum_{i\in I_K} \mathcal{G}_K^i(\boldsymbol{x})(v_i - P_{k,K}(\boldsymbol{x}_i))\right| + |\nabla P_{k,K}(\boldsymbol{x}) - \nabla\varphi(\boldsymbol{x})| \\
&\quad\le C_{11}(\theta+1)^{k+1}\mathrm{diam}(K)^k \|\varphi\|_{W^{k+1,\infty}(\Omega)} \mathrm{diam}(K) \sum_{i\in I_K} |\mathcal{G}_K^i(\boldsymbol{x})|
\end{aligned}$$

$$+ C_{11}\mathrm{diam}(K)^k \left\|\varphi\right\|_{W^{k+1,\infty}(\Omega)} .$$

Taking the $L^p(K)$-norm of this inequality and recalling the definition (7.24) of $\left\|\mathcal{G}_K\right\|_p$ yields

$$\left\|\nabla_{\mathcal{D}} v - \nabla\varphi\right\|_{L^p(K)^d} \leq C_{13} h_{\mathcal{M}}^k |K|^{\frac{1}{p}} \left\|\varphi\right\|_{W^{k+1,\infty}(\Omega)} , \tag{7.45}$$

where C_{13} depends only on θ, k and Ω. Estimate (7.41) follows by raising (7.44) and (7.45) to the power p, summing over $K \in \mathcal{M}$, taking the power $1/p$, and writing $|\Omega|^{\frac{1}{p}} \leq |\Omega| + 1$. ∎

Remark 7.39 (Estimate (7.40))
This estimate is particularly useful for problems in which $S_{\mathcal{D}}$ participates in the error estimates established for the GDM (cf. Theorems 2.28 and 2.38 for example). At the cost of increased technicality, Estimate (7.40) can often be established if φ merely belongs to $W^{k+1,p}(\Omega) \cap W_0^{1,p}(\Omega)$, instead of $W^{k+1,\infty}(\Omega) \cap W_0^{1,p}(\Omega)$. We refer to Section 7.4 for the case $k = 1$ and $p > d/2$.

7.3.3 From local to global basis functions, and matrix assembly

Let $\mathcal{D} = (X_{\mathcal{D},0}, \Pi_{\mathcal{D}}, \nabla_{\mathcal{D}})$ be an LLE GD in the sense of Definition 7.33. The functions $(\pi_K^i)_{K\in\mathcal{M}, i\in I_K}$ and $(\mathcal{G}_K^i)_{K\in\mathcal{M}, i\in I_K}$ can be seen as elementary basis functions, from which global basis functions can be constructed. Each of these global basis functions is associated with one discrete unknown of the GD in the following way. For $i \in I_\Omega$, define $\pi^i \in L^p(\Omega)$ and $\mathcal{G}^i \in L^p(\Omega)^d$ by:

$$\begin{aligned}
&\forall K \in \mathcal{M} \text{ such that } i \in I_K\,,\ (\pi^i)_{|K} = \pi_K^i \text{ and } (\mathcal{G}^i)_{|K} = \mathcal{G}_K^i\,,\\
&\forall K \in \mathcal{M} \text{ such that } i \notin I_K\,,\ (\pi^i)_{|K} = 0 \text{ and } (\mathcal{G}^i)_{|K} = 0.
\end{aligned}$$

Let $(v^{(i)})_{i\in I_\Omega}$ be the canonical basis of $X_{\mathcal{D},0}$, that is: for $i \in I_\Omega$, $v_i^{(i)} = 1$ and $v_j^{(i)} = 0$ for all $j \in I \setminus \{i\}$. It can be checked that, for any $i \in I_\Omega$,

$$\pi^i = \Pi_{\mathcal{D}} v^{(i)} \quad \text{and} \quad \mathcal{G}^i = \nabla_{\mathcal{D}} v^{(i)} \quad \text{on } \Omega.$$

From the definition of $\mathrm{reg}_{\mathrm{LLE}}(\mathcal{D})$ it is expected that, for each $i \in I_\Omega$, the cells $K \in \mathcal{M}$ such that $i \in I_K$ are close to $\boldsymbol{x}_i$. Hence, the global basis functions π^i and $\mathcal{G}^i$ have their support in a neighbourhood of $\boldsymbol{x}_i$, associated with the discrete unknown v_i of a generic $v \in X_{\mathcal{D},0}$.

This construction is illustrated in Figure 7.3, for the special case of basis functions from the $\mathbb{P}_1$ finite element method (for which $I = \mathcal{V}$ and $I_K = \mathcal{V}_K$). As can be seen, the elementary basis function $\pi_K^{\boldsymbol{s}}$ is only defined in K, whereas the global basis function $\pi^{\boldsymbol{s}}$ is defined over all of Ω, and is zero on the cells K' that do not have $\boldsymbol{s}$ as a vertex (i.e., $\boldsymbol{s} \notin \mathcal{V}_{K'}$).

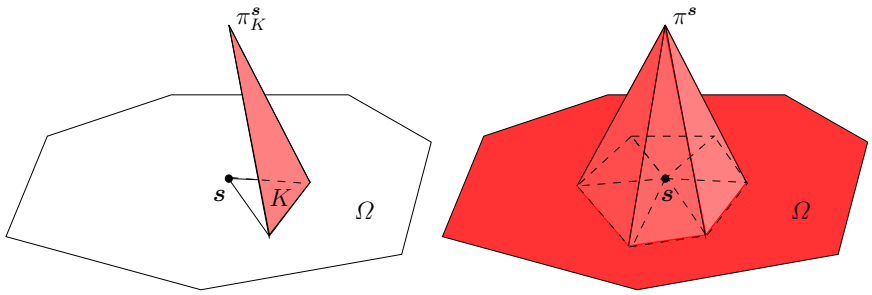


Fig. 7.3. Elementary basis function (left) and global basis functions (right) for $\mathbb{P}_1$ finite elements

Let us now consider the problem (2.22) and its GS approximation (2.23). As seen in Section 2.1.2, this scheme can be re-cast as a linear system $AU = B$ with

$$\begin{aligned}
u &= \sum_{i \in I_\Omega} U_i v^{(i)}, \\
A_{ij} &= \int_\Omega \Lambda(\boldsymbol{x}) \nabla_{\mathcal{D}} v^{(j)}(\boldsymbol{x}) \cdot \nabla_{\mathcal{D}} v^{(i)}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}, \\
B_i &= \int_\Omega f(\boldsymbol{x}) \Pi_{\mathcal{D}} v^{(i)}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} - \int_\Omega \boldsymbol{F}(\boldsymbol{x}) \cdot \nabla_{\mathcal{D}} v^{(i)}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}.
\end{aligned}$$

As in finite element methods, for an LLE GD the matrix A and vector B can be assembled by local computations. Define the elementary matrices and vectors by

$$\begin{aligned}
&\forall i \in I_K,\ \forall j \in I_K, \\
&A_{ij}^K = \int_K \Lambda(\boldsymbol{x}) \mathcal{G}_K^j(\boldsymbol{x}) \cdot \mathcal{G}_K^i(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}, \\
&B_i^K = \int_K f(\boldsymbol{x}) \pi_K^i(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} - \int_K \boldsymbol{F}(\boldsymbol{x}) \cdot \mathcal{G}_K^i(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}.
\end{aligned}$$

Then the global matrix and vectors are assembled by the following operations:

$$A_{ij} = \sum_{K \in \mathcal{M} \text{ s.t. } i,j \in I_K} A_{ij}^K \quad \text{and} \quad B_i = \sum_{K \in \mathcal{M} \text{ s.t. } i \in I_K} B_i^K.$$

7.3.4 Barycentric condensation

The construction of a numerical scheme often requires several interpolation points, the approximation points S of an LLE GD, corresponding to as many discrete unknowns of the scheme. The higher the number of discrete unknowns, the larger the matrix will be and, very likely, the more expensive the scheme

is. A classical way to reduce the computational cost of a scheme is to eliminate some of these unknowns through barycentric combinations. This consists in replacing (and thus, eliminating) certain unknowns by averages of other unknowns.

We describe here a way to perform this reduction in the general context of LLE GDs, while preserving the required properties (coercivity, consistency, limit-conformity and compactness). In the following definition, a subset I^{BA} is selected from the geometrical entities I associated with the discrete unknowns of an LLE GD, and all other discrete unknowns (associated with $I \backslash I^{\text{BA}}$) are eliminated from the GD space by being expressed as local barycentric combinations of the discrete unknowns corresponding to I^{BA}.

Definition 7.40 (Barycentric condensation of an LLE GD). *Let $\mathcal{D}$ be an LLE GD in the sense of Definition 7.33, $S = (\boldsymbol{x}_i)_{i \in I} \subset \mathbb{R}^d$ be its family of approximation points, and $\mathcal{M}$ be its mesh. A gradient discretisation $\mathcal{D}^{\text{BA}}$ is a barycentric condensation of $\mathcal{D}$ if there exists a strict subset $I^{\text{BA}} \subset I$ and, for all $i \in I \backslash I^{\text{BA}}$, a set $H_i \subset I^{\text{BA}}$ and real numbers $(\beta^i_j)_{j \in H_i}$ satisfying*

$$\sum_{j \in H_i} \beta^i_j = 1 \quad \text{and} \quad \sum_{j \in H_i} \beta^i_j \boldsymbol{x}_j = \boldsymbol{x}_i, \tag{7.46}$$

such that

- $I_\partial \subset I^{\text{BA}}$,
- $X_{\mathcal{D}^{\text{BA}},0} = \{u = (u_i)_{i \in I^{\text{BA}}} \,:\, u_i \in \mathbb{R} \text{ for all } i \in I^{\text{BA}},\ u_i = 0 \text{ for all } i \in I_\partial\}$,
- *The function and gradient reconstructions $\Pi_{\mathcal{D}^{\text{BA}}}$ and $\nabla_{\mathcal{D}^{\text{BA}}}$ are given by:*

$$\forall v \in X_{\mathcal{D}^{\text{BA}},0}\,, \quad \Pi_{\mathcal{D}^{\text{BA}}} v = \Pi_{\mathcal{D}} \widetilde{v} \quad \text{and} \quad \nabla_{\mathcal{D}^{\text{BA}}} v = \nabla_{\mathcal{D}} \widetilde{v},$$

where $\widetilde{v} \in X_{\mathcal{D},0}$ is defined by

$$\forall i \in I\,,\ \widetilde{v}_i = \begin{cases} v_i & \text{if } i \in I^{\text{BA}}, \\ \displaystyle\sum_{j \in H_i} \beta^i_j v_j & \text{if } i \in I \setminus I^{\text{BA}}. \end{cases} \tag{7.47}$$

(recall that $X_{\mathcal{D},0} = \{u = (u_i)_{i \in I} \,:\, u_i \in \mathbb{R} \text{ for all } i \in I\,,\ u_i = 0 \text{ for all } i \in I_\partial\}$, and notice that $\widetilde{v}$ indeed belongs to this space since $I_\partial \subset I^{\text{BA}}$ and $v_i = 0$ if $i \in I_\partial$.)

The regularity of the barycentric condensation $\mathcal{D}^{\text{BA}}$ is

$$\operatorname{reg}_{\text{BA}}(\mathcal{D}^{\text{BA}}) = \max_{i \in I \backslash I^{\text{BA}}} \left(\sum_{j \in H_i} |\beta^i_j| + \max_{K \in \mathcal{M} \,|\, i \in I_K} \max_{j \in H_i} \frac{\operatorname{dist}(\boldsymbol{x}_j, \boldsymbol{x}_i)}{\operatorname{diam}(K)} \right).$$

It is clear that $\mathcal{D}^{\text{BA}}$ defined above is a GD. Indeed, if $\nabla_{\mathcal{D}^{\text{BA}}} v = 0$ on Ω then $\nabla_{\mathcal{D}} \widetilde{v} = 0$ on Ω and thus $\widetilde{v}_i = 0$ for all $i \in I$, since $\mathcal{D}$ is a GD and $\|\nabla_{\mathcal{D}} \cdot\|_{L^p(\Omega)^d}$

is a norm on $X_{\mathcal{D},0}$. This shows that $v_i = 0$ for all $i \in I^{\text{BA}}$, and thus that $\|\nabla_{\mathcal{D}^{\text{BA}}} \cdot\|_{L^p(\Omega)^d}$ is a norm on $X_{\mathcal{D}^{\text{BA}},0}$.

Note that $\text{reg}_{\text{BA}}(\mathcal{D}^{\text{BA}})$ is always greater than or equal to 1 (take $i \in I \setminus I^{\text{BA}}$ and write $1 = \sum_{j \in H_i} \beta_j^i \leq \sum_{j \in H_i} |\beta_j^i|$). Let us, for a brief moment, confuse a discrete unknown with the geometrical entity $i \in I$ it is attached to, and with the interpolation point $\boldsymbol{x}_i$ it corresponds to ($\boldsymbol{x}_i$ usually lies on or close to i). Bounding the last term in $\text{reg}_{\text{BA}}(\mathcal{D}^{\text{BA}})$ consists in requiring that, if $i \in I \setminus I^{\text{BA}}$ is involved in the definition (for $\mathcal{D}$) of π_K or $\mathcal{G}_K$ for some $K \in \mathcal{M}$, then i lies within distance $\mathcal{O}(\text{diam}(K))$ of any $j \in H_i$ used to eliminate i. This ensures that, after barycentric condensation, π_K and $\mathcal{G}_K$ are still computed using only discrete unknowns in a neighbourhood of K.

"Barycentric condensation" refers both to the operation consisting in replacing some unknowns with combination of others, and to the resulting gradient discretisation (called, as mentioned above, a barycentric condensation of the initial GD). The combinations performed during a barycentric condensation are linearly exact thanks to (7.46). The LLE property is therefore preserved in the process, as formally stated in the lemma below.

Remark 7.41 (Barycentric condensation vs. static condensation)
A barycentric condensation is not quite the same as a static condensation. A static condensation consists, *after having written a linear scheme*, in expressing some of the unknowns in terms of others *and of the source terms*. Examples of static condensations are given in Remarks 8.19 and 13.9.
A barycentric condensation occurs before a scheme is even written, and can also be performed for non-linear schemes; the replacement of discrete unknowns in a barycentric condensation modifies the space and operators of the scheme independently of the model to which it is applied.

Lemma 7.42 (Barycentric condensation preserves the LLE property). *Let $\mathcal{D}$ be an LLE GD in the sense of Definition 7.33, and let $\mathcal{D}^{\text{BA}}$ be a barycentric condensation of $\mathcal{D}$ in the sense of Definition 7.40. Then $\mathcal{D}^{\text{BA}}$ is an LLE GD on the same mesh as $\mathcal{D}$, and*

$$\text{reg}_{\text{LLE}}(\mathcal{D}^{\text{BA}}) \leq \text{reg}_{\text{BA}}(\mathcal{D}^{\text{BA}})\, \text{reg}_{\text{LLE}}(\mathcal{D}) + \text{reg}_{\text{BA}}(\mathcal{D}^{\text{BA}}). \tag{7.48}$$

Proof. Let $\mathcal{M}$ be the mesh corresponding to $\mathcal{D}$, and let $K \in \mathcal{M}$. Take $v \in X_{\mathcal{D}^{\text{BA}},0}$ and let $\widetilde{v} \in X_{\mathcal{D},0}$ be defined by (7.47). For any $K \in \mathcal{M}$, the values $(\widetilde{v}_i)_{i \in I_K}$ are computed as linear combinations of $(v_i)_{i \in I_K^{\text{BA}}}$, with

$$I_K^{\text{BA}} = (I_K \cap I^{\text{BA}}) \cup \bigcup_{i \in I_K \setminus I^{\text{BA}}} H_i. \tag{7.49}$$

By (7.32) in Definition 7.33 and the definition (7.47) of $\widetilde{v}$, for $K \in \mathcal{M}$ and a.e. $\boldsymbol{x} \in K$,

$$\Pi_{\mathcal{D}^{\mathrm{BA}}} v(\boldsymbol{x}) = \Pi_{\mathcal{D}} \widetilde{v}(\boldsymbol{x}) = \sum_{i \in I_K} \widetilde{v}_i \pi_K^i(\boldsymbol{x})$$
$$= \sum_{i \in I_K \cap I^{\mathrm{BA}}} v_i \pi_K^i(\boldsymbol{x}) + \sum_{i \in I_K \setminus I^{\mathrm{BA}}} \Big(\sum_{j \in H_i} \beta_j^i v_j \Big) \pi_K^i(\boldsymbol{x}) = \sum_{j \in I_K^{\mathrm{BA}}} v_j \widetilde{\pi}_K^j(\boldsymbol{x}), \quad (7.50)$$

where, for $j \in I_K^{\mathrm{BA}}$, the function $\widetilde{\pi}_K^j \in L^p(K)$ is defined by

$$\widetilde{\pi}_K^j = \begin{cases} \pi_K^j + \displaystyle\sum_{i \in I_K \setminus I^{\mathrm{BA}} \,|\, j \in H_i} \beta_j^i \pi_K^i & \text{if } j \in I_K \cap I^{\mathrm{BA}}, \\ \displaystyle\sum_{i \in I_K \setminus I^{\mathrm{BA}} \,|\, j \in H_i} \beta_j^i \pi_K^i & \text{if } j \notin I_K \cap I^{\mathrm{BA}}. \end{cases}$$

Using (7.46) and (7.20) yields, for a.e. $\boldsymbol{x} \in K$,

$$\begin{aligned} \sum_{j \in I_K^{\mathrm{BA}}} \widetilde{\pi}_K^j(\boldsymbol{x}) &= \sum_{j \in I_K \cap I^{\mathrm{BA}}} \pi_K^j(\boldsymbol{x}) + \sum_{j \in I_K^{\mathrm{BA}}} \sum_{i \in I_K \setminus I^{\mathrm{BA}} \,|\, j \in H_i} \beta_j^i \pi_K^i(\boldsymbol{x}) \\ &= \sum_{j \in I_K \cap I^{\mathrm{BA}}} \pi_K^j(\boldsymbol{x}) + \sum_{i \in I_K \setminus I^{\mathrm{BA}}} \pi_K^i(\boldsymbol{x}) \sum_{j \in H_i} \beta_j^i \qquad (7.51) \\ &= \sum_{i \in I_K \cap I^{\mathrm{BA}}} \pi_K^i(\boldsymbol{x}) + \sum_{i \in I_K \setminus I^{\mathrm{BA}}} \pi_K^i(\boldsymbol{x}) \\ &= \sum_{i \in I_K} \pi_K^i(\boldsymbol{x}) = 1. \end{aligned}$$

In the first term of the penultimate line, we simply performed the change of index $j \mapsto i$. The family $(\widetilde{\pi}_K^j)_{j \in I_K^{\mathrm{BA}}}$ is therefore a $\mathbb{P}_0$-exact function reconstruction and, by (7.50), $\Pi_{\mathcal{D}^{\mathrm{BA}}} v$ has the required form (7.32).
In a similar way as above, write

$$\nabla_{\mathcal{D}^{\mathrm{BA}}} v(\boldsymbol{x}) = \nabla_{\mathcal{D}} \widetilde{v}(\boldsymbol{x}) = \sum_{i \in I_K} \widetilde{v}_i \mathcal{G}_K^i(\boldsymbol{x})$$
$$= \sum_{i \in I_K \cap I^{\mathrm{BA}}} v_i \mathcal{G}_K^i(\boldsymbol{x}) + \sum_{i \in I_K \setminus I^{\mathrm{BA}}} \left(\sum_{j \in H_i} \beta_j^i v_j \right) \mathcal{G}_K^i(\boldsymbol{x}) = \sum_{j \in I_K^{\mathrm{BA}}} v_j \widetilde{\mathcal{G}}_K^j(\boldsymbol{x}),$$

where the function $\widetilde{\mathcal{G}}_K^j \in L^p(K)^d$ is defined by

$$\widetilde{\mathcal{G}}_K^j = \begin{cases} \mathcal{G}_K^j + \displaystyle\sum_{i \in I_K \setminus I^{\mathrm{BA}} \,|\, j \in H_i} \beta_j^i \mathcal{G}_K^i & \text{if } j \in I_K \cap I^{\mathrm{BA}}, \\ \displaystyle\sum_{i \in I_K \setminus I^{\mathrm{BA}} \,|\, j \in H_i} \beta_j^i \mathcal{G}_K^i & \text{if } j \notin I_K \cap I^{\mathrm{BA}}. \end{cases}$$

Let A be an affine map. Reproduce similar computations as above for $\widetilde{\pi}_K^j$ and write

$$\begin{aligned}
&\sum_{j\in I_K^{\mathrm{BA}}} A(\boldsymbol{x}_j)\widetilde{\mathcal{G}}_K^j(\boldsymbol{x}) \\
&\quad= \sum_{j\in I_K\cap I^{\mathrm{BA}}} A(\boldsymbol{x}_j)\mathcal{G}_K^j(\boldsymbol{x}) + \sum_{j\in I_K^{\mathrm{BA}}} A(\boldsymbol{x}_j) \sum_{i\in I_K\setminus I^{\mathrm{BA}}\,|\,j\in H_i} \beta_j^i \mathcal{G}_K^i(\boldsymbol{x}) \\
&\quad= \sum_{j\in I_K\cap I^{\mathrm{BA}}} A(\boldsymbol{x}_j)\mathcal{G}_K^j(\boldsymbol{x}) + \sum_{i\in I_K\setminus I^{\mathrm{BA}}} \left(\sum_{j\in H_i} A(\boldsymbol{x}_j)\beta_j^i\right) \mathcal{G}_K^i(\boldsymbol{x}). \qquad (7.52)
\end{aligned}$$

Since A is affine we have $A(\boldsymbol{x}) = A(\boldsymbol{x}_i) + \nabla A\cdot(\boldsymbol{x}-\boldsymbol{x}_i)$. Hence, (7.46) yields

$$\sum_{j\in H_i} \beta_j^i A(\boldsymbol{x}_j) = \sum_{j\in H_i} \beta_j^i A(\boldsymbol{x}_i) + \nabla A\cdot\left(\sum_{j\in H_i}\beta_j^i\boldsymbol{x}_j - \boldsymbol{x}_i\right) = A(\boldsymbol{x}_i).$$

Plugged into (7.52) and using (7.23), this gives

$$\begin{aligned}
\sum_{j\in I_K^{\mathrm{BA}}} A(\boldsymbol{x}_j)\widetilde{\mathcal{G}}_K^j(\boldsymbol{x}) &= \sum_{i\in I_K\cap I^{\mathrm{BA}}} A(\boldsymbol{x}_i)\mathcal{G}_K^i(\boldsymbol{x}) + \sum_{i\in I_K\setminus I^{\mathrm{BA}}} A(\boldsymbol{x}_i)\mathcal{G}_K^i(\boldsymbol{x}) \\
&= \sum_{i\in I_K} A(\boldsymbol{x}_i)\mathcal{G}_K^i(\boldsymbol{x}) = \nabla A.
\end{aligned}$$

The family $(\widetilde{\mathcal{G}}_K^j)_{j\in I_K^{\mathrm{BA}}}$ is therefore a $\mathbb{P}_1$-exact gradient reconstruction, and $\nabla_{\mathcal{D}^{\mathrm{BA}}} v$ has the required form (7.33). This completes the proof that $\mathcal{D}^{\mathrm{BA}}$ is an LLE GD.

Let us now establish the upper bound on $\mathrm{reg}_{\mathrm{LLE}}(\mathcal{D}^{\mathrm{BA}})$. Reproducing the reasoning that leads to (7.51) but using absolute values and inequalities, we see that for any $K\in\mathcal{M}$ and a.e. $\boldsymbol{x}\in K$

$$\begin{aligned}
\sum_{j\in I_K^{\mathrm{BA}}} |\widetilde{\pi}_K^j(\boldsymbol{x})| &\le \sum_{i\in I_K\cap I^{\mathrm{BA}}} |\pi_K^i(\boldsymbol{x})| + \sum_{i\in I_K\setminus I^{\mathrm{BA}}} |\pi_K^i(\boldsymbol{x})| \sum_{j\in H_i} |\beta_j^i| \\
&\le \mathrm{reg}_{\mathrm{BA}}(\mathcal{D}^{\mathrm{BA}}) \sum_{i\in I_K} |\pi_K^i(\boldsymbol{x})|. \qquad (7.53)
\end{aligned}$$

Take the $L^p(K)$ norm, multiply by $|K|^{-\frac{1}{p}}$ and recall the definition (7.21) of the norm of $\mathbb{P}_0$-exact function reconstructions to obtain

$$\|\widetilde{\pi}_K\|_p \le \mathrm{reg}_{\mathrm{BA}}(\mathcal{D}^{\mathrm{BA}})\,\|\pi_K\|_p\,. \qquad (7.54)$$

The estimate on the gradient reconstructions is similar. Using the definition of $(\widetilde{\mathcal{G}}_K^j)_{j\in I_K^{\mathrm{BA}}}$, we see that (7.53) still holds with "$\mathcal{G}$" instead of "π" so that, taking the $L^p(K)$ norm and multiplying by $\mathrm{diam}(K)|K|^{-\frac{1}{p}}$,

$$\left\|\widetilde{\mathcal{G}}_K\right\|_p \le \mathrm{reg}_{\mathrm{BA}}(\mathcal{D}^{\mathrm{BA}})\,\|\mathcal{G}_K\|_p\,. \qquad (7.55)$$

Finally, for $j \in I_K^{\text{BA}}$ we estimate $\frac{\text{dist}(\boldsymbol{x}_j,K)}{\text{diam}(K)}$ by assuming first that $j \notin I_K$. Then, there exists $\ell \in I_K \backslash I^{\text{BA}}$ such that $j \in H_\ell$, and thus $\frac{\text{dist}(\boldsymbol{x}_j,\boldsymbol{x}_\ell)}{\text{diam}(K)} \leq \text{reg}_{\text{BA}}(\mathcal{D}^{\text{BA}})$. This gives

$$\begin{aligned}\frac{\text{dist}(\boldsymbol{x}_j,K)}{\text{diam}(K)} &\leq \frac{\text{dist}(\boldsymbol{x}_j,\boldsymbol{x}_\ell)}{\text{diam}(K)} + \frac{\text{dist}(\boldsymbol{x}_\ell,K)}{\text{diam}(K)} \\ &\leq \text{reg}_{\text{BA}}(\mathcal{D}^{\text{BA}}) + \max_{i\in I_K} \frac{\text{dist}(\boldsymbol{x}_i,K)}{\text{diam}(K)}. \end{aligned} \tag{7.56}$$

This last inequality obviously also holds if $j \in I_K$. The proof of (7.48) is completed by recalling the definition (7.34) of reg_{LLE}, by combining (7.54), (7.55) and (7.56), and by using $\text{reg}_{\text{BA}}(\mathcal{D}^{\text{BA}}) \geq 1$. ∎

The following theorem shows that barycentric condensations of sequences of LLE GDs satisfy the same properties (coercivity, GD-consistency, compactness, limit-conformity) as the original sequence of GDs. The GD-consistency is a consequence of Lemma 7.42 and Proposition 7.37, and the other three properties result from the fact that $X_{\mathcal{D}^{\text{BA}},0}$ is (roughly) a subspace of $X_{\mathcal{D},0}$.

Theorem 7.43 (Properties of barycentric condensations of GDs). *Let $(\mathcal{D}_m)_{m\in\mathbb{N}}$ be a sequence of LLE GDs in the sense of Definition 7.33, that is coercive, GD-consistent, limit-conforming and compact in the sense of Definitions 2.2, 2.4, 2.5 and 2.8. Let $\mathcal{M}_m$ be the mesh associated with $\mathcal{D}_m$. We assume that $h_{\mathcal{M}_m} \to 0$ as $m \to \infty$, and that $(\text{reg}_{\text{LLE}}(\mathcal{D}_m))_{m\in\mathbb{N}}$ is bounded. For any $m \in \mathbb{N}$ we take a barycentric condensation $\mathcal{D}_m^{\text{BA}}$ of $\mathcal{D}_m$ in the sense of Definition 7.40, such that $(\text{reg}_{\text{BA}}(\mathcal{D}_m^{\text{BA}}))_{m\in\mathbb{N}}$ is bounded.*
Then $(\mathcal{D}_m^{\text{BA}})_{m\in\mathbb{N}}$ is also coercive, GD-consistent, limit-conforming, and compact. Moreover, we have

$$C_{\mathcal{D}_m^{\text{BA}}} \leq C_{\mathcal{D}_m} \quad \text{and} \quad W_{\mathcal{D}_m^{\text{BA}}} \leq W_{\mathcal{D}_m}. \tag{7.57}$$

Remark 7.44. Each of the property is transferred to the barycentric condensation independently of the others. This means, for example, that we only need to know that $(\mathcal{D}_m)_{m\in\mathbb{N}}$ is coercive to deduce that $(\mathcal{D}_m^{\text{BA}})_{m\in\mathbb{N}}$ is also coercive.

Proof. For any $v \in X_{\mathcal{D}_m^{\text{BA}},0}$, with $\widetilde{v}$ defined by (7.47) we have

$$\begin{aligned}\left\|\Pi_{\mathcal{D}_m^{\text{BA}}} v\right\|_{L^p(\Omega)} &= \left\|\Pi_{\mathcal{D}_m} \widetilde{v}\right\|_{L^p(\Omega)} \\ &\leq C_{\mathcal{D}_m} \left\|\nabla_{\mathcal{D}_m} \widetilde{v}\right\|_{L^p(\Omega)^d} = C_{\mathcal{D}_m} \left\|\nabla_{\mathcal{D}_m^{\text{BA}}} v\right\|_{L^p(\Omega)^d}.\end{aligned}$$

This shows that $C_{\mathcal{D}_m^{\text{BA}}} \leq C_{\mathcal{D}_m}$ and thus that $(\mathcal{D}_m^{\text{BA}})_{m\in\mathbb{N}}$ is coercive if $(\mathcal{D}_m)_{m\in\mathbb{N}}$ is coercive.
To prove the compactness, we take $(\nabla_{\mathcal{D}_m^{\text{BA}}} v_m)_{m\in\mathbb{N}} = (\nabla_{\mathcal{D}_m} \widetilde{v_m})_{m\in\mathbb{N}}$ bounded in $L^p(\Omega)^d$, and we use the compactness of $(\mathcal{D}_m)_{m\in\mathbb{N}}$ to see that $(\Pi_{\mathcal{D}_m} \widetilde{v_m})_{m\in\mathbb{N}} = (\Pi_{\mathcal{D}_m^{\text{BA}}} v_m)_{m\in\mathbb{N}}$ is relatively compact in $L^p(\Omega)$.

The limit-conformity follows by writing, for $v \in X_{\mathcal{D}_m^{\mathrm{BA}},0}$ and $\boldsymbol{\varphi} \in W_{\mathrm{div}}^{p'}(\Omega)$,

$$\frac{1}{\left\|\nabla_{\mathcal{D}_m^{\mathrm{BA}}} v\right\|_{L^p(\Omega)^d}} \left| \int_\Omega \left(\nabla_{\mathcal{D}_m^{\mathrm{BA}}} v(\boldsymbol{x}) \cdot \boldsymbol{\varphi}(\boldsymbol{x}) + \Pi_{\mathcal{D}_m^{\mathrm{BA}}} v(\boldsymbol{x}) \mathrm{div} \boldsymbol{\varphi}(\boldsymbol{x}) \right) \mathrm{d}\boldsymbol{x} \right|$$
$$= \frac{1}{\left\|\nabla_{\mathcal{D}_m} \widetilde{v}\right\|_{L^p(\Omega)^d}} \left| \int_\Omega \left(\nabla_{\mathcal{D}_m} \widetilde{v}(\boldsymbol{x}) \cdot \boldsymbol{\varphi}(\boldsymbol{x}) + \Pi_{\mathcal{D}_m} \widetilde{v}(\boldsymbol{x}) \mathrm{div} \boldsymbol{\varphi}(\boldsymbol{x}) \right) \mathrm{d}\boldsymbol{x} \right|,$$

which shows that $W_{\mathcal{D}_m^{\mathrm{BA}}}(\boldsymbol{\varphi}) \leq W_{\mathcal{D}_m}(\boldsymbol{\varphi}) \to 0$ as $m \to \infty$.
Finally, by Lemma 7.42 each $\mathcal{D}_m^{\mathrm{BA}}$ is an LLE GD and $(\mathrm{reg}_{\mathrm{LLE}}(\mathcal{D}_m^{\mathrm{BA}}))_{m\in\mathbb{N}}$ is bounded, since $(\mathrm{reg}_{\mathrm{LLE}}(\mathcal{D}_m))_{m\in\mathbb{N}}$ and $(\mathrm{reg}_{\mathrm{BA}}(\mathcal{D}_m^{\mathrm{BA}}))_{m\in\mathbb{N}}$ are bounded. Proposition 7.37 then gives the GD-consistency of $(\mathcal{D}_m^{\mathrm{BA}})_{m\in\mathbb{N}}$. ∎

7.3.5 Mass lumping

"Mass-lumping" is the generic name of the process applied (usually on a case-by-case basis) to modify schemes that do not have a built-in piecewise constant reconstruction, say for instance the $\mathbb{P}_1$ finite element scheme (see Chapter 8 in Part III). In the GDM framework, a generic and rigorous way to perform mass-lumping can be described. It simply consists in modifying the reconstruction operator $\Pi_{\mathcal{D}}$ so that it becomes a piecewise constant reconstruction. Under an assumption easy to verify in practice, this "mass-lumped" GD can be compared with the original GD, which ensures that all properties are preserved.

Note that the notions and results in this section are not limited to LLE GDs, they apply to any kind of gradient discretisation.

Definition 7.45 (Mass-lumped GD). *Let $\mathcal{D} = (X_{\mathcal{D},0}, \Pi_{\mathcal{D}}, \nabla_{\mathcal{D}})$ be a GD in the sense of Definition 2.1. A mass-lumped version of $\mathcal{D}$ is a GD $\mathcal{D}^{\mathrm{ML}} = (X_{\mathcal{D},0}, \Pi_{\mathcal{D}}^{\mathrm{ML}}, \nabla_{\mathcal{D}})$ such that $\Pi_{\mathcal{D}}^{\mathrm{ML}}$ is a piecewise constant reconstruction in the sense of Definition 2.12.*

Example 7.46 (Mass-lumped non-conforming $\mathbb{P}_1$ gradient discretisation)
Consider the special case of an LLE GD $\mathcal{D}$, with I as set of geometrical entities attached to the discrete unknowns. Recalling the notations in Definition 2.12, mass-lumping $\mathcal{D}$ first requires to select disjoint subsets $(\Omega_i)_{i\in I}$ of Ω with each Ω_i lying "around" i. Then, a new function reconstruction $\Pi_{\mathcal{D}}^{\mathrm{ML}}$ is defined such that, if $v = (v_i)_{i\in I}$, for all $i \in I$ we have $\Pi_{\mathcal{D}}^{\mathrm{ML}} v = v_i$ on Ω_i. According to Theorem 7.49 below, this new reconstruction is a valid choice if the $(\Omega_i)_{i\in I}$ are such that $\Pi_{\mathcal{D}} v \approx v_i$ on Ω_i, for all $i \in I$.
For the non-conforming $\mathbb{P}_1$ finite element on a simplicial mesh $\mathfrak{T}$, since $I = \mathcal{F}$ we need to find, for each $\sigma \in \mathcal{F}$, a set Ω_σ that lies "around" σ and is disjoint from all the other sets $(\Omega_{\sigma'})_{\sigma'\neq\sigma}$. There are many possible

choice; one of them is presented in Figure 9.2 page 299, in which each Ω_σ is a diamond D_σ around σ. Then, $(\Pi_{\mathcal{D}}^{\mathrm{ML}} v)_{|D_\sigma} = v_\sigma$ for all $\sigma \in \mathcal{F}$.

Remark 7.47 (Mass-lumping with respect to a canonical basis preserves the LLE property). Let $\mathcal{D}$ be an LLE GD, with I as set of geometrical entities attached to the discrete unknowns. Let $\mathcal{D}^{\mathrm{ML}}$ be a mass-lumping of $\mathcal{D}$ with respect to I, that is, $\Pi_{\mathcal{D}}^{\mathrm{ML}}$ is a piecewise constant reconstruction in the sense of Definition 2.12 with $B = I$ and $(e_i)_j = \delta_{ij}$. Then $\mathcal{D}^{\mathrm{ML}}$ is also an LLE GD, and $\mathrm{reg}_{\mathrm{LLE}}(\mathcal{D}^{\mathrm{ML}}) \leq \mathrm{reg}_{\mathrm{LLE}}(\mathcal{D})$.

Remark 7.48 (Mass lumping with respect to a non-canonical basis)
The basis $(e_i)_{i\in B}$ of $X_{\mathcal{D},0}$ used in Definition 2.12 to perform a mass-lumping of $\mathcal{D}$ is usually a canonical basis, each vector in this basis corresponding to a natural discrete unknown of $\mathcal{D}$. Mass-lumping could be done with respect to a non-standard basis, but this might lead to additional numerical cost if the computation of $\nabla_{\mathcal{D}}$ in this non-standard basis is complex; the scheme implementation might require to perform changes of basis, possibly with full transition matrices, to compute $\Pi_{\mathcal{D}}^{\mathrm{ML}}$ and $\nabla_{\mathcal{D}}$.

Theorem 7.49 (Properties of mass-lumped GDs). *Let $(\mathcal{D}_m)_{m\in\mathbb{N}}$ be a sequence of GDs in the sense of Definition 2.1, that is coercive, GD-consistent, limit-conforming and compact in the sense of Definitions 2.2, 2.4, 2.5 and 2.8. For any $m \in \mathbb{N}$ we take $\mathcal{D}_m^{\mathrm{ML}}$ a mass-lumped version of $\mathcal{D}_m$. If there exists $(\omega_m)_{m\in\mathbb{N}}$ such that $\omega_m \to 0$ as $m \to \infty$ and*

$$\forall m \in \mathbb{N}\,,\ \forall v \in X_{\mathcal{D}_m,0}\,,\ \left\|\Pi_{\mathcal{D}_m}^{\mathrm{ML}} v - \Pi_{\mathcal{D}_m} v\right\|_{L^p(\Omega)} \leq \omega_m \left\|v\right\|_{\mathcal{D}_m}\,, \tag{7.58}$$

then $(\mathcal{D}_m^{\mathrm{ML}})_{m\in\mathbb{N}}$ is coercive, GD-consistent, limit-conforming, and compact. The reconstruction $\Pi_{\mathcal{D}_m}^{\mathrm{ML}}$ is also piecewise constant.

This theorem is a direct consequence of Theorem 7.50 below, which gives a general setting for proving the properties of a GD by comparing it with another GD.

Theorem 7.50 (Comparison of function reconstructions).
Let $(\mathcal{D}_m)_{m\in\mathbb{N}}$ be a sequence of GDs in the sense of Definition 2.1. For any $m \in \mathbb{N}$, let $\mathcal{D}_m^\star$ be a GD defined from $\mathcal{D}_m$ by $\mathcal{D}_m^\star = (X_{\mathcal{D}_m,0}, \Pi_{\mathcal{D}_m}^\star, \nabla_{\mathcal{D}_m})$, where $\Pi_{\mathcal{D}_m}^\star$ is a linear operator from $X_{\mathcal{D}_m,0}$ to $L^p(\Omega)$.

1. *We assume that there exists a sequence $(\omega_m)_{m\in\mathbb{N}}$ such that*

$$\begin{aligned}&\lim_{m\to\infty} \omega_m = 0 \text{ and, for all } m \in \mathbb{N} \text{ and all } v \in X_{\mathcal{D}_m,0},\\ &\left\|\Pi_{\mathcal{D}_m}^\star v - \Pi_{\mathcal{D}_m} v\right\|_{L^p(\Omega)} \leq \omega_m \left\|v\right\|_{\mathcal{D}_m}.\end{aligned} \tag{7.59}$$

If $(\mathcal{D}_m)_{m\in\mathbb{N}}$ is coercive (resp. GD-consistent, limit-conforming, or compact – in the sense of Definitions 2.2, 2.4, 2.5 and 2.8), then $(\mathcal{D}_m^\star)_{m\in\mathbb{N}}$ is also coercive (resp. GD-consistent, limit-conforming, or compact).

2. *Reciprocally, if $(\mathcal{D}_m)_{m\in\mathbb{N}}$ and $(\mathcal{D}_m^\star)_{m\in\mathbb{N}}$ are both limit-conforming and compact in the sense of Definitions 2.5 and 2.8, then there exists $(\omega_m)_{m\in\mathbb{N}}$ such that (7.59) holds.*

Proof.
Step 1: proof of Item 1.
Coercivity: let us assume that $(\mathcal{D}_m)_{m\in\mathbb{N}}$ is coercive with constant C_P. Setting $M = \sup_{m\in\mathbb{N}} \omega_m$, using the triangle inequality and invoking (7.59), we have, for any $v \in X_{\mathcal{D}_m,0}$,

$$\begin{aligned}\left\|\Pi^\star_{\mathcal{D}_m} v\right\|_{L^p(\Omega)} &\le \left\|\Pi^\star_{\mathcal{D}_m} v - \Pi_{\mathcal{D}_m} v\right\|_{L^p(\Omega)} + \left\|\Pi_{\mathcal{D}_m} v\right\|_{L^p(\Omega)} \\ &\le M \left\|\nabla_{\mathcal{D}_m} v\right\|_{L^p(\Omega)^d} + C_{\mathcal{D}_m} \left\|\nabla_{\mathcal{D}_m} v\right\|_{L^p(\Omega)^d}.\end{aligned}$$

The coercivity of $(\mathcal{D}_m^\star)_{m\in\mathbb{N}}$ follows, with $C_{\mathcal{D}_m^\star} \le M + C_{\mathcal{D}_m} \le M + C_P$.

GD-consistency: let us assume that $(\mathcal{D}_m)_{m\in\mathbb{N}}$ is consistent. Using the triangle inequality and (7.59), we write, for $v \in X_{\mathcal{D}_m,0}$ and $\varphi \in W_0^{1,p}(\Omega)$,

$$\begin{aligned}S_{\mathcal{D}_m^\star}(\varphi) \le\ & \left\|\Pi_{\mathcal{D}_m^\star} v - \varphi\right\|_{L^p(\Omega)} + \left\|\nabla_{\mathcal{D}_m} v - \nabla\varphi\right\|_{L^p(\Omega)^d} \\ &\le \omega_m \left\|\nabla_{\mathcal{D}_m} v\right\|_{L^p(\Omega)^d} + \left\|\Pi_{\mathcal{D}_m} v - \varphi\right\|_{L^p(\Omega)} + \left\|\nabla_{\mathcal{D}_m} v - \nabla\varphi\right\|_{L^p(\Omega)^d} \\ &\le \omega_m \left\|\nabla\varphi\right\|_{L^p(\Omega)^d} + \omega_m \left\|\nabla_{\mathcal{D}_m} v - \nabla\varphi\right\|_{L^p(\Omega)^d} \\ &\quad + \left\|\Pi_{\mathcal{D}_m} v - \varphi\right\|_{L^p(\Omega)} + \left\|\nabla_{\mathcal{D}_m} v - \nabla\varphi\right\|_{L^p(\Omega)^d} \\ &\le \omega_m \left\|\nabla\varphi\right\|_{L^p(\Omega)^d} \\ &\quad + (1+M)(\left\|\Pi_{\mathcal{D}_m} v - \varphi\right\|_{L^p(\Omega)} + \left\|\nabla_{\mathcal{D}_m} v - \nabla\varphi\right\|_{L^p(\Omega)^d}).\end{aligned}$$

Hence $S_{\mathcal{D}_m^\star}(\varphi) \le \omega_m \left\|\nabla\varphi\right\|_{L^p(\Omega)^d} + (1+M) S_{\mathcal{D}_m}(\varphi)$ and the consistency of $(\mathcal{D}_m^\star)_{m\in\mathbb{N}}$ follows from the consistency of $(\mathcal{D}_m)_{m\in\mathbb{N}}$ and from $\lim_{m\to\infty} \omega_m = 0$.

Limit-conformity: let us now assume that $(\mathcal{D}_m)_{m\in\mathbb{N}}$ is limit-conforming. By the triangle inequality and (7.59), for any $\boldsymbol{\varphi} \in W^{p'}_{\mathrm{div}}(\Omega)$,

$$\begin{aligned}&\left|\int_\Omega \Big(\nabla_{\mathcal{D}_m} v(\boldsymbol{x}) \cdot \boldsymbol{\varphi}(\boldsymbol{x}) + \Pi^\star_{\mathcal{D}_m} v(\boldsymbol{x}) \mathrm{div}\boldsymbol{\varphi}(\boldsymbol{x})\Big) \mathrm{d}\boldsymbol{x}\right| \\ &\qquad \le \left\|\mathrm{div}\boldsymbol{\varphi}\right\|_{L^{p'}(\Omega)} \omega_m \left\|\nabla_{\mathcal{D}_m} v\right\|_{L^p(\Omega)^d} \\ &\qquad\quad + \left|\int_\Omega \left(\nabla_{\mathcal{D}_m} v(\boldsymbol{x}) \cdot \boldsymbol{\varphi}(\boldsymbol{x}) + \Pi_{\mathcal{D}_m} v(\boldsymbol{x}) \mathrm{div}\boldsymbol{\varphi}(\boldsymbol{x})\right) \mathrm{d}\boldsymbol{x}\right|.\end{aligned}$$

We infer that $W_{\mathcal{D}_m^\star}(\boldsymbol{\varphi}) \le \omega_m \left\|\mathrm{div}\boldsymbol{\varphi}\right\|_{L^{p'}(\Omega)} + W_{\mathcal{D}_m}(\boldsymbol{\varphi}) \to 0$ as $m \to \infty$, and the limit-conformity of $(\mathcal{D}_m^\star)_{m\in\mathbb{N}}$ is established.

Compactness: we now assume that $(\mathcal{D}_m)_{m\in\mathbb{N}}$ is compact. If $(\nabla_{\mathcal{D}_m} v_m)_{m\in\mathbb{N}}$ is bounded in $L^p(\Omega)^d$, then the compactness of $(\mathcal{D}_m)_{m\in\mathbb{N}}$ ensures that $(\Pi_{\mathcal{D}_m} v_m)_{m\in\mathbb{N}}$ is relatively compact in $L^p(\Omega)$. Since $\left\|\Pi^\star_{\mathcal{D}_m} v_m - \Pi_{\mathcal{D}_m} v_m\right\|_{L^p(\Omega)}$ tends to 0 as $m \to \infty$ by (7.59), we deduce that $(\Pi^\star_{\mathcal{D}_m} v_m)_{m\in\mathbb{N}}$ is relatively compact in $L^p(\Omega)$.

Step 2: proof of Item 2.
We reason by way of contradiction, therefore assuming that $(\mathcal{D}_m)_{m\in\mathbb{N}}$ and $(\mathcal{D}^\star_m)_{m\in\mathbb{N}}$ are both compact and limit-conforming, and that

$$\omega_m := \max_{v\in X_{\mathcal{D}_m,0}\setminus\{0\}} \frac{\|\Pi_{\mathcal{D}_m} v - \Pi^\star_{\mathcal{D}_m} v\|_{L^p(\Omega)}}{\|\nabla_{\mathcal{D}_m} v\|_{L^p(\Omega)^d}} \not\to 0 \text{ as } m \to \infty. \tag{7.60}$$

Then there exists $\varepsilon_0 > 0$, a subsequence of $(\mathcal{D}_m, \mathcal{D}^\star_m)_{m\in\mathbb{N}}$ (not denoted differently) and for each $m \in \mathbb{N}$ an element $v_m \in X_{\mathcal{D}_m,0}\setminus\{0\}$ such that $\left\|\Pi^\star_{\mathcal{D}_m} v_m - \Pi_{\mathcal{D}_m} v_m\right\|_{L^p(\Omega)} \geq \varepsilon_0 \left\|\nabla_{\mathcal{D}_m} v_m\right\|_{L^p(\Omega)^d}$. Since $v_m \neq 0$, the element $\widetilde{v}_m = \|\nabla_{\mathcal{D}_m} v_m\|^{-1}_{L^p(\Omega)^d} v_m \in X_{\mathcal{D}_m,0}$ is well defined. It satisfies $\|\nabla_{\mathcal{D}_m} \widetilde{v}_m\|_{L^p(\Omega)^d} = 1$ and

$$\left\|\Pi^\star_{\mathcal{D}_m} \widetilde{v}_m - \Pi_{\mathcal{D}_m} \widetilde{v}_m\right\|_{L^p(\Omega)} \geq \varepsilon_0. \tag{7.61}$$

Extract another subsequence such that $\nabla_{\mathcal{D}_m} \widetilde{v}_m$ weakly converges to some G in $L^p(\Omega)^d$, and, using the compactness of $(\mathcal{D}_m)_{m\in\mathbb{N}}$ and $(\mathcal{D}^\star_m)_{m\in\mathbb{N}}$, $\Pi_{\mathcal{D}_m} \widetilde{v}_m \to v$ in $L^p(\Omega)$ and $\Pi^\star_{\mathcal{D}_m} \widetilde{v}_m \to v^\star$ in $L^p(\Omega)$. Passing to the limit in (7.61) we find $\|v - v^\star\|_{L^p(\Omega)} \geq \varepsilon_0$. Extending the functions $\nabla_{\mathcal{D}_m} \widetilde{v}_m$, $\Pi_{\mathcal{D}_m} \widetilde{v}_m$ and $\Pi^\star_{\mathcal{D}_m} \widetilde{v}_m$ by 0 outside Ω, we see that, for any $\boldsymbol{\varphi} \in W^{p'}_{\mathrm{div}}(\Omega)$,

$$\left|\int_{\mathbb{R}^d} \left(\nabla_{\mathcal{D}_m} \widetilde{v}_m(\boldsymbol{x}) \cdot \boldsymbol{\varphi}(\boldsymbol{x}) + \Pi^\star_{\mathcal{D}_m} \widetilde{v}_m(\boldsymbol{x}) \mathrm{div}\boldsymbol{\varphi}(\boldsymbol{x})\right) \mathrm{d}\boldsymbol{x}\right| \leq W_{\mathcal{D}^\star_m}(\boldsymbol{\varphi}),$$

and

$$\left|\int_{\mathbb{R}^d} \left(\nabla_{\mathcal{D}_m} \widetilde{v}_m(\boldsymbol{x}) \cdot \boldsymbol{\varphi}(\boldsymbol{x}) + \Pi_{\mathcal{D}_m} \widetilde{v}_m(\boldsymbol{x}) \mathrm{div}\boldsymbol{\varphi}(\boldsymbol{x})\right) \mathrm{d}\boldsymbol{x}\right| \leq W_{\mathcal{D}_m}(\boldsymbol{\varphi}).$$

By limit-conformity of both sequences of GDs, let $m \to \infty$ to find

$$\int_{\mathbb{R}^d} \left(G \cdot \boldsymbol{\varphi}(\boldsymbol{x}) + v^\star(\boldsymbol{x}) \mathrm{div}\boldsymbol{\varphi}(\boldsymbol{x})\right) \mathrm{d}\boldsymbol{x} = \int_{\mathbb{R}^d} \left(G \cdot \boldsymbol{\varphi}(\boldsymbol{x}) + v(\boldsymbol{x}) \mathrm{div}\boldsymbol{\varphi}(\boldsymbol{x})\right) \mathrm{d}\boldsymbol{x} = 0.$$

This proves that $v, v^\star \in W^{1,p}_0(\Omega)$ and that $G = \nabla v = \nabla v^\star$. Poincaré's inequality then gives $v = v^\star$, which contradicts $\|v - v^\star\|_{L^p(\Omega)} \geq \varepsilon_0$. Therefore the sequence $(\omega_m)_{m\in\mathbb{N}}$ defined by (7.60) satisfies (7.59). ∎

Remark 7.51. Three estimates obtained in the course of this proof deserve to be put forward. Under Assumption (7.59) and setting $M = \sup_{m\in\mathbb{N}} \omega_m$, we saw that

$$C_{\mathcal{D}_m^\star} \leq M + C_{\mathcal{D}_m}, \tag{7.62}$$

that

$$\forall \varphi \in W_0^{1,p}(\Omega)\,,\ S_{\mathcal{D}_m^\star}(\varphi) \leq \omega_m \left\|\nabla\varphi\right\|_{L^p(\Omega)^d} + (1+M) S_{\mathcal{D}_m}(\varphi), \tag{7.63}$$

and that

$$\forall \boldsymbol{\varphi} \in W_{\mathrm{div}}^{p'}(\Omega)\,,\ W_{\mathcal{D}_m^\star}(\boldsymbol{\varphi}) \leq \omega_m \left\|\mathrm{div}\boldsymbol{\varphi}\right\|_{L^{p'}(\Omega)} + W_{\mathcal{D}_m}(\boldsymbol{\varphi}). \tag{7.64}$$

These estimates are particularly useful in the situation where rates of convergence of $S_{\mathcal{D}_m}$ and $W_{\mathcal{D}_m}$ to 0 are known. Indeed, in this case, (7.63) and (7.64) give some rates of convergence of $S_{\mathcal{D}_m^\star}$ and $W^\star_{\mathcal{D}_m}$ to 0, which in turns provides rates of convergence for $\mathcal{D}_m^\star$ applied to linear, and some non-linear, problems (see, *e.g.*, Theorems 2.28 and 2.38).
An example of this is given for mass-lumped $\mathbb{P}_1$ gradient discretisations in Remark 8.18.

7.3.6 Non-homogeneous Dirichlet, Neumann and Fourier boundary conditions

The (minor) changes that must be made in the definitions and results in the three previous sections in case of non-homogeneous Dirichlet conditions, Neumann conditions or Fourier conditions are now introduced. Mixed boundary conditions being deduced from Dirichlet and Neumann conditions, we do not detail this last case.
Upon trivial changes of the space of discrete unknowns, the definition of a mass-lumped GD (Definition 7.45) does not depend on the considered boundary conditions since it only deals with the reconstruction $\Pi_{\mathcal{D}}$.

Non-homogeneous Dirichlet boundary conditions

LLE gradient discretisation

Definition 7.52 (LLE GD for non-homogeneous Dirichlet BCs). *A gradient discretisation $\mathcal{D} = (X_{\mathcal{D}}, \mathcal{I}_{\mathcal{D},\partial}, \Pi_{\mathcal{D}}, \nabla_{\mathcal{D}})$ for non-homogeneous Dirichlet conditions in the sense of Definition 2.49 is an LLE GD if*

- *There exists a finite set $I = I_\Omega \sqcup I_\partial$ such that*

$$X_{\mathcal{D}} = \{v = (v_i)_{i\in I} \,:\, v_i \in \mathbb{R} \text{ for all } i \in I\} = X_{\mathcal{D},0} \oplus X_{\mathcal{D},\partial}$$

  where

$$X_{\mathcal{D},0} = \{v = (v_i)_{i\in I} \,:\, v_i \in \mathbb{R} \text{ for all } i \in I_\Omega\,,\ v_i = 0 \text{ for all } i \in I_\partial\},$$

  and

$$X_{\mathcal{D},\partial} = \{v = (v_i)_{i\in I} \,:\, v_i \in \mathbb{R} \text{ for all } i \in I_\partial\,,\ v_i = 0 \text{ for all } i \in I_\Omega\},$$

- *$\Pi_{\mathcal{D}}$ and $\nabla_{\mathcal{D}}$ satisfy Item 2 in Definition 7.33,*
- *$\mathcal{I}_{\mathcal{D},\partial} : W^{1-\frac{1}{p},p}(\partial\Omega) \to X_{\mathcal{D},\partial}$ is a linear mapping.*

The regularity factor $\mathrm{reg}_{\mathrm{LLE}}(\mathcal{D})$ is defined by (7.34).

Proposition 7.53 (GD-consistency of LLE GDs for non-homogeneous Dirichlet BCs). *Let $(\mathcal{D}_m)_{m\in\mathbb{N}}$ be a sequence of LLE GDs for non-homogeneous Dirichlet boundary conditions, in the sense of Definition 7.52. Denote by $\mathcal{M}_m$ the mesh associated to $\mathcal{D}_m$. Assume that $(\mathrm{reg}_{\mathrm{LLE}}(\mathcal{D}_m))_{m\in\mathbb{N}}$ is bounded, that $h_{\mathcal{M}_m} \to 0$ as $m \to \infty$, that (2.96) holds, and that*

$$\forall \varphi \in C^\infty(\overline{\Omega}),$$
$$\max_{K\in\mathcal{M}_m} \max_{i\in I_K\cap I_\partial} \frac{|(\mathcal{I}_{\mathcal{D}_m,\partial}\gamma(\varphi))_i - \varphi(\boldsymbol{x}_i)|}{\mathrm{diam}(K)} \to 0 \text{ as } m \to \infty. \quad (7.65)$$

Then $(\mathcal{D}_m)_{m\in\mathbb{N}}$ is GD-consistent in the sense of Definition 2.51.

Here and in the following, to simplify the notation we make the convention that $\max_{i\in I_K\cap I_\partial} Z_i = 0$ if $I_K \cap I_\partial = \emptyset$.

Proof. The property (2.96) enables us to check the GD-consistency only on smooth functions (see Lemma 2.52). Let $\varphi \in C^2(\mathbb{R}^d)$. The function $v^m = (\varphi(\boldsymbol{x}_i^m))_{i\in I^m}$, defined as in the proof of Proposition 7.37, has good approximation properties since $\nabla_{\mathcal{D}_m} v^m \to \nabla\varphi$ in $L^p(\Omega)^d$ and $\Pi_{\mathcal{D}_m} v^m \to \varphi$ in $L^\infty(\Omega)$ as $m \to \infty$ (these properties were established in the proof of Proposition 7.37 without using the zero boundary value of φ). However, v^m does not necessarily satisfy the requirement $v^m - \mathcal{I}_{\mathcal{D}_m,\partial}\gamma(\varphi) \in X_{\mathcal{D}_m,0}$ in Definition 2.51.
Consider therefore $w^m \in X_{\mathcal{D}_m,0} + \mathcal{I}_{\mathcal{D}_m,\partial}\gamma(\varphi)$ defined by $w_i^m = v_i^m = \varphi(\boldsymbol{x}_i)$ if $i \in I_\Omega$ and $w_i^m = (\mathcal{I}_{\mathcal{D}_m,\partial}\gamma(\varphi))_i$ if $i \in I_\partial$. Let, for $K \in \mathcal{M}_m$,

$$\omega_m(K) = \max_{i\in I_K} \frac{|w_i^m - v_i^m|}{\mathrm{diam}(K)}. \quad (7.66)$$

By definition of $\|\mathcal{G}_K\|_p$ (we do not explicitly denote the dependency with respect to m of this $\mathbb{P}_1$-exact gradient reconstruction),

$$\begin{aligned}
\|\mathcal{G}_K[(v_i^m)_{i\in I_K}] - \mathcal{G}_K[(w_i^m)_{i\in I_K}]\|_{L^p(K)^d} &\le \left\| \sum_{i\in I_K} |v_i^m - w_i^m|\, |\mathcal{G}_K^i| \right\|_{L^p(K)^d} \\
&\le \omega_m(K)\mathrm{diam}(K) \left\| \sum_{i\in I_K} |\mathcal{G}_K^i| \right\|_{L^p(K)^d} \\
&\le \|\mathcal{G}_K\|_p\, |K|^{\frac{1}{p}} \omega_m(K) \\
&\le \mathrm{reg}_{\mathrm{LLE}}(\mathcal{D}_m)|K|^{\frac{1}{p}}\omega_m(K). \quad (7.67)
\end{aligned}$$

Raising this estimate the power p and summing the result over $K \in \mathcal{M}_m$ gives

$$\|\nabla_{\mathcal{D}_m} v^m - \nabla_{\mathcal{D}_m} w^m\|_{L^p(\Omega)^d} \leq \mathrm{reg}_{\mathrm{LLE}}(\mathcal{D}_m)|\Omega|^{\frac{1}{p}} \max_{K\in\mathcal{M}_m} \omega_m(K).$$

Since $(\mathrm{reg}_{\mathrm{LLE}}(\mathcal{D}_m))_{m\in\mathbb{N}}$ is bounded, Assumption (7.65) shows that the right-hand side of the previous inequality tends to 0 as $m \to \infty$. Hence, the convergence of $(\nabla_{\mathcal{D}_m} v^m)_{m\in\mathbb{N}}$ gives $\nabla_{\mathcal{D}_m} w^m \to \nabla\varphi$ in $L^p(\Omega)^d$. The convergence of $(\Pi_{\mathcal{D}_m} w^m)_{m\in\mathbb{N}}$ is established similarly. ∎

Barycentric condensation

The change to be made in Definition 7.40, besides considering an LLE GD for non-homogeneous Dirichlet conditions, is the obvious replacement of $X_{\mathcal{D}^{\mathrm{BA}},0} = \{v = (v_i)_{i\in I^{\mathrm{BA}}} : v_i \in \mathbb{R} \text{ for all } i \in I^{\mathrm{BA}}; v_i = 0 \text{ for all } i \in I_\partial\}$ with $X_{\mathcal{D}^{\mathrm{BA}}} = \{v = (v_i)_{i\in I^{\mathrm{BA}}} : v_i \in \mathbb{R} \text{ for all } i \in I^{\mathrm{BA}}\}$. Notice that the boundary unknowns are not eliminated ($I_\partial \subset I^{\mathrm{BA}}$).
Lemma 7.42, that is the preservation of the LLE property, still holds (the proof did not use the zero boundary condition). The properties of barycentric condensations, Theorem 7.43, is also valid provided that we assume (2.96) and (7.65) – to establish the GD-consistency by invoking Proposition 7.53.

Mass-lumping

Since the interpolation operator $\mathcal{I}_{\mathcal{D},\partial}$ is unchanged by the mass lumping of $\mathcal{D}$, it is easy to see that Theorem 7.50, and thus Theorem 7.49, still hold modulo a trivial adjustment of the space of discrete unknowns.

Neumann boundary conditions

LLE gradient discretisation

Definition 7.54 (LLE GD for Neumann BCs). *A gradient discretisation $\mathcal{D} = (X_\mathcal{D}, \Pi_\mathcal{D}, \nabla_\mathcal{D})$ (resp. a $\mathcal{D} = (X_\mathcal{D}, \Pi_\mathcal{D}, \mathbb{T}_\mathcal{D}, \nabla_\mathcal{D})$) for homogeneous Neumann boundary conditions (resp. non-homogeneous Neumann boundary conditions) is an LLE GD if*

- *There is a finite set $I = I_\Omega \sqcup I_\partial$ such that*

$$X_\mathcal{D} = \{v = (v_i)_{i\in I} : v_i \in \mathbb{R} \text{ for all } i \in I\} = X_{\mathcal{D},0} \oplus X_{\mathcal{D},\partial},$$

  where

$$X_{\mathcal{D},0} = \{v = (v_i)_{i\in I} : v_i \in \mathbb{R} \text{ for all } i \in I_\Omega\,,\ v_i = 0 \text{ for all } i \in I_\partial\},$$

  and

$$X_{\mathcal{D},\partial} = \{v = (v_i)_{i\in I} : v_i \in \mathbb{R} \text{ for all } i \in I_\partial\,,\ v_i = 0 \text{ for all } i \in I_\Omega\},$$

- *$\Pi_{\mathcal{D}}$ and $\nabla_{\mathcal{D}}$ satisfy Item 2 in Definition 7.33,*

The regularity factor $\mathrm{reg}_{\mathrm{LLE}}(\mathcal{D})$ *is defined by* (7.34).

The proof of the following proposition is identical to the proof of Proposition 7.37, in which we actually did not use the boundary value of the functions.

Proposition 7.55 (GD-consistency of LLE GDs for Neumann BCs). *Let $(\mathcal{D}_m)_{m\in\mathbb{N}}$ be a sequence of LLE GDs for Neumann boundary conditions, in the sense of Definition 7.54. We denote by $\mathcal{M}_m$ the mesh associated to $\mathcal{D}_m$. If $(\mathrm{reg}_{\mathrm{LLE}}(\mathcal{D}_m))_{m\in\mathbb{N}}$ is bounded and $h_{\mathcal{M}_m} \to 0$ as $m \to \infty$, then $(\mathcal{D}_m)_{m\in\mathbb{N}}$ is GD-consistent in the sense of Definition 3.4.*

Barycentric condensation

Starting from an LLE GD for Neumann conditions as defined above, a barycentric condensation is defined as in Definition 7.40, with the addition that, in the case of non-homogeneous conditions, the trace $\mathbb{T}_{\mathcal{D}^{\mathrm{BA}}}$ of $\mathcal{D}^{\mathrm{BA}}$ is defined by $\mathbb{T}_{\mathcal{D}^{\mathrm{BA}}} v = \mathbb{T}_{\mathcal{D}} \widetilde{v}$, where $\widetilde{v}$ is defined by (7.47). We note that, with the norm (3.8) considered in a GD for Neumann boundary conditions, we have $\|v\|_{\mathcal{D}^{\mathrm{BA}}} = \|\widetilde{v}\|_{\mathcal{D}}$.
The preservation of the LLE property by barycentric condensation (Lemma 7.42) is still valid, as well as Theorem 7.43.

Mass-lumping

There is no change in the definition of a mass-lumped GD. Note that, if $\Pi_{\mathcal{D}}$ and $\Pi^{\star}_{\mathcal{D}}$ are two function reconstructions on $X_{\mathcal{D}}$, by the Hölder inequality (D.6),

$$\left|\int_{\Omega} \Pi^{\star}_{\mathcal{D}} v(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}\right| \leq \left|\int_{\Omega} \Pi_{\mathcal{D}} v(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}\right| + |\Omega|^{1/p'} \left\|\Pi^{\star}_{\mathcal{D}} v - \Pi_{\mathcal{D}} v\right\|_{L^p(\Omega)},$$

and vice versa with $\Pi_{\mathcal{D}}$ and $\Pi^{\star}_{\mathcal{D}}$ switched. This enables to prove the equivalent, for Neumann boundary conditions, of Theorem 7.50 in which the norm $\|v\|_{\mathcal{D}_m}$ in (7.59) is defined by (3.8).
Theorem 7.49 then clearly holds, provided that we use the norm (3.8) in (7.58).

Remark 7.56 (Mass-lumping the trace reconstruction)
In the case of non-homogeneous Neumann conditions, one could also mass-lump the trace reconstruction $\mathbb{T}_{\mathcal{D}}$. This would be useful for problems that are non-linear with respect to the trace, or that involve the trace in a time-stepping. If the trace is mass-lumped, then for Theorems 7.50 and 7.49 to hold one must introduce this trace in (7.59) and (7.58). This latter formula, for example, would therefore become

$$\left\|\Pi^{\mathrm{ML}}_{\mathcal{D}_m} v - \Pi_{\mathcal{D}_m} v\right\|_{L^p(\Omega)} + \left\|\mathbb{T}^{\mathrm{ML}}_{\mathcal{D}_m} v - \mathbb{T}_{\mathcal{D}_m} v\right\|_{L^p(\partial\Omega)} \leq \omega_m \|v\|_{\mathcal{D}_m}. \tag{7.68}$$

Fourier boundary conditions

LLE gradient discretisation

LLE GDs for Fourier boundary conditions are probably those that undergo the major changes with respect to Definition 7.33. Because the consistency of GDs for Fourier boundary condition involves the trace reconstruction, this trace must be dealt with in a similar way as $\Pi_{\mathcal{D}}$.

Definition 7.57 (LLE GD for Fourier BCs). *A gradient discretisation $\mathcal{D} = (X_{\mathcal{D}}, \Pi_{\mathcal{D}}, \mathbb{T}_{\mathcal{D}}, \nabla_{\mathcal{D}})$ for Fourier boundary conditions is an LLE GD if*

- *There is a finite set $I = I_\Omega \sqcup I_\partial$ such that*

$$X_{\mathcal{D}} = \{v = (v_i)_{i\in I} \,:\, v_i \in \mathbb{R} \text{ for all } i \in I\} = X_{\mathcal{D},0} \oplus X_{\mathcal{D},\partial},$$

  where

$$X_{\mathcal{D},0} = \{v = (v_i)_{i\in I} \,:\, v_i \in \mathbb{R} \text{ for all } i \in I_\Omega\,,\ v_i = 0 \text{ for all } i \in I_\partial\},$$

  and

$$X_{\mathcal{D},\partial} = \{v = (v_i)_{i\in I} \,:\, v_i \in \mathbb{R} \text{ for all } i \in I_\partial\,,\ v_i = 0 \text{ for all } i \in I_\Omega\},$$

- *$\Pi_{\mathcal{D}}$ and $\nabla_{\mathcal{D}}$ satisfy Item 2 in Definition 7.33,*
- *There exists a finite mesh $\mathcal{M}_\partial$ of $\partial\Omega$ and, for each $K_\partial \in \mathcal{M}_\partial$, a subset $I_{K_\partial} \subset I$ and a $\mathbb{P}_0$-exact function reconstruction $\pi_{K_\partial} = (\pi^i_{K_\partial})_{i\in I_{K_\partial}}$ on K_∂ such that*

$$\forall v \in X_{\mathcal{D}}\,, \quad \text{for a.e. } \boldsymbol{x} \in K_\partial \text{ (for the } (d-1)\text{-dimensional measure)}$$
$$\mathbb{T}_{\mathcal{D}} v(\boldsymbol{x}) = \pi_{K_\partial}[(v_i)_{i\in I_{K_\partial}}](\boldsymbol{x}) = \sum_{i\in I_{K_\partial}} v_i \pi^i_{K_\partial}(\boldsymbol{x}).$$

The LLE regularity of $\mathcal{D}$ is defined by

$$\begin{aligned}\mathrm{reg}_{\mathrm{LLE}}(\mathcal{D}) = &\max_{K\in\mathcal{M}} \left(\|\pi_K\|_p + \|\mathcal{G}_K\|_p + \max_{i\in I_K} \frac{\mathrm{dist}(\boldsymbol{x}_i, K)}{\mathrm{diam}(K)} \right) \\ &+ \max_{K_\partial\in\mathcal{M}_\partial} \left(\|\pi_{K_\partial}\|_p + \max_{i\in I_{K_\partial}} \frac{\mathrm{dist}(\boldsymbol{x}_i, K_\partial)}{\mathrm{diam}(K_\partial)} \right).\end{aligned} \tag{7.69}$$

The following proposition is then proved as Proposition 7.37, the estimate on $\|\mathbb{T}_{\mathcal{D}_m} v^m - \gamma(\varphi)\|_{L^p(\partial\Omega)}$ is obtained as the estimate on $\|\Pi_{\mathcal{D}_m} v^m - \varphi\|_{L^p(\Omega)}$.

Proposition 7.58 (Consistency of LLE GDs for Fourier BCs). *Let $(\mathcal{D}_m)_{m\in\mathbb{N}}$ be a sequence of LLE GDs for Fourier boundary conditions, in the sense of Definition 7.57. We denote by $\mathcal{M}_m$ the mesh associated to $\mathcal{D}_m$. If $(\mathrm{reg}_{\mathrm{LLE}}(\mathcal{D}_m))_{m\in\mathbb{N}}$ is bounded and $h_{\mathcal{M}_m} \to 0$ as $m \to \infty$ then $(\mathcal{D}_m)_{m\in\mathbb{N}}$ is consistent in the sense of Definition 3.37.*

Barycentric condensation

With the above definition of an LLE GD for Fourier boundary conditions, we use the same definition of barycentric condensation as for non-homogeneous Neumann conditions, *i.e.* Definition 7.40 to which we add the relation $\mathbb{T}_{\mathcal{D}^{\text{BA}}} v = \mathbb{T}_{\mathcal{D}} \widetilde{v}$. With the norm (3.51) of a GD for Fourier boundary conditions, we still have $\|v\|_{\mathcal{D}^{\text{BA}}} = \|\widetilde{v}\|_{\mathcal{D}}$.
The preservation of the LLE property (Lemma 7.42) is still valid, the trace reconstruction $\mathbb{T}_{\mathcal{D}^{\text{BA}}}$ being dealt with as $\Pi_{\mathcal{D}^{\text{BA}}}$. A barycentric condensation for Fourier boundary conditions preserves the properties of a GD (Theorem 7.43 holds).

Mass-lumping

The definition of a mass-lumped GD for Fourier boundary conditions is not different from Definition 7.45. In particular, if the trace reconstruction is not mass-lumped, Theorems 7.50 and 7.49 hold. If the trace reconstruction trace is mass-lumped, Assumption (7.58) must be replaced with (7.68).

7.4 $W^{2,p}$ estimate of $S_{\mathcal{D}}$ for local linearly exact GDs

Estimates on $S_{\mathcal{D}}(\varphi)$ are useful to obtain rates of convergences of GSs for linear (and some non-linear) problems, see, *e.g.*, Theorem 2.28 and Theorem 2.38. The estimate (7.37) on $S_{\mathcal{D}}(\varphi)$ requires $\varphi \in W_0^{1,p}(\Omega) \cap W^{2,\infty}(\mathbb{R}^d)$. Hence, to use for example this estimate in the aforementioned theorems, the solution to the corresponding problem ((2.20) or (2.61)) would need to have a $W^{2,\infty}$ regularity, which is quite restrictive.
The purpose of this section is to write a consistency estimate similar to (7.37) in the case $\varphi \in W^{2,p}(\Omega)$ for $p > d/2$ (this condition ensures the embedding of $W^{2,p}(\Omega)$ into $C(\overline{\Omega})$). This regularity property is much more likely to hold, if φ is the solution of problems (2.20) or (2.61), than the $W^{2,\infty}$ regularity.

7.4.1 Functional estimates in $W^{2,p}$

We start with a lemma that compares in the $L^p(V)$ norm a function $\varphi \in W^{1,p}(V)$ with its average value on a ball in V.

Lemma 7.59. *Let $V \subset \mathbb{R}^d$ be an open bounded set that is star-shaped with respect to all points in a ball $B \subset V$. Let $p \in [1, +\infty)$. There exists C_{14} depending only on d and p such that, for any $\varphi \in W^{1,p}(V)$,*

$$\left\| \varphi - \frac{1}{|B|} \int_B \varphi(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \right\|_{L^p(V)} \leq C_{14} \frac{\mathrm{diam}(V)^{\frac{d}{p}+1}}{\mathrm{diam}(B)^{\frac{d}{p}}} \left\| |\nabla \varphi| \right\|_{L^p(V)}. \tag{7.70}$$

Proof. Since $C^\infty(V) \cap W^{1,p}(V)$ is dense in $W^{1,p}(V)$, we only need to prove the result for $\varphi \in C^\infty(V) \cap W^{1,p}(V)$, and the conclusion follows by density. To simplify the notations we let $h_V = \mathrm{diam}(V)$. For all $(\boldsymbol{x}, \boldsymbol{y}) \in V \times B$, since V is star-shaped with respect to y the segment $[\boldsymbol{x}, \boldsymbol{y}]$ belongs to V and we can write

$$\varphi(\boldsymbol{x}) - \varphi(\boldsymbol{y}) = \int_0^1 \nabla\varphi(t\boldsymbol{x} + (1-t)\boldsymbol{y}) \cdot (\boldsymbol{x} - \boldsymbol{y})\mathrm{d}t.$$

Taking the average value with respect to $\boldsymbol{y} \in B$ and writing $|\boldsymbol{x} - \boldsymbol{y}| \le h_V$ gives

$$\begin{aligned}\left|\varphi(\boldsymbol{x}) - \frac{1}{|B|}\int_B \varphi(\boldsymbol{y})\mathrm{d}\boldsymbol{y}\right| &= \left|\frac{1}{|B|}\int_B\int_0^1 \nabla\varphi(t\boldsymbol{x} + (1-t)\boldsymbol{y}) \cdot (\boldsymbol{x} - \boldsymbol{y})\mathrm{d}t\mathrm{d}y\right| \\ &\le \frac{h_V}{|B|}\int_B\int_0^1 |\nabla\varphi(t\boldsymbol{x} + (1-t)\boldsymbol{y})|\mathrm{d}t\mathrm{d}y.\end{aligned}$$

Taking the power p, using the Jensen inequality (D.10) with the convex function $\Psi = |\cdot|^p$ and $A = B \times (0,1)$, and integrating with respect to $\boldsymbol{x} \in V$, we get

$$\begin{aligned}&\int_V \left|\varphi(\boldsymbol{x}) - \frac{1}{|B|}\int_B \varphi(\boldsymbol{y})\mathrm{d}\boldsymbol{y}\right|^p \mathrm{d}\boldsymbol{x} \\ &\qquad \le \frac{h_V^p}{|B|}\int_V\int_B\int_0^1 |\nabla\varphi(t\boldsymbol{x} + (1-t)\boldsymbol{y})|^p\mathrm{d}t\mathrm{d}y\mathrm{d}\boldsymbol{x}. \quad (7.71)\end{aligned}$$

We then apply the change of variable $\boldsymbol{x} \in V \to \boldsymbol{z} = t\boldsymbol{x} + (1-t)\boldsymbol{y}$, which has values in V since V is star-shaped with respect to all points in B. This gives

$$\begin{aligned}&\int_V\int_B\int_0^1 |\nabla\varphi(t\boldsymbol{x} + (1-t)\boldsymbol{y})|^p\mathrm{d}t\mathrm{d}\boldsymbol{y}\mathrm{d}\boldsymbol{x} \\ &\qquad \le \int_V |\nabla\varphi(\boldsymbol{z})|^p \int_B\int_{I(\boldsymbol{z},\boldsymbol{y})} t^{-d}\mathrm{d}t\mathrm{d}\boldsymbol{y}\mathrm{d}\boldsymbol{z}, \quad (7.72)\end{aligned}$$

where $I(\boldsymbol{z}, \boldsymbol{y}) = \{t \in (0,1) \,:\, \exists \boldsymbol{x} \in V\,,\ t\boldsymbol{x} + (1-t)\boldsymbol{y} = \boldsymbol{z}\}$. For $t \in I(\boldsymbol{z}, \boldsymbol{y})$ we have $t(\boldsymbol{x} - \boldsymbol{y}) = \boldsymbol{z} - \boldsymbol{y}$ for some $\boldsymbol{x} \in V$ and therefore $h_V t \ge |\boldsymbol{z} - \boldsymbol{y}|$. Hence $I(\boldsymbol{z}, \boldsymbol{y}) \subset [\frac{|\boldsymbol{z}-\boldsymbol{y}|}{h_V}, 1]$ and we deduce that (for $d > 1$)

$$\int_{I(\boldsymbol{z},\boldsymbol{y})} t^{-d}\mathrm{d}t \le \int_{\frac{|\boldsymbol{z}-\boldsymbol{y}|}{h_V}}^1 t^{-d}\mathrm{d}t \le \frac{1}{d-1}\frac{h_V^{d-1}}{|\boldsymbol{z} - \boldsymbol{y}|^{d-1}}. \quad (7.73)$$

Thus, denoting by ω_d the area of the unit sphere in $\mathbb{R}^d$, since $B \subset V \subset B(\boldsymbol{z}, h_V)$ for all $\boldsymbol{z} \in V$,

$$\int_B\int_{I(\boldsymbol{z},\boldsymbol{y})} t^{-d}\mathrm{d}t\mathrm{d}\boldsymbol{y} \le \frac{h_V^{d-1}}{d-1}\int_B \frac{1}{|\boldsymbol{z} - \boldsymbol{y}|^{d-1}}\mathrm{d}\boldsymbol{y}$$

$$\leq \frac{h_V^{d-1}}{d-1}\int_{B(\boldsymbol{z},h_V)}|\boldsymbol{z}-\boldsymbol{y}|^{1-d}\mathrm{d}\boldsymbol{y}$$
$$\leq \frac{h_V^{d-1}}{d-1}\omega_d\int_0^{h_V}\rho^{1-d}\rho^{d-1}\mathrm{d}\rho \leq \frac{h_V^d}{d-1}\omega_d. \tag{7.74}$$

The proof is complete by plugging this estimate into (7.72), by using the resulting inequality in (7.71), and by recalling that

$$|B| = |B(0,1)|\left(\frac{\mathrm{diam}(B)}{2}\right)^d.$$

Note that in the case $d=1$, (7.73) is modified and involves $\ln(\frac{h}{|\boldsymbol{z}-\boldsymbol{y}|})$ but the final estimate (7.74) is still in $\mathcal{O}(h^d)$. ■

The following lemma is a simple technical result used in Lemma 7.61 below.

Lemma 7.60. *Let $h>0$, $d\in\mathbb{N}^\star$, $\boldsymbol{x}\in\mathbb{R}^d$ and let us define the function $F_{\boldsymbol{x},h}\ :\ B(\boldsymbol{x},h)\to\mathbb{R}$ by*

$$\forall \boldsymbol{z}\in B(\boldsymbol{x},h),\ F_{\boldsymbol{x},h}(\boldsymbol{z}) = \int_{\frac{|\boldsymbol{x}-\boldsymbol{z}|}{h}}^1 s^{1-d}\mathrm{d}s. \tag{7.75}$$

Let $q\in[1,+\infty]$ if $d=1$, $q\in[1,+\infty)$ if $d=2$, and $q\in[1,\frac{d}{d-2})$ if $d\geq 3$. Then, there exists $C_{15}>0$ depending only on d and q such that

$$\|F_{\boldsymbol{x},h}\|_{L^q(B(\boldsymbol{x},h))} \leq C_{15}h^{d/q}. \tag{7.76}$$

Proof.
Case $d=1$.
We have $|F_{\boldsymbol{x},h}(\boldsymbol{z})|\leq 1$ and therefore (7.76) is satisfied with $C_{15}=2^{1/q}$.

Case $d=2$.
We have $F_{\boldsymbol{x},h}(\boldsymbol{z}) = \ln\left(\frac{h}{|\boldsymbol{x}-\boldsymbol{z}|}\right)$ and therefore, since $q<+\infty$, using a polar change of coordinates,

$$\|F_{\boldsymbol{x},h}\|^q_{L^q(B(\boldsymbol{x},h))} = 2\pi\int_0^h \rho\ln\left(\frac{h}{\rho}\right)^q\mathrm{d}\rho.$$

The function $\rho\mapsto\rho\ln\left(\frac{h}{\rho}\right)^q$ reaches its maximum over $[0,h]$ at $\rho=e^{-q}h$ and thus

$$\|F_{\boldsymbol{x},h}\|^q_{L^q(B(\boldsymbol{x},h))} \leq 2\pi\int_0^h e^{-q}hq^q\mathrm{d}\rho = q^qe^{-q}h^2.$$

This proves (7.76) with $C_{15}=(2\pi)^{1/q}qe^{-1}$.

Case $d\geq 3$.
We write

$$F_{\boldsymbol{x},h}(\boldsymbol{z}) = \frac{1}{d-2}\left[\left(\frac{h}{|\boldsymbol{x}-\boldsymbol{z}|}\right)^{d-2} - 1\right] \le \frac{1}{d-2}\left(\frac{h}{|\boldsymbol{x}-\boldsymbol{z}|}\right)^{d-2}$$

and, using again polar coordinates,

$$\|F_{\boldsymbol{x},h}\|^q_{L^q(B(\boldsymbol{x},h))} \le \frac{\omega_d}{(d-2)^q} h^{(d-2)q} \int_0^h \rho^{(2-d)q+d-1} \mathrm{d}\rho$$

where ω_d is the area of the unit sphere in $\mathbb{R}^d$. The assumption $q < \frac{d}{d-2}$ ensures that $(2-d)q + d - 1 > -1$ and therefore

$$\|F_{\boldsymbol{x},h}\|^q_{L^q(B(\boldsymbol{x},h))} \le \frac{\omega_d}{(d-2)^q((2-d)q+d)} h^d.$$

The proof is complete by choosing $C_{15} = \frac{\omega_d^{1/q}}{(d-2)[(2-d)q+d]^{1/q}}$. ∎

The following lemma states, for a given $\varphi \in W^{2,p}(V)$ with $p > d/2$, the existence of an affine function that approximates φ on V. This approximation is constructed in the spirit of the "averaged Taylor polynomial" of [32].

Lemma 7.61 (First order polynomial approximation of elements of $W^{2,p}$). *Assume that $p > \frac{d}{2}$, and let $B \subset V$ be bounded subsets of $\mathbb{R}^d$, such that B is a ball and V is star-shaped with respect to all points of B. Let $\theta \ge \mathrm{diam}(V)/\mathrm{diam}(B)$.*
Take $\varphi \in W^{2,p}(V) \cap C(\overline{V})$. Then there exists $C_{16} > 0$, depending only on d, p and θ, and an affine function $A_\varphi : V \to \mathbb{R}$ such that

$$\sup_{\boldsymbol{x}\in\overline{V}} |\varphi(\boldsymbol{x}) - A_\varphi(\boldsymbol{x})| \le C_{16}\mathrm{diam}(V)^{2-\frac{d}{p}} \left\| |D^2\varphi| \right\|_{L^p(V)}, \tag{7.77}$$

$$\|\nabla A_\varphi - \nabla\varphi\|_{L^p(V)^d} \le C_{16}\mathrm{diam}(V) \left\| |D^2\varphi| \right\|_{L^p(V)}. \tag{7.78}$$

Remark 7.62. If V is sufficiently regular, $W^{2,p}(V) \subset C(\overline{V})$ and we only need to assume that $\varphi \in W^{2,p}(V)$.

Proof. To simplify the notations, set $h_B = \mathrm{diam}(B)$ and $h_V = \mathrm{diam}(V)$. Let us first assume that $\varphi \in C^2_c(\mathbb{R}^d)$. For a given $\boldsymbol{x} \in \overline{V}$ and any $\boldsymbol{y} \in B$, write the Taylor expansion

$$\varphi(\boldsymbol{x}) = \varphi(\boldsymbol{y}) + \nabla\varphi(\boldsymbol{y})\cdot(\boldsymbol{x}-\boldsymbol{y}) + \int_0^1 sD^2\varphi(\boldsymbol{x}+s(\boldsymbol{y}-\boldsymbol{x}))(\boldsymbol{x}-\boldsymbol{y})\cdot(\boldsymbol{x}-\boldsymbol{y})\mathrm{d}s. \tag{7.79}$$

Denote by $\overline{\boldsymbol{y}}$ the centre of B, and set $\overline{\varphi} = \frac{1}{|B|}\int_B \varphi(\boldsymbol{y})\mathrm{d}\boldsymbol{y}$ and $\overline{\nabla\varphi} = \frac{1}{|B|}\int_B \nabla\varphi(\boldsymbol{y})\mathrm{d}\boldsymbol{y}$. Taking the average of (7.79) over $\boldsymbol{y} \in B$ gives $\varphi(\boldsymbol{x}) = A_\varphi(\boldsymbol{x}) + R_1(\boldsymbol{x}) + R_2(\boldsymbol{x})$ with

$$A_\varphi(\boldsymbol{x}) = \overline{\varphi} + \overline{\nabla\varphi}\cdot(\boldsymbol{x}-\overline{\boldsymbol{y}}),$$

$$R_1(\boldsymbol{x}) = \frac{1}{|B|}\int_B\int_0^1 sD^2\varphi(\boldsymbol{x}+s(\boldsymbol{y}-\boldsymbol{x}))(\boldsymbol{x}-\boldsymbol{y})\cdot(\boldsymbol{x}-\boldsymbol{y})\mathrm{d}s\mathrm{d}\boldsymbol{y},$$

and

$$R_2(\boldsymbol{x}) = \frac{1}{|B|}\int_B(\nabla\varphi(\boldsymbol{y})-\overline{\nabla\varphi})\cdot(\boldsymbol{x}-\boldsymbol{y})\mathrm{d}\boldsymbol{y}.$$

Hence,

$$|\varphi(\boldsymbol{x})-A_\varphi(\boldsymbol{x})| \le |R_1(\boldsymbol{x})|+|R_2(\boldsymbol{x})|. \tag{7.80}$$

It remains to bound R_1 and R_2.

Bound on R_1.
The change of variable $\boldsymbol{y}\in B\to\boldsymbol{z}=\boldsymbol{x}+s(\boldsymbol{y}-\boldsymbol{x})$ has values in V, since V is star-shaped with respect to all points in B. This gives

$$|R_1(\boldsymbol{x})| \le \frac{h_V^2}{|B|}\int_V\int_{I(\boldsymbol{x},\boldsymbol{z})} s^{1-d}|D^2\varphi(\boldsymbol{z})|\mathrm{d}s\mathrm{d}\boldsymbol{z},$$

where $I(\boldsymbol{x},\boldsymbol{z})=\{s\in(0,1)\,:\,\exists\boldsymbol{y}\in B\,,\ \boldsymbol{z}=\boldsymbol{x}+s(\boldsymbol{y}-\boldsymbol{x})\}$. If $s\in I(\boldsymbol{x},\boldsymbol{z})$ then $|\boldsymbol{z}-\boldsymbol{x}|=s|\boldsymbol{y}-\boldsymbol{x}|\le sh_V$ for some $\boldsymbol{y}\in B$, and thus $s\ge\frac{|\boldsymbol{z}-\boldsymbol{x}|}{h_V}$. Hence,

$$|R_1(\boldsymbol{x})| \le \frac{h_V^2}{|B|}\int_V|D^2\varphi(\boldsymbol{z})|\int_{\frac{|\boldsymbol{x}-\boldsymbol{z}|}{h_V}}^1 s^{1-d}\mathrm{d}s\mathrm{d}\boldsymbol{z} = \frac{h_V^2}{|B|}\int_V|D^2\varphi(\boldsymbol{z})|F_{\boldsymbol{x},h_V}(\boldsymbol{z})\mathrm{d}\boldsymbol{z}$$

where $F_{\boldsymbol{x},h_V}$ is defined by (7.75). Using Hölder's inequality, the inclusion $V\subset B(\boldsymbol{x},h_V)$ and Lemma 7.60 we infer

$$|R_1(\boldsymbol{x})| \le \frac{h_V^2}{|B|}\left\|\,|D^2\varphi|\,\right\|_{L^p(V)}\|F_{\boldsymbol{x},h_V}\|_{L^{p'}(B(\boldsymbol{x},h_V))} \le C_{17}\frac{h_V^{2+\frac{d}{p'}}}{|B|}\left\|\,|D^2\varphi|\,\right\|_{L^p(V)}$$

where C_{17} depends only on d and p (notice that $p>d/2$ implies $p'<\frac{d}{d-2}$ if $d\ge 2$). Since $\frac{d}{p'}=d-\frac{d}{p}$ and $|B|=|B(0,1)|h_B^d\ge|B(0,1)|\theta^{-d}h_V^d$, this gives the existence of C_{18} depending only on θ, p and d such that

$$|R_1(\boldsymbol{x})| \le C_{18}h_V^{2-\frac{d}{p}}\left\|\,|D^2\varphi|\,\right\|_{L^p(V)}. \tag{7.81}$$

Bound on R_2.
By Hölder's inequality and $|B|=|B(0,1)|h_B^d$,

$$\begin{aligned}|R_2(\boldsymbol{x})| &\le h_B|B|^{\frac{1}{p'}-1}\left\|\nabla\varphi-\overline{\nabla\varphi}\right\|_{L^p(B)^d}\\ &\le |B(0,1)|^{-\frac{1}{p}}h_B^{1-\frac{d}{p}}\left\|\nabla\varphi-\overline{\nabla\varphi}\right\|_{L^p(B)^d}.\end{aligned}$$

Apply Lemma 7.59 with $V=B$ and φ replaced by $\partial_i\varphi$ (for $i=1,\dots,d$). This gives C_{19} depending only on d and p such that

$$|R_2(\boldsymbol{x})| \leq C_{19} h_B^{2-\frac{d}{p}} \left\| |D^2\varphi| \right\|_{L^p(B)}. \tag{7.82}$$

Conclusion.
Combining (7.80), (7.81) and (7.82) gives (7.77). To prove (7.78), notice that

$$\nabla A_\varphi = \overline{\nabla\varphi} = \frac{1}{|B|}\int_B \nabla\varphi(\boldsymbol{y})\mathrm{d}\boldsymbol{y}$$

and apply Lemma 7.59 with φ replaced by $\partial_i\varphi$, for all $i = 1, \ldots, d$. This gives C_{20} depending only on d an p such that

$$\|\nabla A_\varphi - \nabla\varphi\|_{L^p(V)^d} \leq C_{20}\frac{h_V^{d/p+1}}{h_B^{d/p}} \left\| |D^2\varphi| \right\|_{L^p(V)}.$$

Since $h_B \geq \theta^{-1}h_V$, this completes the proof of (7.78) if $\varphi \in C_c^2(\mathbb{R}^d)$.

All quantities and norms involved in (7.77) and (7.78) are continuous with respect to φ for the norm of $W^{2,p}(V) \cap C(\overline{V})$ (sum of the norms in both spaces). Since V is star-shaped, a classical dilatation and regularisation argument shows that the restrictions of $C_c^2(\mathbb{R}^d)$ functions to V are dense in $W^{2,p}(V) \cap C(\overline{V})$. This density ensures that (7.77) and (7.78) are still valid for $\varphi \in W^{2,p}(V) \cap C(\overline{V})$. ∎

We can now state and prove a local $W^{2,p}$ interpolation estimate for $\mathbb{P}_1$-exact gradient reconstructions.

Lemma 7.63 ($W^{2,p}$ estimates for $\mathbb{P}_1$-exact gradient reconstructions). *Assume that $p > \frac{d}{2}$, and let $B \subset K \subset V$ be bounded subsets of $\mathbb{R}^d$, such that B is a ball and V is star-shaped with respect to all points of B. Let $S = (\boldsymbol{x}_i)_{i\in I} \subset \overline{V}$, and $\mathcal{G} = (\mathcal{G}^i)_{i\in I} \subset L^p(K)^d$ be a $\mathbb{P}_1$-exact gradient reconstruction on K upon S in the sense of Definition 7.28. Let $\theta \geq \mathrm{diam}(V)/\mathrm{diam}(B)$. Take $\varphi \in W^{2,p}(V) \cap C(\overline{V})$ and set $v = (\varphi(\boldsymbol{x}_i))_{i\in I}$. Then, there exists $C_{21} > 0$, depending only on d, p and θ such that*

$$\|\mathcal{G}v - \nabla\varphi\|_{L^p(K)^d} \leq C_{21}\mathrm{diam}(V)(1 + \|\mathcal{G}\|_p) \left\| |D^2\varphi| \right\|_{L^p(V)}. \tag{7.83}$$

Proof. Take A_φ given by Lemma 7.61 and define $\xi = (A_\varphi(\boldsymbol{x}_i))_{i\in I}$. Since $\mathcal{G}$ is a $\mathbb{P}_1$-exact gradient reconstruction upon $(\boldsymbol{x}_i)_{i\in I}$, we have $\mathcal{G}\xi = \nabla A_\varphi$. The definition of $\|\mathcal{G}\|_p$ show that

$$\begin{aligned}\|\mathcal{G}v - \mathcal{G}\xi\|_{L^p(K)^d} &= \left\| \sum_{i\in I}(v_i - \xi_i)\mathcal{G}^i \right\|_{L^p(K)^d} \\ &\leq \left\| \sum_{i\in I} |\mathcal{G}^i| \right\|_{L^p(K)} \max_{i\in I}|v_i - \xi_i|\end{aligned}$$

$$= \|\mathcal{G}\|_p |K|^{\frac{1}{p}} \mathrm{diam}(K)^{-1} \max_{i\in I} |\varphi(\boldsymbol{x}_i) - A_\varphi(\boldsymbol{x}_i)|.$$

Using (7.77) and the inequality $\mathrm{diam}(K)^{-1} \le h_B^{-1} \le \theta h_V^{-1}$, we deduce

$$\|\mathcal{G}v - \mathcal{G}\xi\|_{L^p(K)^d} \le C_{16} \|\mathcal{G}\|_p \theta h_V^{1-\frac{d}{p}} |K|^{\frac{1}{p}} \left\| |D^2\varphi| \right\|_{L^p(V)}.$$

Since $|K| \le |B(0,1)|\mathrm{diam}(K)^d \le |B(0,1)|h_V^d$, this shows that there exists C_{22} depending only on θ, d and p such that

$$\|\mathcal{G}v - \mathcal{G}\xi\|_{L^p(K)^d} \le C_{22} \|\mathcal{G}\|_p h_V \left\| |D^2\varphi| \right\|_{L^p(V)}.$$

The proof of (7.83) is complete by recalling that $\mathcal{G}\xi = \nabla A_\varphi$, by using the triangle inequality, and by invoking (7.78). ∎

7.4.2 Application to local linearly exact GDs

The next proposition states our main bound on $S_{\mathcal{D}}(\varphi)$ for an LLE GD, in the case of homogeneous Dirichlet boundary conditions and $\varphi \in W^{2,p}(\Omega)$ with $p > d/2$. This is established under a slightly restrictive assumption on the points $\boldsymbol{x}_i$, which holds for most of the LLE GDs presented in Chapters 8–14.

Proposition 7.64 ($W^{2,p}$ estimates of $S_{\mathcal{D}}$ for an LLE GD). *Take $p > d/2$ and let $\mathcal{D}$ be an LLE GD in the sense of Definition 7.33. Let $S = (\boldsymbol{x}_i)_{i\in I}$ be the family of approximation points of $\mathcal{D}$, and $\mathcal{M}$ be the mesh associated with $\mathcal{D}$. Assume that*

(i) For all $K \in \mathcal{M}$ and all $i \in I_K$, $\boldsymbol{x}_i \in \overline{K}$,
(ii) For all $K \in \mathcal{M}$, there exists a ball $B_K \subset K$ such that K is star-shaped with respect to all points in B_K. (7.84)

Take $\theta \ge \mathrm{reg}_{\mathrm{LLE}}(\mathcal{D}) + \max_{K\in\mathcal{M}} \frac{\mathrm{diam}(K)}{\mathrm{diam}(B_K)}$. Then, there exists $C_{23} > 0$, depending only on p, d, Ω and θ, such that

$$\forall \varphi \in W^{2,p}(\Omega) \cap W_0^{1,p}(\Omega)\,,\ S_{\mathcal{D}}(\varphi) \le C_{23} h_{\mathcal{M}} \|\varphi\|_{W^{2,p}(\Omega)}\,, \tag{7.85}$$

where $S_{\mathcal{D}}$ is defined by (2.2).

Proof. Remember that, in all this part, Ω is assumed to have a Lipschitz-continuous boundary. Hence, the choice of p ensure that $\varphi \in C(\overline{\Omega})$. The vector $v = (\varphi(\boldsymbol{x}_i))_{i\in I} \in X_{\mathcal{D},0}$ is therefore well defined, and Lemma 7.63 can be applied, for any $K \in \mathcal{M}$, with $V = K$ and $\mathcal{G} = \mathcal{G}_K$. Estimate (7.83) then yields, with $h_K = \mathrm{diam}(K)$,

$$\|\nabla_{\mathcal{D}} v - \nabla\varphi\|_{L^p(K)^d} \le C_{21} h_K (1 + \|\mathcal{G}_K\|_p) \left\| |D^2\varphi| \right\|_{L^p(K)} \tag{7.86}$$

where C_{21} depends only on p, d and θ. Raising to the power p and summing over $K \in \mathcal{M}$ leads to

$$\|\nabla_{\mathcal{D}} v - \nabla\varphi\|_{L^p(\Omega)^d} \le C_{21} h_{\mathcal{M}}(1 + \mathrm{reg}_{\mathrm{LLE}}(\mathcal{D})) \left\| |D^2\varphi| \right\|_{L^p(\Omega)}. \tag{7.87}$$

To estimate $\Pi_{\mathcal{D}} v - \varphi$, we first establish a bound on $\varphi(\boldsymbol{x}) - \varphi(\boldsymbol{y})$ for all $\boldsymbol{x}, \boldsymbol{y} \in K$. Using the affine function A_φ given by Lemma 7.61, write

$$\begin{aligned} |\varphi(\boldsymbol{x}) - \varphi(\boldsymbol{y})| &\le |\varphi(\boldsymbol{x}) - A_\varphi(\boldsymbol{x})| + |A_\varphi(\boldsymbol{x}) - A_\varphi(\boldsymbol{y})| + |A_\varphi(\boldsymbol{y}) - \varphi(\boldsymbol{y})| \\ &\le 2C_{16} h_K^{2-\frac{d}{p}} \left\| |D^2\varphi| \right\|_{L^p(K)} + |\nabla A_\varphi| h_K. \end{aligned} \tag{7.88}$$

Since ∇A_φ is constant, (7.78) gives

$$\begin{aligned} |\nabla A_\varphi| &= |K|^{-\frac{1}{p}} \|\nabla A_\varphi\|_{L^p(K)^d} \\ &\le |K|^{-\frac{1}{p}} \|\nabla\varphi\|_{L^p(K)^d} + |K|^{-\frac{1}{p}} C_{16} h_K \left\| |D^2\varphi| \right\|_{L^p(K)}. \end{aligned}$$

Plugged into (7.88), this yields

$$\begin{aligned} &|\varphi(\boldsymbol{x}) - \varphi(\boldsymbol{y})| \\ &\quad \le \left(2C_{16}\mathrm{diam}(\Omega) h_K^{1-\frac{d}{p}} + (1 + C_{16}\mathrm{diam}(\Omega)) h_K |K|^{-\frac{1}{p}} \right) \|\varphi\|_{W^{2,p}(K)}. \end{aligned}$$

Since $K \subset B(\boldsymbol{z}, h_K)$ for all $\boldsymbol{z} \in K$, we have

$$|K| \le |B(\boldsymbol{z}, h_K)| = |B(0,1)| 2^{-d} h_K^d.$$

Combined with the previous inequality, this provides C_{24} depending only on Ω, p and θ such that

$$|\varphi(\boldsymbol{x}) - \varphi(\boldsymbol{y})| \le C_{24} h_K |K|^{-\frac{1}{p}} \|\varphi\|_{W^{2,p}(K)}. \tag{7.89}$$

Recalling the relation (7.32) between $\Pi_{\mathcal{D}}$ and the elementary basis functions $(\pi_K^i)_{i \in I_K}$, (7.89) gives, for a.e. $\boldsymbol{x} \in K$,

$$\begin{aligned} |\Pi_{\mathcal{D}} v(\boldsymbol{x}) - \varphi(\boldsymbol{x})| &= \left| \sum_{i \in I_K} \pi_K^i(\boldsymbol{x})(v_i - \varphi(\boldsymbol{x})) \right| \\ &\le \sup_{i \in I_K} |\varphi(\boldsymbol{x}_i) - \varphi(\boldsymbol{x})| \sum_{i \in I_K} |\pi_K^i(\boldsymbol{x})| \\ &\le C_{24} h_K \|\varphi\|_{W^{2,p}(K)} |K|^{-\frac{1}{p}} \sum_{i \in I_K} |\pi_K^i(\boldsymbol{x})|. \end{aligned}$$

Take the $L^p(K)$ norm over $\boldsymbol{x} \in K$ and recall the definition (7.21) of $\|\pi_K\|_p$ to deduce

$$\|\Pi_{\mathcal{D}} v - \varphi\|_{L^p(K)} \le C_{24} h_K \|\varphi\|_{W^{2,p}(K)} \|\pi_K\|_p. \tag{7.90}$$

As (7.86), this estimate in K translates into the global estimate

$$\|\Pi_{\mathcal{D}} v - \varphi\|_{L^p(\Omega)} \leq C_{24} h_{\mathcal{M}} \|\varphi\|_{W^{2,p}(\Omega)} \operatorname{reg}_{\mathrm{LLE}}(\mathcal{D}). \tag{7.91}$$

The proof is complete by combining (7.87) and (7.91). ■

Remark 7.65 (Broken $W^{2,p}$ estimates)
An inspection of the proof shows that Proposition 7.64 also holds if we only assume that $\varphi \in C(\overline{\Omega}) \cap W_0^{1,p}(\Omega) \cap W^{2,p}(\mathcal{M})$, where the broken space $W^{2,p}(\mathcal{M})$ is defined by

$$W^{2,p}(\mathcal{M}) = \{\psi \in L^p(\Omega) \,:\, \forall K \in \mathcal{M}\,,\ \psi_{|K} \in W^{2,p}(K)\}.$$

We just have to replace, in (7.85), the term "$h_{\mathcal{M}} \|\varphi\|_{W^{2,p}(\Omega)}$" with

$$\left(\sum_{K \in \mathcal{M}} h_K^p \|\varphi_{|K}\|_{W^{2,p}(K)}^p \right)^{1/p}.$$

Assumption (7.84) ensures that local error estimates can be computed on a *mesh* of the domain (with non-overlapping sets). When added together, the right-hand sides of these estimates directly produce an $L^p(\Omega)$ norm. We can relax this assumption of non-overlapping sets if we impose a control on the overlaps. The following result makes this broad reasoning explicit, and is required to establish $W^{2,p}$ estimates for some methods described in subsequent chapters, noticeably the condensed version of the SUSHI scheme and the VAG scheme, and any other barycentric condensation of an LLE GD, when some discrete unknowns in I_K are eliminated by using other discrete unknowns that lie outside K, see Definition 7.40).

Proposition 7.66 ($W^{2,p}$ estimates of $S_{\mathcal{D}}$ for an LLE GD – generalised form). *Assume that $p > d/2$ and that $\mathcal{D}$ is an LLE GD in the sense of Definition 7.33. Let $S = (\boldsymbol{x}_i)_{i \in I}$ be the family of approximation points of $\mathcal{D}$, and $\mathcal{M}$ be the mesh associated with $\mathcal{D}$.*
For each $K \in \mathcal{M}$, take $V_K \supset K$ a bounded set such that

(i) For all $i \in I_K$, $\boldsymbol{x}_i \in \overline{V_K}$,
(ii) There exists a ball $B_K \subset K$ such that V_K is star-shaped with respect to all points of B_K. (7.92)

Let

$$\theta \geq \operatorname{reg}_{\mathrm{LLE}}(\mathcal{D}) + \max_{K \in \mathcal{M}} \frac{\operatorname{diam}(V_K)}{\operatorname{diam}(B_K)} + \operatorname*{esssup}_{\boldsymbol{x} \in \Omega} \operatorname{Card}(\{K \in \mathcal{M} \,:\, \boldsymbol{x} \in V_K\}). \tag{7.93}$$

Then, there exists $C_{25} > 0$, depending only on p, d, Ω and θ, such that

$$\forall \varphi \in W^{2,p}(\Omega) \cap W_0^{1,p}(\Omega)\,,\ S_{\mathcal{D}}(\varphi) \leq C_{25} h_{\mathcal{M}} \|\varphi\|_{W^{2,p}(\Omega)}\,,$$

where $S_{\mathcal{D}}$ is defined by (2.2).

Remark 7.67. Imposing that $\theta \geq \operatorname{esssup}_{\boldsymbol{x}\in\Omega} \operatorname{Card}(\{K \in \mathcal{M} : \boldsymbol{x} \in V_K\})$ is equivalent to imposing that, almost everywhere on Ω, at most θ sets $(V_K)_{K\in\mathcal{M}}$ overlap.

Proof. Introduce the same $v \in X_{\mathcal{D},0}$ as in the proof of Proposition 7.64 and use Lemma 7.63 with $V = V_K$ to arrive, in a similar way as for (7.86) and (7.90), to

$$\|\Pi_{\mathcal{D}} v - \varphi\|_{L^p(K)} + \|\nabla_{\mathcal{D}} v - \nabla\varphi\|_{L^p(K)^d} \leq C_{26} \operatorname{diam}(V_K) \|\varphi\|_{W^{2,p}(V_K)},$$

where C_{26} depends only on p, d, Ω and θ. Since $\operatorname{diam}(V_K) \leq \theta \operatorname{diam}(B_K) \leq \theta h_{\mathcal{M}}$, raising to the power p gives C_{27} depending only on p, d, Ω and θ such that

$$\|\Pi_{\mathcal{D}} v - \varphi\|^p_{L^p(K)} + \|\nabla_{\mathcal{D}} v - \nabla\varphi\|^p_{L^p(K)^d} \leq C_{27} h^p_{\mathcal{M}} \|\varphi\|^p_{W^{2,p}(V_K)}.$$

Summing over $K \in \mathcal{M}$ yields

$$\|\Pi_{\mathcal{D}} v - \varphi\|^p_{L^p(\Omega)} + \|\nabla_{\mathcal{D}} v - \nabla\varphi\|^p_{L^p(\Omega)^d} \leq C_{27}^p h^p_{\mathcal{M}} \sum_{K\in\mathcal{M}} \|\varphi\|^p_{W^{2,p}(V_K)}. \quad (7.94)$$

We now estimate the sum in this inequality. By the Fubini–Tonelli relation and letting $\mathbf{1}_{V_K}$ be the characteristic function of V_K, for any $g \in L^p(\Omega)$,

$$\begin{aligned}\sum_{K\in\mathcal{M}} \|g\|^p_{L^p(V_K)} &= \sum_{K\in\mathcal{M}} \int_\Omega \mathbf{1}_{V_K}(\boldsymbol{x}) |g(\boldsymbol{x})|^p \mathrm{d}\boldsymbol{x} \\ &= \int_\Omega |g(\boldsymbol{x})|^p \left(\sum_{K\in\mathcal{M}} \mathbf{1}_{V_K}(\boldsymbol{x}) \right) \mathrm{d}\boldsymbol{x}.\end{aligned}$$

The choice of θ ensures that $\sum_{K\in\mathcal{M}} \mathbf{1}_{V_K}(\boldsymbol{x}) = \operatorname{Card}(\{K \in \mathcal{M} : \boldsymbol{x} \in V_K\}) \leq \theta$ for a.e. $\boldsymbol{x} \in \Omega$. Hence,

$$\sum_{K\in\mathcal{M}} \|g\|^p_{L^p(V_K)} \leq \theta \int_\Omega |g|^p \mathrm{d}\boldsymbol{x} = \theta \|g\|^p_{L^p(\Omega)}.$$

The proof is complete by using this estimate in (7.94) with $g = \varphi$, $g = |\nabla\varphi|$ and $g = |D^2\varphi|$. ∎

We now turn to the adaptation of the previous results to other boundary conditions than homogeneous Dirichlet conditions.

Proposition 7.68 ($W^{2,p}$ estimates of $S_{\mathcal{D}}$ for an LLE GD – non-homogeneous Dirichlet BCs). *Assume that $p > d/2$ and that $\mathcal{D}$ is an LLE GD in the sense of Definition 7.52. Let $S = (\boldsymbol{x}_i)_{i\in I}$ be the family of approximation points of $\mathcal{D}$, and $\mathcal{M}$ be the mesh associated with $\mathcal{D}$. Assume that (7.84) holds and take $\theta \geq \operatorname{reg}_{\mathrm{LLE}}(\mathcal{D}) + \max_{K\in\mathcal{M}} \frac{\operatorname{diam}(K)}{\operatorname{diam}(B_K)}$. Take $\varphi \in W^{2,p}(\Omega)$ and assume that*

$$\forall i \in I_\partial\,,\ (\mathcal{I}_{\mathcal{D},\partial}\gamma(\varphi))_i = \varphi(\boldsymbol{x}_i). \tag{7.95}$$

Then, there exists $C_{28} > 0$, depending only on p, d, Ω and θ, such that

$$S_{\mathcal{D}}(\varphi) \le C_{28} h_{\mathcal{M}} \|\varphi\|_{W^{2,p}(\Omega)}\,, \tag{7.96}$$

where $S_{\mathcal{D}}$ is defined by (2.94).

Proof. Assumption (7.95) ensures that the vector $v = (\varphi(\boldsymbol{x}_i))_{i\in I} \in X_{\mathcal{D}}$ satisfies $v - \mathcal{I}_{\mathcal{D},\partial}\gamma(\varphi) \in X_{\mathcal{D},0}$. This vector is therefore suited to the definition (2.94) of $S_{\mathcal{D}}$. Since v satisfies the estimates (7.87) and (7.91) (which have been established without using the boundary value of φ), this completes the proof. ∎

Proposition 7.69 ($W^{2,p}$ estimates on $S_{\mathcal{D}}$ for an LLE GD – non-homogeneous Dirichlet BCs and relaxed assumption on $\mathcal{I}_{\mathcal{D},\partial}$). *Make the same assumptions as in Proposition 7.68, except (7.95) which is replaced by*

$$\begin{aligned}&\forall K \in \mathcal{M}\,,\ \textit{there exists } C_K(\varphi) \ge 0 \textit{ s.t.}\\ &\max_{i\in I_K\cap I_\partial} |(\mathcal{I}_{\mathcal{D},\partial}\gamma\varphi)_i - \varphi(\boldsymbol{x}_i)| \le h_{\mathcal{M}}\mathrm{diam}(K)|K|^{-\frac{1}{p}} C_K(\varphi).\end{aligned} \tag{7.97}$$

Then, there exists C_{29} depending only on p, d, Ω, and θ, such that

$$S_{\mathcal{D}}(\varphi) \le C_{29} h_{\mathcal{M}} \left(\|\varphi\|_{W^{2,p}(\Omega)} + \left(\sum_{K\in\mathcal{M}} C_K(\varphi)^p \right)^{1/p} \right). \tag{7.98}$$

By convention $\max_{i\in\emptyset} |Z_i| = 0$ and the quantity $C_K(\varphi)$ can thus be set to 0 if K is an interior cell (that is, $I_K \cap I_\partial = \emptyset$). For a general K, $C_K(\varphi)$ would usually be the norm on K (or a lower dimensional subset of $\overline{K}$) of some derivatives of φ, and the quantity $\sum_{K\in\mathcal{M}} C_K(\varphi)^p$ would be bounded by some constant depending only on φ (not on $\mathcal{M}$). Notice however that, in practical situations, the regularity imposed on φ in Proposition 7.69 is such that $\mathcal{I}_{\mathcal{D},\partial}\gamma(\varphi)$ is usually re-defined so that (7.95) holds. See Remarks 2.50 and 13.4.

Proof. The estimates established in the proof of Proposition 7.64 are independent of the boundary conditions. Hence, if $v = (\varphi(\boldsymbol{x}_i))_{i\in I} \in X_{\mathcal{D}}$ is defined as in that proof,

$$\|\Pi_{\mathcal{D}} v - \varphi\|_{L^p(\Omega)} + \|\nabla_{\mathcal{D}} v - \nabla\varphi\|_{L^p(\Omega)^d} \le C_{30} h_{\mathcal{M}} \|\varphi\|_{W^{2,p}(\Omega)}\,, \tag{7.99}$$

where C_{30} depends only on d, p, Ω and θ.
Let us now consider $w \in X_{\mathcal{D}}$ as in the proof of Proposition 7.53, that is $w_i = v_i$ if $i \in I_\Omega$ and $w_i = (\mathcal{I}_{\mathcal{D}\partial}\gamma(\varphi))_i$ if $i \in I_\partial$. By (7.97) the quantity $\omega(K)$ defined by (7.66) satisfies $\omega(K) \le h_{\mathcal{M}}|K|^{-\frac{1}{p}} C_K(\varphi)$. Plug this estimate into (7.67), raise the result to the power p and sum over $K \in \mathcal{M}$. This gives

$$\|\nabla_{\mathcal{D}} v - \nabla_{\mathcal{D}} w\|_{L^p(\Omega)^d} \leq \theta h_{\mathcal{M}} C_\Omega(\varphi), \tag{7.100}$$

where $C_\Omega(\varphi) = (\sum_{K\in\mathcal{M}} C_K(\varphi)^p)^{1/p}$. The term $\Pi_{\mathcal{D}} v - \Pi_{\mathcal{D}} w$ is estimated similarly. For $K \in \mathcal{M}$ and a.e. $\boldsymbol{x} \in K$,

$$\begin{aligned}|\Pi_{\mathcal{D}} v(\boldsymbol{x}) - \Pi_{\mathcal{D}} w(\boldsymbol{x})| &\leq \sum_{i\in I_K} |\pi_K^i(\boldsymbol{x})|\,|v_i - w_i| \\ &\leq h_{\mathcal{M}} \mathrm{diam}(K) C_K(\varphi) |K|^{-\frac{1}{p}} \sum_{i\in I_K} |\pi_K^i(\boldsymbol{x})|.\end{aligned}$$

Taking the $L^p(K)$ norm, recalling the definition (7.21) of $\|\pi_K\|_p$, raising to the power p and summing on $K \in \mathcal{M}$ leads to $\|\Pi_{\mathcal{D}} v - \Pi_{\mathcal{D}} w\|_{L^p(\Omega)} \leq \theta h_{\mathcal{M}}^2 C_\Omega(\varphi)$. The proof is complete by combining this estimate with (7.100) and (7.99).

■

Since the estimates (7.87) and (7.91) were obtained without referring to the boundary values of φ, they immediately give the following result.

Proposition 7.70 ($W^{2,p}$ estimates of $S_{\mathcal{D}}$ for an LLE GD – Neumann BCs). *Assume that $p > d/2$ and that $\mathcal{D}$ is an LLE GD in the sense of Definition 7.54. Let $S = (\boldsymbol{x}_i)_{i\in I}$ be the family of approximation points of $\mathcal{D}$, and $\mathcal{M}$ be the mesh associated with $\mathcal{D}$. Assume that (7.84) holds and take $\theta \geq \mathrm{reg}_{\mathrm{LLE}}(\mathcal{D}) + \max_{K\in\mathcal{M}} \frac{\mathrm{diam}(K)}{\mathrm{diam}(B_K)}$. Then, there exists C, depending only on p, d, Ω and θ, such that*

$$\forall \varphi \in W^{2,p}(\Omega)\,,\ S_{\mathcal{D}}(\varphi) \leq C h_{\mathcal{M}} \|\varphi\|_{W^{2,p}(\Omega)}\,,$$

where $S_{\mathcal{D}}$ is defined by (3.3).

The $W^{2,p}$ estimates on $S_{\mathcal{D}}$ for Fourier boundary conditions are notably harder to establish than for the other boundary conditions, since the trace reconstruction $\mathbb{T}_{\mathcal{D}}$ also needs to be handled. The issue is that this trace has values in a lower-dimensional space. If the mesh of $\partial\Omega$ is made of parts of hyperplanes (which is natural if Ω is a polytopal open set) and satisfies the equivalent of (7.84), then the estimates of $\mathbb{T}_{\mathcal{D}}$ can be obtained as the estimates on $\Pi_{\mathcal{D}}$ in the proof of Proposition 7.64.

Proposition 7.71 ($W^{2,p}$ estimates of $S_{\mathcal{D}}$ for an LLE GD – Fourier BCs). *Assume that $p > d/2$ and that $\mathcal{D}$ is an LLE GD in the sense of Definition 7.57. Let $S = (\boldsymbol{x}_i)_{i\in I}$ be the family of approximation points of $\mathcal{D}$, and $\mathcal{M}$ be the mesh associated with $\mathcal{D}$. Assume that (7.84) holds and, with $H_1, \ldots, H_r$ hyperplanes whose union covers $\partial\Omega$, that*

(i) For any $K_\partial \in \mathcal{M}_\partial$ there is $\ell_{K_\partial} \in \{1, \ldots, r\}$ such that $K_\partial \subset H_{\ell_{K_\partial}}$,

(ii) For all $K_\partial \in \mathcal{M}_\partial$ and all $i \in I_{K_\partial}$, $\boldsymbol{x}_i \in \overline{K_\partial}$,

(iii) For all $K_\partial \in \mathcal{M}_\partial$, there exists a ball $B_{K_\partial} \subset K_\partial$ in $H_{\ell_{K_\partial}}$ such that K_∂ is star-shaped with respect to all points of B_{K_∂}.

We take

$$\theta \geq \mathrm{reg}_{\mathrm{LLE}}(\mathcal{D}) + \max_{K\in\mathcal{M}} \frac{\mathrm{diam}(K)}{\mathrm{diam}(B_K)} + \max_{K_\partial\in\mathcal{M}_\partial} \frac{\mathrm{diam}(K_\partial)}{\mathrm{diam}(B_{K_\partial})}.$$

Then, there exists $C_{31} > 0$, depending only on p, d, Ω and θ, such that, for all $\varphi \in W^{2,p}(\Omega)$ satisfying $\gamma(\varphi) \in W^{2,p}(\partial\Omega \cap H_\ell)$ for all $\ell = 1, \ldots, r$,

$$S_{\mathcal{D}}(\varphi) \leq C_{31} h_{\mathcal{M}} \left(\|\varphi\|_{W^{2,p}(\Omega)} + \sum_{\ell=1}^{r} \|\gamma(\varphi)\|_{W^{2,p}(\partial\Omega\cap H_\ell)} \right),$$

where $S_{\mathcal{D}}$ is defined by (3.52).

7.5 Further topics on LLE GDs

7.5.1 LLE GDs with generalised discrete unknowns

The definition 7.28 of $\mathbb{P}_1$-exact gradient reconstructions implicitly assume that the discrete unknowns of the method correspond to the values of functions at given points in the domain (the approximation points S). Some numerical schemes, especially high order methods, use other kinds of discrete unknowns; for example, unknowns that represent moments of functions

$$\int_K \boldsymbol{x}^\alpha f(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}.$$

It is possible to write a more general definition of $\mathbb{P}_1$-exact gradient reconstruction to account for such generalised discrete unknowns. It also makes sense to generalise the definition to higher order reconstructions, as in Proposition 7.38.

Definition 7.72 ($\mathbb{P}_{k+1}$-exact gradient reconstruction with generalised discrete unknowns). *Let K be a bounded subset of $\mathbb{R}^d$, $p \in [1, +\infty]$ and $k \in \mathbb{N}$. A $\mathbb{P}_{k+1}$-exact gradient reconstruction on K with generalised discrete unknowns is $(P, \mathcal{G})$ where:*

- *$P = (P_i)_{i\in I}$ is a finite family of linear mappings $P_i : C^k(\overline{K}) \to \mathbb{R}$,*
- *$\mathcal{G} = (\mathcal{G}^i)_{i\in I}$ is a family of functions in $L^p(K)^d$ such that, for any polynomial function q of degree $k+1$ or less,*

$$\sum_{i\in I} P_i(q)\mathcal{G}^i = \nabla q \text{ on } K.$$

The norm of $(P, \mathcal{G})$ is defined by

$$\|(P,\mathcal{G})\|_p = \mathrm{diam}(K)|K|^{-\frac{1}{p}} \left\| \sum_{i\in I} \|P_i\|_{(C^k)'} |\mathcal{G}^i| \right\|_{L^p(K)^d},$$

where

$$\|P_i\|_{(C^k)'} = \max_{w \in C^k(\overline{K})\setminus\{0\}} \frac{|P_i(w)|}{\|w\|_{C^k(\overline{K})}}.$$

The $\mathbb{P}_1$-exact gradient reconstruction of Definition 7.28 corresponds to $P_i(\varphi) = \varphi(\boldsymbol{x}_i)$ and $k = 0$.

In a similar way as in Definition 7.72, these gradient reconstructions could be used to design a notion of "$\mathbb{P}_{k+1}$-exact GDs with generalised discrete unknowns" and perform most of the analysis done for LLE GDs (using $\|(P, \mathcal{G})\|_p$ instead of $\|\mathcal{G}\|_p$, and with adjustments in some spaces of functions – *e.g.*, in Lemma 7.31 we would work with $\varphi \in W^{k+2,\infty}(\mathbb{R}^d)$). We let the interested reader fill in the details.

7.5.2 Non-linearly exact barycentric condensation

Let us consider a heterogeneous material, with a discontinuous diffusion tensor Λ which is smooth inside subdomains $P_1, \ldots, P_k$ (partition of Ω). The solution to (2.20) is not expected to be smooth over Ω, but rather smooth (at least if we exclude the corners) inside each P_ℓ and with continuous fluxes at the interfaces $P_\ell \cap P_{\ell'}$. LLE GDs are adapted to such solutions provided that all approximation points $(\boldsymbol{x}_i)_{i \in I_K}$, for each $K \in \mathcal{M}$, lie in a single subdomain P_ℓ. Indeed, in this case, the gradient reconstruction $\mathcal{G}_K v$ from the interpolated values $v_i = \overline{u}(\boldsymbol{x}_i)$ of the solution will be a good approximation of $(\nabla \overline{u})_{|K}$ (Lemma 7.31).

When performing a barycentric condensation of an LLE GD, it is common that for some eliminated unknowns $i \in I \setminus I^{\text{BA}}$, the set of approximated points $(\boldsymbol{x}_j)_{j \in H_i}$ spreads over several subdomains P_ℓ, especially if $\boldsymbol{x}_i$ lies at or close to an interface between two such subdomains. It is then clear that a barycentric condensation that is an LLE GD is not the best choice to approximate $\overline{u}$: if we define $v_j = \overline{u}(\boldsymbol{x}_j)$ for all $j \in I^{\text{BA}}$ then, for the unknowns $i \in I \setminus I^{\text{BA}}$ such that H_i spread over several subdomains, the values $\widetilde{v}_i$ defined by (7.47) are no longer good (order $\operatorname{diam}(K)^2$) approximations of $\overline{u}(\boldsymbol{x}_i)$, and therefore $(\nabla_{\mathcal{D}^{\text{BA}}} v)_{|K} = (\nabla_{\mathcal{D}} \widetilde{v})_{|K}$ will not approximate $(\nabla \overline{u})_{|K}$ properly. This does not prevent the corresponding GS from converging, but leads to reduced accuracy on coarse meshes.

Barycentric condensation preserves the LLE property thanks to Assumption (7.46); it is this assumption that ensures that $\sum_{j \in H_i} \beta_j^i A(\boldsymbol{x}_j) = A(\boldsymbol{x}_i)$ for all affine function A. To deal with heterogeneous materials, it might be suitable to relax this assumption and create barycentric condensations that do not satisfy (7.46). Instead, we should aim for relations that ensure that, if v interpolates $\overline{u}$ at $(\boldsymbol{x}_j)_{j \in I^{\text{BA}}}$, the values $\widetilde{v}$ computed through (7.47) give good approximations of the values of $\overline{u}$ at $(\boldsymbol{x}_i)_{i \in I}$. This leads to the notion of $\mathcal{S}$-adapted barycentric condensation.

Definition 7.73 ($\mathcal{S}$-adapted barycentric condensation). *Let $\mathcal{D}$ be an LLE GD in the sense of Definition 7.33, and let $\mathcal{S}$ be a dense subset $W_0^{1,p}(\Omega)$*

such that $\mathcal{S} \subset C(\overline{\Omega})$. An $\mathcal{S}$-adapted barycentric condensation $\mathcal{D}^{\mathcal{S}}$ of $\mathcal{D}$ is a barycentric condensation in the sense of Definition 7.40, without Assumption (7.46) on the barycentric coefficients but such that:

1. *for all $K \in \mathcal{M}$ there exists an open set O_K such that*
   a) O_K is star-shaped with respect to some $\boldsymbol{x}_K$,
   b) $I_K \subset \overline{O_K}$, and
   c) for all $\varphi \in \mathcal{S}$, $\varphi_{|O_K} \in W^{2,\infty}(O_K)$,
2. *for all $\varphi \in \mathcal{S}$, there exists $C_\varphi \geq 0$, depending only on φ, and $R_{\mathcal{D}^{\mathcal{S}}}$, depending only on $\mathcal{D}^{\mathcal{S}}$, such that*

$$\forall K \in \mathcal{M}\,,\ \forall i \in I_K \backslash I^{\text{BA}}\ : \qquad \left|\varphi(\boldsymbol{x}_i) - \sum_{j \in H_i} \beta_j^i \varphi(\boldsymbol{x}_j)\right| \leq C_\varphi R_{\mathcal{D}^{\mathcal{S}}} \operatorname{diam}(K)^2. \quad (7.101)$$

The $\mathcal{S}$-regularity of $\mathcal{D}^{\mathcal{S}}$ is then defined by

$$\operatorname{reg}_{\mathcal{S}}(\mathcal{D}^{\mathcal{S}}) = \operatorname{reg}_{\text{BA}}(\mathcal{D}^{\mathcal{S}}) + R_{\mathcal{D}^{\mathcal{S}}} + \max_{K \in \mathcal{M}} \frac{\operatorname{diam}(O_K)}{\operatorname{diam}(K)}.$$

Linearly exact barycentric condensations (*i.e.* in the sense of Definition 7.40) are $\mathcal{S}$-adapted barycentric condensations with $\mathcal{S} = C_c^\infty(\Omega)$ and O_K the interior of the convex hull of $(\boldsymbol{x}_i)_{i \in I_K^{\text{BA}}}$.
The following theorem is an equivalent of Theorem 7.43 for $\mathcal{S}$-adapted barycentric condensations.

Theorem 7.74 (Properties of $\mathcal{S}$-adapted barycentric condensations). *Let $(\mathcal{D}_m)_{m \in \mathbb{N}}$ be a sequence of LLE GDs in the sense of Definition 7.33, that is coercive, GD-consistent, limit-conforming and compact in the sense of Definition 2.2, 2.4, 2.5 and 2.8. Let $\mathcal{S}$ be a dense subset of $W_0^{1,p}(\Omega)$ and, for each m, take $\mathcal{D}_m^{\mathcal{S}}$ an $\mathcal{S}$-adapted barycentric condensation of $\mathcal{D}_m$. Assume that $(\operatorname{reg}_{\text{LLE}}(\mathcal{D}_m))_{m \in \mathbb{N}}$ and $(\operatorname{reg}_{\mathcal{S}}(\mathcal{D}_m^{\mathcal{S}}))_{m \in \mathbb{N}}$ are bounded, and that $h_{\mathcal{M}_m} \to 0$ as $m \to \infty$ (where $\mathcal{M}_m$ is the mesh associated with $\mathcal{D}_m$).*
Then $(\mathcal{D}_m^{\mathcal{S}})_{m \in \mathbb{N}}$ is coercive, GD-consistent, limit-conforming and compact.

Proof. A close examination of the proof of Theorem 7.43 shows that the transfer of the coercivity, limit-conformity and compactness properties from a sequence of GDs to their barycentric condensations does not require Assumption (7.46). Hence these properties are satisfied by $\mathcal{S}$-adapted barycentric condensations.
Let us now prove the GD-consistency. We drop the index m for legibility and we take $\varphi \in \mathcal{S}$. Analogously to the proof of Proposition 7.37, define the interpolant $v \in X_{\mathcal{D}^{\mathcal{S}},0}$ by $v_i = \varphi(\boldsymbol{x}_i)$ for all $i \in I^{\text{BA}}$. Let $\widetilde{v} \in X_{\mathcal{D},0}$ be given by (7.47), that is $\widetilde{v}_i = v_i = \varphi(\boldsymbol{x}_i)$ if $i \in I^{\text{BA}}$ and

$$\widetilde{v}_i = \sum_{j \in H_i} \beta_j^i v_j = \sum_{j \in H_i} \beta_j^i \varphi(\boldsymbol{x}_i) \quad \text{ if } i \in I \backslash I^{\text{BA}}.$$

By (7.101) we have $|\widetilde{v}_i - \varphi(\boldsymbol{x}_i)| \leq C_\varphi R_{\mathcal{D}^s} \operatorname{diam}(K)^2$ if $i \in I_K$. Hence

$$\forall K \in \mathcal{M}\,,\ \forall i \in I_K\ :\ \widetilde{v}_i = \varphi(\boldsymbol{x}_i) + \mathcal{O}(\operatorname{diam}(K)^2). \tag{7.102}$$

We can then reproduce with this $\widetilde{v}$ the proof of Lemma 7.31, using the $\boldsymbol{x}_K$ with respect to which O_K is star-shaped. This shows that (7.29) holds up to an additional term $\mathcal{O}(\operatorname{diam}(O_K)^2) = \mathcal{O}(\operatorname{diam}(K)^2)$. Still following the computations in the proof of Lemma 7.31, the $W^{2,\infty}(O_K)$-regularity of φ then shows that $\|\mathcal{G}_K \widetilde{v} - \nabla\varphi\|_{L^p(K)^d} = \mathcal{O}(|K|^{1/p}\operatorname{diam}(K))$ on K. This gives

$$\|\nabla_{\mathcal{D}^s} v - \nabla\varphi\|_{L^p(\Omega)^d} = \|\nabla_{\mathcal{D}}\widetilde{v} - \nabla\varphi\|_{L^p(\Omega)^d} = \mathcal{O}(h_{\mathcal{M}}). \tag{7.103}$$

The property (7.102), the definition (7.32) of $\Pi_{\mathcal{D}}$ and the boundedness of $\operatorname{reg}_{\text{LLE}}(\mathcal{D})$ also give

$$\|\Pi_{\mathcal{D}^s} v - \varphi\|_{L^p(\Omega)} = \|\Pi_{\mathcal{D}}\widetilde{v} - \varphi\|_{L^p(\Omega)} = \mathcal{O}(h_{\mathcal{M}}). \tag{7.104}$$

Estimates (7.103) and (7.104) show that $S_{\mathcal{D}^s}(\varphi) = \mathcal{O}(h_{\mathcal{M}})$ for any $\varphi \in \mathcal{S}$. The proof is complete by invoking Lemma 2.16 and the density of $\mathcal{S}$ in $W_0^{1,p}(\Omega)$.

■

8

Conforming approximations

8.1 Conforming Galerkin methods

8.1.1 Homogeneous Dirichlet boundary conditions

Conforming Galerkin methods are probably the simplest GDMs. They simply consist in replacing the infinite-dimensional Sobolev space involved in the weak formulation (*e.g.*, $H_0^1(\Omega)$ in (2.22)) by a finite dimensional subspace; the corresponding GD is defined as follows.
Let $\mathcal{A} = (\varphi_i)_{i\in I}$ be a linearly independent finite family of elements of $W_0^{1,p}(\Omega)$. A conforming Galerkin GD based on $\mathcal{A}$ is defined by:

$$\begin{aligned} &X_{\mathcal{D},0} = \{v = (v_i)_{i\in I} \,:\, v_i \in \mathbb{R} \text{ for all } i \in I\} \text{ and, for } v \in X_{\mathcal{D},0}, \\ &\Pi_{\mathcal{D}} v = \sum_{i\in I} v_i \varphi_i \in W_0^{1,p}(\Omega) \text{ and } \nabla_{\mathcal{D}} v = \nabla(\Pi_{\mathcal{D}} v) = \sum_{i\in I} v_i \nabla\varphi_i. \end{aligned} \tag{8.1}$$

The properties of this GD are straightforward.

Theorem 8.1 (Conforming GDs for hom. Dirichlet BCs). *For all $m \in \mathbb{N}$, take $\mathcal{A}^{(m)} = (\varphi_i^{(m)})_{i\in I^{(m)}}$ a linearly independent finite family of $W_0^{1,p}(\Omega)$ and define $\mathcal{D}_m = (X_{\mathcal{D}_m,0}, \Pi_{\mathcal{D}_m}, \nabla_{\mathcal{D}_m})$ by (8.1) with $\mathcal{A} = \mathcal{A}^{(m)}$. Then $\mathcal{D}_m$ is a GD for homogeneous Dirichlet boundary conditions in the sense of Definition 2.1.*
Furthermore, if

$$\forall \varphi \in W_0^{1,p}(\Omega)\,,\ \lim_{m\to\infty} \min_{v\in X_{\mathcal{D}_m,0}} \|\nabla\varphi - \nabla_{\mathcal{D}_m} v\|_{L^p(\Omega)^d} = 0, \tag{8.2}$$

then the sequence $(\mathcal{D}_m)_{m\in\mathbb{N}}$ is coercive, GD-consistent, limit-conforming and compact in the sense Definitions 2.2, 2.4, 2.5 and 2.8.

Remark 8.2. If $(u_i)_{i\in\mathbb{N}}$ is a dense sequence in $W_0^{1,p}(\Omega)$ and if, for all $m \in \mathbb{N}$, $\mathcal{A}^{(m)}$ is a basis of the space $V^{(m)}$ spanned by $(u_i)_{i=0,\dots,m}$, then the hypotheses of the preceding theorem, in particular (8.2), are satisfied.

Proof. Thanks to the Poincaré inequality in $W_0^{1,p}(\Omega)$, $\|\nabla\cdot\|_{L^p(\Omega)^d}$ is a norm on $W_0^{1,p}(\Omega)$. Let $v \in X_{\mathcal{D},0}$ and assume that $\|\nabla(\Pi_{\mathcal{D}_m} v)\|_{L^p(\Omega)^d} = 0$; then $\Pi_{\mathcal{D}_m} v = \sum_{i\in I} v_i \varphi_i^{(m)} = 0$ in $W_0^{1,p}(\Omega)$. Since the family $(\varphi_i^{(m)})_{i\in I^{(m)}}$ is linearly independent, we infer that $v_i = 0$ for all $i \in I$, which shows that $\|\nabla_{\mathcal{D}_m}\cdot\|_{L^p(\Omega)^d}$ is a norm on $X_{\mathcal{D}_m,0}$. Hence, $\mathcal{D}_m$ is a GD in the sense of Definition 2.1.
The coercivity of $(\mathcal{D}_m)_{m\in\mathbb{N}}$ is an immediate consequence of the continuous Poincaré inequality, since this inequality gives, for all $u \in X_{\mathcal{D}_m,0}$,

$$\|\Pi_{\mathcal{D}_m} u\|_{L^p(\Omega)} \le \operatorname{diam}(\Omega)\, \|\nabla(\Pi_{\mathcal{D}_m} u)\|_{L^p(\Omega)^d} = \operatorname{diam}(\Omega)\, \|\nabla_{\mathcal{D}_m} u\|_{L^p(\Omega)^d}\,.$$

Assumption (8.2) and Poincaré's inequality imply the consistency of $(\mathcal{D}_m)_{m\in\mathbb{N}}$. Indeed, for all $v \in X_{\mathcal{D}_m,0}$ and $\varphi \in W_0^{1,p}(\Omega)$,

$$\begin{aligned}
&\|\Pi_{\mathcal{D}_m} v - \varphi\|_{L^p(\Omega)} + \|\nabla_{\mathcal{D}_m} v - \nabla\varphi\|_{L^p(\Omega)^d} \\
&\qquad = \|\Pi_{\mathcal{D}_m} v - \varphi\|_{L^p(\Omega)} + \|\nabla(\Pi_{\mathcal{D}_m} v) - \nabla\varphi\|_{L^p(\Omega)^d} \\
&\qquad \le (1 + \operatorname{diam}(\Omega))\, \|\nabla(\Pi_{\mathcal{D}_m} v) - \nabla\varphi\|_{L^p(\Omega)^d}\,.
\end{aligned}$$

Hence,

$$S_{\mathcal{D}_m}(\varphi) \le (1 + \operatorname{diam}(\Omega)) \min_{v\in X_{\mathcal{D}_m,0}} \|\nabla_{\mathcal{D}_m} v - \nabla\varphi\|_{L^p(\Omega)^d} \to 0 \text{ as } m \to \infty.$$

The limit-conformity is also straightforward, since $\nabla_{\mathcal{D}_m} u = \nabla(\Pi_{\mathcal{D}_m} u)$ for all $u \in X_{\mathcal{D}_m,0}$, and therefore Stokes' formula in Sobolev spaces shows that $W_{\mathcal{D}_m}(\boldsymbol{\varphi}) = 0$ for all $\boldsymbol{\varphi} \in W_{\mathrm{div}}^{p'}(\Omega)$. The compactness of $(\mathcal{D}_m)_{m\in\mathbb{N}}$ follows from Rellich's theorem. Indeed, if $v_m \in X_{\mathcal{D}_m,0}$ is such that $\|\nabla_{\mathcal{D}_m} v_m\|_{L^p(\Omega)^d} = \|\nabla(\Pi_{\mathcal{D}_m} v_m)\|_{L^p(\Omega)^d}$ is bounded, then by Rellich's compactness theorem, $(\Pi_{\mathcal{D}_m} v_m)_{m\in\mathbb{N}}$ is relatively compact in $L^p(\Omega)$. ∎

Remark 8.3. Dealing with non-homogeneous Dirichlet boundary conditions requires the design of an interpolation operator $\mathcal{I}_{\mathcal{D},\partial}$. This interpolator usually depends on the chosen method and of the expected regularity of the solution. See Section 8.3 for an example.

8.1.2 Non-homogeneous Neumann boundary conditions

The definition of a conforming Galerkin GD for Neumann boundary conditions is pretty straightforward. Take $\mathcal{A} = (\varphi_i)_{i\in I}$ a linearly independent finite family of elements of $W^{1,p}(\Omega)$ and set

$$\begin{aligned}
&X_{\mathcal{D}} = \{v = (v_i)_{i\in I} \,:\, v_i \in \mathbb{R} \text{ for all } i \in I\} \text{ and, for } v \in X_{\mathcal{D}}, \\
&\Pi_{\mathcal{D}} v = \sum_{i\in I} v_i \varphi_i\,,\ \nabla_{\mathcal{D}} v = \nabla(\Pi_{\mathcal{D}} v) = \sum_{i\in I} v_i \nabla\varphi_i \text{ and } \mathbb{T}_{\mathcal{D}} u = \gamma(\Pi_{\mathcal{D}} u),
\end{aligned} \tag{8.3}$$

where γ is the trace on $\partial\Omega$ functions in $W^{1,p}(\Omega)$. The following result can be proved in a similar way as Theorem 8.1.

Theorem 8.4 (Conforming GDs for non-hom. Neumann BCs). *For all $m \in \mathbb{N}$, take $\mathcal{A}^{(m)} = (\varphi_i^{(m)})_{i \in I^{(m)}}$ a linearly independent finite family of $W^{1,p}(\Omega)$ and let $\mathcal{D}_m = (X_{\mathcal{D}_m,0}, \Pi_{\mathcal{D}_m}, \mathbb{T}_{\mathcal{D}_m}, \nabla_{\mathcal{D}_m})$ be defined by (8.3) with $\mathcal{A} = \mathcal{A}^{(m)}$. Then $\mathcal{D}_m$ is a GD for non-homogeneous Neumann problems in the sense of Definition 3.11.*
Furthermore, if

$$\forall \varphi \in W^{1,p}(\Omega)\,,\ \lim_{m\to\infty} \min_{v \in X_{\mathcal{D}_m}} \|\varphi - \Pi_{\mathcal{D}_m} v\|_{W^{1,p}(\Omega)^d} = 0, \tag{8.4}$$

then the sequence $(\mathcal{D}_m)_{m\in\mathbb{N}}$ is coercive, GD-consistent, limit-conforming and compact in the sense of Definitions 3.13, 3.4, 3.14 and 3.15.

Remark 8.5 (Fourier boundary conditions). The relations (8.3) also define a conforming Galerkin GD for Fourier boundary conditions, and the equivalent of Theorem 8.4 holds for sequences of such GDs.

8.2 $\mathbb{P}_k$ finite elements for homogeneous Dirichlet boundary conditions

8.2.1 Definition of $\mathbb{P}_k$ gradient discretisations

$\mathbb{P}_k$ finite elements methods are particular conforming Galerkin methods, and thus are GDMs (Section 8.1). They however deserve to be described in detail, if only because they will give us our first practical example of LLE GD.

Let $\mathfrak{T} = (\mathcal{M}, \mathcal{F}, \mathcal{P}, \mathcal{V})$ be a conforming simplicial mesh of Ω in the sense of Definition 7.4, and let $k \in \mathbb{N} \setminus \{0\}$. We follow Definition 7.33 for the construction of the $\mathbb{P}_k$ LLE gradient discretisation $\mathcal{D} = (X_{\mathcal{D},0}, \nabla_{\mathcal{D}}, \Pi_{\mathcal{D}})$ for homogeneous Dirichlet boundary conditions. We therefore describe the geometrical entities I attached to the discrete unknowns, the set of approximation points S, the $\mathbb{P}_0$-exact function reconstructions π_K the $\mathbb{P}_1$-exact gradients reconstructions $\mathcal{G}_K$ on the elements K of $\mathcal{M}$, and we check that $\|\nabla_{\mathcal{D}}\cdot\|_{L^p(\Omega)^d}$ is a norm on $X_{\mathcal{D},0}$.

1. The set I of geometrical entities attached to the discrete unknowns is $I = \mathcal{V}^{(k)}$, and the set of approximation points is $S = I$, where $\mathcal{V}^{(k)} = \bigcup_{K\in\mathcal{M}} \mathcal{V}_K^{(k)}$ and $\mathcal{V}_K^{(k)}$ is the set of the points $\boldsymbol{x}$ of the form (see Figure 8.1 for examples):

$$\boldsymbol{x} = \sum_{\boldsymbol{s}\in\mathcal{V}_K} \frac{i_{\boldsymbol{s}}}{k}\boldsymbol{s} \quad \text{with } (i_{\boldsymbol{s}})_{\boldsymbol{s}\in\mathcal{V}_K} \in \{0,\dots,k\}^{\mathcal{V}_K} \text{ s.t. } \sum_{\boldsymbol{s}\in\mathcal{V}_K} i_{\boldsymbol{s}} = k. \tag{8.5}$$

(Note that for $k = 1$, $\mathcal{V}^{(1)} = \mathcal{V}$.) Then $I_\Omega = \mathcal{V}_{\text{int}}^{(k)} := \mathcal{V}^{(k)} \cap \Omega$, $I_\partial = \mathcal{V}_{\text{ext}}^{(k)} := \mathcal{V}^{(k)} \cap \partial\Omega$, and thus

$$X_{\mathcal{D},0} = \{v = (v_{\boldsymbol{s}})_{\boldsymbol{s}\in\mathcal{V}} \,:\, v_{\boldsymbol{s}} \in \mathbb{R} \text{ for all } \boldsymbol{s} \in \mathcal{V}_{\text{int}}^{(k)}\,,\ v_{\boldsymbol{s}} = 0 \text{ for all } \boldsymbol{s} \in \mathcal{V}_{\text{ext}}^{(k)}\}.$$

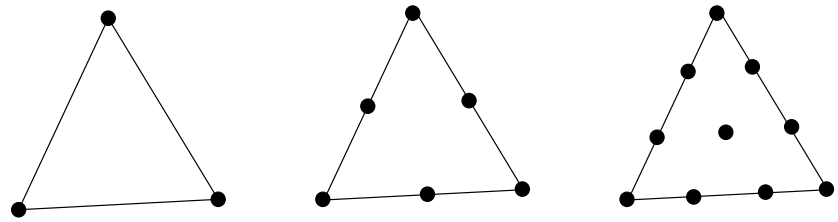


Fig. 8.1. Location of the unknowns in each cell for the $\mathbb{P}_k$ finite element method.

2. For $K \in \mathcal{M}$, we let $I_K = \mathcal{V}_K^{(k)} := \mathcal{V}^{(k)} \cap \overline{K}$. The function reconstruction $\Pi_{\mathcal{D}}$ in (7.32) is defined on K through the local basis functions $(\pi_K^{\boldsymbol{s}})_{\boldsymbol{s}\in\mathcal{V}_K^{(k)}}$, called in this particular case the Lagrange interpolation functions and defined the following way. For each $\boldsymbol{s} \in \mathcal{V}_K^{(k)}$, $\pi_K^{\boldsymbol{s}}$ is polynomial in K of degree k, and satisfies $\pi_K^{\boldsymbol{s}}(\boldsymbol{s}) = 1$ and $\pi_K^{\boldsymbol{s}}(\boldsymbol{s}') = 0$ for all $\boldsymbol{s}' \in \mathcal{V}_K^{(k)} \setminus \{\boldsymbol{s}\}$. This leads to
$$\forall v \in X_{\mathcal{D},0}\,,\ \forall K \in \mathcal{M}\,,\ (\Pi_{\mathcal{D}} v)_{|K} = \sum_{\boldsymbol{s}\in\mathcal{V}_K^{(k)}} v_{\boldsymbol{s}} \pi_K^{\boldsymbol{s}}. \tag{8.6}$$
Since $\sum_{\boldsymbol{s}\in\mathcal{V}_K^{(k)}} \pi_K^{\boldsymbol{s}}$ is a polynomial of degree at most k that has value 1 at each $\boldsymbol{s} \in \mathcal{V}_K^{(k)}$, Lemma 8.6 shows that $\sum_{\boldsymbol{s}\in\mathcal{V}_K^{(k)}} \pi_K^{\boldsymbol{s}} = 1$ on K. Hence, $(\pi_K^{\boldsymbol{s}})_{\boldsymbol{s}\in\mathcal{V}_K^{(k)}}$ is a $\mathbb{P}_0$-exact function reconstruction on K.
For each $v \in X_{\mathcal{D},0}$, $\Pi_{\mathcal{D}} v$ is polynomial of degree k or less in each cell, and satisfies $\Pi_{\mathcal{D}} v(\boldsymbol{s}) = v_{\boldsymbol{s}}$ for all $\boldsymbol{s} \in \mathcal{V}^{(k)}$. By Lemma 8.7, $\Pi_{\mathcal{D}} v$ is therefore continuous over Ω, and thus belong to $W^{1,p}(\Omega)$. Moreover, for any $\sigma \in \mathcal{F}_{\text{ext}} \cap \mathcal{F}_K$, $\Pi_{\mathcal{D}} v$ vanishes at all $\boldsymbol{s} \in \mathcal{V}_K^{(k)} \cap \overline{\sigma}$; since σ is a simplex in dimension $d-1$ and $\mathcal{V}_\sigma^{(k)} = \mathcal{V}_K^{(k)} \cap \overline{\sigma}$, by Lemma 8.6 applied to σ instead of K we deduce that $\Pi_{\mathcal{D}} v = 0$ on the boundary faces, and thus that $\Pi_{\mathcal{D}} v \in W_0^{1,p}(\Omega)$.
3. We define the family $\mathcal{G}_K = (\mathcal{G}_K^{\boldsymbol{s}})_{\boldsymbol{s}\in\mathcal{V}_K^{(k)}}$ of functions in $L^\infty(K)^d$ by
$$\mathcal{G}_K^{\boldsymbol{s}} = \nabla \pi_K^{\boldsymbol{s}}. \tag{8.7}$$
If q is a polynomial of degree less than or equal to k, then $\sum_{\boldsymbol{s}\in\mathcal{V}_K^{(k)}} q(\boldsymbol{s}) \pi_K^{\boldsymbol{s}}$ is a polynomial of degree less than or equal to k, and matches q at all $\boldsymbol{s} \in \mathcal{V}_K^{(k)}$. By Lemma 8.6, these two polynomials coincide. In particular, with $q = A$ affine map,
$$\sum_{\boldsymbol{s}\in\mathcal{V}_K^{(k)}} A(\boldsymbol{s}) \mathcal{G}_K^{\boldsymbol{s}} = \sum_{\boldsymbol{s}\in\mathcal{V}_K^{(k)}} A(\boldsymbol{s}) \nabla \pi_K^{\boldsymbol{s}} = \nabla \sum_{\boldsymbol{s}\in\mathcal{V}_K^{(k)}} A(\boldsymbol{s}) \pi_K^{\boldsymbol{s}} = \nabla A.$$

Hence, $\mathcal{G}_K$ is a $\mathbb{P}_1$-exact gradient reconstruction on K upon $\mathcal{V}_K^{(k)}$.
The gradient reconstruction $\nabla_{\mathcal{D}}$ is given by these local gradients, which means that

$$\forall v \in X_{\mathcal{D},0}\,,\ \forall K \in \mathcal{M}\,,\ (\nabla_{\mathcal{D}} v)_{|K} = \sum_{\boldsymbol{s} \in \mathcal{V}_K^{(k)}} v_{\boldsymbol{s}} \mathcal{G}_K^{\boldsymbol{s}},$$

that is, given (8.6),

$$\forall v \in X_{\mathcal{D},0}\,,\ \nabla_{\mathcal{D}} v = \nabla(\Pi_{\mathcal{D}} v) \text{ a.e. on } \Omega. \tag{8.8}$$

4. Relation (8.8) and the Poincaré inequality in $W_0^{1,p}(\Omega)$ imply that $v \mapsto \|\nabla_{\mathcal{D}} v\|_{L^p(\Omega)^d}$ is a norm on $X_{\mathcal{D},0}$.

The following two lemmas justify the existence and uniqueness of the Lagrange interpolation functions $\pi_K^{\boldsymbol{s}}$, and the above construction. They are classical results, and form together what is called the "unisolvence of $\mathcal{V}^{(k)}$ for conforming $\mathbb{P}_k$-finite elements" [48].

Lemma 8.6 ($\mathcal{V}_K^{(k)}$ is a complete family for $\mathbb{P}_k$). *Let K be a simplex, $k \in \mathbb{N}$ and $\mathcal{V}_K^{(k)}$ be the points defined by (8.5). Then for any choice of values $(a_{\boldsymbol{s}})_{\boldsymbol{s} \in \mathcal{V}_K^{(k)}}$, there exists a unique polynomial function p of degree at most k such that $p(\boldsymbol{s}) = a_{\boldsymbol{s}}$ for all $\boldsymbol{s} \in \mathcal{V}_K^{(k)}$.*

Proof. Let

$$\Phi : \mathbb{P}_k(K) \mapsto X_K := \{(a_{\boldsymbol{s}})_{\boldsymbol{s} \in \mathcal{V}_K^{(k)}} \,:\, a_{\boldsymbol{s}} \in \mathbb{R} \text{ for all } \boldsymbol{s} \in \mathcal{V}_K^{(k)}\}$$

be defined by $\Phi(p) = (p(\boldsymbol{s}))_{\boldsymbol{s} \in \mathcal{V}_K^{(k)}}$. Φ is clearly linear, and X_K is a vector space of dimension $\mathrm{Card}(\mathcal{V}_K^{(k)})$. Let us assume that (i) $\dim(\mathbb{P}_k(K)) = \mathrm{Card}(\mathcal{V}_K^{(k)})$, and (ii) if $\Phi(p) = 0$ then $p \equiv 0$. Then Φ is one-to-one between two vector spaces of same dimension, and therefore Φ is an isomorphism. Hence, for any family of real numbers $(a_{\boldsymbol{s}})_{\boldsymbol{s} \in \mathcal{V}_K^{(k)}} \in X_K$ there exists a unique $p \in \mathbb{P}_k(K)$ such that $\Phi(p) = (a_{\boldsymbol{s}})_{\boldsymbol{s} \in \mathcal{V}_K^{(k)}}$, which is the conclusion of the lemma. It remains to prove (i) and (ii).

Proof of (i): the dimension of $\mathbb{P}_k(K)$ is the number of monomials of the form $\boldsymbol{x}^{\boldsymbol{\alpha}} = x_1^{\alpha_1} \cdots x_d^{\alpha_d}$ with $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_d)$ and $|\boldsymbol{\alpha}| = \alpha_1 + \cdots + \alpha_d \le k$. For such a $\boldsymbol{\alpha}$ we define $\boldsymbol{i} = (i_0, \ldots, i_d)$ by $i_0 = k - (\alpha_1 + \ldots + \alpha_d)$, $i_1 = \alpha_1$, $\ldots$, $i_d = \alpha_d$. This correspondence $\boldsymbol{\alpha} \to \boldsymbol{i}$ clearly creates a bijection between $\{\boldsymbol{\alpha} \in \mathbb{N}^d \,:\, |\boldsymbol{\alpha}| \le k\}$ and $\{\boldsymbol{i} \in \mathbb{N}^{d+1} \,:\, |\boldsymbol{i}| = k\}$. Hence those two sets have the same cardinal. Since $\dim(\mathbb{P}_k(K))$ is the cardinal of the first set and, by (8.5), $\mathrm{Card}(\mathcal{V}_K^{(k)})$ is the cardinal of the second set, the proof of (i) is complete.

Proof of (ii): the proof is done by induction on d.

$\underline{d=1}$: K is then a segment of line, and $\mathcal{V}_K^{(k)}$ are $k+1$ distinct points on $\overline{K}$. It is well-known that if p is a polynomial of one variable, of degree less than or equal to k and that vanishes on $k+1$ distinct points, then $p \equiv 0$ and the case $d=1$ is thus proved.

$\underline{d-1 \Rightarrow d}$: we take $d \geq 2$, we assume that (ii) holds for $d-1$ and we want to prove that it holds for d. The proof is done by induction on k.

- $\underline{k=1}$: the polynomial p is affine and vanishes at the vertices of K. The mapping $p - p(0)$ is linear, and therefore preserves barycentric combinations. This mapping takes the value $-p(0)$ at the vertices of K. Since these vertices form a barycentric basis of $\mathbb{R}^d$, we deduce that $p - p(0)$ is constant equal to $-p(0)$ on $\mathbb{R}^d$, which shows that $p \equiv 0$ on $\mathbb{R}^d$.
- $\underline{k-1 \Rightarrow k}$: up to an affine change of variables, we can assume that one of the faces σ_0 of K lies on the hyperplane $\{x_d = 0\}$. We then denote by $\boldsymbol{s}_0$ the vertex of K opposite to σ_0 (see Figure 8.2). A polynomial p in d variables of degree less than or equal to k can be written

$$p(\boldsymbol{x}) = x_d q(\boldsymbol{x}) + r(x_1, \ldots, x_{d-1})$$

  where q is a polynomial of degree less than or equal to $k-1$, and r a polynomial of degree less than or equal to k. Since p vanishes on $\mathcal{V}_K^{(k)}$ and σ_0 is a $(d-1)$-dimensional simplex that lies on $\{x_d = 0\}$, we see that r vanishes on $\mathcal{V}_K^{(k)} \cap \overline{\sigma_0} = \mathcal{V}_{\sigma_0}^{(k)}$. By the induction hypothesis the result (ii) is valid in dimension $d-1$ and r is therefore the zero polynomial.
  The convex hull of $\mathcal{V}_K^{(k)} \backslash \mathcal{V}_{\sigma_0}^{(k)}$ forms a (closed) simplex K' such that $\mathcal{V}_{K'}^{(k-1)} = \mathcal{V}_K^{(k)} \backslash \mathcal{V}_{\sigma_0}^{(k)}$ (these vertices correspond to (8.5) with the index i_0, corresponding to $\boldsymbol{s}_0$, different from zero). Moreover, since $K' \cap \{x_d = 0\} = \emptyset$, the relation $p(\boldsymbol{x}) = x_d q(\boldsymbol{x})$ shows that q vanishes on $\mathcal{V}_{K'}^{(k-1)}$. Since q has degree $k-1$ or less, the induction hypothesis on k shows that $q \equiv 0$. The proof that $p \equiv 0$ is therefore complete.

■

Lemma 8.7 (Continuity through the faces of piecewise polynomial functions). *Let $k \in \mathbb{N} \backslash \{0\}$, let K and L be two simplices of $\mathbb{R}^d$ with a common face σ, and let the sets $\mathcal{V}_K^{(k)}$ (resp. $\mathcal{V}_L^{(k)}$) be defined by the points (8.5) (resp. with K replaced with L). Let p_K and p_L be polynomial functions on $\overline{K}$ and $\overline{L}$, respectively, such that p_K and p_L have degree at most k and coincide at all points of $\mathcal{V}_K^{(k)} \cap \mathcal{V}_L^{(k)}$. Then p_K and p_L coincide on σ.*

Proof. The functions $(p_K)_{|\sigma}$ and $(p_L)_{|\sigma}$ are polynomial of degree at most k, and are identical at the points of $\mathcal{V}_K^{(k)} \cap \mathcal{V}_L^{(k)}$. Since σ is a simplex in dimension $d-1$, and $\mathcal{V}_\sigma^{(k)} = \mathcal{V}_K^{(k)} \cap \overline{\sigma} = \mathcal{V}_L^{(k)} \cap \overline{\sigma} = \mathcal{V}_K^{(k)} \cap \mathcal{V}_L^{(k)}$, Lemma 8.6 can be applied to σ, and shows that $(p_K)_{|\sigma}$ and $(p_L)_{|\sigma}$ are identical over the whole face σ.

■

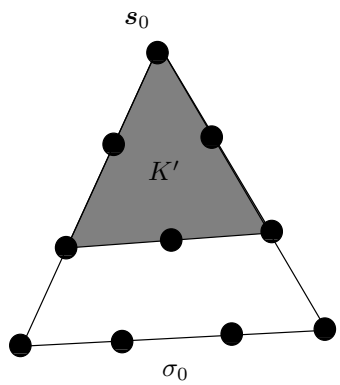


Fig. 8.2. Illustration of the construction in the proof of Lemma 8.6 for $k = 3$.

8.2.2 Properties of $\mathbb{P}_k$ gradient discretisations

The properties of $\mathbb{P}_k$ GDs follow from their conformity and from Proposition 7.37, provided that we establish an estimate on the LLE regularity of $\mathbb{P}_k$ GDs. We first state a classical result, which relates the independence properties of a family of vectors in $\mathbb{R}^d$ with the fact that they enclose a ball of radius comparable to their lengths. This result is then used to bound the LLE regularity of $\mathbb{P}_k$ GDs.

Lemma 8.8. *Let $(\boldsymbol{x}_i)_{i=1,\dots,d}$ be vectors in $\mathbb{R}^d$, and let M be the $d \times d$ matrix with columns $\boldsymbol{x}_i$. We let $\ell = \max_{i=1,\dots,d} |\boldsymbol{x}_i|$ and we assume that the convex hull of $\{0, \boldsymbol{x}_1, \dots, \boldsymbol{x}_d\}$ contains a ball of radius $\varrho\ell$ for some $\varrho > 0$. Then*

$$|M^{-1}| \leq \frac{d^{1/2}}{\omega_d \varrho^d} \ell^{-1}, \tag{8.9}$$

where ω_d is the measure of the unit ball in $\mathbb{R}^d$.

Proof. We first recall that $|\det(M)|$ is the volume of the d-dimensional parallelogram $M[0,1]^d$ defined by $(\boldsymbol{x}_1, \dots, \boldsymbol{x}_d)$. This parallelogram contains the convex hull $\{\sum_{i=1}^d \lambda_i x_i \,:\, \lambda_i \geq 0\,,\ \sum_i \lambda_i \leq 1\}$ of $\{0, \boldsymbol{x}_1, \dots, \boldsymbol{x}_d\}$. Therefore

$$|\det(M)| \geq \mathrm{vol}(B(0, \varrho\ell)) = \omega_d \varrho^d \ell^d. \tag{8.10}$$

Let $\boldsymbol{\xi} = (\xi_1, \dots, \xi_d) \in \mathbb{R}^d$. We have $M\boldsymbol{\xi} = \sum_{i=1}^d \xi_i \boldsymbol{x}_i$. Hence, for all $j = 1, \dots, d$,

$$\begin{aligned} &\det(\boldsymbol{x}_1, \dots, \boldsymbol{x}_{j-1}, M\boldsymbol{\xi}, \boldsymbol{x}_{j+1}, \dots, \boldsymbol{x}_d) \\ &\qquad = \det(\boldsymbol{x}_1, \dots, \boldsymbol{x}_{j-1}, \xi_j \boldsymbol{x}_j, \boldsymbol{x}_{j+1}, \dots, \boldsymbol{x}_d) \\ &\qquad = \xi_j \det(\boldsymbol{x}_1, \dots, \boldsymbol{x}_d) = \xi_j \det(M), \end{aligned} \tag{8.11}$$

where we used the properties of the determinant and created linear combinations to eliminate all vectors except $\boldsymbol{x}_j$ from $M\boldsymbol{\xi}$. The determinant is multi-linear continuous with a norm 1. Using the definition of ℓ and (8.10), Equation (8.11) thus gives

$$\begin{aligned}\ell^{d-1}|M\boldsymbol{\xi}| &\geq |\boldsymbol{x}_1| \ldots |\boldsymbol{x}_{j-1}|\,|M\boldsymbol{\xi}|\,|\boldsymbol{x}_{j+1}| \ldots |\boldsymbol{x}_d| \\ &\geq |\det(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{j-1}, M\boldsymbol{\xi}, \boldsymbol{x}_{j+1}, \ldots, \boldsymbol{x}_d)| \\ &\geq |\xi_j|\,|\det(M)| \geq \omega_d \varrho^d \ell^d |\xi_j|,\end{aligned}$$

that is, $|M\boldsymbol{\xi}| \geq \omega_d \varrho^d \ell |\xi_j|$. We square this relation, sum over $j = 1, \ldots, d$, and take the square root. This leads to $d^{1/2}|M\boldsymbol{\xi}| \geq \omega_d \varrho^d \ell |\boldsymbol{\xi}|$. Applying this to $\boldsymbol{\xi} = M^{-1}\boldsymbol{\eta}$ for a generic vector $\boldsymbol{\eta}$ establishes (8.9). ∎

Lemma 8.9 (Estimate of the LLE regularity of a $\mathbb{P}_k$ GD). *Let $\mathfrak{T}$ be a simplicial mesh of Ω in the sense of Definition 7.4, and $\mathcal{D}$ be a $\mathbb{P}_k$ LLE GD as in Section 8.2.1. Then, if $\varrho \geq \kappa_{\mathfrak{T}}$ (see (7.10)), there exists C_1, depending only on d and ϱ, such that*

$$\mathrm{reg}_{\mathrm{LLE}}(\mathcal{D}) \leq C_1. \tag{8.12}$$

Proof. For any $K \in \mathcal{M}$ and any $i \in I_K = \mathcal{V}_K^{(k)}$ we have $\boldsymbol{x}_i \in \overline{K}$ and thus $\mathrm{dist}(\boldsymbol{x}_i, K) = 0$. To control the first and second terms in $\mathrm{reg}_{\mathrm{LLE}}(\mathcal{D})$, thanks to Remark 7.32 and to (8.7), it is sufficient to prove that

$$\|\pi_K^{\boldsymbol{s}}\|_{L^\infty(K)} \leq C_2 \quad \text{and} \quad \|\nabla \pi_K^{\boldsymbol{s}}\|_{L^\infty(K)^d} \leq C_2 h_K^{-1}, \tag{8.13}$$

where C_2 depends only on d and ϱ. This is done by a classical reference element technique.

Let $K \in \mathcal{M}$. Up to a translation we can assume that one of the vertices of K is 0. Let $(0, \boldsymbol{s}_1, \ldots, \boldsymbol{s}_d)$ be the vertices of K and let S_0 be the reference d-simplex $\{\boldsymbol{\alpha} \in \mathbb{R}^d \,:\, \alpha_i > 0\,,\ \sum_i \alpha_i < 1\}$. Let M be the $d \times d$ matrix with columns $(\boldsymbol{s}_1, \ldots, \boldsymbol{s}_d)$. Each column of M is a vector with length at most h_K. Since K contains a ball of radius $\kappa_{\mathfrak{T}}^{-1} h_K \geq \varrho^{-1} h_K$, Lemma 8.8 shows that $|M^{-1}| \leq C_3 h_K^{-1}$ for some C_3 depending only on ϱ and d. By definition of the simplex K, we have $K = MS_0$, and M maps each approximation point of $\mathcal{V}_{S_0}^{(k)}$ onto the corresponding approximation point of $\mathcal{V}_K^{(k)}$ (because M is linear and these approximation points are defined by barycentric relations).

Hence, if $\boldsymbol{s} \in \mathcal{V}_K^{(k)}$, then $\boldsymbol{x} \mapsto \pi_K^{\boldsymbol{s}}(M\boldsymbol{x})$ is a polynomial of degree k that is 1 at $M^{-1}\boldsymbol{s} \in \mathcal{V}_{S_0}^{(k)}$ and 0 at all other points in $\mathcal{V}_{S_0}^{(k)}$. There are only a finite number of such polynomials – remember that S_0 is fixed and does not depend on K. We can therefore define C_4 as the maximum of the $L^\infty(S_0)$ norms of these polynomials and their gradients. This constant depends only on d, and satisfies

$$\|\pi_K^{\boldsymbol{s}}(M\cdot)\|_{L^\infty(S_0)} \leq C_4 \quad \text{and} \quad \|\nabla(\pi_K^{\boldsymbol{s}}(M\cdot))\|_{L^\infty(S_0)^d} \leq C_4.$$

Estimates (8.13) then follows by recalling that $MS_0 = K$, that

$$(\nabla \pi_K^{\boldsymbol{s}})(M\cdot) = (M^T)^{-1}\nabla(\pi_K^{\boldsymbol{s}}(M\cdot)),$$

and by using the estimate $|(M^T)^{-1}| = |M^{-1}| \leq C_3 h_K^{-1}$. ∎

We can now prove the properties of $\mathbb{P}_k$ GDs.

Theorem 8.10 (Properties of $\mathbb{P}_k$ GDs for homogeneous Dirichlet BCs). *Let $(\mathcal{D}_m)_{m\in\mathbb{N}}$ be a sequence of $\mathbb{P}_k$ GDs, as in Section 8.2.1, based on underlying conforming simplicial meshes $(\mathfrak{T}_m)_{m\in\mathbb{N}}$. Assume that $(\kappa_{\mathfrak{T}_m})_{m\in\mathbb{N}}$ is bounded (see (7.10)), and that $h_{\mathcal{M}_m} \to 0$ as $m \to \infty$.*
Then the sequence $(\mathcal{D}_m)_{m\in\mathbb{N}}$ is coercive, GD-consistent, limit-conforming and compact in the sense of Definitions 2.2, 2.4, 2.5 and 2.8.

Proof. If $v_m \in X_{\mathcal{D}_m}$ then $\Pi_{\mathcal{D}_m} v_m \in W_0^{1,p}(\Omega)$ and $\nabla_{\mathcal{D}_m} v_m = \nabla(\Pi_{\mathcal{D}_m} v_m)$. Thus, as in the proof of Theorem 8.1, Poincaré's inequality and Rellich's theorem in $W_0^{1,p}(\Omega)$ show that $(\mathcal{D}_m)_{m\in\mathbb{N}}$ is coercive and compact. Applying Stokes' formula shows that $W_{\mathcal{D}_m}(\boldsymbol{\varphi}) = 0$ for all $\boldsymbol{\varphi} \in W_{\mathrm{div}}^{p'}(\Omega)$, which gives the limit-conformity. Finally, the consistency is a direct consequence of Proposition 7.37 and Lemma 8.9. ∎

The previous theorem gives all the properties required to apply, for example, Theorem 2.35 to establish the convergence of the conforming $\mathbb{P}_k$ scheme for the quasi-linear model (2.49a). If aiming for error estimates, *e.g.*, by using Theorems 2.28 or 2.38, one needs specific estimates on $S_{\mathcal{D}}$ and $W_{\mathcal{D}}$. The following proposition provides such estimates.

Proposition 8.11 (Estimates on $S_{\mathcal{D}}$ and $W_{\mathcal{D}}$ for $\mathbb{P}_k$ GDs). *Let $\mathfrak{T}$ be a conforming simplicial mesh of Ω in the sense of Definition 7.4, and $\mathcal{D}$ be the $\mathbb{P}_k$ LLE GD on $\mathfrak{T}$ as defined in Section 8.2.1. Let $\varrho \geq \kappa_{\mathfrak{T}}$ (see (7.10)). Then, there exists C_5, depending only on Ω, k and ϱ, such that*

$$\forall \boldsymbol{\varphi} \in W_{\mathrm{div}}^{p'}(\Omega)\,,\ W_{\mathcal{D}}(\boldsymbol{\varphi}) = 0, \tag{8.14}$$

$$\forall \varphi \in W^{k+1,\infty}(\Omega) \cap W_0^{1,p}(\Omega)\,,\ S_{\mathcal{D}}(\varphi) \leq C_5 h_{\mathcal{M}}^k \left\|\varphi\right\|_{W^{k+1,\infty}(\Omega)}. \tag{8.15}$$

Here, $S_{\mathcal{D}}$ and $W_{\mathcal{D}}$ are defined by (2.2) and (2.6), respectively. This means that the space size (see Definition 2.22) of the GD is such that

$$h_{\mathcal{D}}(W^{k+1,\infty}(\Omega) \cap W_0^{1,p}(\Omega); W_{\mathrm{div}}^{p'}(\Omega)) \leq C_5 h_{\mathcal{M}}^k.$$

Proof. The relation (8.14) follows directly from the conformity of the $\mathbb{P}_k$ GD, and was already noticed in the proof of Theorem 8.10. The estimate (8.15) on $S_{\mathcal{D}}$ is as straightforward consequence of Lemma 8.9 and Proposition 7.38, once we notice that $\mathcal{D}$ is of order k (as defined in Proposition 7.38). Indeed,

for any $q \in \mathbb{P}_k$ and any $K \in \mathcal{M}$, $\sum_{\boldsymbol{s}\in\mathcal{V}_K^{(k)}} q(\boldsymbol{s})\pi_K^{\boldsymbol{s}}$ is a polynomial of degree at most k that takes, by definition of $(\pi_K^{\boldsymbol{s}})_{\boldsymbol{s}\in\mathcal{V}_K^{(k)}}$, the value $q(\boldsymbol{s})$ at a generic $\boldsymbol{s} \in \mathcal{V}_K^{(k)}$. The polynomial q also satisfies this property so, by Lemma 8.6, $q = \sum_{\boldsymbol{s}\in\mathcal{V}_K^{(k)}} q(\boldsymbol{s})\pi_K^{\boldsymbol{s}}$. This proves (7.38) for polynomials of order k (not just $k-1$), and the exactness (7.39) of the gradients follows since $\mathcal{G}_K = \nabla\pi_K$.

■

Remark 8.12 (Rates of convergence of the $\mathbb{P}_k$ GS)
Proposition 8.11 and Theorem 2.28 give, as expected, $\mathcal{O}(h_{\mathcal{M}}^k)$ error estimates on the $\mathbb{P}_k$ method applied to the linear diffusion equation (2.20), in the case $\overline{u} \in W^{k+1,\infty}(\Omega)$. We refer to [32, Theorem 4.4.20] for more optimal $W^{m,p}$-error estimates, obtained by taking advantage of the specificities of this conforming method.

8.3 $\mathbb{P}_k$ finite element for non-homogeneous Dirichlet, Neumann and Fourier boundary conditions

We briefly describe here, following the remarks in Section 7.3.6, the modifications to bring to the $\mathbb{P}_k$ GD to deal with non-homogeneous Dirichlet conditions, Neumann conditions or Fourier conditions.

8.3.1 Non-homogeneous Dirichlet conditions

Following Definition 7.52, a $\mathbb{P}_k$ GD for non-homogeneous Dirichlet boundary conditions consists in $(X_{\mathcal{D}}, \mathcal{I}_{\mathcal{D},\partial}, \Pi_{\mathcal{D}}, \nabla_{\mathcal{D}})$ where

$$X_{\mathcal{D}} = \{v = (v_{\boldsymbol{s}})_{\boldsymbol{s}\in\mathcal{V}^{(k)}} \,:\, v_{\boldsymbol{s}} \in \mathbb{R} \text{ for all } \boldsymbol{s} \in \mathcal{V}^{(k)}\},$$

$\Pi_{\mathcal{D}} v$ and $\nabla_{\mathcal{D}} v$ are defined by (8.6) and (8.8) (for all $v \in X_{\mathcal{D}}$), and an interpolation operator $\mathcal{I}_{\mathcal{D},\partial} : W^{1-\frac{1}{p},p}(\partial\Omega) \to X_{\mathcal{D},\partial}$ has to be defined, where

$$X_{\mathcal{D},\partial} = \{v \in X_{\mathcal{D}} \,:\, v_{\boldsymbol{s}} = 0 \text{ for all } \boldsymbol{s} \in \mathcal{V}_{\text{ext}}^{(k)}\}.$$

The definition of such an interpolant on $W^{1-\frac{1}{p},p}(\partial\Omega)$ is somewhat problematic, given that $\mathbb{P}_k$ methods call for nodal interpolants – *i.e.* values of the function at the vertices $\mathcal{V}^{(k)}$. Since functions in $W^{1-\frac{1}{p},p}(\partial\Omega)$ are usually not continuous, their value at a given point is not defined. One could then use the notion of Clément interpolators [50], but this would have to be adapted to interpolate functions only defined on the boundary of Ω.
In practice, in the context of $\mathbb{P}_k$ finite element schemes, the boundary conditions are usually continuous. Following Remark 2.50, we therefore only need to define $\mathcal{I}_{\mathcal{D},\partial} : W^{1-\frac{1}{p},p}(\partial\Omega) \cap C(\partial\Omega) \mapsto X_{\mathcal{D},\partial}$. This can be done by setting, for $g \in W^{1-\frac{1}{p},p}(\partial\Omega) \cap C(\partial\Omega)$ and $\boldsymbol{s} \in \mathcal{V}_{\text{ext}}^{(k)}$,

$$(\mathcal{I}_{\mathcal{D},\partial} g)_{\boldsymbol{s}} = g(\boldsymbol{s}). \tag{8.16}$$

We then have the following result.

Theorem 8.13 (Properties of $\mathbb{P}_k$ GDs for non-homogeneous Dirichlet BCs). *Let $(\mathcal{D}_m)_{m\in\mathbb{N}}$ be a sequence of $\mathbb{P}_k$ GDs for non-homogeneous Dirichlet boundary conditions, as above. We denote by $(\mathfrak{T}_m)_{m\in\mathbb{N}}$ the underlying conforming simplicial meshes, and we assume that $(\kappa_{\mathfrak{T}_m})_{m\in\mathbb{N}}$ is bounded (see (7.10)). We also suppose that $h_{\mathcal{M}_m} \to 0$ as $m \to \infty$.*
Then, the sequence $(\mathcal{D}_m)_{m\in\mathbb{N}}$ is coercive, limit-conforming and compact in the sense of Definitions 2.2, 2.5 and 2.8. Moreover, with $S_{\mathcal{D}}$ defined by (2.94), we have $S_{\mathcal{D}_m}(\varphi) \to 0$ as $m \to \infty$, for all $\varphi \in W^{2,\infty}(\Omega)$.

Proof. Poincaré's inequality, integration-by-parts and Rellich theorem in $W_0^{1,p}(\Omega)$ give the coercivity, limit-conformity and compactness as for homogeneous Dirichlet boundary conditions. Given the definition (8.16) of $\mathcal{I}_{\mathcal{D}_m,\partial}$, the consistency for $\varphi \in W^{2,\infty}(\Omega)$ follows by selecting $v = (v_{\boldsymbol{s}})_{\boldsymbol{s}\in\mathcal{V}^{(k)}}$ defined by $v_{\boldsymbol{s}} = \varphi(\boldsymbol{s})$, and by using Lemmas 7.27 and 7.31 as in the proof of Proposition 7.37. ∎

Remark 8.14 (General GD-consistency property)
The last assertion of Theorem 8.13 states a weaker version of the consistency, required only for regular functions. Checking the consistency in the sense of Definition 2.51, that is for functions in $W^{1,p}(\Omega)$ instead of $W^{2,\infty}(\Omega)$, would require to ascertain that (2.96) holds. This is somewhat technical and requires the usage of a Clément interpolator, with boundary interpolator $\mathcal{I}_{\mathcal{D},\partial}$ defined by (8.16). The literature does not seem to contain clear results in that direction.

8.3.2 Neumann boundary conditions

The modification for Neumann boundary conditions is natural. Following Definition 7.54, we simply enable boundary discrete unknowns to be non-zero, *i.e.* we take

$$X_{\mathcal{D}} = \{v = (v_{\boldsymbol{s}})_{\boldsymbol{s}\in\mathcal{V}^{(k)}} \,:\, v_{\boldsymbol{s}} \in \mathbb{R} \text{ for all } \boldsymbol{s} \in \mathcal{V}^{(k)}\}.$$

$\Pi_{\mathcal{D}}$ and $\nabla_{\mathcal{D}}$ are still defined by (8.6) and (8.8) (for all $v \in X_{\mathcal{D}}$).
The proof that (3.1) is a norm on $X_{\mathcal{D}}$ is straightforward. If $\|v\|_{\mathcal{D}} = 0$ then $\nabla_{\mathcal{D}} v = \nabla(\Pi_{\mathcal{D}} v) = 0$ and thus $\Pi_{\mathcal{D}} v$ is constant. As $\|v\|_{\mathcal{D}} = 0$ also implies $\int_\Omega \Pi_{\mathcal{D}} v(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} = 0$, we infer that $\Pi_{\mathcal{D}} v = 0$. Then, for all $\boldsymbol{s} \in \mathcal{V}$, $v_{\boldsymbol{s}} = \Pi_{\mathcal{D}} v(\boldsymbol{s}) = 0$, which shows that $v = 0$.
Finally, for non-homogeneous Neumann boundary conditions, we define $\mathbb{T}_{\mathcal{D}} : X_{\mathcal{D}} \to L^\infty(\partial\Omega)$ by

$$\mathbb{T}_{\mathcal{D}} v = \gamma(\Pi_{\mathcal{D}} v) = (\Pi_{\mathcal{D}} v)_{|\partial\Omega}. \tag{8.17}$$

The Poincaré–Wirtinger inequality in $W^{1,p}(\Omega)$ gives C depending only on Ω and p such that, for all $v \in X_{\mathcal{D}}$,

$$\|\Pi_{\mathcal{D}} v\|_{L^p(\Omega)} \leq C \left(\|\nabla(\Pi_{\mathcal{D}} v)\|_{L^p(\Omega)^d} + \left| \int_\Omega \Pi_{\mathcal{D}} v(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \right| \right) = C \|v\|_{\mathcal{D}}$$

Combined with the continuity of the trace $\gamma : W^{1,p}(\Omega) \to L^p(\partial\Omega)$, this gives a uniform estimate on $C_{\mathcal{D}}$ (defined by (3.9)) depending only on Ω and p. The choice (8.17) of the trace reconstruction shows that $W_{\mathcal{D}}$, defined by (3.11), is identically zero.

Proposition 7.55 gives the consistency of sequences of $\mathbb{P}_k$ GDs for non-homogeneous Neumann boundary conditions. The compactness of such a sequence follows from Rellich's theorem and from the coercivity property, which gives a bound on $(\mathbb{T}_{\mathcal{D}_m} u_m)_{m\in\mathbb{N}}$ whenever $(\|u_m\|_{\mathcal{D}_m})_{m\in\mathbb{N}}$ is bounded. As a conclusion, we therefore have the following theorem.

Theorem 8.15 (Properties of $\mathbb{P}_k$ GDs for Neumann BCs).
Let $(\mathcal{D}_m)_{m\in\mathbb{N}}$ be a sequence of $\mathbb{P}_k$ GDs for Neumann boundary conditions as above, defined from underlying conforming simplicial meshes $(\mathfrak{T}_m)_{m\in\mathbb{N}}$. Assume that $\sup_{m\in\mathbb{N}} \kappa_{\mathfrak{T}_m} < +\infty$ (see (7.10)) and that $h_{\mathcal{M}_m} \to 0$ as $m \to \infty$. Then the sequence $(\mathcal{D}_m)_{m\in\mathbb{N}}$ is coercive, GD-consistent, limit-conforming and compact in the sense of Definitions 3.13, 3.4, 3.14 and 3.15.

8.3.3 Fourier conditions

For Fourier boundary conditions, the trace is still defined by (8.17) and clearly satisfies the conditions in Definition 7.57, with $\mathcal{M}_\partial = \mathcal{F}_{\text{ext}}$ and $I_\sigma = \mathcal{V}_K^{(k)}$ for all $K \in \mathcal{M}$ and all $\sigma \in \mathcal{F}_K \cap \mathcal{F}_{\text{ext}}$. A bound on the LLE regularity of the obtained GD can be established as in the proof of Lemma 8.9, by transporting the basis functions on the reference simplex S_0 to check that $\left\|\pi_{\boldsymbol{s}}^\partial\right\|_{L^\infty(\sigma)}$ is uniformly bounded for all $\sigma \in \mathcal{F}_K \cap \mathcal{F}_{\text{ext}}$ and all $\boldsymbol{s} \in \mathcal{V}_K^{(k)}$. To bound the quantities $\frac{\text{dist}(\boldsymbol{x}_i, K_\partial)}{\text{diam}(K_\partial)}$ in $\text{reg}_{\text{LLE}}(\mathcal{D})$ defined by (7.69), we also use the fact that $\text{diam}(\sigma) \leq C h_K$ whenever $\sigma \in \mathcal{F}_K$, with C depending only on an upper bound of $\kappa_{\mathfrak{T}}$.

As a conclusion, by Proposition 7.58, Theorem 8.15 remains valid in the context of Fourier boundary conditions (with Definition 3.4 replaced with Definition 3.37).

8.4 Mass-lumped $\mathbb{P}_1$ finite elements

It is obvious from (8.6) that the reconstruction $\Pi_{\mathcal{D}}$ of the $\mathbb{P}_k$ GD is not piecewise constant. To benefit from the advantages of a piecewise constant reconstruction, such as a diagonal mass matrix in time-dependent problems, or

the applicability to non-linear models such as Stefan's or Richards' equations, the $\mathbb{P}_k$ GD needs to be mass-lumped as per Definition 7.45.
Mass-lumping leads to a piecewise constant reconstruction $\Pi_{\mathcal{D}}^{\mathrm{ML}}$, whose best approximation properties are of order 1. There is therefore little interest in using high order methods when mass-lumping is required, which is why we only consider the case $k = 1$ here. Since mass-lumping is essentially independent of the boundary conditions (see Sections 7.3.6), we only present here the case of homogeneous Dirichlet boundary conditions.

Definition 8.16 (Mass-lumped $\mathbb{P}_1$ GD). *Let $\mathfrak{T} = (\mathcal{M}, \mathcal{F}, \mathcal{P}, \mathcal{V})$ be a conforming simplicial mesh of Ω in the sense of Definition 7.4, and let $\mathcal{D} = (X_{\mathcal{D},0}, \Pi_{\mathcal{D}}, \nabla_{\mathcal{D}})$ be the $\mathbb{P}_1$ GD built on $\mathfrak{T}$ as in Section 8.2.1 (with $k = 1$). For each $\boldsymbol{s} \in \mathcal{V}$ and $K \in \mathcal{M}$ such that $\boldsymbol{s} \in \mathcal{V}_K$, let*

$$\Omega_{K,\boldsymbol{s}} = \{\boldsymbol{y} \in K \,:\, \pi_K^{\boldsymbol{s}}(\boldsymbol{y}) > \pi_K^{\boldsymbol{s}'}(\boldsymbol{y}) \text{ for all } \boldsymbol{s}' \in \mathcal{V}_K \backslash \{\boldsymbol{s}\}\}$$

(recall that $(\pi_K^{\boldsymbol{s}})_{\boldsymbol{s}\in\mathcal{V}}$ are the $\mathbb{P}_1$ basis functions, defined in Item 2 of Section 8.2.1). Define then (see Figure 8.4 for an illustration)

$$\Omega_{\boldsymbol{s}} = \bigcup_{K\in\mathcal{M} \,|\, \boldsymbol{s}\in\mathcal{V}_K} \Omega_{K,\boldsymbol{s}}.$$

Then a mass-lumped $\mathbb{P}_1$ GD is defined by $\mathcal{D}^{\mathrm{ML}} = (X_{\mathcal{D},0}, \Pi_{\mathcal{D}}^{\mathrm{ML}}, \nabla_{\mathcal{D}})$ where $\Pi_{\mathcal{D}}^{\mathrm{ML}}$ is the piecewise constant reconstruction built from $(\Omega_{\boldsymbol{s}})_{\boldsymbol{s}\in\mathcal{V}}$, that is

$$\forall v \in X_{\mathcal{D},0}\,,\ \forall \boldsymbol{s} \in \mathcal{V}\,,\ \Pi_{\mathcal{D}}^{\mathrm{ML}} v = v_{\boldsymbol{s}} \text{ on } \Omega_{\boldsymbol{s}}.$$

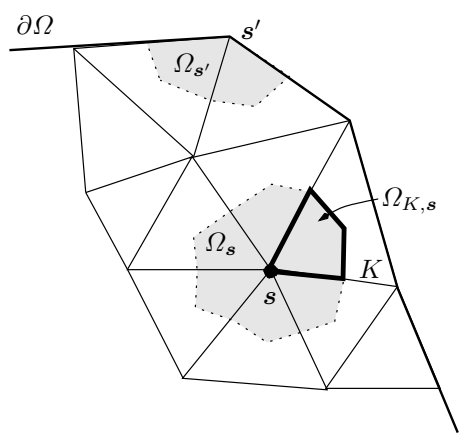


Fig. 8.3. Partitions for the mass-lumping of the $\mathbb{P}_1$ GD.

The mesh $(\Omega_{\boldsymbol{s}})_{\boldsymbol{s}\in\mathcal{V}}$ thus constructed is sometimes called the barycentric dual mesh, or Donald dual mesh, of $\mathfrak{T}$. This is only one possible mesh that can be

used to create a mass-lumped version of the $\mathbb{P}_1$ GD on $\mathfrak{T}$. It is however the most classical one; indeed, for the time-dependent heat equation, this choice leads to the usual mass lumping as known to the finite element community, which consists in replacing the mass matrix $M = (M_{ij})$ (resulting from the discretisation of the time derivative term) by the diagonal matrix D with coefficients $D_{ii} = \sum_j M_{ij}$, see e.g. [136, Chapter 15].

The properties of this mass-lumped $\mathbb{P}_1$ GD are stated in the following theorem.

Theorem 8.17 (Properties of mass-lumped $\mathbb{P}_1$ GDs). *Let $(\mathfrak{T}_m)_{m\in\mathbb{N}}$ be a sequence of conforming simplicial meshes of Ω in the sense of Definition 7.4, and let $(\mathcal{D}_m^{\mathrm{ML}})_{m\in\mathbb{N}}$ be the corresponding mass-lumped $\mathbb{P}_1$ GDs given by Definition 8.16. Assume that $\sup_{m\in\mathbb{N}} \kappa_{\mathfrak{T}_m} < +\infty$ (see (7.10)), and that $h_{\mathcal{M}_m} \to 0$ as $m \to \infty$.*
Then $(\mathcal{D}_m^{\mathrm{ML}})_{m\in\mathbb{N}}$ is coercive, GD-consistent, limit-conforming, compact, and has a piecewise constant reconstruction in the sense of Definitions 2.2, 2.4, 2.5, 2.8 and 2.12.

Proof. Let us assume that

$$\forall v \in X_{\mathcal{D}_m,0}\,,\ \left\|\Pi_{\mathcal{D}_m} v - \Pi_{\mathcal{D}_m}^{\mathrm{ML}} v\right\|_{L^p(\Omega)} \le h_{\mathcal{M}_m} \left\|\nabla_{\mathcal{D}_m} v\right\|_{L^p(\Omega)^d}. \tag{8.18}$$

Then the conclusion of the theorem follows from Theorem 8.10 (which states that the underlying sequence of $\mathbb{P}_1$ GDs $(\mathcal{D}_m)_{m\in\mathbb{N}}$ is coercive, GD-consistent, limit-conforming and compact) and Theorem 7.49.
The proof of (8.18) is done by way of simple Taylor expansion in each $\Omega_{K,\boldsymbol{s}}$. Indeed, since $\Pi_{\mathcal{D}_m} v$ is linear in $K \supset \Omega_{K,\boldsymbol{s}}$ with $\nabla(\Pi_{\mathcal{D}_m} v) = (\nabla_{\mathcal{D}_m} v)_{|K}$, and since $\Pi_{\mathcal{D}_m}^{\mathrm{ML}} v = v(\boldsymbol{s}) = \Pi_{\mathcal{D}_m} v(\boldsymbol{s})$ in $\Omega_{\boldsymbol{s}} \supset \Omega_{K,\boldsymbol{s}}$, we have, for $\boldsymbol{x} \in \Omega_{K,\boldsymbol{s}}$,

$$\begin{aligned}\Pi_{\mathcal{D}_m}^{\mathrm{ML}} v(\boldsymbol{x}) - \Pi_{\mathcal{D}_m} v(\boldsymbol{x}) &= \Pi_{\mathcal{D}_m} v(\boldsymbol{s}) - \Pi_{\mathcal{D}_m} v(\boldsymbol{x}) \\ &= (\nabla_{\mathcal{D}_m} v)_{|K} \cdot (\boldsymbol{s} - \boldsymbol{x}) = \nabla_{\mathcal{D}_m} v(\boldsymbol{x}) \cdot (\boldsymbol{s} - \boldsymbol{x}).\end{aligned}$$

Hence,

$$|\Pi_{\mathcal{D}_m}^{\mathrm{ML}} v(\boldsymbol{x}) - \Pi_{\mathcal{D}_m} v(\boldsymbol{x})| \le h_{\mathcal{M}_m} |\nabla_{\mathcal{D}_m} v(\boldsymbol{x})|. \tag{8.19}$$

This estimate is valid for any $\boldsymbol{x} \in \Omega_{K,\boldsymbol{s}}$, any $K \in \mathcal{M}_m$ and any $\boldsymbol{s} \in \mathcal{V}_K$. Hence, it is valid for any $\boldsymbol{x} \in \Omega$. Raised to the power p and integrated over $\boldsymbol{x} \in \Omega$, (8.19) gives (8.18). ∎

Remark 8.18. If $p > d/2$, by Proposition 7.64 the $\mathbb{P}_1$ gradient discretisations on $(\mathfrak{T}_m)_{m\in\mathbb{N}}$ satisfy $S_{\mathcal{D}_m}(\varphi) \le C h_{\mathcal{M}_m} \|\varphi\|_{W^{2,p}(\Omega)}$ for all $\varphi \in W^{2,p}(\Omega)$ (with C not depending on m or φ), and $W_{\mathcal{D}_m}(\boldsymbol{\varphi}) = 0$ for all $\boldsymbol{\varphi} \in W_{\mathrm{div}}^{p'}(\Omega)$.
Estimate (8.18) shows that (7.59) holds (with $\mathcal{D}_m^\star = \mathcal{D}_m^{\mathrm{ML}}$) with $\omega_m = h_{\mathcal{M}_m}$. Combined with the previous estimates on $S_{\mathcal{D}_m}$ and $W_{\mathcal{D}_m}$, and with (7.63) and (7.64) in Remark 7.51, this proves that the mass-lumped $\mathbb{P}_1$ gradient discretisations satisfy

$$S_{\mathcal{D}_m^{\mathrm{ML}}}(\varphi) \leq C' h_{\mathcal{M}_m} \|\varphi\|_{W^{2,p}(\Omega)}$$

(with C' not depending on m or φ), and

$$W_{\mathcal{D}_m^{\mathrm{ML}}}(\boldsymbol{\varphi}) \leq h_{\mathcal{M}_m} \|\mathrm{div}\boldsymbol{\varphi}\|_{L^{p'}(\Omega)}.$$

Hence, as expected, mass-lumped $\mathbb{P}_1$ GSs are order 1 schemes. More precisely, if the exact solution to the linear elliptic problem (2.20) belongs to H^2 and $d = 1, 2, 3$, then the estimates (2.25) and (2.26) are $\mathcal{O}(h_{\mathcal{M}})$ when the mass-lumped $\mathbb{P}_1$ GD is used in the GS (2.23).

8.5 Vertex approximate gradient (VAG) methods

Successive versions of the VAG schemes have been described in several papers [99, 101]. VAG methods stem from the idea that it is often computationally efficient to have all unknowns located at the vertices of the mesh, especially with tetrahedral meshes (which have much less vertices than cells). It is however known that schemes with discrete unknowns at the vertices may lead to unacceptable results for the transport of a species in a heterogeneous domain, in particular for coarse meshes (one layer of mesh for one homogeneous layer, for example). The VAG schemes are an answer to this conundrum. After all possible local eliminations, the VAG schemes only have vertex unknowns, and have been shown to cure the numerical issues for coarse meshes and heterogeneous media [101, 100, 102]; this is due to a specific mass-lumping that spreads the reconstructed function between the centre of the control volumes and the vertices. Let us remark that the original version of the VAG scheme in [99] uses the same nodal formalism as in Chapter 14, but was shown in the FVCA6 3D Benchmark [104] to be less precise than the version presented here [98].

Starting from a generic polytopal mesh $\mathfrak{T}$, the VAG scheme is defined as a barycentric condensation and a mass-lumping of the $\mathbb{P}_1$ GD on a conforming simplicial sub-mesh of $\mathfrak{T}$. We consider here the situation of homogeneous Dirichlet boundary conditions and space dimension 3; other boundary conditions and/or dimension 2 are easy adaptations.

1. Let $\mathfrak{T} = (\mathcal{M}, \mathcal{F}, \mathcal{P}, \mathcal{V})$ be a polytopal mesh of Ω in the sense of Definition 7.2, except for the removal of the hypothesis that the faces $\sigma \in \mathcal{F}$ are planar. We define a conforming simplicial (tetrahedral in 3D) sub-mesh by the following procedure. For any $K \in \mathcal{M}$, any $\sigma \in \mathcal{F}_K$, and any $\boldsymbol{s}, \boldsymbol{s}' \in \mathcal{V}_\sigma$ such that $[\boldsymbol{s}, \boldsymbol{s}']$ is an edge of σ, we define the simplex $T_{K,\sigma,\boldsymbol{s},\boldsymbol{s}'}$ by its four vertices $\boldsymbol{x}_K, \boldsymbol{x}_\sigma, \boldsymbol{s}, \boldsymbol{s}'$ (see Figure 8.5), where the point $\boldsymbol{x}_\sigma$ corresponding to the face σ is given by

$$\boldsymbol{x}_\sigma = \frac{1}{\mathrm{Card}(\mathcal{V}_\sigma)} \sum_{\boldsymbol{s} \in \mathcal{V}_\sigma} \boldsymbol{s}. \tag{8.20}$$

We denote by $\mathfrak{T}^T$ the conforming simplicial mesh (as per Definition 7.4) defined by the simplices $T_{K,\sigma,\boldsymbol{s},\boldsymbol{s}'}$. More precisely, $\mathfrak{T}^T = (\mathcal{M}^T, \mathcal{F}^T, \mathcal{P}^T, \mathcal{V}^T)$ with

- $\mathcal{M}^T$ is the set

$$\mathcal{M}^T = \{T_{K,\sigma,\boldsymbol{s},\boldsymbol{s}'} : K \in \mathcal{M}\,,\ \sigma \in \mathcal{F}_K\,, \\ (\boldsymbol{s},\boldsymbol{s}') \in \mathcal{V}^2 \text{ such that } [\boldsymbol{s},\boldsymbol{s}'] \text{ is an edge of } \sigma\},$$

- $\mathcal{F}^T$ is the set of all faces of the simplices in $\mathcal{M}^T$,
- $\mathcal{P}^T$ is an arbitrary set of centres of the simplices (they do not play any role in the construction of the scheme),
- $\mathcal{V}^T$ is the set of all vertices of the simplices in $\mathcal{M}^T$; this means that

$$\mathcal{V}^T = \mathcal{P} \cup \mathcal{V} \cup \{\boldsymbol{x}_\sigma \,:\, \sigma \in \mathcal{F}\}. \tag{8.21}$$

2. We let $\overline{\mathcal{D}} = (X_{\overline{\mathcal{D}},0}, \nabla_{\overline{\mathcal{D}}}, \Pi_{\overline{\mathcal{D}}})$ be the $\mathbb{P}_1$ GD defined from $\mathfrak{T}^T$ as in Section 8.2.1 for $k = 1$. Given (8.21), for $\overline{\mathcal{D}}$ we can define the set I of geometrical entities attached to the discrete unknowns by $I = \mathcal{M} \cup \mathcal{V} \cup \mathcal{F}$, and the set of S of approximation points of is $S = ((\boldsymbol{x}_K)_{K\in\mathcal{M}}, (\boldsymbol{s})_{\boldsymbol{s}\in\mathcal{V}}, (\boldsymbol{x}_\sigma)_{\sigma\in\mathcal{F}})$.
3. We define a barycentric condensation $\overline{\mathcal{D}}^{\text{BA}}$ of $\overline{\mathcal{D}}$ (see Definition 7.40) which consists in eliminating the discrete unknowns attached to the internal faces $\mathcal{F}_{\text{int}}$ of $\mathfrak{T}$. Precisely, we let $I^{\text{BA}} = \mathcal{M} \cup \mathcal{V} \cup \mathcal{F}_{\text{ext}}$ and, for $\sigma \in \mathcal{F}_{\text{int}}$, we set $H_\sigma = \mathcal{V}_\sigma$ and we define the coefficients $\beta^\sigma_{\boldsymbol{s}} = 1/\text{Card}(\mathcal{V}_\sigma)$, for all $\boldsymbol{s} \in \mathcal{V}_\sigma$. These coefficients are precisely the ones appearing in (8.20). The mapping $v \in X_{\mathcal{D}^{\text{BA}},0} \mapsto \widetilde{v} \in X_{\mathcal{D},0}$ described by (7.47) is therefore given by $\widetilde{v} = ((\widetilde{v}_K)_{K\in\mathcal{M}}, (\widetilde{v}_{\boldsymbol{s}})_{\boldsymbol{s}\in\mathcal{V}}, (\widetilde{v}_\sigma)_{\sigma\in\mathcal{F}})$ with

$$\begin{aligned} &\forall K \in \mathcal{M}\,, && \widetilde{v}_K = v_K\,, \\ &\forall \boldsymbol{s} \in \mathcal{V}\,, && \widetilde{v}_{\boldsymbol{s}} = v_{\boldsymbol{s}}\,, \\ &\forall \sigma \in \mathcal{F}_{\text{ext}}\,, && \widetilde{v}_\sigma = v_\sigma = 0\,, \\ &\forall \sigma \in \mathcal{F}_{\text{int}}\,, && \widetilde{v}_\sigma = \frac{1}{\text{Card}(\mathcal{V}_\sigma)} \sum_{\boldsymbol{s}\in\mathcal{V}_\sigma} v_{\boldsymbol{s}}. \end{aligned} \tag{8.22}$$

4. The VAG GD is the gradient discretisation $\mathcal{D}$ obtained from $\overline{\mathcal{D}}^{\text{BA}}$ by performing a mass-lumping in the sense of Definition 7.45. We therefore have $I = \mathcal{M} \cup \mathcal{V} \cup \mathcal{F}_{\text{ext}}$, $I_\Omega = \mathcal{M} \cup (\mathcal{V} \cap \Omega)$ and $I_\partial = (\mathcal{V} \cap \partial\Omega) \cup \mathcal{F}_{\text{ext}}$, which gives

$$X_{\mathcal{D},0} = \{v = ((v_K)_{K\in\mathcal{M}}, (v_{\boldsymbol{s}})_{\boldsymbol{s}\in\mathcal{V}}, (v_\sigma)_{\sigma\in\mathcal{F}_{\text{ext}}}) \,:\, v_K \in \mathbb{R} \text{ for all } K \in \mathcal{M}\,, \\ v_{\boldsymbol{s}} \in \mathbb{R} \text{ for all } \boldsymbol{s} \in \mathcal{V} \cap \Omega\,,\ v_{\boldsymbol{s}} = 0 \text{ for all } \boldsymbol{s} \in \mathcal{V} \cap \partial\Omega\,, \\ v_\sigma = 0 \text{ for all } \sigma \in \mathcal{F}_{\text{ext}}\}.$$

To perform the mass-lumping of $\overline{\mathcal{D}}^{\text{BA}}$, we start by splitting each simplex $T_{K,\sigma,\boldsymbol{s},\boldsymbol{s}'}$ into three parts $T^K_{K,\sigma,\boldsymbol{s},\boldsymbol{s}'}$, $T^{\boldsymbol{s}}_{K,\sigma,\boldsymbol{s},\boldsymbol{s}'}$, and $T^{\boldsymbol{s}'}_{K,\sigma,\boldsymbol{s},\boldsymbol{s}'}$ (whose

detailed geometry is not needed), that respectively contain in their closure $\boldsymbol{x}_K$, $\boldsymbol{s}$ and $\boldsymbol{s}'$. We then let, in Definition 2.12, Ω_K be the union of all $(T^K_{K,\sigma,\boldsymbol{s},\boldsymbol{s}'})_{\sigma,\boldsymbol{s},\boldsymbol{s}'}$, and $\Omega_{\boldsymbol{s}}$ be the union of all $(T^{\boldsymbol{s}}_{K,\sigma,\boldsymbol{s},\boldsymbol{s}'})_{K,\sigma,\boldsymbol{s}'}$. This leads to

$$\forall v \in X_{\mathcal{D},0} \,:\, \Pi_{\mathcal{D}} v = \sum_{K\in\mathcal{M}} v_K \mathbf{1}_{\Omega_K} + \sum_{\boldsymbol{s}\in\mathcal{V}} v_{\boldsymbol{s}} \mathbf{1}_{\Omega_{\boldsymbol{s}}}. \tag{8.23}$$

The gradient reconstruction is not modified by the mass-lumping, and therefore $\nabla_{\mathcal{D}} v$ is equal, in a tetrahedron $T_{K,\sigma,\boldsymbol{s},\boldsymbol{s}'}$, to the gradient of the affine functions that takes values $(v_K, \widetilde{v}_\sigma, v_{\boldsymbol{s}}, v_{\boldsymbol{s}'})$ at the vertices $(\boldsymbol{x}_K, \overline{\boldsymbol{x}}_\sigma, \boldsymbol{s}, \boldsymbol{s}')$ of $T_{K,\sigma,\boldsymbol{s},\boldsymbol{s}'}$.

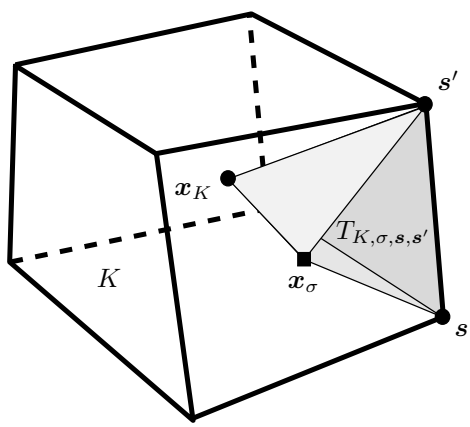


Fig. 8.4. Definition of a simplex $T_{K,\sigma,\boldsymbol{s},\boldsymbol{s}'}$ in a mesh cell K.

Remark 8.19 (Elimination of the cell unknowns in the VAG GS by static condensation)

Apply the VAG GDM to obtain a GS (2.23) (with $\boldsymbol{F} = 0$ to simplify the presentation), and take in this scheme the test function $v \in X_{\mathcal{D},0}$ which satisfies $v_K = 1$ for a given cell K, $v_L = 0$ for all other cells, and $v_{\boldsymbol{s}} = 0$ for all vertices. Then, the integral in the right-hand side of (2.23) can be reduced to K. Let α_K be the $\mathbb{P}_1$ Lagrange interpolator in the tetrahedra $(T_{K,\sigma,\boldsymbol{s},\boldsymbol{s}'})_{\sigma,\boldsymbol{s},\boldsymbol{s}'}$, with value 1 at $\boldsymbol{x}_K$ and 0 at all other vertices of these tetrahedra; then

$$\nabla_{\mathcal{D}} u = u_K \nabla \alpha_K + \sum_{\boldsymbol{s}\in\mathcal{V}_K} u_{\boldsymbol{s}} \Theta_{\boldsymbol{s}} \text{ in } K,$$

for some functions $\Theta_{\boldsymbol{s}}$ (involving the Lagrange interpolators at the vertices of the tetrahedra contained in K). Since $\nabla_{\mathcal{D}} v = \nabla \alpha_K$, we infer that

$$u_K \int_K \Lambda(\boldsymbol{x})\nabla\alpha_K(\boldsymbol{x})\cdot\nabla\alpha_K(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = \int_\Omega f(\boldsymbol{x})\alpha_K(\boldsymbol{x})\mathrm{d}\boldsymbol{x} - \sum_{\boldsymbol{s}\in\mathcal{V}_K} u_{\boldsymbol{s}} \int_K \Lambda(\boldsymbol{x})\Theta_{\boldsymbol{s}}(\boldsymbol{x})\cdot\nabla\alpha_K(\boldsymbol{x})\mathrm{d}\boldsymbol{x}.$$

The coefficient of u_K in the left-hand side is not zero, so this relation yields an expression of u_K in terms of $(u_{\boldsymbol{s}})_{\boldsymbol{s}\in\mathcal{V}_K}$ (and the source term f), without even having to solve a local system.
Hence, when using the VAG GD to obtain GS for a linear elliptic problem, the cell unknowns can be locally eliminated and expressed in terms of the neighbouring vertex unknowns.

Lemma 8.20 (Control of $\mathrm{reg}_{\mathrm{BA}}$ for VAG GD). *Let $\mathfrak{T}$ be a polytopal mesh of Ω in the sense of Definition 7.2, and let $\mathfrak{T}^T$ be the conforming simplicial sub-mesh $\mathfrak{T}$ as in Item 1 above. We take $\varrho \geq \kappa_{\mathfrak{T}^T}$ (see (7.10)). Let $\overline{\mathcal{D}}^{\mathrm{BA}}$ be the barycentric condensation, defined in Item 3 above, of the $\mathbb{P}_1$ GD on $\mathfrak{T}^T$.*
Then there exists C_6 depending only on ϱ such that $\mathrm{reg}_{\mathrm{BA}}(\overline{\mathcal{D}}^{\mathrm{BA}}) \leq C_6$.

Proof. The proof is made in several steps. Here, we write $a \lesssim b$ for $a \leq Cb$ for some $C > 0$ depending only on ϱ. We write $a \approx b$, and we say that a and b are comparable, if $a \lesssim b$ and $b \lesssim a$.

Step 1: *The length of any edge in any tetrahedron $T \in \mathcal{M}^T$ is comparable to the diameter h_T of the tetrahedron.*
Let τ be a face of T and let $\boldsymbol{s}$ be the opposite vertex. Let $B(\boldsymbol{x}_T, \rho_T)$ be the centre of the largest ball included in T; by definition of $\kappa_{\mathfrak{T}^T}$ we have $\rho_T \approx h_T$. Let $(\boldsymbol{s}_i)_{i=1,\ldots,d}$ be the vertices of τ. We write $\boldsymbol{x}_T$ as a convex combination $\boldsymbol{x}_T = \lambda\boldsymbol{s} + \sum_{i=1}^d \lambda_i\boldsymbol{s}_i$. Let $\boldsymbol{n}_{T,\tau}$ be the outer normal to T on τ. For any $\boldsymbol{s}'$ vertex of τ, since $(\boldsymbol{s}' - \boldsymbol{s}_i)\perp\boldsymbol{n}_{T,\tau}$ for all $i = 1,\ldots,d$, and $\lambda + \sum_{i=1}^d \lambda_i = 1$, we have

$$\begin{aligned}(\boldsymbol{x}_T - \boldsymbol{s}')\cdot\boldsymbol{n}_{T,\tau} &= \lambda(\boldsymbol{s} - \boldsymbol{s}')\cdot\boldsymbol{n}_{T,\tau} + \sum_{i=1}^d \lambda_i(\boldsymbol{s}_i - \boldsymbol{s}')\cdot\boldsymbol{n}_{T,\tau}\\ &= \lambda(\boldsymbol{s} - \boldsymbol{s}')\cdot\boldsymbol{n}_{T,\tau}.\end{aligned}$$

We have $(\boldsymbol{x}_T - \boldsymbol{s}')\cdot\boldsymbol{n}_{T,\tau} = \mathrm{dist}(\boldsymbol{x}_T, \tau) \geq \rho_T \approx h_T$, and therefore

$$h_T \lesssim \lambda(\boldsymbol{s} - \boldsymbol{s}')\cdot\boldsymbol{n}_{T,\tau} \leq (\boldsymbol{s} - \boldsymbol{s}')\cdot\boldsymbol{n}_{T,\tau}. \tag{8.24}$$

Therefore, $h_T \lesssim |\boldsymbol{s}' - \boldsymbol{s}|$. Since we also have $|\boldsymbol{s}' - \boldsymbol{s}| \leq h_T$, we infer that

$$\text{The length of any edge of a tetrahedron } T \in \mathcal{M}^T \text{ is comparable to } h_T. \tag{8.25}$$

Step 2: *If $\sigma \in \mathcal{F}$ and h_σ is the maximal distance between two of its vertices, then h_σ is comparable to the diameter of any tetrahedron $T \in \mathcal{M}^T$ having its base on σ.*

Recall that a face σ does not need to be planar. Let T be a tetrahedron with its face on σ, and denote by $K \in \mathcal{M}$ the cell that contains T. We first notice that any two tetrahedra $T_1, T_2 \in \mathcal{M}^T$ in K having their base on σ share the common edge $[\boldsymbol{x}_K, \overline{\boldsymbol{x}}_\sigma]$ and thus, by (8.25),

$$h_{T_1} \approx |\boldsymbol{x}_K - \overline{\boldsymbol{x}}_\sigma| \approx h_{T_2}. \tag{8.26}$$

We have $h_\sigma = |\boldsymbol{s}_1 - \boldsymbol{s}_2|$ for some vertices $\boldsymbol{s}_i$ of σ. Let us take $T_1, T_2 \in \mathcal{M}^T$ tetrahedra in K with base on σ and having respectively $\boldsymbol{s}_1$ and $\boldsymbol{s}_2$ as vertices. Using (8.26) we have

$$h_\sigma = |\boldsymbol{s}_1 - \boldsymbol{s}_2| \le |\boldsymbol{s}_1 - \overline{\boldsymbol{x}}_\sigma| + |\overline{\boldsymbol{x}}_\sigma - \boldsymbol{s}_2| \le h_{T_1} + h_{T_2} \approx h_T. \tag{8.27}$$

Any edge of σ is also an edge of a tetrahedron with base on σ. Properties (8.25) and (8.27) therefore give

$$\text{The length of any edge of } \sigma \text{ is comparable to } h_\sigma. \tag{8.28}$$

Finally, T shares an edge with σ. Hence, (8.25) and (8.28) show that

$$\text{For any tetrahedron } T \in \mathcal{M}^T \text{ having its base on } \sigma,\ h_T \approx h_\sigma. \tag{8.29}$$

Step 3: *conclusion.*
The mesh corresponding to the $\mathbb{P}_1$ gradient discretisation $\overline{\mathcal{D}}$ on $\mathfrak{T}^T$ is $\mathcal{M}^T$. Hence, a cell of this mesh is a tetrahedron T with its base on some $\sigma \in \mathcal{F}$, and the only unknown that is eliminated in I_K (from the $\mathbb{P}_1$ GD) is the unknown at $\overline{\boldsymbol{x}}_\sigma$. This elimination is done by using the vertices of σ and, by (8.29), these vertices all lie within distance $h_\sigma \approx h_T$ of the points in T. Moreover, since $\beta^\sigma_{\boldsymbol{s}} = 1/\mathrm{Card}(\mathcal{V}_\sigma)$ we have $\sum_{\boldsymbol{s}\in\mathcal{V}_\sigma} |\beta^\sigma_{\boldsymbol{s}}| = 1$.
These properties give a bound on $\mathrm{reg}_{\mathrm{BA}}(\overline{\mathcal{D}}^{\mathrm{BA}})$ that depends only on ϱ (through the relations $\approx$). ∎

Remark 8.21 (Comparison between h_T and h_K)
If ϱ is also an upper bound of $\max_{K\in\mathcal{M}} \mathrm{Card}(\mathcal{F}_K)$, then by working neighbour to neighbour it can be shown that any tetrahedra $T \in \mathcal{M}^T$ in a cell $K \in \mathcal{M}$ has a size $h_T \approx h_K$.

Theorem 8.22 (Properties of VAG GDs). *Let $(\mathfrak{T}_m)_{m\in\mathbb{N}}$ be a sequence of polytopal meshes of Ω in the sense of Definition 7.2. For each $m \in \mathbb{N}$ we define the conforming simplicial sub-mesh $\mathfrak{T}^T_m$ of $\mathfrak{T}_m$ as in Item 1 above. Assume that $h_{\mathcal{M}_m} \to 0$ as $m \to \infty$, and that $(\kappa_{\mathfrak{T}^T_m})_{m\in\mathbb{N}}$ is bounded (see (7.10)). Let $\mathcal{D}_m$ be the VAG GD built on $\mathfrak{T}_m$.*
Then $(\mathcal{D}_m)_{m\in\mathbb{N}}$ is coercive, GD-consistent, limit-conforming, compact and has a piecewise constant reconstruction in the sense of Definitions 2.2, 2.4, 2.5, 2.8 and 2.12.

Proof. Let $\overline{\mathcal{D}}_m$ be the $\mathbb{P}_1$ GD on $\mathfrak{T}_m^T$. Since $\mathcal{D}_m$ is the mass-lumping of the barycentric condensation $\overline{\mathcal{D}}_m^{\text{BA}}$ of $\overline{\mathcal{D}}_m$, the result follows from Theorems 7.43 and 7.49 if we can prove that $\text{reg}_{\text{LLE}}(\overline{\mathcal{D}}_m)$ and $\text{reg}_{\text{BA}}(\overline{\mathcal{D}}_m^{\text{BA}})$ remain bounded, and that the following version of (7.58) holds:

$$\forall v \in X_{\mathcal{D}_m,0}\,,\ \left\|\Pi_{\mathcal{D}_m} v - \Pi_{\overline{\mathcal{D}}_m^{\text{BA}}} v\right\|_{L^p(\Omega)} \le h_{\mathcal{M}_m} \left\|\nabla_{\overline{\mathcal{D}}_m^{\text{BA}}} v\right\|_{L^p(\Omega)^d}. \tag{8.30}$$

Since $(\kappa_{\mathfrak{T}_m^T})_{m\in\mathbb{N}}$ is bounded, the boundedness of $\text{reg}_{\text{LLE}}(\overline{\mathcal{D}}_m)$ follows from Lemma 8.9. The bound on $\text{reg}_{\text{BA}}(\overline{\mathcal{D}}_m^{\text{BA}})$ follows from Lemma 8.20. To prove (8.30), we use the same technique as for the mass-lumping of $\mathbb{P}_1$ GDs. In each $T^K_{K,\sigma,\boldsymbol{s},\boldsymbol{s}'}$ (resp. $T^{\boldsymbol{s}}_{K,\sigma,\boldsymbol{s},\boldsymbol{s}'}$), $\Pi_{\overline{\mathcal{D}}_m^{\text{BA}}} v$ is linear, $\nabla_{\overline{\mathcal{D}}_m^{\text{BA}}} v = \nabla(\Pi_{\overline{\mathcal{D}}_m^{\text{BA}}} v)$ and $\Pi_{\mathcal{D}_m} v$ is equal to $v_K = \Pi_{\overline{\mathcal{D}}_m^{\text{BA}}} v(\boldsymbol{x}_K)$ (resp. $v_{\boldsymbol{s}} = \Pi_{\overline{\mathcal{D}}_m^{\text{BA}}} v(\boldsymbol{s})$). Thus, in each $T^K_{K,\sigma,\boldsymbol{s},\boldsymbol{s}'}$ and $T^{\boldsymbol{s}}_{K,\sigma,\boldsymbol{s},\boldsymbol{s}'}$, that is, on the whole of Ω,

$$\left|\Pi_{\mathcal{D}_m} v - \Pi_{\overline{\mathcal{D}}_m^{\text{BA}}} v\right| \le h_{\mathcal{M}_m} \left|\nabla_{\overline{\mathcal{D}}_m^{\text{BA}}} v\right|. \tag{8.31}$$

We then conclude the proof of (8.30) by taking the $L^p(\Omega)$ norms in (8.31). ∎

Theorem 8.23 (Estimates on $S_{\mathcal{D}}$ and $W_{\mathcal{D}}$ for VAG GD). *Let $\mathfrak{T}$ be a polytopal mesh of Ω in the sense of Definition 7.2, and $\mathfrak{T}^T$ be the conforming simplicial sub-mesh of $\mathfrak{T}_m$ as in Item 1 above. We take*

$$\varrho \ge \kappa_{\mathfrak{T}^T} + \max_{K\in\mathcal{M}} \text{Card}(\mathcal{V}_K),$$

and we let $\mathcal{D}$ be the VAG GD built on $\mathfrak{T}$. Then, there exists C_7 depending only on d, p, Ω and ϱ such that

$$C_{\mathcal{D}} \le C_7, \tag{8.32}$$

$$\forall \boldsymbol{\varphi} \in W^{p'}_{\text{div}}(\Omega)\,,\ W_{\mathcal{D}}(\boldsymbol{\varphi}) \le h_{\mathcal{M}} \left\|\text{div}\boldsymbol{\varphi}\right\|_{L^{p'}(\Omega)} \tag{8.33}$$

and

$$\forall \varphi \in W^{1,p}_0(\Omega) \cap W^{2,p}(\Omega)\,,\ S_{\mathcal{D}}(\varphi) \le C_7 h_{\mathcal{M}} \left\|\varphi\right\|_{W^{2,p}(\Omega)} \tag{8.34}$$

Note that, in practice, the uniform bound on the number of vertices of each cell, implied by ϱ, is not a restrictive assumption.

Proof. By (7.62) and (8.30) (which shows that we can take $\omega_m = h_{\mathcal{M}_m} \le \text{diam}(\Omega)$ in (7.59) with $\mathcal{D}_m^\star = \mathcal{D}$ and $\mathcal{D}_m = \overline{\mathcal{D}}^{\text{BA}}$), we have $C_{\mathcal{D}} \le \text{diam}(\Omega) + C_{\overline{\mathcal{D}}^{\text{BA}}}$. We then use (7.57) to get $C_{\overline{\mathcal{D}}^{\text{BA}}} \le C_{\overline{\mathcal{D}}} \le C_P$, where C_P depends only on d, p and Ω (we can actually take $C_P = \text{diam}(\Omega)$, an upper bound of Poincaré's constant in $W^{1,p}_0(\Omega)$ – remember that $\overline{\mathcal{D}}$ is the $\mathbb{P}_1$ GD). This gives (8.32). Similarly, Estimate (8.33) follows from (7.64) (in which we can take $\omega_m = h_{\mathcal{M}_m}$ by (8.30)) and from (7.57), which shows that $W_{\overline{\mathcal{D}}^{\text{BA}}} \le W_{\overline{\mathcal{D}}} = 0$.

Owing to (7.63) with $\mathcal{D}_m^\star = \mathcal{D}$ and $\mathcal{D}_m = \overline{\mathcal{D}}^{\text{BA}}$, and to (8.30), to prove (8.34) it suffices to show that $S_{\overline{\mathcal{D}}^{\text{BA}}}(\varphi) \le C_8 h_{\mathcal{M}} \|\varphi\|_{W^{2,p}(\Omega)}$ with C_8 depends only on p, d, Ω and ϱ. This estimate is obtained by using Proposition 7.66, provided that we find sets $(V_K)_{K\in\mathcal{M}}$ that satisfy (7.92) and (7.93), with $\mathcal{D} = \overline{\mathcal{D}}^{\text{BA}}$ and θ depending only on d, p, Ω and ϱ.

We first notice that the bound on $\text{reg}_{\text{LLE}}(\overline{\mathcal{D}}^{\text{BA}})$ in (7.93) is a consequence of Lemma 7.42, Lemma 8.9 (with $\mathcal{D} = \overline{\mathcal{D}}$ the $\mathbb{P}_1$ GD on $\mathfrak{T}^T$), and Lemma 8.20.

Each cell of the mesh $\mathcal{M}^T$ associated to $\overline{\mathcal{D}}^{\text{BA}}$ is a tetrahedron $T_{K,\sigma,\boldsymbol{s},\boldsymbol{s}'}$ in a certain cell $K \in \mathcal{M}$. In Proposition 7.66, set $V_{T_{K,\sigma,\boldsymbol{s},\boldsymbol{s}'}} = K$. Each $\boldsymbol{x} \in \Omega$ belongs to a single cell $K \in \mathcal{M}$, and can therefore only be in $V_{T_{K,\sigma,\boldsymbol{s},\boldsymbol{s}'}}$ for $T_{K,\sigma,\boldsymbol{s},\boldsymbol{s}'}$ a tetrahedron contained in K. The bound $\text{Card}(\mathcal{V}_K) \le \varrho$ ensures that the number of such tetrahedra, and thus $\text{Card}(\{T_{K,\sigma,\boldsymbol{s},\boldsymbol{s}'} \in \mathcal{M}^T : \boldsymbol{x} \in V_{T_{K,\sigma,\boldsymbol{s},\boldsymbol{s}'}}\})$, is bounded above by some constant depending only on ϱ. This takes care of the last term in (7.93). It therefore remains to prove that each $V_{T_{K,\sigma,\boldsymbol{s},\boldsymbol{s}'}} = K$ is star-shaped with respect to a ball $B_{T_{K,\sigma,\boldsymbol{s},\boldsymbol{s}'}} \subset T_{K,\sigma,\boldsymbol{s},\boldsymbol{s}'}$ such that $\text{diam}(B_{T_{K,\sigma,\boldsymbol{s},\boldsymbol{s}'}}) \ge C_9 \text{diam}(K)$ with C_9 depending only on ϱ. As in the proof of Lemma 8.20, in the following we denote $a \lesssim b$ for $a \le Cb$ with C depending only on ϱ, and $a \approx b$ for $a \lesssim b$ and $b \lesssim a$. From now on, we also set $T = T_{K,\sigma,\boldsymbol{s},\boldsymbol{s}'}$.

In the current setting, the faces $\sigma \in \mathcal{F}_K$ of K may not be planar. However, in the construction of $\mathcal{M}^T$ each of these faces have been split into triangles that are necessarily planar. Hence, we can consider the cell K to be polytopal, with planar faces the bases of the tetrahedra of $\mathcal{M}^T$ contained in K. If τ is the basis on σ of T, applying (8.24) with $\boldsymbol{s} = \boldsymbol{x}_K$ (which is indeed the vertex opposite to τ in T) and any vertex $\boldsymbol{s}'$ of τ shows that

$$h_T \lesssim (\boldsymbol{x}_K - \boldsymbol{s}) \cdot \boldsymbol{n}_{T,\tau} = d_{K,\tau}.$$

Since $\text{Card}(\mathcal{V}_K) \lesssim 1$ we have $\text{Card}(\mathcal{F}_K) \lesssim 1$ and Remark 8.21 can be invoked. This gives

$$h_K \approx h_T \lesssim d_{K,\tau}. \tag{8.35}$$

Then Lemma B.1 shows that K is star-shaped with respect to a ball $B(\boldsymbol{x}_K, r_K)$ with $r_K \approx h_K$ (see Figure 8.5 for an illustration).

Since $\kappa_{\mathfrak{T}^T} \le \varrho$, T contains a ball $B(\boldsymbol{y}_T, \rho_T)$ with, owing to (8.35),

$$\rho_T \approx h_T \approx h_K \approx r_K. \tag{8.36}$$

We now find B_T – mentioned in (7.92) – by an homothetic transformation of $B(\boldsymbol{y}_T, \rho_T)$. Let $\mu \in (0,1)$ and let

$$B_T = (1-\mu)\boldsymbol{x}_K + \mu B(\boldsymbol{y}_T, \rho_T) = B((1-\mu)\boldsymbol{x}_K + \mu \boldsymbol{y}_T, \mu\rho_T).$$

Since K is convex, $B_T \subset K$. If $B_T \subset B(\boldsymbol{x}_K, r_K)$, then $V_T = K$ is indeed star-shaped with respect to B_T. Since $\text{diam}(B_T) = 2\mu\rho_T \approx \mu h_K = \mu \text{diam}(V_T)$, this shows that the second term in the right-hand side of (7.93) is bounded

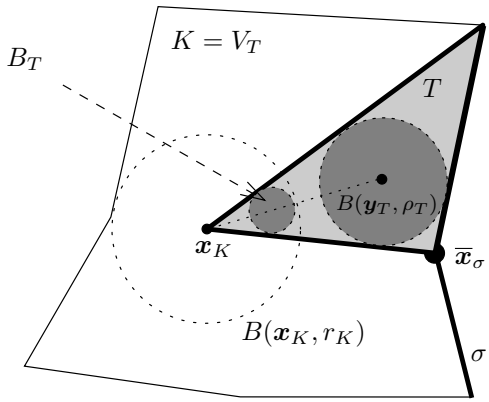


Fig. 8.5. Illustration of the proof of Theorem 8.23. This figure is a planar section of a cell K.

by $\mu^{-1}C_{10}$ with C_{10} depending only on ϱ. Hence, the proof is complete if we can find μ depending only on ϱ such that $B_T \subset B(\boldsymbol{x}_K, r_K)$.
If $\boldsymbol{z} \in B_T$ we have $\boldsymbol{z} = (1-\mu)\boldsymbol{x}_K + \mu\boldsymbol{y}_T + \mu h$ with $|h| < \rho_T$, and therefore

$$|\boldsymbol{z} - \boldsymbol{x}_K| \le \mu|\boldsymbol{y}_T - \boldsymbol{x}_K| + \mu\rho_T < \mu(h_T + \rho_T) \le C_{11}\mu r_K$$

with C_{11} depending only on ϱ (we used (8.36)). Taking $\mu = 1/C_{11}$ ensures that $B_T \subset B(\boldsymbol{x}_K, r_K)$ and concludes the proof. ∎

9

Non-conforming finite element methods

As briefly seen in Chapter 1 and along the examples of Chapter 7, the non-conforming $\mathbb{P}_1$ finite element method on triangular meshes can be recast in the GDM framework. In the present chapter, all standard non-conforming methods are shown to be GDMs. These methods are mesh-based, and their analysis is facilitated by a simple generalisation of the notion of control by a polytopal toolbox from Chapter 7.
A generic presentation of non-conforming methods is first developed for homogeneous Dirichlet boundary conditions and embedded in the GDM framework. The case of non-conforming $\mathbb{P}_k$ finite elements is then considered; high order estimates on $S_{\mathcal{D}}$ and $W_{\mathcal{D}}$ are obtained. Finally, the specific case $k = 1$ is detailed: the non-conforming $\mathbb{P}_1$ finite element method is shown to be an LLE method for various types of boundary conditions; mass-lumping is finally addressed.

9.1 Non-conforming finite element methods for homogeneous Dirichlet BCs

9.1.1 Abstract framework

Definition 9.1 (Non-conforming $W_0^{1,p}$ space). *Let $\mathfrak{T} = (\mathcal{M}, \mathcal{F}, \mathcal{P}, \mathcal{V})$ be a polytopal mesh of Ω in the sense of Definition 7.2. The non-conforming $W_0^{1,p}$ space on $\mathfrak{T}$, denoted by $W^{1,p}_{\mathfrak{T},0}$, is the space of all functions $w \in L^p(\Omega)$ such that:*

1. [$W^{1,p}$-regularity in each cell] *For all $K \in \mathcal{M}$, the restriction $w_{|K}$ of w to K belongs to $W^{1,p}(K)$. The trace of $w_{|K}$ on $\sigma \in \mathcal{F}_K$ is denoted by $w_{|K,\sigma}$.*
2. [Continuity of averages on internal faces] *For all $\sigma \in \mathcal{F}_{\text{int}}$ with $\mathcal{M}_\sigma = \{K, L\}$,*
$$\int_\sigma w_{|K,\sigma}(\boldsymbol{y})\mathrm{d}\gamma(\boldsymbol{y}) = \int_\sigma w_{|L,\sigma}(\boldsymbol{y})\mathrm{d}\gamma(\boldsymbol{y}). \tag{9.1}$$

3. [Homogeneous Dirichlet BC for averages on external faces] *For all* $\sigma \in \mathcal{F}_{\text{ext}}$ *with* $\mathcal{M}_\sigma = \{K\}$,

$$\int_\sigma w_{|K,\sigma}(\boldsymbol{y})\mathrm{d}\gamma(\boldsymbol{y}) = 0. \tag{9.2}$$

If $w \in W^{1,p}_{\mathfrak{T},0}$, *the "broken gradient" of* w *is* $\nabla_{\mathfrak{T}} w$ *defined by*

$$\forall K \in \mathcal{M}\,,\ \nabla_{\mathfrak{T}} w = \nabla(w_{|K}) \text{ in } K$$

and we set

$$\|w\|_{W^{1,p}_{\mathfrak{T},0}} = \|\nabla_{\mathfrak{T}} w\|_{L^p(\Omega)^d} = \left(\sum_{K\in\mathcal{M}} \left\|\nabla(w_{|K})\right\|^p_{L^p(K)^d}\right)^{\frac{1}{p}}. \tag{9.3}$$

Owing to (9.4) in Lemma 9.3 below, $\|\cdot\|_{W^{1,p}_{\mathfrak{T},0}}$ is indeed a norm on $W^{1,p}_{\mathfrak{T},0}$.

Remark 9.2 (Non-conforming approximation). The only continuity requirement for functions in $W^{1,p}_{\mathfrak{T},0}$ is the "0-degree patch test" (9.1). All families of non-conforming finite elements spaces based on the geometric elements of $\mathfrak{T}$ are subspaces of $W^{1,p}_{\mathfrak{T},0}$, and we also observe that $W^{1,p}_0(\Omega) \subset W^{1,p}_{\mathfrak{T},0}$.
Due to this weak continuity requirement (9.1), approximations with functions belonging to $W^{1,p}_{\mathfrak{T},0}$ are "non-conforming", in the sense that functions in $W^{1,p}_0(\Omega)$ are approximated by functions not necessarily in $W^{1,p}_0(\Omega)$. In particular, these approximations in $W^{1,p}_{\mathfrak{T},0}$ do not satisfy the Stokes formula in general.

Let us now prove some properties of the space $W^{1,p}_{\mathfrak{T},0}$, similar to the notions of coercivity, limit-conformity and compactness for GDs.

Lemma 9.3 (Properties of $W^{1,p}_{\mathfrak{T},0}$). *Let* $\mathfrak{T} = (\mathcal{M}, \mathcal{F}, \mathcal{P}, \mathcal{V})$ *be a polytopal mesh of* Ω *in the sense of Definition 7.2. Let* $\varrho \geq \theta_{\mathfrak{T}} + \eta_{\mathfrak{T}}$ *(see (7.8)–(7.9)). Let* $W^{1,p}_{\mathfrak{T},0}$ *be given by Definition 9.1. Then there exists* C_1, *depending only on* Ω, ϱ, d *and* p, *such that*

$$\forall w \in W^{1,p}_{\mathfrak{T},0}\,, \quad \|w\|_{L^p(\Omega)} \leq C_1 \|w\|_{W^{1,p}_{\mathfrak{T},0}} \tag{9.4}$$

and, for all $\boldsymbol{\varphi} \in W^{1,p'}(\Omega)^d$ *and all* $w \in W^{1,p}_{\mathfrak{T},0}$, *the quantity*

$$\widetilde{W}_{\mathfrak{T}}(\boldsymbol{\varphi}, w) = \int_\Omega (\boldsymbol{\varphi}(\boldsymbol{x}) \cdot \nabla_{\mathfrak{T}} w(\boldsymbol{x}) + \mathrm{div}\boldsymbol{\varphi}(\boldsymbol{x}) w(\boldsymbol{x}))\mathrm{d}\boldsymbol{x}$$

satisfies

$$\left|\widetilde{W}_{\mathfrak{T}}(\boldsymbol{\varphi}, w)\right| \leq C_1 h_{\mathcal{M}} \left\| |\nabla\boldsymbol{\varphi}| \right\|_{L^{p'}(\Omega)} \|w\|_{W^{1,p}_{\mathfrak{T},0}}\,. \tag{9.5}$$

Moreover, for any sequence of polytopal meshes $(\mathfrak{T}_m)_{m\in\mathbb{N}}$ *such that* $(\theta_{\mathfrak{T}_m} + \eta_{\mathfrak{T}_m})_{m\in\mathbb{N}}$ *is bounded and* $h_{\mathcal{M}_m} \to 0$ *as* $m \to \infty$, *the sequence* $(W^{1,p}_{\mathfrak{T}_m,0})_{m\in\mathbb{N}}$ *is compactly embedded in* $L^p(\Omega)$ *in the sense of Definition C.4.*

Proof. Step 1: We mimick the notion of control of a GD by a polytopal toolbox (see Chapter 7). Recall that $X_{\mathfrak{T},0}$, $|\cdot|_{\mathfrak{T},p}$, $\Pi_{\mathfrak{T}}$ and $\overline{\nabla}_{\mathfrak{T}}$ are defined by (7.7).
Let $\Psi : W^{1,p}_{\mathfrak{T},0} \to X_{\mathfrak{T},0}$ be the linear mapping defined by: for all $w \in W^{1,p}_{\mathfrak{T},0}$,

$$\begin{aligned} &\forall K \in \mathcal{M}\,,\ \Psi(w)_K = \frac{1}{|K|}\int_K w(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\,, \\ &\forall \sigma \in \mathcal{F}\,,\ \forall K \in \mathcal{M}_\sigma\,,\ \Psi(w)_\sigma = \frac{1}{|\sigma|}\int_\sigma w_{|K,\sigma}(\boldsymbol{x})\mathrm{d}\gamma(\boldsymbol{x}). \end{aligned} \tag{9.6}$$

The definition of $\Psi(w)_\sigma$ is justified by (9.1), which shows that $\Psi(w)_\sigma$ does not depend on $K \in \mathcal{M}_\sigma$. Moreover, owing to (9.2), $\Psi(w)$ indeed belongs to $X_{\mathfrak{T},0}$.
Let us first establish the existence of C_2, depending only on ϱ, d and p, such that

$$\forall w \in W^{1,p}_{\mathfrak{T},0}\,,\ |\Psi(w)|_{\mathfrak{T},p} \le C_2 \left\|w\right\|_{W^{1,p}_{\mathfrak{T},0}}\,, \tag{9.7}$$

$$\forall w \in W^{1,p}_{\mathfrak{T},0}\,,\ \left\|w - \Pi_{\mathfrak{T}}\Psi(w)\right\|_{L^p(\Omega)} \le C_2 h_{\mathcal{M}} \left\|w\right\|_{W^{1,p}_{\mathfrak{T},0}}\,, \tag{9.8}$$

$$\forall w \in W^{1,p}_{\mathfrak{T},0}\,,\ \forall K \in \mathcal{M}\,,\ \int_K \left[\nabla(w_{|K})(\boldsymbol{x}) - \overline{\nabla}_{\mathfrak{T}}\Psi(w)(\boldsymbol{x})\right]\mathrm{d}\boldsymbol{x} = 0. \tag{9.9}$$

Apply first Lemma B.6 to obtain C_1 depending only on d, p and ϱ such that

$$|\Psi(w)_\sigma - \Psi(w)_K|^p \le \frac{C_1 h_K^{p-1}}{|\sigma|}\int_K |\nabla(w_{|K})(\boldsymbol{x})|^p \mathrm{d}\boldsymbol{x} \tag{9.10}$$

and

$$\left\|w - \Pi_{\mathfrak{T}}\Psi(w)\right\|_{L^p(K)} \le C_1 h_K \left\|\nabla(w_{|K})\right\|_{L^p(K)^d}. \tag{9.11}$$

Divide (9.10) by $d_{K,\sigma}^{p-1}$, multiply by $|\sigma|$, use $h_K/d_{K,\sigma} \le \theta_{\mathfrak{T}}$ and sum over $\sigma \in \mathcal{F}_K$ and $K \in \mathcal{M}$ to find

$$\begin{aligned} |\Psi(w)|^p_{\mathfrak{T},p} &\le C_1 \theta_{\mathfrak{T}}^{p-1} \sum_{K\in\mathcal{M}} \sum_{\sigma\in\mathcal{F}_K} \int_K |\nabla(w_{|K})(\boldsymbol{x})|^p \mathrm{d}\boldsymbol{x} \\ &= C_1 \theta_{\mathfrak{T}}^{p-1} \sum_{K\in\mathcal{M}} \mathrm{Card}(\mathcal{F}_K) \int_K |\nabla(w_{|K})(\boldsymbol{x})|^p \mathrm{d}\boldsymbol{x}. \end{aligned}$$

Using $\mathrm{Card}(\mathcal{F}_K) \le \theta_{\mathfrak{T}}$ leads to $|\Psi(w)|^p_{\mathfrak{T},p} \le C_1 \theta_{\mathfrak{T}}^p \left\|w\right\|^p_{W^{1,p}_{\mathfrak{T},0}}$, which proves (9.7).
Estimate (9.8) is obtained by raising (9.11) to the power p and summing over $K \in \mathcal{M}$.
Relation (9.9) follows from the Stokes formula by writing, for any $K \in \mathcal{M}$,

$$\int_K \nabla(w_{|K})(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = \sum_{\sigma\in\mathcal{F}_K} \int_\sigma w_{|K,\sigma}(\boldsymbol{x})\mathrm{d}\gamma(\boldsymbol{x})\, \boldsymbol{n}_{K,\sigma} = \sum_{\sigma\in\mathcal{F}_K} |\sigma|\Psi(w)_\sigma \boldsymbol{n}_{K,\sigma}$$

$$= |K|(\overline{\nabla}_{\mathfrak{T}}\Psi(w))_{|K} = \int_K \overline{\nabla}_{\mathfrak{T}}\Psi(w)(\boldsymbol{x})\mathrm{d}\boldsymbol{x}.$$

Step 2: We mimick the proofs of Theorem 7.11 and Corollary 7.12.
To prove (9.4), use the triangle inequality, Estimate (9.8), the Hölder's inequality (D.7), Lemma B.15 (providing the $q > p$ and C_8 below), and Estimate (9.7) to write, for any $w \in W^{1,p}_{\mathfrak{T},0}$,

$$\begin{aligned}
\|w\|_{L^p(\Omega)} &\le \|w - \Pi_{\mathfrak{T}}\Psi(w)\|_{L^p(\Omega)} + \|\Pi_{\mathfrak{T}}\Psi(w)\|_{L^p(\Omega)} \\
&\le C_2 h_{\mathcal{M}} \|w\|_{W^{1,p}_{\mathfrak{T},0}} + |\Omega|^{\frac{1}{p}-\frac{1}{q}} \|\Pi_{\mathfrak{T}}\Psi(w)\|_{L^q(\Omega)} \\
&\le C_2(\mathrm{diam}(\Omega) + C_8|\Omega|^{\frac{1}{p}-\frac{1}{q}}) \|w\|_{W^{1,p}_{\mathfrak{T},0}}.
\end{aligned}$$

We turn to (9.5). Let $\boldsymbol{\varphi} \in W^{1,p'}(\Omega)^d$ and $w \in W^{1,p}_{\mathfrak{T},0}$. Use the triangle inequality and Estimates (9.8), (B.31) (noticing that $\mathbb{T}_{\mathfrak{T}}\Psi(w) = 0$ here) and (9.7) to obtain

$$\begin{aligned}
&\left|\widetilde{W}_{\mathfrak{T}}(\boldsymbol{\varphi}, w)\right| \\
&\le \left|\int_\Omega [\nabla_{\mathfrak{T}} w(\boldsymbol{x}) - \overline{\nabla}_{\mathfrak{T}}\Psi(w)(\boldsymbol{x})] \cdot \boldsymbol{\varphi}(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\right| \\
&\quad + \left|\int_\Omega [w(\boldsymbol{x}) - \Pi_{\mathfrak{T}}\Psi(w)(\boldsymbol{x})]\mathrm{div}\boldsymbol{\varphi}(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\right| \\
&\quad + \left|\int_\Omega \left(\overline{\nabla}_{\mathfrak{T}}\Psi(w)(\boldsymbol{x}) \cdot \boldsymbol{\varphi}(\boldsymbol{x}) + \Pi_{\mathfrak{T}}\Psi(w)(\boldsymbol{x})\mathrm{div}\boldsymbol{\varphi}(\boldsymbol{x})\right)\mathrm{d}\boldsymbol{x}\right| \\
&\le \left|\int_\Omega [\nabla_{\mathfrak{T}} w(\boldsymbol{x}) - \overline{\nabla}_{\mathfrak{T}}\Psi(w)(\boldsymbol{x})] \cdot \boldsymbol{\varphi}(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\right| + \|\mathrm{div}\boldsymbol{\varphi}\|_{L^{p'}(\Omega)} C_2 h_{\mathcal{M}} \|w\|_{W^{1,p}_{\mathfrak{T},0}} \\
&\quad + C_5 \|\nabla\boldsymbol{\varphi}\|_{L^{p'}(\Omega)^d} |\Psi(w)|_{\mathfrak{T},p} h_{\mathcal{M}} \\
&\le \left|\int_\Omega [\nabla_{\mathfrak{T}} w(\boldsymbol{x}) - \overline{\nabla}_{\mathfrak{T}}\Psi(w)(\boldsymbol{x})] \cdot \boldsymbol{\varphi}(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\right| + \|\mathrm{div}\boldsymbol{\varphi}\|_{L^{p'}(\Omega)} C_2 h_{\mathcal{M}} \|w\|_{W^{1,p}_{\mathfrak{T},0}} \\
&\quad + C_5 \|\nabla\boldsymbol{\varphi}\|_{L^{p'}(\Omega)^d} C_2 \|w\|_{W^{1,p}_{\mathfrak{T},0}} h_{\mathcal{M}}.
\end{aligned} \tag{9.12}$$

Let $\boldsymbol{\varphi}_K = \frac{1}{|K|}\int_K \boldsymbol{\varphi}(\boldsymbol{x})\mathrm{d}\boldsymbol{x}$ and write

$$\begin{aligned}
&\int_\Omega [\nabla_{\mathfrak{T}} w(\boldsymbol{x}) - \overline{\nabla}_{\mathfrak{T}}\Psi(w)(\boldsymbol{x})] \cdot \boldsymbol{\varphi}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \\
&\qquad = \sum_{K\in\mathcal{M}} \int_K [\nabla(w_{|K})(\boldsymbol{x}) - \overline{\nabla}_{\mathfrak{T}}\Psi(w)(\boldsymbol{x})] \cdot \boldsymbol{\varphi}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \\
&\qquad = \sum_{K\in\mathcal{M}} \Big(\int_K \left[\nabla(w_{|K})(\boldsymbol{x}) - (\overline{\nabla}_{\mathfrak{T}}\Psi(w))_{|K}\right] \cdot [\boldsymbol{\varphi}(\boldsymbol{x}) - \boldsymbol{\varphi}_K]\mathrm{d}\boldsymbol{x} \\
&\qquad\qquad + \boldsymbol{\varphi}_K \cdot \int_K [\nabla(w_{|K})(\boldsymbol{x}) - \overline{\nabla}_{\mathfrak{T}}\Psi(w)(\boldsymbol{x})]\mathrm{d}\boldsymbol{x}\Big).
\end{aligned}$$

Use (9.9) and $\int_K (\boldsymbol{\varphi}(\boldsymbol{x}) - \boldsymbol{\varphi}_K)\mathrm{d}\boldsymbol{x} = 0$ to infer

$$\int_\Omega [\nabla_\mathfrak{T} w(\boldsymbol{x}) - \overline{\nabla}_\mathfrak{T} \Psi(w)(\boldsymbol{x})] \cdot \boldsymbol{\varphi}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = \sum_{K\in\mathcal{M}} \int_K \nabla(w_{|K})(\boldsymbol{x}) \cdot [\boldsymbol{\varphi}(\boldsymbol{x}) - \boldsymbol{\varphi}_K]\mathrm{d}\boldsymbol{x}.$$

Applying (B.12) in Lemma B.6 to p' instead of p, we find C_3 depending only on d, p and ϱ such that $\|\boldsymbol{\varphi} - \boldsymbol{\varphi}_K\|_{L^{p'}(K)} \le C_3 h_K \, \| \, |\nabla\boldsymbol{\varphi}| \, \|_{L^{p'}(K)}$. Hence, by the Hölder inequalities (D.5) (in each cell) and (D.1) (for the sum over the cells),

$$\left| \int_\Omega [\nabla_\mathfrak{T} w(\boldsymbol{x}) - \overline{\nabla}_\mathfrak{T} \Psi(w)(\boldsymbol{x})] \cdot \boldsymbol{\varphi}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \right| \le C_3 h_\mathcal{M} \, \|w\|_{W^{1,p}_{\mathfrak{T},0}} \, \| \, |\nabla\boldsymbol{\varphi}| \, \|_{L^{p'}(\Omega)} \, .$$

Plugged into (9.12), this gives (9.5).

It remains to prove the compact embedding. If $w_m \in W^{1,p}_{\mathfrak{T}_m,0}$ is such that $(\|w_m\|_{W^{1,p}_{\mathfrak{T}_m,0}})_{m\in\mathbb{N}}$ is bounded, then (9.7) ensures that $(|\Psi_m(w_m)|_{\mathfrak{T}_m,p})_{m\in\mathbb{N}}$ is bounded. By Lemma B.19, $(\Pi_{\mathfrak{T}_m}\Psi_m(w_m))_{m\in\mathbb{N}}$ converges up to a subsequence as $m \to \infty$ to some w in $L^p(\Omega)$. Estimate (9.8) shows that

$$\|w_m - \Pi_{\mathfrak{T}_m}\Psi_m(w_m)\|_{L^p(\Omega)} \le C_2 h_{\mathcal{M}_m} \, \|w_m\|_{W^{1,p}_{\mathfrak{T}_m,0}} \to 0 \text{ as } m \to \infty.$$

Hence, $w_m \to w$ in $L^p(\Omega)$ and the proof is complete. ∎

Let us describe, in the case $p = 2$, the approximation by a generic non-conforming method of the linear elliptic problem (2.22) under Hypotheses (2.21). To this purpose, a bilinear form $a_\mathfrak{T}$ and a linear form $b_\mathfrak{T}$ are defined by

$$\begin{aligned} \forall w, \widetilde{w} \in W^{1,p}_{\mathfrak{T},0}\,,\ a_\mathfrak{T}(w,\widetilde{w}) &= \int_\Omega \Lambda(\boldsymbol{x})\nabla_\mathfrak{T} w(\boldsymbol{x}) \cdot \nabla_\mathfrak{T}\widetilde{w}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \\ \text{and } b_\mathfrak{T}(\widetilde{w}) &= \int_\Omega (f(\boldsymbol{x})\widetilde{w}(\boldsymbol{x}) - \boldsymbol{F}(\boldsymbol{x}) \cdot \nabla_\mathfrak{T}\widetilde{w}(\boldsymbol{x}))\mathrm{d}\boldsymbol{x}. \end{aligned} \tag{9.13}$$

Let $V \subset W^{1,p}_{\mathfrak{T},0}$ be a finite dimensional space. The non-conforming finite element scheme based on V, for the approximation of (2.22), reads

$$\text{find } u \in V \text{ such that, for all } w \in V,\ a_\mathfrak{T}(u,w) = b_\mathfrak{T}(w). \tag{9.14}$$

Owing to (9.4), $a_\mathfrak{T}(\cdot,\cdot)$ is an inner product in V. The Riesz representation theorem thus shows that Problem (9.14) has a unique solution. It can be proved, see for example [49, Theorem 4.2.2], that the following error estimate is satisfied: let $\overline{u} \in H^1_0(\Omega)$ be the solution to (2.22), then there exists $C > 0$ depending on the regularity of $\mathcal{M}$ (factor $\theta_\mathfrak{T} + \eta_\mathfrak{T}$ as defined by (7.8)–(7.9)) but not on $h_\mathcal{M}$, such that

$$\|\overline{u} - u\|_{W^{1,p}_{\mathfrak{T},0}} \le C \left(\inf_{v\in V} \|\overline{u} - v\|_{W^{1,p}_{\mathfrak{T},0}} + \sup_{w\in V\setminus\{0\}} \frac{|a_\mathfrak{T}(\overline{u},w) - b_\mathfrak{T}(w)|}{\|w\|_{W^{1,p}_{\mathfrak{T},0}}} \right). \tag{9.15}$$

This estimate is based on the second Strang Lemma [132]. As shown below, it can also be recovered from the generic results obtained in the framework of the gradient discretisation method, once we prove that non-conforming finite element methods are GDMs.

9.1.2 GDM formulation of abstract non-conforming finite element methods

Definition 9.4 (Non-conforming finite element gradient discretisation). *Let $\mathfrak{T} = (\mathcal{M}, \mathcal{F}, \mathcal{P}, \mathcal{V})$ be a polytopal mesh of Ω in the sense of Definition 7.2. A non-conforming finite element gradient discretisation (NCFE GD, for short) on $\mathfrak{T}$ is $\mathcal{D} = (X_{\mathcal{D},0}, \Pi_{\mathcal{D}}, \nabla_{\mathcal{D}})$ defined the following way.*

1. *Let V be a finite dimensional subspace of $W^{1,p}_{\mathfrak{T},0}$ (see Definition 9.1). Fixing $(\chi_i)_{i\in I}$ a basis of V, we set*
$$X_{\mathcal{D},0} = \{v = (v_i)_{i\in I} \,:\, v_i \in \mathbb{R} \text{ for all } i \in I\}. \tag{9.16}$$
2. *The operator $\Pi_{\mathcal{D}}$ is the reconstruction in $W^{1,p}_{\mathfrak{T},0}$ from the coordinates described by elements of $X_{\mathcal{D},0}$:*
$$\forall v \in X_{\mathcal{D},0}\,,\ \Pi_{\mathcal{D}} v = \sum_{i\in I} v_i \chi_i. \tag{9.17}$$
For all $K \in \mathcal{M}$ and $v \in X_{\mathcal{D},0}$, we set $\Pi_K v = (\Pi_{\mathcal{D}} v)_{|K} \in W^{1,p}(K)$.
3. *The gradient reconstruction $\nabla_{\mathcal{D}}$ is defined from the broken gradient $\nabla_{\mathfrak{T}}$ by setting, for $v \in X_{\mathcal{D},0}$, $\nabla_{\mathcal{D}} v := \nabla_{\mathfrak{T}}(\Pi_{\mathcal{D}} v)$. In other words,*
$$\forall K \in \mathcal{M}\,,\ \text{for a.e. } \boldsymbol{x} \in K\,,\ \nabla_{\mathcal{D}} v(\boldsymbol{x}) = \nabla(\Pi_K v)(\boldsymbol{x}). \tag{9.18}$$

Observe that $\Pi_{\mathcal{D}} : X_{\mathcal{D},0} \to V$ is an isomorphism and that, by (9.3),
$$\|\nabla_{\mathcal{D}} \cdot\|_{L^p(\Omega)^d} = \|\Pi_{\mathcal{D}} \cdot\|_{W^{1,p}_{\mathfrak{T},0}}\,. \tag{9.19}$$
Hence, $\|\nabla_{\mathcal{D}} \cdot\|_{L^p(\Omega)^d}$ is a norm on $X_{\mathcal{D},0}$.

Proposition 9.5 (Estimates on $C_{\mathcal{D}}$, $S_{\mathcal{D}}$ and $W_{\mathcal{D}}$ for NCFE GDs). *Let $\mathcal{D}$ be a NCFE GD in the sense of Definition 9.4, based on an underlying polytopal mesh $\mathfrak{T}$ and a finite dimensional space $V \subset W^{1,p}_{\mathfrak{T},0}$. Let $\varrho \geq \theta_{\mathfrak{T}} + \eta_{\mathfrak{T}}$ (see (7.8)–(7.9)). Then, with C_1 defined in Lemma 9.3 (and depending only on Ω, ϱ, d and p),*
$$C_{\mathcal{D}} \leq C_1, \tag{9.20}$$
$$\begin{aligned} &\forall \varphi \in W^{1,p}_0(\Omega)\,, \\ &\inf_{w\in V} \|\varphi - w\|_{W^{1,p}_{\mathfrak{T},0}} \leq S_{\mathcal{D}}(\varphi) \leq (1 + C_1) \inf_{w\in V} \|\varphi - w\|_{W^{1,p}_{\mathfrak{T},0}}\,, \end{aligned} \tag{9.21}$$
and
$$\forall \boldsymbol{\varphi} \in W^{1,p'}(\Omega)^d,\ W_{\mathcal{D}}(\boldsymbol{\varphi}) \leq C_1 h_{\mathcal{M}} \|\boldsymbol{\varphi}\|_{W^{1,p'}(\Omega)^d}\,. \tag{9.22}$$
Here, $C_{\mathcal{D}}$ is defined by (2.1), $S_{\mathcal{D}}$ is defined by (2.2) and $W_{\mathcal{D}}$ is defined by (2.6).

Proof. Inequality (9.20) is a consequence of (9.19) and of the Poincaré inequality (9.4) applied to $w = \Pi_{\mathcal{D}} v$ for a generic $v \in X_{\mathcal{D},0}$.
Since $\Pi_{\mathcal{D}} : X_{\mathcal{D},0} \to V$ is onto, the second inequality in (9.21) follows from the same Poincaré inequality (9.4) applied to $w = \Pi_{\mathcal{D}} v - \varphi$, for $\varphi \in W_0^{1,p}(\Omega)$ and a generic $v \in X_{\mathcal{D},0}$. The first inequality in (9.21) comes from the definition of $S_{\mathcal{D}}$.
Estimate (9.22) is established by applying (9.5) to $w = \Pi_{\mathcal{D}} v$ for a generic $v \in X_{\mathcal{D},0}$, and by using (9.19). ∎

In the case $p = 2$, using an NCFE GD $\mathcal{D} = (X_{\mathcal{D},0}, \Pi_{\mathcal{D}}, \nabla_{\mathcal{D}})$ as above, the gradient scheme (2.23) for the discretisation of (2.22) is identical to the non-conforming finite element scheme (9.14). Moreover, since $\Pi_{\mathcal{D}}$ is an isomorphism, (9.19) yields

$$W_{\mathcal{D}}(\Lambda \nabla \overline{u} + \boldsymbol{F}) = \sup_{w \in V \setminus \{0\}} \frac{|a_{\mathfrak{T}}(\overline{u}, w) - b_{\mathfrak{T}}(w)|}{\|w\|_{W^{1,p}_{\mathfrak{T},0}}}.$$

Alongside (9.21), this proves that the error estimate (9.15) is a consequence of (2.25) in Theorem 2.28.

The properties of NCFE GDs are straightforward consequences of the previous results.

Theorem 9.6 (Properties of NCFE GDs for homogeneous Dirichlet BCs). *Let $(\mathcal{D}_m)_{m\in\mathbb{N}}$ be a sequence of NCFE GDs in the sense of Definition 9.4, based on underlying polytopal meshes $(\mathfrak{T}_m)_{m\in\mathbb{N}}$ and finite dimensional spaces $(V_m)_{m\in\mathbb{N}}$. Assume that the sequence $(\theta_{\mathfrak{T}_m} + \eta_{\mathfrak{T}_m})_{m\in\mathbb{N}}$ is bounded (see (7.8)–(7.9)), that $h_{\mathcal{M}_m} \to 0$ as $m \to \infty$, and that*

$$\forall \varphi \in W_0^{1,p}(\Omega)\,, \ \lim_{m\to\infty} \inf_{w \in V_m} \|\varphi - w\|_{W^{1,p}_{\mathfrak{T}_m,0}} = 0. \tag{9.23}$$

Then the sequence $(\mathcal{D}_m)_{m\in\mathbb{N}}$ is coercive, GD-consistent, limit-conforming and compact in the sense of the definitions of Section 2.1.1 in Chapter 2.

Proof. Proposition 9.5 yields the coercivity, limit-conformity, and GD-consistency owing to Hypothesis (9.23). The compactness of $(\mathcal{D}_m)_{m\in\mathbb{N}}$ is a immediate consequence of compact embedding of $(W^{1,p}_{\mathfrak{T}_m,0})_{m\in\mathbb{N}}$ in $L^p(\Omega)$ (see Lemma 9.3) and of (9.19). ∎

9.2 $\mathbb{P}_k$ NCFE GDs for homogeneous Dirichlet BCs

We now turn to a specific example of NCFE GD, corresponding to a simplicial mesh and functions in V that are polynomial of degree k or less on each simplex. This describes the non-conforming $\mathbb{P}_k$ finite element method.

Definition 9.7 ($\mathbb{P}_k$ NCFE gradient discretisation). *Let $\mathfrak{T} = (\mathcal{M}, \mathcal{F}, \mathcal{P}, \mathcal{V})$ be a conforming simplicial mesh of Ω in the sense of Definition 7.4, and let $k \in \mathbb{N} \setminus \{0\}$. A $\mathbb{P}_k$ NCFE gradient discretisation on $\mathfrak{T}$ is given by Definition 9.4 with the finite dimensional space $V = V^{(k)}$ defined as the set functions $w \in L^p(\Omega)$ such that:*

1. *for all $K \in \mathcal{M}$, the restriction $w_{|K}$ of w to K belongs to $\mathbb{P}_k(K)$,*
2. *for all $\sigma \in \mathcal{F}$, there exists a polynomial $R_\sigma w \in \mathbb{P}_{k-1}(\sigma)$ such that*

$$\begin{aligned} &\forall K \in \mathcal{M}_\sigma\,,\ \forall \varphi \in \mathbb{P}_{k-1}(\sigma), \\ &\int_\sigma \varphi(\boldsymbol{y})(w_{|K}(\boldsymbol{y}) - R_\sigma w(\boldsymbol{y}))\mathrm{d}\gamma(\boldsymbol{y}) = 0, \end{aligned} \tag{9.24}$$

3. *for all $\sigma \in \mathcal{F}_{\mathrm{ext}}$, $R_\sigma w = 0$.*

For any $\sigma \in \mathcal{F}$, the constant function $\varphi = 1$ on σ is an element of $\mathbb{P}_{k-1}(\sigma)$. Hence, if $w \in V$ and $\sigma \in \mathcal{F}_{\mathrm{int}}$ with $\mathcal{M}_\sigma = \{K, L\}$,

$$\int_\sigma w_{|K}(\boldsymbol{y})\mathrm{d}\gamma(\boldsymbol{y}) = \int_\sigma R_\sigma w(\boldsymbol{y})\mathrm{d}\gamma(\boldsymbol{y}) = \int_\sigma w_{|L}(\boldsymbol{y})\mathrm{d}\gamma(\boldsymbol{y}).$$

In other words, (9.1) is satisfied. Similarly, if $\sigma \in \mathcal{F}_{\mathrm{ext}}$ and $\mathcal{M}_\sigma = \{K\}$, (9.2) follows by writing

$$\int_\sigma w_{|K}(\boldsymbol{y})\mathrm{d}\gamma(\boldsymbol{y}) = \int_\sigma R_\sigma w(\boldsymbol{y})\mathrm{d}\gamma(\boldsymbol{y}) = 0.$$

Remark 9.8. Relation (9.24) is equivalent to saying that the $L^2(\sigma)$-orthogonal projections on $\mathbb{P}_{k-1}(\sigma)$ of (the restrictions to σ of) $w_{|K}$ and $w_{|L}$ are identical.

Theorem 9.6 applies to $\mathbb{P}_k$ NCFE GDs if we can establish (9.23). Actually, as expected for a method based on local polynomials of arbitrary degree, specific estimates can be established for both $S_\mathcal{D}$ and $W_\mathcal{D}$. Using these estimates, Theorem 2.28 gives optimal orders of convergence for non-conforming $\mathbb{P}_k$ finite elements when applied to linear elliptic problems.

Lemma 9.9 (Estimates on $S_\mathcal{D}$ and $W_\mathcal{D}$ for $\mathbb{P}_k$ NCFE GDs). *Let $\mathcal{D}$ be a $\mathbb{P}_k$ NCFE GD in the sense of Definition 9.7. Take $\varrho \geq \kappa_\mathfrak{T}$ (see (7.10)). Then there exists C_4, depending only on Ω, p, k, ϱ and d, such that*

$$\forall \varphi \in W^{k+1,\infty}(\Omega) \cap W_0^{1,p}(\Omega)\,,\ S_\mathcal{D}(\varphi) \leq C_4 h_\mathcal{M}^k \left\|\varphi\right\|_{W^{k+1,\infty}(\Omega)} \tag{9.25}$$

and

$$\forall \boldsymbol{\varphi} \in W^{k,\infty}(\Omega)^d\,,\ W_\mathcal{D}(\boldsymbol{\varphi}) \leq C_4 h_\mathcal{M}^k \left\|\boldsymbol{\varphi}\right\|_{W^{k,\infty}(\Omega)^d}\,, \tag{9.26}$$

where $S_\mathcal{D}$ and $W_\mathcal{D}$ are defined by (2.2) and (2.6), respectively. This means that the space size (see Definition 2.22) of the GD is such that

$$h_\mathcal{D}(W^{k+1,\infty}(\Omega) \cap W_0^{1,p}(\Omega); W^{k,\infty}(\Omega)^d) \leq C_4 h_\mathcal{M}^k.$$

Proof. Let $\varphi \in W^{k+1,\infty}(\Omega) \cap W_0^{1,p}(\Omega)$. By (8.15) in Proposition 8.11, there is a function w (equal to some $\Pi_{\mathcal{D}_c} v_c$, where $\mathcal{D}_c$ is the conforming $\mathbb{P}_k$ GD on $\mathfrak{T}$ and $v_c \in X_{\mathcal{D}_c,0}$) such that $w \in W_0^{1,p}(\Omega)$, $w_{|K} \in \mathbb{P}_k(K)$ for all $K \in \mathcal{M}$, w is continuous on $\overline{\Omega}$, and

$$\|w - \varphi\|_{L^p(\Omega)} + \|\nabla w - \nabla \varphi\|_{L^p(\Omega)^d} \le C_5 h_{\mathcal{M}}^k \|\varphi\|_{W^{k+1,\infty}(\Omega)}, \tag{9.27}$$

where C_5 is given by Proposition 8.11 and depends only on Ω, k and ϱ. The continuity of w ensures that it belongs to the space $V^{(k)}$ given in Definition 9.7 (since $w_{|K} = w_{|L}$ on σ whenever $K, L \in \mathcal{M}_\sigma$), and (9.25) is then a consequence of (9.27) and (9.21).

We now turn to the proof of (9.26), starting with the case $\boldsymbol{\varphi} \in C^k(\overline{\Omega})^d$. Let $u \in X_{\mathcal{D},0}$. Since $\Pi_{\mathcal{D}} u \in V^{(k)}$, the definition (9.18) of $\nabla_{\mathcal{D}} u$ and an integration-by-parts in each cell yields

$$\begin{aligned}
&\int_\Omega (\nabla_{\mathcal{D}} u(\boldsymbol{x}) \cdot \boldsymbol{\varphi}(\boldsymbol{x}) + \Pi_{\mathcal{D}} u(\boldsymbol{x}) \mathrm{div} \boldsymbol{\varphi}(\boldsymbol{x}))\, \mathrm{d}\boldsymbol{x} \\
&= \sum_{K \in \mathcal{M}} \int_K (\nabla(\Pi_K u)(\boldsymbol{x}) \cdot \boldsymbol{\varphi}(\boldsymbol{x}) + \Pi_K u(\boldsymbol{x}) \mathrm{div} \boldsymbol{\varphi}(\boldsymbol{x}))\, \mathrm{d}\boldsymbol{x} \\
&= \sum_{K \in \mathcal{M}} \sum_{\sigma \in \mathcal{F}_K} \int_\sigma \Pi_K u(\boldsymbol{y}) \boldsymbol{\varphi}(\boldsymbol{y}) \cdot \boldsymbol{n}_{K,\sigma} \mathrm{d}\gamma(\boldsymbol{y}).
\end{aligned} \tag{9.28}$$

Set $\Pi_\sigma u = R_\sigma(\Pi_{\mathcal{D}} u)$ and, gathering the sums by faces, notice that

$$\begin{aligned}
&\sum_{K \in \mathcal{M}} \sum_{\sigma \in \mathcal{F}_K} \int_\sigma \Pi_\sigma u(\boldsymbol{y}) \boldsymbol{\varphi}(\boldsymbol{y}) \cdot \boldsymbol{n}_{K,\sigma} \mathrm{d}\gamma(\boldsymbol{y}) \\
&\qquad = \sum_{\sigma \in \mathcal{F}_{\mathrm{int}},\, \mathcal{M}_\sigma = \{K,L\}} \int_\sigma \Pi_\sigma u(\boldsymbol{y}) \boldsymbol{\varphi}(\boldsymbol{y}) \cdot (\boldsymbol{n}_{K,\sigma} + \boldsymbol{n}_{L,\sigma}) \mathrm{d}\gamma(\boldsymbol{y}) \\
&\qquad\quad + \sum_{\sigma \in \mathcal{F}_{\mathrm{ext}},\, \mathcal{M}_\sigma = \{K\}} \int_\sigma \Pi_\sigma u(\boldsymbol{y}) \boldsymbol{\varphi}(\boldsymbol{y}) \cdot \boldsymbol{n}_{K,\sigma} \mathrm{d}\gamma(\boldsymbol{y}) = 0,
\end{aligned}$$

since $\boldsymbol{n}_{K,\sigma} + \boldsymbol{n}_{L,\sigma} = 0$ if $\mathcal{M}_\sigma = \{K, L\}$ and $\Pi_\sigma u = 0$ if $\sigma \in \mathcal{F}_{\mathrm{ext}}$. Subtracting this quantity from (9.28) yields

$$\begin{aligned}
&\int_\Omega (\nabla_{\mathcal{D}} u(\boldsymbol{x}) \cdot \boldsymbol{\varphi}(\boldsymbol{x}) + \Pi_{\mathcal{D}} u(\boldsymbol{x}) \mathrm{div} \boldsymbol{\varphi}(\boldsymbol{x}))\, \mathrm{d}\boldsymbol{x} \\
&\qquad = \sum_{K \in \mathcal{M}} \sum_{\sigma \in \mathcal{F}_K} \int_\sigma (\Pi_K u(\boldsymbol{y}) - \Pi_\sigma u(\boldsymbol{y})) \boldsymbol{\varphi}(\boldsymbol{y}) \cdot \boldsymbol{n}_{K,\sigma} \mathrm{d}\gamma(\boldsymbol{y}).
\end{aligned}$$

Let $\varphi_{\sigma,k-1}$ be the Taylor expansion of degree $k-1$ of $\boldsymbol{\varphi} \cdot \boldsymbol{n}_{K,\sigma}$ around the point $\overline{\boldsymbol{x}}_\sigma$. Relation (9.24) shows that

$$\int_\Omega (\nabla_{\mathcal{D}} u(\boldsymbol{x}) \cdot \boldsymbol{\varphi}(\boldsymbol{x}) + \Pi_{\mathcal{D}} u(\boldsymbol{x}) \mathrm{div} \boldsymbol{\varphi}(\boldsymbol{x}))\, \mathrm{d}\boldsymbol{x}$$
$$= \sum_{K\in\mathcal{M}} \sum_{\sigma\in\mathcal{F}_K} \int_\sigma (\Pi_K u(\boldsymbol{y}) - \Pi_\sigma u(\boldsymbol{y}))(\boldsymbol{\varphi}(\boldsymbol{y}) \cdot \boldsymbol{n}_{K,\sigma} - \varphi_{\sigma,k-1}(\boldsymbol{y}))\mathrm{d}\gamma(\boldsymbol{y}). \tag{9.29}$$

Let $\sigma \in \mathcal{F}_K$. Since $\Pi_\sigma u$ is the $L^2(\sigma)$-projection of $\Pi_K u$ on $\mathbb{P}_{k-1}(\sigma)$, the projection estimate in [57, Lemma 3.4] applied to $U = \sigma$ gives C_5 depending only on ϱ, p, k and d such that

$$\begin{aligned}\|\Pi_K u - \Pi_\sigma u\|_{L^p(\sigma)} &\le C_5 \mathrm{diam}(\sigma) \|\nabla(\Pi_K u)\|_{L^p(\sigma)^d} \\ &\le C_5 h_K \|\nabla(\Pi_K u)\|_{L^p(\sigma)^d}.\end{aligned}$$

The function $\nabla(\Pi_K u)$ belongs to $\mathbb{P}_{k-1}(K)^d$. Hence, by the discrete trace inequality [61, Lemma 1.52], $\|\nabla(\Pi_K u)\|_{L^p(\sigma)} \le C_6 h_K^{-1/p} \|\nabla(\Pi_K u)\|_{L^p(K)^d}$ for some C_6 depending only on ϱ, p, k and d. We infer that

$$\|\Pi_K u - \Pi_\sigma u\|_{L^p(\sigma)} \le C_5 C_6 h_K^{1-\frac{1}{p}} \|\nabla(\Pi_K u)\|_{L^p(K)^d}. \tag{9.30}$$

Moreover, the smoothness of $\boldsymbol{\varphi}$ provides C_7 depending only on k and d such that

$$\|\boldsymbol{\varphi} \cdot \boldsymbol{n}_{K,\sigma} - \varphi_{\sigma,k-1}\|_{L^{p'}(\sigma)} \le C_7 |\sigma|^{\frac{1}{p'}} h_{\mathcal{M}}^k \|\boldsymbol{\varphi}\|_{W^{k,\infty}(\Omega)^d}.$$

Combined with (9.30) and a Hölder inequality (D.5) on the right-hand side of (9.29), this shows that

$$\begin{aligned}&\left|\int_\Omega (\nabla_{\mathcal{D}} u(\boldsymbol{x}) \cdot \boldsymbol{\varphi}(\boldsymbol{x}) + \Pi_{\mathcal{D}} u(\boldsymbol{x}) \mathrm{div} \boldsymbol{\varphi}(\boldsymbol{x}))\, \mathrm{d}\boldsymbol{x}\right| \\ &\qquad \le C_8 \|\boldsymbol{\varphi}\|_{W^{k,\infty}(\Omega)^d} h_{\mathcal{M}}^k \sum_{K\in\mathcal{M}} \sum_{\sigma\in\mathcal{F}_K} |\sigma|^{\frac{1}{p'}} h_K^{\frac{1}{p'}} \|\nabla(\Pi_K u)\|_{L^p(K)^d}\end{aligned} \tag{9.31}$$

for some C_8 depending only on ϱ, p, k and d. Using $\varrho+d+1 \ge \kappa_{\mathfrak{T}}+d+1 \ge \theta_{\mathfrak{T}}$ (see Lemma B.4) and the second relation in (B.1), we have

$$\sum_{\sigma\in\mathcal{F}_K} |\sigma| h_K \le (\varrho + d + 1) \sum_{\sigma\in\mathcal{F}_K} |\sigma| d_{K,\sigma} = (\varrho + d + 1) d |K|.$$

A discrete Hölder inequality (D.1) in (9.31) therefore provides, since $\mathrm{Card}(\mathcal{F}_K) \le \theta_{\mathfrak{T}} \le \varrho$,

$$\begin{aligned}&\left|\int_\Omega (\nabla_{\mathcal{D}} u(\boldsymbol{x}) \cdot \boldsymbol{\varphi}(\boldsymbol{x}) + \Pi_{\mathcal{D}} u(\boldsymbol{x}) \mathrm{div} \boldsymbol{\varphi}(\boldsymbol{x}))\, \mathrm{d}\boldsymbol{x}\right| \\ &\qquad \le C_8 \|\boldsymbol{\varphi}\|_{W^{k,\infty}(\Omega)^d} h_{\mathcal{M}}^k ((\varrho + d + 1) d |\Omega|)^{\frac{1}{p'}} \varrho^{\frac{1}{p}} \|\nabla_{\mathcal{D}} u\|_{L^p(\Omega)^d}.\end{aligned}$$

This concludes the proof of (9.26) if $\boldsymbol{\varphi} \in C^k(\overline{\Omega})^d$. The general case $\boldsymbol{\varphi} \in W^{k,\infty}(\Omega)^d$ is obtained by an extension and regularisation argument, as in the beginning of the proof of Proposition 7.38. ∎

The properties of $\mathbb{P}_k$ NCFE gradient discretisations follow immediately from Lemma 9.9 and Theorem 9.6.

Theorem 9.10 (Properties of $\mathbb{P}_k$ NCFE GDs for homogeneous Dirichlet BCs). *Let $(\mathcal{D}_m)_{m\in\mathbb{N}}$ be a sequence of $\mathbb{P}_k$ NCFE GDs in the sense of Definition 9.7, defined from underlying conforming simplicial meshes $(\mathfrak{T}_m)_{m\in\mathbb{N}}$. Assume that $(\kappa_{\mathfrak{T}_m})_{m\in\mathbb{N}}$ is bounded (see (7.10)), and that $h_{\mathcal{M}_m} \to 0$ as $m \to \infty$. Then the sequence $(\mathcal{D}_m)_{m\in\mathbb{N}}$ is coercive, GD-consistent, limit-conforming and compact in the sense of the definitions of Section 2.1.1 in Chapter 2.*

9.3 Non-conforming $\mathbb{P}_1$ finite element method for homogeneous Dirichlet boundary conditions

Among the non-conforming finite element methods, the $\mathbb{P}_1$ non-conforming finite element method has a particular importance. This scheme is often called the Crouzeix–Raviart scheme, although this name historically pertains to the usage of the method for the Stokes problem (see [55] for the seminal paper and, for instance, [84, pp.25–26 and 199–201] for a synthetic presentation). This finite element method approximates solutions to PDEs with functions that are piecewise linear on a conforming simplicial mesh of Ω, and continuous at the centres of mass of the faces.
The $\mathbb{P}_1$ NCFE method is nothing but a special case of the $\mathbb{P}_k$ NCFE method and, as such, its inclusion in the GDM and the properties of the corresponding GDs follow from the analysis done in the previous sections. However, it is possible to present the $\mathbb{P}_1$ NCFE GD as an LLE GD, which makes it amenable to the techniques developed for such GDs – in particular, the mass-lumping process. Here, we give this LLE GD presentation of the $\mathbb{P}_1$ NCFE method, and we discuss some consequences.

9.3.1 Definition of the non-conforming $\mathbb{P}_1$ gradient discretisation

Let $\mathfrak{T} = (\mathcal{M}, \mathcal{F}, \mathcal{P}, \mathcal{V})$ be a conforming simplicial mesh of Ω in the sense of Definition 7.4. The non-conforming $\mathbb{P}_1$ gradient discretisation is constructed as an LLE GD, by specifying the objects introduced in Definition 7.33.

1. The set of geometrical entities attached to the discrete unknowns is $I = \mathcal{F}$ and the approximation points are $S = (\overline{\boldsymbol{x}}_\sigma)_{\sigma\in\mathcal{F}}$. Then $I_\Omega = \mathcal{F}_{\rm int}$, $I_\partial = \mathcal{F}_{\rm ext}$, and

$$X_{\mathcal{D},0} = \{v = (v_\sigma)_{\sigma\in\mathcal{F}} \,:\, v_\sigma \in \mathbb{R} \text{ for all } \sigma \in \mathcal{F}_{\rm int}\,, \quad v_\sigma = 0 \text{ for all } \sigma \in \mathcal{F}_{\rm ext}\}. \tag{9.32}$$

   For all $K \in \mathcal{M}$, we let $I_K = \mathcal{F}_K$.

2. The reconstruction $\Pi_{\mathcal{D}}$ in (7.32) is built from the affine non-conforming finite element basis functions $(\pi_K^\sigma)_{\sigma\in\mathcal{F}_K}$ defined, for each $K \in \mathcal{M}$, by

$$\begin{aligned}&\forall\sigma \in \mathcal{F}_K\,,\ \pi_K^\sigma \text{ is affine on } K,\ \pi_K^\sigma(\overline{\boldsymbol{x}}_\sigma) = 1,\\ &\text{and } \pi_K^\sigma(\overline{\boldsymbol{x}}_{\sigma'}) = 0 \text{ for all } \sigma' \in \mathcal{F}_K\backslash\{\sigma\}.\end{aligned} \tag{9.33}$$

This leads to

$$\begin{aligned}&\forall v \in X_{\mathcal{D},0}\,,\ \forall K \in \mathcal{M}\,,\ \text{for a.e. } \boldsymbol{x} \in K\,,\\ &\Pi_{\mathcal{D}} v(\boldsymbol{x}) = \sum_{\sigma\in\mathcal{F}_K} v_\sigma \pi_K^\sigma(\boldsymbol{x}).\end{aligned} \tag{9.34}$$

3. The functions $(\mathcal{G}_K^\sigma)_{K\in\mathcal{M}\,,\,\sigma\in\mathcal{F}_K}$ that define the gradient reconstruction $\nabla_{\mathcal{D}}$ through (7.33) are the constant functions on the cells given by

$$\mathcal{G}_K^\sigma = \nabla\pi_K^\sigma. \tag{9.35}$$

Hence, the reconstructed gradients are piecewise constant on the cells:

$$\begin{aligned}&\forall v \in X_{\mathcal{D},0}\,,\ \forall K \in \mathcal{M}\,,\\ &(\nabla_{\mathcal{D}} v)_{|K} = \sum_{\sigma\in\mathcal{F}_K} v_\sigma \nabla\pi_K^\sigma = \nabla\left[(\Pi_{\mathcal{D}} v)_{|K}\right].\end{aligned} \tag{9.36}$$

The properties of $(\pi_K^\sigma)_{\sigma\in\mathcal{F}_K}$ and $(\mathcal{G}_K^\sigma)_{\sigma\in\mathcal{F}_K}$, ensuring that the relations above indeed define an LLE GD, are given by Lemma 9.12 below.
The link between this non-conforming $\mathbb{P}_1$ LLE GD and the notions in Section 9.2 is described by the following lemma, which also shows that $\|\nabla_{\mathcal{D}}\cdot\|_{L^p(\Omega)^d}$ is a norm on $X_{\mathcal{D},0}$.

Lemma 9.11 (The non-conforming $\mathbb{P}_1$ LLE GD is the $\mathbb{P}_1$ NCFE GD). *Let $\mathfrak{T}$ be a conforming simplicial mesh and $\mathcal{D}$ be the non-conforming $\mathbb{P}_1$ LLE GD defined above. Then $\mathcal{D}$ is the $\mathbb{P}_1$ NCFE GD given by Definition 9.7 with $k = 1$.*

Proof. Let $V^{(1)}$ be the space given in Definition 9.7 for $k = 1$. For all $w \in V^{(1)}$ and $K \in \mathcal{M}$, since $w_{|K}$ is affine on any $\sigma \in \mathcal{F}_K$ we have

$$\int_\sigma w_{|K}(\boldsymbol{y})\mathrm{d}\gamma(\boldsymbol{y}) = |\sigma| w_{|K}(\overline{\boldsymbol{x}}_\sigma).$$

Comparing with (9.24), this shows that $R_\sigma w$ is the constant polynomial $w_{|K}(\overline{\boldsymbol{x}}_\sigma)$. Since this polynomial does not depend on $K \in \mathcal{M}_\sigma$, we infer that w is continuous at $\overline{\boldsymbol{x}}_\sigma$ (its value at this point does not depend on the considered cell on either side of σ).
This allows us to define $v_\sigma = w(\overline{\boldsymbol{x}}_\sigma)$. If $\sigma \in \mathcal{F}_{\text{ext}}$, $v_\sigma = w(\overline{\boldsymbol{x}}_\sigma) = R_\sigma w = 0$ by Definition of $V^{(1)}$. Hence, $v \in X_{\mathcal{D},0}$ defined by (9.32). Since both $w_{|K}$ and $(\Pi_{\mathcal{D}} v)_{|K}$ are affine and take value $v_\sigma = w(\overline{\boldsymbol{x}}_\sigma)$ at any $\sigma \in \mathcal{F}_K$ (see (9.34)),

we have $w_{|K} = (\Pi_{\mathcal{D}} v)_{|K}$ (see, e.g., the start of the proof of Lemma 9.12). In other words, $w = \Pi_{\mathcal{D}} v$ and $V^{(1)} = \Pi_{\mathcal{D}}(X_{\mathcal{D},0})$. Moreover, $\Pi_{\mathcal{D}} : X_{\mathcal{D},0} \to V^{(1)}$ is clearly an isomorphism.
For $\sigma \in \mathcal{F}_{\rm int}$, define $1^\sigma \in X_{\mathcal{D},0}$ by $(1^\sigma)_\sigma = 1$ and $(1^\sigma)_{\sigma'} = 0$ if $\sigma' \neq \sigma$. Then $(\chi_\sigma)_{\sigma\in\mathcal{F}_{\rm int}} = (\Pi_{\mathcal{D}} 1^\sigma)_{\sigma\in\mathcal{F}_{\rm int}}$ forms a basis of $V^{(1)}$ and, using this basis in Definition 9.4, we see that the $\mathbb{P}_1$ NCFE GD given by Definitions 9.7 and 9.4 is the non-conforming $\mathbb{P}_1$ LLE GD. ∎

9.3.2 Estimate on $S_{\mathcal{D}}$ for the non-conforming $\mathbb{P}_1$ LLE gradient discretisation

As mentioned above, the properties of sequences of non-conforming $\mathbb{P}_1$ LLE GDs follow from Theorem 9.10 (and Lemma 9.11). Proposition 9.5 gives a bound on $C_{\mathcal{D}}$, and optimal estimates on $W_{\mathcal{D}}$ (since the $\mathbb{P}_1$ NCFE is a non-conforming order 1 method, $\mathcal{O}(h_{\mathcal{M}})$ estimates on $W_{\mathcal{D}}$ are expected to be optimal). Lemma 9.9 provides an estimate on $S_{\mathcal{D}}(\varphi)$, but it is sub-optimal in the sense that it requires a strong regularity on φ. Using the results of Section 7.4, a better estimate on $S_{\mathcal{D}}$ can be obtained.
We first need to formally establish that the non-conforming $\mathbb{P}_1$ GD defined above is indeed an LLE GD, and to estimate its regularity.

Lemma 9.12. *Let $\mathfrak{T} = (\mathcal{M}, \mathcal{F}, \mathcal{P}, \mathcal{V})$ be a conforming simplicial mesh in the sense of Definition 7.4. Let $K \in \mathcal{M}$, $\pi_K = (\pi_K^\sigma)_{\sigma\in\mathcal{F}_K}$ be given by (9.33), and $\mathcal{G}_K = (\mathcal{G}_K^\sigma)_{\sigma\in\mathcal{F}_K}$ be given by (9.35). Then π_K is a $\mathbb{P}_0$-exact function reconstruction on K, and $\mathcal{G}_K$ is a $\mathbb{P}_1$-exact gradient reconstruction on K upon $(\overline{\boldsymbol{x}}_\sigma)_{\sigma\in\mathcal{F}_K}$. Hence, $\mathcal{D}$ defined in Section 9.3.1 is an LLE GD.*
Moreover, there exists C_9, depending only on d and $\varrho \geq \kappa_{\mathfrak{T}}$ (see (7.10)), such that

$$\mathrm{reg}_{\rm LLE}(\mathcal{D}) \leq C_9. \tag{9.37}$$

Proof. Let $K \in \mathcal{M}$. The convex hull $\widetilde{K}$ of the centres of mass $(\overline{\boldsymbol{x}}_\sigma)_{\sigma\in\mathcal{F}_K}$ of the faces of K is a d-simplex (see Figure 9.1). Applying Lemma 8.6 to $\widetilde{K}$ instead of K and with $k = 1$ shows that, for any given real numbers $(a_\sigma)_{\sigma\in\mathcal{F}_K}$, there exists a unique affine map that takes these values at the face centres $(\overline{\boldsymbol{x}}_\sigma)_{\sigma\in\mathcal{F}_K}$. This proves in particular that (9.33) properly defines the basis functions $(\pi_K^\sigma)_{\sigma\in\mathcal{F}_K}$.
The map $\sum_{\sigma\in\mathcal{F}_K} \pi_K^\sigma$ is affine and takes the value 1 at each of the face centres of K, exactly as the constant function equal to 1. These two affine functions must therefore coincide, which shows that π_K is a $\mathbb{P}_0$-exact function reconstruction on K.
Let A be an affine map. The affine function $\sum_{\sigma\in\mathcal{F}_K} A(\overline{\boldsymbol{x}}_\sigma)\pi_K^\sigma$ has the same values as A at $(\overline{\boldsymbol{x}}_\sigma)_{\sigma\in\mathcal{F}_K}$, and is therefore equal to A. As a consequence, on K,

$$\sum_{\sigma\in\mathcal{F}_K} A(\overline{\boldsymbol{x}}_\sigma)\mathcal{G}_K^\sigma = \sum_{\sigma\in\mathcal{F}_K} A(\overline{\boldsymbol{x}}_\sigma)\nabla\pi_K^\sigma = \nabla \sum_{\sigma\in\mathcal{F}_K} A(\overline{\boldsymbol{x}}_\sigma)\pi_K^\sigma = \nabla A.$$

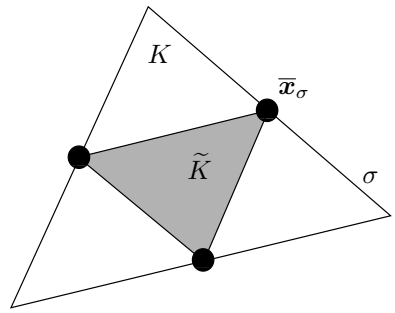


Fig. 9.1. A simplex K and the convex hull $\widetilde{K}$ of its face centres

Hence, $\mathcal{G}_K$ is a $\mathbb{P}_1$-exact gradient reconstruction on K upon $(\overline{\boldsymbol{x}}_\sigma)_{\sigma\in\mathcal{F}_K}$.

We now establish the upper bound on $\mathrm{reg}_{\mathrm{LLE}}(\mathcal{D})$. If $K \in \mathcal{M}$, the simplex $\widetilde{K} \subset K$ created by $(\overline{\boldsymbol{x}}_\sigma)_{\sigma\in\mathcal{F}_K}$ is a symmetry and dilatation by a factor $1/d$ of K. Hence, its regularity factor "*diameter of $\widetilde{K}$ over the radius of the largest ball inscribed in $\widetilde{K}$*" is identical to that of K, which is bounded by $\kappa_{\mathfrak{T}}$. Over $\widetilde{K}$ the functions π_K^σ are defined as affine functions with values 0 or 1 at the vertices of $\widetilde{K}$. Hence, as in the proof of Lemma 8.9 we can use Lemma 8.8 with the vertices of $\widetilde{K}$ as points $\boldsymbol{x}_i$ to see that

$$\|\pi_K^\sigma\|_{L^\infty(\widetilde{K})} \le C_{10} \quad \text{and} \quad \|\nabla\pi_K^\sigma\|_{L^\infty(\widetilde{K})} \le C_{10} h_{\widetilde{K}}^{-1}, \tag{9.38}$$

where C_{10} depends only on d and $\varrho \ge \kappa_{\mathfrak{T}}$. Since $\nabla\pi_K^\sigma$ is constant in K and $h_{\widetilde{K}} = h_K/d$, we deduce that

$$\|\nabla\pi_K^\sigma\|_{L^\infty(K)} \le C_{10} d h_K^{-1}. \tag{9.39}$$

We then write $\pi_K^\sigma(\boldsymbol{x}) = \pi_K^\sigma(\boldsymbol{y}) + (\boldsymbol{x}-\boldsymbol{y})\cdot\nabla\pi_K^\sigma$ for any $\boldsymbol{x} \in K$ and $\boldsymbol{y} \in \widetilde{K}$, and use $|\boldsymbol{x}-\boldsymbol{y}| \le h_K$ to infer from (9.38) and (9.39) that

$$\|\pi_K^\sigma\|_{L^\infty(K)} \le C_{10} + C_{10} d. \tag{9.40}$$

Remark 7.32 and Estimates (9.39) and (9.40) give an upper bound on the first two terms in the definition (7.34) of $\mathrm{reg}_{\mathrm{LLE}}(\mathcal{D})$. This upper bound depends only on d and ϱ. The proof is complete by noticing all points $(\boldsymbol{x}_i)_{i\in I_K} = (\overline{\boldsymbol{x}}_\sigma)_{\sigma\in\mathcal{F}_K}$ involved in the third term of $\mathrm{reg}_{\mathrm{LLE}}(\mathcal{D})$ belong to $\overline{K}$, which shows that this third term vanishes. ∎

Proposition 9.13 (Estimate on $S_{\mathcal{D}}$ for non-conforming $\mathbb{P}_1$ GD). *Let $\mathfrak{T}$ be a conforming simplicial mesh of Ω in the sense of Definition 7.4, and $\mathcal{D}$ be the non-conforming $\mathbb{P}_1$ LLE GD on $\mathfrak{T}$ as in Section 9.3.1. Assume that $p > d/2$ and take $\varrho \ge \kappa_{\mathfrak{T}}$ (see (7.10)). Then there exists $C_{11} > 0$, depending only on p, d, Ω and ϱ, such that*

$$\forall \varphi \in W^{2,p}(\Omega) \cap W_0^{1,p}(\Omega)\,,\ S_{\mathcal{D}}(\varphi) \le C_{11} h_{\mathcal{M}} \left\|\varphi\right\|_{W^{2,p}(\Omega)}\,. \tag{9.41}$$

Here, $S_{\mathcal{D}}$ is the measure of GD-consistency defined by (2.2). This means that the space size (see Definition 2.22) of the GD is such that

$$h_{\mathcal{D}}(W^{2,p}(\Omega) \cap W_0^{1,p}(\Omega); W^{1,p'}(\Omega)^d) \le \max(C_{11}, C_1) h_{\mathcal{M}},$$

where C_1 is defined in Proposition 9.5 (owing to Lemma B.4 page 409, C_1 has the same dependencies as C_{11}).

Proof. For any $K \in \mathcal{M}$, the approximation points $(\boldsymbol{x}_i)_{i\in I_K} = (\overline{\boldsymbol{x}}_\sigma)_{\sigma\in\mathcal{F}_K}$ all belong to $\overline{K}$. Using Lemma 9.12, Lemma B.1 and Lemma B.4, we can invoke Proposition 7.64 and the conclusion follows. ∎

9.4 Non-conforming $\mathbb{P}_1$ methods for Neumann and Fourier BCs

9.4.1 Neumann boundary conditions

Definition 7.54 of LLE GDs for Neumann boundary conditions provides a straightforward definition of non-conforming $\mathbb{P}_1$ GDs for these conditions, by simply using the same I_Ω, I_∂, $\Pi_{\mathcal{D}}$, $\nabla_{\mathcal{D}}$ as in Section 9.3.1.
The proof of Proposition 9.3 shows that Inequality (9.7) is valid even if w does not satisfy (9.2) for boundary edges. Apply this inequality to $w = \Pi_{\mathcal{D}} v$, for $v \in X_{\mathcal{D}}$ such that $\left\|\nabla_{\mathcal{D}} v\right\|_{L^p(\Omega)^d} = 0$. Then the right-hand side vanishes and the definition (7.7f) of $|\cdot|_{\mathfrak{T},p}$ show that all $(v_\sigma)_{\sigma\in\mathcal{F}}$ are identical, equal to some $c \in \mathbb{R}$. As a consequence, $\Pi_{\mathcal{D}} v = c$ over Ω and thus, if $\int_\Omega \Pi_{\mathcal{D}} v(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = 0$, c must be equal to 0. This shows that the the quantity (3.1) is indeed a norm on $X_{\mathcal{D}}$.

For non-homogeneous Neumann boundary conditions, the trace reconstruction $\mathbb{T}_{\mathcal{D}} : X_{\mathcal{D}} \to L^p(\partial\Omega)$ can be naturally defined in a similar way as $\mathbb{T}_{\mathfrak{T}}$ in (7.7d), that is,

$$\forall v \in X_{\mathcal{D}}\,,\ \forall \sigma \in \mathcal{F}_{\text{ext}}\ :\ \mathbb{T}_{\mathcal{D}} v = v_\sigma \text{ on } \sigma. \tag{9.42}$$

Since the regularity factor $\mathrm{reg}_{\text{LLE}}(\mathcal{D})$ for Neumann BCs is defined as for Dirichlet BCs, Lemma 9.12 still applies and shows that this factor remains bounded if $\kappa_{\mathfrak{T}}$ is bounded. Define then the control $\boldsymbol{\Phi} : X_{\mathcal{D}} \to X_{\mathfrak{T}}$ by $\boldsymbol{\Phi} = \Psi \circ \Pi_{\mathcal{D}}$ with Ψ given by (9.6). Since (9.7)–(9.9) are also valid for non-zero boundary conditions, we have $\left\|\boldsymbol{\Phi}\right\|_{\mathcal{D},\mathfrak{T}} \le (C_2^p + 1)^{1/p}$, $\omega^{\Pi}(\mathcal{D}, \mathfrak{T}, \boldsymbol{\Phi}) \le C_2 h_{\mathcal{M}}$ and $\omega^{\nabla}(\mathcal{D}, \mathfrak{T}, \boldsymbol{\Phi}) = 0$. We also obviously have $\omega^{\mathbb{T}}(\mathcal{D}, \mathfrak{T}, \boldsymbol{\Phi}) = 0$. Hence, Corollary 7.18 and Proposition 7.55 give the following theorem.

Theorem 9.14 (Properties of non-conforming $\mathbb{P}_1$ GDs for Neumann BCs). *Let $(\mathcal{D}_m)_{m\in\mathbb{N}}$ be a sequence of non-conforming $\mathbb{P}_1$ GDs for Neumann boundary conditions as above, defined from underlying conforming simplicial meshes $(\mathfrak{T}_m)_{m\in\mathbb{N}}$. Assume that $(\kappa_{\mathfrak{T}_m})_{m\in\mathbb{N}}$ is bounded (see (7.10)), and that $h_{\mathcal{M}_m} \to 0$ as $m \to \infty$.*
Then the sequence $(\mathcal{D}_m)_{m\in\mathbb{N}}$ is coercive, GD-consistent, limit-conforming and compact in the sense of Definitions 3.13, 3.4, 3.14 and 3.15.

Proposition 7.70 and Theorem 7.17 also give estimates on $C_\mathcal{D}$, $W_\mathcal{D}$ and $S_\mathcal{D}$ that are similar to (9.20), (9.22) and (9.41). The constants depend only on Ω, p and an upper bound of $\kappa_\mathfrak{T}$.

Remark 9.15 (Other choice for the trace reconstruction)
Recalling the definition of the global basis functions π^σ, see Section 7.3.3, it is also possible to replace (9.42) by

$$\mathbb{T}^*_\mathcal{D} v = \sum_{\sigma\in\mathcal{F}_{\rm ext}} v_\sigma (\pi^\sigma)_{|\partial\Omega} = (\Pi_\mathcal{D} v)_{|\partial\Omega}.$$

Then, for any $K \in \mathcal{M}$, any $\sigma \in \mathcal{F}_K \cap \mathcal{F}_{\rm ext}$ and any $\boldsymbol{x} \in \sigma$, since $\nabla_\mathcal{D} v = \nabla((\Pi_\mathcal{D} v)_{|K})$ is constant in K,

$$|\mathbb{T}^*_\mathcal{D} v(\boldsymbol{x}) - \mathbb{T}_\mathcal{D} v(\boldsymbol{x})| = |\Pi_\mathcal{D} v(\boldsymbol{x}) - \Pi_\mathcal{D} v(\overline{\boldsymbol{x}}_\sigma)| \le h_K |(\nabla_\mathcal{D} v)_{|K}|.$$

Taking the power p of this estimate and integrating over σ gives

$$\int_\sigma |\mathbb{T}^*_\mathcal{D} v(\boldsymbol{x}) - \mathbb{T}_\mathcal{D} v(\boldsymbol{x})|^p \mathrm{d}\boldsymbol{x} \le h_K^p |\sigma|\, |(\nabla_\mathcal{D} v)_{|K}|^p.$$

Since $h_K|\sigma| \le C_{12}|K|$, where C_{12} depends only on an upper bound on $\theta_\mathfrak{T} \le \kappa_\mathfrak{T} + d + 1$,

$$\int_\sigma |\mathbb{T}^*_\mathcal{D} v(\boldsymbol{x}) - \mathbb{T}_\mathcal{D} v(\boldsymbol{x})|^p \mathrm{d}\boldsymbol{x} \le C_{12} h_\mathcal{M}^{p-1} |K|\, |(\nabla_\mathcal{D} v)_{|K}|^p.$$

Sum this estimate over $\sigma \in \mathcal{F}_{\rm ext}$. A given cell K can have at most $d+1$ boundary faces (and only in the trivial case where $\Omega = K$, otherwise $\mathrm{Card}(\mathcal{F}_K \cap \mathcal{F}_{\rm ext}) \le d$), and thus

$$\begin{aligned}
\|\mathbb{T}^*_\mathcal{D} v - \mathbb{T}_\mathcal{D}\|^p_{L^p(\partial\Omega)} &\le (d+1) C_{12} h_\mathcal{M}^{p-1} \sum_{K\in\mathcal{M},\, \partial K\cap\partial\Omega\neq\emptyset} |K|\, |(\nabla_\mathcal{D} v)_{|K}|^p \\
&\le (d+1) C_{12} h_\mathcal{M}^{p-1} \sum_{K\in\mathcal{M}} |K|\, |(\nabla_\mathcal{D} v)_{|K}|^p \\
&= (d+1) C_{12} h_\mathcal{M}^{p-1} \|\nabla_\mathcal{D} v\|^p_{L^p(\Omega)^d}.
\end{aligned}$$

This estimate enables us to transport the analysis made with $\mathbb{T}_\mathcal{D}$ to the GD based on the trace reconstruction $\mathbb{T}^*_\mathcal{D}$ instead.

9.4.2 Fourier boundary conditions

Starting from the non-conforming $\mathbb{P}_1$ GD for Dirichlet boundary conditions, we follow Definition 7.57 in Section 7.3.6 to define a non-conforming $\mathbb{P}_1$ GD for Fourier boundary conditions.
The boundary mesh $\mathcal{M}_\partial$ is simply $\mathcal{F}_{\text{ext}}$, and the trace reconstruction (9.42) corresponds, for $K_\partial = \sigma \in \mathcal{F}_{\text{ext}}$, to $I_\sigma = \{\sigma\}$ and $\pi_\sigma^\sigma = 1$ on σ, $\pi_\sigma^\sigma = 0$ outside σ. The bound on $\text{reg}_{\text{LLE}}(\mathcal{D})$ for Fourier boundary conditions therefore easily follows from the bound on this quantity for Dirichlet boundary conditions, and the GD-consistency (under boundedness of $\kappa_{\mathfrak{T}_m}$) is therefore a consequence of Proposition 7.58.
As noticed in Remark 7.20, the work done for Neumann boundary conditions then immediately show that Theorem 9.14 also applies for Fourier boundary conditions. Similarly, we could obtain estimates on $S_\mathcal{D}$, $C_\mathcal{D}$ and $W_\mathcal{D}$ as in Propositions 9.13 and 9.5.
We finally remark that, instead of $\mathbb{T}_\mathcal{D}$ defined by (9.42), we can also use $\mathbb{T}^*_\mathcal{D}$ defined in Remark 9.15

9.5 Non-conforming $\mathbb{P}_1$ finite elements for non-homogeneous Dirichlet boundary conditions

For non-homogeneous Dirichlet conditions, the interpolation operator $\mathcal{I}_{\mathcal{D},\partial}$ is defined by

$$\forall g \in W^{1-\frac{1}{p},p}(\partial\Omega)\,,\ \forall \sigma \in \mathcal{F}_{\text{ext}}\ :\ (\mathcal{I}_{\mathcal{D},\partial} g)_\sigma = \frac{1}{|\sigma|}\int_\sigma g(\boldsymbol{x})\mathrm{d}\gamma(\boldsymbol{x}). \tag{9.43}$$

This interpolant clearly satisfies (7.65) since, for any $i = \sigma \in I_\partial = \mathcal{F}_{\text{ext}}$, $\boldsymbol{x}_i = \overline{\boldsymbol{x}}_\sigma$ is the centre of mass of σ and therefore, if $\varphi \in C^\infty(\overline{\Omega})$, $\varphi(\boldsymbol{x}_i) = \frac{1}{|\sigma|}\int_\sigma \varphi(\boldsymbol{x})\mathrm{d}\gamma(\boldsymbol{x}) + \mathcal{O}(\text{diam}(\sigma)^2)$.
We now check that (2.96) holds with C_{12} depending only on Ω, p and an upper bound of $\kappa_\mathfrak{T}$. To this end, take $\varphi \in W^{1,p}(\Omega)$ and let $u = I_\mathfrak{T}\varphi \in X_\mathfrak{T}$, where $I_\mathfrak{T}$ is the interpolant defined by (B.10). Let $v \in X_\mathcal{D}$ be such that $v_\sigma = u_\sigma$ for all $\sigma \in \mathcal{F}$. Since $v - \mathcal{I}_{\mathcal{D},\partial}\gamma\varphi \in X_{\mathcal{D},0}$, (2.96) is proved if we can establish that

$$\|\Pi_\mathcal{D} v\|_{L^p(\Omega)} + \|\nabla_\mathcal{D} v\|_{L^p(\Omega)^d} \le C_{12}\,\|\varphi\|_{W^{1,p}(\Omega)}\,. \tag{9.44}$$

Let $K \in \mathcal{M}$. We have $(\nabla_\mathcal{D} v)_{|K} = \nabla(\Pi_\mathcal{D} v)_{|K}$ and $(\Pi_\mathcal{D} v)_{|K}$ is the affine map on K such that, for all $\sigma \in \mathcal{F}_K$, $(\Pi_\mathcal{D} v)(\overline{\boldsymbol{x}}_\sigma) = u_\sigma$. Since $\overline{\nabla}_K$ is a $\mathbb{P}_1$-exact gradient reconstruction on K upon $(\overline{\boldsymbol{x}}_\sigma)_{\sigma\in\mathcal{F}_K}$ (see Lemma B.10), we infer

$$(\nabla_\mathcal{D} v)_{|K} = \nabla(\Pi_\mathcal{D} v)_{|K} = \overline{\nabla}_K u.$$

Hence, $\nabla_\mathcal{D} v = \overline{\nabla}_\mathfrak{T} u$. Using (B.29) and Proposition B.7 thus shows that

$$\|\nabla_{\mathcal{D}} v\|_{L^p(\Omega)^d} = \left\|\overline{\nabla}_{\mathfrak{T}} u\right\|_{L^p(\Omega)^d} \leq d^{\frac{p-1}{p}} |u|_{\mathfrak{T},p} \leq C_{13} \|\nabla\varphi\|_{L^p(\Omega)^d} \tag{9.45}$$

where C_{13} depends only on Ω, p and an upper bound on $\kappa_{\mathfrak{T}}$ (use Lemma B.4 to get, from the upper bound on $\kappa_{\mathfrak{T}}$, an upper bound on $\theta_{\mathfrak{T}}$ and thus enable the usage of Proposition B.7).

We now turn to the estimate on $\Pi_{\mathcal{D}} v$. Since this function is affine in each cell K, with gradient $(\nabla_{\mathcal{D}} v)_{|K}$ and value u_σ at $\overline{\boldsymbol{x}}_\sigma$, for all $\boldsymbol{x} \in K$ and any $\sigma \in \mathcal{F}_K$ we have

$$|\Pi_{\mathcal{D}} v(\boldsymbol{x}) - u_\sigma| \leq h_K |(\nabla_{\mathcal{D}} v)_{|K}|.$$

Recalling that $u_\sigma = \frac{1}{|\sigma|}\int_\sigma \varphi(\boldsymbol{y})\mathrm{d}\gamma(\boldsymbol{y})$ and using Estimate (B.11), we infer

$$\left|\Pi_{\mathcal{D}} v(\boldsymbol{x}) - \frac{1}{|K|}\int_K \varphi(\boldsymbol{y})\mathrm{d}\boldsymbol{y}\right|^p \leq C_{14} h_K^p |(\nabla_{\mathcal{D}} v)_{|K}|^p + C_{14}\frac{h_K^p}{|K|}\int_K |\nabla\varphi(\boldsymbol{y})|^p \mathrm{d}\boldsymbol{y}$$

with C_{14} depending only on Ω, p and an upper bound on $\kappa_{\mathfrak{T}}$. As a consequence, the power-of-sums estimate (D.12) with $\alpha = p$ and Jensen's inequality yield

$$\begin{aligned}|\Pi_{\mathcal{D}} v(\boldsymbol{x})|^p \leq \frac{C_{15}}{|K|}\int_K |\varphi(\boldsymbol{y})|^p \mathrm{d}\boldsymbol{y} + C_{15} h_K^p |(\nabla_{\mathcal{D}} v)_{|K}|^p \\ + C_{15}\frac{h_K^p}{|K|}\int_K |\nabla\varphi(\boldsymbol{y})|^p \mathrm{d}\boldsymbol{y}\end{aligned}$$

with C_{15} depending only on Ω, p and an upper bound on $\kappa_{\mathfrak{T}}$. Integrate this over $\boldsymbol{x} \in K$ and sum over $K \in \mathcal{M}$ to get

$$\begin{aligned}\|\Pi_{\mathcal{D}} v\|_{L^p(\Omega)}^p \leq C_{15} \|\varphi\|_{L^p(\Omega)}^p + C_{15}\mathrm{diam}(\Omega)^p \|\nabla_{\mathcal{D}} v\|_{L^p(\Omega)^d}^p \\ + C_{15}\mathrm{diam}(\Omega)^p \|\nabla\varphi\|_{L^p(\Omega)^d}^p .\end{aligned}$$

Estimate (9.44) follows from this inequality and (9.45).

Since (7.65) and (2.96) hold, Proposition 7.53 can then be invoked (using Lemma 9.12 to bound $\mathrm{reg}_{\mathrm{LLE}}(\mathcal{D})$) and shows that sequences of non-conforming $\mathbb{P}_1$ GDs for non-homogeneous Dirichlet BCs are GD-consistent, provided that the regularity factors $(\kappa_{\mathfrak{T}_m})_{m\in\mathbb{N}}$ remain bounded and that $h_{\mathcal{M}_m} \to 0$.

The coercivity, limit-conformity and compactness of GDs for non-homogeneous Dirichlet conditions are identical to the same properties for homogeneous Dirichlet conditions, which follow from Theorem 9.10.

9.6 Mass-lumped non-conforming $\mathbb{P}_1$ gradient discretisation

In the case $d = 2$, if $\sigma \neq \sigma'$ are two different faces of the mesh,

$$\int_\Omega \pi^\sigma(\boldsymbol{x})\pi^{\sigma'}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = 0.$$

This property ensures that the non-conforming $\mathbb{P}_1$ method has a diagonal mass matrix. Nevertheless, the properties in Remark 2.13 are not satisfied, which might prevent the usage of the non-conforming $\mathbb{P}_1$ scheme for some non-linear problems. To recover a piecewise constant reconstruction, we apply to the non-conforming $\mathbb{P}_1$ GD the mass lumping process as in Definition 7.45.

Definition 9.16 (Mass-lumped non-conforming $\mathbb{P}_1$ GD). *Take a conforming simplicial mesh $\mathfrak{T} = (\mathcal{M}, \mathcal{F}, \mathcal{P}, \mathcal{V})$ of Ω in the sense of Definition 7.4, and let $\mathcal{D} = (X_{\mathcal{D},0}, \Pi_{\mathcal{D}}, \nabla_{\mathcal{D}})$ be the non-conforming $\mathbb{P}_1$ LLE GD built on $\mathfrak{T}$ as in Section 9.3.1.*
For $\sigma \in \mathcal{F}$, let $\Omega_\sigma = D_\sigma$ be the diamond around σ if $\sigma \in \mathcal{F}_{\rm int}$, and $\Omega_\sigma = D_{K,\sigma}$ be the half-diamond around σ if $\sigma \in \mathcal{F}_{\rm ext}$ with $\mathcal{M}_\sigma = \{K\}$ (see Definition 7.2 for the definitions of these diamond and half-diamond, and Figure 9.2 for an illustration).
A mass-lumped non-conforming $\mathbb{P}_1$ GD is defined by $\mathcal{D}^{\rm ML} = (X_{\mathcal{D},0}, \Pi_{\mathcal{D}}^{\rm ML}, \nabla_{\mathcal{D}})$, where $\Pi_{\mathcal{D}}^{\rm ML}$ is the piecewise constant reconstruction built from $(\Omega_\sigma)_{\sigma\in\mathcal{F}}$, that is

$$\forall v \in X_{\mathcal{D},0}\,,\ \forall \sigma \in \mathcal{F}\,,\ \Pi_{\mathcal{D}}^{\rm ML} v = v_\sigma \ \text{ on } \Omega_\sigma.$$

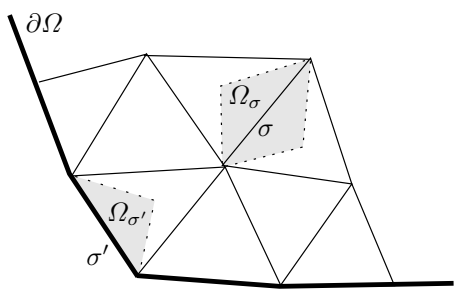


Fig. 9.2. Partition for the mass-lumping of the non-conforming $\mathbb{P}_1$ gradient discretisation.

As for the mass-lumped $\mathbb{P}_1$ GD, the properties of this mass-lumped non-conforming $\mathbb{P}_1$ GD follow directly from Theorem 7.49.

Theorem 9.17 (Properties of mass-lumped non-conforming $\mathbb{P}_1$ GDs). *Let $(\mathfrak{T}_m)_{m\in\mathbb{N}}$ be a sequence of conforming simplicial meshes of Ω in the sense of Definition 7.4, and let $(\mathcal{D}_m^{\rm ML})_{m\in\mathbb{N}}$ be the corresponding mass-lumped non-conforming $\mathbb{P}_1$ GDs given by Definition 9.16. Assume that $\sup_{m\in\mathbb{N}} \kappa_{\mathfrak{T}_m} < +\infty$ (see (7.10)), and that $h_{\mathcal{M}_m} \to 0$ as $m \to \infty$.*
Then $(\mathcal{D}_m^{\rm ML})_{m\in\mathbb{N}}$ is coercive, GD-consistent, limit-conforming, compact, and has a piecewise constant reconstruction in the sense of Definitions 2.2, 2.4, 2.5, 2.8 and 2.12.

Proof. In each $D_{K,\sigma}$, $\Pi_{\mathcal{D}_m} v$ is affine and $\Pi^{\rm ML}_{\mathcal{D}_m} v = \Pi_{\mathcal{D}_m} v(\overline{\boldsymbol{x}}_\sigma)$. Hence, for $\boldsymbol{x} \in D_{K,\sigma}$,

$$\begin{aligned}|\Pi_{\mathcal{D}_m} v(\boldsymbol{x}) - \Pi^{\rm ML}_{\mathcal{D}_m} v(\boldsymbol{x})| &= |\Pi_{\mathcal{D}_m} v(\boldsymbol{x}) - \Pi_{\mathcal{D}_m} v(\overline{\boldsymbol{x}}_\sigma)| \\ &\le h_{\mathcal{M}_m} |(\nabla_{\mathcal{D}} v)_{|D_{K,\sigma}}| = h_{\mathcal{M}_m} |\nabla_{\mathcal{D}} v(\boldsymbol{x})|.\end{aligned}$$

Raising to the power p, integrating over $D_{K,\sigma}$, and summing over $\sigma \in \mathcal{F}_K$ and $K \in \mathcal{M}$ we obtain

$$\left\| \Pi_{\mathcal{D}_m} v - \Pi^{\rm ML}_{\mathcal{D}_m} v \right\|_{L^p(\Omega)} \le h_{\mathcal{M}_m} \left\| \nabla_{\mathcal{D}_m} v \right\|_{L^p(\Omega)^d}. \tag{9.46}$$

The conclusion then follows from the properties of sequences of non-conforming $\mathbb{P}_1$ GDs (Theorem 9.10) and from Theorem 7.49. ∎

Remark 9.18. As in Remark 8.18, Propositions 9.13 and 9.5, Estimate (9.46) and Remark 7.51 show that, for $p > d/2$,

$$S_{\mathcal{D}^{\rm ML}_m}(\varphi) \le C h_{\mathcal{M}_m} \|\varphi\|_{W^{2,p}(\Omega)}$$

and

$$W_{\mathcal{D}^{\rm ML}_m}(\boldsymbol{\varphi}) \le C h_{\mathcal{M}_m} \|\boldsymbol{\varphi}\|_{W^{1,p'}(\Omega)^d},$$

with C not depending on m, φ or $\boldsymbol{\varphi}$. Mass-lumped non-conforming $\mathbb{P}_1$ GSs are thus order 1 schemes: if the exact solution of the linear elliptic problem (2.20) belongs to H^2 and $d = 1, 2, 3$, then the estimates (2.25) and (2.26) are $\mathcal{O}(h_{\mathcal{M}})$ when the mass-lumped non-conforming $\mathbb{P}_1$ GD is used in the gradient scheme (2.23).

10

H_{div} conforming GDs from mixed finite element methods

In this chapter, the mixed finite element (MFE) schemes for an isotropic linear elliptic problem is shown to be a GDM for $p = 2$ and homogeneous Dirichlet boundary conditions; the reconstructed gradient, which belongs to $H_{\mathrm{div}}(\Omega)$ is obtain by solving the non-local subsystem of the MFE scheme corresponding to the first equation of the mixed form ((10.4b) below).
We prove that, under the so-called "inf-sup" or "LBB" condition, the gradient discretisation thus obtained is coercive, GD-consistent and limit-conforming. As a consequence, even though it has been designed only by considering the Laplace problem, this GD can then be used on any of the models presented in Parts I and II, provided that the corresponding convergence analysis only relies on the coercivity, GD-consistency and limit-conformity properties (with $p = 2$).
After considering general MFE schemes, the special case of the popular $\mathbb{RT}_k$ mixed finite elements is considered. For this method, the corresponding GDs are additionally shown to be compact. Turning then to the hybrid form of $\mathbb{RT}_k$, a different GD is constructed in which, contrary to the previous one, the reconstructed gradient is computed locally (as is the case for most GDs studied elsewhere in this book). We prove that, under standard assumptions on the meshes, sequences of hybrid $\mathbb{RT}_k$ GDs are coercive, GD-consistent, limit-conforming and compact. High order estimates on $S_{\mathcal{D}}$ and $W_{\mathcal{D}}$ are also established.

10.1 Mixed finite element schemes

10.1.1 Presentation and error estimate

Let us first recall the formulation of the mixed finite element (MFE) method for the linear isotropic diffusion problem[81]. We consider the problem

$$\begin{aligned}&\text{Find } \overline{u} \in H_0^1(\Omega) \text{ such that, for all } v \in H_0^1(\Omega),\\ &\int_\Omega \nabla \overline{u}(\boldsymbol{x}) \cdot \nabla v(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} = \int_\Omega f(\boldsymbol{x}) v(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}.\end{aligned} \tag{10.1}$$

The assumptions are as usual:

$$\begin{aligned}&\Omega \text{ is an open bounded connected subset of } \mathbb{R}^d \ (d \in \mathbb{N}^\star)\\ &\text{and } f \in L^2(\Omega).\end{aligned} \tag{10.2}$$

Take finite dimensional spaces $\boldsymbol{V}_h^{\rm div} \subset H_{\rm div}(\Omega)$ and $W_h \subset L^2(\Omega)$ such that there exists $\beta_h > 0$ with

$$\forall q \in W_h, \ \sup_{\boldsymbol{w} \in \boldsymbol{V}_h^{\rm div}} \frac{1}{\|\boldsymbol{w}\|_{H_{\rm div}(\Omega)}} \int_\Omega q(\boldsymbol{x}) {\rm div} \boldsymbol{w}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \geq \beta_h \|q\|_{L^2(\Omega)}. \tag{10.3}$$

The formulation of the MFE scheme for Problem (10.1) then reads

$$(\boldsymbol{v}, q) \in \boldsymbol{V}_h^{\rm div} \times W_h, \tag{10.4a}$$

$$\forall \boldsymbol{w} \in \boldsymbol{V}_h^{\rm div}, \qquad \int_\Omega \boldsymbol{w}(\boldsymbol{x}) \cdot \boldsymbol{v}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} - \int_\Omega q(\boldsymbol{x}) {\rm div} \boldsymbol{w}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} = 0, \tag{10.4b}$$

$$\forall \psi \in W_h, \qquad \int_\Omega \psi(\boldsymbol{x}) {\rm div} \boldsymbol{v}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} = \int_\Omega \psi(\boldsymbol{x}) f(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}. \tag{10.4c}$$

There exists a unique $(\boldsymbol{v}, q)$ solution to (10.4), and it moreover satisfies the following stability and error estimates (see [81, Theorem 5.2 p.38 and Theorem 5.3 p.39]): if $0 < \beta \leq \beta_h$, there exists C_1 depending only on Ω and β such that

$$\|q\|_{L^2(\Omega)} + \|\boldsymbol{v}\|_{H_{\rm div}(\Omega)} \leq C_1 \|f\|_{L^2(\Omega)} \tag{10.5}$$

and

$$\begin{aligned}&\|q - \overline{u}\|_{L^2(\Omega)} + \|\boldsymbol{v} + \nabla \overline{u}\|_{H_{\rm div}(\Omega)}\\ &\qquad \leq C_1 \left(\inf_{\psi \in W_h} \|\psi - \overline{u}\|_{L^2(\Omega)} + \inf_{\boldsymbol{w} \in \boldsymbol{V}_h^{\rm div}} \|\boldsymbol{w} + \nabla \overline{u}\|_{H_{\rm div}(\Omega)} \right).\end{aligned} \tag{10.6}$$

10.1.2 Construction and analysis of a mixed finite element GD

We construct a GD, in the sense of Definition 2.1, such that the corresponding gradient scheme (2.23) (with $\boldsymbol{F} = 0$ and $\Lambda = {\rm Id}$) for (10.1) is the MFE scheme (10.4). Let $(\chi_i)_{i \in I}$ be a basis of W_h and define the mixed finite element gradient discretisation (MFEGD) $\mathcal{D} = (X_{\mathcal{D},0}, \Pi_{\mathcal{D}}, \nabla_{\mathcal{D}})$ by:

$$X_{\mathcal{D},0} = \{u = (u_i)_{i \in I} \ : \ u_i \in \mathbb{R} \text{ for all } i \in I\}, \tag{10.7a}$$

$$\forall u \in X_{\mathcal{D},0}, \ \Pi_{\mathcal{D}} u = \sum_{i \in I} u_i \chi_i, \tag{10.7b}$$

$$\forall u \in X_{\mathcal{D},0}\,,\ \nabla_{\mathcal{D}} u \in \boldsymbol{V}_h^{\mathrm{div}} \text{ and} \tag{10.7c}$$

$$\int_\Omega \boldsymbol{w}(\boldsymbol{x}) \cdot \nabla_{\mathcal{D}} u(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} = -\int_\Omega \Pi_{\mathcal{D}} u(\boldsymbol{x}) \mathrm{div} \boldsymbol{w}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}\,,\ \forall \boldsymbol{w} \in \boldsymbol{V}_h^{\mathrm{div}}. \tag{10.7d}$$

In order for (10.7) to define a GD, the system (10.7c)–(10.7d) should define one and only one $\nabla_{\mathcal{D}} u$, and $\|\cdot\|_{\mathcal{D}} := \|\nabla_{\mathcal{D}} \cdot\|_{L^2(\Omega)^d}$ should be a norm on $X_{\mathcal{D},0}$. The existence and uniqueness of $\nabla_{\mathcal{D}} u$ results from the Riesz representation theorem in $\boldsymbol{V}_h^{\mathrm{div}}$ for the $L^2(\Omega)^d$-inner product, which shows that, for a given $\Pi_{\mathcal{D}} u$, (10.7d) has a unique solution $\nabla_{\mathcal{D}} u \in \boldsymbol{V}_h^{\mathrm{div}}$. The norm property of $\|\nabla_{\mathcal{D}} \cdot\|_{L^2(\Omega)^d}$ follows from the coercivity property shown in Theorem 10.3 below.

Remark 10.1 (H_{div} gradient discretisation). The interest of the MFEGD defined by (10.7) is to provide a more regular reconstructed gradient than certain other GDs. Indeed, we have here $\nabla_{\mathcal{D}} v \in H_{\mathrm{div}}(\Omega)$, whereas for most GDs the reconstructed gradient usually only belongs to $L^2(\Omega)^d$.

The following theorem establishes the link between the GD (10.7) and the mixed finite element method.

Theorem 10.2 (The mixed finite element method is a GDM in the isotropic case). *Under Assumption (10.2), if u is the solution of the GS (2.23) (with $\boldsymbol{F} = 0$ and $\Lambda = \mathrm{Id}$) using the GD defined by (10.7), and if $(\boldsymbol{v}, q)$ is the solution of the mixed finite element scheme (10.4), then $(\boldsymbol{v}, q) = (-\nabla_{\mathcal{D}} u, \Pi_{\mathcal{D}} u)$.*

Proof. Let $u \in X_{\mathcal{D},0}$ be the solution to (2.23) and let us prove that $(\boldsymbol{v}, q) = (-\nabla_{\mathcal{D}} u, \Pi_{\mathcal{D}} u) \in \boldsymbol{V}_h^{\mathrm{div}} \times W_h$ is a solution of (10.4). We first observe that (10.7d) ensures (10.4b). Take then $\psi \in W_h$. Since $(\chi_i)_{i \in I}$ is a basis of W_h, there exists a unique $v \in X_{\mathcal{D},0}$ such that $\psi = \sum_{i \in I} v_i \chi_i = \Pi_{\mathcal{D}} v$. The GS (2.23) gives

$$\int_\Omega \nabla_{\mathcal{D}} u(\boldsymbol{x}) \cdot \nabla_{\mathcal{D}} v(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} = \int_\Omega f(\boldsymbol{x}) \psi(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}. \tag{10.8}$$

Write (10.7d) with u replaced by $-v$ to see that, for all $\boldsymbol{w} \in \boldsymbol{V}_h^{\mathrm{div}}$,

$$-\int_\Omega \boldsymbol{w}(\boldsymbol{x}) \cdot \nabla_{\mathcal{D}} v(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} = \int_\Omega \Pi_{\mathcal{D}} v(\boldsymbol{x}) \mathrm{div} \boldsymbol{w}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}.$$

Substitute $\boldsymbol{w} = -\nabla_{\mathcal{D}} u = \boldsymbol{v}$ to obtain

$$\int_\Omega \nabla_{\mathcal{D}} u(\boldsymbol{x}) \cdot \nabla_{\mathcal{D}} v(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} = \int_\Omega \psi(\boldsymbol{x}) \mathrm{div} \boldsymbol{v}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}.$$

Combined with (10.8), this completes the proof of (10.4c).
Reciprocally, consider the solution $(\boldsymbol{v}, q)$ to (10.4). Since $q \in W_h$, as above there exists a unique $u \in X_{\mathcal{D},0}$ such that $q = \Pi_{\mathcal{D}} u$. Comparing (10.4b) and (10.7d) yields $\boldsymbol{v} = -\nabla_{\mathcal{D}} u$. Take $v \in X_{\mathcal{D},0}$, set $\psi = \Pi_{\mathcal{D}} v \in W_h$ and follow the

same computation as above to see, using (10.4c), that u is a solution to the gradient scheme (2.23). ∎

The properties of MFE gradient discretisations are established in the following theorem.

Theorem 10.3 (Properties of MFEGDs).
Let $\beta > 0$ and, for all $m \in \mathbb{N}$, let $\boldsymbol{V}_{h_m}^{\rm div} \subset H_{\rm div}(\Omega)$ and $W_{h_m} \subset L^2(\Omega)$ be such that:

1. *for all $m \in \mathbb{N}$, the inf-sup condition (10.3) holds with $\boldsymbol{V}_h^{\rm div} = \boldsymbol{V}_{h_m}^{\rm div}$, $W_h = W_{h_m}$ and $\beta_h = \beta$,*
2. *for all $\varphi \in L^2(\Omega)$,*
$$\lim_{m\to\infty} \inf_{\psi \in W_{h_m}} \|\psi - \varphi\|_{L^2(\Omega)} = 0,$$
3. *for all $\boldsymbol{\varphi} \in H_{\rm div}(\Omega)$,*
$$\lim_{m\to\infty} \inf_{\boldsymbol{w} \in \boldsymbol{V}_{h_m}^{\rm div}} \|\boldsymbol{\varphi} - \boldsymbol{w}\|_{H_{\rm div}(\Omega)} = 0.$$

Let $\mathcal{D}_m = (X_{\mathcal{D}_m,0}, \Pi_{\mathcal{D}_m}, \nabla_{\mathcal{D}_m})$ be the MFEGD defined by (10.7) from the pair of spaces $(\boldsymbol{V}_{h_m}^{\rm div}, W_{h_m})$. Then $(\mathcal{D}_m)_{m\in\mathbb{N}}$ is coercive, GD-consistent and limit-conforming in the sense of the definitions in Section 2.1.1.

Proof.
Coercivity. Let $m \in \mathbb{N}$ and $u \in X_{\mathcal{D}_m,0}$. Using (10.3) (in which the supremum is actually a maximum) for $q = \Pi_{\mathcal{D}_m} u$, take $\boldsymbol{w} \in \boldsymbol{V}_{h_m}^{\rm div}$ such that $\|\boldsymbol{w}\|_{H_{\rm div}(\Omega)} = 1$ and
$$\int_\Omega \Pi_{\mathcal{D}_m} u(\boldsymbol{x}) {\rm div}\boldsymbol{w}(\boldsymbol{x}) {\rm d}\boldsymbol{x} \geq \beta \left\|\Pi_{\mathcal{D}_m} u\right\|_{L^2(\Omega)}.$$
Invoking (10.7d) and using the Cauchy–Schwarz inequality yields
$$\beta \left\|\Pi_{\mathcal{D}_m} u\right\|_{L^2(\Omega)} \leq -\int_\Omega \boldsymbol{w}(\boldsymbol{x}) \cdot \nabla_{\mathcal{D}_m} u(\boldsymbol{x}) {\rm d}\boldsymbol{x} \leq \left\|\nabla_{\mathcal{D}_m} u\right\|_{L^2(\Omega)^d}.$$
Hence,
$$\|\Pi_{\mathcal{D}_m} u\|_{L^2(\Omega)} \leq \frac{1}{\beta} \|\nabla_{\mathcal{D}_m} u\|_{L^2(\Omega)^d} \tag{10.9}$$
and the coercivity property is proved with $C_P = \frac{1}{\beta}$.

GD-consistency. The GD-consistency follows from Lemma 2.16 if we prove that, for all $\varphi \in C_c^\infty(\Omega)$, $S_{\mathcal{D}_m}(\varphi) \to 0$ as $m \to \infty$. Consider the solution $(\boldsymbol{v}, q) \in \boldsymbol{V}_{h_m}^{\rm div} \times W_{h_m}$ of (10.4) with $f = -\Delta\varphi$. Since $(\chi_i)_{i\in I}$ is a basis of W_{h_m}, there is a unique $u \in X_{\mathcal{D}_m,0}$ such that $q = \sum_{i\in I} u_i \chi_i = \Pi_{\mathcal{D}_m} u$. Comparing (10.4b) and (10.7d) then shows that $\boldsymbol{v} = -\nabla_{\mathcal{D}_m} u$. Since φ is the solution to (10.1) for the chosen f, the error estimate (10.6) leads to

$$\|\Pi_{\mathcal{D}_m} u - \varphi\|_{L^2(\Omega)} + \|-\nabla_{\mathcal{D}_m} u + \nabla\varphi\|_{H_{\rm div}(\Omega)} \\ \leq C_1 \left(\inf_{\psi \in W_{h_m}} \|\psi - \varphi\|_{L^2(\Omega)} + \inf_{\boldsymbol{w} \in \boldsymbol{V}_{h_m}^{\rm div}} \|\boldsymbol{w} - \nabla\varphi\|_{H_{\rm div}(\Omega)} \right).$$

The hypotheses on $\boldsymbol{V}_{h_m}^{\rm div}$ and W_{h_m} ensure that the right hand side of the above inequality tends to 0 as $m \to \infty$, which proves that $S_{\mathcal{D}_m}(\varphi)$ tends to 0 as $m \to \infty$.

Limit-conformity. Take a sequence $(u_m)_{m\in\mathbb{N}}$ such that $u_m \in X_{\mathcal{D}_m,0}$ for all $m \in \mathbb{N}$, and $(\nabla_{\mathcal{D}_m} u_m)_{m\in\mathbb{N}}$ is bounded in $L^2(\Omega)^d$. Let $\boldsymbol{\varphi} \in H_{\rm div}(\Omega)$, and take $\boldsymbol{\varphi}_m \in \boldsymbol{V}_{h_m}^{\rm div}$ such that $\|\boldsymbol{\varphi} - \boldsymbol{\varphi}_m\|_{H_{\rm div}(\Omega)} \to 0$ as $m \to \infty$. Then, recalling the definition (2.9) of $\widetilde{W}_{\mathcal{D}}$ and using (10.7d) with $\boldsymbol{w} = \boldsymbol{\varphi}_m$,

$$\widetilde{W}_{\mathcal{D}_m}(\boldsymbol{\varphi}, u_m) = \int_\Omega (\nabla_{\mathcal{D}_m} u_m(\boldsymbol{x}) \cdot \boldsymbol{\varphi}(\boldsymbol{x}) + \Pi_{\mathcal{D}_m} u_m(\boldsymbol{x}) {\rm div}\boldsymbol{\varphi}(\boldsymbol{x}))\, {\rm d}\boldsymbol{x} = \\ \int_\Omega \Big(\nabla_{\mathcal{D}_m} u_m(\boldsymbol{x}) \cdot (\boldsymbol{\varphi}(\boldsymbol{x}) - \boldsymbol{\varphi}_m(\boldsymbol{x})) + \Pi_{\mathcal{D}_m} u_m(\boldsymbol{x})({\rm div}\boldsymbol{\varphi}(\boldsymbol{x}) - {\rm div}\boldsymbol{\varphi}_m(\boldsymbol{x})) \Big) {\rm d}\boldsymbol{x}.$$

Apply the Cauchy–Schwarz inequality and the coercivity estimate (10.9) to deduce

$$|\widetilde{W}_{\mathcal{D}_m}(\boldsymbol{\varphi}, u_m)| \leq \|\boldsymbol{\varphi} - \boldsymbol{\varphi}_m\|_{H_{\rm div}(\Omega)} \left(1 + \frac{1}{\beta}\right) \|\nabla_{\mathcal{D}_m} u_m\|_{L^2(\Omega)^d}.$$

The boundedness of $(\|\nabla_{\mathcal{D}_m} u_m\|_{L^2(\Omega)^d})_{m\in\mathbb{N}}$ and the choice of $(\boldsymbol{\varphi}_m)_{m\in\mathbb{N}}$ conclude the proof that (2.10) holds, that is $\widetilde{W}_{\mathcal{D}_m}(\boldsymbol{\varphi}, u_m) \to 0$ as $m \to \infty$, which shows the limit-conformity by Lemma 2.7. ∎

10.2 The particular case of the $\mathbb{RT}_k$ mixed finite element

10.2.1 $\mathbb{RT}_k$ mixed finite element gradient discretisation

Take $k \in \mathbb{N}$ and $\mathfrak{T} = (\mathcal{M}, \mathcal{F}, \mathcal{P}, \mathcal{V})$ a conforming simplicial mesh in the sense of Definition 7.4, with $\mathcal{P}$ the set of centres of mass of the cells, and let

$$\boldsymbol{V}_h = \{\boldsymbol{v} \in (L^2(\Omega))^d \,:\, \boldsymbol{v}_{|K} \in \mathbb{RT}_k(K)\,,\ \forall K \in \mathcal{M}\}, \tag{10.10a}$$

$$\boldsymbol{V}_h^{\rm div} = \boldsymbol{V}_h \cap H_{\rm div}(\Omega), \tag{10.10b}$$

$$W_h = \{p \in L^2(\Omega) \,:\, p_{|K} \in \mathbb{P}_k(K)\,,\ \forall K \in \mathcal{M}\}, \tag{10.10c}$$

$$M_h = \left\{ \mu : \bigcup_{\sigma\in\mathcal{F}} \sigma \to \mathbb{R} \,:\, \mu_{|\sigma} \in \mathbb{P}_k(\sigma)\,,\ \forall \sigma \in \mathcal{F} \right\}, \tag{10.10d}$$

$$M_h^0 = \left\{ \mu \in M_h :\ \mu_{|\partial\Omega} = 0 \right\}, \tag{10.10e}$$

where

- $\mathbb{P}_k(K)$ is the space of d-variate polynomials on K having degree less than or equal to k,
- $\mathbb{P}_k(\sigma)$ is the space of $(d-1)$-variate polynomials on σ having degree less than or equal to k,
- $\mathbb{RT}_k(K) = \mathbb{P}_k(K)^d + \boldsymbol{x}\mathbb{P}_k(K)$ is the Raviart–Thomas space on K of order k.

Let us recall some results on $\mathbb{RT}_k$ mixed finite element schemes. The broken Sobolev space $H^1(\mathcal{M})$ is the set of functions whose restriction to each simplex K of the mesh belongs to $H^1(K)$, and $\boldsymbol{H}_{\mathcal{M}} := H_{\mathrm{div}}(\Omega) \cap (H^1(\mathcal{M}))^d$. With $(\boldsymbol{V}_h^{\mathrm{div}}, W_h)$ defined by (10.10b)–(10.10c), the following relations uniquely defines an interpolation operator $P_k : \boldsymbol{H}_{\mathcal{M}} \to \boldsymbol{V}_h^{\mathrm{div}}$ [81, Lemma 3.2]: for all $\boldsymbol{v} \in \boldsymbol{H}_{\mathcal{M}}$ and all $K \in \mathcal{M}$,

$$\begin{aligned}
&\forall \sigma \in \mathcal{F}_K,\ \forall \mu \in \mathbb{P}_k(\sigma),\\
&\qquad \int_\sigma \mu(\boldsymbol{x})\boldsymbol{v}_{|K}(\boldsymbol{x}) \cdot \boldsymbol{n}_{K,\sigma} \mathrm{d}\gamma(\boldsymbol{x}) = \int_\sigma \mu(\boldsymbol{x})(P_k\boldsymbol{v})_{|K}(\boldsymbol{x}) \cdot \boldsymbol{n}_{K,\sigma} \mathrm{d}\gamma(\boldsymbol{x}),\\
&\text{If } k \geq 1,\ \forall q \in \mathbb{P}_{k-1}(K),\\
&\qquad \int_K q(\boldsymbol{x})\boldsymbol{v}_{|K}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = \int_K q(\boldsymbol{x})(P_k\boldsymbol{v})_{|K}(\boldsymbol{x})\mathrm{d}\boldsymbol{x}.
\end{aligned} \tag{10.11}$$

Then P_k satisfies the following properties [81, Lemma 3.5 and Theorem 3.1]:

$$\forall p \in W_h,\ \forall \boldsymbol{v} \in \boldsymbol{H}_{\mathcal{M}},\ \int_\Omega p(\boldsymbol{x})\mathrm{div}(\boldsymbol{v} - P_k\boldsymbol{v})(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = 0 \tag{10.12}$$

and

$$\forall \boldsymbol{v} \in \boldsymbol{H}_{\mathcal{M}},\ \|\boldsymbol{v} - P_k\boldsymbol{v}\|_{L^2(\Omega)^d} \leq \alpha h_{\mathcal{M}} \left(\sum_{K \in \mathcal{M}} \|\boldsymbol{v}\|^2_{H^1(K)} \right)^{1/2}, \tag{10.13}$$

where $\alpha > 0$ depends only on an upper bound of $\kappa_{\mathfrak{T}}$ (see (7.10)).
The "inf-sup" condition (10.3), with β_h depending only on an upper bound of $\kappa_{\mathfrak{T}}$, can be deduced from this property. Let $p \in W_h$. Extend p by 0 outside Ω to a ball B with radius R containing Ω. Then there exists $w \in H_0^1(B)$ solution to $-\Delta w = p$, that is,

$$\forall q \in H_0^1(B), \int_B \nabla w(\boldsymbol{x}) \cdot \nabla q(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = \int_B p(\boldsymbol{x})q(\boldsymbol{x})\mathrm{d}\boldsymbol{x}. \tag{10.14}$$

Moreover, by [33, Theorem 9.25], $w \in H^2(B)$ and, for some $B > 0$ depending only on d and R,

$$\|w\|_{H^2(B)} \leq B\|p\|_{L^2(\Omega)}. \tag{10.15}$$

Therefore, since $\nabla w \in \boldsymbol{H}_{\mathcal{M}}$, Estimate (10.13) yields

$$\|\nabla w - P_k \nabla w\|_{L^2(\Omega)^d} \leq \alpha B h_{\mathcal{M}} \|p\|_{L^2(\Omega)}.$$

Since $h_{\mathcal{M}} \leq \mathrm{diam}(\Omega) \leq 2R$, this shows that

$$\|P_k \nabla w\|_{L^2(\Omega)^d} \leq (2R\alpha + 1)B\|p\|_{L^2(\Omega)}. \tag{10.16}$$

From the definitions (10.10a)–(10.10c) it is easily checked that $\mathrm{div}(P_k \nabla w) \in W_h$. Hence, writing (10.12) with $\boldsymbol{v} = \nabla w$ and $\mathrm{div}(P_k \nabla w)$ instead of p we find

$$\begin{aligned}
\|\mathrm{div}(P_k \nabla w)\|^2_{L^2(\Omega)} &= \int_\Omega \mathrm{div}(P_k \nabla w)(\boldsymbol{x})\mathrm{div}(\nabla w)(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \\
&\leq \|\mathrm{div}(P_k \nabla w)\|_{L^2(\Omega)} \|\Delta w\|_{L^2(\Omega)} \\
&= \|\mathrm{div}(P_k \nabla w)\|_{L^2(\Omega)} \|p\|_{L^2(\Omega)}.
\end{aligned}$$

Combined with (10.16), this shows that

$$\|P_k \nabla w\|_{H_{\mathrm{div}}(\Omega)} \leq [(2R\alpha + 1)B + 1] \|p\|_{L^2(\Omega)}.$$

The inf-sup condition follows by writing, thanks to (10.12) and since $-P_k \nabla w \in \boldsymbol{V}_h^{\mathrm{div}}$,

$$\begin{aligned}
\sup_{\boldsymbol{v} \in \boldsymbol{V}_h^{\mathrm{div}}} \frac{\displaystyle\int_\Omega p(\boldsymbol{x})\mathrm{div}\boldsymbol{v}(\boldsymbol{x})\mathrm{d}\boldsymbol{x}}{\|\boldsymbol{v}\|_{H_{\mathrm{div}}(\Omega)}} &\geq \frac{\displaystyle -\int_\Omega p(\boldsymbol{x})\mathrm{div}(P_k \nabla w)(\boldsymbol{x})\mathrm{d}\boldsymbol{x}}{\|P_k \nabla w\|_{H_{\mathrm{div}}(\Omega)}} \\
&\geq \frac{\displaystyle -\int_\Omega p(\boldsymbol{x})\mathrm{div}(\nabla w)(\boldsymbol{x})\mathrm{d}\boldsymbol{x}}{\|P_k \nabla w\|_{H_{\mathrm{div}}(\Omega)}} \\
&\geq \frac{1}{(2R\alpha + 1)B + 1} \frac{\int_\Omega p(\boldsymbol{x})^2 \mathrm{d}\boldsymbol{x}}{\|p\|_{L^2(\Omega)}} = \frac{1}{(2R\alpha + 1)B + 1} \|p\|_{L^2(\Omega)}.
\end{aligned} \tag{10.17}$$

This shows that the inf-sup condition (10.3) holds with $\beta = \frac{1}{(2R\alpha+1)B+1}$.

The coercivity, GD-consistency and limit-conformity of sequences of $\mathbb{RT}_k$ MFEGDs follow from the generic result of Theorem 10.3, provided that the last two assumptions of this theorem are checked. Here a different approach is chosen: $\mathbb{RT}_k$ MFEGDs can directly shown to be controlled by polytopal toolboxes (Definition 7.9) – yielding the coercivity, limit-conformity and compactness properties – and by establishing high order estimates on $S_{\mathcal{D}}$ from which the GD-consistency can be deduced.
Let us start with a preliminary lemma establishing the existence of the hybrid variables for the $\mathbb{RT}_k$ MFE.

Lemma 10.4 (Hybridisation of the $\mathbb{RT}_k$ MFE). *Let $\mathcal{D}$ be the $\mathbb{RT}_k$ MFEGD, in the sense of* (10.7) *with the spaces* (10.10)*. There exists a unique linear mapping $\Gamma_{\mathcal{D}} : X_{\mathcal{D},0} \to M_h^0$ such that*

$$\forall u \in X_{\mathcal{D},0},\ \forall K \in \mathcal{M},\ \forall \boldsymbol{w} \in \mathbb{RT}_k(K),$$
$$\int_K \boldsymbol{w}(\boldsymbol{x}) \cdot \nabla_{\mathcal{D}} u(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} + \int_K \Pi_{\mathcal{D}} u(\boldsymbol{x}) \mathrm{div} \boldsymbol{w}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} = \sum_{\sigma \in \mathcal{F}_K} \int_\sigma \Gamma_{\mathcal{D}} u(\boldsymbol{x}) \boldsymbol{w}(\boldsymbol{x}) \cdot \boldsymbol{n}_{K,\sigma} \mathrm{d}\gamma(\boldsymbol{x}). \tag{10.18}$$

Proof.
Fix $u \in X_{\mathcal{D},0}$ and let $K \in \mathcal{M}$. Denote by $\mathbb{P}_k(\partial K) = \prod_{\sigma \in \mathcal{F}_K} \mathbb{P}_k(\sigma)$ the space of families $\boldsymbol{q} = (q_\sigma)_{\sigma \in \mathcal{F}_K}$ with $q_\sigma \in \mathbb{P}_k(\sigma)$ for all $\sigma \in \mathcal{F}_K$.
As a consequence of [81, Lemma 3.2], for any $\boldsymbol{q} \in \mathbb{P}_k(\partial K)$, there is a unique $\omega(\boldsymbol{q}) \in \mathbb{RT}_k(K)$ such that

$$\omega(\boldsymbol{q})_{|\sigma} \cdot \boldsymbol{n}_{K,\sigma} = q_\sigma \text{ for all } \sigma \in \mathcal{F}_K, \text{ and}$$
$$\text{the } L^2(K)^d\text{-projection of } \omega(\boldsymbol{q}) \text{ on } \mathbb{P}_{k-1}(K) \text{ vanishes}$$

(if $k = 0$, the second condition is non-existent). This defines a linear map $\omega : \mathbb{P}_k(\partial K) \to \mathbb{RT}_k(K)$. The sum of the L^2-inner product on each face defines an inner product on the finite dimensional space $\mathbb{P}_k(\partial K)$. Since ω is linear, there is therefore a unique $\Gamma_K u \in \mathbb{P}_k(\partial K)$ such that

$$\forall \boldsymbol{q} \in \mathbb{P}_k(\partial K)\,,\ \int_K \omega(\boldsymbol{q})(\boldsymbol{x}) \cdot \nabla_{\mathcal{D}} u(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} + \int_K \Pi_{\mathcal{D}} u(\boldsymbol{x}) \mathrm{div} \omega(\boldsymbol{q})(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} = \sum_{\sigma \in \mathcal{F}_K} \int_\sigma \Gamma_K u(\boldsymbol{x}) q_\sigma(\boldsymbol{x}) \mathrm{d}\gamma(\boldsymbol{x}) \tag{10.19}$$

(the left-hand side is the inner product of $\Gamma_K u$ and $\boldsymbol{q}$). Consider now a generic $\boldsymbol{w} \in \mathbb{RT}_k(K)$ and let $q_\sigma = \boldsymbol{w}_{|\sigma} \cdot \boldsymbol{n}_{K,\sigma}$ for all $\sigma \in \mathcal{F}_K$. Since $\boldsymbol{x} \in \sigma \mapsto \boldsymbol{x} \cdot \boldsymbol{n}_{K,\sigma}$ is constant for all $\sigma \in \mathcal{F}_K$, it can easily be checked that $q_\sigma \in \mathbb{P}_k(\sigma)$ for all $\sigma \in \mathcal{F}_K$ (see [81, Lemma 3.1]). This defines therefore $\boldsymbol{q} = (q_\sigma)_{\sigma \in \mathcal{F}_K} \in \mathbb{P}_k(\partial K)$. The function $\boldsymbol{w} - \omega(\boldsymbol{q}) \in \mathbb{RT}_k(K)$ has a zero normal trace on ∂K, and its extension to Ω by 0 outside K therefore belongs to $\boldsymbol{V}_h^{\mathrm{div}}$ (the normal traces at the interfaces between K and its neighbouring cells are continuous). This function can therefore be used in (10.7d), which shows that

$$\int_K (\boldsymbol{w} - \omega(\boldsymbol{q}))(\boldsymbol{x}) \cdot \nabla_{\mathcal{D}} u(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} + \int_K \Pi_{\mathcal{D}} u(\boldsymbol{x}) \mathrm{div}(\boldsymbol{w} - \omega(\boldsymbol{q}))(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} = 0.$$

Owing to (10.19) and recalling the definition of $(q_\sigma)_{\sigma \in \mathcal{F}_K}$, this yields

$$\int_K \boldsymbol{w}(\boldsymbol{x}) \cdot \nabla_{\mathcal{D}} u(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} + \int_K \Pi_{\mathcal{D}} u(\boldsymbol{x}) \mathrm{div} \boldsymbol{w}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} = \sum_{\sigma \in \mathcal{F}_K} \int_\sigma \Gamma_K u(\boldsymbol{x}) \boldsymbol{w}(\boldsymbol{x}) \cdot \boldsymbol{n}_{K,\sigma} \mathrm{d}\gamma(\boldsymbol{x}). \tag{10.20}$$

We have thus defined a cell-dependent mapping $\Gamma_K : X_{\mathcal{D},0} \to \mathbb{P}_k(\partial K)$ that satisfies (10.18) on K. Take $\sigma \in \mathcal{F}_K \cap \mathcal{F}_{\mathrm{ext}}$ and $q_\sigma \in \mathbb{P}_k(\sigma)$, and consider

$\boldsymbol{w} \in \boldsymbol{V}_h$ such that $\boldsymbol{w}_{|\sigma} \cdot \boldsymbol{n}_{K,\sigma} = q_\sigma$ and all the normal traces of $\boldsymbol{w}$ on the other faces vanish; this defines $\boldsymbol{w} \in \boldsymbol{V}_h^{\mathrm{div}}$ (since the normal traces across the interfaces are continuous) and, used in (10.7d) and combined with (10.20), this shows that

$$\int_\sigma \Gamma_K u(\boldsymbol{x}) q_\sigma(\boldsymbol{x}) \mathrm{d}\gamma(\boldsymbol{x}) = 0.$$

Since this is true for any $q_\sigma \in \mathbb{P}_k(\sigma)$ and since $\Gamma_K u \in \mathbb{P}_k(\sigma)$, this shows that $\Gamma_K u = 0$ on σ, whenever $\sigma \in \mathcal{F}_K \cap \mathcal{F}_{\mathrm{ext}}$. If we prove that

$$\forall \sigma \in \mathcal{F}_{\mathrm{int}} \text{ with } \mathcal{M}_\sigma = \{K, L\}\,,\ (\Gamma_K u)_{|\sigma} = (\Gamma_L u)_{|\sigma}, \tag{10.21}$$

then setting $(\Gamma_{\mathcal{D}} u)_{|\sigma} = (\Gamma_K u)_{|\sigma}$ for any $\sigma \in \mathcal{F}$ and any $K \in \mathcal{M}_\sigma$ provides a mapping $\Gamma_{\mathcal{D}} : X_{\mathcal{D},0} \to M_h^0$ that satisfies (10.18). The uniqueness of this mapping follows from the uniqueness of each Γ_K.
Let σ as in (10.21) and let $q_\sigma \in \mathbb{P}_k(\sigma)$. Take $\boldsymbol{w} \in \boldsymbol{V}_h^{\mathrm{div}}$ such that $\boldsymbol{w}_{|\sigma} \cdot \boldsymbol{n}_{K,\sigma} = -\boldsymbol{w}_{|\sigma} \cdot \boldsymbol{n}_{L,\sigma} = q_\sigma$, and the normal traces of $\boldsymbol{w}$ on all the other faces vanish. Writing (10.20) with this $\boldsymbol{w}$ on all the cells and summing over the cells, the left-hand side vanishes thanks to (10.7d), and only the contributions from K and L on σ remain on the right-hand side, leading to

$$\int_\sigma \Gamma_K u(\boldsymbol{x}) \boldsymbol{w}(\boldsymbol{x}) \cdot \boldsymbol{n}_{K,\sigma} \mathrm{d}\gamma(\boldsymbol{x}) + \int_\sigma \Gamma_L u(\boldsymbol{x}) \boldsymbol{w}(\boldsymbol{x}) \cdot \boldsymbol{n}_{L,\sigma} \mathrm{d}\gamma(\boldsymbol{x}) = 0.$$

This proves that

$$\int_\sigma \Gamma_K u(\boldsymbol{x}) q_\sigma(\boldsymbol{x}) \mathrm{d}\gamma(\boldsymbol{x}) - \int_\sigma \Gamma_L u(\boldsymbol{x}) q_\sigma(\boldsymbol{x}) \mathrm{d}\gamma(\boldsymbol{x}) = 0$$

and thus, since q_σ was arbitrary in $\mathbb{P}_k(\sigma)$, that (10.21) is satisfied. ■

Lemma 10.5 (Control of the $\mathbb{RT}_k$ MFEGD by a polytopal toolbox). *Let $\mathcal{D}$ be the $\mathbb{RT}_k$ MFEGD, in the sense of (10.7), using the spaces defined by (10.10), and let $\varrho \geq \kappa_{\mathfrak{T}}$ (see (7.10)). Let $\Gamma_{\mathcal{D}}$ be given by Lemma 10.4 and define the control $\boldsymbol{\Phi} : X_{\mathcal{D},0} \to X_{\mathfrak{T},0}$ of $\mathcal{D}$ by $\mathfrak{T}$ (see Definition 7.9) by, for all $u \in X_{\mathcal{D},0}$,*

$$\begin{aligned} &\forall \sigma \in \mathcal{F},\ \boldsymbol{\Phi}(u)_\sigma = \frac{1}{|\sigma|} \int_\sigma \Gamma_{\mathcal{D}} u(\boldsymbol{y}) \mathrm{d}\gamma(\boldsymbol{y}), \\ &\forall K \in \mathcal{M}\,,\ \boldsymbol{\Phi}(u)_K = \frac{1}{|K|} \int_K \Pi_{\mathcal{D}} u(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}. \end{aligned} \tag{10.22}$$

Then there exists $C_2 > 0$, depending only on Ω, k, and ϱ, such that

$$\|\boldsymbol{\Phi}\|_{\mathcal{D},\mathfrak{T}} \leq C_2, \tag{10.23}$$

$$\omega^{\Pi}(\mathcal{D}, \mathfrak{T}, \boldsymbol{\Phi}) \leq C_2 h_{\mathcal{M}}, \tag{10.24}$$

and

$$\omega^{\nabla}(\mathcal{D}, \mathfrak{T}, \boldsymbol{\Phi}) = 0. \tag{10.25}$$

Proof.
Step 1: estimate (10.23).
Let $K \in \mathcal{M}$, $\sigma \in \mathcal{F}_K$ and $\boldsymbol{w}_{K,\sigma} \in \mathbb{RT}_0(K) \subset \mathbb{RT}_k(K)$ be defined by $\boldsymbol{w}_{K,\sigma} \cdot \boldsymbol{n}_{K,\sigma} = 1$ on σ, and $\boldsymbol{w}_{K,\sigma} \cdot \boldsymbol{n}_{K,\sigma'} = 0$ on σ' for all $\sigma' \in \mathcal{F}_K \setminus \{\sigma\}$. Then, if $\boldsymbol{s}$ is the vertex opposite to σ in the simplex K, $\boldsymbol{w}_{K,\sigma}$ is given by $\boldsymbol{w}_{K,\sigma}(\boldsymbol{x}) = \frac{\boldsymbol{x}-\boldsymbol{s}}{\mathrm{dist}(\boldsymbol{s},\sigma)}$ for all $\boldsymbol{x} \in K$, where $\mathrm{dist}(\boldsymbol{s},\sigma)$ is the orthogonal distance between $\boldsymbol{s}$ and σ (that is, the distance between $\boldsymbol{s}$ and the hyperplane generated by σ). Hence, $\mathrm{div}(\boldsymbol{w}_{K,\sigma}) = \frac{d}{\mathrm{dist}(\boldsymbol{s},\sigma)} = \frac{|\sigma|}{|K|}$. Taking $\boldsymbol{w}_{K,\sigma}$ as a test function in (10.18) therefore yields

$$\int_K \boldsymbol{w}_{K,\sigma}(\boldsymbol{x}) \cdot \nabla_{\mathcal{D}} u(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} + |\sigma|(\boldsymbol{\Phi}(u)_K - \boldsymbol{\Phi}(u)_\sigma) = 0,$$

which implies

$$\frac{|\sigma|}{d_{K,\sigma}}(\boldsymbol{\Phi}(u)_\sigma - \boldsymbol{\Phi}(u)_K)^2 \leq \frac{1}{d_{K,\sigma}|\sigma|} \|\nabla_{\mathcal{D}} u\|^2_{L^2(K)^d} \|\boldsymbol{w}_{K,\sigma}\|^2_{L^2(K)^d}.$$

The expression of $\boldsymbol{w}_{K,\sigma}$ yields the existence of $C_3 > 0$, depending only on ϱ, such that, for all $\boldsymbol{x} \in K$, $|\boldsymbol{w}_{K,\sigma}(\boldsymbol{x})| \leq C_3$. Moreover, using Lemma B.4 and recalling that $\boldsymbol{x}_K$ is the centre of mass of K, we can see that $\frac{|K|}{|\sigma| d_{K,\sigma}} \leq C_4$ with C_4 depending only on ϱ and d. Hence, since $\mathrm{Card}(\mathcal{F}_K) = d+1$ for all $K \in \mathcal{M}$,

$$\sum_{K\in\mathcal{M}} \sum_{\sigma\in\mathcal{F}_K} \frac{|\sigma|}{d_{K,\sigma}}(\boldsymbol{\Phi}(u)_\sigma - \boldsymbol{\Phi}(u)_K)^2 \leq C_3^2 C_4 (d+1) \|\nabla_{\mathcal{D}} u\|^2_{L^2(\Omega)^d}$$

and (10.23) follows by noticing that the left-hand side of this inequality is $|\boldsymbol{\Phi}(u)|^2_{\mathfrak{T},2}$.

Step 2: Estimate (10.24).
Let $w \in H^1_0(B) \cap H^2(B)$ be defined by (10.14) for p the extension of $\Pi_{\mathcal{D}} u - \Pi_{\mathfrak{T}} \boldsymbol{\Phi}(u)$ to B by 0 outside Ω. Applying (10.15),

$$\|\nabla w\|_{H^1(\Omega)^d} \leq B \|\Pi_{\mathcal{D}} u - \Pi_{\mathfrak{T}} \boldsymbol{\Phi}(u)\|_{L^2(\Omega)}. \tag{10.26}$$

Moreover, from (10.14) with $q = w$,

$$\int_B |\nabla w(\boldsymbol{x})|^2 \mathrm{d}\boldsymbol{x} = \sum_{K\in\mathcal{M}} \int_K (\Pi_{\mathcal{D}} u(\boldsymbol{x}) - \boldsymbol{\Phi}(u)_K) w(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}.$$

Denote by $\overline{w}_K$ the average value of w on K. Since $\int_K (\Pi_{\mathcal{D}} u(\boldsymbol{x}) - \boldsymbol{\Phi}(u)_K) \mathrm{d}\boldsymbol{x} = 0$, we infer

$$\int_B |\nabla w(\boldsymbol{x})|^2 \mathrm{d}\boldsymbol{x} = \sum_{K\in\mathcal{M}} \int_K (\Pi_{\mathcal{D}} u(\boldsymbol{x}) - \boldsymbol{\Phi}(u)_K)(w(\boldsymbol{x}) - \overline{w}_K) \mathrm{d}\boldsymbol{x}.$$

The Hölder inequalities (D.5) and (D.1) (with $p = p' = 2$), and (B.12) in Lemma B.6, show that

$$\begin{aligned}\int_B |\nabla w(\boldsymbol{x})|^2 \mathrm{d}\boldsymbol{x} &\\ &\leq \sum_{K\in\mathcal{M}} \|\Pi_{\mathcal{D}} u - \boldsymbol{\Phi}(u)_K\|_{L^2(K)} \|w - \overline{w}_K\|_{L^2(K)} \\ &\leq \sum_{K\in\mathcal{M}} \|\Pi_{\mathcal{D}} u - \Pi_{\mathfrak{T}}\boldsymbol{\Phi}(u)\|_{L^2(K)} C_5 h_K \|\nabla w\|_{L^2(K)^d} \\ &\leq C_5 h_{\mathcal{M}} \left(\sum_{K\in\mathcal{M}} \|\Pi_{\mathcal{D}} u - \Pi_{\mathfrak{T}}\boldsymbol{\Phi}(u)\|^2_{L^2(K)}\right)^{\frac{1}{2}} \left(\sum_{K\in\mathcal{M}} \|\nabla w\|^2_{L^2(K)^d}\right)^{\frac{1}{2}}\end{aligned}$$

with C_5 depending only on d and ϱ (the assumption of Lemma B.6 is checked by using Lemma B.4 to write $\max_{\sigma\in\mathcal{F}_K} \frac{h_K}{d_{K,\sigma}} \leq \theta_{\mathfrak{T}} \leq \kappa_{\mathfrak{T}} + d + 1 \leq \varrho + d + 1$). As a consequence,

$$\|\nabla w\|_{L^2(\Omega)^d} \leq C_5 h_{\mathcal{M}} \|\Pi_{\mathcal{D}} u - \Pi_{\mathfrak{T}}\boldsymbol{\Phi}(u)\|_{L^2(\Omega)}. \tag{10.27}$$

Set $\boldsymbol{w} = -P_k \nabla w \in \boldsymbol{V}_h^{\mathrm{div}}$ (P_k is defined by (10.11)). Thanks to (10.12) and (10.13),

$$\int_\Omega \Pi_{\mathcal{D}} u(\boldsymbol{x}) \mathrm{div}(\boldsymbol{w} + \nabla w)(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} = 0 \tag{10.28}$$

and

$$\|\nabla w + \boldsymbol{w}\|_{L^2(\Omega)^d} \leq \alpha h_{\mathcal{M}} \|\nabla w\|_{H^1(\Omega)^d}.$$

Combined with (10.26) this yields

$$\|\nabla w + \boldsymbol{w}\|_{L^2(\Omega)^d} \leq \alpha h_{\mathcal{M}} B \|\Pi_{\mathcal{D}} u - \Pi_{\mathfrak{T}}\boldsymbol{\Phi}(u)\|_{L^2(\Omega)},$$

and therefore, with (10.27),

$$\|\boldsymbol{w}\|_{L^2(\Omega)^d} \leq h_{\mathcal{M}}(\alpha B + C_5) \|\Pi_{\mathcal{D}} u - \Pi_{\mathfrak{T}}\boldsymbol{\Phi}(u)\|_{L^2(\Omega)}. \tag{10.29}$$

We now consider $\boldsymbol{w}$ in (10.7d). Using (10.28) and $-\mathrm{div}(\nabla w) = \Pi_{\mathcal{D}} u - \Pi_{\mathfrak{T}}\boldsymbol{\Phi}(u)$, this leads to

$$\int_\Omega \boldsymbol{w}(\boldsymbol{x}) \cdot \nabla_{\mathcal{D}} u(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} + \int_\Omega \Pi_{\mathcal{D}} u(\boldsymbol{x})(\Pi_{\mathcal{D}} u - \Pi_{\mathfrak{T}}\boldsymbol{\Phi}(u))(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} = 0. \tag{10.30}$$

The function $\Pi_{\mathfrak{T}}\boldsymbol{\Phi}(u)$ is the L^2-orthogonal projection of $\Pi_{\mathcal{D}} u$ on piecewise constant functions on $\mathcal{M}$, and thus

$$\int_\Omega \Pi_{\mathcal{D}} u(\boldsymbol{x})(\Pi_{\mathcal{D}} u - \Pi_{\mathfrak{T}}\boldsymbol{\Phi}(u))(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} = \int_\Omega (\Pi_{\mathcal{D}} u - \Pi_{\mathfrak{T}}\boldsymbol{\Phi}(u))^2(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}.$$

Used in (10.30) and owing to (10.29), this gives

$$\begin{aligned}\|\Pi_{\mathcal{D}}u - \Pi_{\mathfrak{T}}\boldsymbol{\Phi}(u)\|^2_{L^2(\Omega)} &\le \|\boldsymbol{w}\|_{L^2(\Omega)^d}\|\nabla_{\mathcal{D}}u\|_{L^2(\Omega)^d}\\ &\le h_{\mathcal{M}}(\alpha B + C_5)\|\Pi_{\mathcal{D}}u - \Pi_{\mathfrak{T}}\boldsymbol{\Phi}(u)\|_{L^2(\Omega)}\|\nabla_{\mathcal{D}}u\|_{L^2(\Omega)^d}.\end{aligned}$$

Hence, $\|\Pi_{\mathcal{D}}u - \Pi_{\mathfrak{T}}\boldsymbol{\Phi}(u)\|_{L^2(\Omega)} \le h_{\mathcal{M}}(\alpha B + C_5)\|\nabla_{\mathcal{D}}u\|_{L^2(\Omega)^d}$ and (10.24) is proved.

Step 3: Relation (10.25).
Let $u \in X_{\mathcal{D},0}$, $K \in \mathcal{M}$, $\boldsymbol{\xi} \in \mathbb{R}^d$, and $\boldsymbol{w} \in \mathbb{RT}_k(K)$ be such that $\boldsymbol{w}(\boldsymbol{x}) = \xi$ for all $\boldsymbol{x} \in K$. Using this function in (10.18) yields

$$\int_K \boldsymbol{\xi}\cdot\nabla_{\mathcal{D}}u(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = \sum_{\sigma\in\mathcal{F}_K}\int_\sigma \Gamma_{\mathcal{D}}u(\boldsymbol{x})\boldsymbol{\xi}\cdot\boldsymbol{n}_{K,\sigma}\mathrm{d}\gamma(\boldsymbol{x}) = \sum_{\sigma\in\mathcal{F}_K}|\sigma|\boldsymbol{\Phi}(u)_\sigma\boldsymbol{\xi}\cdot\boldsymbol{n}_{K,\sigma}.$$

Since $\boldsymbol{\xi}$ is arbitrary, this proves that

$$\int_K \nabla_{\mathcal{D}}u(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = \sum_{\sigma\in\mathcal{F}_K}|\sigma|\boldsymbol{\Phi}(u)_\sigma\boldsymbol{n}_{K,\sigma} = \int_K \overline{\nabla}_{\mathfrak{T}}\boldsymbol{\Phi}(v)(\boldsymbol{x})\mathrm{d}\boldsymbol{x},$$

and (10.25) is proved. ∎

Before stating the properties of $\mathbb{RT}_k$ MFEGDs, we establish high order estimates on $S_{\mathcal{D}}$ and $W_{\mathcal{D}}$ for regular functions. These estimates are essential to obtain optimal $\mathcal{O}(h_{\mathcal{M}}^{k+1})$ error estimates for $\mathbb{RT}_k$ schemes on linear problems (see Theorem 2.28). The estimate on $S_{\mathcal{D}}$ also helps establishing the GD-consistency of $\mathbb{RT}_k$ MFEGDs.

Lemma 10.6 (Higher order estimates for $\mathbb{RT}_k$ MFEGD). *Let $\mathcal{D}$ be the $\mathbb{RT}_k$ MFEGD defined by (10.7) with the choice of spaces (10.10). Take $\varrho \ge \kappa_{\mathfrak{T}}$ (see (7.10)). There exists $C_6 > 0$, depending only on on Ω, k and ϱ, such that*

$$\forall\varphi \in H^{k+2}(\Omega)\cap H_0^1(\Omega)\,,\ S_{\mathcal{D}}(\varphi) \le C_6 h_{\mathcal{M}}^{k+1}\|\varphi\|_{H^{k+2}(\Omega)}\,, \tag{10.31}$$

and

$$\forall\boldsymbol{\varphi} \in H^{k+1}(\Omega)^d,\ W_{\mathcal{D}}(\boldsymbol{\varphi}) \le C_6 h_{\mathcal{M}}^{k+1}\|\boldsymbol{\varphi}\|_{H^{k+1}(\Omega)^d}\,, \tag{10.32}$$

where $S_{\mathcal{D}}$ is defined by (2.2) and $W_{\mathcal{D}}$ is defined by (2.6). In other words, the space size (see Definition 2.22) of the GD satisfies

$$h_{\mathcal{D}}(H^{k+2}(\Omega)\cap H_0^1(\Omega); H^{k+1}(\Omega)^d) \le C_6 h_{\mathcal{M}}^{k+1}.$$

Proof. As in Theorem 10.3, setting $u \in X_{\mathcal{D},0}$ such that $\Pi_{\mathcal{D}}u = q$ with $(\boldsymbol{v}, q)$ solution to (10.4) for $f = -\Delta\varphi$, we have $\boldsymbol{v} = -\nabla_{\mathcal{D}}u$ and [81, Theorem 3.2] thus gives the existence of C_7 depending only on ϱ and k such that

$$\|-\nabla_{\mathcal{D}}u + \nabla\varphi\|_{L^2(\Omega)^d} \le C_7 h_{\mathcal{M}}^{k+1}\|\varphi\|_{H^{k+2}(\Omega)}\,. \tag{10.33}$$

Apply then [81, Theorem 3.3] to find C_8 depending only on ϱ and k such that

$$\|\Pi_{\mathcal{D}} u - \varphi\|_{L^2(\Omega)} \leq C_8 h_{\mathcal{M}}^{k+1} \|\varphi\|_{H^{k+2}(\Omega)} . \tag{10.34}$$

Estimate (10.31) follows from (10.33) and (10.34).

Let us turn to the estimate on $W_{\mathcal{D}}$. Let $\boldsymbol{\varphi} \in H^{k+1}(\Omega)^d$. Thanks to Property (10.12) of the interpolation operator P_k, we have, for any $v \in X_{\mathcal{D},0}$,

$$\begin{aligned}
&\Big| \int_\Omega (\nabla_{\mathcal{D}} v(\boldsymbol{x}) \cdot \boldsymbol{\varphi}(\boldsymbol{x}) + \Pi_{\mathcal{D}} v(\boldsymbol{x}) \mathrm{div} \boldsymbol{\varphi}(\boldsymbol{x}))\, \mathrm{d}\boldsymbol{x} \Big| \\
&= \Big| \int_\Omega (\nabla_{\mathcal{D}} v(\boldsymbol{x}) \cdot \boldsymbol{\varphi}(\boldsymbol{x}) + \Pi_{\mathcal{D}} v(\boldsymbol{x}) \mathrm{div} P_k \boldsymbol{\varphi}(\boldsymbol{x}))\, \mathrm{d}\boldsymbol{x} \Big| \\
&= \Big| \int_\Omega \nabla_{\mathcal{D}} v(\boldsymbol{x}) \cdot (\boldsymbol{\varphi}(\boldsymbol{x}) - P_k \boldsymbol{\varphi}(\boldsymbol{x}))\, \mathrm{d}\boldsymbol{x} \Big| \\
&\leq \|\boldsymbol{\varphi} - P_k \boldsymbol{\varphi}\|_{L^2(\Omega)^d} \|\nabla_{\mathcal{D}} v\|_{L^2(\Omega)^d} ,
\end{aligned} \tag{10.35}$$

using $P_k \boldsymbol{\varphi} \in \boldsymbol{V}_h^{\mathrm{div}}$ and the definition (10.7d) of $\nabla_{\mathcal{D}}$. The estimate in [81, Theorem 3.1] implies

$$\|\boldsymbol{\varphi} - P_k \boldsymbol{\varphi}\|_{L^2(\Omega)^d} \leq C_9 h_{\mathcal{M}}^{k+1} \|\boldsymbol{\varphi}\|_{H^{k+1}(\Omega)^d}$$

with C_9 depending only on Ω, k and ϱ. Used in (10.35) this proves (10.32). ∎

We can now state and prove the properties of $\mathbb{RT}_k$ MFEGDs.

Theorem 10.7 (Properties of $\mathbb{RT}_k$ MFEGDs). *Let $(\mathfrak{T}_m)_{m\in\mathbb{N}}$ be a sequence of conforming simplicial meshes in the sense of Definition 7.4, such that $h_{\mathcal{M}_m} \to 0$ as $m \to \infty$ and $(\kappa_{\mathfrak{T}_m})_{m\in\mathbb{N}}$ is bounded (see (7.10)). Let, for $m \in \mathbb{N}$, $\mathcal{D}_m = (X_{\mathcal{D}_m,0}, \Pi_{\mathcal{D}_m}, \nabla_{\mathcal{D}_m})$ be the $\mathbb{RT}_k$ MFEGD defined by (10.7) with $(\boldsymbol{V}_h^{\mathrm{div}}, W_h) = (\boldsymbol{V}_{h_m}^{\mathrm{div}}, W_{h_m})$ given by (10.10) with $\mathfrak{T} = \mathfrak{T}_m$.*
Then $(\mathcal{D}_m)_{m\in\mathbb{N}}$ is coercive, GD-consistent, limit-conforming and compact in the sense of the definitions in Section 2.1.1.

Proof. Lemma 10.5 and Corollary 7.12 give the coercivity, limit-conformity and compactness properties. The GD-consistency follows from (10.31) and from Lemma 2.16 with $W_s = H^{k+2}(\Omega) \cap H_0^1(\Omega)$. ∎

10.2.2 Gradient discretisation from the $\mathbb{RT}_k$ hybrid mixed finite element formulation

The hybrid, or Arnold–Brezzi, formulation [16] for the approximation of (10.1) corresponds to an hybridisation of the mixed finite element scheme (10.4). Using the discrete spaces defined by (10.10), it reads

$$(\boldsymbol{v}, q, \lambda) \in \boldsymbol{V}_h \times W_h \times M_h^0, \tag{10.36a}$$

$$\int_K \boldsymbol{w}(\boldsymbol{x})\cdot\boldsymbol{v}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} - \int_K q(\boldsymbol{x})\mathrm{div}\boldsymbol{w}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} + \sum_{\sigma\in\mathcal{F}_K}\int_\sigma \lambda(\boldsymbol{x})\ \boldsymbol{w}_{|K}(\boldsymbol{x})\cdot\boldsymbol{n}_{K,\sigma}\mathrm{d}\gamma(\boldsymbol{x}) = 0, \quad \forall \boldsymbol{w}\in\boldsymbol{V}_h\,,\ \forall K\in\mathcal{M} \tag{10.36b}$$

$$\int_K \psi(\boldsymbol{x})\mathrm{div}\boldsymbol{v}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = \int_K \psi(\boldsymbol{x})f(\boldsymbol{x})\mathrm{d}\boldsymbol{x},\ \forall\psi\in W_h\,,\ \forall K\in\mathcal{M}, \tag{10.36c}$$

$$\int_\sigma \mu(\boldsymbol{x})\ \boldsymbol{v}_{|K}(\boldsymbol{x})\cdot\boldsymbol{n}_{K,\sigma}\mathrm{d}\gamma(\boldsymbol{x}) + \int_\sigma \mu(\boldsymbol{x})\ \boldsymbol{v}_{|L}(\boldsymbol{x})\cdot\boldsymbol{n}_{L,\sigma}\mathrm{d}\gamma(\boldsymbol{x}) = 0, \quad \forall\sigma\in\mathcal{F}_{\rm int} \text{ with } \mathcal{M}_\sigma=\{K,L\},\ \forall\mu\in M_h^0. \tag{10.36d}$$

Since the normal traces of $\mathbb{RT}_k$ functions on a face σ belong to $\mathbb{P}_k(\sigma)$, (10.36d) is equivalent to imposing the pointwise continuity of the normal traces of $\boldsymbol{v}$, that is, to imposing $\boldsymbol{v}\in\boldsymbol{V}_h^{\rm div}$. Given that the elements in W_h have independent values in each cell, (10.36c) is then clearly equivalent to (10.4c). Taking $\boldsymbol{w}\in\boldsymbol{V}_h^{\rm div}$ in (10.36b) and summing over the cells yields (10.4b). Conversely, if $(\boldsymbol{v},q)$ satisfies (10.4b) then by introducing $u\in X_{\mathcal{D},0}$ such that $(-\nabla_{\mathcal{D}}u, \Pi_{\mathcal{D}}u) = (\boldsymbol{v},q)$ and by setting $\lambda=\Gamma_{\mathcal{D}}u$, where $\Gamma_{\mathcal{D}}$ is defined in Lemma 10.4, we see that (10.18) is the same equation as (10.36b). All this shows that $(\boldsymbol{v},q)$ is a solution to (10.4) if and only if $(\boldsymbol{v},q,\lambda)$ is a solution to (10.36). As a by-product, this also establishes the existence and uniqueness of a solution to this hybrid formulation.

We now construct a GD (in the sense of Definition 2.1), called HMFEGD, inspired from the hybrid mixed finite element formulation (10.36) of Problem (10.1). Let W_h be defined by (10.10c) and let again $(\chi_i)_{i\in I}$ be a basis of W_h. Let M_h^0 be defined by (10.10d) and let $(\xi_j)_{j\in J}$ be a basis of M_h^0. To avoid confusions in the notations below, the index sets I and J are selected to be disjoint. Recalling that $\boldsymbol{V}_h$ is given by (10.10a), define the HMFEGD $\widetilde{\mathcal{D}} = (X_{\widetilde{\mathcal{D}},0}, \Pi_{\widetilde{\mathcal{D}}}, \nabla_{\widetilde{\mathcal{D}}})$ by:

$$X_{\widetilde{\mathcal{D}},0} = \{v = ((v_i)_{i\in I}, (v_j)_{j\in J})\ :\ v_k\in\mathbb{R} \text{ for all } k\in I\cup J\}, \tag{10.37a}$$

$$\forall u\in X_{\widetilde{\mathcal{D}},0}\,,\ \Pi_{\widetilde{\mathcal{D}}}u = \sum_{i\in I}u_i\chi_i \text{ and } \Gamma_{\widetilde{\mathcal{D}}}u = \sum_{j\in J}u_j\xi_j, \tag{10.37b}$$

$$\forall u\in X_{\widetilde{\mathcal{D}},0}\,,\ \nabla_{\widetilde{\mathcal{D}}}u\in\boldsymbol{V}_h \text{ is such that, for all } K\in\mathcal{M},$$
$$\int_K \boldsymbol{w}(\boldsymbol{x})\cdot\nabla_{\widetilde{\mathcal{D}}}u(\boldsymbol{x})\mathrm{d}\boldsymbol{x} + \int_K \Pi_{\widetilde{\mathcal{D}}}u(\boldsymbol{x})\mathrm{div}\boldsymbol{w}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} - \sum_{\sigma\in\mathcal{F}_K}\int_\sigma \Gamma_{\widetilde{\mathcal{D}}}u(\boldsymbol{x})\ \boldsymbol{w}_{|K}(\boldsymbol{x})\cdot\boldsymbol{n}_{K,\sigma}\mathrm{d}\gamma(\boldsymbol{x}) = 0,\ \forall\boldsymbol{w}\in\boldsymbol{V}_h. \tag{10.37c}$$

Remark 10.8 (Raviart-Thomas mixed finite element and HMM). In the case $k=0$, the hybrid $\mathbb{RT}_0$ scheme is a particular case of the mixed-hybrid mimetic

finite difference [130], which is itself a member of the HMM family of schemes described in Chapter 13. The properties of $\mathbb{RT}_0$ HMFEGDs therefore also follow from the properties of HMM GDs.

As for (10.7), in order for (10.37) to define a GD the system (10.37c) should define one and only one reconstructed gradient $\nabla_{\widetilde{\mathcal{D}}} u$, and $\|\cdot\|_{\widetilde{\mathcal{D}}} := \left\|\nabla_{\widetilde{\mathcal{D}}}\cdot\right\|_{L^2(\Omega)^d}$ should be a norm on $X_{\widetilde{\mathcal{D}},0}$. The existence and uniqueness of $\nabla_{\widetilde{\mathcal{D}}} u$ again results from the Riesz representation theorem (applied in each $\mathbb{RT}_k(K)$ space with the $L^2(K)^d$-inner product). The norm property results, on one hand, from the coercivity property shown in Theorem 10.10 below, and on the other hand, from the reasoning in the proof of Lemma 10.4 (based on the fact that any $\mathbf{q} \in \mathbb{P}_k(\partial K)$ can be written as the normal trace of some $\boldsymbol{w} \in \mathbb{RT}_k(K)$), which shows that, if $\nabla_{\widetilde{\mathcal{D}}} u = 0$ and $\Pi_{\widetilde{\mathcal{D}}} u = 0$, then $\Gamma_{\widetilde{\mathcal{D}}} u$ satisfying (10.37c) must vanish.

Let us check that the GD $\widetilde{\mathcal{D}}$ indeed corresponds to the hybrid $\mathbb{RT}_k$ scheme.

Theorem 10.9 (Hybrid $\mathbb{RT}_k$ is a GDM in the isotropic case). *Using the GD (10.37), u is the solution to the gradient scheme (2.23) for Problem (10.1) if and only if $(\boldsymbol{v}, q, \lambda) = (-\nabla_{\widetilde{\mathcal{D}}} u, \Pi_{\widetilde{\mathcal{D}}} u, \Gamma_{\widetilde{\mathcal{D}}} u)$ is the solution to the Arnold–Brezzi formulation (10.36) of the mixed finite element method.*

Proof. Let $u \in X_{\widetilde{\mathcal{D}},0}$ be a solution to (2.23), and let us show that $(\boldsymbol{v}, q, \lambda) = (-\nabla_{\widetilde{\mathcal{D}}} u, \Pi_{\widetilde{\mathcal{D}}} u, \Gamma_{\widetilde{\mathcal{D}}} u)$ is the solution of (10.36). We first observe that (10.37c) ensures (10.36b). Let $\psi \in W_h$ and $\mu \in M_h^0$, consider a particular $K \in \mathcal{M}$, and take in (2.23) a test function $v \in X_{\widetilde{\mathcal{D}},0}$ such that $\Pi_{\widetilde{\mathcal{D}}} v_{|K} = \psi_{|K}$, $\Pi_{\widetilde{\mathcal{D}}} v_{|L} = 0$ for all $L \in \mathcal{M} \setminus \{K\}$, and $\Gamma_{\widetilde{\mathcal{D}}} v = 0$. Thanks to (10.37c), the support of $\nabla_{\widetilde{\mathcal{D}}} v$ is also reduced to K and the GS (2.23) therefore gives

$$\int_K \nabla_{\widetilde{\mathcal{D}}} u(\boldsymbol{x}) \cdot \nabla_{\widetilde{\mathcal{D}}} v(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} = \int_K f(\boldsymbol{x}) \psi(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}.$$

Setting $\boldsymbol{w} = \boldsymbol{v}$ in (10.37c) with u replaced by v, and using $\Gamma_{\widetilde{\mathcal{D}}} v = 0$, we get

$$\int_K \boldsymbol{v}(\boldsymbol{x}) \cdot \nabla_{\widetilde{\mathcal{D}}} v(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} + \int_K \Pi_{\widetilde{\mathcal{D}}} v(\boldsymbol{x}) \mathrm{div} \boldsymbol{v}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} = 0,$$

which implies

$$\begin{aligned}
\int_K f(\boldsymbol{x}) \psi(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} &= \int_K \nabla_{\widetilde{\mathcal{D}}} u(\boldsymbol{x}) \cdot \nabla_{\widetilde{\mathcal{D}}} v(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \\
&= -\int_K \boldsymbol{v}(\boldsymbol{x}) \cdot \nabla_{\widetilde{\mathcal{D}}} v(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \\
&= \int_K \Pi_{\widetilde{\mathcal{D}}} v(\boldsymbol{x}) \mathrm{div} \boldsymbol{v}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} = \int_K \psi(\boldsymbol{x}) \mathrm{div} \boldsymbol{v}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}.
\end{aligned}$$

This completes the proof of (10.36c). Then, we take $\mu \in M_h^0$ and let $v \in X_{\widetilde{\mathcal{D}},0}$ be such that $\Pi_{\widetilde{\mathcal{D}}} v = 0$ and $\Gamma_{\widetilde{\mathcal{D}}} v_{|\sigma} = \mu_{|\sigma}$ for a given $\sigma = K|L \in \mathcal{F}_{\mathrm{int}}$, and

$\Gamma_{\widetilde{\mathcal{D}}} v_{|\sigma'} = 0$ for all $\sigma' \in \mathcal{F} \setminus \{\sigma\}$. Setting again $\boldsymbol{w} = \boldsymbol{v}$ in (10.37c) with u replaced by v, we get

$$\int_K \boldsymbol{v}(\boldsymbol{x}) \cdot \nabla_{\widetilde{\mathcal{D}}} v(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} - \int_\sigma \mu(\boldsymbol{x})\ \boldsymbol{v}_{|K}(\boldsymbol{x}) \cdot \boldsymbol{n}_{K,\sigma} \mathrm{d}\gamma(\boldsymbol{x}) = 0,$$

and

$$\int_L \boldsymbol{v}(\boldsymbol{x}) \cdot \nabla_{\widetilde{\mathcal{D}}} v(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} - \int_\sigma \mu(\boldsymbol{x})\ \boldsymbol{v}_{|L}(\boldsymbol{x}) \cdot \boldsymbol{n}_{L,\sigma} \mathrm{d}\gamma(\boldsymbol{x}) = 0.$$

Summing these relations and recalling that $\boldsymbol{v} = -\nabla_{\widetilde{\mathcal{D}}} u$ gives

$$\int_{K\cup L} \nabla_{\mathcal{D}} u(\boldsymbol{x}) \cdot \nabla_{\widetilde{\mathcal{D}}} v(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} + \int_\sigma \mu(\boldsymbol{x})\ \boldsymbol{v}_{|K}(\boldsymbol{x}) \cdot \boldsymbol{n}_{K,\sigma} \mathrm{d}\gamma(\boldsymbol{x}) + \int_\sigma \mu(\boldsymbol{x})\ \boldsymbol{v}_{|L}(\boldsymbol{x}) \cdot \boldsymbol{n}_{L,\sigma} \mathrm{d}\gamma(\boldsymbol{x}) = 0. \quad (10.38)$$

Using the GS (2.23), the fact that the support of $\nabla_{\widetilde{\mathcal{D}}} v$ is reduced to $K \cup L$, and that $\Pi_{\widetilde{\mathcal{D}}} v = 0$, we see that the first term in (10.38) vanishes. This proves (10.36d).

Conversely, considering the solution $(\boldsymbol{v}, q, \lambda)$ to (10.36), since $q \in W_h$ and $\lambda \in M_h^0$, there exists a unique $u \in X_{\widetilde{\mathcal{D}},0}$ such that $q = \Pi_{\widetilde{\mathcal{D}}} u$ and $\lambda = \Gamma_{\widetilde{\mathcal{D}}} u$. From (10.36b), we get that $\boldsymbol{v} = -\nabla_{\widetilde{\mathcal{D}}} u$. For any $v \in X_{\widetilde{\mathcal{D}},0}$, letting $\psi = \Pi_{\widetilde{\mathcal{D}}} v$ and $\mu = \Gamma_{\widetilde{\mathcal{D}}} v$, and following the same computation as above, we get that (10.36c) and (10.36d) imply (2.23), using (10.37c) where u is replaced by v.

■

Theorem 10.10 (Properties of the hybrid $\mathbb{RT}_k$ GDs). *Let $(\mathfrak{T}_m)_{m\in\mathbb{N}}$ be a sequence of conforming simplicial meshes in the sense of Definition 7.4, such that $h_{\mathcal{M}_m} \to 0$ as $m \to \infty$ and $(\kappa_{\mathfrak{T}_m})_{m\in\mathbb{N}}$ is bounded (see (7.10)). Let $\widetilde{\mathcal{D}}_m = (X_{\widetilde{\mathcal{D}}_m,0}, \Pi_{\widetilde{\mathcal{D}}_m}, \nabla_{\widetilde{\mathcal{D}}_m})$ be the gradient discretisation defined by (10.37) with the spaces (10.10) built on $\mathfrak{T}_m$.*

Then $(\widetilde{\mathcal{D}}_m)_{m\in\mathbb{N}}$ is coercive, GD-consistent, limit-conforming and compact in the sense of the definitions of Section 2.1.1.

Proof. Let $\mathfrak{T}$ be a conforming simplicial mesh, $\widetilde{\mathcal{D}}$ be the $\mathbb{RT}_k$ HMFEGD defined by (10.37), and $\mathcal{D}$ be the $\mathbb{RT}_k$ MFEGD defined by (10.7). Define the mapping $T : X_{\widetilde{\mathcal{D}},0} \to X_{\mathcal{D},0}$ by $T(\widetilde{u}) = (\widetilde{u}_i)_{i\in I}$. Then $\Pi_{\widetilde{\mathcal{D}}} \widetilde{u} = \Pi_{\mathcal{D}} T(\widetilde{u})$. By selecting $\boldsymbol{w} = \nabla_{\mathcal{D}} T(\widetilde{u}) \in \boldsymbol{V}_h^{\rm div} \subset \boldsymbol{V}_h$ in (10.37c) and summing on $K \in \mathcal{M}$, all the integrals on $\sigma \in \mathcal{F}_{\rm int}$ vanish and we obtain

$$\int_\Omega \nabla_{\mathcal{D}} T(\widetilde{u})(\boldsymbol{x}) \cdot \nabla_{\widetilde{\mathcal{D}}} \widetilde{u}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} + \int_\Omega \Pi_{\widetilde{\mathcal{D}}} \widetilde{u}(\boldsymbol{x}) \mathrm{div}(\nabla_{\mathcal{D}} T(\widetilde{u}))(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} = 0.$$

Using (10.7d) with $\boldsymbol{w} = \nabla_{\mathcal{D}} T(\widetilde{u})$ and $T(\widetilde{u})$ instead of u then yields

$$\int_\Omega \nabla_{\mathcal{D}} T(\widetilde{u})(\boldsymbol{x}) \cdot \nabla_{\widetilde{\mathcal{D}}} \widetilde{u}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} = -\int_\Omega \Pi_{\mathcal{D}} T(\widetilde{u})(\boldsymbol{x}) \mathrm{div}(\nabla_{\mathcal{D}} T(\widetilde{u}))(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}$$
$$= \int_\Omega \nabla_{\mathcal{D}} T(\widetilde{u})(\boldsymbol{x}) \cdot \nabla_{\mathcal{D}} T(\widetilde{u})(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}.$$

The Cauchy–Schwarz inequality then leads to

$$\|\nabla_{\mathcal{D}} T(\widetilde{u})\|_{L^2(\Omega)^d} \leq \|\nabla_{\widetilde{\mathcal{D}}} \widetilde{u}\|_{L^2(\Omega)^d}, \ \forall \widetilde{u} \in X_{\widetilde{\mathcal{D}},0}. \tag{10.39}$$

We can now prove the properties of $(\widetilde{\mathcal{D}}_m)_{m\in\mathbb{N}}$.

Coercivity. The coercivity follows from the relation $\Pi_{\widetilde{\mathcal{D}}_m} \widetilde{u} = \Pi_{\mathcal{D}_m} T(\widetilde{u})$, from the coercivity of $(\mathcal{D}_m)_{m\in\mathbb{N}}$ (Theorem 10.7), and from (10.39).

GD-consistency. Let $u \in X_{\mathcal{D}_m,0}$. Recalling the definition of $\Gamma_{\mathcal{D}_m}$ in Lemma 10.4, there exists $\widetilde{u} \in X_{\widetilde{\mathcal{D}}_m,0}$ such that $T(\widetilde{u}) = u$ (this defines $(\widetilde{u}_i)_{i\in I}$) and $\Gamma_{\widetilde{\mathcal{D}}_m} \widetilde{u} = \Gamma_{\mathcal{D}_m} u$ (this defines $(\widetilde{u}_j)_{j\in J}$). Then (10.18) and (10.37c) show that $\nabla_{\mathcal{D}_m} u = \nabla_{\widetilde{\mathcal{D}}_m} \widetilde{u}$. Take $\varphi \in H_0^1(\Omega)$ and set

$$u = \operatorname*{argmin}_{v\in X_{\mathcal{D}_m,0}} \left(\|\Pi_{\mathcal{D}_m} v - \varphi\|_{L^2(\Omega)} + \|\nabla_{\mathcal{D}_m} v - \nabla\varphi\|_{L^2(\Omega)^d} \right).$$

Then, since $\Pi_{\mathcal{D}_m} u = \Pi_{\mathcal{D}_m} T(\widetilde{u}) = \Pi_{\widetilde{\mathcal{D}}_m} \widetilde{u}$,

$$S_{\widetilde{\mathcal{D}}_m}(\varphi) \leq \left\| \Pi_{\widetilde{\mathcal{D}}_m} \widetilde{u} - \varphi \right\|_{L^2(\Omega)} + \left\| \nabla_{\widetilde{\mathcal{D}}_m} \widetilde{u} - \nabla\varphi \right\|_{L^2(\Omega)^d}$$
$$= \|\Pi_{\mathcal{D}_m} u - \varphi\|_{L^2(\Omega)} + \|\nabla_{\mathcal{D}_m} u - \nabla\varphi\|_{L^2(\Omega)^d} \leq S_{\mathcal{D}_m}(\varphi)$$

and the consistency of $(\widetilde{\mathcal{D}}_m)_{m\in\mathbb{N}}$ follows from the consistency of $(\mathcal{D}_m)_{m\in\mathbb{N}}$ (Theorem 10.7).

Limit-conformity. If $\boldsymbol{w} \in \boldsymbol{V}_{h_m}^{\mathrm{div}}$ then writing (10.37c) over each cell and summing over the cells, the face terms cancel (since $\boldsymbol{w} \cdot \boldsymbol{n}_{K,\sigma} + \boldsymbol{w} \cdot \boldsymbol{n}_{L,\sigma} = 0$ whenever $\sigma \in \mathcal{F}_{\mathrm{int}}$ with $\mathcal{M}_\sigma = \{K, L\}$), and we see that (10.7d) holds with $\widetilde{\mathcal{D}}_m$ instead of $\mathcal{D}$. The limit-conformity can therefore be proved in a similar way as in Theorem 10.3, by taking $\boldsymbol{\varphi}_m \in \boldsymbol{V}_{h_m}^{\mathrm{div}}$ that converges to $\boldsymbol{\varphi}$ in H_{div} and by writing, for $\widetilde{u}_m \in X_{\mathcal{D}_m,0}$,

$$\widetilde{W}_{\widetilde{\mathcal{D}}_m}(\boldsymbol{\varphi}, \widetilde{u}_m) = \int_\Omega \left(\nabla_{\widetilde{\mathcal{D}}_m} \widetilde{u}_m(\boldsymbol{x}) \cdot \boldsymbol{\varphi}(\boldsymbol{x}) + \Pi_{\widetilde{\mathcal{D}}_m} \widetilde{u}_m(\boldsymbol{x}) \mathrm{div}\boldsymbol{\varphi}(\boldsymbol{x}) \right) \mathrm{d}\boldsymbol{x} =$$
$$\int_\Omega \Big(\nabla_{\widetilde{\mathcal{D}}_m} \widetilde{u}_m(\boldsymbol{x}) \cdot (\boldsymbol{\varphi}(\boldsymbol{x}) - \boldsymbol{\varphi}_m(\boldsymbol{x})) + \Pi_{\widetilde{\mathcal{D}}_m} \widetilde{u}_m(\boldsymbol{x}) (\mathrm{div}\boldsymbol{\varphi}(\boldsymbol{x}) - \mathrm{div}\boldsymbol{\varphi}_m(\boldsymbol{x})) \Big) \mathrm{d}\boldsymbol{x}.$$

Compactness. Similarly to the coercivity, this property is an immediate consequence of the compactness of $(\mathcal{D}_m)_{m\in\mathbb{N}}$ (Theorem 10.7), of (10.39) and of $\Pi_{\widetilde{\mathcal{D}}_m} \widetilde{u} = \Pi_{\mathcal{D}_m} T(\widetilde{u})$. ∎

Remark 10.11. The coercivity, limit-conformity and compactness properties of sequences of $\mathbb{RT}_k$ HMFEGDs can also be established through the notion of control by a polytopal toolbox. The control is defined in a similar way as (10.22) in Lemma 10.5, by using $\Gamma_{\widetilde{\mathcal{D}}}$.

11

Discontinuous Galerkin methods

Discontinuous Galerkin (DG) methods have a long history, and have become very popular numerical methods for PDEs of various types. They present the advantage of being applicable on generic meshes and of providing possibly high order approximations of weakly regular functions. DG methods are based on polynomial approximations in the cells, without continuity conditions imposed on these polynomials across the faces. The continuity is weakly enforced in the scheme, by using various possible stabilisation terms controlling the jumps of the polynomials at the interfaces [17]. The convergence of DG methods has been proved for a variety of problems and stabilisation terms (see [61] and references therein).

In this chapter, following [97], we build a discontinuous Galerkin gradient discretisation (DGGD) such that, for linear diffusion problems, the resulting gradient scheme is identical to the symmetric interior penalty discontinuous Galerkin scheme (SIPG).
We then prove that, under standard regularity assumptions on the meshes, sequences of DGGDs satisfy all the core properties of a GDM on general polytopal meshes in any space dimension. This is done via the notion of control by a polytopal toolbox as defined in Chapter 7. Throughout this chapter, we consider the general case $p \in (1, +\infty)$, but we restrict ourselves to homogeneous Dirichlet boundary conditions for the sake of simplicity.

11.1 Discontinuous Galerkin gradient discretisation

11.1.1 Definition of the DGGD

The definition of a DGGD uses some notations presented in Figure 11.1.

Definition 11.1 (Discontinuous Galerkin gradient discretisation). *Let $\mathfrak{T} = (\mathcal{M}, \mathcal{F}, \mathcal{P}, \mathcal{V})$ be a polytopal mesh of Ω in the sense of Definition 7.2, $k \in \mathbb{N}^*$ be a polynomial degree, and $\beta \in (0, 1)$. A discontinuous Galerkin*

gradient discretisation (DGGD) $\mathcal{D} = (X_{\mathcal{D},0}, \Pi_{\mathcal{D}}, \nabla_{\mathcal{D}})$ *of degree* k *on* $\mathfrak{T}$ *is defined by the following elements.*

1. *Let* $(\chi_i)_{i\in I}$ *be a basis of the space defined by* (10.10c)*, that is*

$$W_h = \{w \in L^p(\Omega) \,:\, w_{|K} \in \mathbb{P}_k(K)\,,\ \forall K \in \mathcal{M}\}, \tag{11.1}$$

and set

$$X_{\mathcal{D},0} = \{v = (v_i)_{i\in I} \,:\, v_i \in \mathbb{R} \text{ for all } i \in I\}. \tag{11.2}$$

2. *The operator* $\Pi_{\mathcal{D}}$ *is the reconstruction in* $L^p(\Omega)$ *of the elements of* $X_{\mathcal{D},0}$*:*

$$\forall v \in X_{\mathcal{D},0}\,,\ \Pi_{\mathcal{D}} v = \sum_{i\in I} v_i \chi_i. \tag{11.3}$$

For all $K \in \mathcal{M}$ *and* $v \in X_{\mathcal{D},0}$*, denote by* $\Pi_{\overline{K}} v \in \mathbb{P}_k(\overline{K})$ *the polynomial defined by* $(\Pi_{\mathcal{D}} v)_{|K}$ *and extended to* $\overline{K}$*, and set* $\nabla_{\overline{K}} v = \nabla \Pi_{\overline{K}} v$*.*

3. *For* $v \in X_{\mathcal{D},0}$*,* $K \in \mathcal{M}$*,* $\sigma \in \mathcal{F}_K$ *and a.e.* $\boldsymbol{x} \in D_{K,\sigma}$*, set*

$$\nabla_{\mathcal{D}} v(\boldsymbol{x}) = \nabla_{\overline{K}} v(\boldsymbol{x}) + \psi(s) \frac{[v]_{K,\sigma}(\boldsymbol{y})}{d_{K,\sigma}} \boldsymbol{n}_{K,\sigma} \tag{11.4}$$

where:

- $\boldsymbol{x}$ *is (uniquely) decomposed as* $\boldsymbol{x} = \boldsymbol{x}_K + s(\boldsymbol{y} - \boldsymbol{x}_K)$ *with* $s \in (0,1]$ *and* $\boldsymbol{y} \in \sigma$*,*
- *for all* $\boldsymbol{y} \in \sigma$*,*

$$\begin{aligned} &\text{if } \sigma \in \mathcal{F}_{\text{int}} \text{ and } \mathcal{M}_\sigma = \{K, L\}\,, \\ &\qquad\qquad [v]_{K,\sigma}(\boldsymbol{y}) = \frac{\Pi_{\overline{L}} v(\boldsymbol{y}) - \Pi_{\overline{K}} v(\boldsymbol{y})}{2} \\ &\text{if } \sigma \in \mathcal{F}_{\text{ext}} \text{ and } \mathcal{M}_\sigma = \{K\}\,, \\ &\qquad\qquad [v]_{K,\sigma}(\boldsymbol{y}) = 0 - \Pi_{\overline{K}} v(\boldsymbol{y}), \end{aligned} \tag{11.5}$$

- $\psi \,:\, (0,1) \to \mathbb{R}$ *is the unique function such that* $\psi(s) = 0$ *on* $(0,\beta)$*,* $\psi_{|[\beta,1]} \in \mathbb{P}_{k-1}([\beta,1])$ *and*

$$\int_\beta^1 \psi(s) s^{d-1} \mathrm{d}s = 1, \tag{11.6a}$$

$$\forall i = 1, \ldots, k-1,\ \int_\beta^1 (1-s)^i \psi(s) s^{d-1} \mathrm{d}s = 0. \tag{11.6b}$$

4. *The fact that* $\|\nabla_{\mathcal{D}} \cdot\|_{L^p(\Omega)^d}$ *is a norm on* $X_{\mathcal{D},0}$ *is a consequence of Lemma 11.10 (see Remark 11.11).*

In the case $k = 1$, the function $\psi_{|[\beta,1]} \in \mathbb{P}_0([\beta,1])$ has the constant value $\frac{d}{1-\beta^d}$. For the general case $k \in \mathbb{N}^\star$, considering the basis $\mathcal{B} = (1, (1-s), \ldots, (1-s)^{k-1})$ of $\mathbb{P}_{k-1}([\beta,1])$, and writing the function $\psi_{|[\beta,1]}$ as $\psi(s) = \sum_{j=1}^k \alpha_j (1-$

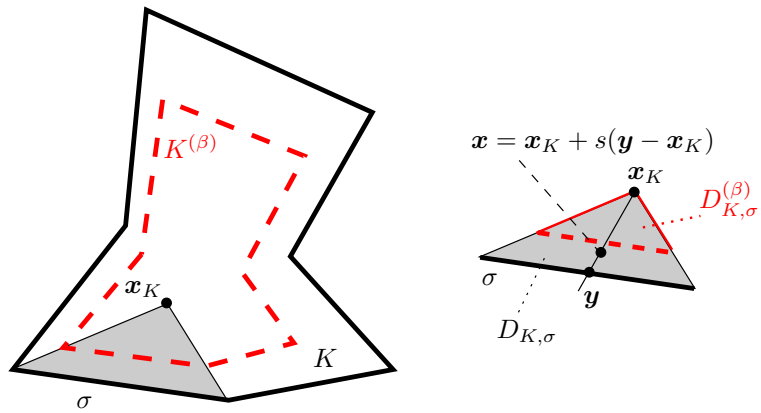


Fig. 11.1. Notations for the definition of a DG gradient discretisation

$s)^{j-1}$, the equations (11.6) boil down to the system $A\boldsymbol{\alpha} = (1, 0, \ldots, 0)^T$ where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_k)^T$ and

$$A_{i,j} = \int_\beta^1 (1-s)^{i+j-2} s^{d-1} \mathrm{d}s.$$

In other words, A is the Gram matrix of $\mathcal{B}$ for the inner product $(f, g) \mapsto \int_\beta^1 f(s)g(s)s^{d-1}\mathrm{d}s$ on $\mathbb{P}_k([\beta, 1])$. Hence, A is symmetric positive definite and there exists a unique $\psi_{|[\beta,1]} \in \mathbb{P}_{k-1}([\beta, 1])$ such that (11.6) holds.

Remark 11.2 (Definition of the jump at the faces of the mesh). In (11.5), the jump across the faces is divided by 2 for interior faces to allow for a seamless definition of $\nabla_{\mathcal{D}}$ on all $D_{K,\sigma}$, no matter if $\sigma \in \mathcal{F}_{\mathrm{int}}$ or $\sigma \in \mathcal{F}_{\mathrm{ext}}$.

Throughout this chapter, given $K \in \mathcal{M}$ and $\sigma \in \mathcal{F}_K$, we use the functions $s : D_{K,\sigma} \to (0, 1)$ and $\boldsymbol{y} : D_{K,\sigma} \to \sigma$ defined such that, for $\boldsymbol{x} \in D_{K,\sigma}$, $\boldsymbol{x} = \boldsymbol{x}_K + s(\boldsymbol{x})(\boldsymbol{y}(\boldsymbol{x}) - \boldsymbol{x}_K)$. In other words,

$$s(\boldsymbol{x}) = \frac{(\boldsymbol{x} - \boldsymbol{x}_K) \cdot \boldsymbol{n}_{K,\sigma}}{d_{K,\sigma}} \quad \text{and} \quad \boldsymbol{y}(\boldsymbol{x}) = \boldsymbol{x}_K + \frac{\boldsymbol{x} - \boldsymbol{x}_K}{s(\boldsymbol{x})}.$$

It will also be useful to split the cone $D_{K,\sigma}$ into $D^{(\beta)}_{K,\sigma}$ and $D_{K,\sigma} \setminus D^{(\beta)}_{K,\sigma}$ with

$$D^{(\beta)}_{K,\sigma} := \{\boldsymbol{x} \in D_{K,\sigma} : \boldsymbol{x} = \boldsymbol{x}_K + s(\boldsymbol{y} - \boldsymbol{x}_K), s \in (0, \beta],\ \boldsymbol{y} \in \sigma\}. \tag{11.7}$$

We then set

$$K^{(\beta)} = \bigcup_{\sigma \in \mathcal{F}_K} D^{(\beta)}_{K,\sigma}.$$

This is illustrated in Figure 11.1. Note that $|D_{K,\sigma} \setminus D^{(\beta)}_{K,\sigma}| = \frac{1-\beta^d}{d} d_{K,\sigma}|\sigma|$.

Remark 11.3. If $v \in X_{\mathcal{D},0}$, the function $\Pi_{\mathcal{D}} v \in W_h$ has discontinuities at the faces of the mesh. Its distributional gradient therefore involves measures concentrated on the faces (and absolutely continuous with respect to the $(d-1)$-dimensional measure on the faces). The gradient $\nabla_{\mathcal{D}} v \in L^p(\Omega)^d$ can be seen as a regularisation of this distributional gradient, in which the measures on the faces are "spread" over a neighbourhood of the faces (specifically, for each $\sigma \in \mathcal{F}$, over $\cup_{M \in \mathcal{M}_\sigma} D_{M,\sigma} \setminus D^{(\beta)}_{M,\sigma}$), and weighted by functions whose purpose is to ensure that the corresponding term has a proper orthogonality property with respect to $\nabla_{\overline{K}} v$, see (11.12).

Remark 11.4. The mathematical analysis of the DGGD is essentially unchanged if the uniform constant β is replaced with local constants $\beta_{K,\sigma} \in (0,1)$.

Remark 11.5 (Piecewise constant reconstruction). A DGGD with piecewise constant reconstruction can be obtained by replacing $\Pi_{\mathcal{D}}$ with $\widehat{\Pi}_{\mathcal{D}}$ defined, for all $K \in \mathcal{M}$ and a.e. $\boldsymbol{x} \in K$, by

$$\widehat{\Pi}_{\mathcal{D}} v(\boldsymbol{x}) = \frac{1}{|K|} \int_K \Pi_{\overline{K}} v(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}.$$

This function reconstruction is piecewise constant in the sense of Definition 2.12 if the basis $(\chi_i)_{i \in I}$ is chosen such that, for each $K \in \mathcal{M}$, there exists $i \in I$ with $v_i = \frac{1}{|K|} \int_K \Pi_{\overline{K}} v(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}$ for all $v \in X_{\mathcal{D},0}$ (in other words, the zero-th moment of $\Pi_{\overline{K}} v$ must be a degree of freedom of the method).

11.1.2 Link with the symmetric interior penalty discontinuous Galerkin (SIPG) method

Theorem 11.6 (SIPG is a gradient scheme). *Let $\mathfrak{T}$ be a polytopal mesh and $\mathcal{D}$ be a discontinuous Galerkin gradient discretisation on $\mathfrak{T}$ as per Definition 11.1. Assume Hypotheses (2.21), $\boldsymbol{F} = 0$ and that Λ is constant in each cell of $\mathfrak{T}$. Then the gradient scheme (2.23) constructed from $\mathcal{D}$ for the linear elliptic problem (2.22) is a symmetric interior penalty Galerkin (SIPG) method as in [83, 61].*

Proof. Recall that the gradient scheme (2.23) is given by

$$\begin{aligned} & u \in X_{\mathcal{D},0}, \ \forall v \in X_{\mathcal{D},0}, \\ & \int_\Omega \Lambda(\boldsymbol{x}) \nabla_{\mathcal{D}} u(\boldsymbol{x}) \cdot \nabla_{\mathcal{D}} v(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} = \int_\Omega f(\boldsymbol{x}) \Pi_{\mathcal{D}} v(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}. \end{aligned} \tag{11.8}$$

Write the integral in the left-hand side as a sum of integrals over the cells and use the definition (11.4) of $\nabla_{\mathcal{D}}$ to develop the dot product. Since Λ is constant in each cell, denoting by Λ_K the value of Λ in $K \in \mathcal{M}$ this yields

$$
\begin{aligned}
&\int_\Omega \Lambda(\boldsymbol{x})\nabla_{\mathcal{D}}u(\boldsymbol{x})\cdot\nabla_{\mathcal{D}}v(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\\
&= \sum_{K\in\mathcal{M}}\Bigg[\int_K \Lambda_K\nabla_{\overline{K}}u(\boldsymbol{x})\cdot\nabla_{\overline{K}}v(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\\
&\quad+\sum_{\sigma\in\mathcal{F}_K}\int_{D_{K,\sigma}}\Lambda_K\psi(s(\boldsymbol{x}))\frac{[u]_{K,\sigma}(\boldsymbol{y}(\boldsymbol{x}))}{d_{K,\sigma}}\boldsymbol{n}_{K,\sigma}\cdot\nabla_{\overline{K}}v(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\\
&\quad+\sum_{\sigma\in\mathcal{F}_K}\int_{D_{K,\sigma}}\Lambda_K\nabla_{\overline{K}}u(\boldsymbol{x})\cdot\psi(s(\boldsymbol{x}))\frac{[v]_{K,\sigma}(\boldsymbol{y}(\boldsymbol{x}))}{d_{K,\sigma}}\boldsymbol{n}_{K,\sigma}\mathrm{d}\boldsymbol{x}\\
&\quad+\sum_{\sigma\in\mathcal{F}_K}\int_{D_{K,\sigma}}\Lambda_K\frac{[u]_{K,\sigma}(\boldsymbol{y}(\boldsymbol{x}))}{d_{K,\sigma}}\boldsymbol{n}_{K,\sigma}\cdot\frac{[v]_{K,\sigma}(\boldsymbol{y}(\boldsymbol{x}))}{d_{K,\sigma}}\boldsymbol{n}_{K,\sigma}\psi(s(\boldsymbol{x}))^2\mathrm{d}\boldsymbol{x}\Bigg].
\end{aligned}
\tag{11.9}
$$

Let us first consider the cross-products (third and fourth lines above). They are symmetric in u, v so understanding only one of them is sufficient. The change of variable $\boldsymbol{x}\in D_{K,\sigma}\mapsto(\boldsymbol{y},s)\in\sigma\times(0,1)$ is defined by $\boldsymbol{x}=\boldsymbol{x}_K+s(\boldsymbol{y}-\boldsymbol{x}_K)$ and satisfies $\mathrm{d}\boldsymbol{x}=d_{K,\sigma}s^{d-1}\mathrm{d}\gamma(\boldsymbol{y})\mathrm{d}s$. Hence,

$$
\begin{aligned}
&\int_{D_{K,\sigma}}\Lambda_K\psi(s(\boldsymbol{x}))\frac{[u]_{K,\sigma}(\boldsymbol{y}(\boldsymbol{x}))}{d_{K,\sigma}}\boldsymbol{n}_{K,\sigma}\cdot\nabla_{\overline{K}}v(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\\
&=\int_\sigma\Lambda_K\frac{[u]_{K,\sigma}(\boldsymbol{y})}{d_{K,\sigma}}\boldsymbol{n}_{K,\sigma}\cdot\int_\beta^1\nabla_{\overline{K}}v(\boldsymbol{x}(\boldsymbol{y},s))\psi(s)s^{d-1}\mathrm{d}s\mathrm{d}\gamma(\boldsymbol{y})d_{K,\sigma}.
\end{aligned}
\tag{11.10}
$$

Since $\nabla_{\overline{K}}v\in\mathbb{P}_k(\overline{K})$, the function $\varphi(s)=\nabla_{\overline{K}}v(\boldsymbol{x}_K+s(\boldsymbol{y}-\boldsymbol{x}_K))\cdot\boldsymbol{n}_{K,\sigma}$ is a polynomial of degree k or less. Expanding it along its Taylor series at $s=1$ leads to

$$
\varphi(s)=\nabla_{\overline{K}}v(\boldsymbol{y})\cdot\boldsymbol{n}_{K,\sigma}+\sum_{m=1}^{k-1}p_m(\boldsymbol{y})(1-s)^m,
\tag{11.11}
$$

where, for $m=1,\ldots,k-1$, $p_m(\boldsymbol{y})$ is a polynomial with degree less or equal to $k-1$ with respect to the coordinates of $\boldsymbol{y}$. Using (11.6) thus shows that

$$
\int_\beta^1\nabla_{\overline{K}}v(\boldsymbol{x}(\boldsymbol{y},s))\cdot\boldsymbol{n}_{K,\sigma}\psi(s)s^{d-1}\mathrm{d}s=\nabla_{\overline{K}}v(\boldsymbol{y})\cdot\boldsymbol{n}_{K,\sigma}.
\tag{11.12}
$$

Plugged in (11.10) this yields the following representation of the first cross-product in (11.9):

$$
\begin{aligned}
&\int_{D_{K,\sigma}}\Lambda_K\psi(s(\boldsymbol{x}))\frac{[u]_{K,\sigma}(\boldsymbol{y}(\boldsymbol{x}))}{d_{K,\sigma}}\boldsymbol{n}_{K,\sigma}\cdot\nabla_{\overline{K}}v(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\\
&\qquad\qquad=\int_\sigma\Lambda_K[u]_{K,\sigma}(\boldsymbol{y})\boldsymbol{n}_{K,\sigma}\cdot\nabla_{\overline{K}}v(\boldsymbol{y})\mathrm{d}\gamma(\boldsymbol{y}).
\end{aligned}
$$

Using the same change of variable as above, the last term in (11.9) is easily re-arranged into

$$\int_{D_{K,\sigma}} \Lambda_K \frac{[u]_{K,\sigma}(\boldsymbol{y}(\boldsymbol{x}))}{d_{K,\sigma}} \boldsymbol{n}_{K,\sigma} \cdot \frac{[v]_{K,\sigma}(\boldsymbol{y}(\boldsymbol{x}))}{d_{K,\sigma}} \boldsymbol{n}_{K,\sigma} \psi(s(\boldsymbol{x}))^2 \mathrm{d}\boldsymbol{x}$$
$$= \frac{\Lambda_K \boldsymbol{n}_{K,\sigma} \cdot \boldsymbol{n}_{K,\sigma}}{d_{K,\sigma}} \int_\beta^1 \psi(s)^2 s^{d-1} \mathrm{d}s \int_\sigma [u]_{K,\sigma}(\boldsymbol{y})[v]_{K,\sigma}(\boldsymbol{y}) \mathrm{d}\gamma(\boldsymbol{y}).$$

The relation (11.9) is therefore re-written as

$$\begin{aligned}
&\int_\Omega \Lambda(\boldsymbol{x}) \nabla_{\mathcal{D}} u(\boldsymbol{x}) \cdot \nabla_{\mathcal{D}} v(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \\
&= \sum_{K \in \mathcal{M}} \Bigg[\int_K \Lambda_K \nabla_{\overline{K}} u(\boldsymbol{x}) \cdot \nabla_{\overline{K}} v(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \\
&\quad + \sum_{\sigma \in \mathcal{F}_K} \Bigg\{ \int_\sigma \Lambda_K \big([u]_{K,\sigma}(\boldsymbol{y}) \nabla_{\overline{K}} v(\boldsymbol{y}) + [v]_{K,\sigma}(\boldsymbol{y}) \nabla_{\overline{K}} u(\boldsymbol{y}) \big) \cdot \boldsymbol{n}_{K,\sigma} \mathrm{d}\gamma(\boldsymbol{y}) \\
&\quad + \frac{\Lambda_K \boldsymbol{n}_{K,\sigma} \cdot \boldsymbol{n}_{K,\sigma}}{d_{K,\sigma}} \int_\beta^1 \psi(s)^2 s^{d-1} \mathrm{d}s \int_\sigma [u]_{K,\sigma}(\boldsymbol{y})[v]_{K,\sigma}(\boldsymbol{y}) \mathrm{d}\gamma(\boldsymbol{y}) \Bigg\} \Bigg].
\end{aligned}$$

This right-hand side corresponds to the bilinear form in the SIPG scheme as presented in [83, 61]. Here, the penalty coefficient τ_σ (term $\frac{\sigma_e}{|e|^{\beta_0}}$ of [83, eqn. (11)], term $\frac{\eta}{h_F}$ of [61, eqn. (4.12)]) is

if $\sigma \in \mathcal{F}_{\text{int}}$ and $\mathcal{M}_\sigma = \{K, L\}$,
$$\tau_\sigma = \left(\frac{1}{4} \int_\beta^1 \psi(s)^2 s^{d-1} \mathrm{d}s \right) \left(\frac{\Lambda_K \boldsymbol{n}_{K,\sigma} \cdot \boldsymbol{n}_{K,\sigma}}{d_{K,\sigma}} + \frac{\Lambda_L \boldsymbol{n}_{L,\sigma} \cdot \boldsymbol{n}_{L,\sigma}}{d_{L,\sigma}} \right),$$
if $\sigma \in \mathcal{F}_{\text{ext}}$ and $\mathcal{M}_\sigma = \{K\}$,
$$\tau_\sigma = \left(\int_\beta^1 \psi(s)^2 s^{d-1} \mathrm{d}s \right) \frac{\Lambda_K \boldsymbol{n}_{K,\sigma} \cdot \boldsymbol{n}_{K,\sigma}}{d_{K,\sigma}}.$$

The proof is complete by noticing that the right hand side of (11.8) is identical to the right-hand side of SIPG methods as in [83, 61]. ∎

Remark 11.7 (Lower bound for τ_σ and mesh regularity). By the Cauchy–Schwarz inequality and (11.6a),

$$1 = \left(\int_\beta^1 \psi(s) s^{d-1} \mathrm{d}s \right)^2 \le \int_\beta^1 \psi(s)^2 s^{d-1} \mathrm{d}s \int_\beta^1 s^{d-1} \mathrm{d}s.$$

Hence,

$$\int_\beta^1 \psi(s)^2 s^{d-1} \mathrm{d}s \ge \frac{d}{1 - \beta^d} \ge d.$$

The coercivity assumption (2.21a) on Λ therefore shows that $\tau_\sigma \ge d\underline{\lambda}/d_{K,\sigma}$ for all $\sigma \in \mathcal{F}$, where K is any mesh in $\mathcal{M}_\sigma$.

Classical analyses of SIPG methods require the stabilisation parameter to be larger than some constant depending on the mesh regularity (in particular, through $\max_{K\in\mathcal{M}} \mathrm{Card}(\mathcal{F}_K)$ and some discrete trace constant, see [61, Lemma 4.12]). For the DGGD analysis, no such assumption is made. The lower bound on τ_σ only depends on $d_{K,\sigma}$, and no "large enough" parameter has to be selected by the user.

11.2 Mathematical properties of DG gradient discretisations

11.2.1 Preliminary results

Definition 11.8 (Comparison of norms on $\mathbb{R}^{n+1}$). *Let $n \in \mathbb{N}$ and $q > 0$ be given. We denote by $C_{q,n} > 0$ the largest number, depending only on n, q and d, such that*

$$\forall (a_0, \ldots, a_n) \in \mathbb{R}^{n+1},\ \int_0^1 \left| \sum_{i=0}^n a_i t^i \right|^q t^{d-1} \mathrm{d}s \geq C_{q,n} \sum_{i=0}^n |a_i|^q. \tag{11.13}$$

Note that the existence of $C_{q,n}$ follows from the homogeneity of each side of (11.13), and from the fact that, if the left-hand side vanishes, then the polynomial $\sum_{i=0}^n a_i t^i$ is zero on $[0,1]$, and thus $a_i = 0$ for all $i = 0, \ldots, n$.

The following lemma is instrumental to the study of $\|\nabla_{\mathcal{D}} \cdot\|_{L^p(\Omega)^d}$.

Lemma 11.9. *Let $n \in \mathbb{N}$ and $\beta \in (0,1)$ be given. Let $\mathfrak{T}$ be a polytopal mesh in the sense of Definition 7.2. Then*

$$\forall v \in \mathbb{P}_n(\mathbb{R}^d),\ \forall K \in \mathcal{M},\ \forall \sigma \in \mathcal{F}_K,$$
$$\int_{D_{K,\sigma}} |v(\boldsymbol{x})|^p \mathrm{d}\boldsymbol{x} \leq \frac{(n+1)^{p-1}}{\beta^{d+pn} C_{p,n}} \int_{D_{K,\sigma}^{(\beta)}} |v(\boldsymbol{x})|^p \mathrm{d}\boldsymbol{x},$$

where $C_{p,n}$ is defined in Definition 11.8 with $q = p$, and $D_{K,\sigma}^{(\beta)}$ is defined by (11.7).

Proof. For $K \in \mathcal{M}$ and $\sigma \in \mathcal{F}_K$, compute $\int_{D_{K,\sigma}^{(\beta)}} |v(\boldsymbol{x})|^p \mathrm{d}\boldsymbol{x}$ by making the change of variable $\boldsymbol{x} = \boldsymbol{x}_K + s(\boldsymbol{y} - \boldsymbol{x}_K)$, where $\boldsymbol{y} \in \sigma$ and $s \in (0, \beta)$. Recalling that $\mathrm{d}\boldsymbol{x} = d_{K,\sigma} s^{d-1} \mathrm{d}\gamma(\boldsymbol{y}) \mathrm{d}s$, this yields

$$\int_{D_{K,\sigma}^{(\beta)}} |v(\boldsymbol{x})|^p \mathrm{d}\boldsymbol{x} = \int_\sigma \int_0^\beta |v(\boldsymbol{x}_K + s(\boldsymbol{y} - \boldsymbol{x}_K))|^p d_{K,\sigma} s^{d-1} \mathrm{d}s \mathrm{d}\gamma(\boldsymbol{y}).$$

For a given $\boldsymbol{y} \in \sigma$, $s \mapsto v(\boldsymbol{x}_K + s(\boldsymbol{y} - \boldsymbol{x}_K))$ is a polynomial of degree at most n, and we can thus write $v(\boldsymbol{x}_K + s(\boldsymbol{y} - \boldsymbol{x}_K)) = \sum_{i=0}^n a_i(\boldsymbol{y}) s^i$. Then, using

the notation introduced in Definition 11.8, performing the change of variable $s = \beta t$, and recalling that $\beta \in (0,1)$,

$$\int_0^\beta \left|\sum_{i=0}^n a_i(\boldsymbol{y}) s^i\right|^p s^{d-1}\mathrm{d}s = \beta^d \int_0^1 \left|\sum_{i=0}^n a_i(\boldsymbol{y})\beta^i t^i\right|^p t^{d-1}\mathrm{d}t$$
$$\geq \beta^d C_{p,n} \sum_{i=0}^n |a_i(\boldsymbol{y})\beta^i|^p \geq \beta^{d+pn} C_{p,n} \sum_{i=0}^n |a_i(\boldsymbol{y})|^p.$$

This leads to

$$\int_{D_{K,\sigma}^{(\beta)}} |v(\boldsymbol{x})|^p \mathrm{d}\boldsymbol{x} \geq \beta^{d+pn} C_{p,n} \int_\sigma d_{K,\sigma} \left(\sum_{i=0}^n |a_i(\boldsymbol{y})|^p\right) \mathrm{d}\gamma(\boldsymbol{y}). \tag{11.14}$$

On the other hand, since $(\sum_{i=0}^n |a_i(\boldsymbol{y})|)^p \leq (n+1)^{p-1} \sum_{i=0}^n |a_i(\boldsymbol{y})|^p$ (see (D.1) with $b_i = 1$ for all i),

$$\begin{aligned}\int_{D_{K,\sigma}} |v(\boldsymbol{x})|^p \mathrm{d}\boldsymbol{x} &= \int_\sigma \int_0^1 |v(\boldsymbol{x}_K + s(\boldsymbol{y} - \boldsymbol{x}_K))|^p d_{K,\sigma} s^{d-1}\mathrm{d}s\mathrm{d}\gamma(\boldsymbol{y}) \\ &= \int_\sigma \int_0^1 \left|\sum_{i=0}^n a_i(\boldsymbol{y}) s^i\right|^p d_{K,\sigma} s^{d-1}\mathrm{d}s\mathrm{d}\gamma(\boldsymbol{y}) \\ &\leq (n+1)^{p-1} \int_\sigma \left(\sum_{i=0}^n |a_i(\boldsymbol{y})|^p\right) d_{K,\sigma}\mathrm{d}\gamma(\boldsymbol{y}).\end{aligned}$$

The lemma is proved by using (11.14) in this last equation. ∎

Norms well-suited to the analysis of DG methods involve the L^p norm of the gradient in each cell, and jump terms at the faces. The following lemma relates such a norm with the norm of the DGGD reconstructed gradient.

Lemma 11.10. *Let $\mathcal{D}$ be a DGGD in the sense of Definition 11.1. Then there exists $A > 0$, depending only on β, p, k and d, such that*

$$\forall v \in X_{\mathcal{D},0},\ \frac{1}{A}\|v\|_{\mathrm{DG},p} \leq \|\nabla_{\mathcal{D}} v\|_{L^p(\Omega)^d} \leq A\|v\|_{\mathrm{DG},p}, \tag{11.15}$$

where

$$\|v\|_{\mathrm{DG},p}^p = \sum_{K\in\mathcal{M}} \left(\int_K |\nabla_{\overline{K}} v(\boldsymbol{x})|^p \mathrm{d}\boldsymbol{x} + \sum_{\sigma\in\mathcal{F}_K} \frac{1}{d_{K,\sigma}^{p-1}} \int_\sigma |[v]_{K,\sigma}(\boldsymbol{y})|^p \mathrm{d}\gamma(\boldsymbol{y})\right). \tag{11.16}$$

Remark 11.11 (DG norm). It is easy to see that $\|\cdot\|_{\mathrm{DG},p}$ defined by (11.16) is a norm on $X_{\mathcal{D},0}$. Indeed, if $\|v\|_{\mathrm{DG},p} = 0$ then the jumps at the faces (both

interior and exterior) are equal to 0, which implies that $\Pi_{\mathcal{D}} v \in W_0^{1,p}(\Omega)$. Since $\|v\|_{\mathrm{DG},p} = 0$ implies that $\nabla \Pi_{\mathcal{D}} v = 0$ a.e. in Ω, we get that $\Pi_{\mathcal{D}} v = 0$ a.e. in Ω. It is then deduced from the definition (11.3) of $\Pi_{\mathcal{D}} v$ that $v = 0$.
Note that this DG norm is slightly different from [60, eqn. (5)] or [61, eqn. (5.1)], mostly through the usage of $d_{K,\sigma}$ instead of $\mathrm{diam}(\sigma)$ in the jump term.

Proof. Let $K \in \mathcal{M}$ and $\sigma \in \mathcal{F}_K$. Writing $\boldsymbol{x} = \boldsymbol{x}_K + s(\boldsymbol{y} - \boldsymbol{x}_K)$ and using, for some $c > 0$ to be chosen later, $|a+b|^p \le (1 + c^{p'})^{p-1}(|a|^p + |\frac{b}{c}|^p)$ with $\frac{1}{p} + \frac{1}{p'} = 1$, $a = \nabla_{\overline{K}} v(\boldsymbol{x}) + \frac{[v]_{K,\sigma}(\boldsymbol{y}(\boldsymbol{x}))}{d_{K,\sigma}} \psi(s(\boldsymbol{x}))\boldsymbol{n}_{K,\sigma}$ and $b = -\nabla_{\overline{K}} v(\boldsymbol{x})$, we have

$$\begin{aligned}
&\int_K |\nabla_{\mathcal{D}} v(\boldsymbol{x})|^p \mathrm{d}\boldsymbol{x} \\
&= \int_{K^{(\beta)}} |\nabla_{\overline{K}} v(\boldsymbol{x})|^p \mathrm{d}\boldsymbol{x} \\
&\quad + \sum_{\sigma \in \mathcal{F}_K} \int_{D_{K,\sigma} \setminus D_{K,\sigma}^{(\beta)}} \left| \nabla_{\overline{K}} v(\boldsymbol{x}) + \frac{[v]_{K,\sigma}(\boldsymbol{y}(\boldsymbol{x}))}{d_{K,\sigma}} \psi(s(\boldsymbol{x}))\boldsymbol{n}_{K,\sigma} \right|^p \mathrm{d}\boldsymbol{x} \\
&\ge \int_{K^{(\beta)}} |\nabla_{\overline{K}} v(\boldsymbol{x})|^p \mathrm{d}\boldsymbol{x} - \frac{1}{c^p} \int_{K \setminus K^{(\beta)}} |\nabla_{\overline{K}} v(\boldsymbol{x})|^p \mathrm{d}\boldsymbol{x} \\
&\quad + \frac{1}{(1+c^{p'})^{p-1}} \sum_{\sigma \in \mathcal{F}_K} \int_{D_{K,\sigma} \setminus D_{K,\sigma}^{(\beta)}} \left| \frac{[v]_{K,\sigma}(\boldsymbol{y}(\boldsymbol{x}))}{d_{K,\sigma}} \psi(s(\boldsymbol{x}))\boldsymbol{n}_{K,\sigma} \right|^p \mathrm{d}\boldsymbol{x}. \quad (11.17)
\end{aligned}$$

We have, by equivalence of the l^2 and l^p norms on $\mathbb{R}^d$ (iterate (D.13) with $\alpha = d/2$),

$$\frac{1}{d} \sum_{i=1}^d |\partial_i \Pi_{\overline{K}} v|^p \le |\nabla_{\overline{K}} v|^p \le d^{p/2} \sum_{i=1}^d |\partial_i \Pi_{\overline{K}} v|^p. \quad (11.18)$$

Applying Lemma 11.9 to the polynomial $\partial_i \Pi_{\overline{K}} v \in \mathbb{P}_{k-1}(\mathbb{R}^d)$ for $i = 1, \ldots, d$ then shows that

$$\begin{aligned}
\int_{K^{(\beta)}} |\nabla_{\overline{K}} v(\boldsymbol{x})|^p \mathrm{d}\boldsymbol{x} &\ge \frac{\beta^{d+p(k-1)} C_{p,k-1}}{d k^{p-1}} \sum_{i=1}^d \int_K |\partial_i \Pi_{\overline{K}} v(\boldsymbol{x})|^p \mathrm{d}\boldsymbol{x} \\
&\ge C' \int_K |\nabla_{\overline{K}} v(\boldsymbol{x})|^p \mathrm{d}\boldsymbol{x},
\end{aligned}$$

where by $C' = \frac{\beta^{d+p(k-1)} C_{p,k-1}}{d^{1+p/2} k^{p-1}}$. We then define c by $\frac{1}{c^p} = \frac{1}{2} C'$. Using $|\boldsymbol{n}_{K,\sigma}| = 1$ and coming back to (11.17) yields

$$\begin{aligned}
\int_K |\nabla_{\mathcal{D}} v(\boldsymbol{x})|^p \mathrm{d}\boldsymbol{x} &\ge \frac{1}{2} C' \int_K |\nabla_{\overline{K}} v(\boldsymbol{x})|^p \mathrm{d}\boldsymbol{x} \\
&\quad + \frac{1}{(1+c^{p'})^{p-1}} \sum_{\sigma \in \mathcal{F}_K} \int_{D_{K,\sigma} \setminus D_{K,\sigma}^{(\beta)}} \left| \frac{[v]_{K,\sigma}(\boldsymbol{y}(\boldsymbol{x}))}{d_{K,\sigma}} \psi(s(\boldsymbol{x})) \right|^p \mathrm{d}\boldsymbol{x}. \quad (11.19)
\end{aligned}$$

Using once more the change of variable $\boldsymbol{x} = \boldsymbol{x}_K + s(\boldsymbol{y} - \boldsymbol{x}_K)$, we obtain

$$\begin{aligned}
&\int_{D_{K,\sigma}\setminus D_{K,\sigma}^{(\beta)}} \left| \frac{[v]_{K,\sigma}(\boldsymbol{y}(\boldsymbol{x}))}{d_{K,\sigma}} \psi(s(\boldsymbol{x})) \right|^p \mathrm{d}\boldsymbol{x} \\
&\qquad = d_{K,\sigma} \int_\sigma \frac{|[v]_{K,\sigma}(\boldsymbol{y})|^p}{d_{K,\sigma}^p} \mathrm{d}\gamma(\boldsymbol{y}) \int_\beta^1 |\psi(s)|^p s^{d-1} \mathrm{d}s \\
&\qquad = \frac{1}{d_{K,\sigma}^{p-1}} \int_\sigma |[v]_{K,\sigma}(\boldsymbol{y})|^p \mathrm{d}\gamma(\boldsymbol{y}) \int_\beta^1 |\psi(s)|^p s^{d-1} \mathrm{d}s.
\end{aligned}$$

Used in (11.19), this completes the proof of the left-most inequality in (11.15). To prove the other inequality, simple use the definition of $\nabla_{\mathcal{D}} v$ and (D.12) to write

$$\begin{aligned}
\int_{D_{K,\sigma}} |\nabla_{\mathcal{D}} v(\boldsymbol{x})|^p \mathrm{d}\boldsymbol{x} &\le 2^{p-1} \int_{D_{K,\sigma}} |\nabla_{\overline{K}} v(\boldsymbol{x})|^p \mathrm{d}\boldsymbol{x} \\
&\quad + 2^{p-1} \int_{D_{K,\sigma}\setminus D_{K,\sigma}^{(\beta)}} \left| \frac{[v]_{K,\sigma}(\boldsymbol{y}(\boldsymbol{x}))}{d_{K,\sigma}} \psi(s(\boldsymbol{x})) \right|^p \mathrm{d}\boldsymbol{x} \\
&= 2^{p-1} \int_{D_{K,\sigma}} |\nabla_{\overline{K}} v(\boldsymbol{x})|^p \mathrm{d}\boldsymbol{x} \\
&\quad + \frac{2^{p-1}}{d_{K,\sigma}^{p-1}} \int_\sigma |[v]_{K,\sigma}(\boldsymbol{y})|^p \mathrm{d}\gamma(\boldsymbol{y}) \int_\beta^1 |\psi(s)|^p s^{d-1} \mathrm{d}s.
\end{aligned}$$

■

We now define, and state estimates on, a control of a DGGD by a polytopal toolbox.

Lemma 11.12 (Control of a DGGD by a polytopal toolbox). *Let $\mathfrak{T}$ be a polytopal mesh and $\mathcal{D}$ be a DGGD in the sense of Definition 11.1. Define the control $\boldsymbol{\Phi} : X_{\mathcal{D},0} \to X_{\mathfrak{T},0}$ of $\mathcal{D}$ by $\mathfrak{T}$ (see Definition 7.9) by: for all $v \in X_{\mathcal{D},0}$,*

$$\begin{aligned}
&\forall \sigma \in \mathcal{F}_{\mathrm{int}} \text{ with } \mathcal{M}_\sigma = \{K, L\}, \\
&\qquad \boldsymbol{\Phi}(v)_\sigma = \frac{1}{|\sigma|} \int_\sigma \frac{\Pi_{\overline{K}} v(\boldsymbol{y}) + \Pi_{\overline{L}} v(\boldsymbol{y})}{2} \mathrm{d}\gamma(\boldsymbol{y}), \\
&\forall \sigma \in \mathcal{F}_{\mathrm{ext}},\ \boldsymbol{\Phi}(v)_\sigma = 0, \\
&\forall K \in \mathcal{M}\,,\ \boldsymbol{\Phi}(v)_K = \frac{1}{|K|} \int_K \Pi_{\overline{K}} v(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}.
\end{aligned} \tag{11.20}$$

Take $\varrho \ge \theta_{\mathfrak{T}} + \eta_{\mathfrak{T}}$ (see (7.8)-(7.9)). Then, there exists $C_1 > 0$ depending only on Ω, k, d, β, p and ϱ, such that

$$\|\boldsymbol{\Phi}\|_{\mathcal{D},\mathfrak{T}} \le C_1, \tag{11.21}$$

$$\omega^{\Pi}(\mathcal{D}, \mathfrak{T}, \boldsymbol{\Phi}) \le C_1 h_{\mathcal{M}}, \tag{11.22}$$

and

$$\omega^{\nabla}(\mathcal{D}, \mathfrak{T}, \boldsymbol{\Phi}) = 0. \tag{11.23}$$

Proof. Apply first Lemma B.6 to obtain C_2 depending only on d, p and ϱ such that, for all $K \in \mathcal{M}$ and $\sigma \in \mathcal{F}_K$,

$$\left| \frac{1}{|\sigma|} \int_\sigma \Pi_{\overline{K}} v(\boldsymbol{y}) \mathrm{d}\gamma(\boldsymbol{y}) - \boldsymbol{\Phi}(v)_K \right|^p \leq \frac{C_2 h_K^{p-1}}{|\sigma|} \left\| |\nabla_{\overline{K}} v| \right\|^p_{L^p(K)^d} \tag{11.24}$$

and

$$\left\| \Pi_{\overline{K}} v - \boldsymbol{\Phi}(v)_K \right\|_{L^p(K)} \leq C_2 h_K \left\| |\nabla_{\overline{K}} v| \right\|_{L^p(K)^d}. \tag{11.25}$$

By Definition (11.5) of $[v]_{K,\sigma}$, in all cases $\sigma \in \mathcal{F}_K \cap \mathcal{F}_{\mathrm{int}}$ or $\sigma \in \mathcal{F}_K \cap \mathcal{F}_{\mathrm{ext}}$ there holds

$$\boldsymbol{\Phi}(v)_\sigma = \frac{1}{|\sigma|} \int_\sigma [v]_{K,\sigma}(\boldsymbol{y}) \mathrm{d}\gamma(\boldsymbol{y}) + \frac{1}{|\sigma|} \int_\sigma \Pi_{\overline{K}} v(\boldsymbol{y}) \mathrm{d}\gamma(\boldsymbol{y}). \tag{11.26}$$

Therefore, using (D.12), Jensen's inequality (D.10) (with $\Psi(s) = s^p$) and (11.24),

$$\begin{aligned} &|\boldsymbol{\Phi}(v)_\sigma - \boldsymbol{\Phi}(v)_K|^p \\ &\qquad \leq 2^{p-1} \left(\frac{1}{|\sigma|} \int_\sigma |[v]_{K,\sigma}(\boldsymbol{y})|^p \mathrm{d}\gamma(\boldsymbol{y}) + \frac{C_2 h_K^{p-1}}{|\sigma|} \left\| |\overline{\nabla}_K v| \right\|^p_{L^p(K)^d} \right). \end{aligned}$$

Divide this inequality by $d_{K,\sigma}^{p-1}$, multiply by $|\sigma|$, use $h_K / d_{K,\sigma} \leq \theta_{\mathfrak{T}}$, and sum over $\sigma \in \mathcal{F}_K$. Then use $\mathrm{Card}(\mathcal{F}_K) \leq \theta_{\mathfrak{T}}$ and sum over $K \in \mathcal{M}$ to see that there exists $C_3 > 0$ depending only on Ω, p and ϱ, such that

$$|\boldsymbol{\Phi}(v)|^p_{\mathfrak{T},p} \leq C_3 \|v\|^p_{\mathrm{DG},p}.$$

Conclude (11.21) by invoking Lemma 11.10.

Estimate (11.22) is obtained by raising (11.25) to the power p, by summing over $K \in \mathcal{M}$ and by using Lemma 11.10 to see that $\sum_{K \in \mathcal{M}} \|\nabla_{\overline{K}} v\|^p_{L^p(K)^d} \leq A^p \|\nabla_{\mathcal{D}} v\|^p_{L^p(\Omega)^d}$.

To prove (11.23), notice that the change of variable $\boldsymbol{x} = \boldsymbol{x}_K + s(\boldsymbol{y} - \boldsymbol{x}_K)$, (11.6a) and Stokes formula on $\nabla_{\overline{K}} v = \nabla(\Pi_{\overline{K}} v)$ give

$$\begin{aligned} \int_K \nabla_{\mathcal{D}} v(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} &= \int_K \nabla_{\overline{K}} v(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} + \sum_{\sigma \in \mathcal{F}_K} d_{K,\sigma} \int_\sigma \frac{[v]_{K,\sigma}(\boldsymbol{y})}{d_{K,\sigma}} \boldsymbol{n}_{K,\sigma} \mathrm{d}\gamma(\boldsymbol{y}) \\ &= \sum_{\sigma \in \mathcal{F}_K} \int_\sigma (\Pi_{\overline{K}} v(\boldsymbol{y}) + [v]_{K,\sigma}(\boldsymbol{y})) \boldsymbol{n}_{K,\sigma} \mathrm{d}\gamma(\boldsymbol{y}). \end{aligned}$$

Accounting for (11.26) leads to

$$\int_K \nabla_{\mathcal{D}} v(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} = \sum_{\sigma \in \mathcal{F}_K} |\sigma| \boldsymbol{\Phi}(v)_\sigma \boldsymbol{n}_{K,\sigma} = \int_K \overline{\nabla}_{\mathfrak{T}} \boldsymbol{\Phi}(v)(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}$$

and the proof of (11.23) is complete. ∎

11.2.2 Properties of DGGDs

Most of the properties of sequence of DGGDs are immediate consequences of their control by polytopal toolboxes.

Theorem 11.13 (Properties of DGGDs). *Let $(\mathcal{D}_m)_{m\in\mathbb{N}}$ be a sequence of DGGDs in the sense of Definition 11.1, defined from underlying polytopal meshes $(\mathfrak{T}_m)_{m\in\mathbb{N}}$. Assume that $(\theta_{\mathfrak{T}_m} + \eta_{\mathfrak{T}_m})_{m\in\mathbb{N}}$ is bounded (see (7.8) and (7.9)), and that $h_{\mathcal{M}_m} \to 0$ as $m \to \infty$.*
Then the sequence $(\mathcal{D}_m)_{m\in\mathbb{N}}$ is coercive, GD-consistent, limit-conforming and compact in the sense of the definitions of Section 2.1.1 in Chapter 2.

Proof. The limit-conformity, coercivity and compactness are obtained by applying Corollary 7.12, thanks to Lemma 11.12.
The GD-consistency is obtained by applying Lemma 11.14 below with $\ell = 1$, and by using Lemma 2.16 and the density of $W^{2,p}(\Omega) \cap W_0^{1,p}(\Omega)$ in $W_0^{1,p}(\Omega)$. ∎

As expected, DGGD is a high order method, which means that high order estimates should be obtained for $S_{\mathcal{D}}$ and $W_{\mathcal{D}}$. These estimates are given in the following two lemmas.

Lemma 11.14 (Estimate on $S_{\mathcal{D}}$ for DGGD). *Let $\mathcal{D}$ be a DGGD in the sense of Definition 11.1, with underlying polytopal mesh $\mathfrak{T}$. Take $\varrho \geq \theta_{\mathfrak{T}} + \eta_{\mathfrak{T}}$ (see (7.8) and (7.9)) and let $\ell \in \{1, \ldots, k\}$. Then there exists $C_4 > 0$, depending only on on Ω, β, p, k, ℓ, d and ϱ, such that*

$$\forall \varphi \in W^{\ell+1,p}(\Omega) \cap W_0^{1,p}(\Omega)\,,\ S_{\mathcal{D}}(\varphi) \leq C_4 h_{\mathcal{M}}^{\ell} \|\varphi\|_{W^{\ell+1,p}(\Omega)}\,, \tag{11.27}$$

where $S_{\mathcal{D}}$ is defined by (2.2).

Proof. In this proof, C is a generic notation for various positive numbers depending only on Ω, β, p, k, ℓ, d and ϱ. Let $\varphi \in W^{\ell+1,p}(\Omega) \cap W_0^{1,p}(\Omega)$ and, for $K \in \mathcal{M}$, denote by $\pi_K^k : L^1(K) \to \mathbb{P}_k(K)$ the $L^2(K)$-projection on polynomials over K of degree at most k. By [57, Lemmata 3.4 and 3.6],

$$\left\|\varphi - \pi_K^k \varphi\right\|_{L^p(K)} \leq C h_K^{\ell+1} \|\varphi\|_{W^{\ell+1,p}(K)}\,, \tag{11.28}$$

$$\left\|\nabla\varphi - \nabla(\pi_K^k \varphi)\right\|_{L^p(K)^d} \leq C h_K^{\ell} \|\varphi\|_{W^{\ell+1,p}(K)}\,, \tag{11.29}$$

$$\forall \sigma \in \mathcal{F}_K\,,\ \left\|\varphi - \pi_K^k \varphi\right\|_{L^p(\sigma)} \leq C h_K^{\ell+1-\frac{1}{p}} \|\varphi\|_{W^{\ell+1,p}(K)}\,. \tag{11.30}$$

The functions $(\pi_K^k \varphi)_{K\in\mathcal{M}}$ define an element of W_h. Since $\Pi_{\mathcal{D}} : X_{\mathcal{D},0} \to W_h$ is an isomorphism (see its definition (11.3)), there exists $v \in X_{\mathcal{D},0}$ such that $(\Pi_{\mathcal{D}} v)_{|K} = \Pi_{\overline{K}} v = \pi_K^k \varphi$ for all $K \in \mathcal{M}$. Then, raising (11.28) to the power p and summing over $K \in \mathcal{M}$ yields

$$\|\varphi - \Pi_{\mathcal{D}} v\|_{L^p(\Omega)} \leq C h_{\mathcal{M}}^{\ell+1} \|\varphi\|_{W^{\ell+1,p}(\Omega)}\,. \tag{11.31}$$

Let us now analyse the jump terms in $\nabla_{\mathcal{D}} v$. Let $\sigma \in \mathcal{F}_{\rm int}$ with $\mathcal{M}_\sigma = \{K, L\}$, and $\boldsymbol{y} \in \sigma$. Writing

$$\begin{aligned}[v]_{K,\sigma}(\boldsymbol{y}) &= \frac{1}{2}(\Pi_{\overline{L}} v(\boldsymbol{y}) - \varphi(\boldsymbol{y}) + \varphi(\boldsymbol{y}) - \Pi_{\overline{K}} v(\boldsymbol{y})) \\ &= \frac{1}{2}(\pi_L^k \varphi(\boldsymbol{y}) - \varphi(\boldsymbol{y})) + \frac{1}{2}(\varphi(\boldsymbol{y}) - \pi_K^k \varphi(\boldsymbol{y}))\end{aligned}$$

and using (11.30) in cells K and L yields

$$\begin{aligned}\|[v]_{K,\sigma}\|_{L^p(\sigma)} &\le C h_L^{\ell+1-\frac{1}{p}} \|\varphi\|_{W^{\ell+1,p}(L)} + C h_K^{\ell+1-\frac{1}{p}} \|\varphi\|_{W^{\ell+1,p}(K)} \\ &\le C h_{\mathcal{M}}^{\ell} \left(h_L^{1-\frac{1}{p}} \|\varphi\|_{W^{\ell+1,p}(L)} + h_K^{1-\frac{1}{p}} \|\varphi\|_{W^{\ell+1,p}(K)} \right).\end{aligned}$$

By definition of ϱ, $d_{K,\sigma} \ge \varrho^{-1} h_K$ and $d_{K,\sigma} \ge \varrho^{-1} d_{L,\sigma} \ge \varrho^{-2} h_L$, so $d_{K,\sigma}^{\frac{1}{p}-1} \le C h_K^{\frac{1}{p}-1}$ and $d_{K,\sigma}^{\frac{1}{p}-1} \le C h_L^{\frac{1}{p}-1}$. Hence,

$$d_{K,\sigma}^{\frac{1}{p}-1} \|[v]_{K,\sigma}\|_{L^p(\sigma)} \le C h_{\mathcal{M}}^{\ell} \left(\|\varphi\|_{W^{\ell+1,p}(L)} + \|\varphi\|_{W^{\ell+1,p}(K)} \right).$$

Using the same change of variable $\boldsymbol{x} = \boldsymbol{x}_K + s(\boldsymbol{y} - \boldsymbol{x}_K)$ as in the proof of Lemma 11.10, we infer

$$\begin{aligned}&\int_{D_{K,\sigma}} |\nabla_{\mathcal{D}} v(\boldsymbol{x}) - \nabla_{\overline{K}} v(\boldsymbol{x})|^p \mathrm{d}\boldsymbol{x} \\ &\qquad = \int_{D_{K,\sigma}} \left| \psi(s(\boldsymbol{x})) \frac{[v]_{K,\sigma}(\boldsymbol{y}(\boldsymbol{x}))}{d_{K,\sigma}} \right|^p \mathrm{d}\boldsymbol{x} \\ &\qquad = \int_\beta^1 |\psi(s)|^p s^{d-1} \mathrm{d}s \int_\sigma d_{K,\sigma}^{1-p} |[v]_{K,\sigma}(\boldsymbol{y})|^p \mathrm{d}\gamma(\boldsymbol{y}) \\ &\qquad \le C h_{\mathcal{M}}^{p\ell} \left(\|\varphi\|_{W^{\ell+1,p}(L)}^p + \|\varphi\|_{W^{\ell+1,p}(K)}^p \right). \qquad (11.32)\end{aligned}$$

Since $\varphi = 0$ on $\partial\Omega$, performing the same steps as above shows that (11.32) also holds if $\sigma \in \mathcal{F}_K \cap \mathcal{F}_{\rm ext}$, by simply removing the term involving L. Sum (11.32) over $\sigma \in \mathcal{F}_K$ and $K \in \mathcal{M}$, recall that $\nabla_{\overline{K}} v = \nabla(\Pi_{\overline{K}} v) = \nabla(\pi_K^k \varphi)$, and use then (11.29) and the triangle inequality to infer

$$\|\nabla_{\mathcal{D}} v - \nabla\varphi\|_{L^p(\Omega)^d} \le C h_{\mathcal{M}}^{\ell} \|\varphi\|_{W^{\ell+1,p}(\Omega)}.$$

Combined with (11.31), this completes the proof of (11.27). ∎

Lemma 11.15 (Estimate on $W_{\mathcal{D}}(\boldsymbol{\varphi})$ for DGGD). *Let $\mathcal{D}$ be a DGGD in the sense of Definition 11.1, with underlying polytopal mesh $\mathfrak{T}$. Take $\varrho \ge \theta_{\mathfrak{T}} + \eta_{\mathfrak{T}}$ (see (7.8) and (7.9)) and let $\ell \in \{1, \ldots, k\}$. Then there exists C_5, depending only on Ω, β, p, k, ℓ, d and ϱ, such that*

$$\forall \boldsymbol{\varphi} \in W^{\ell,p'}(\Omega)^d,\ W_{\mathcal{D}}(\boldsymbol{\varphi}) \le C_5 h_{\mathcal{M}}^{\ell} \|\boldsymbol{\varphi}\|_{W^{\ell,p'}(\Omega)^d}, \qquad (11.33)$$

where $W_{\mathcal{D}}$ is defined by (2.6).

Remark 11.16 (Estimate on the space size of a DGGD). Lemmata 11.14 and 11.15 give the following estimate on the space size of a DGGD (as in Definition 2.22):

$$h_{\mathcal{D}}(W^{\ell+1,p}(\Omega)\cap W_0^{1,p}(\Omega);W^{\ell,p'}(\Omega)^d)\le \max(C_4,C_5)h_{\mathcal{M}}^{\ell}.$$

Proof. In this proof, C denotes various constants having the same dependencies as C_5 in the lemma. Let $\boldsymbol{\varphi}\in W^{\ell,p'}(\Omega)^d$. Using the definition of $\nabla_{\mathcal{D}}v$ and $\Pi_{\mathcal{D}}v$ yields

$$\begin{aligned}
&\int_\Omega (\nabla_{\mathcal{D}}v(\boldsymbol{x})\cdot\boldsymbol{\varphi}(\boldsymbol{x})+\Pi_{\mathcal{D}}v(\boldsymbol{x})\mathrm{div}\boldsymbol{\varphi}(\boldsymbol{x}))\,\mathrm{d}\boldsymbol{x} && (11.34)\\
&=\sum_{K\in\mathcal{M}}\int_K(\nabla(\Pi_{\overline{K}}v)(\boldsymbol{x})\cdot\boldsymbol{\varphi}(\boldsymbol{x})+\Pi_{\overline{K}}v(\boldsymbol{x})\mathrm{div}\boldsymbol{\varphi}(\boldsymbol{x}))\,\mathrm{d}\boldsymbol{x}\\
&\quad+\sum_{K\in\mathcal{M}}\sum_{\sigma\in\mathcal{F}_K}\int_{D_{K,\sigma}}\psi(s(\boldsymbol{x}))\frac{[v]_{K,\sigma}(\boldsymbol{y}(\boldsymbol{x}))}{d_{K,\sigma}}\boldsymbol{n}_{K,\sigma}\cdot\boldsymbol{\varphi}(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\\
&=T_1+T_2. && (11.35)
\end{aligned}$$

Stokes formula in each cell K yields

$$T_1=\sum_{K\in\mathcal{M}}\sum_{\sigma\in\mathcal{F}_K}\int_\sigma \Pi_{\overline{K}}v(\boldsymbol{y})\boldsymbol{\varphi}(\boldsymbol{y})\cdot\boldsymbol{n}_{K,\sigma}\mathrm{d}\gamma(\boldsymbol{y}). \tag{11.36}$$

Let $\pi_K^{k-1}:L^p(K)^d\to\mathbb{P}_{k-1}(K)^d$ be the component-wise $L^2(K)$-projection over polynomial vectors on K of degree at most $k-1$, and write

$$\begin{aligned}
T_2&=\sum_{K\in\mathcal{M}}\sum_{\sigma\in\mathcal{F}_K}\int_{D_{K,\sigma}}\psi(s(\boldsymbol{x}))\frac{[v]_{K,\sigma}(\boldsymbol{y}(\boldsymbol{x}))}{d_{K,\sigma}}\boldsymbol{n}_{K,\sigma}\cdot(\boldsymbol{\varphi}(\boldsymbol{x})-\pi_K^{k-1}\boldsymbol{\varphi}(\boldsymbol{x}))\mathrm{d}\boldsymbol{x}\\
&\quad+\sum_{K\in\mathcal{M}}\sum_{\sigma\in\mathcal{F}_K}\int_{D_{K,\sigma}}\psi(s(\boldsymbol{x}))\frac{[v]_{K,\sigma}(\boldsymbol{y}(\boldsymbol{x}))}{d_{K,\sigma}}\boldsymbol{n}_{K,\sigma}\cdot\pi_K^{k-1}\boldsymbol{\varphi}(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\\
&=T_{2,1}+T_{2,2}. && (11.37)
\end{aligned}$$

Similarly to (11.28) we have $\left\|\boldsymbol{\varphi}-\pi_K^{k-1}\boldsymbol{\varphi}\right\|_{L^{p'}(K)^d}\le Ch_K^{\ell}\left\|\boldsymbol{\varphi}\right\|_{W^{\ell,p'}(K)^d}$. Hence, the Hölder inequalities (D.1), (D.5) and the now classical change of variable $\boldsymbol{x}=\boldsymbol{x}_K+s(\boldsymbol{y}-\boldsymbol{x}_K)$ show that

$$\begin{aligned}
|T_{2,1}|&\le Ch_{\mathcal{M}}^{\ell}\left\|\boldsymbol{\varphi}\right\|_{W^{\ell,p'}(\Omega)^d}\\
&\qquad\times\left(\sum_{K\in\mathcal{M}}\sum_{\sigma\in\mathcal{F}_K}\int_{D_{K,\sigma}}|\psi(s)|^p s^{d-1}\mathrm{d}s\int_\sigma d_{K,\sigma}^{1-p}|[v]_{K,\sigma}(\boldsymbol{y})|^p\mathrm{d}\gamma(\boldsymbol{y})\right)^{1/p}\\
&\le Ch_{\mathcal{M}}^{\ell}\left\|\boldsymbol{\varphi}\right\|_{W^{\ell,p'}(\Omega)^d}\left\|v\right\|_{\mathrm{DG},p}\\
&\le CAh_{\mathcal{M}}^{\ell}\left\|\boldsymbol{\varphi}\right\|_{W^{\ell,p'}(\Omega)^d}\left\|\nabla_{\mathcal{D}}v\right\|_{L^p(\Omega)^d}. && (11.38)
\end{aligned}$$

Lemma 11.10 was invoked in the last line. We now turn to $T_{2,2}$. Since $\pi_K^{k-1}\boldsymbol{\varphi}$ is a polynomial of degree $k-1$ or less, using the change of variable $\boldsymbol{x} = \boldsymbol{x}_K + s(\boldsymbol{y} - \boldsymbol{x}_K)$ and a Taylor expansion in s about $s = 1$ we have, similarly to (11.11),

$$\pi_K^{k-1}\boldsymbol{\varphi}(\boldsymbol{x}) \cdot \boldsymbol{n}_{K,\sigma} = \pi_K^{k-1}\boldsymbol{\varphi}(\boldsymbol{y}) \cdot \boldsymbol{n}_{K,\sigma} + \sum_{m=1}^{k-1} p_m(\boldsymbol{y})(1-s)^m$$

where p_m is a polynomial in $\boldsymbol{y}$. Hence, by (11.6),

$$T_{2,2} = \sum_{K\in\mathcal{M}} \sum_{\sigma\in\mathcal{F}_K} \int_\sigma [v]_{K,\sigma}(\boldsymbol{y}) \pi_K^{k-1}\boldsymbol{\varphi}(\boldsymbol{y}) \cdot \boldsymbol{n}_{K,\sigma} \mathrm{d}\boldsymbol{y}.$$

Therefore, with (11.36)

$$\begin{aligned} T_1 + T_{2,2} &= \sum_{K\in\mathcal{M}} \sum_{\sigma\in\mathcal{F}_K} \int_\sigma \left(\Pi_{\overline{K}} v(\boldsymbol{y})\boldsymbol{\varphi}(\boldsymbol{y}) + [v]_{K,\sigma}(\boldsymbol{y})\pi_K^{k-1}\boldsymbol{\varphi}(\boldsymbol{y})\right) \cdot \boldsymbol{n}_{K,\sigma} \mathrm{d}\gamma(\boldsymbol{y}) \\ &= \sum_{K\in\mathcal{M}} \sum_{\sigma\in\mathcal{F}_K} \int_\sigma \left(\Pi_{\overline{K}} v(\boldsymbol{y}) + [v]_{K,\sigma}(\boldsymbol{y})\right) \boldsymbol{\varphi}(\boldsymbol{y}) \cdot \boldsymbol{n}_{K,\sigma} \mathrm{d}\gamma(\boldsymbol{y}) \qquad (11.39) \\ &\quad + \sum_{K\in\mathcal{M}} \sum_{\sigma\in\mathcal{F}_K} \int_\sigma [v]_{K,\sigma}(\boldsymbol{y})(\pi_K^{k-1}\boldsymbol{\varphi}(\boldsymbol{y}) - \boldsymbol{\varphi}(\boldsymbol{y})) \cdot \boldsymbol{n}_{K,\sigma} \mathrm{d}\gamma(\boldsymbol{y}). \end{aligned}$$

If $\sigma \in \mathcal{F}_{\mathrm{int}}$ with $\mathcal{M}_\sigma = \{K, L\}$, $\Pi_{\overline{K}} v(\boldsymbol{y}) + [v]_{K,\sigma}(\boldsymbol{y}) = \frac{1}{2}(\Pi_{\overline{K}} v(\boldsymbol{y}) + \Pi_{\overline{L}} v(\boldsymbol{y}))$ and thus, since $\boldsymbol{n}_{K,\sigma} + \boldsymbol{n}_{L,\sigma} = 0$,

$$\begin{aligned} &\int_\sigma \left(\Pi_{\overline{K}} v(\boldsymbol{y}) + [v]_{K,\sigma}(\boldsymbol{y})\right) \boldsymbol{\varphi}(\boldsymbol{y}) \cdot \boldsymbol{n}_{K,\sigma} \mathrm{d}\gamma(\boldsymbol{y}) \\ &\qquad + \int_\sigma \left(\Pi_{\overline{L}} v(\boldsymbol{y}) + [v]_{L,\sigma}(\boldsymbol{y})\right) \boldsymbol{\varphi}(\boldsymbol{y}) \cdot \boldsymbol{n}_{L,\sigma} \mathrm{d}\gamma(\boldsymbol{y}) = 0. \end{aligned}$$

If $\sigma \in \mathcal{F}_{\mathrm{ext}}$ with $\mathcal{M}_\sigma = \{K\}$, then $\Pi_{\overline{K}} v(\boldsymbol{y}) + [v]_{K,\sigma}(\boldsymbol{y}) = 0$. These arguments show that the term (11.39) vanishes, and thus that

$$T_1 + T_{2,2} = \sum_{K\in\mathcal{M}} \sum_{\sigma\in\mathcal{F}_K} \int_\sigma [v]_{K,\sigma}(\boldsymbol{y})(\pi_K^{k-1}\boldsymbol{\varphi}(\boldsymbol{y}) - \boldsymbol{\varphi}(\boldsymbol{y})) \cdot \boldsymbol{n}_{K,\sigma} \mathrm{d}\gamma(\boldsymbol{y}).$$

Similarly to (11.30), we have $\|\boldsymbol{\varphi} - \pi_K^{k-1}\boldsymbol{\varphi}\|_{L^{p'}(\sigma)^d} \le C h_K^{\ell-1/p'} \|\boldsymbol{\varphi}\|_{W^{\ell,p'}(K)^d} \le C h_{\mathcal{M}}^\ell d_{K,\sigma}^{-1/p'} \|\boldsymbol{\varphi}\|_{W^{\ell,p'}(K)^d}$ (use $d_{K,\sigma} \le h_K$). The continuous and discrete Hölder inequalities (D.5) and (D.2) (the latter with parameters $d_i = d_{K,\sigma}^{\frac{1}{p}-1} = d_{K,\sigma}^{-1/p'}$) thus give

$$|T_1 + T_{2,2}| \le C \left(\sum_{K\in\mathcal{M}} \sum_{\sigma\in\mathcal{F}_K} \int_\sigma d_{K,\sigma}^{1-p} |[v]_{K,\sigma}(\boldsymbol{y})|^p \mathrm{d}\gamma(\boldsymbol{y}) \right)^{1/p} h_{\mathcal{M}}^\ell \|\boldsymbol{\varphi}\|_{W^{\ell,p'}(\Omega)^d}$$

$$\leq C \|v\|_{\mathrm{DG},p} h_{\mathcal{M}}^{\ell} \|\boldsymbol{\varphi}\|_{W^{\ell,p'}(\Omega)^d}$$
$$\leq CAh_{\mathcal{M}}^{\ell} \|\nabla_{\mathcal{D}} v\|_{L^p(\Omega)^d} \|\boldsymbol{\varphi}\|_{W^{\ell,p'}(\Omega)^d}.$$

Combined with (11.38) and plugged alongside (11.37) into (11.35), this concludes the proof of (11.33). ∎

11.3 Average DG gradient discretisation

This section presents an alternative gradient discretisation associated with discontinuous Galerkin methods. This GD, called the Average Discontinuous Galerkin Gradient Discretisation (ADGGD), leads to simpler computations, as shown in Chapter E in the case $k = 1$. Let $X_{\mathcal{D},0}$ and $\Pi_{\mathcal{D}}$ be defined as in Section 11.1. We substitute to (11.4) the following definition: for $v \in X_{\mathcal{D},0}$, $K \in \mathcal{M}$, $\sigma \in \mathcal{F}_K$ and a.e. $\boldsymbol{x} \in D_{K,\sigma}$, set

$$\nabla_{\mathcal{D}} v(\boldsymbol{x}) = \nabla_{\overline{K}} v(\boldsymbol{x}) + \psi(\boldsymbol{x}) \frac{[v]_{K,\sigma}^a}{d_{K,\sigma}} \boldsymbol{n}_{K,\sigma}, \tag{11.40}$$

where $\psi(\boldsymbol{x}) = 0$ for $\boldsymbol{x} \in D_{K,\sigma}^{(\beta)}$, $\psi(\boldsymbol{x}) = \frac{d}{1-\beta^d}$ for $\boldsymbol{x} \in D_{K,\sigma} \setminus D_{K,\sigma}^{(\beta)}$ and the average jump $[v]_{K,\sigma}^a$, which replaces the pointwise jump in (11.4), is defined by

$$\begin{aligned} &\text{if } \sigma \in \mathcal{F}_{\mathrm{int}} \text{ and } \mathcal{M}_\sigma = \{K, L\}, \\ &\qquad [v]_{K,\sigma}^a = \frac{1}{|\sigma|} \int_\sigma \frac{\Pi_{\overline{L}} v(\boldsymbol{y}) - \Pi_{\overline{K}} v(\boldsymbol{y})}{2} \mathrm{d}\gamma(\boldsymbol{y}), \\ &\text{if } \sigma \in \mathcal{F}_{\mathrm{ext}} \text{ and } \mathcal{M}_\sigma = \{K\}, \\ &\qquad [v]_{K,\sigma}^a = \frac{1}{|\sigma|} \int_\sigma (0 - \Pi_{\overline{K}} v(\boldsymbol{y})) \mathrm{d}\gamma(\boldsymbol{y}). \end{aligned} \tag{11.41}$$

The same analysis as in the previous sections can be carried out, following the items below.

- Lemma 11.10 holds for the ADGGD, replacing $\|v\|_{\mathrm{DG},p}$ by $\|v\|_{\mathrm{ADG},p}$ defined by

$$\|v\|_{\mathrm{ADG},p}^p = \sum_{K \in \mathcal{M}} \left(\int_K |\nabla_{\overline{K}} v(\boldsymbol{x})|^p \mathrm{d}\boldsymbol{x} + \sum_{\sigma \in \mathcal{F}_K} \frac{|\sigma|}{d_{K,\sigma}^{p-1}} \left| [v]_{K,\sigma}^a \right|^p \right). \tag{11.42}$$

- The conclusions of Lemma 11.12 hold, with the same definition of the control $\boldsymbol{\Phi} : X_{\mathcal{D},0} \to X_{\mathfrak{T},0}$ of $\mathcal{D}$ by $\mathfrak{T}$. Hence, Theorem 11.13 is also verified by ADGGDs, by noticing that Lemma 11.14 holds and by replacing the pointwise jump by the average jump.

- The proof of Lemma 11.15 cannot be immediately adapted since the term (11.39) does no longer vanish.

The idea of using jumps of mean values (instead of jumps of functions) appears in [114], and in [31] with a non-symmetric scheme. Following the same steps as in the proof of Theorem 11.6, we check that using the ADGGD in the GS (2.23) consists in applying to the linear diffusion model (2.22) the method described in [114] on a model of linear elasticity.

12

The multi-point flux approximation MPFA-O scheme

The two-point flux approximation (TPFA) method was introduced in Section 1.1.3. This scheme has been widely used in industry because of its simplicity in the case of scalar diffusion operators, since it leads (in 2D) to a 5-point approximation for the Laplace operator, with no face unknowns. As we saw 1.1.3 the TPFA-CG scheme is a GDM. The GDM version can then be written with any full diffusion matrices, with the drawback that the face unknowns can no longer be eliminated from the flux conservation equations at the faces. The multi-point flux approximation-O [2] scheme mitigates this drawback, but leads to a symmetric definite positive matrix only on certain meshes. The aim of this chapter is to show that, in two such particular cases of meshes, the MPFA O-scheme is a GDM.

12.1 MPFA methods, Dirichlet boundary conditions

12.1.1 Definition of the MPFA gradient discretisation

We consider the MPFA-O scheme on particular polytopal meshes $\mathfrak{T} = (\mathcal{M}, \mathcal{F}, \mathcal{P}, \mathcal{V})$ of Ω: Cartesian (each $K \in \mathcal{M}$ is a parallelepipedic polyhedron with faces parallel to the axes), or simplicial (in the sense of Definition 7.4). In each of these cases, $\mathcal{P}$ is the set of centres of mass of the cells. We define a partition $(V_{K,\boldsymbol{s}})_{\boldsymbol{s}\in\mathcal{V}_K}$ of each $K \in \mathcal{M}$ the following way (see Figure 12.1):

- *Cartesian meshes*: $V_{K,\boldsymbol{s}}$ is the parallelepipedic polyhedron whose faces are parallel to the faces of K, and that has $\boldsymbol{x}_K$ and $\boldsymbol{s}$ as vertices. For $\sigma \in \mathcal{F}$ and $\boldsymbol{s} \in \mathcal{V}_\sigma$, we let $\boldsymbol{x}_{\sigma,\boldsymbol{s}} = \overline{\boldsymbol{x}}_\sigma$. We have $\mathrm{Card}(\mathcal{V}_K) = 2^d$ and $\mathrm{Card}(\mathcal{V}_\sigma) = 2^{d-1}$.
- *Simplicial mesh*: We denote by $(\beta^K_{\boldsymbol{s}}(\boldsymbol{x}))_{\boldsymbol{s}\in\mathcal{V}_K}$ the barycentric coordinates of $\boldsymbol{x}$ in K, that is,

$$\boldsymbol{x} - \boldsymbol{x}_K = \sum_{\boldsymbol{s}\in\mathcal{V}_K} \beta^K_{\boldsymbol{s}}(\boldsymbol{x})(\boldsymbol{s} - \boldsymbol{x}_K) \text{ with } \beta^K_{\boldsymbol{s}}(\boldsymbol{x}) \geq 0 \text{ and } \sum_{\boldsymbol{s}\in\mathcal{V}_K} \beta^K_{\boldsymbol{s}}(\boldsymbol{x}) = 1.$$

The set $V_{K,\boldsymbol{s}}$ is made of the points $\boldsymbol{x} \in K$ whose barycentric coordinates $(\beta^K_{\boldsymbol{s}'}(\boldsymbol{x}))_{\boldsymbol{s}' \in \mathcal{V}_K}$ satisfy $\beta^K_{\boldsymbol{s}}(\boldsymbol{x}) > \beta^K_{\boldsymbol{s}'}(\boldsymbol{x})$ for all $\boldsymbol{s}' \in \mathcal{V}_K \setminus \{\boldsymbol{s}\}$. For $\sigma \in \mathcal{F}$ and $\boldsymbol{s} \in \mathcal{V}_\sigma$, $\boldsymbol{x}_{\sigma,\boldsymbol{s}}$ is the point of σ whose barycentric coordinates in σ are $\beta^\sigma_{\boldsymbol{s}'}(\boldsymbol{x}_{\sigma,\boldsymbol{s}}) = 1/(d+1)$ for all $\boldsymbol{s}' \in \mathcal{V}_\sigma \setminus \{\boldsymbol{s}\}$, and $\beta^\sigma_{\boldsymbol{s}}(\boldsymbol{x}_{\sigma,\boldsymbol{s}}) = 2/(d+1)$. Then, denoting by $\overline{\boldsymbol{s}}$ the vertex opposed to σ in K, the barycentric coordinates in K of $\boldsymbol{x}_{\sigma,\boldsymbol{s}}$ are given by $\beta^K_{\boldsymbol{s}'}(\boldsymbol{x}_{\sigma,\boldsymbol{s}}) = 1/(d+1)$ for all $\boldsymbol{s}' \in \mathcal{V}_\sigma \setminus \{\boldsymbol{s}\}$, $\beta^K_{\boldsymbol{s}}(\boldsymbol{x}_{\sigma,\boldsymbol{s}}) = 2/(d+1)$ and $\beta^K_{\overline{\boldsymbol{s}}}(\boldsymbol{x}_{\sigma,\boldsymbol{s}}) = 0$. We have $\mathrm{Card}(\mathcal{V}_K) = d+1$ and $\mathrm{Card}(\mathcal{V}_\sigma) = d$.

In both cases, we denote by $\mathcal{F}_{K,\boldsymbol{s}}$ the set of all elements $\sigma \in \mathcal{F}_K$ such that $\boldsymbol{s} \in \mathcal{V}_\sigma$, and we denote by $\tau_{\sigma,\boldsymbol{s}}$ the external face of $V_{K,\boldsymbol{s}}$ defined by

$$\tau_{\sigma,\boldsymbol{s}} = \overline{V_{K,\boldsymbol{s}}} \cap \sigma.$$

Observe that

$$|V_{K,\boldsymbol{s}}| = \frac{|K|}{\mathrm{Card}(\mathcal{V}_K)} \quad \text{and} \quad |\tau_{\sigma,\boldsymbol{s}}| = \frac{|\sigma|}{\mathrm{Card}(\mathcal{V}_\sigma)}. \tag{12.1}$$

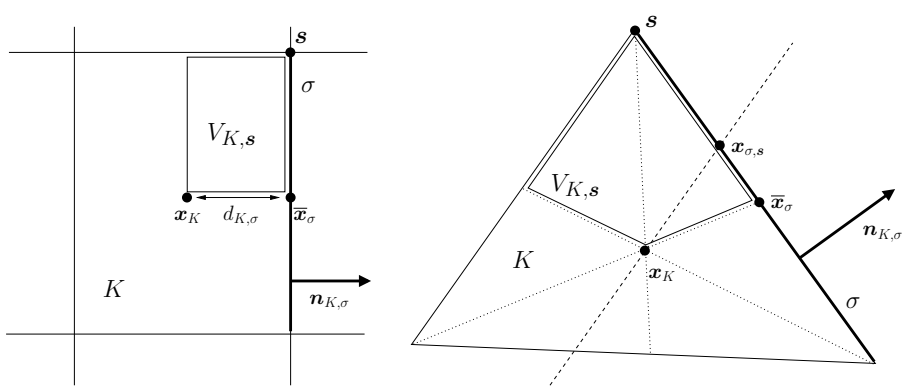


Fig. 12.1. Notations for MPFA-O schemes defined on Cartesian (left) and simplicial (right) meshes.

We follow the notations in Definition 7.33 to construct the MPFA-O LLE GD in both cases:

1. The set of geometrical entities attached to the discrete unknowns is $I = \mathcal{M} \cup \{\tau_{\sigma,\boldsymbol{s}} : \sigma \in \mathcal{F},\ \boldsymbol{s} \in \mathcal{V}_\sigma\}$ and the family of approximation points is $S = ((\boldsymbol{x}_K)_{K\in\mathcal{M}}, (\boldsymbol{x}_{\sigma,\boldsymbol{s}})_{\sigma\in\mathcal{F},\, \boldsymbol{s}\in\mathcal{V}_\sigma})$. We define $I_\Omega = \mathcal{M} \cup \{\tau_{\sigma,\boldsymbol{s}} : \sigma \in \mathcal{F}_{\mathrm{int}},\ \boldsymbol{s} \in \mathcal{V}_\sigma\}$ and $I_\partial = \{\tau_{\sigma,\boldsymbol{s}} : \sigma \in \mathcal{F}_{\mathrm{ext}},\ \boldsymbol{s} \in \mathcal{V}_\sigma\}$. This gives, with a slight abuse of notation (we should write $v_{\tau_{\sigma,\boldsymbol{s}}}$ instead of $v_{\sigma,\boldsymbol{s}}$),

$$X_{\mathcal{D},0} = \{v = ((v_K)_{K\in\mathcal{M}}, (v_{\sigma,\boldsymbol{s}})_{\sigma\in\mathcal{F},\, \boldsymbol{s}\in\mathcal{V}_\sigma}) : \\ v_K \in \mathbb{R} \text{ for all } K \in \mathcal{M},\ v_{\sigma,\boldsymbol{s}} \in \mathbb{R} \text{ for all } \sigma \in \mathcal{F}_{\mathrm{int}} \text{ and } \boldsymbol{s} \in \mathcal{V}_\sigma,$$

$v_{\sigma,\boldsymbol{s}} = 0$ for all $\sigma \in \mathcal{F}_{\text{ext}}$ and $\boldsymbol{s} \in \mathcal{V}_\sigma\}$.

For any $K \in \mathcal{M}$, we set $I_K = \{K\} \cup \{\tau_{\sigma,\boldsymbol{s}} : \sigma \in \mathcal{F}_K,\ \boldsymbol{s} \in \mathcal{V}_\sigma\}$.

2. The functions $\pi_K = (\pi_K^i)_{i \in I_K}$ of $L^p(K)$ are defined by

$$\pi_K^i = 1 \text{ for } i = K, \text{ and } \pi_K^i = 0 \text{ for } i = \tau_{\sigma,\boldsymbol{s}}, \tag{12.2}$$

which means that

$$\forall v \in X_{\mathcal{D},0}\,,\ \forall K \in \mathcal{M}\,,\ \forall \boldsymbol{x} \in K\,,\ \Pi_{\mathcal{D}} v(\boldsymbol{x}) = v_K. \tag{12.3}$$

3. The functions $\mathcal{G}_K = (\mathcal{G}_K^i)_{i \in I_K}$ of $L^p(K)^d$ are defined by: for all $\boldsymbol{s} \in \mathcal{V}_K$ and a.e. $\boldsymbol{x} \in V_{K,\boldsymbol{s}}$,

$$\begin{aligned}
&\mathcal{G}_K^K(\boldsymbol{x}) = -\frac{1}{|V_{K,\boldsymbol{s}}|} \sum_{\sigma \in \mathcal{V}_{K,\boldsymbol{s}}} |\tau_{\sigma,\boldsymbol{s}}| \boldsymbol{n}_{K,\sigma}, \\
&\forall \sigma \in \mathcal{F}_{K,\boldsymbol{s}}\,,\ \mathcal{G}_K^{\sigma,\boldsymbol{s}}(\boldsymbol{x}) = \frac{1}{|V_{K,\boldsymbol{s}}|} |\tau_{\sigma,\boldsymbol{s}}| \boldsymbol{n}_{K,\sigma}\,, \\
&\forall \sigma \in \mathcal{F}_{K,\boldsymbol{s}}\,,\ \mathcal{G}_K^{\sigma,\boldsymbol{s}} = 0 \text{ on } K \text{ outside } V_{K,\boldsymbol{s}}.
\end{aligned} \tag{12.4}$$

Hence, for all $v \in X_{\mathcal{D},0}$,

$$\begin{aligned}
&\forall K \in \mathcal{M}\,,\ \forall \boldsymbol{s} \in \mathcal{V}_K\,,\ \text{for a.e. } \boldsymbol{x} \in V_{K,\boldsymbol{s}}\,, \\
&\nabla_{\mathcal{D}} v(\boldsymbol{x}) = \frac{1}{|V_{K,\boldsymbol{s}}|} \sum_{\sigma \in \mathcal{F}_{K,\boldsymbol{s}}} |\tau_{\sigma,\boldsymbol{s}}| (v_{\sigma,\boldsymbol{s}} - v_K) \boldsymbol{n}_{K,\sigma}.
\end{aligned} \tag{12.5}$$

4. The exactness of the reconstructions π_K and $\mathcal{G}_K$, as well as the fact that $\|\nabla_{\mathcal{D}} \cdot\|_{L^p(\Omega)^d}$ is a norm on $X_{\mathcal{D},0}$, are proved in Lemma 12.3 below.

Remark 12.1 (Identical approximation points). Note that, in the case of a Cartesian mesh, for a given $\sigma \in \mathcal{F}$ all the approximation points $(\boldsymbol{x}_{\sigma,\boldsymbol{s}})_{\boldsymbol{s} \in \mathcal{V}_\sigma}$ are identical. This is allowed in the definition of an LLE GD, see Definition 7.33

For such a GD, the GS (2.23) is a finite volume scheme. Indeed, by selecting a test function with only non-zero value $v_K = 1$ in (2.23), we obtain the flux balance

$$\begin{aligned}
&\sum_{\sigma \in \mathcal{F}_K} \sum_{\boldsymbol{s} \in \mathcal{V}_\sigma} F_{K,\sigma,\boldsymbol{s}}(u) = \int_K f(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}, \\
&\text{where } F_{K,\sigma,\boldsymbol{s}}(u) = \int_{\sigma_{\boldsymbol{s}}} \mathcal{G}_K u(\boldsymbol{x}) \cdot \boldsymbol{n}_{K,\sigma} \mathrm{d}\gamma(\boldsymbol{x}).
\end{aligned} \tag{12.6}$$

Selecting a test function with only non-zero value $v_{\sigma,\boldsymbol{s}} = 1$ in (2.23) leads to the conservativity of the fluxes:

$$\begin{aligned}
&F_{K,\sigma,\boldsymbol{s}}(u) + F_{L,\sigma,\boldsymbol{s}}(u) = 0 \\
&\text{for all } \sigma \in \mathcal{F}_{\text{int}} \text{ with } \mathcal{M}_\sigma = \{K, L\}, \text{ and all } \boldsymbol{s} \in \mathcal{V}_\sigma.
\end{aligned} \tag{12.7}$$

For a given $\boldsymbol{s} \in \mathcal{V}$, the unknowns $(u_{\sigma,\boldsymbol{s}})_{\sigma|\boldsymbol{s}\in\mathcal{V}_\sigma}$ can be locally expressed in terms of $(u_K)_{K|\boldsymbol{s}\in\mathcal{V}_K}$. This is done by solving the local linear system issued from (12.7) written for all σ such that $\boldsymbol{s} \in \mathcal{V}_\sigma$. After these local eliminations of $u_{\sigma,\boldsymbol{s}}$, the resulting linear system only involves the cell unknowns. This discretisation of (2.20) obtained by writing the balance and conservativity of half-fluxes $F_{K,\sigma,\boldsymbol{s}}$, constructed via $\mathbb{P}_1$-exact gradients reconstructions, is identical to the construction of the MPFA-O method in [2]. The GD constructed above therefore gives indeed the MPFA-O scheme when used in the GS (2.23).

Remark 12.2 (Other meshes)
The identification of MPFA-O schemes as GSs is, to our knowledge, restricted to the two cases considered here (Cartesian and simplicial meshes). In the case of more general meshes for the approximation of (2.20), the gradient reconstruction defined by the MPFA-O scheme can be used in the finite volume scheme (12.6)–(12.7); however, the GS (2.23) built upon this gradient reconstruction cannot be expected to always converge, since the corresponding GD may fail to be limit-conforming and coercive.

12.1.2 Preliminary lemmas

Let us first prove that the GD constructed above is indeed an LLE GD, and let us estimate its regularity.

Lemma 12.3 (Estimate on $\mathrm{reg}_{\mathbf{LLE}}$, MPFA-O). *Let $\mathfrak{T}$ be a polytopal mesh in the sense of Definition 7.2, which is either Cartesian or simplicial. For $K \in \mathcal{M}$, let $\pi_K = (\pi_K^i)_{i\in I_K}$ be defined by (12.2), and $\mathcal{G}_K = (\mathcal{G}_K^i)_{i\in I_K}$ be defined by (12.4). Then π_K is a $\mathbb{P}_0$-exact function reconstruction on K, $\mathcal{G}_K$ is a $\mathbb{P}_1$-exact gradient reconstruction on K upon $(\boldsymbol{x}_K, (\boldsymbol{x}_{\sigma,\boldsymbol{s}})_{\sigma\in\mathcal{F}_K,\,\boldsymbol{s}\in\mathcal{V}_\sigma})$, and*

$$\forall \xi = (\xi_K, (\xi_{\sigma,\boldsymbol{s}})_{\sigma\in\mathcal{F}_K,\,\boldsymbol{s}\in\mathcal{V}_\sigma}),\ (\mathcal{G}_K\xi)_{|V_{K,\boldsymbol{s}}} \cdot (\boldsymbol{x}_{\sigma,\boldsymbol{s}} - \boldsymbol{x}_K) = \xi_{\sigma,\boldsymbol{s}} - \xi_K. \quad (12.8)$$

Moreover, $\mathcal{D}$ is an LLE GE and there exists C_1, depending only on d and $\theta \geq \theta_{\mathfrak{T}}$ (see (7.8)), such that

$$\mathrm{reg}_{\mathrm{LLE}}(\mathcal{D}) \leq C_1. \quad (12.9)$$

Proof.
Step 1: properties of π_K and $\mathcal{G}_K$.
We have $\sum_{i\in I_K} \pi_K^i = \pi_K^K = 1$ so π_K is a $\mathbb{P}_0$-exact function reconstruction. Let $\boldsymbol{s} \in \mathcal{V}_K$ and assume that we can prove the following two properties:

$$\forall \widetilde{\sigma} \in \mathcal{F}_{K,\boldsymbol{s}} \backslash \{\sigma\},\ (\boldsymbol{x}_{\sigma,\boldsymbol{s}} - \boldsymbol{x}_K) \perp \boldsymbol{n}_{K,\widetilde{\sigma}}, \text{ and} \quad (12.10)$$

$$\frac{1}{|V_{K,\boldsymbol{s}}|} |\tau_{\sigma,\boldsymbol{s}}| \boldsymbol{n}_{K,\sigma} \cdot (\boldsymbol{x}_{\sigma,\boldsymbol{s}} - \boldsymbol{x}_K) = 1. \quad (12.11)$$

Then the expression

$$(\mathcal{G}_K\xi)_{|V_{K,\boldsymbol{s}}} = \frac{1}{|V_{K,\boldsymbol{s}}|}\sum_{\sigma\in\mathcal{F}_{K,\boldsymbol{s}}} |\tau_{\sigma,\boldsymbol{s}}|(\xi_{\sigma,\boldsymbol{s}} - \xi_K)\boldsymbol{n}_{K,\sigma}$$

shows that

$$\begin{aligned}(\mathcal{G}_K\xi)_{|V_{K,\boldsymbol{s}}}\cdot(\boldsymbol{x}_{\sigma,\boldsymbol{s}} - \boldsymbol{x}_K) &= \frac{1}{|V_{K,\boldsymbol{s}}|}|\tau_{\sigma,\boldsymbol{s}}|(\xi_{\sigma,\boldsymbol{s}} - \xi_K)\boldsymbol{n}_{K,\sigma}\cdot(\boldsymbol{x}_{\sigma,\boldsymbol{s}} - \boldsymbol{x}_K)\\ &= \xi_{\sigma,\boldsymbol{s}} - \xi_K,\end{aligned}$$

which proves (12.8). Take an affine function A and apply this relation to $\xi = (A(\boldsymbol{x}_K), (A(\boldsymbol{x}_{\sigma,\boldsymbol{s}}))_{\sigma\in\mathcal{F}_K,\,\boldsymbol{s}\in\mathcal{V}_\sigma})$. Then

$$(\mathcal{G}_K\xi)_{|V_{K,\boldsymbol{s}}}\cdot(\boldsymbol{x}_{\sigma,\boldsymbol{s}} - \boldsymbol{x}_K) = A(\boldsymbol{x}_{\sigma,\boldsymbol{s}}) - A(\boldsymbol{x}_K) = \nabla A\cdot(\boldsymbol{x}_{\sigma,\boldsymbol{s}} - \boldsymbol{x}_K).$$

Since the family $(\boldsymbol{x}_{\sigma,\boldsymbol{s}} - \boldsymbol{x}_K)_{\sigma\in\mathcal{F}_{K,\boldsymbol{s}}}$ spans the whole space $\mathbb{R}^d$, this shows that the two vectors $(\mathcal{G}_K\xi)_{|V_{K,\boldsymbol{s}}}$ and ∇A are identical, which concludes the proof that $\mathcal{G}_K$ is a $\mathbb{P}_1$-exact gradient reconstruction on K upon the approximation points $(\boldsymbol{x}_K, (\boldsymbol{x}_{\sigma,\boldsymbol{s}})_{\sigma\in\mathcal{F}_K,\,\boldsymbol{s}\in\mathcal{V}_\sigma})$. We now have to establish (12.10) and (12.11).

- *Cartesian mesh.* For a Cartesian mesh, (12.10) and (12.11) are rather straightforward by inspecting Figure 12.1, left.
- *Simplicial mesh.* Let $\sigma\in\mathcal{F}_{K,\boldsymbol{s}}$ and $\overline{\boldsymbol{s}}$ be the vertex opposed to σ in K. Recall that the barycentric coordinates in K of $\boldsymbol{x}_{\sigma,\boldsymbol{s}}$ are given by

  $$\begin{aligned}&\beta^K_{\boldsymbol{s}'}(\boldsymbol{x}_{\sigma,\boldsymbol{s}}) = 1/(d+1) \text{ for all } \boldsymbol{s}'\in\mathcal{V}_\sigma\setminus\{\boldsymbol{s}\},\\ &\beta^K_{\boldsymbol{s}}(\boldsymbol{x}_{\sigma,\boldsymbol{s}}) = 2/(d+1)\\ &\beta^K_{\overline{\boldsymbol{s}}}(\boldsymbol{x}_{\sigma,\boldsymbol{s}}) = 0.\end{aligned}$$

  Since the barycentric coordinate of $\boldsymbol{x}_K = \overline{\boldsymbol{x}}_K$ are all $1/(d+1)$, this shows that

  $$\boldsymbol{x}_{\sigma,\boldsymbol{s}} - \boldsymbol{x}_K = \frac{1}{d+1}(\boldsymbol{s} - \overline{\boldsymbol{s}}).$$

  For any face $\widetilde{\sigma}\in\mathcal{F}_{K,\boldsymbol{s}}\backslash\{\sigma\}$, the vertices $\boldsymbol{s}$ and $\overline{\boldsymbol{s}}$ both belong to $\widetilde{\sigma}$, and $\boldsymbol{s} - \overline{\boldsymbol{s}}$ is thus orthogonal to $\boldsymbol{n}_{K,\widetilde{\sigma}}$. This proves (12.10).
  Since $\boldsymbol{n}_{K,\sigma}\cdot(\boldsymbol{x}_{\sigma,\boldsymbol{s}} - \boldsymbol{x}_K)$ is the orthogonal distance between $\boldsymbol{x}_K$ and σ, $|\tau_{\sigma,\boldsymbol{s}}|\boldsymbol{n}_{K,\sigma}\cdot(\boldsymbol{x}_{\sigma,\boldsymbol{s}} - \boldsymbol{x}_K)$ is equal to d times the measure of the cone with basis $\tau_{\sigma,\boldsymbol{s}}$ and vertex $\boldsymbol{x}_K$. We therefore have $|\tau_{\sigma,\boldsymbol{s}}|\boldsymbol{n}_{K,\sigma}\cdot(\boldsymbol{x}_{\sigma,\boldsymbol{s}} - \boldsymbol{x}_K) = |K|/(d+1)$, which concludes (12.11) due to (12.1).

Step 2: proof that $\mathcal{D}$ is an LLE GD, and estimate on $\mathrm{reg}_{\mathrm{LLE}}(\mathcal{D})$.
To prove that $\mathcal{D}$ is an LLE GD, it remains to show that $\|\nabla_{\mathcal{D}}\cdot\|_{L^p(\Omega)^d}$ is a norm on $X_{\mathcal{D},0}$. If $\nabla_{\mathcal{D}}v = 0$ then (12.8) shows that $v_K = v_{\sigma,\boldsymbol{s}}$ for all $\sigma\in\mathcal{F}_K$ and $\boldsymbol{s}\in\mathcal{V}_\sigma$. Reasoning from neighbour to neighbour, this shows that v is the constant vector. Since $v_{\sigma,\boldsymbol{s}} = 0$ whenever $\sigma\in\mathcal{F}_{\mathrm{ext}}$, we infer that $v = 0$.

Let us now bound $\mathrm{reg}_{\mathrm{LLE}}(\mathcal{D})$. Since all π^i_K are non-negative, $\sum_{i\in I_K}|\pi^i_K| = 1$ and thus $\|\pi_K\|_p \le 1$. All points $(\boldsymbol{x}_i)_{i\in I_K}$ are in the closure of K, so

$\mathrm{dist}(\boldsymbol{x}_i, K) = 0$ and the third term in $\mathrm{reg}_{\mathrm{LLE}}(\mathcal{D})$ vanishes. To bound $\|\mathcal{G}_K\|_p$, we simply use $|\tau_{\sigma,\boldsymbol{s}}| \le C_2 h_K^{d-1}$ and $h_K^d \le C_2 |V_{K,\boldsymbol{s}}|$ for some C_2 depending only on θ and d, so that, by (12.4),

$$|\mathcal{G}_K^K| \le d C_2^2 h_K^{-1} \text{ and } |\mathcal{G}_K^{\sigma,\boldsymbol{s}}| \le C_2^2 h_K^{-1}.$$

The bound on $\|\mathcal{G}_K\|_p$ follows from Remark 7.32. ∎

Let us show how the generic tools presented in Chapter B apply.

Lemma 12.4 (Control of the MPFA GD by a polytopal toolbox). *Let $\mathfrak{T} = (\mathcal{M}, \mathcal{F}, \mathcal{P}, \mathcal{V})$ be a Cartesian or simplicial polytopal mesh. Define the polytopal mesh $\mathfrak{T}' = (\mathcal{M}, \mathcal{F}', \mathcal{P}, \mathcal{V}')$ such that the cells and centres $(\mathcal{M}, \mathcal{P})$ are those $\mathfrak{T}$,*

$$\mathcal{F}' = \{\tau_{\sigma,\boldsymbol{s}} \,:\, \sigma \in \mathcal{F}, \boldsymbol{s} \in \mathcal{V}_\sigma\},$$

and $\mathcal{V}'$ is the set of all vertices of the elements of $\mathcal{F}'$. We define a control of $\mathcal{D}$ by $\mathfrak{T}'$ (in the sense of Definition 7.9) as the isomorphism $\boldsymbol{\Phi} : X_{\mathcal{D},0} \to X_{\mathfrak{T}',0}$ given by $\boldsymbol{\Phi}(u)_K = u_K$ and $\boldsymbol{\Phi}(u)_{\tau_{\sigma,\boldsymbol{s}}} = u_{\sigma,\boldsymbol{s}}$.
Then, there exists C_3, constant if $\mathfrak{T}$ is Cartesian and depending only on $\theta \ge \kappa_{\mathfrak{T}}$ if $\mathfrak{T}$ is simplicial, such that

$$\|\boldsymbol{\Phi}\|_{\mathcal{D},\mathfrak{T}'} \le C_3, \tag{12.12}$$

$$\omega^{\Pi}(\mathcal{D}, \mathfrak{T}', \boldsymbol{\Phi}) = 0, \tag{12.13}$$

$$\omega^{\nabla}(\mathcal{D}, \mathfrak{T}', \boldsymbol{\Phi}) = 0. \tag{12.14}$$

Proof. Let $u \in X_{\mathcal{D},0}$ and apply (12.8) to $\xi = (u_K, (u_{\sigma,\boldsymbol{s}})_{\sigma\in\mathcal{F}_K,\, \boldsymbol{s}\in\mathcal{V}_\sigma})$ to deduce

$$\sum_{\sigma\in\mathcal{F}_K} \sum_{\boldsymbol{s}\in\mathcal{V}_\sigma} |\tau_{\sigma,\boldsymbol{s}}| d_{K,\sigma} \left| \frac{u_{\sigma,\boldsymbol{s}} - u_K}{d_{K,\sigma}} \right|^p \le C_4 \int_K |\nabla_{\mathcal{D}} u(\boldsymbol{x})|^p \mathrm{d}\boldsymbol{x},$$

with $C_4 = 1$ for parallelepipedic meshes, and $C_4 > 0$ depends on $\theta \ge \kappa_{\mathfrak{T}}$ for simplicial meshes. Therefore $\|\boldsymbol{\Phi}(u)\|_{\mathfrak{T}',p}^p \le C_4 \|\nabla_{\mathcal{D}} u\|_{L^p(\Omega)^d}^p$ and (12.12) is proved.
Relation (12.13) follows immediately from $\Pi_{\mathcal{D}} u = \Pi_{\mathfrak{T}'} \boldsymbol{\Phi}(u)$. Finally, we have

$$\begin{aligned}\int_K \nabla_{\mathcal{D}} u(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} &= \sum_{\sigma\in\mathcal{F}_K} \sum_{\boldsymbol{s}\in\mathcal{V}_\sigma} |\tau_{\sigma,\boldsymbol{s}}| (u_{\sigma,\boldsymbol{s}} - u_K) \boldsymbol{n}_{K,\sigma} \\ &= \sum_{\sigma'\in\mathcal{F}'_K} |\sigma'| (\boldsymbol{\Phi}(u)_{\sigma'} - \boldsymbol{\Phi}(u)_K) \boldsymbol{n}_{K,\sigma'} = |K| (\overline{\nabla}_{\mathfrak{T}'} \boldsymbol{\Phi}(u))_{|K}.\end{aligned}$$

This shows that $\omega^{\nabla}(\mathcal{D}, \mathfrak{T}', \boldsymbol{\Phi}) = 0$, which establishes (12.14). ∎

12.1.3 Properties of the MPFA-O gradient discretisation

Thanks to the previous lemmas, the proof of the properties of MPFA-O GDs is straightforward.

Theorem 12.5 (Properties of MPFA-O GDs). *Let $(\mathcal{D}_m)_{m\in\mathbb{N}}$ be a sequence of MPFA-O GDs, as in Section 12.1.1, defined from underlying polytopal meshes $(\mathfrak{T}_m)_{m\in\mathbb{N}}$ that are either Cartesian or simplicial. Assume that $(\theta_{\mathfrak{T}_m} + \eta_{\mathfrak{T}_m})_{m\in\mathbb{N}}$ is bounded (see (7.8) and (7.9)), and that $h_{\mathcal{M}_m} \to 0$ as $m \to \infty$.*
Then the sequence $(\mathcal{D}_m)_{m\in\mathbb{N}}$ is coercive, GD-consistent, limit-conforming and compact in the sense of the definitions of Section 2.1.1 in Chapter 2. Each $\mathcal{D}_m$ also has a piecewise constant reconstruction.

Proof. The limit-conformity, coercivity and compactness are obtained by applying Corollary 7.12, thanks to Lemma 12.4. The consistency is obtained by applying Proposition 7.37, thanks to Lemma 12.3. The piecewise constant reconstruction property is obvious from (12.3). ∎

The following two propositions, also direct consequences of results in the previous sections, are useful to establish precise error estimates for MPFA-O GSs.

Proposition 12.6 (Estimate on $W_{\mathcal{D}}$ for MPFA-O). *Let $\mathfrak{T}$ be a polytopal mesh of Ω in the sense of Definition 7.2, that is either Cartesian or simplicial. Let $\mathcal{D}$ be the MPFA-O GD as in Section 12.1.1. For Cartesian meshes we take $\varrho \geq \theta_{\mathfrak{T}} + \eta_{\mathfrak{T}}$ (see (7.8) and (7.9)), and for simplicial meshes we take $\varrho \geq \kappa_{\mathfrak{T}}$ (see (7.10)). Then, there exists C_5 depending only on Ω, p, and ϱ, such that*

$$C_{\mathcal{D}} \leq C_5 \tag{12.15}$$

and, for all $\boldsymbol{\varphi} \in W^{1,p'}(\Omega)^d$,

$$W_{\mathcal{D}}(\boldsymbol{\varphi}) \leq C_5 \left\|\boldsymbol{\varphi}\right\|_{W^{1,p'}(\Omega)^d} h_{\mathcal{M}}. \tag{12.16}$$

Here, $C_{\mathcal{D}}$ and $W_{\mathcal{D}}$ are the coercivity constant and limit-conformity measure defined by (2.1) and (2.6).

Proof. The conclusion follows immediately from Theorem 7.11 and Lemma 12.4. ∎

Proposition 12.7 (Estimate on $S_{\mathcal{D}}$ for MPFA-O). *Let $\mathfrak{T}$ be a polytopal mesh of Ω in the sense of Definition 7.2, that is either Cartesian or simplicial. Let $\mathcal{D}$ be the MPFA-O GD as in Section 12.1.1. Assume $p > d/2$ and take $\varrho \geq \theta_{\mathfrak{T}}$ (see (7.8)). Then there exists $C_6 > 0$, depending only on Ω, p and ϱ, such that*

$$\forall \varphi \in W^{2,p}(\Omega) \cap W_0^{1,p}(\Omega)\,,\ S_{\mathcal{D}}(\varphi) \leq C_6 \left\|\varphi\right\|_{W^{2,p}(\Omega)} h_{\mathcal{M}},$$

where $S_{\mathcal{D}}$ is defined by (2.2). This means that the space size (see Definition 2.22) of the GD satisfies

$$h_{\mathcal{D}}(W^{2,p}(\Omega) \cap W_0^{1,p}(\Omega); W^{1,p'}(\Omega)^d) \leq \max(C_6, C_5) h_{\mathcal{M}},$$

where C_5 is defined in Proposition 12.6.

Proof. For all $K \in \mathcal{M}$ and all $i \in I_K$, we have $\boldsymbol{x}_i \in \overline{K}$. By Lemmas 12.3 and B.1, the hypotheses of Proposition 7.64 are satisfied with θ depending only on ϱ. This proposition yields the expected estimate on $S_{\mathcal{D}}$. ∎

The application of Lemmas 12.3 and 12.6 to the estimates (2.25) and (2.26) in Theorem 2.28 provides an error in $\mathcal{O}(h_{\mathcal{M}})$ in the case of a linear elliptic problem in one, two or three space dimensions, when the exact solution belongs to $H^2(\Omega)$.

12.2 MPFA-O methods, Neumann and Fourier boundary conditions

12.2.1 Neumann boundary conditions

We refer to Definition 7.54 for the construction of an MPFA-O GD for Neumann boundary conditions, with the same I_Ω, I_∂, $\Pi_{\mathcal{D}}$, $\nabla_{\mathcal{D}}$ as in Section 12.1.1.

Defining $\mathfrak{T}'$ as in Lemma 12.4, for $v \in X_{\mathcal{D}} = X_{\mathfrak{T}'}$ such that $\|\nabla_{\mathcal{D}} v\|_{L^p(\Omega)^d} = 0$, Inequality (12.12) (still valid for non-zero boundary values) and the definition (7.7f) of $|\cdot|_{\mathfrak{T}',p}$ show that all $(v_K)_{K\in\mathcal{M}}$ and all $(v_{\sigma,\boldsymbol{s}})_{\sigma\in\mathcal{F},\boldsymbol{s}\in\mathcal{V}_\sigma}$ are identical. Hence, by definition of $\Pi_{\mathcal{D}}$ the quantity (3.1) is indeed a norm on $X_{\mathcal{D}}$.

For non-homogeneous Neumann boundary conditions, the trace reconstruction $\mathbb{T}_{\mathcal{D}} : X_{\mathcal{D}} \to L^p(\partial\Omega)$ can be defined as $\mathbb{T}_{\mathfrak{T}'}$ (see (7.7d) with $\mathfrak{T} = \mathfrak{T}'$):

$$\forall v \in X_{\mathcal{D}}\,,\ \forall \sigma \in \mathcal{F}_{\text{ext}}\,,\ \forall \boldsymbol{s} \in \mathcal{V}_\sigma\ :\ \mathbb{T}_{\mathcal{D}} v = v_{\sigma,\boldsymbol{s}} \text{ on } \tau_{\sigma,\boldsymbol{s}}. \tag{12.17}$$

Since the regularity factor $\text{reg}_{\text{LLE}}(\mathcal{D})$ is defined as for Dirichlet boundary conditions, Lemma 12.3 still applies and show that this factor remains bounded if $\theta_{\mathfrak{T}}$ is bounded. Defining the control $\boldsymbol{\Phi} = \text{Id} : X_{\mathcal{D}} \to X_{\mathfrak{T}'}$ as in Lemma 12.4, we see that this lemma still holds and that $\omega^{\mathbb{T}}(\mathcal{D}, \mathfrak{T}', \boldsymbol{\Phi}) = 0$. Hence, Corollary 7.18 and Proposition 7.55 give the following theorem.

Theorem 12.8 (Properties of MPFA-O GDs for Neumann BCs). *Let $(\mathcal{D}_m)_{m\in\mathbb{N}}$ be a sequence of MPFA-O GDs for Neumann boundary conditions as above, defined from underlying polytopal meshes $(\mathfrak{T}_m)_{m\in\mathbb{N}}$ that are either Cartesian or simplicial. Assume that the sequence $(\theta_{\mathfrak{T}_m} + \eta_{\mathfrak{T}_m})_{m\in\mathbb{N}}$ is bounded (see (7.8) and (7.9)), and that $h_{\mathcal{M}_m} \to 0$ as $m \to \infty$.*

Then the sequence $(\mathcal{D}_m)_{m\in\mathbb{N}}$ is coercive, GD-consistent, limit-conforming and compact in the sense of Definitions 3.13, 3.4, 3.14 and 3.15. Moreover, each $\mathcal{D}_m$ has a piecewise constant reconstruction in the sense of Definition 2.12.

Proposition 7.70 and Theorem 7.17 also give estimates on $S_{\mathcal{D}}$, $C_{\mathcal{D}}$ and $W_{\mathcal{D}}$ that are similar to those in Lemma 12.3 and Proposition 12.6. The constants depend only on an upper bound of $\theta_{\mathfrak{T}} + \eta_{\mathfrak{T}}$ (for Cartesian meshes) or $\kappa_{\mathfrak{T}}$ (for simplicial meshes, due to Lemma B.4).

12.2.2 Fourier boundary conditions

Starting from an MPFA-O GD for Dirichlet boundary conditions, we follow Definition 7.57 in Section 7.3.6 to define an MPFA-O GD for Fourier boundary conditions.

The boundary mesh $\mathcal{M}_\partial$ is simply $\{\tau_{\sigma,\boldsymbol{s}} : \sigma \in \mathcal{F}_{\text{ext}}, \boldsymbol{s} \in \mathcal{V}_\sigma\}$, and the trace reconstruction (12.17) corresponds to $I_{\sigma,\boldsymbol{s}} = \{\tau_{\sigma,\boldsymbol{s}}\}$ and $\pi^{\sigma,\boldsymbol{s}}_{\sigma,\boldsymbol{s}} = 1$ on $\tau_{\sigma,\boldsymbol{s}}$. The bound on $\text{reg}_{\text{LLE}}(\mathcal{D})$ for Fourier boundary conditions therefore easily follows from the bound on this quantity for Dirichlet boundary conditions, and the consistency (under boundedness of $\theta_{\mathfrak{T}}$) is a consequence of Proposition 7.58.

As noticed in Remark 7.20, the work done for Neumann boundary conditions then immediately shows that Theorem 12.8 also applies to Fourier boundary conditions. Similarly, we could obtain estimates on $S_{\mathcal{D}}$, $C_{\mathcal{D}}$ and $W_{\mathcal{D}}$ as in Propositions 12.7 and 12.6.

13

Hybrid mimetic mixed schemes

Since the 50's, several schemes have been developed which aim to satisfy some form of calculus formula at the discrete level. These schemes are called mimetic finite difference (MFD) or compatible discrete operator (CDO) schemes. Contrary to DDFV methods (see Section 14.2 and [77]) for which the discrete operators and duality products are designed to satisfy fully discrete calculus formula, the discrete operators of MFD/CDO methods satisfy a Stokes formula that involves both continuous and discrete functions. Depending on the choice of the location of the main unknowns (faces or vertices), two different MFD/CDO families exist. We refer to [128] for a review on MFD methods, and to [27, 26] (and reference therein) for CDO methods.

A first MFD method, that we call mixed/hybrid MFD or hMFD here, is designed by using the fluxes through the mesh faces as initial unknowns [36, 35]. This requires to recast (2.20) in a mixed form, *i.e.* to write $\overline{q} = \Lambda \nabla \overline{u}$ and $-\mathrm{div}(\overline{q}) = f + \mathrm{div}(\boldsymbol{F})$, and to discretise this set of two equations. The resulting scheme takes a form that is apparently far from the GS (2.23). It was however proved in [75] that this form of hMFD can be actually embedded in a slightly larger family also containing the hybrid finite volume (HFV) scheme, which is the hybrid version of the SUSHI (Scheme Using Stabilisation and Hybrid Interfaces scheme [96], and mixed finite volume (MFV) methods [70, 71]. This family has been called hybrid mimetic mixed (HMM) schemes; each scheme in this family can be written in three different ways, depending on the considered approach (hMFD, HFV or MFV). The HFV formulation of an HMM scheme is very close to the weak formulation (2.22) of the elliptic PDE; it actually consists in writing this weak formulation with a reconstructed gradient and a stabilisation term (bilinear form on (u, v)). It was proved in [99] that this specific stabilisation term could be included in an augmented gradient, and thus that the HFV scheme is a GS. More surprisingly perhaps, [76] managed to prove that *all* possible stabilisations in the HMM families can be embedded in a gradient, and thus that all HMM methods (and thus all hMFD, HFV and MFV schemes) are GDMs.

In the following sections we detail the GD that leads to HMM methods when applied to linear diffusion equations, and we establish its properties. HMM methods correspond to LLE GDs. Following the nomenclature in Section 7.3.4, the condensed SUSHI methods (including elimination of some or all of the interface unknowns) are nothing else than barycentric condensation of HMM methods; they are therefore also GDMs.
We conclude this section by presenting some considerations on the fluxes associated to the HMM and SUSHI methods.
Note that some schemes adapting HMM ideas and variants to non-linear equations and systems have already been proposed and analysed in [68, 43, 95], but they are not GDMs and do not fully take advantage of the coercive gradient provided by HMM methods.

13.1 HMM methods for Dirichlet boundary conditions

We consider here the case of non-homogeneous Dirichlet boundary conditions, which includes as a special case homogeneous Dirichlet conditions.

13.1.1 Definition of HMM gradient discretisations

The discrete elements that define an HMM GD are the following. We take $\mathfrak{T} = (\mathcal{M}, \mathcal{F}, \mathcal{P}, \mathcal{V})$ a polytopal mesh of Ω as in Definition 7.2, and we refer to the notions in Definition 7.52.

1. The geometrical entities attached to the discrete unknowns are $I = \mathcal{M} \cup \mathcal{F}$ and the approximation points are $S = ((\boldsymbol{x}_K)_{K\in\mathcal{M}}, (\overline{\boldsymbol{x}}_\sigma)_{\sigma\in\mathcal{F}})$. We let $I_\Omega = \mathcal{M} \cup \mathcal{F}_{\rm int}$ and $I_\partial = \mathcal{F}_{\rm ext}$. Hence, recalling the definitions (7.7a) and (7.7b) of $X_{\mathfrak{T}}$ and $X_{\mathfrak{T},0}$,
$$X_{\mathcal{D}} = X_{\mathfrak{T}} = \{v = ((v_K)_{K\in\mathcal{M}}, (v_\sigma)_{\sigma\in\mathcal{F}}) \,:\, v_K \in \mathbb{R} \text{ for all } K \in \mathcal{M}\,, \\ v_\sigma \in \mathbb{R} \text{ for all } \sigma \in \mathcal{F}\},$$
and
$$X_{\mathcal{D},0} = X_{\mathfrak{T},0} = \{v \in X_{\mathfrak{T}} \,:\, v_\sigma = 0 \text{ for all } \sigma \in \mathcal{F}_{\rm ext}\}.$$
For $K \in \mathcal{M}$, we set $I_K = \{K\} \cup \mathcal{F}_K$.
2. The function reconstructions $\pi_K = (\pi_K^K, (\pi_K^\sigma)_{\sigma\in\mathcal{F}_K})$ of $L^p(K)$ are defined by
$$\pi_K^K = 1 \text{ and } \pi_K^\sigma = 0 \text{ for all } \sigma \in \mathcal{F}_K. \tag{13.1}$$
Recalling the definition (7.7c) of $\Pi_{\mathfrak{T}}$, (7.32) therefore reads
$$\forall v \in X_{\mathcal{D}}\,,\ \forall K \in \mathcal{M}\,,\ \text{for a.e. } \boldsymbol{x} \in K\,,\ \Pi_{\mathcal{D}} v(\boldsymbol{x}) = \Pi_{\mathfrak{T}} v(\boldsymbol{x}) = v_K. \tag{13.2}$$

3. The gradient reconstruction $\mathcal{G}_K$ is best initially described through its action $\mathcal{G}_K v$ on families of real numbers, than through explicit formulas for the functions $(\mathcal{G}_K^i)_{i\in I_K}$. As already mentioned, the polytopal gradient defined by (7.7e), that is,

$$\overline{\nabla}_K v = \frac{1}{|K|}\sum_{\sigma\in\mathcal{F}_K} |\sigma| v_\sigma \boldsymbol{n}_{K,\sigma}, \tag{13.3}$$

is $\mathbb{P}_1$-exact (Lemma B.10), but not "strong enough" to control all the discrete unknowns in $X_{\mathfrak{T},0}$ since it does not involve the unknowns $(u_K)_{K\in\mathcal{M}}$ (see Remark 7.8). The HMM gradient is built by adding to this polytopal gradient a stabilisation term that is constant in each half-diamond in K. Let $X_K = \{v = (v_K, (v_\sigma)_{\sigma\in\mathcal{F}_K}) \,:\, v_K \in \mathbb{R}\,,\ v_\sigma \in \mathbb{R}\}$ be the space of discrete unknowns in K and define, for $v \in X_K$, the function $\mathcal{G}_K v \in L^p(K)^d$ by

$$\begin{aligned} &\forall \sigma \in \mathcal{F}_K\,,\ \text{for a.e. } \boldsymbol{x} \in D_{K,\sigma}\,,\\ &\mathcal{G}_K v(\boldsymbol{x}) = \overline{\nabla}_K v + \frac{\sqrt{d}}{d_{K,\sigma}}[\mathcal{L}_K R_K(v)]_\sigma \boldsymbol{n}_{K,\sigma}, \end{aligned} \tag{13.4}$$

where, denoting by $X_{\mathcal{F}_K} = \{\xi = (\xi_\sigma)_{\sigma\in\mathcal{F}_K} \,:\, \xi_\sigma \in \mathbb{R}\}$ the space of face values around K,

- $R_K : X_K \to X_{\mathcal{F}_K}$ is the linear mapping given by

$$\begin{aligned} &R_K(v) = (R_{K,\sigma}(v))_{\sigma\in\mathcal{F}_K} \text{ with}\\ &R_{K,\sigma}(v) = v_\sigma - v_K - \overline{\nabla}_K v\cdot(\overline{\boldsymbol{x}}_\sigma - \boldsymbol{x}_K), \end{aligned} \tag{13.5}$$

- $\mathcal{L}_K$ is an isomorphism of the vector space $\mathrm{Im}(R_K)$.

The gradient reconstruction $\nabla_{\mathcal{D}}$ is then defined by (7.33), which simply gives

$$\begin{aligned} &\forall v \in X_{\mathcal{D}}\,,\ \forall K \in \mathcal{M}\,,\ \forall \sigma \in \mathcal{F}_K\,,\ \text{for a.e. } \boldsymbol{x} \in D_{K,\sigma}\,,\\ &\nabla_{\mathcal{D}} v(\boldsymbol{x}) = \overline{\nabla}_K v + \frac{\sqrt{d}}{d_{K,\sigma}}[\mathcal{L}_K R_K(v)]_\sigma \boldsymbol{n}_{K,\sigma}. \end{aligned} \tag{13.6}$$

Remark 13.1. If all values v_σ are equal to 0 and all values v_K are equal to 1, then $\overline{\nabla}_K v = 0$ and the non-zero part of $\nabla_{\mathcal{D}} v$ comes from the components of $R_K(v)$, all equal to -1.

Remark 13.2 (The SUSHI scheme). A natural choice of $\mathcal{L}_K$ is $\mathcal{L}_K = \mathrm{Id}_{\mathrm{Im}(R_K)}$. The corresponding HMM method is then the hybrid finite volume (HFV) scheme, *i.e.* the full hybrid version of the SUSHI scheme [96]. The coefficient $\sqrt{d}$ in the expression (13.6) of the gradient may be modified to obtain other schemes; replacing it by d, for instance, leads to the so-called "DGA" scheme, see [26, Remark 7.34].

The functions $(\mathcal{G}_K^i)_{i\in I_K}$ of $L^p(K)^d$ can be recovered through $\mathcal{G}_K$ defined by (13.4). Let $v^K \in X_K$ (resp. $v^\sigma \in X_K$) be the vectors with value 1 at K (resp. at σ) and 0 at all other positions. Then,

$$\mathcal{G}_K^K = \mathcal{G}_K v^K \text{ and } \mathcal{G}_K^\sigma = \mathcal{G}_K v^\sigma \text{ for all } \sigma \in \mathcal{F}_K. \tag{13.7}$$

4. The trace interpolation operator $\mathcal{I}_{\mathcal{D},\partial} : W^{1-\frac{1}{p},p}(\partial\Omega) \to X_{\mathcal{D},\partial}$ is defined by

$$\forall g \in W^{1-\frac{1}{p},p}(\partial\Omega)\,,\ \forall \sigma \in \mathcal{F}_{\text{ext}}\,,\ (\mathcal{I}_{\mathcal{D},\partial} g)_\sigma = \frac{1}{|\sigma|}\int_\sigma g(\boldsymbol{x})\mathrm{d}\gamma(\boldsymbol{x}). \tag{13.8}$$

5. Lemma 13.10 below establishes the exactness of π_K and $\mathcal{G}_K$, and the fact that $\|\nabla_{\mathcal{D}}\cdot\|_{L^p(\Omega)^d}$ is a norm on $X_{\mathcal{D},0}$.

Remark 13.3 (Hybrid method)
The face unknowns $(v_\sigma)_{\sigma\in\mathcal{F}}$ correspond to the hybridisation of the hMFD methods.

Remark 13.4 (Simpler trace interpolation)
As explained in Remark 2.50, simpler trace interpolations can be used if the boundary condition g of the considered problem (*e.g.*, in (2.98b)) is more regular than $W^{1-\frac{1}{p},p}(\partial\Omega)$. For example, if $g \in C(\overline{\Omega})$ we can define $(\mathcal{I}_{\mathcal{D},\partial} g)_\sigma = g(\overline{\boldsymbol{x}}_\sigma)$.

We now want to prove that all hMFD, HFV and MFV methods, as presented in the literature, are GDMs with gradient discretisations as above for suitable choices of $(\mathcal{L}_K)_{K\in\mathcal{M}}$. As explained in the introduction of this chapter, hMFD, HFV and MFV schemes are three different presentations of the same method [75]. The presentation that is the closest to a GS is that of the HFV scheme. With the notations above, any HMM method for the weak form (2.100) of the linear problem (2.98) with $\mathbf{F} = 0$ can be written (see [75] in the case $g = 0$):

$$\begin{aligned}&\text{Find } u \in \mathcal{I}_{\mathcal{D},\partial} g + X_{\mathcal{D},0} \text{ such that, for all } v \in X_{\mathcal{D},0},\\ &\sum_{K\in\mathcal{M}} |K|\Lambda_K \overline{\nabla}_K u \cdot \overline{\nabla}_K v + \sum_{K\in\mathcal{M}} R_K(v)^T \mathbb{B}_K R_K(u)\\ &\qquad = \sum_{K\in\mathcal{M}} v_K \int_K f(\boldsymbol{x})\mathrm{d}\boldsymbol{x},\end{aligned} \tag{13.9}$$

where Λ_K is the constant value of Λ on K (we assume that Λ is piecewise constant on $\mathcal{M}$ – see Remark 13.17 below for a discussion on this assumption), $\mathbb{B}_K = ((\mathbb{B}_K)_{\sigma,\sigma'})_{\sigma,\sigma\in\mathcal{F}_K}$ is a symmetric positive definite matrix, and $R_K(v)^T$ the transpose of the vector $R_K(v)$.

Remark 13.5 ($\mathbb{R}^{\mathrm{Card}(\mathcal{F}_K)}$ *vs.* $X_{\mathcal{F}_K}$)
There is a slight abuse of notation here. We write $R_K(v)$ as a column vector as if it belonged to $\mathbb{R}^{\mathrm{Card}(\mathcal{F}_K)}$, while it actually belongs to $X_{\mathcal{F}_K}$. Implicitly, when switching from elements w of $X_{\mathcal{F}_K}$ to column vectors, we have chosen a numbering $(\sigma_1, \ldots, \sigma_\ell)$ of the faces of K, and we set $w(\sigma_i) = w_i$ for all $i = 1, \ldots, \ell$. The same abuse of notation is made when considering $\mathbb{B}_K$ as a matrix and writing $R_K(v)^T \mathbb{B}_K R_K(v)$, or further below in (13.15) when considering $\mathbb{D}_K$ as a matrix.

The following lemma will be useful both to establish that all HMM methods are GDMs, and to analyse the properties of HMM GDs.

Lemma 13.6. *Let $\mathfrak{T} = (\mathcal{M}, \mathcal{F}, \mathcal{P}, \mathcal{V})$ be a polytopal mesh of Ω in the sense of Definition 7.2, and let $\mathcal{D}$ be an HMM GD as defined above, for certain choices of $(\mathcal{L}_K)_{K\in\mathcal{M}}$. Then*

1. *For all $K \in \mathcal{M}$, $\beta \in \mathrm{Im}(R_K)$ if and only if $\sum_{\sigma\in\mathcal{F}_K} |\sigma|\beta_\sigma \boldsymbol{n}_{K,\sigma} = 0$.*
2. *For all $v \in X_{\mathcal{D}}$ and all $K \in \mathcal{M}$,*

$$\overline{\nabla}_K v = \frac{1}{|K|} \int_K \nabla_{\mathcal{D}} v(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}. \tag{13.10}$$

Proof.
Item 1. Let us first introduce the mapping $\widetilde{R}_K : X_{\mathcal{F}_K} \to X_{\mathcal{F}_K}$ defined, for $\xi \in X_{\mathcal{F}_K}$ by $\widetilde{R}_K(\xi) = (\widetilde{R}_{K,\sigma}(\xi))_{\sigma\in\mathcal{F}_K}$ with

$$\widetilde{R}_{K,\sigma}(\xi) = \xi_\sigma - X_\xi \cdot (\overline{\boldsymbol{x}}_\sigma - \boldsymbol{x}_K) \quad \text{with} \quad X_\xi = \frac{1}{|K|} \sum_{\sigma'\in\mathcal{F}_K} |\sigma'| \xi_{\sigma'} \boldsymbol{n}_{K,\sigma'}.$$

By noting that $R_K(v) = \widetilde{R}_K((v_\sigma - v_K)_{\sigma\in\mathcal{F}_K})$ we see that $\mathrm{Im}(R_K) = \mathrm{Im}(\widetilde{R}_K)$. Let $\beta \in \mathrm{Im}(\widetilde{R}_K)$. Taking $\xi \in X_{\mathcal{F}_K}$ such that $\beta_\sigma = \xi_\sigma - X_\xi \cdot (\overline{\boldsymbol{x}}_\sigma - \boldsymbol{x}_K)$, and using Lemma B.3, we see that

$$\begin{aligned}
\sum_{\sigma\in\mathcal{F}_K} |\sigma|\beta_\sigma \boldsymbol{n}_{K,\sigma} &= \sum_{\sigma\in\mathcal{F}_K} |\sigma|\xi_\sigma \boldsymbol{n}_{K,\sigma} - \sum_{\sigma\in\mathcal{F}_K} |\sigma| X_\xi \cdot (\overline{\boldsymbol{x}}_\sigma - \boldsymbol{x}_K) \boldsymbol{n}_{K,\sigma} \\
&= \sum_{\sigma\in\mathcal{F}_K} |\sigma|\xi_\sigma \boldsymbol{n}_{K,\sigma} - \left(\sum_{\sigma\in\mathcal{F}_K} |\sigma| \boldsymbol{n}_{K,\sigma} (\overline{\boldsymbol{x}}_\sigma - \boldsymbol{x}_K)^T \right) X_\xi \\
&= \sum_{\sigma\in\mathcal{F}_K} |\sigma|\xi_\sigma \boldsymbol{n}_{K,\sigma} - |K| X_\xi = 0.
\end{aligned}$$

Setting

$$G_K : \beta \in X_{\mathcal{F}_K} \mapsto \sum_{\sigma\in\mathcal{F}_K} |\sigma|\beta_\sigma \boldsymbol{n}_{K,\sigma} \in \mathbb{R}^d,$$

we just showed that $\mathrm{Im}(\widetilde{R}_K) \subset \ker(G_K)$. Since $(\boldsymbol{n}_{K,\sigma})_{\sigma\in\mathcal{F}_K}$ spans $\mathbb{R}^d$, the linear mapping G_K has rank d and therefore $\dim(\ker G_K) = \mathrm{Card}(\mathcal{F}_K) - d$. It is easy to see that

$$\ker(\widetilde{R}_K) = \{\xi \in X_{\mathcal{F}_K}\,;\, \exists Z_\xi \in \mathbb{R}^d \text{ such that } \xi_\sigma = Z_\xi \cdot (\overline{\boldsymbol{x}}_\sigma - \boldsymbol{x}_K)\},$$

and thus that $Z \in \mathbb{R}^d \mapsto (Z \cdot (\overline{\boldsymbol{x}}_\sigma - \boldsymbol{x}_K))_{\sigma\in\mathcal{F}_K} \in \ker(\widetilde{R}_K)$ is an isomorphism (the one-to-one property comes from the fact that $(\overline{\boldsymbol{x}}_\sigma - \boldsymbol{x}_K)_{\sigma\in\mathcal{F}_K}$ spans $\mathbb{R}^d$). Hence, $\dim(\mathrm{Im}(\widetilde{R}_K)) = \mathrm{Card}(\mathcal{F}_K) - d = \dim(\ker(G_K))$. Since $\mathrm{Im}(\widetilde{R}_K) \subset \ker(G_K)$, the equality of dimensions therefore gives $\mathrm{Im}(\widetilde{R}_K) = \ker(G_K)$ and completes the proof of Item 1.

Item 2. By (13.6), since $\nabla_{\mathcal{D}} v$ is constant in each half-diamond inside K, using (B.1) to write $|D_{K,\sigma}| = \frac{|\sigma| d_{K,\sigma}}{d}$ gives

$$\begin{aligned}\int_K \nabla_{\mathcal{D}} v(\boldsymbol{x})\mathrm{d}\boldsymbol{x} &= \sum_{\sigma\in\mathcal{F}_K} |D_{K,\sigma}| (\nabla_{\mathcal{D}} v)_{|D_{K,\sigma}} \\ &= |K|\overline{\nabla}_K v + \frac{1}{\sqrt{d}} \sum_{\sigma\in\mathcal{F}_K} |\sigma| [\mathcal{L}_K(R_K(v))]_\sigma \boldsymbol{n}_{K,\sigma}. \end{aligned} \tag{13.11}$$

But $\mathcal{L}_K(R_K(v)) \in \mathrm{Im}(R_K)$ since $\mathcal{L}_K$ is an isomorphism of this space, and by Item 1 the last term in (13.11) vanishes. This proves that $\int_K \nabla_{\mathcal{D}} v(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = |K|\overline{\nabla}_K v$ as claimed. ∎

Theorem 13.7 (HMM methods are GDMs). *Let $\mathfrak{T}$ be a polytopal mesh of Ω in the sense of Definition 7.2, and for each $K \in \mathcal{M}$ take $\mathbb{B}_K$ a symmetric positive definite matrix of size* $\mathrm{Card}(\mathcal{F}_K)$. *Then there exists a choice of isomorphisms $\mathcal{L}_K : \mathrm{Im}(R_K) \to \mathrm{Im}(R_K)$ such that, if $\mathcal{D}$ is the GD defined above using these isomorphisms, the GS (2.23) (with $\boldsymbol{F} = 0$) is the HMM scheme (13.9) for the choice of matrices $(\mathbb{B}_K)_{K\in\mathcal{M}}$.*

The proof also shows that any choice of isomorphisms $(\mathcal{L}_K)_{K\in\mathcal{M}}$ leads to an HMM method. In other words, there is a perfect equivalence between the HMM family of methods and the family of GDs defined above.

Proof. Given the definition (13.2) of $\Pi_{\mathcal{D}}$, the right-hand sides of (2.23) and (13.9) clearly coincide. Since the space for the unknown and the test functions are the same in both schemes, it simply remains to prove that the left-hand sides also coincide for a proper choice of the isomorphisms $(\mathcal{L}_K)_{K\in\mathcal{M}}$.
Let $K \in \mathcal{M}$. We will prove that there exists an isomorphism $\mathcal{L}_K$ such that, for all $(u, v) \in X_{\mathcal{D}}^2$,

$$\begin{aligned}&|K|\Lambda_K \overline{\nabla}_K u \cdot \overline{\nabla}_K v + R_K(v)^T \mathbb{B}_K R_K(u) \\ &\qquad = \int_K \Lambda_K \nabla_{\mathcal{D}} u(\boldsymbol{x}) \cdot \nabla_{\mathcal{D}} v(\boldsymbol{x})\mathrm{d}\boldsymbol{x}. \end{aligned} \tag{13.12}$$

Summing (13.12) over $K \in \mathcal{M}$ then shows that the left-hand sides of (2.23) and (13.9) are identical.
Recall the definition (13.6) of $\nabla_{\mathcal{D}}$ and use $\sum_{\sigma\in\mathcal{F}_K} |D_{K,\sigma}| = |K|$ to write, by developing the scalar product,

$$\int_K \Lambda_K \nabla_{\mathcal{D}} u(\boldsymbol{x}) \cdot \nabla_{\mathcal{D}} v(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} = \sum_{\sigma \in \mathcal{F}_K} |D_{K,\sigma}| \Lambda_K (\nabla_{\mathcal{D}} u)_{|K} \cdot (\nabla_{\mathcal{D}} v)_{|K}$$

$$= |K| \Lambda_K \overline{\nabla}_K u \cdot \overline{\nabla}_K v$$

$$+ \Lambda_K \overline{\nabla}_K u \cdot \sum_{\sigma \in \mathcal{F}_K} |D_{K,\sigma}| \frac{\sqrt{d}}{d_{K,\sigma}} [\mathcal{L}_K R_K(v)]_\sigma \boldsymbol{n}_{K,\sigma} \tag{13.13}$$

$$+ \overline{\nabla}_K v \cdot \Lambda_K \sum_{\sigma \in \mathcal{F}_K} |D_{K,\sigma}| \frac{\sqrt{d}}{d_{K,\sigma}} [\mathcal{L}_K R_K(u)]_\sigma \boldsymbol{n}_{K,\sigma} \tag{13.14}$$

$$+ \sum_{\sigma \in \mathcal{F}_K} |D_{K,\sigma}| \frac{d}{d_{K,\sigma}^2} \Lambda_K \boldsymbol{n}_{K,\sigma} \cdot \boldsymbol{n}_{K,\sigma} [\mathcal{L}_K R_K(u)]_\sigma [\mathcal{L}_K R_K(v)]_\sigma.$$

By (B.1), $\frac{|D_{K,\sigma}|}{d_{K,\sigma}} = \frac{|\sigma|}{d}$ and thus, since $\mathcal{L}_K$ has values in $\mathrm{Im}(R_K)$, Item 1 in Lemma 13.6 shows that (13.13) and (13.14) vanish. Hence

$$\int_K \Lambda_K \nabla_{\mathcal{D}} u(\boldsymbol{x}) \cdot \nabla_{\mathcal{D}} v(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}$$
$$= |K| \Lambda_K \overline{\nabla}_K u \cdot \overline{\nabla}_K v + [\mathcal{L}_K R_K(v)]^T \mathbb{D}_K [\mathcal{L}_K R_K(u)] \tag{13.15}$$

with $\mathbb{D}_K = \mathrm{diag}(\frac{|\sigma|}{d_{K,\sigma}} \Lambda_K \boldsymbol{n}_{K,\sigma} \cdot \boldsymbol{n}_{K,\sigma})$ a diagonal definite positive matrix. Relation (13.12) therefore holds provided that, for all $(\xi, \eta) \in (\mathrm{Im}(R_K))^2$,

$$\xi^T \mathbb{B}_K \eta = (\mathcal{L}_K(\xi))^T \mathbb{D}_K (\mathcal{L}_K(\eta)). \tag{13.16}$$

Consider the vector space $E = \mathrm{Im}(R_K) \subset X_{\mathcal{F}_K}$, endowed with the two inner products $\langle \xi, \eta \rangle_1 = \xi^T \mathbb{B}_K \eta$ and $\langle \xi, \eta \rangle_2 = \xi^T \mathbb{D}_K \eta$. The isomorphism $\mathcal{L}_K : \mathrm{Im}(R_K) \mapsto \mathrm{Im}(R_K)$ given by Lemma 13.8 below then satisfies (13.17), which is precisely (13.16) with $x = \xi$ and $y = \eta$. ■

Lemma 13.8. *Let E be a finite-dimensional vector space endowed with two inner products $\langle\ ,\ \rangle_1$ and $\langle\ ,\ \rangle_2$. There exists an isomorphism $\mathcal{L} : E \mapsto E$ such that*

$$\text{for all } (x, y) \in E^2,\ \langle x, y \rangle_1 = \langle \mathcal{L}x, \mathcal{L}y \rangle_2. \tag{13.17}$$

Proof. Let e be an orthonormal basis for $\langle\ ,\ \rangle_2$ and M_e be the (symmetric definite positive) matrix of $\langle\ ,\ \rangle_1$ in this basis. If X_e and Y_e are the coordinates of x and y in e then $\langle x, y \rangle_1 = Y_e^T M_e X_e$. Let $\mathcal{L}_e = \sqrt{M_e}$ and define $\mathcal{L}$ as the isomorphism whose matrix relative to the basis e is $\mathcal{L}_e$. Since e is orthonormal for $\langle\ ,\ \rangle_2$, the relation $Y_e^T M_e X_e = (\mathcal{L}_e Y_e)^T (\mathcal{L}_e X_e)$ translates into $\langle x, y \rangle_1 = \langle \mathcal{L}x, \mathcal{L}y \rangle_2$. ■

Remark 13.9 (Elimination of the cell unknowns in the HMM GS by static condensation)

By static condensation, the cell unknowns can be eliminated when an HMM method is applied to a linear elliptic equation. This is done by taking, in (13.9),the test function v such that $v_K = 1$ for one cell $K \in \mathcal{M}$, $v_L = 0$ for all other cells L, and $v_\sigma = 0$ for all $\sigma \in \mathcal{F}_K$. Then (13.3) shows that $\overline{\nabla}_L v = 0$ for all $L \in \mathcal{M}$, which gives $R_K(v) = -(1)_{\sigma\in\mathcal{F}_K} =: -\mathbf{1}_K$, and $R_L(v) = 0$ for all $L \neq K$. Hence, (13.9) leads to

$$-\mathbf{1}_K^T \mathbb{B}_K R_K(u) = \int_K f.$$

Since $R_K(u) = M_K(u_\sigma)_{\sigma\in\mathcal{F}_K} - \mathbf{1}_K u_K$, with M_K a linear operator, we infer that

$$(\mathbf{1}_K^T \mathbb{B}_K \mathbf{1}_K) u_K = \int_K f + \mathbf{1}_K^T \mathbb{B}_K M_K(u_\sigma)_{\sigma\in\mathcal{F}_K}.$$

The matrix $\mathbb{B}_K$ being symmetric definite positive, $\mathbf{1}_K^T \mathbb{B}_K \mathbf{1}_K > 0$ and therefore

$$u_K = (\mathbf{1}_K^T \mathbb{B}_K \mathbf{1}_K)^{-1} \left(\int_K f + \mathbf{1}_K^T \mathbb{B}_K M_K(u_\sigma)_{\sigma\in\mathcal{F}_K} \right).$$

Hence, the unknown u_K can be locally computed from the source term f and the face unknowns $(u_\sigma)_{\sigma\in\mathcal{F}_K}$, without even having to invert a local system. This expression for u_K can be plugged back into (13.9) and provides a symmetric positive definite system only on $(u_\sigma)_{\sigma\in\mathcal{F}_{\rm int}}$.

13.1.2 Preliminary lemmas

To prove that HMM GD satisfy the properties defined in Part I, preliminary results must first be established. If $\mathcal{D}$ is an HMM GD as in Section 13.1.1, we define the following measure of the invertibility properties of the isomorphisms $(\mathcal{L}_K)_{K\in\mathcal{M}}$:

$$\begin{aligned}\zeta_{\mathcal{D}} = \min \Bigg\{ \zeta > 0 \,:\, \forall K \in \mathcal{M},\ \forall v \in X_K\,, \\ \zeta^{-1} \sum_{\sigma\in\mathcal{F}_K} |D_{K,\sigma}| \left| \frac{R_{K,\sigma}(v)}{d_{K,\sigma}} \right|^p \le \sum_{\sigma\in\mathcal{F}_K} |D_{K,\sigma}| \left| \frac{[\mathcal{L}_K R_K(v)]_\sigma}{d_{K,\sigma}} \right|^p \\ \le \zeta \sum_{\sigma\in\mathcal{F}_K} |D_{K,\sigma}| \left| \frac{R_{K,\sigma}(v)}{d_{K,\sigma}} \right|^p \Bigg\}.\end{aligned} \tag{13.18}$$

In the case of the SUSHI scheme, see Remark 13.2, it is clear that $\zeta_{\mathcal{D}} = 1$. The following lemma states that HMM GDs are LLE GDs, and gives a control of their regularity $\mathrm{reg}_{\rm LLE}$ in terms of $\zeta_{\mathcal{D}}$ and geometric regularity factors.

Lemma 13.10 (Estimate on $\mathrm{reg}_{\rm LLE}(\mathcal{D})$ for the HMM GD). *Let $\mathfrak{T}$ be a polytopal mesh of Ω in the sense of Definition 7.2, and $\mathcal{D}$ be an HMM GD as in Section 13.1.1. Then, for all $K \in \mathcal{M}$, π_K is a $\mathbb{P}_0$-exact function*

reconstruction on K, and $\mathcal{G}_K$ is $\mathbb{P}_1$-exact gradient reconstruction on K upon I_K.
Moreover, $\mathcal{D}$ is an LLE GD and, if $\varrho \geq \theta_{\mathfrak{T}} + \zeta_{\mathcal{D}}$ (see (7.8) and (13.18)), there exists C_1, depending only on p, d and ϱ, such that

$$\mathrm{reg}_{\mathrm{LLE}}(\mathcal{D}) \leq C_1. \tag{13.19}$$

Proof. Let $K \in \mathcal{M}$. According to (13.1), $\sum_{i \in I_K} \pi_K^i = \pi_K^K = 1$ so π_K is a $\mathbb{P}_0$-exact function reconstruction.
Lemma B.10 shows that $\overline{\nabla}_K$ is $\mathbb{P}_1$-exact gradient reconstruction. Hence, if v interpolates an affine mapping A at the approximation points $(\boldsymbol{x}_K, (\overline{\boldsymbol{x}}_\sigma)_{\sigma \in \mathcal{F}_K})$, (13.5) gives $R_{K,\sigma}(v) = A(\overline{\boldsymbol{x}}_\sigma) - A(\boldsymbol{x}_K) - \nabla A \cdot (\overline{\boldsymbol{x}}_\sigma - \boldsymbol{x}_K) = 0$. Therefore, $\mathcal{G}_K v_{|D_{K,\sigma}} = \overline{\nabla}_K v = \nabla A$ and $\mathcal{G}_K$ is a $\mathbb{P}_1$-exact gradient reconstruction on K upon I_K.
To prove that $\mathcal{D}$ is an LLE GD, we need to show that $v = 0$ whenever $\nabla_{\mathcal{D}} v = 0$. If the latter equality holds, then (13.10) shows that $\overline{\nabla}_K v = 0$ for all $K \in \mathcal{M}$ and thus, by (13.4) and the fact that $\mathcal{L}_K$ is an isomorphism, $R_K(v) = 0$. Combined with $\overline{\nabla}_K v = 0$ this establishes that $v_\sigma - v_K = 0$ for all $\sigma \in \mathcal{F}_K$. Reasoning from neighbour to neighbour we infer that v is the constant vector, which means that it is zero since $v_\sigma = 0$ for all $\sigma \in \mathcal{F}_{\mathrm{ext}}$.

Let us now estimate $\mathrm{reg}_{\mathrm{LLE}}(\mathcal{D})$. The first and last terms in the definition of this regularity factor are easy to bound since, for all $i \in I_K$, $\mathrm{dist}(\boldsymbol{x}_i, K) = 0$ and $\sum_{i \in I_K} |\pi_K^i(\boldsymbol{x})| = 1$. Let us estimate the term $\|\mathcal{G}_K\|_p$.
Take $v = (v_K, (v_\sigma)_{\sigma \in \mathcal{F}_K})$ and write, using the power-of-sums inequality (D.12) and the definitions (13.4) and (13.18) of $\mathcal{G}_K$ and $\zeta_{\mathcal{D}}$,

$$\begin{aligned}
\|\mathcal{G}_K v\|_{L^p(K)^d}^p &= \sum_{\sigma \in \mathcal{F}_K} |D_{K,\sigma}| \, |(\mathcal{G}_K v)_{|D_{K,\sigma}}|^p \\
&\leq 2^{p-1} \left(|K| \left|\overline{\nabla}_K v\right|^p + d^{\frac{p}{2}} \sum_{\sigma \in \mathcal{F}_K} |D_{K,\sigma}| \left| \frac{[\mathcal{L}_K R_K(v)]_\sigma}{d_{K,\sigma}} \right|^p \right) \\
&\leq 2^{p-1} \left(|K| \left|\overline{\nabla}_K v\right|^p + \zeta_{\mathcal{D}} d^{\frac{p}{2}} \sum_{\sigma \in \mathcal{F}_K} |D_{K,\sigma}| \left| \frac{R_{K,\sigma}(v)}{d_{K,\sigma}} \right|^p \right).
\end{aligned}$$

The definition (13.5) of $R_{K,\sigma}$, the power-of-sums inequality (D.12), the estimate $|\overline{\boldsymbol{x}}_\sigma - \boldsymbol{x}_K| \leq h_K \leq \theta_{\mathfrak{T}} d_{K,\sigma}$, and the relation $\sum_{\sigma \in \mathcal{F}_K} |D_{K,\sigma}| = |K|$ then give

$$\begin{aligned}
\|\mathcal{G}_K v\|_{L^p(K)^d}^p \leq 2^{p-1} \Bigg(&|K| \left|\overline{\nabla}_K v\right|^p + 2^{p-1} \zeta_{\mathcal{D}} d^{\frac{p}{2}} \sum_{\sigma \in \mathcal{F}_K} |D_{K,\sigma}| \left| \frac{v_\sigma - v_K}{d_{K,\sigma}} \right|^p \\
&+ 2^{p-1} \zeta_{\mathcal{D}} d^{\frac{p}{2}} \theta_{\mathfrak{T}}^p |K| \left|\overline{\nabla}_K v\right|^p \Bigg).
\end{aligned} \tag{13.20}$$

Integrating (B.30) over $\boldsymbol{x} \in K$ yields

$$|K|\,|\overline{\nabla}_K v|^p \le d^{p-1} \sum_{\sigma\in\mathcal{F}_K} |\sigma| d_{K,\sigma} \left|\frac{v_\sigma - v_K}{d_{K,\sigma}}\right|^p .$$

Plugging this estimate into (13.20) and recalling that $|D_{K,\sigma}| = |\sigma| d_{K,\sigma}/d$ (see (B.1)), we obtain

$$\|\mathcal{G}_K v\|^p_{L^p(K)^d} \le C_2 \sum_{\sigma\in\mathcal{F}_K} |\sigma| d_{K,\sigma} \left|\frac{v_\sigma - v_K}{d_{K,\sigma}}\right|^p . \tag{13.21}$$

where C_2 depends only on ϱ, p and d. Apply this estimate to $v = v^K$ or $v = v^\sigma$ (defined in Item 3 of Section 13.1.1) and use $|\sigma| d_{K,\sigma} = d|D_{K,\sigma}| \le d|K|$ and $d_{K,\sigma}^{-1} \le \theta_{\mathfrak{T}} h_K^{-1}$ for all $\sigma \in \mathcal{F}_K$ to obtain $\left\|\mathcal{G}_K^i\right\|_{L^p(K)^d} \le (C_2 d)^{1/p} \theta_{\mathfrak{T}} |K|^{1/p} h_K^{-1}$ for all $i \in I_K$. Recalling the definition (7.24) of $\|\mathcal{G}_K\|_p$ and the fact that $\mathrm{Card}(I_K) = 1 + \mathrm{Card}(\mathcal{F}_K)$, we infer that

$$\|\mathcal{G}_K\|_p \le (C_2 d)^{1/p} \theta_{\mathfrak{T}} (1 + \mathrm{Card}(\mathcal{F}_K)) \le (C_2 d)^{1/p} \theta_{\mathfrak{T}} (1 + \varrho).$$

The proof of (13.19) is complete. ∎

Lemma 13.11. *Let $\mathfrak{T}$ be a polytopal mesh of Ω in the sense of Definition 7.2, and let $\mathcal{D}$ be an HMM GD on $\mathfrak{T}$ as in Section 13.1.1. We take $\varrho \ge \theta_{\mathfrak{T}} + \zeta_{\mathcal{D}}$ (see (7.8) and (13.18)). Then, there exists $C_3 > 0$ depending only on Ω, p and ϱ, such that*

$$\forall v \in X_{\mathcal{D}}\,,\ \frac{1}{C_3} \|\nabla_{\mathcal{D}} v\|_{L^p(\Omega)^d} \le |v|_{\mathfrak{T},p} \le C_3 \|\nabla_{\mathcal{D}} v\|_{L^p(\Omega)^d} . \tag{13.22}$$

Remark 13.12. The first inequality in (13.22) is used in Theorem 13.14 to check Condition (2.96) and thus establish the GD-consistency of sequences of HMM gradient discretisations. The second inequality in (13.22) is used in Lemma 13.13 to control such sequences by polytopal toolboxes, and thus prove their coercivity, limit-conformity and compactness.

Proof. In this proof, $A \lesssim B$ means that $A \le MB$ for some M depending only on Ω, p and ϱ. Let $v \in X_{\mathcal{D}}$. The first inequality in (13.22) follows simply by summing (13.21) over $K \in \mathcal{M}$. Let us therefore turn to the second inequality in (13.22). Relation (13.10) and Jensen's inequality give

$$|\overline{\nabla}_K v|^p \le \frac{1}{|K|} \int_K |\nabla_{\mathcal{D}} v(\boldsymbol{x})|^p \mathrm{d}\boldsymbol{x}. \tag{13.23}$$

By definition (13.6) of $\nabla_{\mathcal{D}} v$ and by the power-of-sums inequality (D.12), we deduce that, for a.e. $\boldsymbol{y} \in D_{K,\sigma}$,

$$\left|\frac{\sqrt{d}}{d_{K,\sigma}} [\mathcal{L}_K R_K(v)]_\sigma \boldsymbol{n}_{K,\sigma}\right|^p = \left|\nabla_{\mathcal{D}} v(\boldsymbol{y}) - \overline{\nabla}_K v\right|^p$$

$$\lesssim |\nabla_{\mathcal{D}} v(\boldsymbol{y})|^p + \frac{1}{|K|}\int_K |\nabla_{\mathcal{D}} v(\boldsymbol{x})|^p \mathrm{d}\boldsymbol{x}.$$

Integrating over $\boldsymbol{y} \in D_{K,\sigma}$ and summing over $\sigma \in \mathcal{F}_K$ leads to

$$\sum_{\sigma\in\mathcal{F}_K} |D_{K,\sigma}| \left|\frac{[\mathcal{L}_K R_K(v)]_\sigma}{d_{K,\sigma}}\right|^p \lesssim \int_K |\nabla_{\mathcal{D}} v(\boldsymbol{x})|^p \mathrm{d}\boldsymbol{x}. \tag{13.24}$$

Use then the definition (13.18) of $\zeta_{\mathcal{D}}$ to write

$$\sum_{\sigma\in\mathcal{F}_K} |D_{K,\sigma}| \left|\frac{R_{K,\sigma}(v)}{d_{K,\sigma}}\right|^p \lesssim \int_K |\nabla_{\mathcal{D}} v(\boldsymbol{x})|^p \mathrm{d}\boldsymbol{x}. \tag{13.25}$$

By definition (13.5) of $R_{K,\sigma}$, and since $|\overline{\boldsymbol{x}}_\sigma - \boldsymbol{x}_K| \le h_K \le \theta_{\mathfrak{T}} d_{K,\sigma}$, we have $|v_\sigma - v_K| \lesssim |R_{K,\sigma}(v)| + |\overline{\nabla}_K v| d_{K,\sigma}$. Hence, recalling (13.23) and using (13.25),

$$\sum_{\sigma\in\mathcal{F}_K} |D_{K,\sigma}| \left|\frac{v_\sigma - v_K}{d_{K,\sigma}}\right|^p \lesssim \int_K |\nabla_{\mathcal{D}} v(\boldsymbol{x})|^p \mathrm{d}\boldsymbol{x}.$$

Since $|D_{K,\sigma}| = \frac{|\sigma| d_{K,\sigma}}{d}$ (cf. (B.1)), summing the above relation over $K \in \mathcal{M}$ and recalling the definition (7.7f) of $|\cdot|_{\mathfrak{T},p}$, the second inequality in (13.22) is proved. ∎

We can now define, and state estimates on, a control of an HMM GD by a polytopal toolbox.

Lemma 13.13 (Control of an HMM GD by a polytopal toolbox). *Let $\mathfrak{T}$ be a polytopal mesh of Ω in the sense of Definition 7.2, and let $\mathcal{D}$ be an HMM GD on $\mathfrak{T}$ as in Section 13.1.1. Take $\varrho \ge \theta_{\mathfrak{T}} + \zeta_{\mathcal{D}}$ (see (7.8) and (13.18)) and define the control $\boldsymbol{\Phi} = \mathrm{Id} : X_{\mathcal{D},0} \to X_{\mathfrak{T},0}$ of $\mathcal{D}$ by $\mathfrak{T}$ (see Definition 7.9). Then, there exists $C_3 > 0$ depending only on Ω, p and ϱ, such that*

$$\|\boldsymbol{\Phi}\|_{\mathcal{D},\mathfrak{T}} \le C_3, \tag{13.26}$$

and

$$\omega^{\Pi}(\mathcal{D},\mathfrak{T},\boldsymbol{\Phi}) = 0\,, \quad \omega^{\nabla}(\mathcal{D},\mathfrak{T},\boldsymbol{\Phi}) = 0. \tag{13.27}$$

Proof. Estimate (13.26) is given by Lemma 13.11. The first relation in (13.27) follows from $\Pi_{\mathcal{D}} v = \Pi_{\mathfrak{T}} v = \Pi_{\mathfrak{T}} \boldsymbol{\Phi}(v)$ (see (13.2)). The second relation in (13.27) is a straightforward consequence of (13.10). ∎

13.1.3 Properties of HMM gradient discretisations

Thanks to the previous lemmas, the proof of the properties of HMM GDs is straightforward.

Theorem 13.14 (Properties of HMM GDs for Dirichlet BCs). *Let $(\mathcal{D}_m)_{m\in\mathbb{N}}$ be a sequence of HMM GDs, as in Section 13.1.1, defined from underlying polytopal meshes $(\mathfrak{T}_m)_{m\in\mathbb{N}}$. Assume that $(\theta_{\mathfrak{T}_m}+\eta_{\mathfrak{T}_m})_{m\in\mathbb{N}}$ and $(\zeta_{\mathcal{D}_m})_{m\in\mathbb{N}}$ are bounded (see (7.8), (7.9) and (13.18)), and that $h_{\mathcal{M}_m}\to 0$ as $m\to\infty$.*
Then, the sequence $(\mathcal{D}_m)_{m\in\mathbb{N}}$ is coercive, GD-consistent, limit-conforming and compact in the sense of Definitions 2.2, 2.51, 2.5 and 2.8. Moreover, each $\mathcal{D}_m$ has a piecewise constant reconstruction in the sense of Definition 2.12.

Proof. The limit-conformity, coercivity and compactness are obtained by applying Corollary 7.12, thanks to Lemma 13.13. The property of piecewise constant reconstruction is also straightforward from (13.2) (using the notations in Definition 2.12, one simply chooses $\Omega_K=K$ if $K\in\mathcal{M}$ and $\Omega_\sigma=\emptyset$ if $\sigma\in\mathcal{F}$).
To prove the GD-consistency, we aim at applying Proposition 7.53. The bound on $\mathrm{reg}_{\mathrm{LLE}}(\mathcal{D}_m)$ being provided by Lemma 13.10, we just have to check that (7.65) and (2.96) hold. We drop the index m in $\mathcal{D}_m$.
Let $\varphi\in C^\infty(\overline{\Omega})$ and $\sigma\in\mathcal{F}_K$. We have, for $\boldsymbol{x}\in\sigma$, by Taylor's expansion

$$\varphi(\boldsymbol{x})=\varphi(\overline{\boldsymbol{x}}_\sigma)+\nabla\varphi(\overline{\boldsymbol{x}}_\sigma)\cdot(\boldsymbol{x}-\overline{\boldsymbol{x}}_\sigma)+\mathcal{R}_\sigma(\boldsymbol{x})$$

where $|\mathcal{R}_\sigma(\boldsymbol{x})|\le\frac{1}{2}\mathrm{diam}(\sigma)^2\sup_{\overline{\Omega}}|D^2\varphi|$. Hence, taking the average over $\boldsymbol{x}\in\sigma$ and recalling the definition (13.8) of $\mathcal{I}_{\mathcal{D},\partial}$, since $\frac{1}{|\sigma|}\int_\sigma\boldsymbol{x}\mathrm{d}\gamma(\boldsymbol{x})=\overline{\boldsymbol{x}}_\sigma$,

$$|(\mathcal{I}_{\mathcal{D},\partial}\gamma(\varphi))_\sigma-\varphi(\overline{\boldsymbol{x}}_\sigma)|\le\frac{1}{2}\mathrm{diam}(\sigma)^2\sup_{\overline{\Omega}}|D^2\varphi|.$$

Since $\mathrm{diam}(\sigma)\le\mathrm{diam}(K)$ for any $\sigma\in\mathcal{F}_K$, this proves that (7.65) holds.
To prove (2.96), take $\varphi\in W^{1,p}(\Omega)$ and consider $v=I_{\mathfrak{T}}\varphi$, where $I_{\mathfrak{T}}$ is the interpolant given by (B.10). By definition of the face values of v and by choice (13.8) of the interpolant $\mathcal{I}_{\mathcal{D},\partial}$, we have $v-\mathcal{I}_{\mathcal{D},\partial}\gamma\varphi\in X_{\mathcal{D},0}$. Moreover, by Lemma 13.11 and Proposition B.7,

$$\begin{aligned}\|\Pi_{\mathcal{D}}v\|_{L^p(\Omega)}+\|\nabla_{\mathcal{D}}v\|_{L^p(\Omega)^d}&\le\|\varphi\|_{L^p(\Omega)}+C_3\,|I_{\mathfrak{T}}\varphi|_{\mathfrak{T},p}\\&\le\|\varphi\|_{L^p(\Omega)}+C_4\,\|\nabla\varphi\|_{L^p(\Omega)^d}\end{aligned}$$

with C_4 depending only on d, p and an upper bound of $\theta_{\mathfrak{T}}+\zeta_{\mathcal{D}}$. This concludes the proof of (2.96). ∎

The following two propositions, also easy consequences of the preliminary results in the preceding section, are useful to establish error estimates for HMM GSs.

Proposition 13.15 (Estimate on $C_{\mathcal{D}}$ and $W_{\mathcal{D}}$ for HMM GD – Dirichlet BCs). *Let $\mathfrak{T}$ be a polytopal mesh of Ω in the sense of Definition 7.2, and let $\mathcal{D}$ be an HMM GD on $\mathfrak{T}$ as in Section 13.1.1. Take $\varrho\ge\theta_{\mathfrak{T}}+\eta_{\mathfrak{T}}+\zeta_{\mathcal{D}}$ (see*

(7.8), (7.9) and (13.18)). Then, there exists C_5 depending only on Ω, p, any ϱ, such that

$$C_{\mathcal{D}} \leq C_5 \tag{13.28}$$

and

$$\forall \boldsymbol{\varphi} \in W^{1,p'}(\Omega)^d\,,\ W_{\mathcal{D}}(\boldsymbol{\varphi}) \leq C_5 \left\|\boldsymbol{\varphi}\right\|_{W^{1,p'}(\Omega)^d} h_{\mathcal{M}}. \tag{13.29}$$

Here, $C_{\mathcal{D}}$ and $W_{\mathcal{D}}$ are the coercivity constant and limit-conformity measure defined by (2.1) and (2.6).

Proof. The conclusion immediately follows from Theorem 7.11 and Lemma 13.13. ∎

Proposition 13.16 (Estimate on $S_{\mathcal{D}}$ for HMM GD – Dirichlet BCs). *Let $\mathfrak{T}$ be a polytopal mesh of Ω in the sense of Definition 7.2, and $\mathcal{D}$ be an HMM GD on $\mathfrak{T}$ as in Section 13.1.1. Assume that $p > d/2$ and take $\varrho \geq \theta_{\mathfrak{T}} + \zeta_{\mathcal{D}}$ (see (7.8) and (13.18)). Let $\varphi \in W^{2,p}(\Omega)$ and, as in Remark 13.4, re-define $\mathcal{I}_{\mathcal{D},\partial}\gamma(\varphi)$ by: $(\mathcal{I}_{\mathcal{D},\partial}\gamma(\varphi))_\sigma = \varphi(\overline{\boldsymbol{x}}_\sigma)$ for all $\sigma \in \mathcal{F}_{\rm ext}$ (this makes sense since $\varphi \in C(\overline{\Omega})$). Then, there exists $C_6 > 0$, depending only on Ω, p and ϱ, such that*

$$S_{\mathcal{D}}(\varphi) \leq C_6 \left\|\varphi\right\|_{W^{2,p}(\Omega)} h_{\mathcal{M}},$$

where $S_{\mathcal{D}}$ is defined by (2.94). This means that the space size (obtained by an easy adaptation of Definition 2.22 to non-zero Dirichlet BCs) of the GD is such that

$$h_{\mathcal{D}}(W^{2,p}(\Omega); W^{1,p'}(\Omega)^d) \leq \max(C_6, C_5) h_{\mathcal{M}},$$

where C_5 is defined in Proposition 13.15.

Proof. By Lemma B.1, each cell K is star-shaped with respect to a ball of radius $\min_{\sigma\in\mathcal{F}_K} d_{K,\sigma} \geq \theta_{\mathfrak{T}}^{-1} h_K \geq \varrho^{-1} h_K$. Moreover, for all $K \in \mathcal{M}$ and all $i \in I_K$ we have $\boldsymbol{x}_i \in \overline{K}$, which shows that (7.84) holds. Using Lemma 13.10, Proposition 7.68 can be applied and the result follows immediately. ∎

The application of Propositions 13.16 and 13.15 to the estimates (2.25) and (2.26) in Theorem 2.28 provides an $\mathcal{O}(h_{\mathcal{M}})$ error in the case of a linear elliptic problem in one, two or three space dimensions, when the solution belongs to $H^2(\Omega)$.

Remark 13.17 (Non piecewise constant diffusion tensor)
If Λ is not piecewise constant on $\mathcal{M}$, then (13.9) is the GS (2.23) for the problem (2.22) with Λ is replaced with its piecewise projection on the mesh, *i.e.* (13.9) is the GS for

$$\begin{aligned}
&\overline{u}_{\mathcal{M}} \in H_0^1(\Omega),\ \forall v \in H_0^1(\Omega),\\
&\int_\Omega \Lambda_{\mathcal{M}}(\boldsymbol{x}) \nabla \overline{u}_{\mathcal{M}}(\boldsymbol{x}) \cdot \nabla v(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} = \int_\Omega f(\boldsymbol{x}) v(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} - \int_\Omega \boldsymbol{F}(\boldsymbol{x}) \cdot \nabla v(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}
\end{aligned}$$

where $(\Lambda_{\mathcal{M}})_{|K} = \frac{1}{|K|}\int_K \Lambda(\boldsymbol{x})\mathrm{d}\boldsymbol{x}$ for all $K \in \mathcal{M}$. Assuming that Λ is Lipschitz-continuous inside each cell, we have $\|\Lambda - \Lambda_{\mathcal{M}}\|_{L^\infty(\Omega)} \le Ch_{\mathcal{M}}$ and thus, denoting by $\overline{u}$ the solution to (2.22), subtracting the equations satisfied by $\overline{u}_{\mathcal{M}}$ and $\overline{u}$ and taking $v = \overline{u}_{\mathcal{M}} - \overline{u}$ as a test function, we obtain

$$\begin{aligned}\underline{\lambda}\,\|\nabla\overline{u}_{\mathcal{M}} - \nabla\overline{u}\|^2_{L^2(\Omega)^d} &\le \int_\Omega \Lambda_{\mathcal{M}}(\boldsymbol{x})(\nabla\overline{u}_{\mathcal{M}} - \nabla\overline{u})(\boldsymbol{x})\cdot(\nabla\overline{u}_{\mathcal{M}} - \nabla\overline{u})(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\\ &= \int_\Omega (\Lambda(\boldsymbol{x}) - \Lambda_{\mathcal{M}}(\boldsymbol{x}))\nabla\overline{u}(\boldsymbol{x})\cdot(\nabla\overline{u}_{\mathcal{M}} - \nabla\overline{u})(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\\ &\le Ch_{\mathcal{M}}\,\|\nabla\overline{u}\|_{L^2(\Omega)}\,\|\nabla\overline{u}_{\mathcal{M}} - \nabla\overline{u}\|_{L^2(\Omega)^d}\,.\end{aligned}$$

This shows that $\|\overline{u}_{\mathcal{M}} - \overline{u}\|_{H^1_0(\Omega)} = \mathcal{O}(h_{\mathcal{M}})$. If u is the solution to the HMM scheme (13.9), the estimates in Section 2.2.2 and in Propositions 13.16 and 13.15 show that $\|\overline{u}_{\mathcal{M}} - \Pi_{\mathcal{D}}u\|_{L^2(\Omega)} + \|\nabla\overline{u}_{\mathcal{M}} - \nabla_{\mathcal{D}}u\|_{L^2(\Omega)^d} = \mathcal{O}(h_{\mathcal{M}})$. Hence, we see that $\|\overline{u} - \Pi_{\mathcal{D}}u\|_{L^2(\Omega)} + \|\nabla\overline{u} - \nabla_{\mathcal{D}}u\|_{L^2(\Omega)^d} = \mathcal{O}(h_{\mathcal{M}})$.
In other words, the replacement of Λ by its piecewise constant approximation in (2.22), and the approximation of this latter equation by an HMM GS, does not impact the expected rates of convergence. Assuming that Λ is piecewise constant is therefore not extremely restrictive, especially since it is the case in many practical applications.

13.2 HMM methods, Neumann and Fourier boundary conditions

13.2.1 Neumann boundary conditions

Following Definition 7.54, an HMM GD for homogeneous Neumann boundary conditions simply consists in defining $X_{\mathcal{D}}$, $\Pi_{\mathcal{D}}$ and $\nabla_{\mathcal{D}}$ as in Items 1, 2 and 3 in Section 13.1.1.
If $v \in X_{\mathcal{D}} = X_{\mathfrak{T}}$ and $\|\nabla_{\mathcal{D}}v\|_{L^p(\Omega)^d} = 0$, then Inequality (13.22) and the definition (7.7f) of $|\cdot|_{\mathfrak{T},p}$ show that all $(v_K)_{K\in\mathcal{M}}$ and all $(v_\sigma)_{\sigma\in\mathcal{F}_K}$ are identical. Hence, the definition of $\Pi_{\mathcal{D}}$ shows that the quantity (3.1) is indeed a norm on $X_{\mathcal{D}}$.

For non-homogeneous Neumann boundary conditions, we take as trace reconstruction $\mathbb{T}_{\mathcal{D}} : X_{\mathcal{D}} \to L^p(\partial\Omega)$ the operator $\mathbb{T}_{\mathfrak{T}}$ (see (7.7d)), that is,

$$\forall v \in X_{\mathcal{D}}\,,\ \forall \sigma \in \mathcal{F}_{\mathrm{ext}}\ :\ \mathbb{T}_{\mathcal{D}}v = \mathbb{T}_{\mathfrak{T}}v = v_\sigma \text{ on } \sigma. \tag{13.30}$$

Since the regularity factor $\mathrm{reg}_{\mathrm{LLE}}(\mathcal{D})$ is defined as for Dirichlet boundary conditions, Lemma 13.10 still applies and shows that this factor remains bounded if $\theta_{\mathfrak{T}}$ and $\zeta_{\mathcal{D}}$ are bounded. Defining the control $\boldsymbol{\Phi} = \mathrm{Id} : X_{\mathcal{D}} \to X_{\mathfrak{T}}$ of an HMM GD $\mathcal{D}$ for Neumann boundary conditions by $\mathfrak{T}$, we see that Lemma 13.13 still holds and that $\omega^{\mathbb{T}}(\mathcal{D},\mathfrak{T},\boldsymbol{\Phi}) = 0$. Hence, Corollary 7.18 and Proposition 7.55 give the following theorem.

Theorem 13.18 (Properties of HMM GDs for Neumann BCs). *Let $(\mathcal{D}_m)_{m\in\mathbb{N}}$ be a sequence of HMM GDs for Neumann boundary conditions as above, defined from underlying polytopal meshes $(\mathfrak{T}_m)_{m\in\mathbb{N}}$. Assume that $(\theta_{\mathfrak{T}_m} + \eta_{\mathfrak{T}_m})_{m\in\mathbb{N}}$ and $(\zeta_{\mathcal{D}_m})_{m\in\mathbb{N}}$ are bounded (see (7.8), (7.9) and (13.18)), and that $h_{\mathcal{M}_m} \to 0$ as $m \to \infty$.*
Then the sequence $(\mathcal{D}_m)_{m\in\mathbb{N}}$ is coercive, GD-consistent, limit-conforming and compact in the sense of Definitions 3.13, 3.4, 3.14 and 3.15. Moreover, each $\mathcal{D}_m$ has a piecewise constant reconstruction in the sense of Definition 2.12.

Proposition 7.70 and Theorem 7.17 also give estimates on $S_{\mathcal{D}}$, $C_{\mathcal{D}}$ and $W_{\mathcal{D}}$ that are similar to those in Propositions 13.16 and 13.15. The constants depend only on an upper bound of $\theta_{\mathfrak{T}} + \zeta_{\mathcal{D}}$ (for $S_{\mathcal{D}}$), or of $\theta_{\mathfrak{T}} + \eta_{\mathfrak{T}} + \zeta_{\mathcal{D}}$ (for $C_{\mathcal{D}}$ and $W_{\mathcal{D}}$).

13.2.2 Fourier boundary conditions

Starting from an HMM GD for Dirichlet boundary conditions, we follow Definition 7.57 in Section 7.3.6 to define an HMM GD for Fourier boundary conditions.
The boundary mesh $\mathcal{M}_\partial$ is simply $\mathcal{F}_{\text{ext}}$, and the reconstructed trace (13.30) corresponds to $I_\sigma = \{\sigma\}$ and $\pi_\sigma^\sigma = 1$. The bound on $\text{reg}_{\text{LLE}}(\mathcal{D})$ for Fourier boundary conditions therefore easily follows from the bound on this quantity for Dirichlet boundary conditions, and the consistency (under boundedness of $\theta_{\mathfrak{T}_m} + \zeta_{\mathcal{D}_m}$) is therefore a consequence of Proposition 7.58.
As noticed in Remark 7.20, the work done for Neumann boundary conditions then immediately show that Theorem 13.18 also applies to Fourier boundary conditions. Similarly, we could obtain estimates on $S_{\mathcal{D}}$, $C_{\mathcal{D}}$ and $W_{\mathcal{D}}$ as in Propositions 13.16 and 13.15.

13.3 HMM fluxes and link with the two-point finite volume method

Let us define the family of fluxes $(F_{K,\sigma})_{K\in\mathcal{M},\sigma\in\mathcal{F}_K}$ as the linear mappings on $X_{\mathcal{D}}$ such that

$$\forall u, v \in X_{\mathcal{D}},\ \forall K \in \mathcal{M}\ :\ \sum_{\sigma\in\mathcal{F}_K} F_{K,\sigma}(u)(v_K - v_\sigma) = \int_K \Lambda(\boldsymbol{x})\nabla_{\mathcal{D}} u(\boldsymbol{x}) \cdot \nabla_{\mathcal{D}} v(\boldsymbol{x})\mathrm{d}\boldsymbol{x}. \tag{13.31}$$

The existence and uniqueness of these fluxes is ensured by the following proposition.

Proposition 13.19 (Existence and uniqueness of the fluxes). *There exists a unique family of linear mappings $(F_{K,\sigma})_{K\in\mathcal{M},\,\sigma\in\mathcal{F}_K}$ that satisfy (13.31).*

Proof. Let $u \in X_{\mathcal{D}}$ and assume that $(F_{K,\sigma}(u))_{K,\sigma}$ a solution of (13.31). Take $K \in \mathcal{M}$, $\sigma \in \mathcal{F}_K$, and let $w^\sigma \in X_{\mathfrak{T}}$ be such that $w^\sigma_\sigma = 1$, $w^\sigma_{\sigma'} = 0$ for all $\sigma \neq \sigma'$, and $w^\sigma_L = 0$ for all $L \in \mathcal{M}$. Substituting $v = w^\sigma$ in (13.31) gives

$$F_{K,\sigma}(u) = -\int_K \Lambda(\boldsymbol{x})\nabla_{\mathcal{D}} u(\boldsymbol{x}) \cdot \nabla_{\mathcal{D}} w^\sigma(\boldsymbol{x})\mathrm{d}\boldsymbol{x}, \tag{13.32}$$

which determines uniquely $F_{K,\sigma}(u)$, since w^σ depends only on σ. This formula also clearly shows that $u \in X_{\mathcal{D}} \to F_{K,\sigma}(u)$ is linear.
We now prove that the fluxes defined by (13.32) satisfy (13.31). Fix a cell $K \in \mathcal{M}$ and let $v \in X_{\mathfrak{T}}$. Multiplying (13.32) by $v_K - v_\sigma$ and summing on $\sigma \in \mathcal{F}_K$ gives

$$\begin{aligned}
&\sum_{\sigma \in \mathcal{F}_K} F_{K,\sigma}(u)(v_K - v_\sigma) \\
&\qquad = \int_K \Lambda(\boldsymbol{x})\nabla_{\mathcal{D}} u(\boldsymbol{x}) \cdot \nabla_{\mathcal{D}} \left(\sum_{\sigma \in \mathcal{F}_K} (v_\sigma - v_K) w^\sigma \right)(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \\
&\qquad = \int_K \Lambda(\boldsymbol{x})\nabla_{\mathcal{D}} u(\boldsymbol{x}) \cdot \nabla_{\mathcal{D}} V(\boldsymbol{x})\mathrm{d}\boldsymbol{x},
\end{aligned} \tag{13.33}$$

where $V = \sum_{\sigma \in \mathcal{F}_K} (v_\sigma - v_K) w^\sigma \in X_{\mathcal{D}}$. V has components $V_{\sigma'} = v_{\sigma'} - v_K$ for all $\sigma' \in \mathcal{F}_K$, and $V_{\sigma''} = V_L = 0$ for all $\sigma'' \notin \mathcal{F}_K$ and all $L \in \mathcal{M}$. We therefore have, by definition (7.7e) of $\overline{\nabla}_K$,

$$\overline{\nabla}_K V = \frac{1}{|K|} \sum_{\sigma \in \mathcal{F}_K} |\sigma| V_\sigma \boldsymbol{n}_{K,\sigma} = \frac{1}{|K|} \sum_{\sigma \in \mathcal{F}_K} |\sigma| (v_\sigma - v_K) \boldsymbol{n}_{K,\sigma} = \overline{\nabla}_K v.$$

Moreover, for any $\sigma \in \mathcal{F}_K$, $V_\sigma - V_K = v_\sigma - v_K$. Hence, by (13.4) and (13.5) we see that $\nabla_{\mathcal{D}} V = \nabla_{\mathcal{D}} v$ on K. Equation (13.33) therefore shows that (13.31) is satisfied. ∎

The GS for (2.20) (with $\boldsymbol{F} = 0$) then corresponds to writing the flux conservativity and flux balances (see [75]):

$$\forall \sigma \in \mathcal{F}_{\mathrm{int}} \,:\, F_{K,\sigma}(u) + F_{L,\sigma}(u) = 0, \tag{13.34}$$

$$\forall K \in \mathcal{M} \,:\, \sum_{\sigma \in \mathcal{F}_K} F_{K,\sigma}(u) = \int_K f(\boldsymbol{x})\mathrm{d}\boldsymbol{x}. \tag{13.35}$$

The HMM method is therefore a finite volume scheme (more precisely, the mixed finite volume, see [75]).

For specific meshes and with $\Lambda = \mathrm{Id}$, the flux $F_{K,\sigma}(u)$ actually depends only on the values u_K and u_σ.

Lemma 13.20 (Superadmissible mesh and two-point flux). *Let $\mathfrak{T}$ be a polytopal mesh of Ω in the sense of Definition 7.2. We assume that the following superadmissibility condition is satisfied:*

$$\forall K \in \mathcal{M}\,,\ \forall \sigma \in \mathcal{F}_K\ :\ \boldsymbol{n}_{K,\sigma} = \frac{\overline{\boldsymbol{x}}_\sigma - \boldsymbol{x}_K}{d_{K,\sigma}} \tag{13.36}$$

*(*i.e. *the orthogonal projection of* $\boldsymbol{x}_K$ *on each face* $\sigma \in \mathcal{F}_K$ *is the centre of mass* $\overline{\boldsymbol{x}}_\sigma$ *of* σ*). Assume that, for each* $K \in \mathcal{M}$*,* $\Lambda_K = \mathrm{Id}$ *and* $\mathcal{L}_K = \mathrm{Id}$*. Then*

$$\int_K \nabla_{\mathcal{D}} u(\boldsymbol{x}) \cdot \nabla_{\mathcal{D}} v(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} = \sum_{\sigma \in \mathcal{F}_K} \frac{|\sigma|}{d_{K,\sigma}} (u_K - u_\sigma)(v_K - v_\sigma) \tag{13.37}$$

and the fluxes defined by (13.31) *are given by*

$$F_{K,\sigma}(u) = \frac{|\sigma|}{d_{K,\sigma}} (u_K - u_\sigma).$$

A similar lemma can be proved [75] for isotropic Λ, *i.e.* $\Lambda(\boldsymbol{x}) = \lambda(\boldsymbol{x})\mathrm{Id}$.
The superadmissibility condition is satisfied by rectangles (with $\boldsymbol{x}_K$ the centre of mass of K) and acute triangles (with $\boldsymbol{x}_K$ the circumcenter of K) in 2D, and by rectangular parallelepipeds (with $\boldsymbol{x}_K$ the centre of mass of K) in 3D. It is unfortunately not satisfied by tetrahedra in general.

Proof. Since $\Lambda_K = \mathrm{Id}$, the choice $\mathcal{L}_K = \mathrm{Id}$ and Equation (13.15) give

$$\begin{aligned}\int_K \nabla_{\mathcal{D}} u(\boldsymbol{x}) \cdot \nabla_{\mathcal{D}} v(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \\ = |K| \overline{\nabla}_K u \cdot \overline{\nabla}_K v + \sum_{\sigma \in \mathcal{F}_K} \frac{|\sigma|}{d_{K,\sigma}} R_{K,\sigma}(u) R_{K,\sigma}(v).\end{aligned} \tag{13.38}$$

Thanks to Assumption (13.36), the reconstructed gradient may be written

$$\overline{\nabla}_K v = \frac{1}{|K|} \sum_{\sigma \in \mathcal{F}_K} \frac{|\sigma|}{d_{K,\sigma}} (v_\sigma - v_K)(\overline{\boldsymbol{x}}_\sigma - \boldsymbol{x}_K).$$

Using again (13.36), Equation (B.2) gives $\sum_{\sigma \in \mathcal{F}_K} \frac{|\sigma|}{d_{K,\sigma}} (\overline{\boldsymbol{x}}_\sigma - \boldsymbol{x}_K)(\overline{\boldsymbol{x}}_\sigma - \boldsymbol{x}_K)^T = |K|\mathrm{Id}$ and therefore, recalling the definition (13.5) of $R_{K,\sigma}$,

$$\begin{aligned}\sum_{\sigma \in \mathcal{F}_K} \frac{|\sigma|}{d_{K,\sigma}} R_{K,\sigma}(u) R_{K,\sigma}(v) \\ = \sum_{\sigma \in \mathcal{F}_K} \frac{|\sigma|}{d_{K,\sigma}} (u_\sigma - u_K)(v_\sigma - v_K) - |K| \overline{\nabla}_K u \cdot \overline{\nabla}_K v.\end{aligned}$$

Plugged into (13.38), this yields (13.37). The expressions of $F_{K,\sigma}$ are then obtained by comparing (13.37) and (13.31). ∎

13.4 A cell-centred variant of HMM schemes on Δ-admissible meshes

Let us consider a Δ-admissible mesh in the sense of [91]. We recall that in the case of a Δ-admissible polytopal mesh, the line $(\boldsymbol{x}_K, \boldsymbol{x}_L)$ is orthogonal to the interface σ. A discrete gradient was constructed in [93] to deal with anisotropic problems on such meshes. Let us set $X_{\mathcal{D},0} = \{(v_K)_{K\in\mathcal{M}} : v_K \in \mathbb{R}\}$ and define d_σ and $\delta_{K,\sigma}u$, for all $u \in X_{\mathcal{D},0}$, by

$$\begin{array}{lll} d_\sigma = d_{K,\sigma} + d_{L,\sigma} & \text{and } \delta_{K,\sigma}u = u_L - u_K, & \forall \sigma = K|L \in \mathcal{F}_{\text{int}}, \\ d_\sigma = d_{K,\sigma} & \text{and } \delta_{K,\sigma}u = -u_K, & \forall \sigma \in \mathcal{F}_K \cap \mathcal{F}_{\text{ext}}. \end{array} \tag{13.39}$$

Let as before $\Pi_{\mathcal{D}}u \in L^2(\Omega)$ be the piecewise constant function equal to u_K in K. The gradient reconstruction $\nabla_{\mathcal{D}}u \in L^2(\Omega)^d$ is constructed in the following way. We start, as in HMM methods, by defining a constant gradient in each cell K, using a formula that accounts for the Δ-admissibility of the mesh:

$$\nabla_K u = \frac{1}{|K|} \sum_{\sigma\in\mathcal{F}_K} |\sigma|(\overline{\boldsymbol{x}}_\sigma - \boldsymbol{x}_K)\frac{\delta_{K,\sigma}u}{d_\sigma}. \tag{13.40}$$

We then let

$$R_{K,\sigma}(u) = \frac{\delta_{K,\sigma}u}{d_\sigma} - \nabla_K u \cdot \boldsymbol{n}_{K,\sigma} \tag{13.41}$$

and

$$\nabla_{K,\sigma}u = \nabla_K u + \sqrt{d}R_{K,\sigma}(u)(\overline{\boldsymbol{x}}_\sigma - \boldsymbol{x}_K). \tag{13.42}$$

Then, as in HMM methods, $\nabla_{\mathcal{D}}u \in L^2(\Omega)^d$ is the piecewise constant function defined by the value $\nabla_{K,\sigma}u$ in $D_{K,\sigma}$.
The mathematical analysis of the consistency and limit-conformity follows similar steps to that of standard HMM schemes. As in the case of an HMM method with $\mathcal{L}_K = \text{Id}$, this variant gives back the standard 2-point scheme for superadmissible meshes when $\Lambda = \text{Id}$.

13.5 The harmonic averaging points for heterogeneous domains

The SUSHI scheme [96] is probably the starting point in the development of the GDM method. In its hybrid version, it is an HMM scheme, but it also features a possible barycentric condensation, in which some of the face unknowns are eliminated. In its simplest form, a SUSHI GD is given by Definition 7.40 with $\mathcal{D}$ an HMM GD and $I^{\text{BA}} = \mathcal{M} \cup \mathcal{F}_{\text{hyb}}$ for some $\mathcal{F}_{\text{hyb}} \subset \mathcal{F}$. The face unknowns that are eliminated correspond to $\mathcal{F}_{\text{bary}} = \mathcal{F}\backslash\mathcal{F}_{\text{hyb}}$. If $\sigma \in \mathcal{F}_{\text{bary}}$, the points $(\boldsymbol{x}_i)_{i\in H_\sigma}$ used to eliminate the unknown associated with σ are located around σ. If σ is on or around a discontinuity of Λ, as discussed in Section

7.5.2 a linearly exact barycentric condensation as in Definition 7.40 leads to a poor approximation of the solution. The notion of $\mathcal{S}$-adapted barycentric condensation introduced in Definition 7.73 relaxes this requirement of a linearly exact condensation and is therefore particularly useful for the SUSHI GD.

13.5.1 Harmonic interpolation coefficients

We consider here $p = 2$, since this construction is mostly meaningful for linear problems. If Λ is discontinuous, the solution $\overline{u}$ to (2.20) is smooth in the regions where Λ is smooth, and has continuous fluxes where Λ is discontinuous. This describes a subset $\mathcal{S}$ of $H_0^1(\Omega)$. Here a $\mathcal{S}$-adapted SUSHI GD is presented; it produces better approximation results in the case of heterogeneous material. The construction of the interpolation families is based on the following result.

Lemma 13.21. *Let $K = \mathbb{R}^{d-1} \times (-\infty, 0)$ and $L = \mathbb{R}^{d-1} \times (0, \infty)$ be two half-spaces, and $\sigma = \mathbb{R}^{d-1} \times \{0\}$ be their interface. We consider a diffusion tensor Λ which is constant equal to Λ_K in K and constant equal to Λ_L in L. The vector $\boldsymbol{n}_{KL}$ is the unit vector in the direction $x_d > 0$. We take $\boldsymbol{x}_K \in K$ and $\boldsymbol{x}_L \in L$ and define $\boldsymbol{y}_K$ and $\boldsymbol{y}_L$ as the respective projections of $\boldsymbol{x}_K$ and $\boldsymbol{x}_L$ on σ. We let $d_{K,\sigma} = \mathrm{dist}(\boldsymbol{x}_K, \sigma)$ and $d_{L,\sigma} = \mathrm{dist}(\boldsymbol{x}_L, \sigma)$ and we define the dioptrical point $\boldsymbol{y}_\sigma \in \sigma$ by*

$$\boldsymbol{y}_\sigma = \frac{\lambda_L d_{K,\sigma} \boldsymbol{y}_L + \lambda_K d_{L,\sigma} \boldsymbol{y}_K}{\lambda_L d_{K,\sigma} + \lambda_K d_{L,\sigma}} + \frac{d_{K,\sigma} d_{L,\sigma}}{\lambda_L d_{K,\sigma} + \lambda_K d_{L,\sigma}} (\boldsymbol{\lambda}_L^t - \boldsymbol{\lambda}_K^t), \tag{13.43}$$

where $\lambda_K = \boldsymbol{n}_{KL} \cdot \Lambda_K \boldsymbol{n}_{KL}$, $\boldsymbol{\lambda}_K^t = (\Lambda_K - \lambda_K \mathrm{Id}) \boldsymbol{n}_{KL}$, $\lambda_L = \boldsymbol{n}_{KL} \cdot \Lambda_L \boldsymbol{n}_{KL}$ and $\boldsymbol{\lambda}_L^t = (\Lambda_L - \lambda_L \mathrm{Id}) \boldsymbol{n}_{KL}$.
Let u be a continuous function on $\mathbb{R}^d$, affine in both sets K and L and such that $\Lambda_K \nabla u_{|K} . \boldsymbol{n}_{KL} = \Lambda_L \nabla u_{|L} . \boldsymbol{n}_{KL}$. Then we have

$$u(\boldsymbol{y}_\sigma) = \frac{\lambda_L d_{K,\sigma} u(\boldsymbol{x}_L) + \lambda_K d_{L,\sigma} u(\boldsymbol{x}_K)}{\lambda_L d_{K,\sigma} + \lambda_K d_{L,\sigma}}. \tag{13.44}$$

Proof. Let us first notice that y_σ indeed belongs to σ. This is a consequence of $\boldsymbol{y}_K \in \sigma$, $\boldsymbol{y}_L \in \sigma$ and $(\boldsymbol{\lambda}_L^t - \boldsymbol{\lambda}_K^t) \perp \boldsymbol{n}_{KL}$. This ensures that $\boldsymbol{\lambda}_L^t - \boldsymbol{\lambda}_K^t$ is a vector in σ.
Let us now take u as in the lemma, and let $\boldsymbol{G}_K$ and $\boldsymbol{G}_L$ be its gradients in K and L. We decompose these gradients in their normal and tangential part relative to σ: $\boldsymbol{G}_K = g_K \boldsymbol{n}_{KL} + \boldsymbol{G}_K^t$ with $\boldsymbol{G}_K^t \cdot \boldsymbol{n}_{KL} = 0$, $\boldsymbol{G}_L = g_L \boldsymbol{n}_{KL} + \boldsymbol{G}_L^t$ with $\boldsymbol{G}_L^t \cdot \boldsymbol{n}_{KL} = 0$. We set $u_K = u(\boldsymbol{x}_K)$ and $u_L = u(\boldsymbol{x}_L)$. Since $\boldsymbol{y} - \boldsymbol{x}_K = \boldsymbol{y} - \boldsymbol{y}_K + d_{K,\sigma} \boldsymbol{n}_{KL}$ and $\boldsymbol{y} - \boldsymbol{x}_L = \boldsymbol{y} - \boldsymbol{y}_L - d_{L,\sigma} \boldsymbol{n}_{KL}$, the continuity of u along the σ writes

$$\begin{aligned} \forall \boldsymbol{y} \in \sigma \,:\, u(\boldsymbol{y}) &= u_K + d_{K,\sigma} g_K + (\boldsymbol{y} - \boldsymbol{y}_K) \cdot \boldsymbol{G}_K^t \\ &= u_L - d_{L,\sigma} g_L + (\boldsymbol{y} - \boldsymbol{y}_L) \cdot \boldsymbol{G}_L^t. \end{aligned} \tag{13.45}$$

This is equivalent to the two conditions $\boldsymbol{G}_K^t = \boldsymbol{G}_L^t =: \boldsymbol{g}^t$ and

$$d_{K,\sigma} g_K + d_{L,\sigma} g_L = u_L - u_K + (\boldsymbol{y}_K - \boldsymbol{y}_L) \cdot \boldsymbol{g}^t. \tag{13.46}$$

The condition $\Lambda_K \boldsymbol{G}_K \cdot \boldsymbol{n}_{KL} = \Lambda_L \boldsymbol{G}_L \cdot \boldsymbol{n}_{KL}$ can be written

$$g_K \lambda_K - g_L \lambda_L = \boldsymbol{g}^t \cdot (\boldsymbol{\lambda}_L^t - \boldsymbol{\lambda}_K^t). \tag{13.47}$$

From (13.46) and (13.47) we deduce

$$g_K = \frac{\lambda_L [u_L - u_K + (\boldsymbol{y}_K - \boldsymbol{y}_L) \cdot \boldsymbol{g}^t] + d_{L,\sigma} \boldsymbol{g}^t \cdot (\boldsymbol{\lambda}_L^t - \boldsymbol{\lambda}_K^t)}{\lambda_L d_{K,\sigma} + \lambda_K d_{L,\sigma}}.$$

Plugged into (13.45), this formula gives, for any $\boldsymbol{y} \in \sigma$,

$$\begin{aligned} u(\boldsymbol{y}) = u_K &+ d_{K,\sigma} \frac{\lambda_L (u_L - u_K)}{\lambda_L d_{K,\sigma} + \lambda_K d_{L,\sigma}} \\ &+ d_{K,\sigma} \frac{\lambda_L (\boldsymbol{y}_K - \boldsymbol{y}_L) \cdot \boldsymbol{g}^t + d_{L,\sigma} \boldsymbol{g}^t \cdot (\boldsymbol{\lambda}_L^t - \boldsymbol{\lambda}_K^t)}{\lambda_L d_{K,\sigma} + \lambda_K d_{L,\sigma}} + (\boldsymbol{y} - \boldsymbol{y}_K) \cdot \boldsymbol{g}^t. \end{aligned} \tag{13.48}$$

We then just need to define the point $\boldsymbol{y}_\sigma$ as the point $\boldsymbol{y} \in \sigma$ which eliminates the unknown term $\boldsymbol{g}^t$ from this expression, that is

$$d_{K,\sigma} \frac{\lambda_L (\boldsymbol{y}_K - \boldsymbol{y}_L) + d_{L,\sigma} (\boldsymbol{\lambda}_L^t - \boldsymbol{\lambda}_K^t)}{\lambda_L d_{K,\sigma} + \lambda_K d_{L,\sigma}} + (\boldsymbol{y}_\sigma - \boldsymbol{y}_K) = 0,$$

which corresponds to (13.43). Equation (13.48) then reads

$$u(\boldsymbol{y}_\sigma) = u_K + d_{K,\sigma} \frac{\lambda_L (u_L - u_K)}{\lambda_L d_{K,\sigma} + \lambda_K d_{L,\sigma}},$$

which is equivalent to (13.44). ∎

This lemma justifies the following construction of interpolation families.

13.5.2 Construction of interpolation families

We recall that $\mathcal{F}$ is split into $\mathcal{F}_{\text{hyb}}$, corresponding to unknowns that remain in the SUSHI GD, and $\mathcal{F}_{\text{bary}}$, corresponding to unknowns that are eliminated. We first compute, for any face $\tau \in \mathcal{F}$, a point $\boldsymbol{y}_\tau$ on the hyperplane containing τ and a value w_τ by the following method:

1. if $\tau \in \mathcal{F}_{\text{hyb}}$, then $\boldsymbol{y}_\tau = \overline{\boldsymbol{x}}_\tau$ and $w_\tau = u_\tau$;
2. if $\tau \in \mathcal{F}_{\text{bary}}$ is a common face to grid cells M and N, then

$$\boldsymbol{y}_\tau = \frac{\lambda_N d_{M,\tau} \boldsymbol{y}_N + \lambda_M d_{N,\tau} \boldsymbol{y}_M + d_{M,\tau} d_{N,\tau} (\boldsymbol{\lambda}_N^\tau - \boldsymbol{\lambda}_M^\tau)}{\lambda_N d_{M,\tau} + \lambda_M d_{N,\tau}}, \tag{13.49}$$

$$w_\tau = \frac{\lambda_N d_{M,\tau} u_N + \lambda_M d_{N,\tau} u_M}{\lambda_N d_{M,\tau} + \lambda_M d_{N,\tau}}, \tag{13.50}$$

where: $\boldsymbol{y}_M$ and $\boldsymbol{y}_N$ are the orthogonal projections of $\boldsymbol{x}_M$ and $\boldsymbol{x}_N$ on the hyperplane containing τ; $d_{M,\tau} = \mathrm{dist}(\boldsymbol{x}_M, \boldsymbol{y}_M)$ and $d_{N,\tau} = \mathrm{dist}(\boldsymbol{x}_N, \boldsymbol{y}_N)$; and

$$\begin{aligned} \lambda_M &= \boldsymbol{n}_{MN} \cdot \Lambda_M \boldsymbol{n}_{MN}, \quad \boldsymbol{\lambda}_M^\tau = \Lambda_M \boldsymbol{n}_{MN} - \lambda_M \boldsymbol{n}_{MN}, \\ \lambda_N &= \boldsymbol{n}_{MN} \cdot \Lambda_N \boldsymbol{n}_{MN}, \quad \boldsymbol{\lambda}_N^\tau = \Lambda_N \boldsymbol{n}_{MN} - \lambda_N \boldsymbol{n}_{MN} \end{aligned}$$

with $\boldsymbol{n}_{MN}$ the unit normal vector orthogonal to τ and oriented from M to N.

We can now construct the interpolation families for any $\sigma \in \mathcal{F}_{\mathrm{bary}}$. Let K and L be the cells on each side of σ. We select $d-1$ faces $\tau_i \in \mathcal{F}_K \cup \mathcal{F}_L$, different from σ but sharing a common cell with σ, such that there exists a unique function w satisfying:

$$\begin{gathered} w \text{ is affine in } K \text{ and in } L,\ w \text{ is continuous on } \sigma, \\ \Lambda_K (\nabla w)_K \cdot \boldsymbol{n}_{KL} = \Lambda_L (\nabla w)_L \cdot \boldsymbol{n}_{KL} \text{ and} \\ u_K = w(\boldsymbol{x}_K),\ u_L = w(\boldsymbol{x}_L),\ w_{\tau_i} = w(\boldsymbol{y}_{\tau_i}) \text{ for all } i = 1, \ldots, d. \end{gathered} \tag{13.51}$$

By construction of the values w_{τ_i}, the function w is entirely determined by the cell values $(u_M)_{M \in \mathcal{M}}$. We then set $u_\sigma = w(\overline{\boldsymbol{x}}_\sigma)$, which defines u_σ as a linear combination of cell values u_M for $M \in H_\sigma$ (a certain set of cells close to σ), that is

$$u_\sigma = \sum_{M \in H_\sigma} \beta_M^\sigma u_M. \tag{13.52}$$

This defines the family of barycentric coefficients $(\beta_i^\sigma)_{i \in H_\sigma}$. The unknowns corresponding to $\mathcal{F}_{\mathrm{hyb}}$ are therefore not used to eliminate the unknowns on $\mathcal{F}_{\mathrm{bary}}$.

The computation of the linear combination defining u_σ can be simplified by adopting the following algorithm:

1. The continuity of w forces the tangential components of the gradients $\nabla w_{|K}$ and $\nabla w_{|L}$ to be equal, say to $\boldsymbol{g}^t$, on τ. The gradients of w are therefore entirely determined by $\boldsymbol{g}^t$ and their two normal components g_K and g_L to τ in K and L, that is, by $d+1$ scalar unknowns $G = (\boldsymbol{g}^t, g_K, g_L)$.
2. Given that w is affine in K and L, we can write a linear relation between the gradient components G and the increments $X = (u_K - u_\sigma, u_L - u_\sigma, (w_{\tau_i} - u_\sigma)_{i=1,\ldots,d-1})$ of w, that is, $MG = X$ for some matrix M.
3. We then invert M to get $G = M^{-1}X$, which defines all the gradients of w in terms of the increments X.
4. The flux conservativity $\Lambda_K (\nabla w)_K \cdot \boldsymbol{n}_{KL} - \Lambda_L (\nabla w)_L \cdot \boldsymbol{n}_{KL} = 0$ is then imposed and, given the construction of X, gives a linear relation between u_σ and $(u_K, u_L, (w_{\tau_i})_{i=1,\ldots,d-1})$ as expected.

In practice, the selection of the faces $(\tau_i)_{i=1,\ldots,d-1}$ is done by selecting those who produce the most invertible matrix M in the previous algorithm.

GD-consistency of the method. We assume that Λ is piecewise constant on a polytopal mesh $\Omega = \cup_{\ell=1}^k P_\ell$ of the domain Ω, that the polytopal mesh

$\mathfrak{T}$ are adapted to this mesh (*i.e.* each cell of the mesh is fully contained into only one P_ℓ), and that the sets of barycentric faces $\mathcal{F}_{\rm bary}$ are chosen such that

$$\forall \tau \in \mathcal{F}_{\rm bary} \,:\, \boldsymbol{y}_\tau \text{ defined by (13.49) belongs to } \tau \tag{13.53}$$

We consider the set $\mathcal{S}$ of continuous functions φ on $\overline{\Omega}$ that are equal to 0 on $\partial\Omega$, belong to $W^{2,\infty}(P_\ell)$ for each P_ℓ, and that have continuous fluxes through the jumps of Λ (that is, for all ℓ, ℓ' such that $\overline{P_\ell} \cap \overline{P_{\ell'}}$ has a non-zero $(d-1)$-dimensional measure, $\Lambda_{|P_\ell} \nabla \varphi_{|P_\ell} \cdot \boldsymbol{n}_{\ell\ell'} = \Lambda_{|P'_\ell} \nabla \varphi_{|P_{\ell'}} \boldsymbol{n}_{\ell\ell'}$ on $\overline{P_\ell} \cap \overline{P_{\ell'}}$, where $\boldsymbol{n}_{\ell\ell'}$ is a fixed unit normal to $\overline{P_\ell} \cap \overline{P_{\ell'}}$). The following lemma is an enabler of Theorem 7.74 and therefore shows that SUSHI GDs constructed using the coefficients (13.52) are coercive, consistent, limit-conforming and compact. SUSHI GDs also obviously have piecewise constant reconstructions. We refer to Definition 7.73 for the definition of the quantity $\mathrm{reg}_{\mathcal{S}}$ used in the next lemma.

Lemma 13.22. *Let $\mathcal{S} \subset H_0^1(\Omega)$ be constructed as above, let $\mathcal{D}$ be an HMM GD, and let $\mathcal{D}^{\mathcal{S}}$ be a SUSHI GD constructed from $\mathcal{D}$ by using the coefficients* (13.52).
Then $\mathcal{S}$ is dense in $H_0^1(\Omega)$ and, under Assumption (13.53)*, $\mathcal{S}$ satisfies: for all $\varphi \in \mathcal{S}$, there exists $C_\varphi \geq 0$ and $R_{\mathcal{D}^{\mathcal{S}}}$ (depending only on an upper bound of $\zeta_{\mathcal{D}}$ and $\mathrm{reg}_{\mathcal{S}}(\mathcal{D}^{\mathcal{S}})$) such that*

$$\forall \sigma \in \mathcal{F}_{\rm bary} \,:\, \left|\varphi(\overline{\boldsymbol{x}}_\sigma) - \sum_{K \in H_\sigma} \beta_K^\sigma \varphi(\boldsymbol{x}_K)\right| \leq C_\varphi R_{\mathcal{D}^{\mathcal{S}}} \mathrm{diam}(\sigma)^2. \tag{13.54}$$

Proof.
The density of $\mathcal{S}$ is established in [8, Lemma 3.2]. We consider therefore Property (13.54). For any $\tau = M|N \in \mathcal{F}_{\rm bary}$, define a piecewise linear approximation $\overline{\varphi}$ of φ in $M \cup N$ by:

$$\begin{aligned} \forall \boldsymbol{x} \in M \,:\, \overline{\varphi}(\boldsymbol{x}) &= \varphi(\boldsymbol{y}_\tau) + \nabla\varphi_{|M}(\boldsymbol{y}_\tau) \cdot (\boldsymbol{x} - \boldsymbol{y}_\tau), \\ \forall \boldsymbol{x} \in N \,:\, \overline{\varphi}(\boldsymbol{x}) &= \varphi(\boldsymbol{y}_\tau) + \nabla\varphi_{|N}(\boldsymbol{y}_\tau) \cdot (\boldsymbol{x} - \boldsymbol{y}_\tau). \end{aligned}$$

Then $\overline{\varphi}$ is continuous on through τ (because φ is continuous on τ, so the tangential parts, with respect to τ, of $\nabla\varphi_{|M}(\boldsymbol{y}_\tau)$ and of $\nabla\varphi_{|N}(\boldsymbol{y}_\tau)$ coincide), and the continuity of the fluxes of φ ensure that $\overline{\varphi}$ also has a continuous flux through τ. Therefore, by Lemma 13.21,

$$\overline{\varphi}(\boldsymbol{y}_\tau) = \frac{\lambda_N d_{M,\tau} \overline{\varphi}(\boldsymbol{x}_N) + \lambda_M d_{N,\tau} \overline{\varphi}(\boldsymbol{x}_M)}{\lambda_N d_{M,\tau} + \lambda_M d_{N,\tau}}$$

Since $\varphi - \overline{\varphi} = \mathcal{O}((h_M + h_N)^2)$ in $M \cup N$ (because φ is smooth in M and in N), we infer that

$$\varphi(\boldsymbol{y}_\tau) = \frac{\lambda_N d_{M,\tau} \varphi(\boldsymbol{x}_N) + \lambda_M d_{N,\tau} \varphi(\boldsymbol{x}_M)}{\lambda_N d_{M,\tau} + \lambda_M d_{N,\tau}} + \mathcal{O}((h_M + h_N)^2). \tag{13.55}$$

Let us then consider the values w_τ constructed as above from the values $u_\tau = \varphi(\overline{\boldsymbol{x}}_\tau)$ if $\tau \in \mathcal{F}_{\rm hyb}$, and $u_N = \varphi(\boldsymbol{x}_N)$, $u_M = \varphi(\boldsymbol{x}_M)$ if $\tau = M|N \in \mathcal{F}_{\rm bary}$. Using the bound on $\zeta_{\mathcal{D}}$, the preceding reasoning shows that for any $\tau \in \mathcal{F}$, $\varphi(y_\tau) = w_\tau + \mathcal{O}(\mathrm{diam}(\tau)^2)$. Hence, for a given face $\sigma = K|L$, any piecewise linear function w constructed as in (13.51) from the values $u_K = \varphi(\boldsymbol{x}_K)$, $u_L = \varphi(\boldsymbol{x}_K)$ and $(w_{\tau_i})_{i=1,\ldots,d-1}$ satisfies

$$w - \varphi = \mathcal{O}(\mathrm{diam}(\sigma)^2) \tag{13.56}$$

at the points $\boldsymbol{x}_K$, $\boldsymbol{x}_L$ and $(\boldsymbol{y}_{\tau_i})_{i=1,\ldots,d-1}$. This shows in particular that the gradients of w in K and L (entirely computable from the values at the preceding points) are within distance $\mathcal{O}(\mathrm{diam}(\sigma))$ of the gradients of φ in these cells, and therefore that (13.56) actually holds uniformly in $\overline{K \cup L}$. Applied at $\overline{\boldsymbol{x}}_\sigma \in \sigma$, this estimate gives $w(\overline{\boldsymbol{x}}_\sigma) = \varphi(\overline{\boldsymbol{x}}_\sigma) + \mathcal{O}(\mathrm{diam}(\sigma)^2)$, which is precisely (13.54). ∎

14

Nodal mimetic finite difference methods

Nodal mimetic finite differences (nMFD) methods form the second family of MFD methods, after hMFD, that we study in this book. The analysis of nMFD is relatively similar to that of hMFD, but several changes have to be made since the discrete unknowns of nMFD are located at the vertices of the mesh, rather than the cells and edges as in hMFD.
We only consider here homogeneous Dirichlet boundary conditions, but we briefly address the questions of other boundary conditions in Remark 14.1.

14.1 Definition and properties of nMFD gradient discretisations

We first define the GD, and then prove that the corresponding GS (2.23) is indeed the nMFD scheme as defined in [34].
Let $\mathfrak{T} = (\mathcal{M}, \mathcal{F}, \mathcal{P}, \mathcal{V})$ be a polytopal mesh of Ω in the sense of Definition 7.2. For each $K \in \mathcal{M}$ we choose non-negative weights $(\omega_K^{\boldsymbol{s}})_{\boldsymbol{s}\in\mathcal{V}_K}$ such that the quadrature

$$\int_K w(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \approx \sum_{\boldsymbol{s}\in\mathcal{V}_K} \omega_K^{\boldsymbol{s}} w(\boldsymbol{s}) \tag{14.1}$$

is exact for constant functions w, which means that

$$\sum_{\boldsymbol{s}\in\mathcal{V}_K} \omega_K^{\boldsymbol{s}} = |K|. \tag{14.2}$$

For each face $\sigma \in \mathcal{F}_K \cap \mathcal{F}_{\mathrm{int}}$, we also choose non-negative weights $(\omega_\sigma^{\boldsymbol{s}})_{\boldsymbol{s}\in\mathcal{V}_\sigma}$ such that the quadrature

$$\int_\sigma w(\boldsymbol{x})\mathrm{d}\gamma(\boldsymbol{x}) \approx \sum_{\boldsymbol{s}\in\mathcal{V}_\sigma} \omega_\sigma^{\boldsymbol{s}} w(\boldsymbol{s}) \tag{14.3}$$

is exact for affine functions w. This is equivalent to

$$\sum_{\boldsymbol{s}\in\mathcal{V}_\sigma} \omega_\sigma^{\boldsymbol{s}} = |\sigma| \quad \text{and} \quad \sum_{\boldsymbol{s}\in\mathcal{V}_\sigma} \omega_\sigma^{\boldsymbol{s}} \boldsymbol{s} = |\sigma| \overline{\boldsymbol{x}}_\sigma. \tag{14.4}$$

We also assume the following property on these weights.

$$\forall K \in \mathcal{M}\,,\ \forall \boldsymbol{s} \in \mathcal{V}_K\,,\ \exists \sigma \in \mathcal{F}_{K,\boldsymbol{s}} \text{ such that } \omega_\sigma^{\boldsymbol{s}} \neq 0, \tag{14.5}$$

where $\mathcal{F}_{K,\boldsymbol{s}} = \{\sigma \in \mathcal{F}_K \,:\, \boldsymbol{s} \in \mathcal{V}_\sigma\}$ is the set of faces of K that have $\boldsymbol{s}$ as one of their vertices. This assumption, not very restrictive in practice, states that each vertex of each cell K is genuinely involved in at least one of the quadrature rules (14.3) on the faces of K. (14.5) is not required in the construction of the nMFD, but it is used to identify the nMFD method with a GDM.
For each cell $K \in \mathcal{M}$, we re-define its centre $\boldsymbol{x}_K \in \mathcal{P}$ by setting

$$\boldsymbol{x}_K = \frac{1}{|K|} \sum_{\boldsymbol{s}\in\mathcal{V}_K} \omega_K^{\boldsymbol{s}} \boldsymbol{s} \tag{14.6}$$

(which is assumed to belong to K) and we select a partition $(V_{K,\boldsymbol{s}})_{\boldsymbol{s}\in\mathcal{V}_K}$ of K such that

$$\forall \boldsymbol{s} \in \mathcal{V}_K\,,\ |V_{K,\boldsymbol{s}}| = \sum_{\sigma\in\mathcal{F}_{K,\boldsymbol{s}}} \omega_\sigma^{\boldsymbol{s}} \frac{|D_{K,\sigma}|}{|\sigma|} = \frac{1}{d} \sum_{\sigma\in\mathcal{F}_{K,\boldsymbol{s}}} \omega_\sigma^{\boldsymbol{s}} d_{K,\sigma}. \tag{14.7}$$

The second equality follows from (B.1), and we note that (14.4) yields $\sum_{\boldsymbol{s}\in\mathcal{V}_K} |V_{K,\boldsymbol{s}}| = |K|$, which is compatible with the requirement that $(V_{K,\boldsymbol{s}})_{\boldsymbol{s}\in\mathcal{V}_K}$ is a partition of K.

The nMFD LLE gradient discretisation is constructed by following the notations in Definition 7.33.

1. The set of geometrical entities attached to the discrete unknowns is $I = \mathcal{V}$, and the set of approximation points is $S = I$. We set $I_\Omega = \mathcal{V} \cap \Omega$ and $I_\partial = \mathcal{V} \cap \partial\Omega$. Hence,

$$X_{\mathcal{D},0} = \{v = (v_{\boldsymbol{s}})_{\boldsymbol{s}\in\mathcal{V}} \,:\, v_{\boldsymbol{s}} \in \mathbb{R} \text{ for all } \boldsymbol{s} \in \mathcal{V} \cap \Omega\,, \\ v_{\boldsymbol{s}} = 0 \text{ for all } \boldsymbol{s} \in \mathcal{V} \cap \partial\Omega\}.$$

   For $K \in \mathcal{M}$, we let $I_K = \mathcal{V}_K$.
2. For all $K \in \mathcal{M}$, the functions $\pi_K = (\pi_K^{\boldsymbol{s}})_{\boldsymbol{s}\in\mathcal{V}_K}$ are defined by

$$\forall \boldsymbol{s} \in \mathcal{V}_K\,,\ \text{for a.e. } \boldsymbol{x} \in K\,,\ \pi_K^{\boldsymbol{s}}(\boldsymbol{x}) := \frac{\omega_K^{\boldsymbol{s}}}{|K|}. \tag{14.8}$$

   Relation (7.32) gives

$$\forall v \in X_{\mathcal{D},0},\ \forall K \in \mathcal{M},\ \forall \boldsymbol{x} \in K,\ \Pi_{\mathcal{D}} v(\boldsymbol{x}) = v_K := \frac{1}{|K|} \sum_{\boldsymbol{s}\in\mathcal{V}_K} \omega_K^{\boldsymbol{s}} v_{\boldsymbol{s}}. \tag{14.9}$$

3. In a similar way as for the HMM method, the reconstructed gradient is the sum of a constant gradient in each cell, and of stabilisation terms in each $V_{K,\boldsymbol{s}}$. It is also best defined by first giving an expression of $\mathcal{G}_K v$. Let $X_{\mathcal{V}_K} = \{v = (v_{\boldsymbol{s}})_{\boldsymbol{s}\in\mathcal{V}_K} \,:\, v_{\boldsymbol{s}} \in \mathbb{R}\}$ be the space of discrete unknowns in K, and

$$\forall K \in \mathcal{M}\,,\ \forall v \in X_{\mathcal{V}_K}\,,\ \nabla_K v = \frac{1}{|K|}\sum_{\sigma\in\mathcal{F}_K}\left(\sum_{\boldsymbol{s}\in\mathcal{V}_\sigma}\omega_\sigma^{\boldsymbol{s}} v_{\boldsymbol{s}}\right)\boldsymbol{n}_{K,\sigma}. \tag{14.10}$$

Define then, for $v \in X_{\mathcal{V}_K}$, the function $\mathcal{G}_K v \in L^p(K)^d$ by

$$\begin{aligned}&\forall \boldsymbol{s} \in \mathcal{V}_K\,,\ \text{for a.e. } \boldsymbol{x} \in V_{K,\boldsymbol{s}}\,,\\ &\mathcal{G}_K v(\boldsymbol{x}) = \nabla_K v + \frac{1}{h_K}[\mathcal{L}_K R_K(v)]_{\boldsymbol{s}}\boldsymbol{N}_{K,\boldsymbol{s}}\end{aligned} \tag{14.11}$$

where
- $\boldsymbol{N}_{K,\boldsymbol{s}} = \frac{h_K}{d|V_{K,\boldsymbol{s}}|}\sum_{\sigma\in\mathcal{F}_{K,\boldsymbol{s}}}\omega_\sigma^{\boldsymbol{s}}\boldsymbol{n}_{K,\sigma}$,
- $R_K : X_{\mathcal{V}_K} \mapsto X_{\mathcal{V}_K}$ is the linear mapping described by $R_K(v) = (R_{K,\boldsymbol{s}}(v))_{\boldsymbol{s}\in\mathcal{V}_K}$ with

$$R_{K,\boldsymbol{s}}(v) = v_{\boldsymbol{s}} - v_K - \nabla_K v\cdot(\boldsymbol{s} - \boldsymbol{x}_K), \tag{14.12}$$

  where v_K is defined in (14.9) and $\boldsymbol{x}_K$ is given by (14.6),
- $\mathcal{L}_K$ is an isomorphism of the space $\mathrm{Im}(R_K)$.

By (7.33), we then have

$$\begin{aligned}&\forall v \in X_{\mathcal{D}}\,,\ \forall K \in \mathcal{M}\,,\ \forall \boldsymbol{s} \in \mathcal{V}_K\,,\ \text{for a.e. } \boldsymbol{x} \in V_{K,\sigma}\,,\\ &\nabla_{\mathcal{D}} v(\boldsymbol{x}) = \nabla_K v + \frac{1}{h_K}[\mathcal{L}_K R_K(v)]_{\boldsymbol{s}}\boldsymbol{N}_{K,\boldsymbol{s}}.\end{aligned} \tag{14.13}$$

The functions $(\mathcal{G}_K^{\boldsymbol{s}})_{\boldsymbol{s}\in\mathcal{V}_K}$ of $L^p(K)^d$ are recovered from the definition (14.11) of $\mathcal{G}_K v$ by considering, for each $\boldsymbol{s} \in \mathcal{V}_K$, the vector $v^{\boldsymbol{s}} \in X_{\mathcal{V}_K}$ with value 1 at $\boldsymbol{s}$ and 0 at all other vertices of K, and by setting

$$\mathcal{G}_K^{\boldsymbol{s}} = \mathcal{G}_K v^{\boldsymbol{s}} \text{ for all } \boldsymbol{s} \in \mathcal{V}_K. \tag{14.14}$$

4. The proof that π_K and $\mathcal{G}_K$ are $\mathbb{P}_1$-exact reconstructions and that $\|\nabla_{\mathcal{D}}\cdot\|_{L^p(\Omega)^d}$ is a norm on $X_{\mathcal{D},0}$, is provided in Lemma 14.8 below.

Remark 14.1 (Other boundary conditions)
The adaptation of nMFD to non-homogeneous Dirichlet conditions raises the same interpolation issues as for $\mathbb{P}_1$ finite element methods (see Section 8.3.1), and essentially requires a boundary condition smoother than $W^{1-1/p,p}$. Other boundary conditions (Neumann, Fourier) are rather straightforward to deal with, using the value $u_\sigma = \frac{1}{|\sigma|}\sum_{\boldsymbol{s}\in\mathcal{V}_\sigma}\omega_\sigma^{\boldsymbol{s}} u_{\boldsymbol{s}}$ to define the trace reconstruction $\mathbb{T}_{\mathcal{D}}$.

We prove that an nMFD scheme is the GS (2.23) corresponding the GD defined above, for suitable choice of $(\mathcal{L}_K)_{K\in\mathcal{M}}$. Let us first recall the definition of an nMFD scheme from [34]. The space of discrete unknowns at the interior vertices of the mesh, denoted by $\mathcal{N}_0$ in [34], is simply $X_{\mathcal{D},0}$ defined above. The nMFD for (2.20) (with $\boldsymbol{F}=0$) is then written under the general form

$$\text{Find } u \in X_{\mathcal{D},0} \text{ such that, for all } v \in X_{\mathcal{D},0},\ [u,v]_{X_{\mathcal{D},0}} = \widetilde{f}(v), \tag{14.15}$$

where $[\cdot,\cdot]_{X_{\mathcal{D},0}}$ is an inner product on $X_{\mathcal{D},0}$ and $\widetilde{f}$ a linear form on $X_{\mathcal{D},0}$. Using the quadrature rule (14.1), the linear form $\widetilde{f}$ is defined as

$$\widetilde{f}(v) = \sum_{K\in\mathcal{M}} \left(\frac{1}{|K|}\int_K f\right) \sum_{\boldsymbol{s}\in\mathcal{V}_K} \omega_K^{\boldsymbol{s}} v_{\boldsymbol{s}}. \tag{14.16}$$

The inner product $[\cdot,\cdot]_{X_{\mathcal{D},0}} = \sum_{K\in\mathcal{M}}[\cdot,\cdot]_{\mathcal{V}_K}$ is designed cell-by-cell to ensure that a discrete Stokes formula is satisfied for interpolants of linear functions. It is shown in [34] that this leads to the following generic form:

$$\begin{aligned} &\forall u \in X_{\mathcal{V}_K},\ \forall v \in X_{\mathcal{V}_K}\ : \\ &\qquad [u,v]_{\mathcal{V}_K} = u^T \mathbb{M}_K v \ \text{ with } \ \mathbb{M}_{\mathbb{K}} = \frac{1}{|K|}\mathbb{C}_K \Lambda_K^{-1}\mathbb{C}_K^T + \mathbb{D}_K\mathbb{K}_K\mathbb{D}_K^T, \end{aligned} \tag{14.17}$$

where

- Λ_K is the constant value of Λ on K (as in HMM methods, we assume that Λ is piecewise constant on $\mathcal{M}$),
- $\mathbb{C}_K$ is the $\mathrm{Card}(\mathcal{V}_K)\times d$ matrix with rows $(\sum_{\sigma\in\mathcal{F}_{K,\boldsymbol{s}}} \omega_\sigma^{\boldsymbol{s}}(\Lambda_K \boldsymbol{n}_{K,\sigma})^T)_{\boldsymbol{s}\in\mathcal{V}_K}$, where, as before, $\mathcal{F}_{K,\boldsymbol{s}}$ is the set of faces of K that have $\boldsymbol{s}$ as one of their vertices.
- $\mathbb{D}_K$ is a $\mathrm{Card}(\mathcal{V}_K)\times(\mathrm{Card}(\mathcal{V}_K)-d)$ matrix whose columns span the orthogonal space in $X_{\mathcal{V}_K}$ of E_K, where

$$E_K = \{(A(\boldsymbol{s}))_{\boldsymbol{s}\in\mathcal{V}_K}\ :\ A:\mathbb{R}^d\to\mathbb{R} \text{ affine mapping}\}$$

  is the vector space of the values of affine mappings at the vertices of K.
- $\mathbb{K}_K$ is a symmetric positive definite matrix of size $\mathrm{Card}(\mathcal{V}_K)-d$.

Remark 14.2 *($\mathbb{R}^{\mathcal{V}_K}$ vs. $\mathbb{R}^{\mathrm{Card}(\mathcal{V}_K)}$)*
As in Remark 13.5, we make abuses of notation when we consider $\mathbb{C}_K$, $\mathbb{D}_K$ and $\mathbb{K}_K$ as matrices. Formally, this supposes that a numbering of the vertices $\mathcal{V}_K$ of K has been chosen.

Before proving that nMFD methods are GDMs, two technical results are required. The first one contains in particular results similar to those in Lemma 13.6, and the second one describes the kernel of $\mathbb{D}_K$.

Lemma 14.3. *Let $\mathfrak{T} = (\mathcal{M}, \mathcal{F}, \mathcal{P}, \mathcal{V})$ be a polytopal mesh in the sense of Definition 7.2, and let $\mathcal{D}$ be an nMFD GD as defined above, for some choices of $(\mathcal{L}_K)_{K\in\mathcal{M}}$. Then,*

1. *For all $K \in \mathcal{M}$, $\beta \in \mathrm{Im}(R_K)$ if and only if*

$$\sum_{\boldsymbol{s}\in\mathcal{V}_K} \frac{|V_{K,\boldsymbol{s}}|}{h_K}\beta_{\boldsymbol{s}}\boldsymbol{N}_{K,\boldsymbol{s}} = 0. \tag{14.18}$$

2. *For all $v \in X_{\mathcal{D},0}$ and all $K \in \mathcal{M}$,*

$$\nabla_K v = \frac{1}{|K|}\int_K \nabla_{\mathcal{D}} v(\boldsymbol{x})\mathrm{d}\boldsymbol{x}. \tag{14.19}$$

3. *∇_K is a $\mathbb{P}_1$-exact gradient reconstruction on K upon $\mathcal{V}_K$, in the sense of Definition 7.28.*
4. *For all $K \in \mathcal{M}$ and all $\boldsymbol{s} \in \mathcal{V}_K$, $|\boldsymbol{N}_{K,\boldsymbol{s}}| \geq 1$.*

Proof.
Item 1. If $\beta \in \mathrm{Im}(R_K)$ then, for some $v \in X_{\mathcal{V}_K}$,

$$\beta_{\boldsymbol{s}} = v_{\boldsymbol{s}} - v_K - \nabla_K v \cdot (\boldsymbol{s} - \boldsymbol{x}_K) \text{ for all } \boldsymbol{s} \in \mathcal{V}_K.$$

Set $w_{\boldsymbol{s}} = v_{\boldsymbol{s}} - v_K$. Using (14.4) and $\sum_{\sigma\in\mathcal{F}_K} |\sigma|\boldsymbol{n}_{K,\sigma} = 0$ (see (B.4)) shows that $\nabla_K w = \nabla_K v$. Hence, $\beta_{\boldsymbol{s}} = w_{\boldsymbol{s}} - \nabla_K w \cdot (\boldsymbol{s} - \boldsymbol{x}_K)$ Given the definition of $\boldsymbol{N}_{K,\boldsymbol{s}}$, this yields

$$\begin{aligned}
\sum_{\boldsymbol{s}\in\mathcal{V}_K} \frac{|V_{K,\boldsymbol{s}}|}{h_K}\beta_{\boldsymbol{s}}\boldsymbol{N}_{K,\boldsymbol{s}} &= \frac{1}{d}\sum_{\boldsymbol{s}\in\mathcal{V}_K}\sum_{\sigma\in\mathcal{F}_{K,\boldsymbol{s}}} \beta_{\boldsymbol{s}}\omega_\sigma^{\boldsymbol{s}}\boldsymbol{n}_{K,\sigma} \\
&= \frac{1}{d}\sum_{\sigma\in\mathcal{F}_K}\left(\sum_{\boldsymbol{s}\in\mathcal{V}_\sigma}\omega_\sigma^{\boldsymbol{s}}\beta_{\boldsymbol{s}}\right)\boldsymbol{n}_{K,\sigma} \\
&= \frac{1}{d}\sum_{\sigma\in\mathcal{F}_K}\left(\sum_{\boldsymbol{s}\in\mathcal{V}_\sigma}\omega_\sigma^{\boldsymbol{s}} w_{\boldsymbol{s}}\right)\boldsymbol{n}_{K,\sigma} - \frac{1}{d}\sum_{\sigma\in\mathcal{F}_K}\left(\sum_{\boldsymbol{s}\in\mathcal{V}_\sigma}\omega_\sigma^{\boldsymbol{s}}\nabla_K w\cdot(\boldsymbol{s}-\boldsymbol{x}_K)\right)\boldsymbol{n}_{K,\sigma} \\
&= \frac{1}{d}\left(|K|\nabla_K w - T_1\right).
\end{aligned} \tag{14.20}$$

We then use (14.4) and Lemma B.3 to write

$$\begin{aligned}
T_1 &= \sum_{\sigma\in\mathcal{F}_K}\left[\nabla_K w\cdot\left(\sum_{\boldsymbol{s}\in\mathcal{V}_\sigma}\omega_\sigma^{\boldsymbol{s}}(\boldsymbol{s}-\boldsymbol{x}_K)\right)\right]\boldsymbol{n}_{K,\sigma} \\
&= \sum_{\sigma\in\mathcal{F}_K}|\sigma|\left[\nabla_K w\cdot(\overline{\boldsymbol{x}}_\sigma - \boldsymbol{x}_K)\right]\boldsymbol{n}_{K,\sigma} = |K|\nabla_K w.
\end{aligned} \tag{14.21}$$

Substituted in (14.20) this shows that β satisfies (14.18). Defining

$$G_K : \beta \in X_{\mathcal{V}_K} \to \sum_{\boldsymbol{s}\in\mathcal{V}_K} \frac{|V_{K,\boldsymbol{s}}|}{h_K}\beta_{\boldsymbol{s}}\boldsymbol{N}_{K,\boldsymbol{s}} \in \mathbb{R}^d,$$

we just showed that $\mathrm{Im}(R_K) \subset \ker(G_K)$. The vectors $(\boldsymbol{N}_{K,\boldsymbol{s}})_{\boldsymbol{s}\in\mathcal{V}_K}$ span $\mathbb{R}^d$. Indeed, for any vector $\boldsymbol{\xi} \in \mathbb{R}^d$, using (14.4) and Lemma B.3 (with $\boldsymbol{x}_K = 0$),

$$\begin{aligned}\sum_{\boldsymbol{s}\in\mathcal{V}_K} \frac{|V_{K,\boldsymbol{s}}|d}{h_K}(\boldsymbol{s}\cdot\boldsymbol{\xi})\boldsymbol{N}_{K,\boldsymbol{s}} &= \sum_{\sigma\in\mathcal{F}_K}\left(\sum_{\boldsymbol{s}\in\mathcal{V}_\sigma}\omega_\sigma^{\boldsymbol{s}}\boldsymbol{s}\cdot\boldsymbol{\xi}\right)\boldsymbol{n}_{K,\sigma}\\ &= \sum_{\sigma\in\mathcal{F}_K}|\sigma|(\overline{\boldsymbol{x}}_\sigma\cdot\boldsymbol{\xi})\boldsymbol{n}_{K,\sigma} = |K|\boldsymbol{\xi}.\end{aligned}$$

By Assumption (14.5), none of the $(V_{K,\boldsymbol{s}})_{\boldsymbol{s}\in\mathcal{V}_K}$ has a zero measure. Hence, $\mathrm{Im}(G_K) = \mathbb{R}^d$ and $\dim(\ker G_K) = \mathrm{Card}(\mathcal{V}_K) - d$. Using similar computations as in (14.21), it can be seen that $Z \in \mathbb{R}^d \mapsto (Z\cdot(\boldsymbol{s}-\boldsymbol{x}_K))_{\boldsymbol{s}\in\mathcal{V}_K} \in \ker(R_K)$ is an isomorphism (the one-to-one property comes from the fact that $(\boldsymbol{s}-\boldsymbol{x}_K)_{\sigma\in\mathcal{F}_K}$ spans $\mathbb{R}^d$). Hence, $\dim(\mathrm{Im}(R_K)) = \mathrm{Card}(\mathcal{V}_K) - d = \dim(\ker(G_K))$. Since $\mathrm{Im}(R_K) \subset \ker(G_K)$, the equality of dimensions therefore gives $\mathrm{Im}(R_K) = \ker(G_K)$ and completes the proof of Item 1.

Item 2. By Definition (14.13) of $\nabla_{\mathcal{D}}$,

$$\int_K \nabla_{\mathcal{D}} v(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = |K|\nabla_K v + \sum_{\boldsymbol{s}\in\mathcal{V}_K}\frac{|V_{K,\boldsymbol{s}}|}{h_K}[\mathcal{L}_K R_K(v)]_{\boldsymbol{s}}\boldsymbol{N}_{K,\boldsymbol{s}}.$$

Since $\mathcal{L}_K R_K(v) \in \mathrm{Im}(R_K)$, Item 1 shows that the last term in this relation vanishes, which concludes the proof of (14.19).

Item 3. If $v = (A(\boldsymbol{s}))_{\boldsymbol{s}\in\mathcal{V}_K}$ for some affine map A, then (14.4) shows that

$$\sum_{\boldsymbol{s}\in\mathcal{V}_\sigma}\omega_\sigma^{\boldsymbol{s}} v_{\boldsymbol{s}} = \sum_{\boldsymbol{s}\in\mathcal{V}_\sigma}\omega_\sigma^{\boldsymbol{s}} A(\boldsymbol{s}) = |\sigma|A(\overline{\boldsymbol{x}}_\sigma).$$

Hence, setting $u = (A(\boldsymbol{x}_K), A(\overline{\boldsymbol{x}}_\sigma)_{\sigma\in\mathcal{F}_K})$, recalling the definition (7.7e) of $\overline{\nabla}_K$ and using Lemma B.10,

$$\nabla_K v = \frac{1}{|K|}\sum_{\sigma\in\mathcal{F}_K}|\sigma|A(\overline{\boldsymbol{x}}_\sigma)\boldsymbol{n}_{K,\sigma} = \overline{\nabla}_K u = \nabla A.$$

Item 4. By definition (7.4) of $d_{K,\sigma}$, for $\sigma \in \mathcal{F}_K$ and $\boldsymbol{s} \in \mathcal{V}_\sigma$ we have $(\boldsymbol{s}-\boldsymbol{x}_K)\cdot\boldsymbol{n}_{K,\sigma} = d_{K,\sigma}$. Hence, by definition (14.7) of $|V_{K,\boldsymbol{s}}|$,

$$\begin{aligned}(\boldsymbol{s}-\boldsymbol{x}_K)\cdot\boldsymbol{N}_{K,\boldsymbol{s}} &= \frac{h_K}{d|V_{K,\boldsymbol{s}}|}\sum_{\sigma\in\mathcal{F}_{K,\boldsymbol{s}}}\omega_\sigma^{\boldsymbol{s}}(\boldsymbol{s}-\boldsymbol{x}_K)\cdot\boldsymbol{n}_{K,\sigma}\\ &= \frac{h_K}{d|V_{K,\boldsymbol{s}}|}\sum_{\sigma\in\mathcal{F}_{K,\boldsymbol{s}}}\omega_\sigma^{\boldsymbol{s}} d_{K,\sigma} = h_K.\end{aligned}$$

Since $(\boldsymbol{s}-\boldsymbol{x}_K)\cdot\boldsymbol{N}_{K,\boldsymbol{s}} \le |\boldsymbol{s}-\boldsymbol{x}_K|\,|\boldsymbol{N}_{K,\boldsymbol{s}}| \le h_K|\boldsymbol{N}_{K,\boldsymbol{s}}|$, it follows that $|\boldsymbol{N}_{K,\boldsymbol{s}}| \ge 1$.

■

Lemma 14.4. *Let $\mathfrak{T}$ be a polytopal mesh in the sense of Definition 7.2, let $K \in \mathcal{M}$, and let $\mathbb{D}_K$ and R_K be defined as above. Then, the mappings $\mathbb{D}_K^T : X_{\mathcal{V}_K} \mapsto \mathbb{R}^{\mathrm{Card}(\mathcal{V}_K)-d}$ and $R_K : X_{\mathcal{V}_K} \mapsto X_{\mathcal{V}_K}$ have the same kernel.*

Proof. The kernel of $\mathbb{D}_K^T$ is the orthogonal (for the dot product in $X_{\mathcal{V}_K}$) of the columns of $\mathbb{D}_K$, that is to say, according to the definition of $\mathbb{D}_K$, the space E_K of values at the vertices of K of affine mappings.
We have $v \in \ker(R_K)$ if and only if

$$\forall \boldsymbol{s} \in \mathcal{V}_K \,:\, v_{\boldsymbol{s}} = v_K + \nabla_K v \cdot (\boldsymbol{s} - \boldsymbol{x}_K). \tag{14.22}$$

If there exists A affine such that $v_{\boldsymbol{s}} = A(\boldsymbol{s})$ for all $\boldsymbol{s} \in \mathcal{V}_K$ then $\nabla_K v = \nabla A$ by Item 3 in Lemma 14.3. The definitions (14.9) and (14.6) of v_K and $\boldsymbol{x}_K$ show that $v_K = A(\boldsymbol{x}_K)$. Hence, since A is affine,

$$v_{\boldsymbol{s}} = A(\boldsymbol{s}) = A(\boldsymbol{x}_K) + \nabla A \cdot (\boldsymbol{s} - \boldsymbol{x}_K) = v_K + \nabla_K v \cdot (\boldsymbol{s} - \boldsymbol{x}_K)$$

and (14.22) holds. Conversely, if (14.22) holds then, defining the affine mapping $A(\boldsymbol{x}) = v_K + \nabla_K v \cdot (\boldsymbol{x} - \boldsymbol{x}_K)$, we have $v_{\boldsymbol{s}} = A(\boldsymbol{s})$ for all $\boldsymbol{s} \in \mathcal{V}_K$. We just established that the kernel of R_K is made of the values at the vertices of K of affine mappings. This kernel is therefore identical to $E_K = \ker(\mathbb{D}_K^T)$ and the proof is complete. ∎

We can now prove that the GD constructed above corresponds to the nMFD scheme.

Theorem 14.5 (nMFD methods are GDMs). *Let $\mathfrak{T}$ be a polytopal mesh of Ω in the sense of Definition 7.2. Assume that Λ is piecewise constant on $\mathcal{M}$. Take weights that satisfy (14.2), (14.4) and (14.5), and let (14.15) be an nMFD method constructed from these weights. Then, there exists isomorphisms $(\mathcal{L}_K)_{K\in\mathcal{M}}$ such that, if $\mathcal{D}$ is the GD defined as at the start of this section, the corresponding GS (2.23) is identical to (14.15).*

Remark 14.6 (Non piecewise constant diffusion tensor)
As for the HMM method (see Remark 13.17), if Λ is not piecewise constant on $\mathcal{M}$, then (14.15) is the GS (2.23) in which Λ is replaced with a piecewise constant approximation. We already noticed that this modification does not impact in practice the rates of convergence provided by the theorems in Section 2.1.2.

Proof. Given the definitions (14.9) of $\Pi_{\mathcal{D}}$ and (14.16) of $\widetilde{f}$, the right-hand sides of (2.23) and (14.15) clearly coincide. We therefore just have to prove that the left-hand sides coincide. Since the inner product $[\cdot,\cdot]_{X_{\mathcal{D},0}}$ and the gradient (14.11) are constructed cell-wise, it suffices to show that, for any $u, v \in X_{\mathcal{D},0}$ and any cell K, we can find $\mathcal{L}_K$ such that

$$\int_K \Lambda(\boldsymbol{x}) \nabla_{\mathcal{D}} u(\boldsymbol{x}) \cdot \nabla_{\mathcal{D}} v(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} = u^T \mathbb{M}_K v. \tag{14.23}$$

Let $S_K(u) = \nabla_{\mathcal{D}} u - \nabla_K u$ be the stabilisation part of $\nabla_{\mathcal{D}} u$ on K. By (14.19), we have $\int_K S_K(u)(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = \int_K S_K(v)(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = 0$ and thus, since $\Lambda = \Lambda_K$ is constant on K,

$$\begin{aligned}
&\int_K \Lambda(\boldsymbol{x})\nabla_{\mathcal{D}} u(\boldsymbol{x}) \cdot \nabla_{\mathcal{D}} v(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \\
&\qquad = |K|\Lambda_K \nabla_K u \cdot \nabla_K v + \int_K \Lambda_K \nabla_K u \cdot S_K(v)(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \\
&\qquad\quad + \int_K \Lambda_K S_K(u)(\boldsymbol{x}) \cdot \nabla_K v \mathrm{d}\boldsymbol{x} + \int_K \Lambda_K S_K(u)(\boldsymbol{x}) \cdot S_K(v)(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \\
&\qquad = |K|\Lambda_K \nabla_K u \cdot \nabla_K v + \int_K \Lambda_K S_K(u)(\boldsymbol{x}) \cdot S_K(v)(\boldsymbol{x})\mathrm{d}\boldsymbol{x}. \qquad (14.24)
\end{aligned}$$

By definition of $\mathbb{C}_K$, for all $\xi \in X_{\mathcal{V}_K}$,

$$\begin{aligned}
\mathbb{C}_K^T \xi &= \sum_{\boldsymbol{s} \in \mathcal{V}_K} \left(\sum_{\sigma \in \mathcal{F}_{K,\boldsymbol{s}}} \omega_\sigma^{\boldsymbol{s}} \Lambda_K \boldsymbol{n}_{K,\sigma} \right) \xi_{\boldsymbol{s}} \\
&= \Lambda_K \sum_{\sigma \in \mathcal{F}_K} \left(\sum_{\boldsymbol{s} \in \mathcal{V}_\sigma} \omega_\sigma^{\boldsymbol{s}} \xi_{\boldsymbol{s}} \right) \boldsymbol{n}_{K,\sigma} = |K|\Lambda_K \nabla_K \xi.
\end{aligned}$$

Hence,

$$\frac{1}{|K|} u^T \mathbb{C}_K \Lambda_K^{-1} \mathbb{C}_K^T v = |K|(\Lambda_K \nabla_K u)^T \Lambda_K^{-1} (\Lambda_K \nabla_K v) = |K|\Lambda_K \nabla_K u \cdot \nabla_K v.$$

The first term in the right-hand side of (14.24) therefore corresponds to the first term in the expression (14.17) of $u^T \mathbb{M}_K v$. To complete the proof of the theorem, we therefore only have to show that, for any symmetric positive definite $(n_K - d) \times (n_K - d)$ matrix $\mathbb{K}_K$, there exists an isomorphism $\mathcal{L}_K$ of $\mathrm{Im}(R_K)$ such that, for all $u, v \in X_{\mathcal{V}_K}$,

$$u^T \mathbb{D}_K \mathbb{K}_K \mathbb{D}_K^T v = \int_K \Lambda_K S_K(u)(\boldsymbol{x}) \cdot S_K(v)(\boldsymbol{x})\mathrm{d}\boldsymbol{x}. \qquad (14.25)$$

By Lemma 14.4 we have $\ker(\mathbb{D}_K^T) = \ker(R_K)$. Let $\{\cdot,\cdot\}_1$ be the inner product on $\mathbb{R}^{\mathrm{Card}(\mathcal{V}_K)-d}$ defined by $\mathbb{K}_K$, and apply Lemma 14.7 to produce an inner product $\{\cdot,\cdot\}_2$ on $X_{\mathcal{V}_K}$ such that $\{\mathbb{D}_K^T u, \mathbb{D}_K^T v\}_1 = \{R_K(u), R_K(v)\}_2$. Then (14.25) follows if we can establish the existence of an isomorphism $\mathcal{L}_K$ of $\mathrm{Im}(R_K)$ such that, for all $u, v \in X_{\mathcal{D},0}$,

$$\{R_K(v), R_K(v)\}_2 = \int_K \Lambda_K S_K(u)(\boldsymbol{x}) \cdot S_K(v)(\boldsymbol{x})\mathrm{d}\boldsymbol{x}. \qquad (14.26)$$

By definition of $S_K(u)$ (see (14.13)), we have

$$\int_K \Lambda_K S_K(u)(\boldsymbol{x}) \cdot S_K(v)(\boldsymbol{x})\mathrm{d}\boldsymbol{x}$$

$$= \sum_{\boldsymbol{s}\in\mathcal{V}_K} \frac{|V_{K,\boldsymbol{s}}|}{h_K^2} [\mathcal{L}_K R_K(u)]_{\boldsymbol{s}} [\mathcal{L}_K R_K(v)]_{\boldsymbol{s}} \Lambda_K \boldsymbol{N}_{K,\boldsymbol{s}} \cdot \boldsymbol{N}_{K,\boldsymbol{s}}$$
$$= \langle \mathcal{L}_K R_K(u), \mathcal{L}_K R_K(v) \rangle \tag{14.27}$$

where $\langle \cdot, \cdot \rangle$ is the scalar product on $\mathrm{Im}(R_K)$ defined by

$$\langle \xi, \beta \rangle = \sum_{\boldsymbol{s}\in\mathcal{V}_K} \frac{|V_{K,\boldsymbol{s}}|}{h_K^2} \Lambda_K \boldsymbol{N}_{K,\boldsymbol{s}} \cdot \boldsymbol{N}_{K,\boldsymbol{s}} \xi_{\boldsymbol{s}} \beta_{\boldsymbol{s}}$$

(notice that $\Lambda_K \boldsymbol{N}_{K,\boldsymbol{s}} \cdot \boldsymbol{N}_{K,\boldsymbol{s}} > 0$ by assumption on Λ and Item 4 in Lemma 14.3). Since $\{\cdot,\cdot\}_2$ and $\langle \cdot, \cdot \rangle$ are two scalar products on $\mathrm{Im}(R_K)$, Lemma 13.8 provides an isomorphism $\mathcal{L}_K$ of $\mathrm{Im}(R_K)$ such that $\{\xi, \beta\}_2 = \langle \mathcal{L}_K(\xi), \mathcal{L}_K(\beta) \rangle$ for all $\xi, \beta \in \mathrm{Im}(R_K)$. Applying this relation to $\xi = R_K(u)$ and $\beta = R_K(v)$ and plugging the result in (14.27) shows that (14.26) holds for this choice of $\mathcal{L}_K$. ∎

The following lemma, used the in the above proof, is taken from [75].

Lemma 14.7. *Let X, Y and Z be finite dimensional vector spaces and $A : X \to Y$, $B : X \to Z$ be two linear mappings with identical kernel. Then, for any inner product $\{\cdot,\cdot\}_Y$ on Y, there exists an inner product $\{\cdot,\cdot\}_Z$ on Z such that, for all $(x, x') \in X^2$, $\{Bx, Bx'\}_Z = \{Ax, Ax'\}_Y$.*

Proof. Let $N = \ker(A) = \ker(B)$. The mappings A and B define one-to-one mappings $\bar{A} : X/N \to Y$ and $\bar{B} : X/N \to Z$ such that, if $\bar{x}$ is the class of x, $Ax = \bar{A}\bar{x}$ and $Bx = \bar{B}\bar{x}$. We can therefore work with $\bar{A}$ and $\bar{B}$ on X/N rather than with A and B on X, and assume in fact that A and B are one-to-one. Then $A : X \to \mathrm{Im}(A)$ and $B : X \to \mathrm{Im}(B)$ are isomorphisms. If $\{\cdot,\cdot\}_Y$ is an inner product on Y, we can define the inner product $\{\cdot,\cdot\}_{\mathrm{Im}(B)}$ on $\mathrm{Im}(B)$ the following way: for all $z, z' \in \mathrm{Im}(B)$, $\{z, z'\}_{\mathrm{Im}(B)} = \{AB^{-1}z, AB^{-1}z'\}_Y$, which precisely means that $\{Bx, Bx'\}_{\mathrm{Im}(B)} = \{Ax, Ax'\}_Y$ for all $x, x' \in X$. This inner product is only defined on $\mathrm{Im}(B)$, but we extend it to Z by choosing W such that $\mathrm{Im}(B) \oplus W = Z$, by selecting any inner product $\{\cdot,\cdot\}_W$ on W, and by letting $\{z, z'\}_Z = \{z_B, z'_B\}_{\mathrm{Im}(B)} + \{z_W, z'_W\}_W$ for all $z = z_B + z'_W \in Z = \mathrm{Im}(B) \oplus W$ and $z' = z'_B + z'_W \in Z$. This extension of $\{\cdot,\cdot\}_{\mathrm{Im}(B)}$ preserves the property $\{Bx, Bx'\}_Z = \{Ax, Ax'\}_Y$. ∎

14.1.1 Preliminary lemmas

We now turn to prove the properties of nMFD GDs, starting with preliminary results. In a similar way as for HMM GDs, we define the following factor which measures the invertibility properties of the isomorphisms $(\mathcal{L}_K)_{K\in\mathcal{M}}$:

$$\zeta_{\mathcal{D}} = \min \Bigg\{ \zeta > 0 \,:\, \forall K \in \mathcal{M},\ \forall v \in X_{\mathcal{V}_K},$$
$$\zeta^{-1} \sum_{\boldsymbol{s}\in\mathcal{V}_K} |V_{K,\boldsymbol{s}}| \left| \frac{R_{K,\boldsymbol{s}}(v)}{h_K} \right|^p \le \sum_{\boldsymbol{s}\in\mathcal{V}_K} |V_{K,\boldsymbol{s}}| \left| \frac{[\mathcal{L}_K R_K(v)]_{\boldsymbol{s}}}{h_K} \right|^p \tag{14.28}$$
$$\le \zeta \sum_{\boldsymbol{s}\in\mathcal{V}_K} |V_{K,\boldsymbol{s}}| \left| \frac{R_{K,\boldsymbol{s}}(v)}{h_K} \right|^p \Bigg\}.$$

The boundedness of $\zeta_{\mathcal{D}}$ is a weaker assumption than the classical coercivity assumption of nMFD methods, see, *e.g.*, [34, Eq. (5.15)]. Choosing $\mathcal{L}_K = \beta_K \mathrm{Id}$ with $\beta_K \in [\zeta^{-1}, \zeta]$ ensures that the inequalities within (14.28) is satisfied.

Lemma 14.8 (Estimate of the LLE regularity of an nMFD GD). *Let $\mathfrak{T}$ be a polytopal mesh in the sense of Definition 7.2, and let $\mathcal{D}$ be a nMFD GD on $\mathfrak{T}$ as defined in Section 14.1. Then, for any $K \in \mathcal{M}$, $\pi_K = (\pi_K^{\boldsymbol{s}})_{\boldsymbol{s}\in\mathcal{V}_K}$ is a $\mathbb{P}_0$-exact function reconstruction on K, and $\mathcal{G}_K = (\mathcal{G}_K^{\boldsymbol{s}})_{\boldsymbol{s}\in\mathcal{V}_K}$ is a $\mathbb{P}_1$-exact gradient reconstruction on K upon $\mathcal{V}_K$.*
Moreover, $\mathcal{D}$ is an LLE GD and, if $\varrho \ge \theta_{\mathfrak{T}} + \zeta_{\mathcal{D}}$ (see (7.8) and (14.28)) and

$$\varrho \ge \max_{K\in\mathcal{M}} \mathrm{Card}(\mathcal{V}_K), \tag{14.29}$$

then there exists C_1 depending only on p, d and ϱ such that $\mathrm{reg}_{\mathrm{LLE}}(\mathcal{D}) \le C_1$.

Proof. By choice (14.2) of the weights and definition (14.8) of the functions $(\pi_K^{\boldsymbol{s}})_{\boldsymbol{s}\in\mathcal{V}_K}$, $\sum_{\boldsymbol{s}\in\mathcal{V}} \pi_K^{\boldsymbol{s}} = 1$ on K and thus π_K is a $\mathbb{P}_0$-exact function reconstruction on K.
We proved in Lemma 14.3 that ∇_K is a $\mathbb{P}_1$-exact gradient reconstruction upon $\mathcal{V}_K$. Assume that $v = (A(\boldsymbol{s}))_{\boldsymbol{s}\in\mathcal{V}_K}$ interpolates an affine mapping A. As in the proof of Lemma 14.4, $v_K = A(\boldsymbol{x}_K)$ and thus $R_{K,\boldsymbol{s}}(v) = A(\boldsymbol{s}) - A(\boldsymbol{x}_K) - \nabla A \cdot (\boldsymbol{s} - \boldsymbol{x}_K) = 0$. Hence, $\mathcal{G}_K v = \nabla_K v = \nabla A$, which proves that $\mathcal{G}_K$ is a $\mathbb{P}_1$-exact gradient reconstruction on K upon $\mathcal{V}_K$.
Let us now show that $\mathcal{D}$ is an LLE GD, i.e that $\|\nabla_{\mathcal{D}}\cdot\|_{L^p(\Omega)^d}$ is a norm on $X_{\mathcal{D},0}$. If $\nabla_{\mathcal{D}} v = 0$ then (14.19) shows that $\nabla_K v = 0$ for all $K \in \mathcal{M}$ and thus, by (14.13), $R_K(v) = 0$. The definition (14.12) of $R_{K,\boldsymbol{s}}$ and the fact that $\nabla_K v = 0$ then implies $v_K = v_{\boldsymbol{s}}$ for all $\boldsymbol{s} \in \mathcal{V}_K$. Reasoning from neighbour to neighbour, we see v is a constant vector. Since $v_{\boldsymbol{s}} = 0$ for $\boldsymbol{s} \in \mathcal{V} \cap \partial\Omega$, this shows that $v = 0$.

Let us now estimate $\mathrm{reg}_{\mathrm{LLE}}(\mathcal{D})$. For any $K \in \mathcal{M}$ and any $i = \boldsymbol{s} \in I_K = \mathcal{V}_K$, we have $\boldsymbol{x}_{\boldsymbol{s}} \in \overline{K}$ and thus $\mathrm{dist}(\boldsymbol{x}_i, K) = 0$. Moreover, since all functions π_K^i are non-negative, $\sum_{i\in I_K} |\pi_K^i| = \sum_{i\in I_K} \pi_K^i = 1$ and thus $\|\pi_K\|_p = 1$. The bound on $\mathrm{reg}_{\mathrm{LLE}}(\mathcal{D})$ will therefore follow from estimating $\|\mathcal{G}_K\|_p$.
For any $\boldsymbol{s} \in \mathcal{V}_K$, by choice (14.7) of $|V_{K,\boldsymbol{s}}|$,

$$|\boldsymbol{N}_{K,\sigma}| \le \frac{h_K}{d|V_{K,\boldsymbol{s}}|} \sum_{\sigma\in\mathcal{F}_{K,\boldsymbol{s}}} \omega_\sigma^{\boldsymbol{s}} \le \frac{\theta_{\mathfrak{T}}}{d|V_{K,\boldsymbol{s}}|} \sum_{\sigma\in\mathcal{F}_{K,\boldsymbol{s}}} \omega_\sigma^{\boldsymbol{s}} d_{K,\sigma} = \theta_{\mathfrak{T}}.$$

Hence, if $v \in X_{\mathcal{V}_K}$, the definition (14.11) of $\mathcal{G}_K$ gives

$$\begin{aligned}\|\mathcal{G}_K v\|^p_{L^p(K)^d} &\le 2^{p-1}\left(|K|\,|\nabla_K v|^p + \sum_{\boldsymbol{s}\in\mathcal{V}_K}|V_{K,\boldsymbol{s}}|\left|\frac{[\mathcal{L}_K R_K(v)]_{\boldsymbol{s}}}{h_K}\right|^p|\boldsymbol{N}_{K,\sigma}|^p\right)\\ &\le 2^{p-1}\left(|K|\,|\nabla_K v|^p + \theta^p_{\mathfrak{T}}\sum_{\boldsymbol{s}\in\mathcal{V}_K}|V_{K,\boldsymbol{s}}|\left|\frac{[\mathcal{L}_K R_K(v)]_{\boldsymbol{s}}}{h_K}\right|^p\right)\\ &\le 2^{p-1}\left(|K|\,|\nabla_K v|^p + \theta^p_{\mathfrak{T}}\zeta_{\mathcal{D}}\sum_{\boldsymbol{s}\in\mathcal{V}_K}|V_{K,\boldsymbol{s}}|\left|\frac{R_{K,\boldsymbol{s}}(v)}{h_K}\right|^p\right). \quad (14.30)\end{aligned}$$

Let $V = \max_{\boldsymbol{s}\in\mathcal{V}_K}|v_{\boldsymbol{s}}|$. The definition (14.13) of $\nabla_K v$ yields, thanks to (B.1),

$$|\nabla_K v| \le \frac{V}{|K|}\sum_{\sigma\in\mathcal{F}_K}\sum_{\boldsymbol{s}\in\mathcal{V}_\sigma}\omega^{\boldsymbol{s}}_\sigma \le \frac{\theta_{\mathfrak{T}}V}{h_K|K|}\sum_{\sigma\in\mathcal{F}_K}|\sigma|d_{K,\sigma} = \frac{d\theta_{\mathfrak{T}}V}{h_K}. \quad (14.31)$$

Using the definition (14.9) of v_K, we infer

$$|R_{K,\boldsymbol{s}}(v)| \le \frac{1}{|K|}\sum_{\boldsymbol{s}\in\mathcal{V}_K}\omega^{\boldsymbol{s}}_K V + V + |\nabla_K v|h_K \le (2+d\theta_{\mathfrak{T}})V.$$

Hence,

$$\sum_{\boldsymbol{s}\in\mathcal{V}_K}|V_{K,\boldsymbol{s}}|\left|\frac{R_{K,\boldsymbol{s}}(v)}{h_K}\right|^p \le \frac{(2+d\theta_{\mathfrak{T}})^pV^p}{h_K^p}\sum_{\boldsymbol{s}\in\mathcal{V}_K}|V_{K,\boldsymbol{s}}| = \frac{(2+d\theta_{\mathfrak{T}})^pV^p}{h_K^p}|K|.$$

Substituted alongside (14.31) into (14.30), this estimate gives

$$\|\mathcal{G}_K v\|^p_{L^p(K)^d} \le 2^{p-1}\left[(d\theta_{\mathfrak{T}})^p + \theta^p_{\mathfrak{T}}\zeta_{\mathcal{D}}(2+d\theta_{\mathfrak{T}})^p\right]h_K^{-p}|K|V^p.$$

Applied to $v = v^\sigma$ for all $\sigma \in \mathcal{F}_K$, and recalling the definition (13.7) of the functions $(\mathcal{G}^{\boldsymbol{s}}_K)_{\boldsymbol{s}\in\mathcal{V}_K}$, we deduce $\|\mathcal{G}^{\boldsymbol{s}}_K\|_{L^p(K)^d} \le C_2 h_K^{-1}|K|^{1/p}$ with C_2 depending only on d, p and ϱ. The definition (7.24) of $\|\mathcal{G}_K\|_p$ and (14.29) then yield a bound on $\|\mathcal{G}_K\|_p$ that depends only on d, p and ϱ. ∎

Lemma 14.9 (Norm on $X_{\mathcal{D},0}$). *Let $\mathfrak{T}$ be a polytopal mesh in the sense of Definition 7.2, and $\mathcal{D}$ be a nMFD GD on $\mathfrak{T}$ as in Section 14.1. We take $\varrho \ge \theta_{\mathfrak{T}} + \zeta_{\mathcal{D}}$ (see (7.8) and (14.28)). Then, there exists C_3 depending only on Ω, p and ϱ such that*

$$\forall v \in X_{\mathcal{D},0},\ \sum_{K\in\mathcal{M}}\sum_{\boldsymbol{s}\in\mathcal{V}_K}|V_{K,\boldsymbol{s}}|\left|\frac{v_{\boldsymbol{s}} - v_K}{h_K}\right|^p \le C_3\|\nabla_{\mathcal{D}}v\|^p_{L^p(\Omega)}. \quad (14.32)$$

Proof.
In this proof, $A \lesssim B$ means that $A \le CB$ for some C depending only on Ω, p and ϱ. Let $v \in X_{\mathcal{D},0}$ and $K \in \mathcal{M}$. By (14.19) and Jensen's inequality,

$$|\nabla_K v|^p \leq \frac{1}{|K|}\int_K |\nabla_{\mathcal{D}} v(\boldsymbol{x})|^p \mathrm{d}\boldsymbol{x}. \tag{14.33}$$

Using Item 4 in Lemma 14.3 and the definition (14.13) of $\nabla_{\mathcal{D}}$, we infer that, for all $\boldsymbol{s} \in \mathcal{V}_K$ and a.e. $\boldsymbol{y} \in V_{K,\boldsymbol{s}}$,

$$\begin{aligned}\left|\frac{1}{h_K}[\mathcal{L}_K R_K(v)]_{\boldsymbol{s}}\right|^p &\leq \left|\frac{1}{h_K}[\mathcal{L}_K R_K(v)]_{\boldsymbol{s}} \boldsymbol{N}_{K,\boldsymbol{s}}\right|^p \\ &\lesssim |\nabla_{\mathcal{D}} v(\boldsymbol{y})|^p + \frac{1}{|K|}\int_K |\nabla_{\mathcal{D}} v(\boldsymbol{x})|^p \mathrm{d}\boldsymbol{x}.\end{aligned}$$

Integrate over $\boldsymbol{y} \in V_{K,\boldsymbol{s}}$, sum over $\boldsymbol{s} \in \mathcal{V}_K$ and use the definition (14.28) of $\zeta_{\mathcal{D}}$ to deduce

$$\sum_{\boldsymbol{s}\in\mathcal{V}_K} |V_{K,\boldsymbol{s}}| \left|\frac{R_{K,\boldsymbol{s}}(v)}{h_K}\right|^p \lesssim \int_K |\nabla_{\mathcal{D}} v(\boldsymbol{x})|^p \mathrm{d}\boldsymbol{x}.$$

Write $|v_{\boldsymbol{s}} - v_K| \leq |R_{K,\boldsymbol{s}}(v)| + h_K|\nabla_K v|$ and use (14.33) to obtain

$$\sum_{\boldsymbol{s}\in\mathcal{V}_K} |V_{K,\boldsymbol{s}}| \left|\frac{v_{\boldsymbol{s}} - v_K}{h_K}\right|^p \lesssim \int_K |\nabla_{\mathcal{D}} v(\boldsymbol{x})|^p \mathrm{d}\boldsymbol{x}.$$

Summing this estimate over $K \in \mathcal{M}$ proves (14.32). ∎

We now define a control of an nMFD GD by a polytopal toolbox, and we establish some estimates on this control.

Lemma 14.10 (Control of an nMFD GD by a polytopal toolbox). *Let $\mathfrak{T}$ be a polytopal mesh in the sense of Definition 7.2, and $\mathcal{D}$ be a nMFD GD on $\mathfrak{T}$ as in Section 14.1. Let $\boldsymbol{\Phi} : X_{\mathcal{D},0} \to X_{\mathfrak{T},0}$ be the control of $\mathcal{D}$ by the polytopal toolbox $\mathfrak{T}$ (see Definition 7.9) defined by: for $v \in X_{\mathcal{D},0}$,*

$$\begin{aligned}&\forall \sigma \in \mathcal{F}_K\,,\ \boldsymbol{\Phi}(v)_\sigma = \frac{1}{|\sigma|}\sum_{\boldsymbol{s}\in\mathcal{V}_\sigma} \omega_\sigma^{\boldsymbol{s}} v_{\boldsymbol{s}}\,, \quad \textit{and} \\ &\forall K \in \mathcal{M}\,,\ \boldsymbol{\Phi}(v)_K = v_K = \frac{1}{|K|}\sum_{\boldsymbol{s}\in\mathcal{V}_K} \omega_K^{\boldsymbol{s}} v_{\boldsymbol{s}}.\end{aligned} \tag{14.34}$$

Let $\varrho \geq \theta_{\mathfrak{T}} + \zeta_{\mathcal{D}}$ (see (7.8) and (14.28)). Then, there exists C_4 depending only on Ω, p and ϱ such that

$$\|\boldsymbol{\Phi}\|_{\mathcal{D},\mathfrak{T}} \leq C_4, \tag{14.35}$$

and

$$\omega^{\Pi}(\mathcal{D},\mathfrak{T},\boldsymbol{\Phi}) = 0\,, \quad \omega^{\nabla}(\mathcal{D},\mathfrak{T},\boldsymbol{\Phi}) = 0. \tag{14.36}$$

Proof. In this proof, $A \lesssim B$ means again that $A \leq CB$ for some C depending only on Ω, p and ϱ. By definition of $\boldsymbol{\Phi}$, for $\sigma \in \mathcal{F}_K$,

$$|\boldsymbol{\Phi}(v)_\sigma - \boldsymbol{\Phi}(v)_K| = \left|\frac{1}{|\sigma|}\sum_{\boldsymbol{s}\in\mathcal{V}_\sigma} \omega_\sigma^{\boldsymbol{s}}(v_{\boldsymbol{s}} - v_K)\right| \leq \frac{1}{|\sigma|}\sum_{\boldsymbol{s}\in\mathcal{V}_\sigma} \omega_\sigma^{\boldsymbol{s}}|v_{\boldsymbol{s}} - v_K|$$

Hence, the discrete Jensen inequality (D.11) (with the convex function $\Psi(s) = |s|^p$) and the definition of $\theta_{\mathfrak{T}}$ give

$$\left|\frac{\boldsymbol{\Phi}(v)_\sigma - \boldsymbol{\Phi}(v)_K}{d_{K,\sigma}}\right|^p \leq \theta_{\mathfrak{T}}^p \left|\frac{\boldsymbol{\Phi}(v)_\sigma - \boldsymbol{\Phi}(v)_K}{h_K}\right|^p \lesssim \frac{1}{|\sigma|}\sum_{\boldsymbol{s}\in\mathcal{V}_\sigma} \omega_\sigma^{\boldsymbol{s}} \left|\frac{v_{\boldsymbol{s}} - v_K}{h_K}\right|^p.$$

We multiply this by $|\sigma| d_{K,\sigma}$, sum over $\sigma \in \mathcal{F}_K$, and swap the sums over the vertices and edges in the right-hand side to find

$$\sum_{\sigma\in\mathcal{F}_K} |\sigma| d_{K,\sigma} \left|\frac{\boldsymbol{\Phi}(v)_\sigma - \boldsymbol{\Phi}(v)_K}{d_{K,\sigma}}\right|^p \lesssim \sum_{\boldsymbol{s}\in\mathcal{V}_K}\sum_{\sigma\in\mathcal{F}_{K,\boldsymbol{s}}} \omega_\sigma^{\boldsymbol{s}} d_{K,\sigma} \left|\frac{v_{\boldsymbol{s}} - v_K}{h_K}\right|^p.$$

Since $\sum_{\sigma\in\mathcal{F}_{K,\boldsymbol{s}}} \omega_\sigma^{\boldsymbol{s}} d_{K,\sigma} = d|V_{K,\boldsymbol{s}}|$, summing the above relation over $K \in \mathcal{M}$ yields

$$\sum_{K\in\mathcal{M}}\sum_{\sigma\in\mathcal{F}_K} |\sigma| d_{K,\sigma} \left|\frac{\boldsymbol{\Phi}(v)_\sigma - \boldsymbol{\Phi}(v)_K}{d_{K,\sigma}}\right|^p \lesssim d \sum_{K\in\mathcal{M}}\sum_{\boldsymbol{s}\in\mathcal{V}_K} |V_{K,\boldsymbol{s}}| \left|\frac{v_{\boldsymbol{s}} - v_K}{h_K}\right|^p.$$

The proof of (14.35) is completed by using (14.32) in Lemma 14.9 and by recalling the definition (7.12) of $\|\boldsymbol{\Phi}\|_{\mathcal{D},\mathfrak{T}}$.
We now turn to (14.36). By definitions (7.7c) of $\Pi_{\mathfrak{T}}$, (14.9) of $\Pi_{\mathcal{D}}$, and (14.34) of $\boldsymbol{\Phi}$, we have $\Pi_{\mathcal{D}} v = v_K = \Pi_{\mathfrak{T}}\boldsymbol{\Phi}(v)$ on K, for all $K \in \mathcal{M}$. Hence, $\omega^{\Pi}(\mathcal{D}, \mathfrak{T}, \boldsymbol{\Phi}) = 0$. We then notice that

$$\overline{\nabla}_K \boldsymbol{\Phi}(v) = \frac{1}{|K|}\sum_{\sigma\in\mathcal{F}_K} |\sigma| \boldsymbol{\Phi}(v)_\sigma \boldsymbol{n}_{K,\sigma} = \frac{1}{|K|}\sum_{\sigma\in\mathcal{F}_K}\left(\sum_{\boldsymbol{s}\in\mathcal{V}_\sigma} \omega_\sigma^{\boldsymbol{s}} v_{\boldsymbol{s}}\right)\boldsymbol{n}_{K,\sigma} = \nabla_K v.$$

Hence, (14.19) shows that $\int_K \overline{\nabla}_{\mathfrak{T}}\boldsymbol{\Phi}(v)(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = \int_K \nabla_{\mathcal{D}} v(\boldsymbol{x})\mathrm{d}\boldsymbol{x}$, and thus that $\omega^{\nabla}(\mathcal{D}, \mathfrak{T}, \boldsymbol{\Phi}) = 0$. ∎

14.1.2 Properties of nMFD gradient discretisations

The theorems presented here follow immediately, as for HMM GDs, from the preliminary results above and from Propositions 7.37 and 7.64, Theorem 7.11 and Corollary 7.12.

Theorem 14.11 (Properties of nMFD GDs). *Let $(\mathcal{D}_m)_{m\in\mathbb{N}}$ be a sequence of nMFD GDs, as in Section 14.1, defined from underlying polytopal meshes $(\mathfrak{T}_m)_{m\in\mathbb{N}}$. Assume that the sequences $(\theta_{\mathfrak{T}_m} + \eta_{\mathfrak{T}_m})_{m\in\mathbb{N}}$, $(\zeta_{\mathcal{D}_m})_{m\in\mathbb{N}}$ and $(\max_{K\in\mathcal{M}_m} \mathrm{Card}(\mathcal{V}_K))_{m\in\mathbb{N}}$ are bounded (see (7.8), (7.9) and (14.28)), and that $h_{\mathcal{M}_m} \to 0$ as $m \to \infty$.*
Then $(\mathcal{D}_m)_{m\in\mathbb{N}}$ is coercive, GD-consistent, limit-conforming and compact in the sense of Definitions 2.2, 2.4, 2.5 and 2.8.

Remark 14.12. Contrary to HMM gradient discretisations, nMFD gradient discretisations do not have a piecewise constant reconstruction for the natural choice of unknowns, nor for any obvious choice of unknowns. The nMFD GDs should therefore be modified, *e.g.*, by mass-lumping as in Section 7.3.5, to be applicable in practice to certain non-linear models.

Proposition 14.13 (Estimate on $C_{\mathcal{D}}$ and $W_{\mathcal{D}}$ for nMFD GD). *Let $\mathfrak{T}$ be a polytopal mesh of Ω in the sense of Definition 7.2, and let $\mathcal{D}$ be a nMFD GD on $\mathfrak{T}$ as in Section 14.1. We take $\varrho \geq \theta_{\mathfrak{T}} + \eta_{\mathfrak{T}} + \zeta_{\mathcal{D}}$ (see (7.8), (7.9) and (14.28)) that also satisfies (14.29). Then, there exists C_5 depending only on Ω, p and ϱ such that*

$$C_{\mathcal{D}} \leq C_5 \tag{14.37}$$

and

$$\forall \boldsymbol{\varphi} \in W^{1,p'}(\Omega)^d\,,\ W_{\mathcal{D}}(\boldsymbol{\varphi}) \leq C_5 \left\|\boldsymbol{\varphi}\right\|_{W^{1,p'}(\Omega)^d} h_{\mathcal{M}}. \tag{14.38}$$

Here, $C_{\mathcal{D}}$ and $W_{\mathcal{D}}$ are the coercivity constant and limit-conformity measure defined by (2.1) and (2.6).

Proposition 14.14 (Estimate on $S_{\mathcal{D}}$ for nMFD GD). *Let $\mathfrak{T}$ be a polytopal mesh of Ω in the sense of Definition 7.2, and $\mathcal{D}$ be a nMFD GD on $\mathfrak{T}$ as in Section 14.1. Assume that $p > d/2$ and take $\varrho \geq \theta_{\mathfrak{T}} + \zeta_{\mathcal{D}}$ (see (7.8) and (14.28)) that also satisfies (14.29). Then, there exists $C_6 > 0$, depending only on Ω, p and ϱ, such that*

$$\forall \varphi \in W^{2,p}(\Omega) \cap W_0^{1,p}(\Omega)\,,\ S_{\mathcal{D}}(\varphi) \leq C_6 \left\|\varphi\right\|_{W^{2,p}(\Omega)} h_{\mathcal{M}},$$

where $S_{\mathcal{D}}$ is defined by (2.2). This means that the space size (see Definition 2.22) of the GD satisfies

$$h_{\mathcal{D}}(W^{2,p}(\Omega) \cap W_0^{1,p}(\Omega); W^{1,p'}(\Omega)^d) \leq \max(C_6, C_5) h_{\mathcal{M}},$$

where C_5 is defined in Proposition 14.13.

Remark 14.15 (Assumption on the weights)
As already mentioned, Assumption (14.5) is not very restrictive as most natural weights will satisfy it. We emphasise that no lower bound on $\sum_{\sigma \in \mathcal{F}_{\boldsymbol{s}}} \omega_\sigma^{\boldsymbol{s}}$ is required, only that this quantity is non-zero for any $\boldsymbol{s} \in \mathcal{V}$. Even if this quantity becomes extremely small for some vertex, no component of the GDs becomes extremely small or large (we have $1 \leq |\boldsymbol{N}_{K,\boldsymbol{s}}| \leq \theta_{\mathfrak{T}}$) and all estimates on $S_{\mathcal{D}}$ or $W_{\mathcal{D}}$ remain uniform with respect to the weights.

14.2 Link with discrete duality finite volume methods (DDFV)

Let us consider the special case, in dimension $d = 3$, of an *octahedral* mesh, *i.e.* a polytopal mesh $\mathfrak{T}$ such that the elements of $\mathcal{M}$ are octahedra (open

polyhedra with eight triangular faces and six vertices, not necessarily convex; five vertices may be coplanar), and the elements of $\mathcal{F}$ are the triangular faces of the elements of $\mathcal{M}$. Each $\mathcal{F}_K$ has 8 elements, each $\mathcal{V}_K$ has 6 elements, and each $\mathcal{V}_\sigma$ has 3 elements (see Figure 14.1, left). For any $K \in \mathcal{M}$, the centre of K is defined by $\boldsymbol{x}_K = \frac{1}{6}\sum_{\boldsymbol{s}\in\mathcal{V}_K} \boldsymbol{s}$.

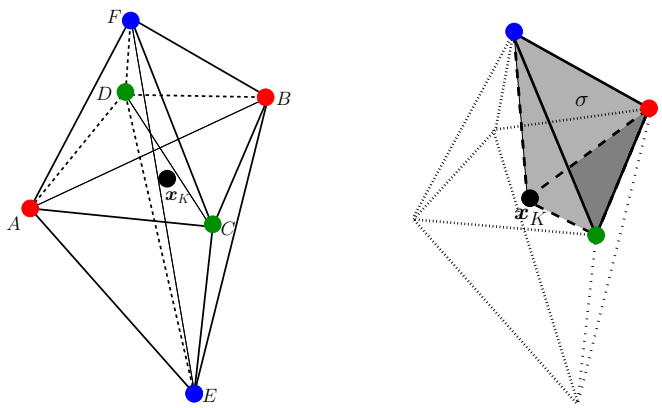


Fig. 14.1. Left: octahedral cell K. Right: illustration of $T_{K,\sigma}$ (greyed domain).

We consider a modification of an nMFD GD $\mathcal{D} = (X_{\mathcal{D},0}, \Pi_{\mathcal{D}}, \nabla_{\mathcal{D}})$ on $\mathfrak{T}$, in which the space of discrete unknowns is unchanged, the gradient reconstruction is only built from the consistent part (14.10) of $\nabla_{\mathcal{D}}$, and the reconstructed functions are piecewise constant on sub-tetrahedra. Precisely, we take for each triangle $\sigma \in \mathcal{F}$ the order 1 quadrature rule (14.3) with equal weights $\omega_\sigma^{\boldsymbol{s}} = \frac{|\sigma|}{3}$, and we define $\mathcal{D}^\star = (X_{\mathcal{D},0}, \Pi_{\mathcal{D}^\star}, \nabla_{\mathcal{D}^\star})$ the following way.

1. $\nabla_{\mathcal{D}^\star} \ : \ X_{\mathcal{D},0} \to L^p(\Omega)^d$ is given by

$$\forall v \in X_{\mathcal{D},0}\,,\ \forall K \in \mathcal{M}\,,\ \text{for a.e. } \boldsymbol{x} \in K\,,$$
$$\nabla_{\mathcal{D}^\star} v(\boldsymbol{x}) = \nabla_K v = \frac{1}{|K|}\sum_{\sigma\in\mathcal{F}_K} |\sigma| \left(\frac{1}{3}\sum_{\boldsymbol{s}\in\mathcal{V}_K} v_{\boldsymbol{s}}\right) \boldsymbol{n}_{K,\sigma}.$$

   (given the equal-weights quadrature rule chosen for each face, this expression of $\nabla_K v$ corresponds to (14.10)).
2. $\Pi_{\mathcal{D}^\star} \ : \ X_{\mathcal{D},0} \to L^p(\Omega)$ is given by

$$\forall v \in X_{\mathcal{D},0}\,,\ \forall K \in \mathcal{M}\,,\ \forall \sigma \in \mathcal{F}_K\,,\ \text{for a.e. } \boldsymbol{x} \in T_{K,\sigma}\,,$$
$$\Pi_{\mathcal{D}^\star} v(\boldsymbol{x}) = \frac{1}{3}\sum_{\boldsymbol{s}\in\mathcal{V}_\sigma} v_{\boldsymbol{s}},$$

   where $T_{K,\sigma}$ is the tetrahedra formed by $\boldsymbol{x}_K$ and σ (see Figure 14.1, right).

The following lemma characterises the reconstructed gradient.

Lemma 14.16. *For any $v \in X_{\mathcal{D},0}$ and any $K \in \mathcal{M}$, the constant vector $(\nabla_{\mathcal{D}^\star} v)_{|K}$ is the unique vector $\boldsymbol{\xi} \in \mathbb{R}^3$ such that*

$$\text{For all opposite vertices } (\boldsymbol{s}_0, \boldsymbol{s}_1) \text{ of } K,\ \boldsymbol{\xi} \cdot (\boldsymbol{s}_0 - \boldsymbol{s}_1) = v_{\boldsymbol{s}_0} - v_{\boldsymbol{s}_1}. \tag{14.39}$$

Remark 14.17. The opposite vertices in the octahedra in Figure 14.1 are (A, B), (C, D) and (E, F).

Proof. First note that, since the three directions defined by the three pairs of opposite vertices in K are linearly independent, (14.39) indeed characterises one and only one vector $\boldsymbol{\xi} \in \mathbb{R}^3$. We therefore just have to show that $(\nabla_{\mathcal{D}^\star} v)_{|K}$ satisfies (14.39). We have

$$(\nabla_{\mathcal{D}^\star} v)_{|K} = \frac{1}{|K|}\frac{1}{3} \sum_{\boldsymbol{s} \in \mathcal{V}_K} v_{\boldsymbol{s}} \sum_{\sigma \in \mathcal{F}_K | \boldsymbol{s} \in \mathcal{V}_\sigma} |\sigma| \boldsymbol{n}_{K,\sigma}. \tag{14.40}$$

Let us consider for example the case where $\boldsymbol{s} = A$ in Figure 14.1. For a triangular face σ, the outer normal $|\sigma|\boldsymbol{n}_{K,\sigma}$ can be written as the exterior product of two of the edges of σ (with proper orientation). This gives

$$\begin{aligned}\sum_{\sigma \in \mathcal{F}_K | \boldsymbol{s} \in \mathcal{V}_\sigma} |\sigma| \boldsymbol{n}_{K,\sigma} &= \frac{1}{2}(\overrightarrow{AC} \times \overrightarrow{AF} + \overrightarrow{AF} \times \overrightarrow{AD} + \overrightarrow{AD} \times \overrightarrow{AE} + \overrightarrow{AE} \times \overrightarrow{AC}) \\ &= \frac{1}{2}(\overrightarrow{DC} \times \overrightarrow{AF} + \overrightarrow{CD} \times \overrightarrow{AE}) = -\frac{1}{2}\overrightarrow{CD} \times \overrightarrow{EF}.\end{aligned}$$

Applying this to all vertices of K, and since $|K| = \frac{1}{6}\Delta_K$ with $\Delta_K = \det(\overrightarrow{AB}, \overrightarrow{CD}, \overrightarrow{EF})$, we deduce from (14.40) that

$$\begin{aligned}(\nabla_{\mathcal{D}^\star} v)_{|K} = \frac{1}{\Delta_K}\Big(&(v_B - v_A)\overrightarrow{CD} \times \overrightarrow{EF} + (v_D - v_C)\overrightarrow{EF} \times \overrightarrow{AB} \\ &+ (v_F - v_E)\overrightarrow{AB} \times \overrightarrow{CD}\Big).\end{aligned}$$

Property (14.39) is then straightforward. Considering for example the case $(\boldsymbol{s}_0, \boldsymbol{s}_1) = (B, A)$, the formula follows from $(\overrightarrow{EF} \times \overrightarrow{AB}) \cdot \overrightarrow{AB} = (\overrightarrow{AB} \times \overrightarrow{CD}) \cdot \overrightarrow{AB} = 0$ and $(\overrightarrow{CD} \times \overrightarrow{EF}) \cdot \overrightarrow{AB} = \det(\overrightarrow{CD}, \overrightarrow{EF}, \overrightarrow{AB}) = \Delta_K$. ∎

This lemma proves that $\|\nabla_{\mathcal{D}^\star} \cdot\|_{L^p(\Omega)^d}$ is a norm on $X_{\mathcal{D},0}$. Moreover, (14.39) is a well-known characterisation of the reconstructed gradient in the CeVeFE discrete duality finite volume (DDFV) method [52, 53], which is piecewise constant on the so-called "diamond cells". The function reconstruction $\Pi_{\mathcal{D}^\star}$ has been defined to match the function reconstruction used in the CeVeFE-DDFV; this reconstruction is what ensures the discrete duality (Stokes) formula, that gave the name to DDFV methods. Hence, the CeVeFE-DDFV scheme can be

considered as an nMFD scheme on octahedral meshes, without the need for a stabilisation and with a different function reconstruction.
A complete analysis of the CeVeFE-DDFV method as a GDM may be found in [77]. The same analysis also applies to the case $d = 2$, in which case the mesh is now quadrangular (its cells still correspond to the "diamond cells" in the DDFV terminology).

Part IV

Appendix

The first chapter of this appendix presents gradient discretisations (GDs), and some of their properties, in an abstract setting. This setting is shown to cover most boundary conditions considered in Chapters 2 and 3, which enables us to present unified proofs for some results on gradient discretisations.

Parts I (for elliptic problems) and II (for parabolic problems) introduced the properties (coercivity, GD-consistency, etc.) needed on GDs to generate convergent gradient schemes (GSs). Chapters B and C now introduce the technical tools which are used in Part III to prove that a given GD satisfies these core properties.

Chapter B is devoted to discrete functional analysis tools, that is, the translation to the discrete setting of classical results of functional analysis (Poincaré inequality, compactness theorems, etc.). These tools are used, in Section 7.6 in conjunction with the notion of control of a GD by a polytopal toolbox, to establish the coercivity, limit-conformity and compactness of gradient discretisations. They also provide explicit estimates on $C_{\mathcal{D}}$ and $W_{\mathcal{D}}$. Most of the results and notions presented in this chapter expand results that originally appeared in [96].

Chapter C covers generic compactness results for time-dependent functions, with abstract Banach spaces E as co-domains. Both averaged-in-time (i.e., in $L^p(0,T;E)$) and uniform-in-time (i.e., in $L^\infty(0,T;E)$) situations are considered, and the focus is on piecewise-in-time functions, usually encountered in numerical schemes for time-dependent problems.

In Chapter D, classical technical results are presented; these results are used throughout the book.

Finally, numerical examples are provided in Chapter E to illustrate the behaviour of particular GDs applied to specific problems.

A

Gradient discretisations – abstract setting

Chapters 2 and 3 introduced the notions and properties of gradient discretisation for various boundary conditions. Here, GDs are developed in a generic setting that is shown to cover most boundary conditions. This enables unified proofs of many results on GDs, independently of the specific boundary conditions they are designed for.
Throughout this chapter we take $p \in (1, +\infty)$ and let p' be such that $1/p + 1/p' = 1$.

A.1 Continuous abstract setting

Let L and $\boldsymbol{L}$ be reflexive Banach spaces, with respective topological dual spaces L' and $\boldsymbol{L}'$. Let $\boldsymbol{H}$ be a dense subspace of $\boldsymbol{L}'$; this implies (and is actually equivalent to) the following property.

$$\text{For all } \boldsymbol{u} \in \boldsymbol{L},\ (\forall \boldsymbol{v} \in \boldsymbol{H}, \langle \boldsymbol{v}, \boldsymbol{u} \rangle_{\boldsymbol{L}',\boldsymbol{L}} = 0) \Rightarrow \boldsymbol{u} = 0. \tag{A.1}$$

Take a linear operator $\mathrm{D} : \boldsymbol{H} \to L'$, such that the graph of D is closed in $\boldsymbol{L}' \times L'$. Endowed with the graph norm $\|\boldsymbol{v}\|_{\boldsymbol{H}} = \|\boldsymbol{v}\|_{\boldsymbol{L}'} + \|\mathrm{D}\boldsymbol{v}\|_{L'}$, $\boldsymbol{H}$ is a Banach space continuously embedded in $\boldsymbol{L}'$. Define

$$B = \{u \in L \,:\, \exists \boldsymbol{u} \in \boldsymbol{L}, \forall \boldsymbol{v} \in \boldsymbol{H}, \langle \boldsymbol{v}, \boldsymbol{u} \rangle_{\boldsymbol{L}',\boldsymbol{L}} + \langle \mathrm{D}\boldsymbol{v}, u \rangle_{L',L} = 0\}. \tag{A.2}$$

Thanks to (A.1), for all $u \in B$ the element $\boldsymbol{u} \in \boldsymbol{L}$ in the definition of B is unique. This defines a linear operator $\mathrm{G} : B \to \boldsymbol{L}$, adjoint operator of $-\mathrm{D}$ in the sense of [124, p. 167], such that $\boldsymbol{u} = \mathrm{G}u$, that is,

$$\forall u \in B,\ \forall \boldsymbol{v} \in \boldsymbol{H},\ \langle \boldsymbol{v}, \mathrm{G}u \rangle_{\boldsymbol{L}',\boldsymbol{L}} + \langle \mathrm{D}\boldsymbol{v}, u \rangle_{L',L} = 0. \tag{A.3}$$

This definition easily shows that the graph of G is closed in $L \times \boldsymbol{L}$. As a consequence, B endowed with the graph norm $\|u\|_{B,\mathcal{G}} = \|u\|_L + \|\mathrm{G}u\|_{\boldsymbol{L}}$ is a Banach space continuously embedded in L. By [124, Theorem 5.29], B is also dense in L.

Remark A.1. Equivalently, we could define the same abstract setting by starting from a dense subspace B of L and a linear operator $\mathrm{G} : B \to \boldsymbol{L}$ with closed graph, and by defining then $\boldsymbol{H}$ and $\mathrm{D} : \boldsymbol{H} \to L'$ by:

$$\boldsymbol{H} = \{\boldsymbol{v} \in \boldsymbol{L}' \,:\, \exists w \in L', \forall u \in B, \langle \boldsymbol{v}, \mathrm{G}u\rangle_{\boldsymbol{L}',\boldsymbol{L}} + \langle w, u\rangle_{L',L} = 0\}$$

and, for $\boldsymbol{v} \in \boldsymbol{H}$, $\mathrm{D}\boldsymbol{v}$ is the element w in the definition of $\boldsymbol{H}$.

Let V be a subspace of L' such that

$$L' = \mathrm{Im}(\mathrm{D}) + V. \tag{A.4}$$

We denote by $|\cdot|_L$ the semi-norm on L defined by

$$\forall u \in L,\ |u|_L = \begin{cases} \sup\limits_{\mu \in V\setminus\{0\}} \dfrac{|\langle \mu, u\rangle_{L',L}|}{\|\mu\|_{L'}} & \text{if } V \neq \{0\}, \\ 0 & \text{if } V = \{0\}. \end{cases} \tag{A.5}$$

Lemma A.2. *With the definitions and notations above, for $u \in B$ we set*

$$\|u\|_B = (|u|_L^p + \|\mathrm{G}u\|_{\boldsymbol{L}}^p)^{1/p}. \tag{A.6}$$

Then $\|\cdot\|_B$ and $\|\cdot\|_{B,\mathcal{G}}$ are two equivalent norms.

Proof. Since $\|\cdot\|_{B,\mathcal{G}}$ is a norm, proving its equivalence with $\|\cdot\|_B$ establishes that this later semi-norm is also a norm.
For any $u \in L$ and $\mu \in V$, we have $|\langle \mu, u\rangle_{L',L}| \leq \|\mu\|_{L'} \|u\|_L$. Hence, $|u|_L \leq \|u\|_L$ and thus $\|u\|_B \leq 2^{1/p} \|u\|_{B,\mathcal{G}}$. This proves half of the equivalence. To prove the other half, we just need to show that

$$E = \{u \in B \,:\, \|u\|_B = 1\}$$

is bounded in L. Indeed, this establishes the existence of $M \geq 0$ such that, for all $u \in E$, $\|u\|_L \leq M$ and thus, since $\|\mathrm{G}u\|_{\boldsymbol{L}} \leq \|u\|_B = 1$,

$$\|u\|_{B,\mathcal{G}} \leq M + 1 = (M+1)\|u\|_B\,.$$

By homogeneity of the semi-norms, this concludes the proof that $\|\cdot\|_{B,\mathcal{G}}$ and $\|\cdot\|_B$ are equivalent on B.
To prove that E is bounded, take $f \in L'$ and apply (A.4) to get $\boldsymbol{v}_f \in \boldsymbol{H}$ and $\mu_f \in V$ such that $f = \mathrm{D}\boldsymbol{v}_f + \mu_f$. Then, for any $u \in E$, by definition of the semi-norm $|\cdot|_L$ and since $\|\mathrm{G}u\|_{\boldsymbol{L}} \leq 1$ and $|u|_L \leq 1$,

$$\begin{aligned} |\langle f, u\rangle_{L',L}| &= |\langle \mathrm{D}\boldsymbol{v}_f, u\rangle_{L',L} + \langle \mu_f, u\rangle_{L',L}| \\ &= |-\langle \boldsymbol{v}_f, \mathrm{G}u\rangle_{\boldsymbol{L}',\boldsymbol{L}} + \langle \mu_f, u\rangle_{L',L}| \\ &\leq \|\boldsymbol{v}_f\|_{\boldsymbol{L}'} \|\mathrm{G}u\|_{\boldsymbol{L}} + \|\mu_f\|_{L'} |u|_L \\ &\leq \|\boldsymbol{v}_f\|_{\boldsymbol{L}'} + \|\mu_f\|_{L'}\,. \end{aligned}$$

This shows that $\{\langle f, u\rangle_{L',L} : u \in E\}$ is bounded by some constant depending on f. Since this is valid for any $f \in L'$, the Banach–Steinhaus theorem [33, Theorem 2.2] shows that E is bounded in L. ∎

In the following sections, we make explicit correspondences between this abstract setting and the specific settings of Dirichlet, Neumann and Fourier boundary conditions.

A.1.1 Homogeneous Dirichlet BCs

We consider in this case the following spaces and operator D:

- $L = L^p(\Omega)$, so that $L' = L^{p'}(\Omega)$.
- $\boldsymbol{L} = L^p(\Omega)^d$, so that $\boldsymbol{L}' = L^{p'}(\Omega)^d$.
- $\boldsymbol{H} = W^{p'}_{\mathrm{div}}(\Omega)$, $\mathrm{D} = \mathrm{div}$ and $V = \{0\}$.
- $B = W^{1,p}_0(\Omega)$.

The choice of V ensures (A.4) since $\mathrm{div} : W^{p'}_{\mathrm{div}}(\Omega) \to L^{p'}(\Omega)$ is surjective (see the proof of Lemma 2.6).

The operator $\mathrm{G} : W^{1,p}_0(\Omega) \to L^p(\Omega)^d$ is then the standard gradient, $\mathrm{G} = \nabla$. Moreover, (A.5) shows that $|\cdot|_L = 0$, and thus $\|u\|_B = \|\nabla u\|_{L^p(\Omega)^d}$.

A.1.2 Homogeneous Neumann BCs

We consider in this case the following spaces and operator D:

- $L = L^p(\Omega)$, so that $L' = L^{p'}(\Omega)$.
- $\boldsymbol{L} = L^p(\Omega)^d$, so that $\boldsymbol{L}' = L^{p'}(\Omega)^d$.
- $\boldsymbol{H} = W^{p'}_{\mathrm{div},0}(\Omega)$, $\mathrm{D} = \mathrm{div}$ and $V = \mathbb{R}1_\Omega$.
- $B = W^{1,p}(\Omega)$.

To see that this V satisfies (A.4), take $f \in L^{p'}(\Omega)$ and write $f = f_0 + \mu_f 1_\Omega$ with $\mu_f = \frac{1}{|\Omega|}\int_\Omega f(\boldsymbol{x})\mathrm{d}\boldsymbol{x}$. Since f_0 has zero average, there exists $\overline{u} \in W^{1,p}(\Omega)$ solution of $-\mathrm{div}(|\nabla\overline{u}|^{p-2}\nabla\overline{u}) = f_0$ in Ω with homogeneous Neumann boundary conditions $|\nabla\overline{u}|^{p-2}\nabla\overline{u}\cdot\boldsymbol{n}_{\partial\Omega} = 0$ on $\partial\Omega$. Set $\boldsymbol{\varphi} = -|\nabla\overline{u}|^{p-2}\nabla\overline{u} \in W^{p'}_{\mathrm{div},0}(\Omega)$ and notice that $\mathrm{div}\boldsymbol{\varphi} = f_0$, so that $f = \mathrm{div}\boldsymbol{\varphi} + \mu_f 1_\Omega \in \mathrm{Im}(\mathrm{D}) + V$.

The operator $\mathrm{G} : W^{1,p}(\Omega) \to L^p(\Omega)^d$ is the standard gradient, $\mathrm{G} = \nabla$. The definition (A.5) gives $|u|_L = |\Omega|^{-1/p'}|\int_\Omega u(\boldsymbol{x})\mathrm{d}\boldsymbol{x}|$. The factor $|\Omega|^{-1/p'}$ can be dropped without changing anything to the analysis, and the norm on B can therefore be defined by

$$\|u\|_B = \left(\left|\int_\Omega u(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\right|^p + \|\nabla u\|^p_{L^p(\Omega)^d}\right)^{1/p}.$$

A.1.3 Non-homogeneous Neumann BCs

We consider in this case the following spaces and operator D:

- $L = L^p(\Omega) \times L^p(\partial\Omega)$, so that $L' = L^{p'}(\Omega) \times L^{p'}(\partial\Omega)$.
- $\boldsymbol{L} = L^p(\Omega)^d$, so that $\boldsymbol{L}' = L^{p'}(\Omega)^d$.
- $\boldsymbol{H} = W^{p'}_{\mathrm{div},\partial}(\Omega)$, $\mathrm{D}\boldsymbol{\varphi} = (\mathrm{div}\boldsymbol{\varphi}, -\gamma_{\mathbf{n}}\boldsymbol{\varphi})$ and $V = \mathbb{R}(1_\Omega, 0)$.
- $B = \{(u, \gamma u) \in L^p(\Omega) \times L^p(\partial\Omega) \,:\, u \in W^{1,p}(\Omega)\}$.

To prove that (A.4) holds, write any $(f, g) \in L^{p'}(\Omega) \times L^{p'}(\partial\Omega)$ as $(f, g) = (f_0 + \mu_{f,g} 1_\Omega, g)$ where

$$\mu_{f,g} = \frac{1}{|\Omega|}\int_\Omega f(\boldsymbol{x})\mathrm{d}\boldsymbol{x} + \frac{1}{|\Omega|}\int_{\partial\Omega} g(\boldsymbol{x})\mathrm{d}\gamma(\boldsymbol{x}).$$

Then

$$\int_\Omega f_0(\boldsymbol{x})\mathrm{d}\boldsymbol{x} + \int_{\partial\Omega} g(\boldsymbol{x})\mathrm{d}\gamma(\boldsymbol{x}) = \int_\Omega f(\boldsymbol{x})\mathrm{d}\boldsymbol{x} + \int_{\partial\Omega} g(\boldsymbol{x})\mathrm{d}\gamma(\boldsymbol{x}) - |\Omega|\mu_{f,g} = 0.$$

Hence, (f_0, g) satisfies the compatibility condition to be source and boundary terms in a non-homogeneous Neumann problem. There exists thus $\overline{u} \in W^{1,p}(\Omega)$ solution of $-\mathrm{div}(|\nabla\overline{u}|^{p-2}\nabla\overline{u}) = f_0$ in Ω with boundary conditions $|\nabla\overline{u}|^{p-2}\nabla\overline{u} \cdot \boldsymbol{n}_{\partial\Omega} = g$ on $\partial\Omega$. Set $\boldsymbol{\varphi} = -|\nabla\overline{u}|^{p-2}\nabla\overline{u} \in W^{p'}_{\mathrm{div},\partial}(\Omega)$. We have $\mathrm{div}\boldsymbol{\varphi} = f_0$ and $-\gamma_{\mathbf{n}}\boldsymbol{\varphi} = g$, that is, $\mathrm{D}\boldsymbol{\varphi} = (f_0, g)$. Hence, $(f, g) = \mathrm{D}\boldsymbol{\varphi} + \mu_{f,g}(1_\Omega, 0) \in \mathrm{Im}(\mathrm{D}) + V$.

The operator $\mathrm{G} : B \to L^p(\Omega)^d$ is then given by $\mathrm{G}(u, \gamma u) = \nabla u$, and Relation (A.3) is the standard Stokes formula

$$\begin{aligned}
&\forall u \in W^{1,p}(\Omega)\,,\ \forall \boldsymbol{\varphi} \in W^{p'}_{\mathrm{div},\partial}(\Omega),\\
&\int_\Omega \nabla u(\boldsymbol{x}) \cdot \boldsymbol{\varphi}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} + \int_\Omega u(\boldsymbol{x})\mathrm{div}\boldsymbol{\varphi}(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\\
&\qquad\qquad - \int_{\partial\Omega} \gamma u(\boldsymbol{x})\gamma_{\mathbf{n}}\boldsymbol{\varphi}(\boldsymbol{x})\mathrm{d}\gamma(\boldsymbol{x}) = 0.
\end{aligned}$$

Moreover, (A.5) gives $|(u, w)|_L = |\Omega|^{-1/p'}|\int_\Omega u(\boldsymbol{x})\mathrm{d}\boldsymbol{x}|$, and thus, after dropping the factor $|\Omega|^{-1/p'}$, the norm on B is

$$\|(u, \gamma u)\|_B = \left(\left|\int_\Omega u(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\right|^p + \|\nabla u\|^p_{L^p(\Omega)^d}\right)^{1/p}.$$

A.1.4 Fourier BCs

The spaces and operator D are exactly the same as for non-homogeneous Neumann BCs, except that we now take $V = \{0\} \times L^{p'}(\partial\Omega)$. Any $(f, g) \in L^{p'}(\Omega) \times L^{p'}(\partial\Omega)$ can be written $(f, g) = (f, g_0) + (0, \nu_{f,g} 1_{\partial\Omega})$, where

$$\nu_{f,g} = \frac{1}{|\partial\Omega|}\int_\Omega f(\boldsymbol{x})\mathrm{d}\boldsymbol{x} + \frac{1}{|\partial\Omega|}\int_{\partial\Omega} g(\boldsymbol{x})\mathrm{d}\gamma(\boldsymbol{x}).$$

Then (f, g_0) satisfies the compatibility condition to be source and boundary terms in a non-homogeneous Neumann problem, which allows us, as in the case of non-homogeneous Neumann BCs, to find $\boldsymbol{\varphi} \in W^{p'}_{\mathrm{div},\partial}(\Omega)$ such that $(f, g) = \mathrm{D}\boldsymbol{\varphi} + (0, \nu_{f,g}1_{\partial\Omega}) \in \mathrm{Im}(\mathrm{D}) + V$.
The definition (A.5) gives $|(u, w)|_L = \|w\|_{L^p(\partial\Omega)}$, which leads to

$$\|(u, \gamma u)\|_B = \left(\|\gamma u\|^p_{L^p(\partial\Omega)} + \|\nabla u\|^p_{L^p(\Omega)^d}\right)^{1/p}. \tag{A.7}$$

Remark A.3. The decomposition of (f, g) made above shows that we could take $V = \mathbb{R}(0, 1_{\partial\Omega})$, and thus that the norm on B could be weakened into

$$\|(u, \gamma u)\|_B = \left(\left|\int_{\partial\Omega} \gamma u(\boldsymbol{x})\mathrm{d}\gamma(\boldsymbol{x})\right|^p + \|\nabla u\|^p_{L^p(\Omega)^d}\right)^{1/p}.$$

This norm is actually equivalent to (A.7). In the context of Fourier boundary conditions, (A.7) is the standard norm in which estimates on solutions to PDEs are obtained.

A.2 Gradient discretisation in the abstract setting

Based on the abstract setting described in Section A.1, we define a notion of gradient discretisation, with corresponding properties and consequences. It can be checked that, with the particular choices described in Sections A.1.1 to A.1.4, the following theory gives the concepts and results mentioned in Section 2.1 (GDs for homogeneous Dirichlet BCs), Section 3.1.1 (GDs for homogeneous and non-homogeneous Neumann BCs – with a variant definition of consistency in the latter case) and Section 3.2.1 (GDs for Fourier BCs).

Definition A.4 (GD, abstract setting). *In the context described in Section A.1, a gradient discretisation $\mathcal{D}$ is defined by $\mathcal{D} = (X_\mathcal{D}, \mathrm{P}_\mathcal{D}, \mathrm{G}_\mathcal{D})$, where:*

1. *The set of discrete unknowns $X_\mathcal{D}$ is a finite dimensional vector space on $\mathbb{R}$.*
2. *The "function" reconstruction $\mathrm{P}_\mathcal{D} : X_\mathcal{D} \to L$ is a linear mapping that reconstructs, from an element of $X_\mathcal{D}$, an element in L.*
3. *The "gradient" reconstruction $\mathrm{G}_\mathcal{D} : X_\mathcal{D} \to \boldsymbol{L}$ is a linear mapping that reconstructs, from an element of $X_\mathcal{D}$, an element in $\boldsymbol{L}$.*
4. *The mappings $\mathrm{P}_\mathcal{D}$ and $\mathrm{G}_\mathcal{D}$ are such that*

$$\|u\|_\mathcal{D} := (|\mathrm{P}_\mathcal{D}u|^p_L + \|\mathrm{G}_\mathcal{D}u\|^p_{\boldsymbol{L}})^{1/p}$$

is a norm on $X_\mathcal{D}$.

Remark A.5 ($\mathrm{P}_{\mathcal{D}}$ and $\mathrm{G}_{\mathcal{D}}$ for various boundary conditions). Definition A.4 is translated in Chapters 2 and 3 to the contexts of homogeneous Dirichlet BCs, homogeneous and non-homogeneous Neumann BCs, and Fourier BCs. Using the notations in these chapters, we always have $\mathrm{G}_{\mathcal{D}} = \nabla_{\mathcal{D}}$. The operator $\mathrm{P}_{\mathcal{D}}$ however depends upon the boundary conditions:

- For homogeneous Dirichlet BCs (Definition 2.1) and homogeneous Neumann BCs (Definition 3.1): $\mathrm{P}_{\mathcal{D}} = \Pi_{\mathcal{D}}$,
- For non-homogeneous Neumann BCs (Definition 3.11) and Fourier BCs (Definition 3.36): $\mathrm{P}_{\mathcal{D}} = (\Pi_{\mathcal{D}}, \mathbb{T}_{\mathcal{D}})$.

Definition A.6 (Coercivity, abstract setting)

If $\mathcal{D}$ is a gradient discretisation in the sense of Definition A.4, let $C_{\mathcal{D}}$ be the norm of $\mathrm{P}_{\mathcal{D}}$:

$$C_{\mathcal{D}} = \max_{v \in X_{\mathcal{D}} \setminus \{0\}} \frac{\|\mathrm{P}_{\mathcal{D}} v\|_L}{\|v\|_{\mathcal{D}}}. \tag{A.8}$$

A sequence $(\mathcal{D}_m)_{m\in\mathbb{N}}$ of gradient discretisations is **coercive** if there exists $C_P \in \mathbb{R}_+$ such that $C_{\mathcal{D}_m} \leq C_P$ for all $m \in \mathbb{N}$.

Definition A.7 (Limit-conformity,definition)

If $\mathcal{D}$ is a gradient discretisation in the sense of Definition A.4, let $W_{\mathcal{D}} : \boldsymbol{H} \to [0, +\infty)$ be given by

$$\forall \boldsymbol{\varphi} \in \boldsymbol{H}\,,$$
$$W_{\mathcal{D}}(\boldsymbol{\varphi}) = \sup_{u \in X_{\mathcal{D}} \setminus \{0\}} \frac{|\langle \boldsymbol{\varphi}, \mathrm{G}_{\mathcal{D}} u\rangle_{\boldsymbol{L}',\boldsymbol{L}} + \langle \mathrm{D}\boldsymbol{\varphi}, \mathrm{P}_{\mathcal{D}} u\rangle_{L',L}|}{\|u\|_{\mathcal{D}}}. \tag{A.9}$$

A sequence $(\mathcal{D}_m)_{m\in\mathbb{N}}$ of gradient discretisations is **limit-conforming** if

$$\forall \boldsymbol{\varphi} \in \boldsymbol{H}\,, \quad \lim_{m\to\infty} W_{\mathcal{D}_m}(\boldsymbol{\varphi}) = 0. \tag{A.10}$$

The following lemma shows that the limit-conformity is stronger than the coercivity.

Lemma A.8 (Limit-conformity implies coercivity, abstract setting). *Let $(\mathcal{D}_m)_{m\in\mathbb{N}}$ be a sequence of gradient discretisations that is limit-conforming in the sense of Definition A.7. Then $(\mathcal{D}_m)_{m\in\mathbb{N}}$ is also coercive in the sense of Definition A.6.*

Proof. Set

$$E = \left\{ \frac{\mathrm{P}_{\mathcal{D}_m} v}{\|v\|_{\mathcal{D}_m}} \in L \,:\, m \in \mathbb{N}\,,\ v \in X_{\mathcal{D}_m} \backslash \{0\} \right\}.$$

Proving the coercivity of $(\mathcal{D}_m)_{m\in\mathbb{N}}$ consists in proving that E is bounded in L. Let $f \in L'$. By (A.4), there exists $\boldsymbol{v}_f \in \boldsymbol{H}$ and $\mu_f \in V$ such that $f = \mathrm{D}\boldsymbol{v}_f + \mu_f$. The definition of $|\cdot|_L$ shows that $|\langle \mu_f, \cdot\rangle_{L',L}| \le \|\mu_f\|_{L'} |\cdot|_L$. For $z \in E$, take $m \in \mathbb{N}$ and $v \in X_{\mathcal{D}_m}\backslash\{0\}$ such that $z = \frac{\mathrm{P}_{\mathcal{D}_m} v}{\|v\|_{\mathcal{D}_m}}$ and write

$$\begin{aligned}
|\langle f, z\rangle_{L',L}| &\le \frac{1}{\|v\|_{\mathcal{D}_m}} |\langle \mathrm{D}\boldsymbol{v}_f, \mathrm{P}_{\mathcal{D}_m} v\rangle_{L',L}| + \frac{1}{\|v\|_{\mathcal{D}_m}} |\langle \mu_f, \mathrm{P}_{\mathcal{D}_m} v\rangle_{L',L}| \\
&\le \frac{1}{\|v\|_{\mathcal{D}_m}} |\langle \mathrm{D}\boldsymbol{v}_f, \mathrm{P}_{\mathcal{D}_m} v\rangle_{L',L} + \langle \boldsymbol{v}_f, \mathrm{G}_{\mathcal{D}_m} v\rangle_{\boldsymbol{L}',\boldsymbol{L}}| \\
&\quad + \frac{1}{\|v\|_{\mathcal{D}_m}} |\langle \boldsymbol{v}_f, \mathrm{G}_{\mathcal{D}_m} v\rangle_{\boldsymbol{L}',\boldsymbol{L}}| + \frac{1}{\|v\|_{\mathcal{D}_m}} \|\mu_f\|_{L'} |\mathrm{P}_{\mathcal{D}_m} v|_L \\
&\le W_{\mathcal{D}_m}(\boldsymbol{v}_f) + \|\boldsymbol{v}_f\|_{\boldsymbol{L}'} + \|\mu_f\|_{L'}\,. \qquad \text{(A.11)}
\end{aligned}$$

In the last inequality we used $|\mathrm{P}_{\mathcal{D}_m} v|_L \le \|v\|_{\mathcal{D}_m}$ and $\|\mathrm{G}_{\mathcal{D}_m} v\|_{\boldsymbol{L}} \le \|v\|_{\mathcal{D}_m}$. Since $(\mathcal{D}_m)_{m\in\mathbb{N}}$ is limit-conforming, $(W_{\mathcal{D}_m}(\boldsymbol{v}_f))_{m\in\mathbb{N}}$ converges to 0 and is therefore bounded. Estimate (A.11) thus shows that $\{\langle f, z\rangle_{L',L} \,:\, z \in E\}$ is bounded by some constant depending on f. Since this is valid for any $f \in L'$, we infer from the Banach–Steinhaus theorem [33, Theorem 2.2] that E is bounded in L. ∎

Checking limit-conformity is made easier by the following result, which reduces the set of elements $\boldsymbol{\varphi}$ on which the convergence in (A.10) has to be asserted.

Lemma A.9 (Equivalent condition for limit-conformity, abstract setting). *Let $(\mathcal{D}_m)_{m\in\mathbb{N}}$ be a sequence of gradient discretisations in the sense of Definition A.6. Then $(\mathcal{D}_m)_{m\in\mathbb{N}}$ is limit-conforming in the sense of Definition A.7 if and only if it is coercive in the sense of Definition A.6, and there exists a dense subset $\boldsymbol{H}_d$ of $\boldsymbol{H}$ such that*

$$\forall \boldsymbol{\psi} \in \boldsymbol{H}_d,\ \lim_{m\to\infty} W_{\mathcal{D}_m}(\boldsymbol{\psi}) = 0. \qquad \text{(A.12)}$$

Proof. If $(\mathcal{D}_m)_{m\in\mathbb{N}}$ is limit-conforming, then it is coercive by Lemma A.8, and (A.12) is satisfied with $\boldsymbol{H}_d = \boldsymbol{H}$ (this is (A.10)).
Conversely, assume that $(\mathcal{D}_m)_{m\in\mathbb{N}}$ is coercive and that (A.12) holds. Let $C_P \in \mathbb{R}_+$ be an upper bound of $(C_{\mathcal{D}_m})_{m\in\mathbb{N}}$. To prove (A.10), let $\boldsymbol{\varphi} \in \boldsymbol{H}$, $\varepsilon > 0$ and take $\boldsymbol{\psi} \in \boldsymbol{H}_d$ such that $\|\boldsymbol{\varphi} - \boldsymbol{\psi}\|_{\boldsymbol{H}} \le \varepsilon$. By definition of the norm in $\boldsymbol{H}$, this means that

$$\|\boldsymbol{\varphi} - \boldsymbol{\psi}\|_{\boldsymbol{L}'} + \|\mathrm{D}\boldsymbol{\varphi} - \mathrm{D}\boldsymbol{\psi}\|_{L'} \le \varepsilon.$$

Hence, for any $u \in X_{\mathcal{D}_m}\backslash\{0\}$,

$$\frac{|\langle \boldsymbol{\varphi}-\boldsymbol{\psi}, \mathrm{G}_{\mathcal{D}_m} u\rangle_{\boldsymbol{L}',\boldsymbol{L}} + \langle \mathrm{D}\boldsymbol{\varphi}-\mathrm{D}\boldsymbol{\psi}, \mathrm{P}_{\mathcal{D}_m} u\rangle_{L',L}|}{\|u\|_{\mathcal{D}_m}}$$
$$\leq \|\boldsymbol{\varphi}-\boldsymbol{\psi}\|_{\boldsymbol{L}'} \frac{\|\mathrm{G}_{\mathcal{D}_m} u\|_{\boldsymbol{L}}}{\|u\|_{\mathcal{D}_m}} + \|\mathrm{D}\boldsymbol{\varphi}-\mathrm{D}\boldsymbol{\psi}\|_{L'} \frac{\|\mathrm{P}_{\mathcal{D}_m} u\|_{L}}{\|u\|_{\mathcal{D}_m}} \leq \max(1, C_P)\varepsilon.$$

Introducing $\boldsymbol{\psi}$ and $\mathrm{D}\boldsymbol{\psi}$ in the definition (A.9) of $W_{\mathcal{D}_m}(\boldsymbol{\varphi})$, we infer

$$W_{\mathcal{D}_m}(\boldsymbol{\varphi}) \leq \sup_{u\in X_{\mathcal{D}_m}\setminus\{0\}} \frac{|\langle \boldsymbol{\psi}, \mathrm{G}_{\mathcal{D}_m} u\rangle_{\boldsymbol{L}',\boldsymbol{L}} + \langle \mathrm{D}\boldsymbol{\psi}, \mathrm{P}_{\mathcal{D}_m} u\rangle_{L',L}|}{\|u\|_{\mathcal{D}_m}} + \max(1, C_P)\varepsilon$$
$$= W_{\mathcal{D}_m}(\boldsymbol{\psi}) + \max(1, C_P)\varepsilon.$$

Invoking (A.12) we deduce that $\limsup_{m\to\infty} W_{\mathcal{D}_m}(\boldsymbol{\varphi}) \leq \max(1, C_P)\varepsilon$, and the proof is concluded by letting $\varepsilon \to 0$. ∎

The lemma of regularity of the limit (Lemma A.11 below) is an essential tool to use compactness techniques in the convergence analysis of numerical methods for non-linear models. We start by a preliminary result that facilitates the proof of the regularity of the limit.

Lemma A.10 (On limit-conformity, abstract setting). *Let $\mathcal{D}$ be a gradient discretisation in the sense of Definition A.4. Define $\widetilde{W}_{\mathcal{D}} : \boldsymbol{H} \times X_{\mathcal{D}} \to [0, +\infty)$ by*

$$\forall(\boldsymbol{\varphi}, u) \in \boldsymbol{H} \times X_{\mathcal{D}}\,,\ \widetilde{W}_{\mathcal{D}}(\boldsymbol{\varphi}, u) = \langle \boldsymbol{\varphi}, \mathrm{G}_{\mathcal{D}} u\rangle_{\boldsymbol{L}',\boldsymbol{L}} + \langle \mathrm{D}\boldsymbol{\varphi}, \mathrm{P}_{\mathcal{D}} u\rangle_{L',L}. \tag{A.13}$$

A sequence $(\mathcal{D}_m)_{m\in\mathbb{N}}$ of gradient discretisations is limit-conforming in the sense of Definition A.7 if and only if, for any sequence $u_m \in X_{\mathcal{D}_m}$ such that $(\|u_m\|_{\mathcal{D}_m})_{m\in\mathbb{N}}$ is bounded,

$$\forall \boldsymbol{\varphi} \in \boldsymbol{H},\ \lim_{m\to\infty} \widetilde{W}_{\mathcal{D}_m}(\boldsymbol{\varphi}, u_m) = 0. \tag{A.14}$$

Proof. Remark that

$$W_{\mathcal{D}}(\boldsymbol{\varphi}) = \sup_{u\in X_{\mathcal{D}}\setminus\{0\}} \frac{|\widetilde{W}_{\mathcal{D}}(\boldsymbol{\varphi}, u)|}{\|u\|_{\mathcal{D}}}.$$

The proof that (A.10) implies (A.14) is straightforward, since $|\widetilde{W}_{\mathcal{D}_m}(\boldsymbol{\varphi}, u_m)| \leq \|u_m\|_{\mathcal{D}_m} W_{\mathcal{D}_m}(\boldsymbol{\varphi})$. Let us prove the converse by way of contradiction. If (A.10) does not hold then there exists $\boldsymbol{\varphi} \in \boldsymbol{H}$, $\varepsilon > 0$ and a subsequence of $(\mathcal{D}_m)_{m\in\mathbb{N}}$, still denoted by $(\mathcal{D}_m)_{m\in\mathbb{N}}$, such that $W_{\mathcal{D}_m}(\boldsymbol{\varphi}) \geq \varepsilon$ for all $m \in \mathbb{N}$. We can then find $v_m \in X_{\mathcal{D}_m} \setminus \{0\}$ such that

$$|\widetilde{W}_{\mathcal{D}}(\boldsymbol{\varphi}, v_m)| \geq \frac{1}{2}\varepsilon \|v_m\|_{\mathcal{D}_m}.$$

Set $u_m = v_m/\|v_m\|_{\mathcal{D}_m}$. Then, for all $m \in \mathbb{N}$, $\|u_m\|_{\mathcal{D}_m} = 1$ and

$$\widetilde{W}_{\mathcal{D}}(\boldsymbol{\varphi}, u_m) = \frac{1}{\|v_m\|_{\mathcal{D}_m}} \widetilde{W}_{\mathcal{D}}(\boldsymbol{\varphi}, v_m) \geq \frac{1}{2}\varepsilon.$$

This leads to a contradiction with (A.14). ∎

Lemma A.11 (Regularity of the limit, abstract setting). *Let $(\mathcal{D}_m)_{m\in\mathbb{N}}$ be a limit-conforming sequence of gradient discretisations, in the sense of Definition A.7. For any $m \in \mathbb{N}$, take $u_m \in X_{\mathcal{D}_m}$ and assume that $(\|u_m\|_{\mathcal{D}_m})_{m\in\mathbb{N}}$ is bounded. Then there exists $u \in B$ such that, along a subsequence as $m \to \infty$, $\mathrm{P}_{\mathcal{D}_m} u_m$ converges weakly in L to u, and $\mathrm{G}_{\mathcal{D}_m} u_m$ converges weakly in $\boldsymbol{L}$ to $\mathrm{G}u$.*

Proof. By definition of $\|\cdot\|_{\mathcal{D}_m}$, $(\mathrm{G}_{\mathcal{D}_m} u_m)_{m\in\mathbb{N}}$ is bounded in $\boldsymbol{L}$. By Lemma A.8, $(\mathcal{D}_m)_{m\in\mathbb{N}}$ is coercive and therefore $(\mathrm{P}_{\mathcal{D}_m} u_m)_{m\in\mathbb{N}}$ is bounded in L. The reflexivity of L and $\boldsymbol{L}$ thus gives a subsequence of $(\mathcal{D}_m, u_m)_{m\in\mathbb{N}}$, denoted in the same way, and elements $u \in L$ and $\boldsymbol{u} \in \boldsymbol{L}$ such that $\mathrm{P}_{\mathcal{D}_m} u_m$ converges weakly in L to u and $\mathrm{G}_{\mathcal{D}_m} u_m$ converges weakly in $\boldsymbol{L}$ to $\boldsymbol{u}$. These weak convergences, the limit-conformity of $(\mathcal{D}_m)_{m\in\mathbb{N}}$ and the boundedness of $(\|u_m\|_{\mathcal{D}_m})_{m\in\mathbb{N}}$ enable us to identify the limit in (A.14) to see that

$$\forall \boldsymbol{\varphi} \in \boldsymbol{H}\,,\ \langle \boldsymbol{\varphi}, \boldsymbol{u}\rangle_{\boldsymbol{L}',\boldsymbol{L}} + \langle \mathrm{D}\boldsymbol{\varphi}, u\rangle_{L',L} = 0.$$

This relation simultaneously proves that $u \in B$ and that $\boldsymbol{u} = \mathrm{G}u$. ∎

We conclude this appendix by the notions of GD-consistency and compactness in the abstract setting. Note that, once L, $\boldsymbol{L}$, $\boldsymbol{H}$ and D are chosen, the definition A.7 of limit-conformity is constrained by the continuous duality formula (A.3); as a consequence of Lemma A.8, the definition of coercivity is also constrained by this formula. These two notions therefore naturally follow from the continuous abstract setting.
On the contrary, the definitions of GD-consistency and compactness are disconnected from the duality formula. In the absence of a specific problem to analyse in the abstract setting, these definitions therefore remain rather open. Particular choices for these notions are presented here, but variants are possible – see Remark 3.12 for GD-consistency and Remark 3.16 for compactness.

Definition A.12 (GD-consistency, abstract setting)

If $\mathcal{D}$ is a gradient discretisation in the sense of Definition A.4, let $S_{\mathcal{D}} : B \to [0, +\infty)$ be given by

$$\forall \varphi \in B\,, \quad S_{\mathcal{D}}(\varphi) = \min_{v\in X_{\mathcal{D}}} \Big(\|\mathrm{P}_{\mathcal{D}} v - \varphi\|_L + \|\mathrm{G}_{\mathcal{D}} v - \mathrm{G}\varphi\|_{\boldsymbol{L}} \Big). \quad \text{(A.15)}$$

A sequence $(\mathcal{D}_m)_{m\in\mathbb{N}}$ of gradient discretisations is **GD-consistent**, or consistent for short, if

$$\forall \varphi \in B\,,\ \lim_{m\to\infty} S_{\mathcal{D}_m}(\varphi) = 0. \tag{A.16}$$

Lemma A.13 (Equivalent condition for GD-consistency, abstract setting). *A sequence $(\mathcal{D}_m)_{m\in\mathbb{N}}$ of gradient discretisations is GD-consistent in the sense of Definition A.12 if and only if there exists a dense subset B_d of B such that*

$$\forall \psi \in B_d\,,\ \lim_{m\to\infty} S_{\mathcal{D}_m}(\psi) = 0. \tag{A.17}$$

Proof. Let us assume that (A.17) holds and let us prove (A.16) (the converse is straightforward, take $B_d = B$). Observe first that, since B is continuously embedded in L, there exists $C_B > 0$ such that

$$\forall \varphi \in B\,,\ \|\varphi\|_L \le C_B \|\varphi\|_B\,.$$

Let $\varphi \in B$. Take $\varepsilon > 0$ and $\psi \in B_d$ such that $\|\varphi - \psi\|_B \le \varepsilon$. For $v \in X_{\mathcal{D}_m}$, the triangle inequality and the definition of the norm in B yield

$$\begin{aligned}
&\|\mathrm{P}_{\mathcal{D}_m} v - \varphi\|_L + \|\mathrm{G}_{\mathcal{D}_m} v - \mathrm{G}\varphi\|_{\boldsymbol{L}} \\
&\quad \le \|\mathrm{P}_{\mathcal{D}_m} v - \psi\|_L + \|\psi - \varphi\|_L + \|\mathrm{G}_{\mathcal{D}_m} v - \mathrm{G}\psi\|_{\boldsymbol{L}} + \|\mathrm{G}\psi - \mathrm{G}\varphi\|_{\boldsymbol{L}} \\
&\quad \le \|\mathrm{P}_{\mathcal{D}_m} v - \psi\|_L + \|\mathrm{G}_{\mathcal{D}_m} v - \mathrm{G}\psi\|_{\boldsymbol{L}} + (C_B + 1)\|\psi - \varphi\|_B\,.
\end{aligned}$$

Taking the infimum over $v \in X_{\mathcal{D}_m}$ leads to $S_{\mathcal{D}_m}(\varphi) \le S_{\mathcal{D}_m}(\psi) + (C_B+1)\varepsilon$. Assumption (A.17) then yields $\limsup_{m\to\infty} S_{\mathcal{D}_m}(\varphi) \le (C_B+1)\varepsilon$, and letting $\varepsilon \to 0$ concludes the proof that $S_{\mathcal{D}_m}(\varphi) \to 0$ as $m \to \infty$. ∎

Definition A.14 (Compactness, abstract setting)

A sequence $(\mathcal{D}_m)_{m\in\mathbb{N}}$ of gradient discretisations in the sense of Definition A.4 is **compact** if, for any sequence $u_m \in X_{\mathcal{D}_m}$ such that $(\|u_m\|_{\mathcal{D}_m})_{m\in\mathbb{N}}$ is bounded, the sequence $(\mathrm{P}_{\mathcal{D}_m} u_m)_{m\in\mathbb{N}}$ is relatively compact in L.

Remark A.15. The compactness of $(\mathcal{D}_m)_{m\in\mathbb{N}}$ often follows from some compactness property of B – perhaps translated in a discrete setting. The typical example is the case of "conforming Galerkin" gradient discretisations, defined by $\mathcal{D}_m = (X_{\mathcal{D}_m}, \mathrm{P}_{\mathcal{D}_m} = \mathrm{Id}, \mathrm{G}_{\mathcal{D}_m} = \mathrm{G})$, where $X_{\mathcal{D}_m}$ is a finite dimensional subspace of B. Then, if B is compactly embedded in L, $(\mathcal{D}_m)_{m\in\mathbb{N}}$ is compact in the sense of Definition A.14.

Lemma A.16 (Compactness implies coercivity, abstract setting). *Let $(\mathcal{D}_m)_{m\in\mathbb{N}}$ be a sequence of gradient discretisations that is compact in the sense of Definition A.14. Then $(\mathcal{D}_m)_{m\in\mathbb{N}}$ is also coercive in the sense of Definition A.6.*

Proof. Assume that $(\mathcal{D}_m)_{m\in\mathbb{N}}$ is not coercive. Then there exists a subsequence of $(\mathcal{D}_m)_{m\in\mathbb{N}}$ (denoted in the same way) such that, for all $m \in \mathbb{N}$, we can find $v_m \in X_{\mathcal{D}_m} \setminus \{0\}$ satisfying

$$\lim_{m\to\infty} \frac{\|\mathrm{P}_{\mathcal{D}_m} v_m\|_L}{\|v_m\|_{\mathcal{D}_m}} = +\infty.$$

Setting $u_m = v_m / \|v_m\|_{\mathcal{D}_m}$, this gives $\lim_{m\to\infty} \|\mathrm{P}_{\mathcal{D}_m} u_m\|_L = +\infty$. But $\|u_m\|_{\mathcal{D}_m} = 1$ for all $m \in \mathbb{N}$ and the compactness of the sequence of gradient discretisations therefore implies that $(\mathrm{P}_{\mathcal{D}_m} u_m)_{m\in\mathbb{N}}$ is relatively compact in L, which is a contradiction. ∎

B

Discrete functional analysis

Because the GDM encompasses non-conforming schemes, the functional spaces where the approximate solutions live are not included in the classical Sobolev spaces. Therefore, the usual Poincaré inequalities, Sobolev embeddings or trace inequalities cannot be directly used. This chapter introduces a number of tools, referred to as "discrete functional analysis tools", which are the equivalent of the aforementioned inequalities/embeddings in discrete spaces (made of vectors gathering cell and face unknowns). These tools are combined, in Section 7.2, with the notion of polytopal toolbox to establish the coercivity, limit-conformity and compactness of sequences of gradient discretisations that are controlled by such toolboxes. As shown in Chapters 8–14, many conforming and non-conforming schemes can be analysed through controls by polytopal toolboxes and thus, indirectly, through the discrete functional analysis tools presented here.

In Section B.1, technical results on polytopal meshes and related reconstruction operators are presented. The three subsequent sections are devoted to discrete functional analysis results for diffusion problems with, respectively, Dirichlet, Neumann/Fourier, and mixed boundary conditions. Some of the tools developed here are inspired by previous works; this is in particular the case for discrete Poincaré and Sobolev inequalities, see, *e.g.*, [51, 91, 106, 96, 110, 111, 25, 70] to cite a few.

In this chapter, unless otherwise specified we take $p \in (1, \infty)$ and Ω is an open bounded connected subset of $\mathbb{R}^d$ ($d \in \mathbb{N}^\star$) with Lipschitz-continuous boundary $\partial\Omega$.

B.1 Preliminary results

We state here a few technical results on polytopal meshes and associated discrete elements.

B.1.1 Geometrical properties of cells

The lemmas in this section state simple geometrical properties and formulas associated with a cell.

Lemma B.1. *Let $\mathfrak{T}$ be a polytopal mesh in the sense of Definition 7.2. Take $K \in \mathcal{M}$ and let $\varrho_K = \min_{\sigma \in \mathcal{F}_K} d_{K,\sigma}$. Then, the open ball $B(\boldsymbol{x}_K, \varrho_K)$ of centre $\boldsymbol{x}_K$ and radius ϱ_K is contained in K, and K is star-shaped with respect to all points in this ball.*

Proof. For $\sigma \in \mathcal{F}_K$ we let H_σ be the affine hyperplane generated by σ and $H_\sigma^- = \{\boldsymbol{x} \in \mathbb{R}^d : (\boldsymbol{x} - \boldsymbol{z}) \cdot \boldsymbol{n}_{K,\sigma} < 0 \text{ for all } \boldsymbol{z} \in H_\sigma\}$ be the half space, opposite to $\boldsymbol{n}_{K,\sigma}$, corresponding to σ (see Figure B.1).

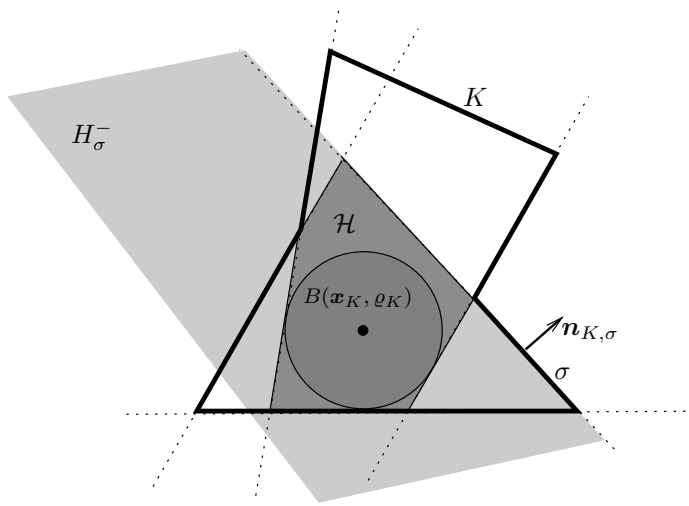


Fig. B.1. Illustration of the proof of Lemma B.1.

By definition, $d_{K,\sigma}$ is the (usual) distance from $\boldsymbol{x}_K$ to H_σ. Hence $B(\boldsymbol{x}_K, \varrho_K)$ is contained in H_σ^-; otherwise, we would have a point in this ball which is at a greater distance from $\boldsymbol{x}_K$ than $d_{K,\sigma}$, which contradicts $\varrho_K \leq d_{K,\sigma}$. Hence $B(\boldsymbol{x}_K, \varrho_K) \subset \cap_{\sigma \in \mathcal{F}_K} H_\sigma^- =: \mathcal{H}$. The proof is concluded if we show that K is star-shaped with respect to any point in $\mathcal{H}$.
Let $\boldsymbol{x} \in \mathcal{H}$ and $\boldsymbol{y} \in K$. If $[\boldsymbol{x}, \boldsymbol{y}]$ is not contained in K, then by convexity of $[\boldsymbol{x}, \boldsymbol{y}]$ we have $(\boldsymbol{x}, \boldsymbol{y}) \cap \partial K \neq \emptyset$. Let $\boldsymbol{z}$ be the last point, towards $\boldsymbol{y}$, in $(\boldsymbol{x}, \boldsymbol{y}) \cap \partial K$. Then $(\boldsymbol{z}, \boldsymbol{y}) \subset K$ and, if σ is the face of K on which $\boldsymbol{z}$ lies, $(\boldsymbol{z} - \boldsymbol{y}) \cdot \boldsymbol{n}_{K,\sigma} > 0$. But $\boldsymbol{x} - \boldsymbol{z} = \alpha(\boldsymbol{z} - \boldsymbol{y})$ for some positive α since $\boldsymbol{z}$ lies between $\boldsymbol{x}$ and $\boldsymbol{y}$, and thus $(\boldsymbol{x} - \boldsymbol{z}) \cdot \boldsymbol{n}_{K,\sigma} = \alpha[(\boldsymbol{z} - \boldsymbol{y}) \cdot \boldsymbol{n}_{K,\sigma}] > 0$. On the other hand, since $\boldsymbol{x} \in \mathcal{H} \subset H_\sigma^-$ and $z \in \sigma$, $(\boldsymbol{x} - \boldsymbol{z}) \cdot \boldsymbol{n}_{K,\sigma} < 0$. This is a contradiction and the proof is complete. ∎

Lemma B.2. *Let $\mathfrak{T}$ be a polytopal mesh in the sense of Definition 7.2, $K \in \mathcal{M}$ and $\sigma \in \mathcal{F}_K$. Then*

$$|D_{K,\sigma}| = \frac{1}{d}|\sigma| d_{K,\sigma} \quad \textit{and} \quad \sum_{\sigma \in \mathcal{F}_K} |\sigma| d_{K,\sigma} = d|K|. \tag{B.1}$$

Proof. We first compute $|D_{K,\sigma}| = \int_{D_{K,\sigma}} \mathrm{d}t\mathrm{d}\boldsymbol{x}$. Since the integral is invariant by translation and change of orthonormal axis system, there is no loss of generality in supposing that σ lies on the hyperplane $x^{(1)} = 0$, and that $\boldsymbol{x}_K$ on the line orthogonal to it. Then $\boldsymbol{x}_K = (d_{K,\sigma}, 0, \ldots, 0)$, see Figure B.2.

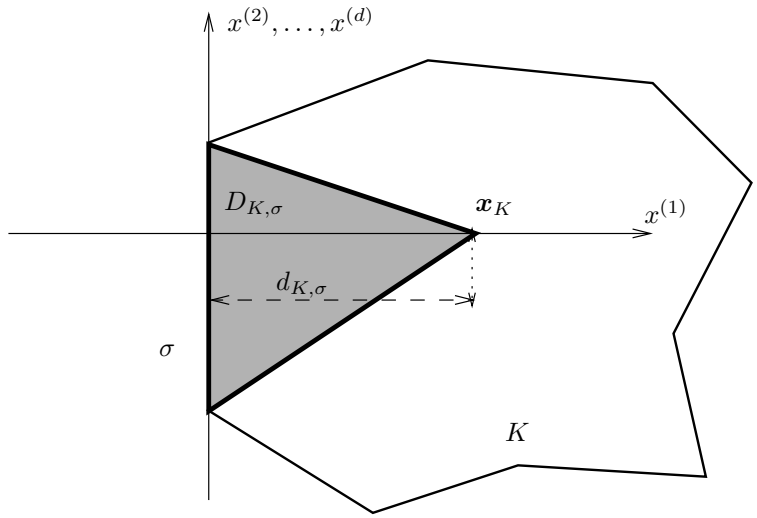


Fig. B.2. Illustration of the proof of Lemma B.2

Consider the change of variable $(t, \boldsymbol{y}) \in (0,1) \times \sigma \mapsto \boldsymbol{x} \in D_{K,\sigma}$ defined by $\boldsymbol{x} = (1-t)\boldsymbol{x}_K + t\boldsymbol{y} = ((1-t)d_{K,\sigma}, ty^{(2)}, \ldots, ty^{(d)})$ (note that $y^{(1)} = 0$). Its Jacobian determinant is $J(t, \boldsymbol{y}) = d_{K,\sigma} \times t^{d-1}$ so

$$|D_{K,\sigma}| = \int_0^1 \int_\sigma t^{d-1} d_{K,\sigma} \mathrm{d}t\mathrm{d}\gamma(\boldsymbol{y}) = \frac{1}{d} d_{K,\sigma}|\sigma|,$$

as announced in the lemma. The second equation in (B.1) follows immediately from the fact that $(D_{K,\sigma})_{\sigma \in \mathcal{F}_K}$ forms a partition of K (up to a set of zero measure). ∎

The following lemma and corollary are extremely useful to construct $\mathbb{P}_1$-exact gradient reconstructions.

Lemma B.3. *Let K be a polytopal subset of $\mathbb{R}^d$ with faces $\mathcal{F}_K$ and, for $\sigma \in \mathcal{F}_K$, denote by $\overline{\boldsymbol{x}}_\sigma$ the centre of mass of σ. Let $\boldsymbol{x}_K$ be any point of $\mathbb{R}^d$. Then,*

$$\sum_{\sigma\in\mathcal{F}_K} |\sigma|\boldsymbol{n}_{K,\sigma}(\overline{\boldsymbol{x}}_\sigma - \boldsymbol{x}_K)^T = |K|\mathrm{Id}, \tag{B.2}$$

where $(\overline{\boldsymbol{x}}_\sigma - \boldsymbol{x}_K)^T$ is the transpose of $\overline{\boldsymbol{x}}_\sigma - \boldsymbol{x}_K \in \mathbb{R}^d$, and Id *is the $d \times d$ identity matrix.*

Proof. Since $\overline{\boldsymbol{x}}_\sigma$ is the centre of mass of σ, for any $i = 1, \ldots, d$,

$$\overline{\boldsymbol{x}}_\sigma^{(i)} = \frac{1}{|\sigma|}\int_\sigma \boldsymbol{x}^{(i)} ds(\boldsymbol{x})$$

(where $\boldsymbol{x}^{(i)}$ denotes the i-th component of $\boldsymbol{x}$), and therefore

$$\sum_{\sigma\in\mathcal{F}_K} |\sigma|\overline{\boldsymbol{x}}_\sigma^{(i)}\boldsymbol{n}_{K,\sigma} = \sum_{\sigma\in\mathcal{F}_K}\int_\sigma \boldsymbol{x}^{(i)}\boldsymbol{n}_{K,\sigma} ds(\boldsymbol{x}).$$

The divergence (or Stokes') formula then gives

$$\sum_{\sigma\in\mathcal{F}_K} |\sigma|\overline{\boldsymbol{x}}_\sigma^{(i)}\boldsymbol{n}_{K,\sigma} = \int_K \nabla(\boldsymbol{x}^{(i)})\mathrm{d}\boldsymbol{x} = |K|\mathbf{e}_i$$

where $\mathbf{e}_i$ is the i-th vector of the canonical basis of $\mathbb{R}^d$. Since $\overline{\boldsymbol{x}}_\sigma^T\mathbf{e}_i = \overline{\boldsymbol{x}}_\sigma^{(i)}$, this shows that

$$\left(\sum_{\sigma\in\mathcal{F}_K} |\sigma|\boldsymbol{n}_{K,\sigma}\overline{\boldsymbol{x}}_\sigma^T\right)\mathbf{e}_i = (|K|\mathrm{Id})\mathbf{e}_i.$$

This relation being valid for any $i = 1, \ldots, d$, we infer that

$$\sum_{\sigma\in\mathcal{F}_K} |\sigma|\boldsymbol{n}_{K,\sigma}\overline{\boldsymbol{x}}_\sigma^T = |K|\mathrm{Id}. \tag{B.3}$$

Apply now divergence formula to a constant field $\boldsymbol{\xi} \in \mathbb{R}^d$:

$$\left(\sum_{\sigma\in\mathcal{F}_K} |\sigma|\boldsymbol{n}_{K,\sigma}\right)\cdot\boldsymbol{\xi} = \sum_{\sigma\in\mathcal{F}_K}\int_\sigma \boldsymbol{\xi}\cdot\boldsymbol{n}_{K,\sigma}\mathrm{d}\gamma(\boldsymbol{x}) = \int_K \mathrm{div}(\boldsymbol{\xi})\mathrm{d}\boldsymbol{x} = 0.$$

Since this relation is true for any $\boldsymbol{\xi} \in \mathbb{R}^d$, it shows that

$$\sum_{\sigma\in\mathcal{F}_K} |\sigma|\boldsymbol{n}_{K,\sigma} = 0. \tag{B.4}$$

(B.2) is proved by adding (B.3) and (B.4) multiplied on the right by $-\boldsymbol{x}_K^T$. ∎

For simplicial meshes, the next lemma shows that the regularity factor $\kappa_{\mathfrak{T}}$ defined by (7.10) controls all the other ones.

Lemma B.4. *Let K be a simplex of $\mathbb{R}^d$, $\overline{\boldsymbol{x}}_K$ be the centre of mass of K, and ρ_K be the maximum radius of the balls centred at $\overline{\boldsymbol{x}}_K$ and contained in K. For $\sigma \in \mathcal{F}_K$, let $d_{K,\sigma}$ be defined by (7.4) with $\boldsymbol{x}_K = \overline{\boldsymbol{x}}_K$. Then*

$$\rho_K = \min_{\sigma \in \mathcal{F}_K} d_{K,\sigma}\,, \tag{B.5}$$

$$\forall \boldsymbol{s}_0 \neq \boldsymbol{s}_1 \text{ in } \mathcal{V}_K\,,\ \rho_K \leq \frac{1}{d+1}\mathrm{dist}(\boldsymbol{s}_0, \boldsymbol{s}_1)\,, \tag{B.6}$$

$$\forall \sigma \in \mathcal{F}_K\,,\ \rho_K \leq \frac{1}{d+1}\mathrm{diam}(\sigma). \tag{B.7}$$

As a consequence, if $\mathfrak{T}$ is a conforming simplicial mesh with $\mathcal{P}$ the centres of mass of the cells, then, recalling the definitions (7.8)–(7.10),

$$\eta_{\mathfrak{T}} \leq \frac{2\kappa_{\mathfrak{T}}^2}{d+1} \quad \text{and} \quad \theta_{\mathfrak{T}} \leq \kappa_{\mathfrak{T}} + d + 1.$$

Proof. The inequality $\geq$ in (B.5) is a consequence of Lemma B.1. The other inequality actually only relies on the convexity of K. If $\sigma \in \mathcal{F}_K$, as in the proof of Lemma B.1 denote by H_σ the affine hyperplane containing σ, and by H_σ^- the half space $H_\sigma + \mathbb{R}^- \boldsymbol{n}_{K,\sigma}$. Since K is convex, $K \subset H_\sigma^-$ and $d_{K,\sigma}$ is the (positive) distance from $\overline{\boldsymbol{x}}_K$ to H_σ. We have $B(\overline{\boldsymbol{x}}_K, \rho_K) \subset K \subset H_\sigma^-$ and ρ_K must therefore be less than $\mathrm{dist}(\overline{\boldsymbol{x}}_K, H_\sigma) = d_{K,\sigma}$.
Let us now prove (B.6). Let σ be the face of K opposite to $\boldsymbol{s}_1$. Write $\overline{\boldsymbol{x}}_K = \frac{1}{d+1}\sum_{\boldsymbol{s}\in\mathcal{V}_K} \boldsymbol{s}$, so that

$$\begin{aligned}\boldsymbol{s}_0 - \overline{\boldsymbol{x}}_K &= \frac{1}{d+1}\sum_{\boldsymbol{s}\in\mathcal{V}_K}(\boldsymbol{s}_0 - \boldsymbol{s}) \\ &= \frac{1}{d+1}\sum_{\boldsymbol{s}\in\mathcal{V}_K,\,\boldsymbol{s}\neq\boldsymbol{s}_1}(\boldsymbol{s}_0 - \boldsymbol{s}) + \frac{1}{d+1}(\boldsymbol{s}_0 - \boldsymbol{s}_1).\end{aligned} \tag{B.8}$$

If $\boldsymbol{s} \neq \boldsymbol{s}_1$ then $\boldsymbol{s}, \boldsymbol{s}_0 \in \overline{\sigma}$ and thus $(\boldsymbol{s}_0 - \boldsymbol{s}) \cdot \boldsymbol{n}_{K,\sigma} = 0$. Taking the scalar product of (B.8) with $\boldsymbol{n}_{K,\sigma}$ therefore gives, since $\boldsymbol{s}_0 \in \overline{\sigma}$,

$$d_{K,\sigma} = (\boldsymbol{s}_0 - \overline{\boldsymbol{x}}_K) \cdot \boldsymbol{n}_{K,\sigma} = \frac{1}{d+1}(\boldsymbol{s}_0 - \boldsymbol{s}_1) \cdot \boldsymbol{n}_{K,\sigma} \leq \frac{1}{d+1}\mathrm{dist}(\boldsymbol{s}_0, \boldsymbol{s}_1).$$

Equation (B.6) follows since $\rho_K \leq d_{K,\sigma}$ by (B.5). Estimate (B.7) is a consequence of (B.6) since, for any face $\sigma \in \mathcal{F}_K$ and any two vertices $\boldsymbol{s}_0 \neq \boldsymbol{s}_1$ of σ, $\mathrm{dist}(\boldsymbol{s}_0, \boldsymbol{s}_1) \leq \mathrm{diam}(\sigma)$.

Let us turn to the upper bound on $\eta_{\mathfrak{T}}$. For any neighbouring cells K and L, denoting by σ their common face, by (B.5) applied to K and (B.7) applied to L,

$$\begin{aligned}d_{K,\sigma} \geq \rho_K \geq \kappa_{\mathfrak{T}}^{-1} h_K \geq \kappa_{\mathfrak{T}}^{-1}\mathrm{diam}(\sigma) &\geq \kappa_{\mathfrak{T}}^{-1}(d+1)\rho_L \\ &\geq \kappa_{\mathfrak{T}}^{-2}(d+1)h_L \geq \kappa_{\mathfrak{T}}^{-2}(d+1)d_{L,\sigma}.\end{aligned}$$

Hence $\frac{d_{L,\sigma}}{d_{K,\sigma}} \leq \frac{\kappa_{\mathfrak{T}}^2}{d+1}$ which gives, by reversing the roles of K and L, the upper bound on $\eta_{\mathfrak{T}}$.
The bound on $\theta_{\mathfrak{T}}$ is trivial since any simplex K has $d+1$ faces and, by (B.5),

$$\frac{h_K}{d_{K,\sigma}} \leq \kappa_{\mathfrak{T}} \frac{\rho_K}{d_{K,\sigma}} \leq \kappa_{\mathfrak{T}}. \tag{B.9}$$

■

Remark B.5 (Generalisation to $\boldsymbol{x}_K$ not located at the centre of mass)
The proof shows that (B.5) holds with $\overline{\boldsymbol{x}}_K$ replaced by any $\boldsymbol{x}_K \in K$. Writing $\boldsymbol{x}_K = \sum_{\boldsymbol{s}\in\mathcal{V}_K} \alpha_{\boldsymbol{s}} \boldsymbol{s}$ as a convex combination and reproducing the previous proof with these coefficients $\alpha_{\boldsymbol{s}} \in [0,1]$ instead of $1/(d+1)$, we see that (B.6) and (B.7) holds with 1 instead of $1/(d+1)$.

B.1.2 Interpolant on $X_{\mathfrak{T}}$

For $\mathfrak{T}$ a polytopal mesh of Ω in the sense of Definition 7.2 and $p \in [1,\infty)$, define the interpolant $I_{\mathfrak{T}} : W^{1,p}(\Omega) \to X_{\mathfrak{T}}$ by

$$\begin{aligned} &\forall \varphi \in W^{1,p}(\Omega)\,,\ I_{\mathfrak{T}}\varphi = ((\varphi_K)_{K\in\mathcal{M}}, (\varphi_\sigma)_{\sigma\in\mathcal{F}}) \text{ with} \\ &\quad \forall K \in \mathcal{M}\,,\ \varphi_K = \frac{1}{|K|}\int_K \varphi(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\,, \\ &\quad \forall \sigma \in \mathcal{F}\,,\ \varphi_\sigma = \frac{1}{|\sigma|}\int_\sigma \varphi(\boldsymbol{x})\mathrm{d}\gamma(\boldsymbol{x}). \end{aligned} \tag{B.10}$$

This interpolant enjoys essential stability and approximation properties. Before establishing them, let us start with a preliminary lemma.

Lemma B.6. *Let $\mathfrak{T}$ be a polytopal mesh of Ω in the sense of Definition 7.2, $p \in [1,\infty)$ and θ be such that*

$$\max\left\{\frac{h_K}{d_{K,\sigma}}\ :\ K \in \mathcal{M}\,,\ \sigma \in \mathcal{F}_K\right\} \leq \theta.$$

Then there exists C_1, depending only on d, p and θ, such that, for any $K \in \mathcal{M}$ and any $\varphi \in W^{1,p}(K)$, with the notations in (B.10),

$$|\varphi_\sigma - \varphi_K|^p \leq \frac{C_1 h_K^{p-1}}{|\sigma|}\int_K |\nabla\varphi(\boldsymbol{x})|^p \mathrm{d}\boldsymbol{x} \tag{B.11}$$

and

$$\|\varphi - \varphi_K\|_{L^p(K)} \leq C_1 h_K \|\nabla\varphi\|_{L^p(K)^d}\,. \tag{B.12}$$

Proof. Let us assume the existence of C_2 depending only on d, p and θ such that, for all $K \in \mathcal{M}$ and all $\sigma \in \mathcal{F}_K$, setting $B_K = B(\boldsymbol{x}_K, \theta^{-1}h_K/2)$,

$$\left|\varphi_\sigma - \frac{1}{|B_K|}\int_{B_K}\varphi(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\right|^p \le C_2\frac{h_K^{p-1}}{|\sigma|}\int_K |\nabla\varphi(\boldsymbol{x})|^p\mathrm{d}\boldsymbol{x}, \tag{B.13}$$

$$\left|\frac{1}{|B_K|}\int_{B_K}\varphi(\boldsymbol{x})\mathrm{d}\boldsymbol{x} - \varphi_K\right|^p \le C_2\frac{h_K^{p}}{|K|}\int_K |\nabla\varphi(\boldsymbol{x})|^p\mathrm{d}\boldsymbol{x}, \tag{B.14}$$

and

$$\left\|\varphi - \frac{1}{|B_K|}\int_{B_K}\varphi(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\right\|_{L^p(K)} \le C_2 h_K \left\|\nabla\varphi\right\|_{L^p(K)^d}. \tag{B.15}$$

Then (B.11) follows from (B.13) and (B.14) by using the triangle inequality, the power-of-sums inequality (D.12), and, in (B.14), the estimate $|K| \ge |D_{K,\sigma}| = \frac{|\sigma|d_{K,\sigma}}{d} \ge \theta^{-1}d^{-1}|\sigma|h_K$. Similarly, Estimate (B.12) follows from (B.14), (B.15) and the triangle inequality.
To prove the existence of C_2 such that (B.13)–(B.15) hold, notice first that, since the restrictions to K of functions in $C^\infty(\mathbb{R}^d)$ are dense in $W^{1,p}(K)$ (K is a polytopal set), these estimates only need to be established for $\varphi \in C^\infty(\mathbb{R}^d)$.

Proof of (B.13)
For $\boldsymbol{z} \in B_K$ and $\boldsymbol{y} \in \sigma$, write $\varphi(\boldsymbol{y}) - \varphi(\boldsymbol{z}) = \int_0^1 \nabla\varphi(\boldsymbol{z} + t(\boldsymbol{y}-\boldsymbol{z}))\cdot(\boldsymbol{z}-\boldsymbol{y})\mathrm{d}t$. Taking the mean value for $\boldsymbol{z} \in B_K$ and $\boldsymbol{y} \in \sigma$ and using Jensen's inequality yields

$$L_{(\text{B.13})} \le \frac{h_K^p}{|\sigma|\,|B_K|}\int_0^1\int_\sigma\int_{B_K}|\nabla\varphi(\boldsymbol{z}+t(\boldsymbol{y}-\boldsymbol{z}))|^p\mathrm{d}\boldsymbol{z}\mathrm{d}\gamma(\boldsymbol{y})\mathrm{d}t, \tag{B.16}$$

where $L_{(\text{B.13})}$ is the left-hand side of (B.13). Since $\theta^{-1}h_K/2 \le d_{K,\sigma}$ for all $\sigma \in \mathcal{F}_K$, by Lemma B.1 the cell K is star-shaped with respect to all points in B_K. Hence, for all $\boldsymbol{z} \in B_K$ the change of variable $\psi : (t, \boldsymbol{y}) \in (0,1)\times\sigma \to \boldsymbol{x} = \boldsymbol{z} + t(\boldsymbol{y}-\boldsymbol{z})$ has values in K. By the same reasoning as in the proof of Lemma B.2, the Jacobian determinant of this change of variable is $J\psi = t^{d-1}|(\boldsymbol{y}-\boldsymbol{z})\cdot\boldsymbol{n}_{K,\sigma}|$. Since $|\boldsymbol{x}-\boldsymbol{z}| = t|\boldsymbol{y}-\boldsymbol{z}| \le th_K$, we have $t \ge \frac{|\boldsymbol{x}-\boldsymbol{z}|}{h_K}$. Moreover,

$$|(\boldsymbol{y}-\boldsymbol{z})\cdot\boldsymbol{n}_{K,\sigma}| \ge |(\boldsymbol{y}-\boldsymbol{x}_K)\cdot\boldsymbol{n}_{K,\sigma}| - |\boldsymbol{z}-\boldsymbol{x}_K| \ge d_{K,\sigma} - \frac{\theta^{-1}h_K}{2} \ge \frac{\theta^{-1}h_K}{2}.$$

Hence,

$$J\psi \ge \left(\frac{|\boldsymbol{x}-\boldsymbol{z}|}{h_K}\right)^{d-1}\frac{\theta^{-1}}{2}h_K \ge (2\theta)^{-1}h_K^{2-d}|\boldsymbol{x}-\boldsymbol{z}|^{d-1}.$$

Using ψ in (B.16) therefore leads to

$$L_{(\text{B.13})} \le \frac{2\theta h_K^{p+d-2}}{|\sigma|\,|B_K|}\int_K|\nabla\varphi(\boldsymbol{x})|^p\int_{B_K}|\boldsymbol{x}-\boldsymbol{z}|^{1-d}\mathrm{d}\boldsymbol{z}\mathrm{d}\boldsymbol{x}. \tag{B.17}$$

Since $B_K \subset K \subset B(\boldsymbol{x}, h_K)$ for any $\boldsymbol{x} \in K$, denoting by ω_d the surface of the unit sphere in $\mathbb{R}^d$,

$$\int_{B_K} |\boldsymbol{x}-\boldsymbol{z}|^{1-d}\mathrm{d}\boldsymbol{z} \le \int_{B(\boldsymbol{x},h_K)} |\boldsymbol{x}-\boldsymbol{z}|^{1-d}\mathrm{d}\boldsymbol{z} = \omega_d \int_0^{h_K} \rho^{1-d}\rho^{d-1}\mathrm{d}\rho = \omega_d h_K.$$

Plugged into (B.17), this estimate gives (B.13) since $|B_K| = |B(0,1)|(2\theta)^{-d}h_K^d$.

Proof of (B.14)
We follow similar ideas as in the proof of Lemma 7.59. For all $(\boldsymbol{x},\boldsymbol{y}) \in B_K \times K$, we have

$$\varphi(\boldsymbol{x}) - \varphi(\boldsymbol{y}) = \int_0^1 \nabla\varphi(t\boldsymbol{x} + (1-t)\boldsymbol{y}) \cdot (\boldsymbol{x}-\boldsymbol{y})\mathrm{d}t. \tag{B.18}$$

Taking the mean values for $\boldsymbol{x} \in B_K$ and $\boldsymbol{y} \in K$ and denoting by $L_{(\text{B.14})}$ the left-hand side of (B.14), Jensen's inequality gives

$$L_{(\text{B.14})} \le \frac{h_K^p}{|B_K|\,|K|} \int_{B_K}\int_K\int_0^1 |\nabla\varphi(t\boldsymbol{x} + (1-t)\boldsymbol{y})|^p \mathrm{d}t\mathrm{d}\boldsymbol{y}\mathrm{d}\boldsymbol{x}. \tag{B.19}$$

Applying the change of variable $\boldsymbol{x} \in B_K \to \boldsymbol{z} = t\boldsymbol{x} + (1-t)\boldsymbol{y}$, which has values in K since K is star-shaped with respect to all points in B_K, we have

$$\begin{aligned}\int_{B_K}\int_K\int_0^1 &|\nabla\varphi(t\boldsymbol{x} + (1-t)\boldsymbol{y})|^p \mathrm{d}t\mathrm{d}\boldsymbol{y}\mathrm{d}\boldsymbol{x} \\ &\le \int_K |\nabla\varphi(\boldsymbol{z})|^p \int_K \int_{I(\boldsymbol{z},\boldsymbol{y})} t^{-d}\mathrm{d}t\mathrm{d}\boldsymbol{y}\mathrm{d}\boldsymbol{z}\end{aligned} \tag{B.20}$$

where, as in the proof of Lemma 7.59 with $V = B_K$, $I(\boldsymbol{z},\boldsymbol{y}) = \{t \in (0,1) : \exists \boldsymbol{x} \in B_K\,,\ t\boldsymbol{x} + (1-t)\boldsymbol{y} = \boldsymbol{z}\}$. Using $B_K \subset K$ and following estimates (7.73) and (7.74), we arrive at

$$\int_K \int_{I(\boldsymbol{z},\boldsymbol{y})} t^{-d}\mathrm{d}t\mathrm{d}\boldsymbol{y} \le \frac{h_K^d}{d-1}\omega_d. \tag{B.21}$$

Substituting this inequality into (B.20) and coming back to (B.19) completes the proof of (B.14), since $|B_K| = |B(0,1)|(2\theta)^{-d}h_K^d$.

Proof of (B.15)
This estimate follows immediately from Lemma 7.59 since $V = K$ is star-shaped with respect to $B = B_K$, and $\mathrm{diam}(V) = h_K = \theta\mathrm{diam}(B)$. ∎

The following stability property of $I_{\mathfrak{T}}$ is useful to check the condition (2.96) in Lemma 2.52. It enabled us (in Section 9.5 and in the proof of Theorem 13.14) to establish the GD-consistency of non-conforming $\mathbb{P}_1$ GDs and of HMM GDs in the case of non-homogeneous Dirichlet boundary conditions.

Proposition B.7 (Stability of the interpolant $I_{\mathfrak{T}}$). *Let $\mathfrak{T}$ be a polytopal mesh of Ω in the sense of Definition 7.2, let $I_{\mathfrak{T}}$ be defined by (B.10), and let $\theta \geq \theta_{\mathfrak{T}}$ (see (7.8)). Then, there exists C_3 depending only on d, p and θ such that, for all $\varphi \in W^{1,p}(\Omega)$,*

$$\begin{aligned} &\|\Pi_{\mathfrak{T}}(I_{\mathfrak{T}}\varphi)\|_{L^p(\Omega)} \leq \|\varphi\|_{L^p(\Omega)}\,, \quad \|\mathbb{T}_{\mathfrak{T}}(I_{\mathfrak{T}}\varphi)\|_{L^p(\partial\Omega)} \leq \|\gamma\varphi\|_{L^p(\partial\Omega)} \\ &\textit{and} \quad |I_{\mathfrak{T}}\varphi|_{\mathfrak{T},p} \leq C_3 \|\nabla\varphi\|_{L^p(\Omega)^d}\,. \end{aligned} \tag{B.22}$$

Proof. Using the notations in (B.10), by Jensen's inequality,

$$|\varphi_K|^p \leq \frac{1}{|K|}\int_K |\varphi(\boldsymbol{x})|^p \mathrm{d}\boldsymbol{x}.$$

Multiplying this inequality by $|K|$ and summing over $K \in \mathcal{M}$ gives

$$\|\Pi_{\mathfrak{T}}(I_{\mathfrak{T}}\varphi)\|_{L^p(\Omega)} \leq \|\varphi\|_{L^p(\Omega)}\,.$$

A similar reasoning gives the estimate on $\mathbb{T}_{\mathfrak{T}}(I_{\mathfrak{T}}\varphi)$. To estimate $|I_{\mathfrak{T}}\varphi|_{\mathfrak{T},p}$, we apply (B.11) in Lemma B.6 to find C_4 depending only on d, p and θ such that

$$|\varphi_\sigma - \varphi_K|^p \leq \frac{C_4 h_K^{p-1}}{|\sigma|}\int_K |\nabla\varphi(\boldsymbol{x})|^p \mathrm{d}\boldsymbol{x}.$$

Multiply this inequality by $|\sigma| d_{K,\sigma}^{1-p}$ and sum over $\sigma \in \mathcal{F}_K$ and $K \in \mathcal{M}$. Since, for all $K \in \mathcal{M}$, $\mathrm{Card}(\mathcal{F}_K) \leq \theta$ and $h_K/d_{K,\sigma} \leq \theta$ for all $\sigma \in \mathcal{F}_K$, this yields $|I_{\mathfrak{T}}\varphi|_{\mathfrak{T},p}^p \leq C_4\theta^p \|\nabla\varphi\|_{L^p(\Omega)^d}^p$. ∎

To prove the approximation properties of $I_{\mathfrak{T}}$ (Proposition B.9 below), let us first state a preliminary lemma. We establish this lemma in the context of partitions of Ω, but it actually holds in a more general setting of measurable spaces; in particular, it is also valid if we replace Ω with $\partial\Omega$ (endowed with its $(d-1)$-dimensional measure).

Lemma B.8 (Approximation properties of projections on partitions). *Let $(\mathcal{M}_m)_{m\in\mathbb{N}}$ be a sequence of families of measurable subsets of Ω such that, for each $m \in \mathbb{N}$,*

- *$\overline{\Omega} \setminus \left(\bigcup_{K\in\mathcal{M}_m} \overline{K}\right)$ has a zero measure,*
- *each $K \in \mathcal{M}_m$ has a non-zero measure,*
- *if K and L are two distinct elements of $\mathcal{M}_m$, then $K \cap L = \emptyset$.*

For $m \in \mathbb{N}$, define $\mathrm{Pr}_{\mathcal{M}_m} : L^p(\Omega) \to L^p(\Omega)$ as the projection on piecewise constant functions on $\mathcal{M}_m$, that is,

$$\forall\varphi \in L^p(\Omega)\,,\ \forall K \in \mathcal{M}_m\,,\ (\mathrm{Pr}_{\mathcal{M}_m}\varphi)_{|K} = \frac{1}{|K|}\int_K \varphi(\boldsymbol{x})\mathrm{d}\boldsymbol{x}. \tag{B.23}$$

Assume that $\max_{K\in\mathcal{M}_m} \mathrm{diam}(K) \to 0$ as $m \to \infty$. Then, for all $\varphi \in L^p(\Omega)$, $\mathrm{Pr}_{\mathcal{M}_m}\varphi \to \varphi$ in $L^p(\Omega)$ as $m \to \infty$.

Proof. Take $\varepsilon > 0$ and let $\varphi_\varepsilon \in C^1(\mathbb{R}^d)$ be such that $\|\varphi - \varphi_\varepsilon\|_{L^p(\Omega)} \leq \varepsilon$. A triangle inequality yields

$$\begin{aligned}\|\mathrm{Pr}_{\mathcal{M}_m}\varphi - \varphi\|_{L^p(\Omega)} \leq\ & \|\mathrm{Pr}_{\mathcal{M}_m}(\varphi - \varphi_\varepsilon)\|_{L^p(\Omega)} + \|\mathrm{Pr}_{\mathcal{M}_m}\varphi_\varepsilon - \varphi_\varepsilon\|_{L^p(\Omega)} \\ & + \|\varphi_\varepsilon - \varphi\|_{L^p(\Omega)}\,.\end{aligned} \tag{B.24}$$

By Jensen's inequality, for all $\psi \in L^p(\Omega)$ and $K \in \mathcal{M}_m$,

$$|(\mathrm{Pr}_{\mathcal{M}_m}\psi)_{|K}|^p \leq \frac{1}{|K|}\int_K |\psi(\boldsymbol{x})|^p \mathrm{d}\boldsymbol{x}.$$

Multiply this by $|K|$ and sum over $K \in \mathcal{M}_m$ to obtain $\|\mathrm{Pr}_{\mathcal{M}_m}\psi\|_{L^p(\Omega)} \leq \|\psi\|_{L^p(\Omega)}$. Using this estimate in (B.24) with $\psi = \varphi - \varphi_\varepsilon$ and recalling that $\|\varphi - \varphi_\varepsilon\|_{L^p(\Omega)} \leq \varepsilon$ leads to

$$\|\mathrm{Pr}_{\mathcal{M}_m}\varphi - \varphi\|_{L^p(\Omega)} \leq 2\varepsilon + \|\mathrm{Pr}_{\mathcal{M}_m}\varphi_\varepsilon - \varphi_\varepsilon\|_{L^p(\Omega)}\,. \tag{B.25}$$

Then, for all $K \in \mathcal{M}_m$ and $\boldsymbol{x} \in K$,

$$\begin{aligned}|\mathrm{Pr}_{\mathcal{M}_m}\varphi_\varepsilon(\boldsymbol{x}) - \varphi_\varepsilon(\boldsymbol{x})| &= \left|\frac{1}{|K|}\int_K (\varphi_\varepsilon(\boldsymbol{y}) - \varphi_\varepsilon(\boldsymbol{x}))\mathrm{d}y\right| \\ &\leq \mathrm{diam}(K)\,\|\nabla\varphi_\varepsilon\|_{L^\infty(\mathbb{R}^d)^d}\,.\end{aligned}$$

Using Hölder's inequality and taking the supremum of the above inequality over $\boldsymbol{x} \in K$ and $K \in \mathcal{M}_m$, we obtain

$$\begin{aligned}\|\mathrm{Pr}_{\mathcal{M}_m}\varphi_\varepsilon - \varphi_\varepsilon\|_{L^p(\Omega)} &\leq |\Omega|^{1/p}\,\|\mathrm{Pr}_{\mathcal{M}_m}\varphi_\varepsilon - \varphi_\varepsilon\|_{L^\infty(\Omega)} \\ &\leq |\Omega|^{1/p}\left(\max_{K\in\mathcal{M}_m}\mathrm{diam}(K)\right)\|\nabla\varphi_\varepsilon\|_{L^\infty(\mathbb{R}^d)^d}\,.\end{aligned}$$

Plugged into (B.25), this gives

$$\|\mathrm{Pr}_{\mathcal{M}_m}\varphi - \varphi\|_{L^p(\Omega)} \leq 2\varepsilon + |\Omega|^{1/p}\left(\max_{K\in\mathcal{M}_m}\mathrm{diam}(K)\right)\|\nabla\varphi_\varepsilon\|_{L^\infty(\mathbb{R}^d)^d}\,.$$

Take now the superior limit as $m \to \infty$ and use $\max_{K\in\mathcal{M}_m}\mathrm{diam}(K) \to 0$ to get $\limsup_{m\to\infty}\|\mathrm{Pr}_{\mathcal{M}_m}\varphi - \varphi\|_{L^p(\Omega)} \leq 2\varepsilon$. Letting $\varepsilon \to 0$ concludes the proof that $\mathrm{Pr}_{\mathcal{M}_m}\varphi \to \varphi$ in $L^p(\Omega)$ as $m \to \infty$. ∎

We can now state the approximation properties of $I_{\mathfrak{T}}$.

Proposition B.9 (Approximation properties of the interpolant $I_{\mathfrak{T}}$). *Let $(\mathfrak{T}_m)_{m\in\mathbb{N}}$ be a sequence of polytopal meshes of Ω in the sense of Definition 7.2, such that $h_{\mathcal{M}_m} \to 0$ as $m \to \infty$. Then, for all $\varphi \in W^{1,p}(\Omega)$, as $m \to \infty$,*

$$\Pi_{\mathfrak{T}_m}(I_{\mathfrak{T}_m}\varphi) \to \varphi \text{ in } L^p(\Omega), \tag{B.26}$$

$$\mathbb{T}_{\mathfrak{T}_m}(I_{\mathfrak{T}_m}\varphi) \to \gamma\varphi \text{ in } L^p(\partial\Omega), \tag{B.27}$$

$$\overline{\nabla}_{\mathfrak{T}_m}(I_{\mathfrak{T}_m}\varphi) \to \nabla\varphi \text{ in } L^p(\Omega)^d. \tag{B.28}$$

Proof. Let $\mathrm{Pr}_{\mathcal{M}_m}$ be the projection on $\mathcal{M}_m$ as defined by (B.23). By definitions (B.10) and (7.7c) of $I_{\mathfrak{T}_m}$ and $\Pi_{\mathfrak{T}_m}$, $\Pi_{\mathfrak{T}_m}(I_{\mathfrak{T}_m}\varphi) = \mathrm{Pr}_{\mathcal{M}_m}\varphi$ and the convergence (B.26) follows from Lemma B.8. The convergence (B.27) of the reconstructed traces follows by the same argument, using a variant of Lemma B.8 for partitions of $\partial\Omega$ instead of Ω.
By Stokes' formula, the definition (B.10) of $I_{\mathfrak{T}_m}$, and the definition (7.7e) of $\overline{\nabla}_{\mathfrak{T}_m}$, for all $K \in \mathcal{M}_m$,

$$(\overline{\nabla}_{\mathfrak{T}_m}(I_{\mathfrak{T}_m}\varphi))_{|K} = \frac{1}{|K|}\sum_{\sigma\in\mathcal{F}_K}\int_\sigma \varphi(\boldsymbol{x})\mathrm{d}\gamma(\boldsymbol{x})\boldsymbol{n}_{K,\sigma} = \frac{1}{|K|}\int_K \nabla\varphi(\boldsymbol{x})\mathrm{d}\boldsymbol{x}.$$

Hence, $\overline{\nabla}_{\mathfrak{T}_m}(I_{\mathfrak{T}_m}\varphi) = \mathrm{Pr}_{\mathcal{M}_m}(\nabla\varphi)$, where $\mathrm{Pr}_{\mathcal{M}_m}$ acts on $\nabla\varphi$ component-by-component. Hence, Lemma B.8 shows that $\overline{\nabla}_{\mathfrak{T}_m}(I_{\mathfrak{T}_m}\varphi) \to \nabla\varphi$ in $L^p(\Omega)^d$ as $m \to \infty$, and the proof is complete. ∎

B.1.3 Approximation properties of $\overline{\boldsymbol{\nabla}}_{\mathfrak{T}}$

The following result is the key to proving that several classical gradient discretisations are LLE GDs.

Lemma B.10 ($\mathbb{P}_1$-exactness of $\overline{\nabla}_{\mathfrak{T}}$ and stability). *Under Hypothesis (7.2), let $p \in [1, +\infty)$ and $\mathfrak{T}$ be a polytopal mesh of Ω in the sense of Definition 7.2. Define $X_{\mathfrak{T}}$, $\overline{\nabla}_{\mathfrak{T}}$, $\overline{\nabla}_K$ and $|\cdot|_{\mathfrak{T},p}$ as in (7.7). Then*

1. *$\overline{\nabla}_K$ is a $\mathbb{P}_1$-exact gradient reconstruction on K upon $(\boldsymbol{x}_K, (\overline{\boldsymbol{x}}_\sigma)_{\sigma\in\mathcal{F}_K})$, in the sense of Definition 7.28. In other words, if A is an affine function and $u = (A(\boldsymbol{x}_K), (A(\overline{\boldsymbol{x}}_\sigma))_{\sigma\in\mathcal{F}_K})$ are the values at $\boldsymbol{x}_K$ and $(\overline{\boldsymbol{x}}_\sigma)_{\sigma\in\mathcal{F}_K}$ of A, then $\overline{\nabla}_K u = \nabla A$.*
2. *For all $v \in X_{\mathfrak{T}}$,*
$$\left\|\overline{\nabla}_{\mathfrak{T}} v\right\|_{L^p(\Omega)^d} \le d^{\frac{p-1}{p}}\,|v|_{\mathfrak{T},p}\,. \tag{B.29}$$

Proof. The proof of Item 1 follows by multiplying both sides of (B.2) by the constant vector ∇A, and by noticing that, since A is affine,

$$(\overline{\boldsymbol{x}}_\sigma - \boldsymbol{x}_K)^T\nabla A = (\overline{\boldsymbol{x}}_\sigma - \boldsymbol{x}_K)\cdot\nabla A = A(\overline{\boldsymbol{x}}_\sigma) - A(\boldsymbol{x}_K) = u_\sigma - u_K.$$

To prove Item 2, write, for $\boldsymbol{x} \in K$,

$$|\overline{\nabla}_{\mathfrak{T}} v(\boldsymbol{x})| \le \frac{1}{|K|}\sum_{\sigma\in\mathcal{F}_K}|\sigma|\,|v_\sigma - v_K| \le d\sum_{\sigma\in\mathcal{F}_K}\frac{|\sigma|d_{K,\sigma}}{d|K|}\left|\frac{v_\sigma - v_K}{d_{K,\sigma}}\right|.$$

By (B.1) we have $\sum_{\sigma\in\mathcal{F}_K}\frac{|\sigma|d_{K,\sigma}}{d|K|} = 1$ and the convexity of $s \mapsto s^p$ for $s \ge 0$ therefore gives

$$|\overline{\nabla}_{\mathfrak{T}} v(\boldsymbol{x})|^p \le d^p\sum_{\sigma\in\mathcal{F}_K}\frac{|\sigma|d_{K,\sigma}}{d|K|}\left|\frac{v_\sigma - v_K}{d_{K,\sigma}}\right|^p$$

$$= \frac{d^{p-1}}{|K|} \sum_{\sigma \in \mathcal{F}_K} |\sigma| d_{K,\sigma} \left| \frac{v_\sigma - v_K}{d_{K,\sigma}} \right|^p . \tag{B.30}$$

Integrate this estimate over $\boldsymbol{x} \in K$, sum over $K \in \mathcal{M}$ and recall the definition (7.7f) of $|\cdot|_{\mathfrak{T},p}$ to obtain (B.29). ∎

The following lemma helps proving the limit-conformity of a GD controlled by a polytopal toolbox.

Lemma B.11 (Discrete Stokes' formula). *Let $\mathfrak{T}$ be a polytopal mesh of Ω in the sense of Definition 7.2, $p \in [1, +\infty)$ and $\theta \geq \theta_{\mathfrak{T}}$ (see (7.8)). We define $X_{\mathfrak{T}}$, $\Pi_{\mathfrak{T}}$, $\mathbb{T}_{\mathfrak{T}}$, $\overline{\nabla}_{\mathfrak{T}}$ and $|\cdot|_{\mathfrak{T},p}$ as in (7.7). Then, there exists C_5 depending only on d, p and θ such that, for all $\boldsymbol{\varphi} \in W^{1,p'}(\Omega)^d$ and all $v \in X_{\mathfrak{T}}$,*

$$\Big| \int_\Omega \big(\overline{\nabla}_{\mathfrak{T}} v(\boldsymbol{x}) \cdot \boldsymbol{\varphi}(\boldsymbol{x}) + \Pi_{\mathfrak{T}} v(\boldsymbol{x}) \mathrm{div} \boldsymbol{\varphi}(\boldsymbol{x}) \big) \, \mathrm{d}\boldsymbol{x} \\ - \int_{\partial\Omega} \mathbb{T}_{\mathfrak{T}} v(\boldsymbol{x}) \gamma_{\mathbf{n}}(\boldsymbol{\varphi})(\boldsymbol{x}) \mathrm{d}\gamma(\boldsymbol{x}) \Big| \leq C_5 \, \| \, |\nabla \boldsymbol{\varphi}| \, \|_{L^{p'}(\Omega)} \, |v|_{\mathfrak{T},p} \, h_{\mathcal{M}}, \tag{B.31}$$

where $\gamma_{\mathbf{n}}(\boldsymbol{\varphi}) = \gamma(\boldsymbol{\varphi}) \cdot \boldsymbol{n}_{\partial\Omega}$ is the normal trace of $\boldsymbol{\varphi}$.

Remark B.12 (Broken $W^{1,p'}$ estimate)
The proof actually shows that the result still holds if we take $\boldsymbol{\varphi} \in W^{p'}_{\mathrm{div}}(\Omega) \cap W^{1,p'}(\mathcal{M})^d$, where the broken space $W^{1,p'}(\mathcal{M})^d$ is defined by

$$W^{1,p}(\mathcal{M}) = \{\psi \in L^{p'}(\Omega) \, : \, \forall K \in \mathcal{M} \, , \, \psi \in W^{1,p'}(K)\}.$$

In (B.31), the factor "$\| \, |\nabla \boldsymbol{\varphi}| \, \|_{L^{p'}(\Omega)} \, h_{\mathcal{M}}$" must simply be replaced with

$$\left(\sum_{K \in \mathcal{M}} \| \, |\nabla \boldsymbol{\varphi}| \, \|^{p'}_{L^{p'}(K)} \, h_K^{p'} \right)^{1/p'} ,$$

or $|\boldsymbol{\varphi}|_{W^{1,\infty}(\mathcal{M})} = \max_{K \in \mathcal{M}} (\| \, |\nabla \boldsymbol{\varphi}| \, \|_{L^\infty(K)} \, h_K)$ if $p = 1$.

Proof. Set $\boldsymbol{\varphi}_\sigma = \frac{1}{|\sigma|} \int_\sigma \boldsymbol{\varphi}(\boldsymbol{x}) \mathrm{d}\gamma(\boldsymbol{x})$. Since $\boldsymbol{n}_{K,\sigma} = -\boldsymbol{n}_{L,\sigma}$ whenever σ is a face between K and L, gathering by faces shows that

$$\begin{aligned} \sum_{K \in \mathcal{M}} \sum_{\sigma \in \mathcal{F}_K} & v_\sigma |\sigma| \boldsymbol{\varphi}_\sigma \cdot \boldsymbol{n}_{K,\sigma} \\ &= \sum_{\sigma \in \mathcal{F}_{\mathrm{int}}, \, \mathcal{M}_\sigma = \{K,L\}} v_\sigma |\sigma| \left(\boldsymbol{\varphi}_\sigma \cdot \boldsymbol{n}_{K,\sigma} + \boldsymbol{\varphi}_\sigma \cdot \boldsymbol{n}_{L,\sigma} \right) \\ &\quad + \sum_{\sigma \in \mathcal{F}_{\mathrm{ext}}, \, \mathcal{M}_\sigma = \{K\}} v_\sigma \int_\sigma \boldsymbol{\varphi}(\boldsymbol{x}) \cdot \boldsymbol{n}_{K,\sigma} \mathrm{d}\gamma(\boldsymbol{x}) \\ &= \int_{\partial\Omega} \mathbb{T}_{\mathfrak{T}} v(\boldsymbol{x}) \boldsymbol{\varphi}(\boldsymbol{x}) \cdot \boldsymbol{n}_{\partial\Omega}(\boldsymbol{x}) \mathrm{d}\gamma(\boldsymbol{x}). \end{aligned}$$

By Stokes' formula, $\int_K \mathrm{div}\boldsymbol{\varphi}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = \sum_{\sigma\in\mathcal{F}_K} |\sigma|\boldsymbol{\varphi}_\sigma \cdot \boldsymbol{n}_{K,\sigma}$. Therefore,

$$\begin{aligned}\int_\Omega \Pi_{\mathfrak{T}} v(\boldsymbol{x})\mathrm{div}\boldsymbol{\varphi}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} &= \sum_{K\in\mathcal{M}} v_K \sum_{\sigma\in\mathcal{F}_K} |\sigma|\boldsymbol{\varphi}_\sigma \cdot \boldsymbol{n}_{K,\sigma} \\ &= \sum_{K\in\mathcal{M}}\sum_{\sigma\in\mathcal{F}_K} (v_K - v_\sigma)|\sigma|\boldsymbol{\varphi}_\sigma \cdot \boldsymbol{n}_{K,\sigma} + \int_{\partial\Omega} \mathbb{T}_{\mathfrak{T}} v(\boldsymbol{x})\gamma_{\boldsymbol{n}}(\boldsymbol{\varphi})(\boldsymbol{x})\mathrm{d}\gamma(\boldsymbol{x}). \quad \text{(B.32)}\end{aligned}$$

Introduce $\boldsymbol{\varphi}_K = \frac{1}{|K|}\int_K \boldsymbol{\varphi}(\boldsymbol{x})\mathrm{d}\boldsymbol{x}$ and write, since $\sum_{\sigma\in\mathcal{F}_K} |\sigma|(v_\sigma - v_K)\boldsymbol{n}_{K,\sigma} = |K|\overline{\nabla}_K v$,

$$\begin{aligned}\int_\Omega \Pi_{\mathfrak{T}} v(\boldsymbol{x})\mathrm{div}\boldsymbol{\varphi}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} &- \int_{\partial\Omega} \mathbb{T}_{\mathfrak{T}} v(\boldsymbol{x})\gamma_{\mathbf{n}}(\boldsymbol{\varphi})(\boldsymbol{x})\mathrm{d}\gamma(\boldsymbol{x}) \\ &= \sum_{K\in\mathcal{M}}\sum_{\sigma\in\mathcal{F}_K} |\sigma|(v_K - v_\sigma)\boldsymbol{n}_{K,\sigma}\cdot\boldsymbol{\varphi}_K \\ &\quad + \sum_{K\in\mathcal{M}}\sum_{\sigma\in\mathcal{F}_K} |\sigma|(v_K - v_\sigma)(\boldsymbol{\varphi}_\sigma - \boldsymbol{\varphi}_K)\cdot\boldsymbol{n}_{K,\sigma} \\ &= -\int_\Omega \overline{\nabla}_{\mathfrak{T}} v(\boldsymbol{x})\cdot\boldsymbol{\varphi}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \\ &\quad + \sum_{K\in\mathcal{M}}\sum_{\sigma\in\mathcal{F}_K} |\sigma|(v_K - v_\sigma)(\boldsymbol{\varphi}_\sigma - \boldsymbol{\varphi}_K)\cdot\boldsymbol{n}_{K,\sigma}. \quad \text{(B.33)}\end{aligned}$$

Let T be the left-hand side of (B.31). Equation (B.33) and Hölder's inequality (D.3) show that, for $p > 1$,

$$\begin{aligned}T &\le \sum_{K\in\mathcal{M}}\sum_{\sigma\in\mathcal{F}_K} |\sigma| d_{K,\sigma}\left|\frac{v_\sigma - v_K}{d_{K,\sigma}}\right| |\boldsymbol{\varphi}_\sigma - \boldsymbol{\varphi}_K| \quad \text{(B.34)} \\ &\le \left(\sum_{K\in\mathcal{M}}\sum_{\sigma\in\mathcal{F}_K} |\sigma| d_{K,\sigma}\left|\frac{v_\sigma - v_K}{d_{K,\sigma}}\right|^p\right)^{\frac{1}{p}} \\ &\qquad\qquad \times \left(\sum_{K\in\mathcal{M}}\sum_{\sigma\in\mathcal{F}_K} |\sigma| d_{K,\sigma}|\boldsymbol{\varphi}_\sigma - \boldsymbol{\varphi}_K|^{p'}\right)^{\frac{1}{p'}}.\end{aligned}$$

Apply (B.11) in Lemma B.6 to each component of $\boldsymbol{\varphi}$, with p' instead of p. Since $d_{K,\sigma} \le h_K$, this gives C_6 depending only on d, p and θ such that

$$\begin{aligned}T &\le C_6\, |v|_{\mathfrak{T},p} \left(\sum_{K\in\mathcal{M}}\sum_{\sigma\in\mathcal{F}_K} h_K^{p'} \int_K |\nabla\boldsymbol{\varphi}(\boldsymbol{x})|^{p'}\mathrm{d}\boldsymbol{x}\right)^{\frac{1}{p'}} \\ &\le C_6\theta^{\frac{1}{p'}}\, |v|_{\mathfrak{T},p}\, h_{\mathcal{M}} \,\|\, |\nabla\boldsymbol{\varphi}| \,\|_{L^{p'}(\Omega)}.\end{aligned}$$

This completes the proof in the case $p > 1$. If $p = 1$, simply write $|\boldsymbol{\varphi}_K - \boldsymbol{\varphi}_\sigma| \le \|\, |\nabla\boldsymbol{\varphi}| \,\|_{L^\infty(\Omega)}\, h_{\mathcal{M}}$ in (B.34). ∎

B.2 Discrete functional analysis for Dirichlet boundary conditions

We establish discrete functional analysis results in the case of Dirichlet boundary conditions. We first consider discrete Sobolev embeddings, starting with the case $p = 1$ and then generalising to the case $p > 1$. Then we study a Rellich compactness result, also looking at the case $p = 1$ first. All these results apply to functions reconstructed, through $\Pi_{\mathfrak{T}}$, from elements in $X_{\mathfrak{T},0}$.

B.2.1 Discrete Sobolev embeddings

Let us first recall the Sobolev embedding, due to L. Nirenberg, of $W^{1,1}(\mathbb{R}^d)$ into $L^{1^\star}(\mathbb{R}^d)$, where $1^\star = \frac{d}{d-1}$:

$$\forall w \in W^{1,1}(\mathbb{R}^d)\,,\ \|w\|_{L^{1^\star}(\mathbb{R}^d)} \leq \frac{1}{2d}\sum_{i=1}^{d} \|\partial_i w\|_{L^1(\mathbb{R}^d)}\,. \tag{B.35}$$

Recall that the $BV(\mathbb{R}^d)$ norm of functions in $L^1(\mathbb{R}^d)$ is defined by

$$\|w\|_{BV(\mathbb{R}^d)} = \sup\Big\{ \int_{\mathbb{R}^d} w(\boldsymbol{x})\mathrm{div}\boldsymbol{\varphi}(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\ :\ \boldsymbol{\varphi} \in C_c^\infty(\mathbb{R}^d, \mathbb{R}^d), \\ \|\boldsymbol{\varphi}\|_{L^\infty(\mathbb{R}^d)^d} \leq 1\Big\},$$

with $\boldsymbol{\varphi} = (\varphi_1, \ldots, \varphi_d)$ and $\|\boldsymbol{\varphi}\|_{L^\infty(\mathbb{R}^d)^d} = \sup_{i=1,\ldots,d} \|\varphi_i\|_{L^\infty(\mathbb{R}^d)}$. The space $BV(\mathbb{R}^d)$ is defined as the set of functions $w \in L^1(\Omega)$ such that $\|w\|_{BV(\mathbb{R}^d)} < \infty$. The Sobolev embedding (B.35) can be extended to $BV(\mathbb{R}^d)$, by using a regularisation technique.
Precisely, let $w \in BV(\mathbb{R}^d)$ and take $(\rho_n)_{n\geq 1}$ a smoothing kernel, that is, $\rho_1 \in C_c^\infty(B(0,1))$, $\rho_1 \geq 0$, $\int_{B(0,1)} \rho_1(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = 1$, and $\rho_n(\boldsymbol{x}) = n^d \rho_1(n\boldsymbol{x})$. Then, $w_n = w \star \rho_n$ belongs to $W^{1,1}(\mathbb{R}^d)$, and $w_n \to w$ in $L^1(\mathbb{R}^d)$ as $n \to \infty$ (and thus a.e. up to a subsequence). Moreover, $\sum_{i=1}^d \|\partial_i w_n\|_{L^1(\mathbb{R}^d)} \leq \|w\|_{BV(\mathbb{R}^d)}$. Apply then (B.35) to $w = w_n$ to obtain

$$\|w_n\|_{L^{1^\star}(\mathbb{R}^d)} \leq \frac{1}{2d}\sum_{i=1}^{d} \|\partial_i w_n\|_{L^1(\mathbb{R}^d)} \leq \frac{1}{2d}\|w\|_{BV(\mathbb{R}^d)}\,.$$

Taking now the inferior limit and using Fatou's lemma in the left-hand side yields that

$$\forall w \in BV(\mathbb{R}^d)\,,\ \|w\|_{L^{1^\star}(\mathbb{R}^d)} \leq \frac{1}{2d}\|w\|_{BV(\mathbb{R}^d)}\,. \tag{B.36}$$

Let us now state the discrete Sobolev embedding for $p = 1$.

Lemma B.13 (Discrete embedding of $W_0^{1,1}(\Omega)$ in $L^{1^\star}(\Omega)$). *Let $\mathfrak{T}$ be a polytopal mesh of Ω. Setting $1^\star = \frac{d}{d-1}$ and recalling the notations (7.7), we have*

$$\forall u \in X_{\mathfrak{T},0}\,,\ \|\Pi_{\mathfrak{T}} u\|_{L^{1^\star}(\Omega)} \le \frac{1}{2\sqrt{d}}\,|u|_{\mathfrak{T},1}\,. \tag{B.37}$$

Proof. Let $u \in X_{\mathfrak{T},0}$, and extend $\Pi_{\mathfrak{T}} u$ by 0 outside Ω. We have $\Pi_{\mathfrak{T}} u \in L^1(\mathbb{R}^d)$. Let $\boldsymbol{\varphi} \in C_c^\infty(\mathbb{R}^d, \mathbb{R}^d)$ such that $\|\boldsymbol{\varphi}\|_{L^\infty(\mathbb{R}^d)} \le 1$. This implies $|\boldsymbol{\varphi}| \le \sqrt{d}$. Write (B.32) for $v = u$ and take into account the boundary conditions $u_\sigma = 0$ for all $\sigma \in \mathcal{F}_{\rm ext}$ (which implies $\mathbb{T}_{\mathfrak{T}} u = 0$) to obtain

$$\begin{aligned}\int_{\mathbb{R}^d} \Pi_{\mathfrak{T}} u(\boldsymbol{x}) \mathrm{div}\boldsymbol{\varphi}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} &= \int_\Omega \Pi_{\mathfrak{T}} u(\boldsymbol{x}) \mathrm{div}\boldsymbol{\varphi}(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\\ &= \sum_{K\in\mathcal{M}} \sum_{\sigma\in\mathcal{F}_K} |\sigma|(u_K - u_\sigma)\frac{1}{|\sigma|}\int_\sigma \boldsymbol{\varphi}(\boldsymbol{x})\cdot \boldsymbol{n}_{K,\sigma}\mathrm{d}\gamma(\boldsymbol{x})\\ &\le \sqrt{d}\sum_{K\in\mathcal{M}} \sum_{\sigma\in\mathcal{F}_K} |\sigma||u_K - u_\sigma| = \sqrt{d}\,|u|_{\mathfrak{T},1}\,.\end{aligned} \tag{B.38}$$

Hence, $\|\Pi_{\mathfrak{T}} u\|_{BV(\mathbb{R}^d)} \le \sqrt{d}\,|u|_{\mathfrak{T},1}$ and (B.36) leads to (B.37). ∎

Lemma B.14 (Discrete embedding of $W_0^{1,p}(\Omega)$ in $L^{p^\star}(\Omega)$, $1 < p < d$). *Let $\mathfrak{T}$ be a polytopal mesh of Ω, $p \in (1,d)$ and $p^\star = \frac{pd}{d-p}$. Then, there exists C_7, depending only on d, p and $\eta \ge \eta_{\mathfrak{T}}$ (see (7.9)), such that*

$$\forall u \in X_{\mathfrak{T},0}\,,\ \|\Pi_{\mathfrak{T}} u\|_{L^{p^\star}(\Omega)} \le C_7\,|u|_{\mathfrak{T},p}\,. \tag{B.39}$$

Proof. We follow again L. Nirenberg's ideas. Let α be such that $\alpha 1^\star = p^\star$, that is, $\alpha = p(d-1)/(d-p) > 1$. Take $u \in X_{\mathfrak{T},0}$ and define $\widehat{u} = ((|u_K|^\alpha)_{K\in\mathcal{M}}, (\widehat{u}_\sigma)_{\sigma\in\mathcal{F}})$ with

$$\begin{aligned}&\widehat{u}_\sigma = \frac{1}{2}(|u_K|^\alpha + |u_L|^\alpha) \text{ for all } \sigma \in \mathcal{F}_{\rm int} \text{ with } \mathcal{M}_\sigma = \{K, L\},\\ &\widehat{u}_\sigma = 0 \text{ if } \sigma \in \mathcal{F}_{\rm ext}.\end{aligned}$$

Since $|\Pi_{\mathfrak{T}}\widehat{u}|^{\frac{d}{d-1}} = |\Pi_{\mathfrak{T}} u|^{p^\star}$, applying (B.37) to $\widehat{u}$ and gathering the sums by edges gives

$$\Big(\int_\Omega |\Pi_{\mathfrak{T}} u(\boldsymbol{x})|^{p^\star}\mathrm{d}\boldsymbol{x}\Big)^{\frac{d-1}{d}} \le \frac{1}{2\sqrt{d}}\sum_{K\in\mathcal{M}}\sum_{\sigma\in\mathcal{F}_K} |\sigma|\,\big||u_K|^\alpha - \widehat{u}_\sigma\big| \tag{B.40}$$

$$\le \frac{1}{2\sqrt{d}}\sum_{\sigma\in\mathcal{F}_{\rm ext},\,\mathcal{M}_\sigma=\{K\}} |\sigma||u_K|^\alpha + \frac{1}{2\sqrt{d}}\sum_{\sigma\in\mathcal{F}_{\rm int},\,\mathcal{M}_\sigma=\{K,L\}} |\sigma|\,\big||u_K|^\alpha - |u_L|^\alpha\big|.$$

Since $f : s \mapsto s^\alpha$ is differentiable on $[0,\infty)$ and $\sup_{[a,b]}|f'| \le \alpha(a^{\alpha-1} + b^{\alpha-1})$ for all $0 \le a \le b$, the mean value theorem yields

$$\big|\,|u_K|^\alpha - |u_L|^\alpha\big| \le \alpha(|u_K|^{\alpha-1} + |u_L|^{\alpha-1})|u_K - u_L|. \tag{B.41}$$

Hence, setting $\delta_\sigma u = |u_K|$ if $\sigma \in \mathcal{F}_{\rm ext}$ and $\delta_\sigma u = |u_K - u_L|$ if $\sigma \in \mathcal{F}_{\rm int}$, gathering back by cells,

$$\begin{aligned}\left(\int_\Omega |\Pi_{\mathfrak{T}} u(\boldsymbol{x})|^{p^\star}\,\mathrm{d}\boldsymbol{x}\right)^{\frac{d-1}{d}} &\le \frac{\alpha}{2\sqrt{d}} \sum_{K\in\mathcal{M}} \sum_{\sigma\in\mathcal{F}_K} |\sigma||u_K|^{\alpha-1}\delta_\sigma u \\ &= \frac{\alpha}{2\sqrt{d}} \sum_{K\in\mathcal{M}} \sum_{\sigma\in\mathcal{F}_K} |\sigma| d_{K,\sigma}|u_K|^{\alpha-1}\frac{\delta_\sigma u}{d_{K,\sigma}}.\end{aligned}$$

The Hölder inequality (D.3) then yields

$$\begin{aligned}\left(\int_\Omega |\Pi_{\mathfrak{T}} u(\boldsymbol{x})|^{p^\star}\,\mathrm{d}\boldsymbol{x}\right)^{\frac{d-1}{d}} \le \frac{\alpha}{2\sqrt{d}} &\left(\sum_{K\in\mathcal{M}} \sum_{\sigma\in\mathcal{F}_K} |\sigma| d_{K,\sigma}|u_K|^{(\alpha-1)p'}\right)^{\frac{1}{p'}} \\ &\times \left(\sum_{K\in\mathcal{M}} \sum_{\sigma\in\mathcal{F}_K} |\sigma| d_{K,\sigma}\left|\frac{\delta_\sigma u}{d_{K,\sigma}}\right|^p\right)^{\frac{1}{p}}.\end{aligned} \tag{B.42}$$

Since $(\alpha-1)p' = p^\star$ and $\sum_{\sigma\in\mathcal{F}_K} |\sigma| d_{K,\sigma} = d|K|$ (see (B.1)),

$$\begin{aligned}\sum_{K\in\mathcal{M}} \sum_{\sigma\in\mathcal{F}_K} |\sigma| d_{K,\sigma}|u_K|^{(\alpha-1)p'} &= \sum_{K\in\mathcal{M}} d|K|\,|u_K|^{p^\star} \\ &= d\int_\Omega |\Pi_{\mathfrak{T}} u(\boldsymbol{x})|^{p^\star}\,\mathrm{d}\boldsymbol{x}.\end{aligned}$$

Plugging this into (B.42) and noticing that $\frac{d-1}{d} - \frac{1}{p'} = \frac{1}{p^\star}$, this shows that

$$\|\Pi_{\mathfrak{T}} u\|_{L^{p^\star}(\Omega)} \le \frac{\alpha d^{1/p'}}{2\sqrt{d}} \left(\sum_{K\in\mathcal{M}} \sum_{\sigma\in\mathcal{F}_K} |\sigma| d_{K,\sigma}\left|\frac{\delta_\sigma u}{d_{K,\sigma}}\right|^p\right)^{\frac{1}{p}}. \tag{B.43}$$

For $\mathcal{M}_\sigma = \{K, L\}$, using the definition of η,

$$\begin{aligned}d_{K,\sigma}\left|\frac{\delta_\sigma u}{d_{K,\sigma}}\right|^p &\le \frac{1}{d_{K,\sigma}^{p-1}}(|u_K - u_\sigma| + |u_\sigma - u_L|)^p \\ &\le 2^{p-1}\left(\frac{|u_K - u_\sigma|^p}{d_{K,\sigma}^{p-1}} + \frac{|u_L - u_\sigma|^p}{d_{L,\sigma}^{p-1}}\right)\frac{d_{K,\sigma}^{p-1} + d_{L,\sigma}^{p-1}}{d_{K,\sigma}^{p-1}} \\ &\le 2^{p-1}\left(d_{K,\sigma}\left|\frac{u_K - u_\sigma}{d_{K,\sigma}}\right|^p + d_{L,\sigma}\left|\frac{u_L - u_\sigma}{d_{L,\sigma}}\right|^p\right)(1 + \eta^{p-1}).\end{aligned}$$

The same holds, with $u_L = 0$, if $\sigma \in \mathcal{F}_K \cap \mathcal{F}_{\rm ext}$. Hence,

$$\sum_{K\in\mathcal{M}} \sum_{\sigma\in\mathcal{F}_K} |\sigma| d_{K,\sigma}\left|\frac{\delta_\sigma u}{d_{K,\sigma}}\right|^p$$

$$\leq 2^{p-1}(1+\eta^{p-1}) \sum_{K\in\mathcal{M}} \sum_{\sigma\in\mathcal{F}_K} |\sigma| \left(d_{K,\sigma} \left| \frac{u_K - u_\sigma}{d_{K,\sigma}} \right|^p + d_{L,\sigma} \left| \frac{u_L - u_\sigma}{d_{L,\sigma}} \right|^p \right)$$

$$\leq 2^{p}(1+\eta^{p-1}) \sum_{K\in\mathcal{M}} \sum_{\sigma\in\mathcal{F}_K} |\sigma| d_{K,\sigma} \left| \frac{u_K - u_\sigma}{d_{K,\sigma}} \right|^p . \tag{B.44}$$

To write the last line, we noticed that each contribution involving $u_K - u_\sigma$ appears twice for interior edges (once when summing over $\sigma \in \mathcal{F}_K$, and another one when summing over $\sigma \in \mathcal{F}_L$). The Sobolev inequality (B.39) is deduced from (B.43), (B.44) and the definition of $|u|_{\mathfrak{T},p}$. ∎

To prove the final result of this section, we first need to establish a natural inequality on discrete Sobolev norms. Let $1 \leq q < p < +\infty$. Using Hölder's inequality (D.3) with exponents $\frac{p}{q} > 1$ and $\frac{p}{p-q}$, we have

$$\begin{aligned} |u|_{\mathfrak{T},q} &= \left(\sum_{K\in\mathcal{M}} \sum_{\sigma\in\mathcal{F}_K} |\sigma| d_{K,\sigma} \left| \frac{u_\sigma - u_K}{d_{K,\sigma}} \right|^q \right)^{\frac{1}{q}} \\ &\leq \left(\sum_{K\in\mathcal{M}} \sum_{\sigma\in\mathcal{F}_K} |\sigma| d_{K,\sigma} \left| \frac{u_\sigma - u_K}{d_{K,\sigma}} \right|^p \right)^{\frac{1}{p}} \left(\sum_{K\in\mathcal{M}} \sum_{\sigma\in\mathcal{F}_K} |\sigma| d_{K,\sigma} \right)^{\frac{1}{q}-\frac{1}{p}} \\ &= |u|_{\mathfrak{T},p} \, (d|\Omega|)^{\frac{1}{q}-\frac{1}{p}}. \end{aligned} \tag{B.45}$$

In the last line, we invoked (B.1).

Lemma B.15 (Discrete embedding of $W_0^{1,p}(\Omega)$ in $L^q(\Omega)$, for some $q > p$). *Let $\mathfrak{T}$ be a polytopal mesh of Ω, $p \in [1, +\infty)$ and $\eta \geq \eta_{\mathfrak{T}}$. Then, there exists $q > p$, depending only on p and d, and there exists C_8, depending only on Ω, p, q and η, such that*

$$\forall u \in X_{\mathfrak{T},0}\,, \ \|\Pi_{\mathfrak{T}} u\|_{L^q(\Omega)} \leq C_8 \, |u|_{\mathfrak{T},p} \,. \tag{B.46}$$

If $p < d$ we can take $q = p^\star = \frac{pd}{d-p}$ and, if $p \geq d$, we can take any $q < +\infty$.

Remark B.16 (Discrete Poincaré inequality). Combining (B.46) and the Hölder inequality (D.7) yields the discrete Poincaré inequality

$$\|\Pi_{\mathfrak{T}} u\|_{L^p(\Omega)} \leq C_8 |\Omega|^{\frac{1}{p}-\frac{1}{q}} \, |u|_{\mathfrak{T},p} \,.$$

This has been established here for polytopal meshes, using in particular the assumption that each cell is star-shaped with respect to a ball. For some numerical methods, appropriate discrete Poincaré inequalities can be proved with a milder assumption [137, 86].

Proof. If $p = 1$, take $q = 1^\star$ and the result follows from Lemma B.13 (in this case, C_8 does not depend on η). If $1 < p < d$, take $q = p^\star$ and the result is given by Lemma B.14.

If $p \geq d$, choose any $q \in (p, \infty)$ and take $p_1 < d$ such that $p_1^\star = q$ (this is possible since $p_1^\star$ tends to $+\infty$ as p_1 tends to d). The choice of p_1 depends only on q and d, and Lemma B.14 gives

$$\|\Pi_{\mathfrak{T}} u\|_{L^q(\Omega)} \leq C_7 \, |u|_{\mathfrak{T},p_1}$$

for some C_7 depending only on p_1, d and η. Inequality (B.46) follows from this estimate and (B.45) with $q = p_1$. ∎

B.2.2 Compactness of $\Pi_{\mathfrak{T}}$

The continuous Rellich theorem states that bounded families in $W_0^{1,p}(\Omega)$ are relatively compact in $L^p(\Omega)$. We prove here a discrete version of this result, involving the discrete $W_0^{1,p}(\Omega)$ norm $|\cdot|_{\mathfrak{T},p}$ and the function reconstruction operator $\Pi_{\mathfrak{T}}$. As for Sobolev embeddings, with start with the case $p = 1$, which requires less assumptions on the mesh and from which we deduce the case $p > 1$.

Lemma B.17 (Estimates on the translates in L^1). *Let $\mathfrak{T}$ be a polytopal mesh of Ω in the sense of Definition 7.2. Let $u \in X_{\mathfrak{T},0}$ and extend $\Pi_{\mathfrak{T}} u$ to $\mathbb{R}^d$ by 0 outside Ω. Then,*

$$\forall \boldsymbol{h} \in \mathbb{R}^d \,, \; \|\Pi_{\mathfrak{T}} u(\cdot + \boldsymbol{h}) - \Pi_{\mathfrak{T}} u\|_{L^1(\mathbb{R}^d)} \leq |\boldsymbol{h}| \sqrt{d} \, |u|_{\mathfrak{T},1} \,. \tag{B.47}$$

Proof. Since $p = 1$, the proof can be done by following the technique in [91], which yields (B.47) without $\sqrt{d}$. A more direct proof based on the BV space is chosen here, as in Lemma B.13.

Let $w \in C_c^\infty(\mathbb{R}^d)$. For $\boldsymbol{x}, \boldsymbol{h} \in \mathbb{R}^d$, write

$$|w(\boldsymbol{x} + \boldsymbol{h}) - w(\boldsymbol{x})| = \left| \int_0^1 \nabla w(\boldsymbol{x} + t\boldsymbol{h}) \cdot \boldsymbol{h} \mathrm{d}t \right| \leq |\boldsymbol{h}| \int_0^1 |\nabla w(\boldsymbol{x} + t\boldsymbol{h})| \mathrm{d}t.$$

Integrating with respect to $\boldsymbol{x} \in \mathbb{R}^d$ and using Fubini's theorem gives the well known result

$$\|w(\cdot + \boldsymbol{h}) - w\|_{L^1(\mathbb{R}^d)} \leq |\boldsymbol{h}| \int_{\mathbb{R}^d} |\nabla w(\boldsymbol{x})| \mathrm{d}\boldsymbol{x} \leq |\boldsymbol{h}| \sum_{i=1}^d \|\partial_i w\|_{L^1(\mathbb{R}^d)} \,. \tag{B.48}$$

By density of $C_c^\infty(\mathbb{R}^d)$ in $W^{1,1}(\mathbb{R}^d)$, Inequality (B.48) is also true for $w \in W^{1,1}(\mathbb{R}^d)$ and, proceeding as at the start of Section B.2.1, leads to the following estimate for $BV(\mathbb{R}^d)$ functions:

$$\forall w \in BV(\mathbb{R}^d) \,, \; \forall \boldsymbol{h} \in \mathbb{R}^d \,, \; \|w(\cdot + \boldsymbol{h}) - w\|_{L^1(\mathbb{R}^d)} \leq |\boldsymbol{h}| \, \|w\|_{BV(\mathbb{R}^d)} \,. \tag{B.49}$$

Take now $u \in X_{\mathfrak{T},0}$ and, as in the statement of the lemma, set $\Pi_{\mathfrak{T}} u = 0$ outside Ω. Then $\Pi_{\mathfrak{T}} u \in L^1(\mathbb{R}^d)$ and it was proved in lemma B.13 that $\|\Pi_{\mathfrak{T}} u\|_{BV(\mathbb{R}^d)} \leq \sqrt{d}\,|u|_{\mathfrak{T},1}$. The proof is therefore complete by applying (B.49) to $w = \Pi_{\mathfrak{T}} u$. ∎

The following compactness result in L^1 results from Lemmas B.13 and B.17, and the Kolmogorov compactness criterion.

Lemma B.18 (Discrete Rellich theorem, $p = 1$). *Let $(\mathfrak{T}_m)_{m\in\mathbb{N}}$ be a sequence of polytopal meshes of Ω. Then, for any $u_m \in X_{\mathfrak{T}_m,0}$ such that $(|u_m|_{\mathfrak{T}_m,1})_{m\in\mathbb{N}}$ is bounded, the sequence $(\Pi_{\mathfrak{T}_m} u_m)_{m\in\mathbb{N}}$ is relatively compact in $L^1(\Omega)$.*

Proof. Lemma B.13 shows that $(\Pi_{\mathcal{D}_m} u_m)_{m\in\mathbb{N}}$ is bounded in $L^{1^\star}(\Omega)$, and thus also in $L^1(\Omega)$ since Ω is bounded. Extending the functions $\Pi_{\mathfrak{T}_m} u_m$ by 0 outside Ω, they remain bounded in $L^1(\mathbb{R}^d)$. The Kolmogorov compactness theorem [33, Theorem 4.26] and Lemma B.17 then show that $(\Pi_{\mathcal{D}_m} u_m)_{m\in\mathbb{N}}$ is relatively compact in $L^1(\Omega)$. ∎

As for discrete Sobolev embeddings, establishing a compactness result for $p > 1$ requires some an additional hypothesis on the meshes.

Lemma B.19 (Discrete Rellich theorem, $p > 1$). *Let $p \in [1, +\infty)$ and $(\mathfrak{T}_m)_{m\in\mathbb{N}}$ be a sequence of polytopal meshes of Ω, such that $\sup_{m\in\mathbb{N}} \eta_{\mathfrak{T}_m} < +\infty$. Then, for any $u_m \in X_{\mathfrak{T}_m,0}$ such that $(|u_m|_{\mathfrak{T}_m,p})_{m\in\mathbb{N}}$ is bounded, the sequence $(\Pi_{\mathfrak{T}_m} u_m)_{m\in\mathbb{N}}$ is relatively compact in $L^p(\Omega)$.*

Proof. Using (B.45) with $q = 1$ shows that $(|u_m|_{\mathfrak{T}_m,1})_{m\in\mathbb{N}}$ is bounded. By Lemma B.18, $(\Pi_{\mathcal{D}_m} u_m)_{m\in\mathbb{N}}$ is thus relatively compact in $L^1(\Omega)$ and, up to a subsequence denoted the same way, converges in this space. By Lemma B.15, $(\Pi_{\mathcal{D}_m} u_m)_{m\in\mathbb{N}}$ is also bounded by some C_9 in $L^q(\Omega)$ for some $q > p$.
Recall now the interpolation inequality, consequence of Hölder's inequality (D.5) applied to $|f|^p = |f|^{\alpha p}|f|^{(1-\alpha)p}$ with $\alpha = \frac{q-p}{(q-1)p}$ and exponents $(r, r') = (\frac{1}{p\alpha}, \frac{q}{(1-\alpha)p})$:

$$\|f\|_{L^p(\Omega)} \leq \|f\|_{L^1(\Omega)}^{\frac{q-p}{(q-1)p}} \|f\|_{L^q(\Omega)}^{\frac{q(p-1)}{(q-1)p}}.$$

Apply this estimate to $f = \Pi_{\mathfrak{T}_m} u_m - \Pi_{\mathfrak{T}_\ell} u_\ell$ and use $\|f\|_{L^q(\Omega)} \leq 2C_9$. This gives

$$\|\Pi_{\mathfrak{T}_m} u_m - \Pi_{\mathfrak{T}_\ell} u_\ell\|_{L^p(\Omega)} \leq \|\Pi_{\mathfrak{T}_m} u_m - \Pi_{\mathfrak{T}_\ell} u_\ell\|_{L^1(\Omega)}^{\frac{q-p}{(q-1)p}} (2C_9)^{\frac{q(p-1)}{(q-1)p}}. \tag{B.50}$$

Since $\frac{q-p}{(q-1)p} > 0$ and $(\Pi_{\mathfrak{T}_m} u_m)_{m\in\mathbb{N}}$ is a Cauchy sequence in $L^1(\Omega)$, (B.50) shows that $(\Pi_{\mathfrak{T}_m} u_m)_{m\in\mathbb{N}}$ is also a Cauchy sequence in $L^p(\Omega)$, and thus that it converges in this space. ∎

B.3 Discrete functional analysis for Neumann and Fourier BCs

We develop here discrete functional analysis results for Neumann and Fourier boundary conditions.

B.3.1 Estimates involving the reconstructed trace

Let us start with discrete versions of classical trace estimates, stated for the case $p = 1$ in Lemma B.20 and for the case $p > 1$ in Lemma B.21.

Lemma B.20 (Discrete trace inequality, $p = 1$). *Let $\mathfrak{T}$ be a polytopal mesh of Ω in the sense of Definition 7.2, and $\varrho \geq \theta_{\mathfrak{T}} + \eta_{\mathfrak{T}}$ (see (7.8) and (7.9)). Then, there exists $C_{10} > 0$, depending only on Ω, d and ϱ, such that*

$$\forall u \in X_{\mathfrak{T}}\,,\ \|\mathbb{T}_{\mathfrak{T}} u\|_{L^1(\partial\Omega)} \leq C_{10} \left(|u|_{\mathfrak{T},1} + \|\Pi_{\mathfrak{T}} u\|_{L^1(\Omega)} \right). \tag{B.51}$$

Proof.
Step 1: we prove the existence of a finite family $(\tau_i, \boldsymbol{\xi}_i)_{i=1,\dots,M}$ such that:

1. for $i = 1, \dots, M$, $\tau_i \subset \partial\Omega$ is an open connected subset of an external face of Ω, with outward unit normal vector $\boldsymbol{n}_{\tau_i}$,
2. $\boldsymbol{\xi}_i \in \mathbb{R}^d \setminus \{0\}$ and the cylinder $\mathcal{C}(\tau_i, \boldsymbol{\xi}_i) = \{\boldsymbol{x} + t\boldsymbol{\xi}_i \,:\, t \in (0,1)\,,\ \boldsymbol{x} \in \tau_i\}$ is contained in Ω,
3. there exists $\alpha > 0$ such that $-\boldsymbol{\xi}_i \cdot \boldsymbol{n}_{\tau_i} \geq \alpha |\boldsymbol{\xi}_i|$,
4. $\partial\Omega \subset \bigcup_{i=1,\dots,M} \overline{\tau}_i$.

To establish the existence of this family, recall that $\overline{\Omega}$ can be defined as a finite union of simplices of $\mathbb{R}^d$. Take one of these simplices $S = \mathcal{S}((\boldsymbol{x}_i)_{i=1,\dots,d+1})$ (see (7.1)), that touches the boundary of Ω and whose interior S^o is contained in Ω. Assume that the face $F = \mathcal{S}((\boldsymbol{x}_\ell)_{\ell=1,\dots,d})$ of S is an external face of Ω and define

$$\tau_i = \left\{ \sum_{j=1}^d \alpha_j \boldsymbol{x}_j \,:\, \sum_{j=1}^d \alpha_j = 1\,,\ \alpha_j > 0 \text{ for all } j, \text{ and } \alpha_i > \frac{1}{d+1} \right\}.$$

For any family of real numbers $(\alpha_i)_{i=1,\dots,d}$ such that $\sum_{j=1}^d \alpha_j = 1$, by way of contradiction we can find $i \in \{1, \dots, d\}$ such that $\alpha_i > \frac{1}{d+1}$. Hence,

$$F = \mathcal{S}((\boldsymbol{x}_\ell)_{\ell=1,\dots,d}) = \bigcup_{i=1}^d \overline{\tau}_i.$$

Let $\boldsymbol{n}_{\tau_i}$ be the unit normal to τ_i (that is, to F) outside S, and set $\boldsymbol{\xi}_i = \frac{1}{d+1}(\boldsymbol{x}_{d+1} - \boldsymbol{x}_i)$. If $\boldsymbol{x} \in \mathcal{C}(\tau_i, \boldsymbol{\xi}_i)$ then there exists $t \in (0,1)$ and $(\alpha_i)_{i=1,\dots,d}$, with $\alpha_j > 0$ for all j and $\alpha_i > \frac{1}{d+1}$, such that

$$\boldsymbol{x} = \sum_{j=1}^{d} \alpha_j \boldsymbol{x}_j + \frac{t}{d+1}(\boldsymbol{x}_{d+1} - \boldsymbol{x}_i)$$
$$= \sum_{j=1, j\neq i}^{d} \alpha_j \boldsymbol{x}_j + \left(\alpha_i - \frac{t}{d+1}\right)\boldsymbol{x}_i + \frac{t}{d+1}\boldsymbol{x}_{d+1}.$$

Since $\alpha_i - \frac{t}{d+1} > 0$, all the coefficients in this convex combination of the vertices of S are strictly positive, so $\boldsymbol{x} \in S^o \subset \Omega$. Hence, $\mathcal{C}(\tau_i, \boldsymbol{\xi}_i) \subset \Omega$.
Finally, since $\boldsymbol{x}_i \in F$, $-\boldsymbol{\xi}_i \cdot \boldsymbol{n}_{\tau_i} = \frac{1}{d+1}(\boldsymbol{x}_i - \boldsymbol{x}_{d+1}) \cdot \boldsymbol{n}_{\tau_i}$ is strictly positive, since it is $\frac{1}{d+1}$ times the orthogonal distance between $\boldsymbol{x}_{d+1}$ and F. We are working with a global finite number (depending only on Ω) of indices $i = 1, \ldots, M$, so $\alpha = \min_{i=1,\ldots,M}(-\boldsymbol{\xi}_i \cdot \boldsymbol{n}_{\tau_i}/|\boldsymbol{\xi}_i|)$ is strictly positive.

Step 2: proof of the trace inequality (B.51).

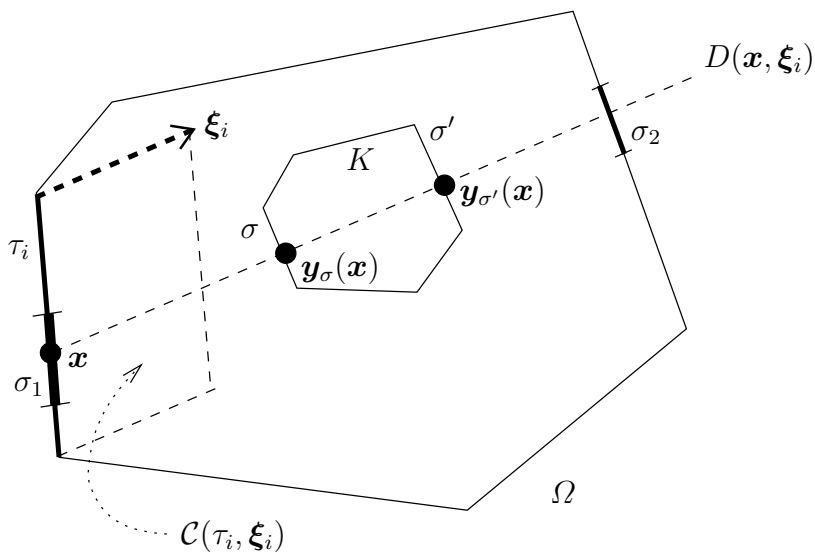


Fig. B.3. Illustration of Step 2 in the proof of Lemma B.20. Here, $\chi_{K,\sigma}(\boldsymbol{x}) = 1$, $\chi_{K,\sigma'}(\boldsymbol{x}) = -1$, $\beta_{\sigma_1}(\boldsymbol{x}) = 1$ and $\beta_{\sigma_2}(\boldsymbol{x}) = 0$.

Fix $i \in \{1, \ldots, M\}$ and denote by $D(\boldsymbol{x}, \boldsymbol{\xi}_i)$ the half line starting from $\boldsymbol{x}$ and with direction $\boldsymbol{\xi}_i$. For $K \in \mathcal{M}$ and $\sigma \in \mathcal{F}_K$, take $\boldsymbol{x} \in \tau_i$ such that (see Figure B.3 for an illustration):

- either $D(\boldsymbol{x}, \boldsymbol{\xi}_i)$ does not intersect σ, in which case set $\boldsymbol{y}_\sigma(\boldsymbol{x}) = \boldsymbol{x}$ and $\chi_{K,\sigma}(\boldsymbol{x}) = 0$,
- or $D(\boldsymbol{x}, \boldsymbol{\xi}_i)$ intersect σ at only one point, in which case set $\boldsymbol{y}_\sigma(\boldsymbol{x})$ as this point and

- ⋆ $\chi_{K,\sigma}(\boldsymbol{x}) = 1$ if, starting from $\boldsymbol{x}$, $D(\boldsymbol{x}, \boldsymbol{\xi}_i)$ intersects σ while entering into K,
- ⋆ $\chi_{K,\sigma}(\boldsymbol{x}) = -1$ if, starting from $\boldsymbol{x}$, $D(\boldsymbol{x}, \boldsymbol{\xi}_i)$ intersects σ while exiting K.

In other words, $\chi_{K,\sigma}(\boldsymbol{x}) = -\mathrm{sgn}(\boldsymbol{\xi}_i \cdot \boldsymbol{n}_{K,\sigma})$.

Note that a.e. $\boldsymbol{x} \in \tau_i$ fall into one or the other of these two categories. The half line $D(\boldsymbol{x}, \boldsymbol{\xi}_i)$ always exits a cell after entering it and thus

$$\forall K \in \mathcal{M}\,, \sum_{\sigma \in \mathcal{F}_K} \chi_{K,\sigma}(\boldsymbol{x}) = 0. \tag{B.52}$$

Define

$$\begin{aligned}
\forall \sigma \in \mathcal{F}\,,\ \beta_\sigma(\boldsymbol{x}) &= \max\left(1 - \frac{(\boldsymbol{y}_\sigma(\boldsymbol{x}) - \boldsymbol{x}) \cdot \boldsymbol{\xi}_i}{|\boldsymbol{\xi}_i|^2}, 0\right), \\
\forall K \in \mathcal{M}\,,\ \beta_K(\boldsymbol{x}) &= \max\left(1 - \frac{(\boldsymbol{x}_K - \boldsymbol{x}) \cdot \boldsymbol{\xi}_i}{|\boldsymbol{\xi}_i|^2}, 0\right).
\end{aligned}$$

Let $\sigma \in \mathcal{F}_{\mathrm{ext}}$ be such that $\chi_{K,\sigma}(\boldsymbol{x}) \neq 0$. If $\boldsymbol{x} \in \sigma$ then $\boldsymbol{y}_\sigma(\boldsymbol{x}) = \boldsymbol{x}$ and thus $\beta_\sigma(\boldsymbol{x}) = 1$. If $\boldsymbol{x} \notin \sigma$, then the inclusion $\mathcal{C}(\tau_i, \boldsymbol{\xi}_i) \subset \Omega$ shows that $\boldsymbol{y}_\sigma(\boldsymbol{x}) \notin \mathcal{C}(\tau_i, \boldsymbol{\xi}_i)$ and thus that $(\boldsymbol{y}_\sigma(\boldsymbol{x}) - \boldsymbol{x}) \cdot \boldsymbol{\xi}_i \geq |\boldsymbol{\xi}_i|^2$, which implies $\beta_\sigma(\boldsymbol{x}) = 0$. If $\sigma \in \mathcal{F}_{\mathrm{int}}$ with $\mathcal{M}_\sigma = \{K, L\}$ and $D(\boldsymbol{x}, \boldsymbol{\xi}_i)$ crosses σ, then if it exits K (for example) it must enter L and thus $\chi_{K,\sigma}(\boldsymbol{x}) = -\chi_{L,\sigma}(\boldsymbol{x})$. As a consequence of this reasoning, for a.e. $\boldsymbol{x} \in \tau_i$ and for all $\sigma \in \mathcal{F}$,

$$\begin{aligned}
&\text{If } \boldsymbol{x} \notin \sigma \text{ then } \sum_{K \in \mathcal{M}_\sigma} \chi_{K,\sigma}(\boldsymbol{x}) \beta_\sigma(\boldsymbol{x}) = 0\,, \\
&\text{If } \boldsymbol{x} \in \sigma \text{ then } \sum_{K \in \mathcal{M}_\sigma} \chi_{K,\sigma}(\boldsymbol{x}) \beta_\sigma(\boldsymbol{x}) = 1
\end{aligned} \tag{B.53}$$

(note that the second situation only happens for a single $\sigma \in \mathcal{F}_{\mathrm{ext}}$ since $\boldsymbol{x} \in \partial\Omega$). Relations (B.52) and (B.53) show that

$$\begin{aligned}
&\sum_{K \in \mathcal{M}} \sum_{\sigma \in \mathcal{F}_K} \chi_{K,\sigma}(\boldsymbol{x})(\beta_\sigma(\boldsymbol{x}) u_\sigma - \beta_K(\boldsymbol{x}) u_K) \\
&\quad = \sum_{\sigma \in \mathcal{F}} u_\sigma \sum_{K \in \mathcal{M}_\sigma} \chi_{K,\sigma}(\boldsymbol{x}) \beta_\sigma(\boldsymbol{x}) - \sum_{K \in \mathcal{M}} \beta_K(\boldsymbol{x}) u_K \sum_{\sigma \in \mathcal{F}_K} \chi_{K,\sigma}(\boldsymbol{x}) \\
&\quad = u_{\sigma_{\boldsymbol{x}}}
\end{aligned}$$

where $\sigma_{\boldsymbol{x}}$ is the unique boundary edge that contains $\boldsymbol{x}$. We have $\mathbb{T}_{\mathfrak{T}} u(\boldsymbol{x}) = u_{\sigma_{\boldsymbol{x}}}$ and thus

$$|\mathbb{T}_{\mathfrak{T}} u(\boldsymbol{x})| = \left| \sum_{K \in \mathcal{M}} \sum_{\sigma \in \mathcal{F}_K} \chi_{K,\sigma}(\boldsymbol{x})(\beta_\sigma(\boldsymbol{x}) u_\sigma - \beta_K(\boldsymbol{x}) u_K) \right|$$

$$= \left| \sum_{K\in\mathcal{M}} \sum_{\sigma\in\mathcal{F}_K} \chi_{K,\sigma}(\boldsymbol{x}) \Big[\beta_\sigma(\boldsymbol{x})(u_\sigma - u_K) + (\beta_\sigma(\boldsymbol{x}) - \beta_K(\boldsymbol{x}))u_K \Big] \right|$$
$$\leq \sum_{K\in\mathcal{M}} \sum_{\sigma\in\mathcal{F}_K} |\chi_{K,\sigma}(\boldsymbol{x})| \Big[\beta_\sigma(\boldsymbol{x})|u_\sigma - u_K| + |\beta_\sigma(\boldsymbol{x}) - \beta_K(\boldsymbol{x})||u_K| \Big].$$

Integrating over τ_i gives

$$\|\mathbb{T}_{\mathfrak{T}} u\|_{L^1(\tau_i)} \leq \sum_{K\in\mathcal{M}} \sum_{\sigma\in\mathcal{F}_K} |u_\sigma - u_K| \int_{\tau_i} |\chi_{K,\sigma}(\boldsymbol{x})|\beta_\sigma(\boldsymbol{x})\mathrm{d}\gamma(\boldsymbol{x})$$
$$+ \sum_{K\in\mathcal{M}} |u_K| \sum_{\sigma\in\mathcal{F}_K} \int_{\tau_i} |\chi_{K,\sigma}(\boldsymbol{x})|\,|\beta_\sigma(\boldsymbol{x}) - \beta_K(\boldsymbol{x})|\mathrm{d}\gamma(\boldsymbol{x}). \tag{B.54}$$

For any $\boldsymbol{x} \in \tau_i$ such that $|\chi_{K,\sigma}(\boldsymbol{x})| > 0$, there exists $\boldsymbol{y} \in \sigma$ such that $\boldsymbol{x} \in D(\boldsymbol{y}, -\boldsymbol{\xi}_i)$. The measure of $\{\boldsymbol{x} \in \tau_i : |\chi_{K,\sigma}(\boldsymbol{x})| > 0\}$ is thus bounded by the measure of the trace on τ_i of the cylinder $\mathcal{C}(\sigma, -\boldsymbol{\xi}_i)$. This measure is less that $|\sigma|/|\widehat{\boldsymbol{\xi}_i} \cdot \boldsymbol{n}_{\tau_i}|$, where $\widehat{\boldsymbol{\xi}_i} = \boldsymbol{\xi}_i/|\boldsymbol{\xi}_i|$. Since $|\boldsymbol{\xi}_i \cdot \boldsymbol{n}_{\tau_i}| \geq \alpha|\boldsymbol{\xi}_i|$, we have $|\sigma|/|\widehat{\boldsymbol{\xi}_i} \cdot \boldsymbol{n}_{\tau_i}| \leq |\sigma|/\alpha$. Hence, using $\beta_\sigma(\boldsymbol{x}) \leq 1$,

$$\int_{\tau_i} |\chi_{K,\sigma}(\boldsymbol{x})|\beta_\sigma(\boldsymbol{x})\mathrm{d}\gamma(\boldsymbol{x}) \leq \frac{|\sigma|}{\alpha}. \tag{B.55}$$

Noticing that $|\beta_\sigma(\boldsymbol{x}) - \beta_K(\boldsymbol{x})| \leq \frac{|(\boldsymbol{y}_\sigma(\boldsymbol{x}) - \boldsymbol{x}_K)\cdot\boldsymbol{\xi}_i|}{|\boldsymbol{\xi}_i|^2} \leq \frac{h_K}{|\boldsymbol{\xi}_i|} \leq \frac{\varrho d_{K,\sigma}}{|\boldsymbol{\xi}_i|}$, we also have

$$\int_{\tau_i} |\chi_{K,\sigma}(\boldsymbol{x})|\,|\beta_\sigma(\boldsymbol{x}) - \beta_K(\boldsymbol{x})|\mathrm{d}\gamma(\boldsymbol{x}) \leq \frac{|\sigma|}{\alpha}\frac{\varrho d_{K,\sigma}}{|\boldsymbol{\xi}_i|}. \tag{B.56}$$

Plugging (B.55) and (B.56) into (B.54), and recalling (B.1), provides C_{11} depending only on α, ϱ, $\boldsymbol{\xi}_i$ and d such that

$$\|\mathbb{T}_{\mathfrak{T}} u\|_{L^1(\tau_i)} \leq C_{11}(|u|_{\mathfrak{T},1} + \|\Pi_{\mathfrak{T}} u\|_{L^1(\Omega)}).$$

The trace inequality (B.51) follows by summing these estimates over $i = 1, \ldots, M$. ∎

Lemma B.21 (Discrete trace inequality, $p > 1$). *Let $p \in (1, +\infty)$, $\mathfrak{T}$ be a polytopal mesh of Ω in the sense of Definition 7.2, and $\varrho \geq \theta_{\mathfrak{T}} + \eta_{\mathfrak{T}}$ (see (7.8) and (7.9)). Then, there exists $C_{12} > 0$, depending only on Ω, d, p and ϱ, such that*

$$\forall u \in X_{\mathfrak{T}}\,,\ \|\mathbb{T}_{\mathfrak{T}} u\|^p_{L^p(\partial\Omega)} \leq C_{12}\Big[|u|_{\mathfrak{T},p}\, \|\Pi_{\mathfrak{T}} u\|^{p-1}_{L^p(\Omega)} + \|\Pi_{\mathfrak{T}} u\|^p_{L^p(\Omega)} + |u|^p_{\mathfrak{T},p}\, h^{p-1}_{\mathcal{M}} \Big]. \tag{B.57}$$

As a consequence, there exists $C_{13} > 0$, depending only on Ω, d, p and ϱ, such that

$$\forall u \in X_{\mathfrak{T}}\,,\ \|\mathbb{T}_{\mathfrak{T}} u\|_{L^p(\partial\Omega)} \leq C_{13}\left(|u|_{\mathfrak{T},p} + \|\Pi_{\mathfrak{T}} u\|_{L^p(\Omega)} \right). \tag{B.58}$$

Proof. To deduce (B.58) from (B.57), start by using in the latter estimate the bounds $h_{\mathcal{M}} \le \operatorname{diam}(\Omega)$ and, owing to Young's inequality (D.8),

$$|u|_{\mathfrak{T},p} \, \|\Pi_{\mathfrak{T}} u\|_{L^p(\Omega)}^{p-1} \le \frac{1}{p} \, |u|_{\mathfrak{T},p}^p + \frac{1}{p'} \, \|\Pi_{\mathfrak{T}} u\|_{L^p(\Omega)}^p \, .$$

Take then the power $1/p$ of the resulting inequality and conclude by applying the power-of-sums estimate (D.12) with $\alpha = p$.
Let us now prove (B.57). Take $u \in X_{\mathfrak{T}}$ and, in a similar way as in the proof of Lemma B.14, apply (B.51) in Lemma B.20 to $\widehat{u} = ((|u_K|^p)_{K\in\mathcal{M}}, (\widehat{u}_\sigma)_{\sigma\in\mathcal{F}})$ with

$$\begin{aligned} &\widehat{u}_\sigma = \frac{1}{2}(|u_K|^p + |u_L|^p) && \text{if } \sigma \in \mathcal{F}_{\text{int}} \text{ with } \mathcal{M}_\sigma = \{K, L\}, \\ &\widehat{u}_\sigma = |u_\sigma|^p && \text{if } \sigma \in \mathcal{F}_{\text{ext}}. \end{aligned}$$

Since $\Pi_{\mathfrak{T}} \widehat{u} = |\Pi_{\mathfrak{T}} u|^p$ and $\mathbb{T}_{\mathfrak{T}} \widehat{u} = |\mathbb{T}_{\mathfrak{T}} u|^p$, this gives

$$\|\mathbb{T}_{\mathfrak{T}} u\|_{L^p(\partial\Omega)}^p \le C_{10}(|\widehat{u}|_{\mathfrak{T},1} + \|\Pi_{\mathfrak{T}} u\|_{L^p(\Omega)}^p). \tag{B.59}$$

Suppose that we establish the existence of C_{14}, depending only on Ω, d, p and ϱ, such that

$$|\widehat{u}|_{\mathfrak{T},1} \le C_{14} \, |u|_{\mathfrak{T},p} \left(h_{\mathcal{M}}^{\frac{1}{p'}} \, \|\mathbb{T}_{\mathfrak{T}} u\|_{L^p(\partial\Omega)}^{p-1} + \|\Pi_{\mathfrak{T}} u\|_{L^p(\Omega)}^{p-1} \right). \tag{B.60}$$

Then, by Young's inequality (D.9),

$$\begin{aligned} &|\widehat{u}|_{\mathfrak{T},1} \\ &\le C_{14} \left(\frac{1}{p\varepsilon^{p/p'}} \, |u|_{\mathfrak{T},p}^p \, h_{\mathcal{M}}^{p-1} + \frac{\varepsilon}{p'} \, \|\mathbb{T}_{\mathfrak{T}} u\|_{L^p(\partial\Omega)}^p + |u|_{\mathfrak{T},p} \, \|\Pi_{\mathfrak{T}} u\|_{L^p(\Omega)}^{p-1} \right). \end{aligned} \tag{B.61}$$

Taking $\varepsilon > 0$ such that $C_{10} C_{14} \frac{\varepsilon}{p'} = \frac{1}{2}$ and plugging the result in (B.59) gives (B.57).
Let us now prove (B.60). If $\mathcal{M}_\sigma = \{K, L\}$, owing to (B.41),

$$|\widehat{u}_K - \widehat{u}_\sigma| = \frac{1}{2} \big| |u_K|^p - |u_L|^p \big| \le \frac{p}{2}(|u_K|^{p-1} + |u_L|^{p-1})|u_K - u_L|.$$

Similarly, if $\mathcal{M}_\sigma = \{K\}$,

$$|\widehat{u}_K - \widehat{u}_\sigma| = \big| \, |u_K|^p - |u_\sigma|^p \big| \le p(|u_K|^{p-1} + |u_\sigma|^{p-1})|u_K - u_\sigma|.$$

Hence, setting $\delta_\sigma u = |u_K - u_\sigma|$ if $\mathcal{M}_\sigma = \{K\}$ and $\delta_\sigma u = |u_K - u_L|$ if $\mathcal{M}_\sigma = \{K, L\}$,

$$|\widehat{u}|_{\mathfrak{T},1} \le p \sum_{\sigma\in\mathcal{F}_{\text{ext}}} |\sigma| \, |u_\sigma|^{p-1} \delta_\sigma u + p \sum_{K\in\mathcal{M}} \sum_{\sigma\in\mathcal{F}_K} |\sigma| \, |u_K|^{p-1} \delta_\sigma u. \tag{B.62}$$

Let $w_i, F_i, G_i \geq 0$ and $H_i > 0$. Applying the Hölder inequality (D.4) to $a_i = G_i$, $b_i = F_i^{p-1}$ and $d_i = H_i^{-(p-1)/p} = H_i^{-1/p'}$, we find

$$\sum_{i\in I} w_i F_i^{p-1} G_i \leq \left(\sum_{i\in I} w_i \frac{G_i^p}{H_i^{p-1}}\right)^{\frac{1}{p}} \left(\sum_{i\in I} w_i H_i F_i^p\right)^{\frac{p-1}{p}}.$$

Applied with $w_i = |\sigma|$, $H_i = 1$, $F_i = |u_\sigma|$ and $G_i = \delta_\sigma u$ in the first term of (B.62), and with $w_i = |\sigma|$, $H_i = d_{K,\sigma}$, $F_i = |u_K|$ and $G_i = \delta_\sigma u$ in the second term of (B.62), this gives

$$\begin{aligned}
|\widehat{u}|_{\mathfrak{T},1} &\leq p\left(\sum_{\sigma\in\mathcal{F}_{\text{ext}}} |\sigma|(\delta_\sigma u)^p\right)^{\frac{1}{p}} \left(\sum_{\sigma\in\mathcal{F}_{\text{ext}}} |\sigma|\,|u_\sigma|^p\right)^{\frac{p-1}{p}} \\
&\quad + p\left(\sum_{K\in\mathcal{M}}\sum_{\sigma\in\mathcal{F}_K} |\sigma|\frac{(\delta_\sigma u)^p}{d_{K,\sigma}^{p-1}}\right)^{\frac{1}{p}} \left(\sum_{K\in\mathcal{M}}\sum_{\sigma\in\mathcal{F}_K} |\sigma| d_{K,\sigma}|u_K|^p\right)^{\frac{p-1}{p}} \\
&= T_1 + T_2. \qquad \text{(B.63)}
\end{aligned}$$

For $\sigma \in \mathcal{F}_{\text{ext}}$ with $\mathcal{M}_\sigma = \{K\}$, write $(\delta_\sigma u)^p = d_{K,\sigma}^p \frac{(\delta_\sigma u)^p}{d_{K,\sigma}^p} \leq h_{\mathcal{M}}^{p-1} d_{K,\sigma} \frac{(\delta_\sigma u)^p}{d_{K,\sigma}^p}$ to obtain

$$T_1 \leq p h_{\mathcal{M}}^{\frac{1}{p'}} |u|_{\mathfrak{T},p} \|\mathbb{T}_{\mathfrak{T}} u\|_{L^p(\partial\Omega)}^{p-1}. \qquad \text{(B.64)}$$

To estimate T_2, first use the triangle inequality to write, if $\sigma \in \mathcal{F}_{\text{ext}}$ with $\mathcal{M}_\sigma = \{K, L\}$, $\delta_\sigma u \leq |u_K - u_\sigma| + |u_L - u_\sigma|$. Then, by definition of $\varrho \geq \eta_{\mathfrak{T}}$ and invoking the power-of-sums inequality (D.12),

$$\frac{(\delta_\sigma u)^p}{d_{K,\sigma}^{p-1}} \leq 2^{p-1} d_{K,\sigma} \left|\frac{u_K - u_\sigma}{d_{K,\sigma}}\right|^p + 2^{p-1}\varrho^{p-1} d_{L,\sigma} \left|\frac{u_L - u_\sigma}{d_{L,\sigma}}\right|^p.$$

This also holds, dropping the second addend, if $\sigma \in \mathcal{F}_{\text{ext}}$ with $\mathcal{M}_\sigma = \{K\}$. Using this estimate in the first factor in T_2, the term $d_{K,\sigma}|\frac{u_K - u_\sigma}{d_{K,\sigma}}|^p$ appears twice, once with a factor 2^{p-1} and another time with a factor $2^{p-1}\varrho^{p-1}$ (when summing on the faces of the cell L on the other side of K with respect to σ). Hence,

$$\begin{aligned}
\sum_{K\in\mathcal{M}}\sum_{\sigma\in\mathcal{F}_K} |\sigma|\frac{(\delta_\sigma u)^p}{d_{K,\sigma}^{p-1}} &\leq 2^{p-1}(1+\varrho^{p-1}) \sum_{K\in\mathcal{M}}\sum_{\sigma\in\mathcal{F}_K} |\sigma| d_{K,\sigma} \left|\frac{u_K - u_\sigma}{d_{K,\sigma}}\right|^p \\
&= 2^{p-1}(1+\varrho^{p-1})\,|u|_{\mathfrak{T},p}^p.
\end{aligned}$$

Invoke then (B.1) to re-write the second factor in T_2 and obtain

$$T_2 \leq p(2d)^{\frac{1}{p'}}(1+\varrho^{p-1})^{\frac{1}{p}}\,|u|_{\mathfrak{T},p}\,\|\Pi_{\mathfrak{T}} u\|_{L^p(\Omega)}^{p-1}. \qquad \text{(B.65)}$$

Estimates (B.63), (B.64) and (B.65) complete the proof of (B.60). ■

The following lemma is particularly useful when dealing with Fourier boundary conditions.

Lemma B.22. *Let $p \in [1, +\infty)$, $\mathfrak{T}$ be a polytopal mesh of Ω in the sense of Definition 7.2 and $\varrho \geq \theta_{\mathfrak{T}} + \eta_{\mathfrak{T}}$. Then, there exists $C_{15} > 0$, depending only on Ω, d, p and ϱ, such that*

$$\forall u \in X_{\mathfrak{T}},\ \|\Pi_{\mathfrak{T}} u\|_{L^p(\Omega)} \leq C_{15} \left(|u|_{\mathfrak{T},p} + \|\mathbb{T}_{\mathfrak{T}} u\|_{L^p(\partial\Omega)} \right).$$

Proof. Let $\mathbf{e}$ be a unit vector (say, for example, corresponding to the first coordinate in $\mathbb{R}^d$). As in the proof of Lemma B.20, define $\chi_{K,\sigma} : \Omega \to \{-1, 0, +1\}$ by $\chi_{K,\sigma}(\boldsymbol{x}) = \mathrm{sgn}(\mathbf{e} \cdot \boldsymbol{n}_{K,\sigma})$ if the half-line $D(\boldsymbol{x}, \mathbf{e}) = \boldsymbol{x} + \mathbb{R}^+ \mathbf{e}$ intersects σ at one point, and $\chi_{K,\sigma}(\boldsymbol{x}) = 0$ otherwise. Contrary to the proof of Lemma B.20, $\chi_{K,\sigma}(\boldsymbol{x})$ is here defined for all $\boldsymbol{x} \in \Omega$. Since $\chi_{K,\sigma}$ is non-zero (and equal to ± 1) only in the cylinder with base σ and axis $\mathbf{e}$,

$$\int_\Omega |\chi_{K,\sigma}(\boldsymbol{x})| \mathrm{d}\boldsymbol{x} \leq |\sigma| \mathrm{diam}(\Omega). \tag{B.66}$$

Drawing the half-line $D(\boldsymbol{x}, \mathbf{e})$ and writing $\Pi_{\mathfrak{T}} u(\boldsymbol{x})$ as the sum of jumps between $\boldsymbol{x}$ and the faces $\sigma \in \mathcal{F}$ that intersect $D(\boldsymbol{x}, \mathbf{e})$ leads to

$$\Pi_{\mathfrak{T}} u(\boldsymbol{x}) = \sum_{K \in \mathcal{M}} \sum_{\sigma \in \mathcal{F}_K} \chi_{K,\sigma}(\boldsymbol{x})(u_K - u_\sigma) + \sum_{\sigma \in \mathcal{F}_{\mathrm{ext}},\, \mathcal{M}_\sigma = \{K\}} \chi_{K,\sigma}(\boldsymbol{x}) u_\sigma.$$

Take the absolute value, integrate over $\boldsymbol{x} \in \Omega$ and use (B.66) to deduce

$$\begin{aligned} \|\Pi_{\mathfrak{T}} u\|_{L^1(\Omega)} &\leq \mathrm{diam}(\Omega) \sum_{K \in \mathcal{M}} \sum_{\sigma \in \mathcal{F}_K} |\sigma| \, |u_K - u_\sigma| + \mathrm{diam}(\Omega) \sum_{\sigma \in \mathcal{F}_{\mathrm{ext}}} |\sigma| \, |u_\sigma| \\ &= \mathrm{diam}(\Omega)(|u|_{\mathfrak{T},1} + \|\mathbb{T}_{\mathfrak{T}} u\|_{L^1(\partial\Omega)}). \end{aligned} \tag{B.67}$$

This concludes the proof in the case $p = 1$.
In the case $p > 1$, apply (B.67) to $\widehat{u}$ defined as in the proof of Lemma B.21. Using the estimate (B.60) on $|\widehat{u}|_{\mathfrak{T},p}$ and writing $h_{\mathcal{M}} \leq \mathrm{diam}(\Omega)$ then yields

$$\|\Pi_{\mathfrak{T}} u\|^p_{L^p(\Omega)} \leq C_{16} \left(|u|_{\mathfrak{T},p} \|\mathbb{T}_{\mathfrak{T}} u\|^{p-1}_{L^p(\partial\Omega)} + |u|_{\mathfrak{T},p} \|\Pi_{\mathfrak{T}} u\|^{p-1}_{L^p(\Omega)} + \|\mathbb{T}_{\mathfrak{T}} u\|^p_{L^p(\partial\Omega)} \right)$$

where C_{16} depends only on Ω, d, p and ϱ. The proof is concluded by using Young's inequalities (D.8) and (D.9) to write

$$\begin{aligned} |u|_{\mathfrak{T},p} \|\mathbb{T}_{\mathfrak{T}} u\|^{p-1}_{L^p(\partial\Omega)} &\leq \frac{1}{p} |u|^p_{\mathfrak{T},p} + \frac{1}{p'} \|\mathbb{T}_{\mathfrak{T}} u\|^p_{L^p(\partial\Omega)}, \\ |u|_{\mathfrak{T},p} \|\Pi_{\mathfrak{T}} u\|^{p-1}_{L^p(\Omega)} &\leq \frac{1}{p \varepsilon^{p/p'}} |u|^p_{\mathfrak{T},p} + \frac{\varepsilon}{p'} \|\Pi_{\mathfrak{T}} u\|^p_{L^p(\Omega)}, \end{aligned}$$

by choosing ε such that $C_{16} \frac{\varepsilon}{p'} = \frac{1}{2}$, and by using the power-of-sums estimate (D.12) with $\alpha = p$. ∎

B.3.2 Discrete Sobolev embeddings

Lemma B.23 (Discrete embedding of $W^{1,1}(\Omega)$, with zero average, in $L^{1^\star}(\Omega)$). *Let $\mathfrak{T}$ be a polytopal mesh of Ω in the sense of Definition 7.2, and recall the notations (7.7). There exists C_{17} depending only on Ω and d such that*

$$\forall u \in X_{\mathfrak{T}}\,,\ \left\|\Pi_{\mathfrak{T}} u - \overline{\Pi_{\mathfrak{T}} u}\right\|_{L^{1^\star}(\Omega)} \leq C_{17}\,|u|_{\mathfrak{T},1}\,, \tag{B.68}$$

where $1^\star = \frac{d}{d-1}$ and $\overline{\Pi_{\mathfrak{T}} u} = \frac{1}{|\Omega|}\int_\Omega \Pi_{\mathfrak{T}} u(\boldsymbol{x})\mathrm{d}\boldsymbol{x}$.

Proof. The Sobolev embedding and the Poincaré-Wirtinger inequality show that $\|w - \overline{w}\|_{L^{1^\star}(\Omega)} \leq C_{18}\,\|\nabla w\|_{L^1(\Omega)^d}$ for all $w \in W^{1,1}(\Omega)$, where C_{18} depends only on Ω. By approximating $\Pi_{\mathfrak{T}} u$, strongly in $L^1(\Omega)$ and weakly in $BV(\Omega)$, by functions in $W^{1,1}(\Omega)$, the "mean" Nirenberg inequality can be deduced:

$$\left\|\Pi_{\mathfrak{T}} u - \overline{\Pi_{\mathfrak{T}} u}\right\|_{L^{1^\star}(\Omega)} \leq C_{18}\,|\Pi_{\mathfrak{T}} u|_{BV(\Omega)}\,, \tag{B.69}$$

where

$$|w|_{BV(\Omega)} = \sup\left\{\int_\Omega w(\boldsymbol{x})\mathrm{div}\boldsymbol{\varphi}(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\ :\ \boldsymbol{\varphi} \in C_c^\infty(\Omega,\mathbb{R}^d), \|\boldsymbol{\varphi}\|_{L^\infty(\Omega)^d} \leq 1\right\}.$$

Write (B.32) with $v = u$. The integral term on $\partial\Omega$ can be dropped since $\boldsymbol{\varphi}$ vanishes on the boundary. Reason then as in (B.38) in Lemma B.13 to obtain $|\Pi_{\mathfrak{T}} u|_{BV(\Omega)} \leq \sqrt{d}\,|u|_{\mathfrak{T},1}$, and the conclusion follows from (B.69). ∎

Lemma B.24 (Discrete embedding of $W^{1,p}(\Omega)$, with zero average, in $L^{p^\star}(\Omega)$, $1 < p < d$). *Let $\mathfrak{T}$ be a polytopal mesh of Ω in the sense of Definition 7.2. Let $p \in (1,d)$ and $\varrho \geq \theta_{\mathfrak{T}} + \eta_{\mathfrak{T}}$. Then, there exists C_{19}, depending only on Ω, d, p and ϱ, such that*

$$\forall u \in X_{\mathfrak{T}}\,,\ \left\|\Pi_{\mathfrak{T}} u - \overline{\Pi_{\mathfrak{T}} u}\right\|_{L^{p^\star}(\Omega)} \leq C_{19}\,|u|_{\mathfrak{T},p}\,,$$

where $p^\star = \frac{pd}{d-p}$ and $\overline{\Pi_{\mathfrak{T}} u} = \frac{1}{|\Omega|}\int_\Omega \Pi_{\mathfrak{T}} u(\boldsymbol{x})\mathrm{d}\boldsymbol{x}$.

Proof. Let $u \in X_{\mathfrak{T}}$. Upon translating by $\overline{\Pi_{\mathfrak{T}} u}$ all the values of $u = ((u_K)_{K\in\mathcal{M}}, (u_\sigma)_{\sigma\in\mathcal{F}})$, which does not change $|u|_{\mathfrak{T},p}$, we can assume that $\overline{\Pi_{\mathfrak{T}} u} = 0$. In the following, $A \lesssim B$ means that $A \leq MB$ with M depending only on Ω, d, p and ϱ.
Let $\alpha > 1$ and consider $\widehat{u} = ((|u_K|^\alpha)_{K\in\mathcal{M}}, (\widehat{u}_\sigma)_{\sigma\in\mathcal{F}})$ with

$$\widehat{u}_\sigma = \frac{1}{2}(|u_K|^\alpha + |u_L|^\alpha) \quad \text{if } \mathcal{M}_\sigma = \{K,L\},$$
$$\widehat{u}_\sigma = |u_K|^\alpha \quad \text{if } \mathcal{M}_\sigma = \{K\}.$$

Since $|\overline{\Pi_{\mathfrak{T}}\widehat{u}}| \leq \frac{1}{|\Omega|}\,\|\Pi_{\mathfrak{T}} u\|^\alpha_{L^\alpha(\Omega)}$, Inequality (B.68) applied to $\widehat{u}$ yields

$$\begin{aligned}\|\Pi_{\mathfrak{T}} u\|^{\alpha}_{L^{\alpha 1^\star}(\Omega)} &= \|\Pi_{\mathfrak{T}} \widehat{u}\|_{L^{1^\star}(\Omega)} \\ &\leq \left\|\Pi_{\mathfrak{T}} \widehat{u} - \overline{\Pi_{\mathfrak{T}} \widehat{u}}\right\|_{L^{1^\star}(\Omega)} + |\Omega|^{\frac{1}{1^\star}} |\overline{\Pi_{\mathfrak{T}} \widehat{u}}| \\ &\lesssim |\widehat{u}|_{\mathfrak{T},1} + \|\Pi_{\mathfrak{T}} u\|^{\alpha}_{L^{\alpha}(\Omega)} . \end{aligned} \tag{B.70}$$

The definition of $\widehat{u}_\sigma$ ensures that the terms in $|\widehat{u}|_{\mathfrak{T},1}$ corresponding to boundary faces vanish. Hence, for any $r \in (1,\infty)$, a similar reasoning as in the proof of Lemma B.14 (passage from (B.40) to (B.42)) shows that

$$|\widehat{u}|_{\mathfrak{T},1} \lesssim |u|_{\mathfrak{T},r} \left\| |\Pi_{\mathfrak{T}} u|^{\alpha-1} \right\|_{L^{r'}(\Omega)} .$$

Plugging this estimate into (B.70) and taking the power $1/\alpha$ (thanks to the power-of-sums inequality (D.13)) yields

$$\|\Pi_{\mathfrak{T}} u\|_{L^{\alpha 1^\star}(\Omega)} \lesssim |u|^{\frac{1}{\alpha}}_{\mathfrak{T},r} \left\| |\Pi_{\mathfrak{T}} u|^{\alpha-1} \right\|^{\frac{1}{\alpha}}_{L^{r'}(\Omega)} + \|\Pi_{\mathfrak{T}} u\|_{L^{\alpha}(\Omega)} . \tag{B.71}$$

Take $r > 1$ such that $(\alpha-1)r' = \alpha 1^\star$ (since $\alpha 1^\star/(\alpha-1) > 1^\star > 1$, this defines $r' \in (1,\infty)$ and thus $r \in (1,\infty)$). This choice gives

$$\left\| |\Pi_{\mathfrak{T}} u|^{\alpha-1} \right\|^{\frac{1}{\alpha}}_{L^{r'}(\Omega)} = \|\Pi_{\mathfrak{T}} u\|^{\frac{1}{\alpha'}}_{L^{\alpha 1^\star}(\Omega)} .$$

Use Young's inequality (D.9) with exponent α, and ε small enough (depending only on the constants hidden in $\lesssim$), to deduce from (B.71) that

$$\|\Pi_{\mathfrak{T}} u\|_{L^{\alpha 1^\star}(\Omega)} \lesssim |u|_{\mathfrak{T},r} + \|\Pi_{\mathfrak{T}} u\|_{L^{\alpha}(\Omega)} .$$

If $r \leq p$, that is if $r' = \frac{\alpha 1^\star}{\alpha-1} \geq p'$, then (B.45) shows that

$$\|\Pi_{\mathfrak{T}} u\|_{L^{\alpha 1^\star}(\Omega)} \lesssim |u|_{\mathfrak{T},p} + \|\Pi_{\mathfrak{T}} u\|_{L^{\alpha}(\Omega)} . \tag{B.72}$$

The estimate (B.68) and the fact that $\overline{\Pi_{\mathfrak{T}} u} = 0$ give $\|\Pi_{\mathfrak{T}} u\|_{L^{1^\star}(\Omega)} \lesssim |u|_{\mathfrak{T},1} \lesssim |u|_{\mathfrak{T},p}$. An induction based on (B.72) applied with $\alpha = 1^\star$, $(1^\star)^2$, ... then establishes that, for any $k \in \mathbb{N}$ such that $\frac{(1^\star)^{k+1}}{(1^\star)^k - 1} \geq p'$,

$$\|\Pi_{\mathfrak{T}} u\|_{L^{(1^\star)^{k+1}}(\Omega)} \lesssim |u|_{\mathfrak{T},p} . \tag{B.73}$$

Select k as the largest integer such that $\frac{(1^\star)^{k+1}}{(1^\star)^k - 1} \geq p'$. Such a k exists since $k = 0$ satisfies this inequality and, as $k \to \infty$, $\frac{(1^\star)^{k+1}}{(1^\star)^k - 1} \to 1^\star = d' > p'$ (we have $p < d$). Let $\alpha = \frac{p^\star}{1^\star} > 1$ and assume that

$$\frac{\alpha 1^\star}{\alpha - 1} = \frac{1^\star p^\star}{p^\star - 1^\star} \geq p' \tag{B.74}$$

and

$$\alpha = \frac{p^\star}{1^\star} \leq (1^\star)^{k+1}. \tag{B.75}$$

Inequality (B.74) allows us to apply (B.72), which gives

$$\|\Pi_{\mathfrak{T}} u\|_{L^{p^\star}(\Omega)} \lesssim |u|_{\mathfrak{T},p} + \|\Pi_{\mathfrak{T}} u\|_{L^{\alpha}(\Omega)}.$$

By (B.75), $\|\Pi_{\mathfrak{T}} u\|_{L^{\alpha}(\Omega)} \lesssim \|\Pi_{\mathfrak{T}} u\|_{L^{(1^\star)^{k+1}}(\Omega)}$ and (B.73) then concludes the proof.

It remains to check (B.74) and (B.75). We have $p = \frac{dp^\star}{d+p^\star}$ so (B.74) boils down to $\frac{1^\star p^\star}{p^\star - 1^\star} \geq \frac{dp^\star}{dp^\star - d - p^\star}$, that is to say $1^\star(dp^\star - d - p^\star) \geq d(p^\star - 1^\star)$, or $1^\star(d-1)p^\star \geq dp^\star$. This last relation is obvious since $1^\star(d-1) = d$ (we thus even have equality in (B.74)). To check (B.75), we start by writing that, by definition of k, $\frac{(1^\star)^{k+2}}{(1^\star)^{k+1}-1} \leq p'$, which can be recast as $1 - \frac{1}{p} = \frac{1}{p'} \leq \frac{1}{1^\star} - \frac{1}{(1^\star)^{k+2}}$. But $\frac{1}{p^\star} = \frac{1}{p} - \frac{1}{d}$ and $\frac{1}{1^\star} = 1 - \frac{1}{d}$, so

$$\frac{1}{p^\star} \geq 1 - \frac{1}{1^\star} + \frac{1}{(1^\star)^{k+2}} - \frac{1}{d} = \frac{1}{(1^\star)^{k+2}},$$

which is equivalent to (B.75). ∎

The proof of the following lemma is similar to the proof of Lemma B.15, using Lemmas B.23 and B.24.

Lemma B.25 (Discrete embedding of $W^{1,p}(\Omega)$, with zero average, in $L^q(\Omega)$, for some $q > p$). *Let $p \in [1, +\infty)$, $\mathfrak{T}$ be a polytopal mesh of Ω in the sense of Definition 7.2, and $\varrho \geq \theta_{\mathfrak{T}} + \eta_{\mathfrak{T}}$. Then, there exists $q > p$, depending only on d and p, and there exists C_{20}, depending only on Ω, d, p and ϱ, such that*

$$\forall u \in X_{\mathfrak{T}}\,,\ \left\|\Pi_{\mathfrak{T}} u - \overline{\Pi_{\mathfrak{T}} u}\right\|_{L^q(\Omega)} \leq C_{20}\, |u|_{\mathfrak{T},p}\,,$$

where $\overline{\Pi_{\mathfrak{T}} u} = \frac{1}{|\Omega|}\int_\Omega \Pi_{\mathfrak{T}} u(\boldsymbol{x})\mathrm{d}\boldsymbol{x}$.

If $p < d$ we can take $q = p^\star = \frac{pd}{d-p}$ and, if $p \geq d$, we can take any $q < +\infty$.

B.3.3 Compactness of $\Pi_{\mathfrak{T}}$ and $\mathbb{T}_{\mathfrak{T}}$

Lemma B.26. *Let $\mathfrak{T}$ be a polytopal mesh of Ω in the sense of Definition 7.2, and $\varrho \geq \theta_{\mathfrak{T}} + \eta_{\mathfrak{T}}$. Then, there exists C_{21}, depending only on Ω and ϱ, such that*

$$\forall u \in X_{\mathfrak{T}}\,,\ \forall \boldsymbol{h} \in \mathbb{R}^d\,,\ \|\Pi_{\mathfrak{T}} u(\cdot + \boldsymbol{h}) - \Pi_{\mathfrak{T}} u\|_{L^1(\mathbb{R}^d)} \leq |\boldsymbol{h}| C_{21}(|u|_{\mathfrak{T},1} + |\overline{\Pi_{\mathfrak{T}} u}|),$$

where $\Pi_{\mathfrak{T}} u$ has been extended by 0 outside Ω, and $\overline{\Pi_{\mathfrak{T}} u} = \frac{1}{|\Omega|}\int_\Omega \Pi_{\mathfrak{T}} u(\boldsymbol{x})\mathrm{d}\boldsymbol{x}$.

Proof. Writing (B.32) with $v = u$ yields, for any $\boldsymbol{\varphi} \in C_c^\infty(\mathbb{R}^d, \mathbb{R}^d)$ such that $\|\boldsymbol{\varphi}\|_{L^\infty(\mathbb{R}^d)^d} \leq 1$ (so that $|\boldsymbol{\varphi}| \leq \sqrt{d}$),

$$\int_{\mathbb{R}^d} \Pi_{\mathfrak{T}} u(\boldsymbol{x}) \mathrm{div}\boldsymbol{\varphi}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}$$
$$\leq \sqrt{d} \sum_{K\in\mathcal{M}} \sum_{\sigma\in\mathcal{F}_K} |\sigma||u_K - u_\sigma| + \sqrt{d} \int_{\partial\Omega} |\mathbb{T}_{\mathfrak{T}} u(\boldsymbol{x})| \mathrm{d}\gamma(\boldsymbol{x})$$
$$\leq \sqrt{d}\, |u|_{\mathfrak{T},1} + \sqrt{d}\, \|\mathbb{T}_{\mathfrak{T}} u\|_{L^1(\partial\Omega)}.$$

Hence,

$$\|\Pi_{\mathfrak{T}} u\|_{BV(\mathbb{R}^d)} \leq \sqrt{d}\, |u|_{\mathfrak{T},1} + \sqrt{d}\, \|\mathbb{T}_{\mathfrak{T}} u\|_{L^1(\partial\Omega)}.$$

Lemma B.20 and B.23 then provide C_{22} depending only on Ω, d and ϱ such that

$$\|\Pi_{\mathfrak{T}} u\|_{BV(\mathbb{R}^d)} \leq C_{22}(|u|_{\mathfrak{T},1} + |\overline{\Pi_{\mathfrak{T}} u}|).$$

The inequality (B.49) concludes the proof. ∎

We can now state the compactness of the function and trace reconstructions.

Lemma B.27 (Discrete Rellich theorem and compactness of the trace from a bound on the mean value). *Let $(\mathfrak{T}_m)_{m\in\mathbb{N}}$ be a sequence of polytopal meshes of Ω and $p \in [1, +\infty)$. Assume that $\sup_{m\in\mathbb{N}}(\theta_{\mathfrak{T}_m} + \eta_{\mathfrak{T}_m}) < +\infty$. Then, for any $u_m \in X_{\mathfrak{T}_m}$ such that $(|u_m|_{\mathfrak{T}_m,p})_{m\in\mathbb{N}}$ and $(\int_\Omega \Pi_{\mathfrak{T}_m} u_m(\boldsymbol{x}) \mathrm{d}\boldsymbol{x})_{m\in\mathbb{N}}$ are bounded, the sequence $(\Pi_{\mathfrak{T}_m} u_m)_{m\in\mathbb{N}}$ is relatively compact in $L^p(\Omega)$.*
Moreover, if $p > 1$ and $h_{\mathcal{M}_m} \to 0$ as $m \to \infty$, then the sequence $(\mathbb{T}_{\mathfrak{T}_m} u_m)_{m\in\mathbb{N}}$ is relatively compact in $L^p(\partial\Omega)$.

Proof. The relative compactness of $(\Pi_{\mathfrak{T}_m} u_m)_{m\in\mathbb{N}}$ follows from Lemmas B.25 and B.26 in a similar way as for Dirichlet boundary conditions. We now assume that $p > 1$ and $h_{\mathcal{M}_m} \to 0$, and we establish the relative compactness of the traces. By (B.58) in Lemma B.21 and the boundedness of $(\Pi_{\mathfrak{T}_m} u_m)_{m\in\mathbb{N}}$ in $L^p(\Omega)$, $(\mathbb{T}_{\mathfrak{T}_m} u_m)_{m\in\mathbb{N}}$ is bounded in $L^p(\partial\Omega)$. The estimate (B.29) on $\overline{\nabla}_{\mathfrak{T}_m}$ and the boundedness of $(|u_m|_{\mathfrak{T}_m,p})_{m\in\mathbb{N}}$ show that $(\overline{\nabla}_{\mathfrak{T}_m} u_m)_{m\in\mathbb{N}}$ is bounded in $L^p(\Omega)^d$. Upon extracting subsequences, we can therefore assume that there exists $\psi \in L^p(\Omega)$, $\chi \in L^p(\partial\Omega)$ and $\boldsymbol{\xi} \in L^p(\Omega)^d$ such that $\Pi_{\mathfrak{T}_m} u_m \to \psi$ strongly in $L^p(\Omega)$, $\mathbb{T}_{\mathfrak{T}_m} u_m \to \chi$ weakly in $L^p(\partial\Omega)$ and $\overline{\nabla}_{\mathfrak{T}_m} u_m \to \boldsymbol{\xi}$ weakly in $L^p(\Omega)^d$.

Using the same ideas as for Lemma 2.15 (regularity of the limit), we analyse ψ. Take $\boldsymbol{\varphi} \in C^\infty(\overline{\Omega})^d$ and apply the discrete Stokes formula (B.31) to $\mathfrak{T} = \mathfrak{T}_m$ and $v = u_m$. The aforementioned convergences enable us to pass to the limit $m \to \infty$ to see that

$$\int_\Omega (\boldsymbol{\xi}(\boldsymbol{x}) \cdot \boldsymbol{\varphi}(\boldsymbol{x}) + \psi(\boldsymbol{x}) \mathrm{div}\boldsymbol{\varphi}(\boldsymbol{x}))\, \mathrm{d}\boldsymbol{x} - \int_{\partial\Omega} \chi(\boldsymbol{x}) \gamma_{\mathbf{n}} \boldsymbol{\varphi}(\boldsymbol{x}) \mathrm{d}\gamma(\boldsymbol{x}) = 0. \quad \text{(B.76)}$$

Applied to $\boldsymbol{\varphi} \in C_c^\infty(\Omega)^d$ this shows that $\boldsymbol{\xi} = \nabla\psi$, and thus that $\psi \in W^{1,p}(\Omega)$. Using then an integration-by-parts in (B.76) with a generic $\boldsymbol{\varphi} \in C^\infty(\overline{\Omega})^d$ shows that $\gamma\psi = \chi$.

We now prove that $\mathbb{T}_{\mathfrak{T}_m} u_m \to \gamma\psi$ strongly in $L^p(\partial\Omega)$, which will conclude the proof of the lemma. Let $I_{\mathfrak{T}_m} : W^{1,p}(\Omega) \to X_{\mathfrak{T}_m}$ be the interpolant defined by (B.10). Applying (B.57) in Lemma B.21 to $\mathfrak{T} = \mathfrak{T}_m$ and $u = u_m - I_{\mathfrak{T}_m}\psi$, and using the boundedness of $(|u_m|_{\mathfrak{T}_m,p})_{m\in\mathbb{N}}$ and $(|I_{\mathfrak{T}_m}\psi|_{\mathfrak{T}_m,p})_{m\in\mathbb{N}}$ (see (B.22)), we find C_{23} not depending on m such that

$$\begin{aligned}\|\mathbb{T}_{\mathfrak{T}_m} u_m - \mathbb{T}_{\mathfrak{T}_m}(I_{\mathfrak{T}_m}\psi)\|^p_{L^p(\partial\Omega)} &\le C_{23}\,\|\Pi_{\mathfrak{T}_m} u_m - \Pi_{\mathfrak{T}_m}(I_{\mathfrak{T}_m}\psi)\|^{p-1}_{L^p(\Omega)} \\ &\quad + C_{23}\,\|\Pi_{\mathfrak{T}_m} u_m - \Pi_{\mathfrak{T}_m}(I_{\mathfrak{T}_m}\psi)\|^p_{L^p(\Omega)} + C_{23} h^{p-1}_{\mathcal{M}_m}.\end{aligned}$$

Since $p > 1$, $h_{\mathcal{M}_m} \to 0$, and $(\Pi_{\mathfrak{T}_m} u_m)_{m\in\mathbb{N}}$ and $(\Pi_{\mathfrak{T}_m}(I_{\mathfrak{T}_m}\psi))_{m\in\mathbb{N}}$ both converge strongly to ψ in $L^p(\Omega)$ (see (B.26)), the right-hand side of the above inequality tends to 0 as $m \to \infty$. Hence, $\mathbb{T}_{\mathfrak{T}_m} u_m - \mathbb{T}_{\mathfrak{T}_m}(I_{\mathfrak{T}_m}\psi) \to 0$ in $L^p(\partial\Omega)$. The strong convergence of $(\mathbb{T}_{\mathfrak{T}_m} u_m)_{m\in\mathbb{N}}$ follows by using (B.27) to see that $\mathbb{T}_{\mathfrak{T}_m}(I_{\mathfrak{T}_m}\psi) \to \gamma\psi$ in $L^p(\partial\Omega)$ as $m \to \infty$. ∎

B.4 Discrete functional analysis for mixed boundary condition

We consider here that Assumption (7.2) on Ω and Assumption (3.60) on Γ_d and Γ_n hold. If $\mathfrak{T}$ is a polytopal mesh of Ω in the sense of Definition 7.2, we recall the notations in (7.7) and we additionally define

$$\begin{aligned} X_{\mathfrak{T},\Gamma_d} &= \{v \in X_{\mathfrak{T},\partial} \,:\, v_\sigma = 0 \text{ for all } \sigma \in \mathcal{F}_{\text{ext}} \text{ such that } \\ &\qquad\qquad \sigma \cap \Gamma_d = \emptyset.\}, \\ X_{\mathfrak{T},\Omega,\Gamma_n} &= \{v \in X_{\mathfrak{T}} \,:\, v_\sigma = 0 \text{ for all } \sigma \in \mathcal{F}_{\text{ext}} \text{ such that } \\ &\qquad\qquad \sigma \cap \Gamma_d \neq \emptyset\}. \end{aligned} \tag{B.77}$$

Note that $X_{\mathfrak{T}} = X_{\mathfrak{T},\Omega,\Gamma_n} \oplus X_{\mathfrak{T},\Gamma_d}$, and that $\mathbb{T}_{\mathfrak{T}} u = 0$ on Γ_d for any $u \in X_{\mathfrak{T},\Omega,\Gamma_n}$.

B.4.1 Discrete Sobolev embeddings

Discrete functional analysis tools for mixed conditions are a consequence of the two following lemmas, and of the techniques used in the previous sections for Dirichlet and Neumann boundary conditions.

Lemma B.28. *Let $\widetilde{\Omega}$ be a bounded connected open subset of $\mathbb{R}^d$ with Lipschitz boundary and let $A \subset \widetilde{\Omega}$ be a set of non-zero measure. Then, there exists C_{24} depending only on $\widetilde{\Omega}$ and A such that, for all $w \in BV(\widetilde{\Omega})$ satisfying $\int_A w(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = 0$,*

$$\|w\|_{L^{1^\star}(\widetilde{\Omega})} \le C_{24}\,|w|_{BV(\widetilde{\Omega})}\,, \tag{B.78}$$

where we recall that

$$|w|_{BV(\widetilde{\Omega})} = \sup\left\{\int_{\widetilde{\Omega}} w(\boldsymbol{x})\mathrm{div}\boldsymbol{\varphi}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \,:\, \boldsymbol{\varphi} \in C_c^\infty(\widetilde{\Omega},\mathbb{R}^d), \|\boldsymbol{\varphi}\|_{L^\infty(\widetilde{\Omega})^d} \le 1\right\}.$$

Proof. Let us start by recalling the Sobolev embedding, which can be obtained by passing to the limit on the similar embedding in $W^{1,1}(\widetilde{\Omega})$: there exists C_{25} depending only on $\widetilde{\Omega}$ such that

$$\forall w \in BV(\widetilde{\Omega})\,,\ \|w\|_{L^{1^\star}(\widetilde{\Omega})} \leq C_{25}(|w|_{BV(\widetilde{\Omega})} + \|w\|_{L^1(\widetilde{\Omega})}).$$

Estimate (B.78) is proved if we establish the following Poincaré inequality: there exists C_{26} depending only on $\widetilde{\Omega}$ and A such that, for any $w \in BV(\widetilde{\Omega})$ satisfying $\int_A w(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = 0$,

$$\|w\|_{L^1(\widetilde{\Omega})} \leq C_{26}\,|w|_{BV(\widetilde{\Omega})}\,. \tag{B.79}$$

The proof of (B.79) is done by way of contradiction, using a classical compactness technique. If this inequality does not hold, there exists a sequence $(w_m)_{m\in\mathbb{N}}$ in $BV(\widetilde{\Omega})$ such that $\int_A w_m(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = 0$ for all m and $\|w_m\|_{L^1(\widetilde{\Omega})} \geq m\,|w_m|_{BV(\widetilde{\Omega})}$. Dividing throughout by $\|w_m\|_{L^1(\widetilde{\Omega})}$ we can assume that $\|w_m\|_{L^1(\widetilde{\Omega})} = 1$ for all m. Then $(w_m)_{m\in\mathbb{N}}$ is bounded in $L^1(\widetilde{\Omega}) \cap BV(\widetilde{\Omega})$ and therefore, up to a subsequence, converges strongly in $L^1(\widetilde{\Omega})$ to some w such that $\|w\|_{L^1(\widetilde{\Omega})} = 1$. As $|w_m|_{BV(\widetilde{\Omega})} \leq 1/m \to 0$, we have $\nabla w_m \to 0$ in the sense of distributions on $\widetilde{\Omega}$ and therefore $\nabla w = 0$ on $\widetilde{\Omega}$. Since $\widetilde{\Omega}$ is connected, this shows that w is constant on $\widetilde{\Omega}$, equal to $\frac{1}{|\widetilde{\Omega}|}$ since its norm in $L^1(\widetilde{\Omega})$ is equal to 1.
But, passing to the limit in $\int_A w_m(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = 0$ gives $0 = \int_A w(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = \frac{|A|}{|\widetilde{\Omega}|}$, which is a contradiction with the fact that A has a non-zero measure. Hence (B.79) holds and so does (B.78). ∎

Under Assumptions (7.2) and (3.60), it is easy to construct a bounded connected open set $\widetilde{\Omega}$ with Lipschitz boundary which contains Ω, such that $A = \widetilde{\Omega}\backslash\Omega$ has a non-zero measure and $\overline{A} \cap \widetilde{\Omega} \subset \Gamma_d$. This can for example be done by gluing to Ω a small hypercube A along a planar subset of Γ_d, see Figure B.4. $\widetilde{\Omega}$ and A depend only on Ω and Γ_d.

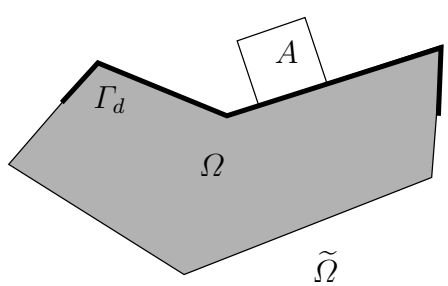


Fig. B.4. Extension of Ω.

Lemma B.29. *Under Assumptions* (7.2) *and* (3.60)*, let* $\widetilde{\Omega}$ *be constructed as above. Take* $\mathfrak{T}$ *a polytopal mesh of* Ω *in the sense of Definition 7.2 and, if* $u \in X_{\mathcal{D}}$*, define* $\widetilde{\Pi}_{\mathfrak{T}} u \in L^1(\widetilde{\Omega})$ *as the extension of* $\Pi_{\mathfrak{T}} u$ *by* 0 *outside* Ω*. Then*

$$\forall u \in X_{\mathfrak{T},\Omega,\Gamma_n}\,,\ \left|\widetilde{\Pi}_{\mathcal{D}} u\right|_{BV(\widetilde{\Omega})} \leq \sqrt{d}\,|u|_{\mathfrak{T},1}\,. \tag{B.80}$$

Proof. Let $\boldsymbol{\varphi} \in C_c^\infty(\widetilde{\Omega}, \mathbb{R}^d)$ be such that $\|\boldsymbol{\varphi}\|_{L^\infty(\widetilde{\Omega})} \leq 1$. We have

$$\int_{\widetilde{\Omega}} \widetilde{\Pi}_{\mathfrak{T}} u(\boldsymbol{x}) \mathrm{div}\boldsymbol{\varphi}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} = \int_{\Omega} \Pi_{\mathfrak{T}} u(\boldsymbol{x}) \mathrm{div}\boldsymbol{\varphi}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}.$$

Since $u_\sigma = 0$ whenever $\sigma \in \mathcal{F}_{\mathrm{ext}}$ is such that $\sigma \cap \Gamma_d \neq \emptyset$, and since $\boldsymbol{\varphi} = 0$ on $\partial\Omega \backslash \Gamma_d$, the boundary integral in (B.32) written for $v = u$ vanishes, and the same computations as in (B.38) lead to (B.80). ∎

The following Sobolev embeddings are a straightforward consequence of Lemma B.28 and B.29.

Lemma B.30 (Discrete embedding of $W^{1,1}(\Omega)$ in $L^{1^\star}(\Omega)$, mixed BCs). *Under Assumptions* (7.2) *and* (3.60)*, let* $\mathfrak{T}$ *be a polytopal mesh of* Ω *in the sense of Definition 7.2. Then, there exists* C_{27} *depending only on* Ω *and* Γ_d *such that*

$$\forall u \in X_{\mathfrak{T},\Omega,\Gamma_n}\,,\ \|\Pi_{\mathfrak{T}} u\|_{L^{1^\star}(\Omega)} \leq C_{27}\,|u|_{\mathfrak{T},1}\,.$$

The following results can be then proved from Lemma B.30 by using the same trick as in the proof of Lemma B.14 and Lemma B.15.

Lemma B.31 (Discrete embedding of $W^{1,p}(\Omega)$ in $L^{p^\star}(\Omega)$, mixed BCs, $p \in (1,d)$). *Under Assumptions* (7.2) *and* (3.60)*, let* $\mathfrak{T}$ *be a polytopal mesh of* Ω *in the sense of Definition 7.2,* $p \in (1,d)$ *and* $\eta \geq \eta_{\mathfrak{T}}$*. Then, there exists* C_{28} *depending only on* Ω*,* Γ_d *and* η *such that*

$$\forall u \in X_{\mathfrak{T},\Omega,\Gamma_n}\,,\ \|\Pi_{\mathfrak{T}} u\|_{L^{p^\star}(\Omega)} \leq C_{28}\,|u|_{\mathfrak{T},p}\,,$$

where $p^\star = \frac{dp}{d-p}$.

Lemma B.32 (Discrete embedding of $W^{1,p}(\Omega)$ in $L^q(\Omega)$ for some $q > p$, mixed BCs). *Under Assumptions* (7.2) *and* (3.60)*, let* $\mathfrak{T}$ *be a polytopal mesh of* Ω *in the sense of Definition 7.2,* $p \in [1,+\infty)$ *and* $\eta \geq \eta_{\mathfrak{T}}$*. Then, there exists* $q > p$*, depending only on* p *and* d*, and* C_{29}*, depending only on* Ω*,* d*,* p*,* Γ_d *and* η*, such that*

$$\forall u \in X_{\mathfrak{T},\Omega,\Gamma_n}\,,\ \|\Pi_{\mathfrak{T}} u\|_{L^{q}(\Omega)} \leq C_{29}\,|u|_{\mathfrak{T},p}\,.$$

If $p < d$ *we can take* $q = p^\star = \frac{pd}{d-p}$*. If* $p \geq d$*, we can take any* $q < +\infty$*.*

B.4.2 Compactness of $\Pi_{\mathfrak{T}}$ and $\mathbb{T}_{\mathfrak{T}}$

Lemma B.33 (Discrete Rellich theorem and compactness of the traces, mixed BCs). *Under Assumptions (7.2) and (3.60), let $p \in [1, +\infty)$ and $(\mathfrak{T}_m)_{m\in\mathbb{N}}$ be a sequence of polytopal meshes of Ω, such that $\sup_{m\in\mathbb{N}}(\theta_{\mathfrak{T}_m} + \eta_{\mathfrak{T}_m}) < +\infty$. Then, for any $u_m \in X_{\mathfrak{T}_m,\Omega,\Gamma_n}$ such that $(|u_m|_{\mathfrak{T}_m,p})_{m\in\mathbb{N}}$ is bounded, the sequence $(\Pi_{\mathfrak{T}_m} u_m)_{m\in\mathbb{N}}$ is relatively compact in $L^p(\Omega)$.*
Moreover, if $p > 1$ and $h_{\mathcal{M}_m} \to 0$ as $m \to \infty$, then $(\mathbb{T}_{\mathfrak{T}_m} u_m)_{m\in\mathbb{N}}$ is relatively compact in $L^p(\partial\Omega)$.

Proof. By Lemma B.32, the sequence $(\|\Pi_{\mathfrak{T}_m} u_m\|_{L^p(\Omega)})_{m\in\mathbb{N}}$ is bounded. Hence, the sequence $(\int_\Omega \Pi_{\mathfrak{T}_m} u_m(\boldsymbol{x})\mathrm{d}\boldsymbol{x})_{m\in\mathbb{N}}$ is also bounded and Lemma B.27 gives the relative compactness of $(\Pi_{\mathfrak{T}_m} u_m)_{m\in\mathbb{N}}$ in $L^p(\Omega)$, and of $(\mathbb{T}_{\mathfrak{T}_m} u_m)_{m\in\mathbb{N}}$ in $L^p(\partial\Omega)$ if $p > 1$ and $h_{\mathcal{M}_m} \to 0$ as $m \to \infty$. ∎

C

Discrete functional analysis for time-dependent problems

This chapter is devoted to compactness results for sequences of functions with domain $[0, T]$ and abstract co-domains (generic vector spaces). We focus on functions that are discrete-in-time, as they are classically encountered in numerical methods for time-dependent problems. The results established here apply to a range of numerical schemes for such problems, and are used in Section 4.2 to establish specific compactness properties of space–time gradient discretisations.

In Section C.1, we first consider compactness results based on the L^p norm on $[0, T]$, with $p < \infty$. There are called "averaged-in-time" because they only apply to norms that involve a time integral. A number of such compactness results, for piecewise-constant-in-time functions, can be found in the literature – see, *e.g.*, [12, 66, 45]. Much more scarce are uniform-in-time compactness results for discontinuous functions, i.e., results that apply to the supremum norm on $[0, T]$. The second section of this chapter, Section C.2, is devoted to establishing such uniform-in-time compactness theorems.

C.1 Averaged-in-time compactness results

The first two theorems are generalisations to vector-valued Lebesgue spaces of the classical Kolmogorov compactness theorem for L^p spaces [33]. If E is a measured space and B a Banach space, we denote by $L^p(E; B)$ the Lebesgue space of p-integrable functions $E \to B$, see, *e.g.*, [67, 88] for a definition and some properties of these spaces.

Theorem C.1 (Kolmogorov (1)). *Let B be a Banach space, $1 \le p < +\infty$, $T > 0$ and $A \subset L^p(0, T; B)$. Then A is relatively compact in $L^p(0, T; B)$ if it satisfies the following conditions:*

1. *For all $f \in A$, there exists $Pf \in L^p(\mathbb{R}; B)$ such that $Pf = f$ a.e. on $(0, T)$ and $\|Pf\|_{L^p(\mathbb{R};B)} \le C$, where C depends only on A.*

2. *For all* $\varphi \in C_c^\infty(\mathbb{R})$*, the set* $\{\int_{\mathbb{R}}(Pf)\varphi \mathrm{d}t,\ f \in A\}$ *is relatively compact in* B*.*
3. $\|Pf(\cdot + h) - Pf\|_{L^p(\mathbb{R};B)} \to 0$ *as* $h \to 0$*, uniformly with respect to* $f \in A$*.*

Remark C.2 (Necessary conditions)
The conditions 1, 2 and 3 are actually also necessary for A to be relatively compact in $L^p(0,T;B)$.

Proof. Let $(\rho_m)_{m\geq 1}$ be a sequence of mollifiers constructed by scaling a given smooth function ρ, that is:

$$\begin{aligned}&\rho \in C_c^\infty(-1,1),\ \int_{\mathbb{R}} \rho \mathrm{d}t = 1,\ \rho \geq 0,\ \rho(-t) = \rho(t) \text{ for all } t \in \mathbb{R}\\ &\text{and, for all } m \geq 1 \text{ and } t \in \mathbb{R},\ \rho_m(t) = m\rho(mt).\end{aligned} \tag{C.1}$$

Set $K = [0,T]$ and $A_m = \{(Pf \star \rho_m)_{|K} : f \in A\}$, where $\star$ denotes the convolution product in $\mathbb{R}$.
The proof is divided in two steps. In Step 1 we prove, using the Arzelà–Ascoli theorem and Assumption 2, that, for $m \geq 1$, the set A_m is relatively compact in $C(K;B)$ endowed with its usual topology of the supremum norm. This easily gives the relative compactness of A_m in $L^p(0,T;B)$. In Step 2, we show that Assumptions 1 and 3 give $Pf \star \rho_m \to Pf$ in $L^p(\mathbb{R};B)$ as $m \to +\infty$, uniformly with respect to $f \in A$. This allows to conclude that the set A is relatively compact in $L^p(0,T;B)$.

Step 1. Let $m \geq 1$. In order to prove that A_m is relatively compact in $C(K;B)$, we use the Ascoli–Arzelà theorem C.10. Hence, we need to prove that:

(AA1) for all $t \in K$, the set $\{Pf \star \rho_m(t),\ f \in A\}$ is relatively compact in B;
(AA2) the sequence $\{Pf \star \rho_m,\ f \in A\}$ is equicontinuous from K to B (*i.e.* the continuity of $Pf \star \rho_m : K \to B$ is uniform with respect to $f \in A$).

We first prove Property (AA1). For $t \in K$ we have, with $\varphi_t = \rho_m(t - \cdot) \in C_c^\infty(\mathbb{R})$,

$$Pf \star \rho_m(t) = \int_{\mathbb{R}} Pf(s)\rho_m(t-s)\mathrm{d}s = \int_{\mathbb{R}} Pf(s)\varphi_t(s)\mathrm{d}s.$$

Then, Assumption 2 applied to $\varphi = \varphi_t$ gives Property (AA1).

We now prove Property (AA2). Let $t_1, t_2 \in K$ and recall that $p' = \frac{p}{p-1}$. By Hölder's inequality,

$$\begin{aligned}\|Pf \star \rho_m(t_2) - Pf \star \rho_m(t_1)\|_B \\ \leq \int_{\mathbb{R}} \|Pf(s)\|_B \left|\rho_m(t_2 - s) - \rho_m(t_1 - s)\right| \mathrm{d}s\end{aligned}$$

$$\leq \|Pf\|_{L^p(\mathbb{R};B)} \|\rho_m(t_2 - \cdot) - \rho_m(t_1 - \cdot)\|_{L^{p'}(\mathbb{R})}.$$

Since $t_1, t_2 \in K = [0,T]$, the functions $\rho_m(t_2 - \cdot)$ and $\rho_m(t_1 - \cdot)$ vanish outside $[-1, T+1]$. Hence, using the mean value theorem and Assumption 3, we infer

$$\|Pf \star \rho_m(t_2) - Pf \star \rho_m(t_1)\|_B \leq C|t_1 - t_2| \left(\sup_{t\in\mathbb{R}} |\rho_m'(t)|\right) (T+2)^{\frac{1}{p'}}.$$

This shows that $Pf \star \rho_m$ is uniformly continuous on $\mathbb{R}$, with a modulus of continuity which does not depend on f. Hence, Property (AA2) is proved.

As a consequence, A_m is indeed relatively compact in $C(K;B)$. This is equivalent to saying that, for any $\varepsilon > 0$, there exists a finite number of balls of radius ε (for the supremum norm of $C(K;B)$) whose union cover the set A_m. Then, since $\|\cdot\|_{L^p(0,T;B)} \leq T^{1/p} \|\cdot\|_{C(K;B)}$, we also obtain the relative compactness of A_m in $L^p(0,T;B)$.

Step 2. Let $t \in \mathbb{R}$, we have, using $\int_{\mathbb{R}} \rho_m(s)\mathrm{d}s = 1$ and setting $\overline{s} = ms$,

$$\begin{aligned} Pf \star \rho_m(t) - Pf(t) &= \int_{\mathbb{R}} [Pf(t-s) - Pf(t)]\, \rho_m(s)\mathrm{d}s \\ &= \int_{-1}^{1} \left[Pf\left(t - \frac{\overline{s}}{m}\right) - Pf(t)\right] \rho(\overline{s})\mathrm{d}\overline{s}. \end{aligned}$$

Then, by Hölder's inequality,

$$\|Pf \star \rho_m(t) - Pf(t)\|_B^p \leq \|\rho\|_{L^{p'}}^p \int_{-1}^{1} \left\|Pf\left(t - \frac{\overline{s}}{m}\right) - Pf(t)\right\|_B^p \mathrm{d}\overline{s}.$$

Integrating with respect to $t \in \mathbb{R}$ and using the Fubini-Tonelli theorem to swap the integrals on t and $\overline{s}$ leads to

$$\begin{aligned} &\|Pf \star \rho_m - Pf\|_{L^p(0,T;B)}^p \\ &\qquad \leq \|\rho\|_{L^{p'}}^p \int_{-1}^{1} \left\|Pf\left(\cdot - \frac{\overline{s}}{m}\right) - Pf\right\|_{L^p(0,T;B)}^p \mathrm{d}\overline{s} \\ &\qquad \leq 2\,\|\rho\|_{L^{p'}}^p \sup\left\{\|Pf(\cdot + h) - Pf\|_{L^p(\mathbb{R};B)}^p \,:\, |h| \leq \frac{1}{m}\right\}. \end{aligned}$$

Using Assumption 3 then gives $\|Pf \star \rho_m - Pf\|_{L^p(0,T;B)} \to 0$ as $m \to +\infty$, uniformly with respect to $f \in A$.

We can now conclude the proof. Let $\varepsilon > 0$ and pick $m(\varepsilon)$ large enough such that

$$\left\|Pf \star \rho_{m(\varepsilon)} - Pf\right\|_{L^p(0,T;B)} \leq \varepsilon/2 \quad \text{for all } f \in A. \tag{C.2}$$

By Step 1, we can cover $A_{m(\varepsilon)} = \{(Pf \star \rho_{m(\varepsilon)})_{|[0,T]} \,:\, f \in A\}$ by a finite number of balls in $L^p(0,T;B)$ of radius $\varepsilon/2$. Property (C.2) then shows that

$\{(Pf)_{|[0,T]} : f \in A\} = A$ is covered by the same finite number of balls of radius ε. This concludes the proof that A is relatively compact in $L^p(0,T;B)$.

■

Theorem C.3 (Kolmogorov (2)). *Let B be a Banach space, $1 \le p < +\infty$, $T > 0$ and $A \subset L^p(0,T;B)$. Then A is relatively compact in $L^p(0,T;B)$ if it satisfies the following conditions:*

1. *A is bounded in $L^p(0,T;B)$.*
2. *For all $\varphi \in C_c^\infty(\mathbb{R})$, the set $\{\int_0^T f\varphi \mathrm{d}t : f \in A\}$ is relatively compact in B.*
3. *There exists a function $\eta : (0,T) \to [0,\infty)$ such that $\lim_{h\to 0^+} \eta(h) = 0$ and, for all $h \in (0,T)$ and $f \in A$,*

$$\int_0^{T-h} \|f(t+h) - f(t)\|_B^p \,\mathrm{d}t \le \eta(h).$$

Proof.
The proof uses Theorem C.1 with P defined the following way: for $f \in A$, $Pf = f$ on $[0,T]$ and $Pf = 0$ on $\mathbb{R}\backslash[0,T]$. Owing to this definition and to Assumption 1 in Theorem C.3, Items 1 and 2 of Theorem C.1 are clearly satisfied.
We now prove, in two steps, Item 3 of Theorem C.1. Notice first that, replacing η with $\widetilde{\eta}(h) = \sup_{(0,h]} \eta$ (which still satisfies $\lim_{h\to 0^+} \widetilde{\eta}(h) = 0$), we can assume without loss of generality that η is non-decreasing.

Step 1. In this step, we prove that $\int_0^\tau \|f(t)\|_B^p \,\mathrm{d}t \to 0$ as $\tau \to 0^+$, uniformly with respect to $f \in A$.

Let $\tau, h \in (0,T)$ such that $\tau + h \le T$. For all $t \in (0,\tau)$ one has $\|f(t)\|_B \le \|f(t+h)\|_B + \|f(t+h) - f(t)\|_B$ and thus, by the power-of-sums inequality (D.12),

$$\|f(t)\|_B^p \le 2^{p-1} \|f(t+h)\|_B^p + 2^{p-1} \|f(t+h) - f(t)\|_B^p .$$

Integrating this inequality for $t \in (0,\tau)$ gives

$$\int_0^\tau \|f(t)\|_B^p \,\mathrm{d}t \le 2^{p-1} \int_0^\tau \|f(t+h)\|_B^p \,\mathrm{d}t + 2^{p-1} \int_0^\tau \|f(t+h) - f(t)\|_B^p \,\mathrm{d}t. \quad \text{(C.3)}$$

Now let $h_0 \in (0,T)$ and $\tau \in (0, T - h_0)$. For all $h \in (0,h_0)$, Inequality (C.3) gives, using $\eta(h) \le \eta(h_0)$,

$$\int_0^\tau \|f(t)\|_B^p \,\mathrm{d}t \le 2^{p-1} \int_0^\tau \|f(t+h)\|_B^p \,\mathrm{d}t + 2^{p-1}\eta(h_0).$$

Integrating this inequality over $h \in (0, h_0)$ leads to

$$h_0 \int_0^\tau \|f(t)\|_B^p \,\mathrm{d}t \le 2^{p-1} \int_0^{h_0} \left(\int_0^\tau \|f(t+h)\|_B^p \,\mathrm{d}t \right) dh + 2^{p-1} h_0 \eta(h_0). \quad \text{(C.4)}$$

Using the Fubini-Tonelli Theorem,

$$\begin{aligned} \int_0^{h_0} \left(\int_0^\tau \|f(t+h)\|_B^p \,\mathrm{d}t \right) dh &= \int_0^\tau \left(\int_0^{h_0} \|f(t+h)\|_B^p \, dh \right) \mathrm{d}t \\ &\le \int_0^\tau \left(\int_0^T \|f(s)\|_B^p \, ds \right) \le \tau \, \|f\|_{L^p(0,T;B)}^p , \end{aligned}$$

from which one deduces, owing to (C.4),

$$\int_0^\tau \|f(t)\|_B^p \,\mathrm{d}t \le \frac{\tau 2^{p-1}}{h_0} \|f\|_{L^p(0,T;B)}^p + 2^{p-1} \eta(h_0).$$

We can now conclude this step. Let $\varepsilon > 0$ and choose $h_0 \in (0, T)$ such that $2^{p-1}\eta(h_0) \le \varepsilon$. Then, with $C = \sup_{f \in A} \|f\|_{L^p(0,T;B)}^p$, take $\overline{\tau} = \min(T - h_0, \varepsilon h_0/(2^{p-1}C))$. This gives, for all $f \in A$ and all $\tau \le \overline{\tau}$,

$$\int_0^\tau \|f(t)\|_B^p \le 2\varepsilon.$$

The proof that $\int_0^\tau \|f(t)\|_B^p \,\mathrm{d}t \to 0$ as $\tau \to 0^+$, uniformly with respect to $f \in A$, is complete.

A similar proof gives $\int_{T-\tau}^T \|f(t)\|_B^p \,\mathrm{d}t \to 0$ as $\tau \to 0^+$, uniformly with respect to $f \in A$ (this can for example be obtained by working on $g(t) = f(T - t)$ instead of f).

Step 2. We now prove that Item 3 in Theorem C.1 is satisfied, and thus conclude the proof of Theorem C.3.
Recall that $Pf(t) = 0$ if $t \notin [0, T]$ so that, for all $h \in (0, T)$ and $f \in A$,

$$\begin{aligned} &\int_{\mathbb{R}} \|Pf(t+h) - Pf(t)\|_B^p \mathrm{d}t \\ &\quad \le \int_{-h}^0 \|f(t+h)\|_B^p \,\mathrm{d}t + \int_0^{T-h} \|f(t+h) - f(t)\|_B^p \,\mathrm{d}t + \int_{T-h}^T \|f(t)\|_B^p \,\mathrm{d}t \\ &\quad \le \int_0^h \|f(t)\|_B^p \,\mathrm{d}t + \eta(h) + \int_{T-h}^T \|f(t)\|_B^p \,\mathrm{d}t. \end{aligned} \quad \text{(C.5)}$$

Let $\varepsilon > 0$ and take $h_1 > 0$ such $\eta(h_1) \le \varepsilon$. Owing to Step 1, there exists $h_2 > 0$ such that, for all $f \in A$ and $h \le h_2$,

$$\int_0^h \|f(t)\|_B^p \,\mathrm{d}t \le \varepsilon \quad \text{and} \quad \int_{T-h}^T \|f(t)\|_B^p \,\mathrm{d}t \le \varepsilon.$$

Hence, by (C.5), for all $f \in A$ and $h \leq \min(h_1, h_2)$,

$$\int_{\mathbb{R}} \|Pf(t+h) - Pf(t)\|_B^p \, \mathrm{d}t \leq 3\varepsilon.$$

This concludes the proof that Assumption 3 in Theorem C.1 is satisfied.

■

We now turn to compactness theorems involving sequences of spaces as codomains of the functions. This typically occurs in numerical schemes, when we consider sequences of functions that are piecewise constant on varying meshes. We first state a notion of "compact embedding" of a sequence of spaces in a fixed Banach space.

Definition C.4 (Compactly embedded sequence). *Let B be a Banach space and $(X_m, \|\cdot\|_{X_m})_{m\in\mathbb{N}}$ be a sequence of Banach spaces included in B. The sequence $(X_m)_{m\in\mathbb{N}}$ is compactly embedded in B if any sequence $(u_m)_{m\in\mathbb{N}}$ such that*

$$u_m \in X_m \text{ for all } m \in \mathbb{N}, \text{ and } (\|u_m\|_{X_m})_{m\in\mathbb{N}} \text{ is bounded,}$$

is relatively compact in B.

The first compactness result for sequences of subspaces is a straightforward translation in that setting of the second Kolmogorov theorem above.

Proposition C.5 (Time compactness with a sequence of subspaces). *Let $1 \leq p < +\infty$, $T > 0$, B be a Banach space, and $(X_m)_{m\in\mathbb{N}}$ be compactly embedded in B (see Definition C.4). Let $(f_m)_{m\in\mathbb{N}}$ be a sequence in $L^p(0,T;B)$ satisfying the following conditions:*

1. *The sequence $(f_m)_{m\in\mathbb{N}}$ is bounded in $L^p(0,T;B)$.*
2. *The sequence $(\|f_m\|_{L^1(0,T;X_m)})_{m\in\mathbb{N}}$ is bounded.*
3. *There exists a function $\eta : (0,T) \to [0,\infty)$ such that $\lim_{h\to 0^+} \eta(h) = 0$ and, for all $h \in (0,T)$ and $m \in \mathbb{N}$,*

$$\int_0^{T-h} \|f_m(t+h) - f_m(t)\|_B^p \, \mathrm{d}t \leq \eta(h).$$

Then, the sequence $(f_m)_{m\in\mathbb{N}}$ is relatively compact in $L^p(0,T;B)$.

Proof. We aim at applying Theorem C.3 with $A = \{f_m : m \in \mathbb{N}\}$. We only have to prove Assumption 2 in this theorem, the other two assumptions being already stated as assumptions of the proposition.
Let $\varphi \in C_c^\infty(\mathbb{R})$. We need to prove that the sequence $(\int_0^T f_m \varphi \mathrm{d}t)_{m\in\mathbb{N}}$ is relatively compact in B. We have, with $\|\varphi\|_\infty = \sup_{t\in\mathbb{R}} |\varphi(t)|$,

$$\left\| \int_0^T f_m \varphi \mathrm{d}t \right\|_{X_m} \leq \|\varphi\|_\infty \|f_m\|_{L^1(0,T;X_m)}.$$

The sequence $(\|f_m\|_{L^1(0,T;X_m)})_{m\in\mathbb{N}}$ being bounded, this shows that the sequence

$$\left(\left\|\int_0^T f_m\varphi \mathrm{d}t\right\|_{X_m}\right)_{m\in\mathbb{N}}$$

is also bounded. Since $(X_m)_{m\in\mathbb{N}}$ is compactly embedded in B, this concludes the proof that $(\int_0^T f_m\varphi \mathrm{d}t)_{m\in\mathbb{N}}$ in relatively compact in B. ∎

We then turn to the statement and proof of a discrete Aubin–Simon theorem, which was first used in [109] and generalised in [108], see also [107].
In the continuous setting, the Aubin–Simon compactness theorem establishes a strong compactness property of sequences of functions in $L^p(0,T;B)$, based on their boundedness in $L^q(0,T;A)$ and the boundedness of their derivatives in $L^r(0,T;C)$, where A is compactly embedded in B and B is continuously embedded in C. We first define a notion of triplets (A,B,C) having these compact–continuous embedding properties, in the case where A and C are replaced by sequences of spaces.

Definition C.6 (Compactly–continuously embedded sequence). *Let B be a Banach space, $(X_m,\|\cdot\|_{X_m})_{m\in\mathbb{N}}$ be a sequence of Banach spaces included in B, and $(Y_m,\|\cdot\|_{Y_m})_{m\in\mathbb{N}}$ be a sequence of Banach spaces. We say that the sequence $(X_m,Y_m)_{m\in\mathbb{N}}$ is compactly–continuously embedded in B if the following conditions are satisfied:*

1. *The sequence $(X_m)_{m\in\mathbb{N}}$ is compactly embedded in B (see Definition C.4).*
2. *$X_m\subset Y_m$ for all $m\in\mathbb{N}$ and, for any sequence $(u_m)_{m\in\mathbb{N}}$ such that*
   a) $u_m\in X_m$ for all $m\in\mathbb{N}$ and $(\|u_m\|_{X_m})_{m\in\mathbb{N}}$ is bounded,
   b) $\|u_m\|_{Y_m}\to 0$ as $n\to+\infty$,
   c) $(u_m)_{m\in\mathbb{N}}$ converges in B,
   it holds $u_m\to 0$ in B.

Lemma C.7. *Let B be a Banach space and $(X_m,Y_m)_{m\in\mathbb{N}}$ be compactly–continuously embedded in B (see Definition C.6). Then, for any $\varepsilon>0$, there exists $m_0\in\mathbb{N}$ and $C_\varepsilon\geq 0$ such that, for any $m\geq m_0$ and $w\in X_m$, one has*

$$\|w\|_B\leq \varepsilon\|w\|_{X_m}+C_\varepsilon\|w\|_{Y_m}.$$

Proof. We prove the result by contradiction. Let us therefore assume the existence of $\varepsilon>0$ such that, for any $m_0\in\mathbb{N}$, we can find $m=\varphi(m_0)\geq m_0$ and $w_{\varphi(m_0)}\in X_{\varphi(m_0)}$ such that

$$\left\|w_{\varphi(m_0)}\right\|_B>\varepsilon\left\|w_{\varphi(m_0)}\right\|_{X_{\varphi(m_0)}}+m_0\left\|w_{\varphi(m_0)}\right\|_{Y_{\varphi(m_0)}}.$$

There is no loss of generality in also selecting, by induction, each $m=\varphi(m_0)$ greater than $\varphi(m_0-1)$; then $\varphi:\mathbb{N}\to\mathbb{N}$ is a strictly increasing mapping. Since $w_{\varphi(m_0)}\neq 0$, we can then set $u_{\varphi(m_0)}=\frac{w_{\varphi(m_0)}}{\left\|w_{\varphi(m_0)}\right\|_B}\in X_{\varphi(m_0)}$. We then have, for any $m\in\varphi(\mathbb{N})$,

$$1 = \|u_m\|_B \geq \varepsilon \|u_m\|_{X_m} + \psi(m) \|u_m\|_{Y_m}, \tag{C.6}$$

where $\psi = \varphi^{-1} : \varphi(\mathbb{N}) \to \mathbb{N}$ satisfies $\psi(m) \to \infty$ as $m \to \infty$. To define u_m for all $m \in \mathbb{N}$, we let $u_m = 0$ whenever $n \notin \varphi(\mathbb{N})$ and, defining $\psi(m) = m$ in that case, we see that (C.6) still holds. This definition also preserves the property $\psi(m) \to \infty$ as $m \to \infty$.
The sequence $(u_m)_{m\in\mathbb{N}}$ is such that $u_m \in X_m$ for all $m \in \mathbb{N}$ and, owing to (C.6), $(\|u_m\|_{X_m})_{m\in\mathbb{N}}$ is bounded by $1/\varepsilon$. By the compact embedding of $(X_m)_{m\in\mathbb{N}}$ in B, we infer that there exists a subsequence, still denoted $(u_m)_{m\in\mathbb{N}}$, that converges in B. Then, using (C.6) again, $\|u_m\|_{Y_m} \leq 1/\psi(m) \to 0$ as $m \to +\infty$, and thus, by Definition C.6, the limit of $(u_m)_{m\in\mathbb{N}}$ in B must be 0. This contradicts (C.6) which states that, since each u_m has norm 1 in B, the limit in this space of these vectors should also have norm 1. ■

We can now state a discrete Aubin–Simon theorem with sequences of spaces.

Theorem C.8 (Aubin–Simon with sequences of spaces and discrete derivative). *Let $p \in [1, +\infty)$. Let B be a Banach space and $(X_m, Y_m)_{m\in\mathbb{N}}$ be compactly–continuously embedded in B (see Definition C.6). Let $T > 0$, $\theta \in [0, 1]$, and $(f_m)_{m\in\mathbb{N}}$ be a sequence of $L^p(0, T; B)$ satisfying the following properties:*

1. *For all $m \in \mathbb{N}$, there exists*
   - $N \in \mathbb{N}^\star$,
   - $0 = t^{(0)} < t^{(1)} < \cdots < t^{(N)} = T$, *and*
   - $(v^{(n)})_{n=0,\ldots,N} \in X_m^{N+1}$

   such that, for all $n \in \{0, \ldots, N-1\}$ and a.e. $t \in (t^{(n)}, t^{(n+1)})$, $f_m(t) = \theta v^{(n+1)} + (1-\theta) v^{(n)}$.
   We then define almost everywhere the discrete derivative $\delta_m f_m$ by setting, with $\mathfrak{d}t^{(n+\frac{1}{2})} = t^{(n+1)} - t^{(n)}$,

$$\delta_m f_m(t) = \frac{v^{(n+1)} - v^{(n)}}{\mathfrak{d}t^{(n+\frac{1}{2})}} \text{ for } n \in \{0, \ldots, N-1\} \text{ and } t \in (t^{(n)}, t^{(n+1)}).$$

2. *The sequence $(f_m)_{m\in\mathbb{N}}$ is bounded in $L^p(0, T; B)$.*
3. *The sequence $(\|f_m\|_{L^p(0,T;X_m)})_{m\in\mathbb{N}}$ is bounded.*
4. *The sequence $(\|\delta_m f_m\|_{L^1(0,T;Y_m)})_{m\in\mathbb{N}}$ is bounded.*

Then $(f_m)_{m\in\mathbb{N}}$ is relatively compact in $L^p(0, T; B)$.

Proof. We apply Proposition C.5. The only assumption in this proposition that needs to be established in order to conclude is the third one, that is

$$\int_0^{T-h} \|f_m(t+h) - f_m(t)\|_B^p \, \mathrm{d}t \to 0 \text{ as } h \to 0,$$

uniformly w.r.t. $m \in \mathbb{N}$.

Note that, without the "uniformly with respect to $m \in \mathbb{N}$", this convergence is known since each f_m belongs to $L^p(0,T;B)$. As a consequence, we only have to prove that, for all $\eta > 0$, there exist $m_0 \in \mathbb{N}$ and $0 < h_0 < T$ such that

$$\forall m \geq m_0,\, \forall h \in (0,h_0)\,,\ \int_0^{T-h} \|f_m(\cdot + h) - f_m\|_B^p \,\mathrm{d}t \leq \eta. \tag{C.7}$$

Indeed, once this is proved, upon reducing h_0 we can ensure that this estimate also holds for $f_1, \ldots, f_{m_0-1}$.
Let $\varepsilon > 0$. Lemma C.7 gives the existence of $m_0 \in \mathbb{N}$ and $C_\varepsilon \in \mathbb{R}$ such that, for all $m \geq m_0$ and $u \in X_m$, $\|u\|_B \leq \varepsilon \|u\|_{X_m} + C_\varepsilon \|u\|_{Y_m}$. Then, for $m \geq m_0$, $0 < h < T$ and $t \in (0, T-h)$,

$$\begin{aligned}
&\|f_m(t+h) - f_m(t)\|_B \\
&\qquad \leq \varepsilon \|f_m(t+h) - f_m(t)\|_{X_m} + C_\varepsilon \|f_m(t+h) - f_m(t)\|_{Y_m} \\
&\qquad \leq \varepsilon \|f_m(t+h)\|_{X_m} + \varepsilon \|f_m(t)\|_{X_m} + C_\varepsilon \|f_m(t+h) - f_m(t)\|_{Y_m}\,.
\end{aligned}$$

Take the power p of this inequality and use the power-of-sums inequality (D.14) to obtain

$$\begin{aligned}
\|f_m(t+h) - f_m(t)\|_B^p \leq\ & 3^{p-1}\varepsilon^p \|f_m(t+h)\|_{X_m}^p \\
& + 3^{p-1}\varepsilon^p \|f_m(t)\|_{X_m}^p + 3^{p-1}C_\varepsilon^p \|f_m(t+h) - f_m(t)\|_{Y_m}^p\,.
\end{aligned}$$

Integrating this inequality with respect to $t \in (0, T-h)$ leads to

$$\begin{aligned}
\int_0^{T-h} \|f_m(t+h) - f_m(t)\|_B^p \,\mathrm{d}t \leq\ & 2 \times 3^{p-1}\varepsilon^p \|f_m\|_{L^p(0,T;X_m)}^p \\
& + 3^{p-1}C_\varepsilon^p \int_0^{T-h} \|f_m(t+h) - f_m(t)\|_{Y_m}^p \,\mathrm{d}t.
\end{aligned} \tag{C.8}$$

We now estimate the last term in this inequality by using the discrete derivative of f_m. This function is piecewise constant in time so, for a.e. $t \in (0, T-h)$, writing $f_m(t+h) - f_m(t)$ as the sum of the jumps of f_m at its discontinuities gives

$$\begin{aligned}
f_m(t+h) - f_m(t) &= \sum_{n:\, t^{(n)} \in (t,t+h)} (f_m)_{|(t^{(n)}, t^{(n+1)})} - (f_m)_{|(t^{(n-1)}, t^{(n)})} \\
&= \sum_{n:\, t^{(n)} \in (t,t+h)} (\theta v^{(n+1)} + (1-\theta)v^{(n)}) - (\theta v^{(n)} + (1-\theta)v^{(n-1)}) \\
&= \sum_{n:\, t^{(n)} \in (t,t+h)} \left[\theta(v^{(n+1)} - v^{(n)}) + (1-\theta)(v^{(n)} - v^{(n-1)})\right] \\
&= \sum_{n=1}^{N-1} \left[\theta(v^{(n+1)} - v^{(n)}) + (1-\theta)(v^{(n)} - v^{(n-1)})\right] \mathbf{1}_{(t,t+h)}(t^{(n)})
\end{aligned}$$

$$= \theta \sum_{n=1}^{N-1} \frac{v^{(n+1)} - v^{(n)}}{\eth t^{(n+\frac{1}{2})}} \eth t^{(n+\frac{1}{2})} \mathbf{1}_{(t,t+h)}(t^{(n)})$$
$$+ (1-\theta) \sum_{n=1}^{N-1} \frac{v^{(n)} - v^{(n-1)}}{\eth t^{(n-\frac{1}{2})}} \eth t^{(n-\frac{1}{2})} \mathbf{1}_{(t,t+h)}(t^{(n)}), \tag{C.9}$$

where $\mathbf{1}_{(t,t+h)}(t^{(n)}) = 1$ if $t^{(n)} \in (t, t+h)$ and $\mathbf{1}_{(t,t+h)}(t^{(n)}) = 0$ if $t^{(n)} \notin (t, t+h)$. Let M be a bound of $\|\delta_m f_m\|_{L^1(0,T;Y_m)}$, which means that, for all $m \in \mathbb{N}$,

$$\sum_{n=0}^{N-1} \left\| \frac{v^{(n+1)} - v^{(n)}}{\eth t^{(n+\frac{1}{2})}} \right\|_{Y_m} \eth t^{(n+\frac{1}{2})} \leq M.$$

Taking the Y_m-norm of (C.9), then the power p, and using the convexity of $s \to s^p$ gives

$$\begin{aligned}
&\|f_m(t+h) - f_m(t)\|_{Y_m}^p \\
&\leq \theta \left(\sum_{n=1}^{N-1} \left\| \frac{v^{(n+1)} - v^{(n)}}{\eth t^{(n+\frac{1}{2})}} \right\|_{Y_m} \eth t^{(n+\frac{1}{2})} \mathbf{1}_{(t,t+h)}(t^{(n)}) \right)^p \\
&\quad + (1-\theta) \left(\sum_{n=1}^{N-1} \left\| \frac{v^{(n)} - v^{(n-1)}}{\eth t^{(n-\frac{1}{2})}} \right\|_{Y_m} \eth t^{(n-\frac{1}{2})} \mathbf{1}_{(t,t+h)}(t^{(n)}) \right)^p \\
&\leq \theta M^{p-1} \left(\sum_{n=1}^{N-1} \left\| \frac{v^{(n+1)} - v^{(n)}}{\eth t^{(n+\frac{1}{2})}} \right\|_{Y_m} \eth t^{(n+\frac{1}{2})} \mathbf{1}_{(t,t+h)}(t^{(n)}) \right) \\
&\quad + (1-\theta) M^{p-1} \left(\sum_{n=1}^{N-1} \left\| \frac{v^{(n)} - v^{(n-1)}}{\eth t^{(n-\frac{1}{2})}} \right\|_{Y_m} \eth t^{(n-\frac{1}{2})} \mathbf{1}_{(t,t+h)}(t^{(n)}) \right).
\end{aligned} \tag{C.10}$$

Writing $\mathbf{1}_{(t,t+h)}(t^{(n)}) = \mathbf{1}_{(t^{(n)}-h, t^{(n)})}(t)$ and integrating this inequality over $t \in (0, T-h)$ leads to

$$\int_0^{T-h} \|f_m(t+h) - f_m(t)\|_{Y_m}^p \, \mathrm{d}t \leq M^p h. \tag{C.11}$$

Plugging this inequality into (C.8), we obtain

$$\int_0^{T-h} \|f_m(t+h) - f_m(t)\|_B^p \, \mathrm{d}t \leq 2 \times 3^{p-1} \varepsilon^p \|f_m\|_{L^p(0,T;X_m)}^p + 3^{p-1} C_\varepsilon^p M^p h. \tag{C.12}$$

We can now conclude the proof. Let $\eta > 0$. Since $(\|f_m\|_{L^p(0,T;X_m)})_{m \in \mathbb{N}}$ is bounded, we can fix ε (and thus also m_0) such that, for all $m \geq m_0$,

$$2 \times 3^{p-1} \varepsilon^p \|f_m\|_{L^p(0,T;X_m)}^p \leq \frac{\eta}{2}.$$

We can then select $h_0 \in (0, T)$ such that $3^{p-1} C_\varepsilon^p M^p h_0 \leq \eta/2$. Estimate (C.12) then shows that (C.7) holds, which proves the theorem. ∎

C.2 Uniform-in-time compactness

Solutions of numerical schemes for parabolic equations are usually piecewise constant in time, and therefore not continuous. Their jumps nevertheless tend to become small with the time step, and it is possible to establish some uniform-in-time convergence results. These results, most of which were first published in [72], are based on a generalisation to non-continuous functions of the classical Arzelà–Ascoli theorem for continuous functions, which is recalled for the sake of completeness.

Definition C.9. *If (K, d_K) and (E, d_E) are metric spaces, we denote by $\mathcal{F}(K, E)$ the space of functions $K \to E$, endowed with the uniform metric $d_{\mathcal{F}}(v, w) = \sup_{s \in K} d_E(v(s), w(s))$ (note that this metric may take infinite values).*

Theorem C.10 (Arzelà–Ascoli). *Let (K, d_K) be a compact metric space, (E, d_E) be a complete metric space, and let $C(K, E)$ be the set of continuous functions from (K, d_K) to (E, d_E), endowed with its usual norm. Let $(v_m)_{m \in \mathbb{N}}$ be a sequence in $C(K, E)$. The sequence $(v_m)_{m \in \mathbb{N}}$ is relatively compact if it is pointwise bounded and equicontinuous, that is to say:*

$$\begin{aligned}
&\forall x \in K,\ \text{the sequence } (v_m(x))_{m \in \mathbb{N}} \text{ is bounded in } E,\\
&\forall x \in K,\ \forall \varepsilon > 0,\ \exists \delta > 0\ :\\
&\qquad [y \in K,\ d_K(x, y) < \delta,\ m \in \mathbb{N}] \Rightarrow [d_E(v_m(x), v_m(y)) < \varepsilon]\,.
\end{aligned}$$

Theorem C.11 (Discontinuous Ascoli–Arzelà's theorem). *Let (K, d_K) be a compact metric space, (E, d_E) be a complete metric space, and let $(\mathcal{F}(K, E), d_{\mathcal{F}})$ be as in Definition C.9. Let $(v_m)_{m \in \mathbb{N}}$ be a sequence in $\mathcal{F}(K, E)$ such that there exists a function $\omega : K \times K \to [0, \infty]$ and a sequence $(\tau_m)_{m \in \mathbb{N}} \subset [0, \infty)$ satisfying*

$$\lim_{d_K(s,s') \to 0} \omega(s, s') = 0\,, \quad \lim_{m \to \infty} \tau_m = 0\,, \tag{C.13}$$

$$\forall (s, s') \in K^2\,,\ \forall m \in \mathbb{N}\,,\ d_E(v_m(s), v_m(s')) \le \omega(s, s') + \tau_m. \tag{C.14}$$

We also assume that, for all $s \in K$, $\{v_m(s) : m \in \mathbb{N}\}$ is relatively compact in (E, d_E).
Then $(v_m)_{m \in \mathbb{N}}$ is relatively compact in $(\mathcal{F}(K, E), d_{\mathcal{F}})$, and any adherence value of $(v_m)_{m \in \mathbb{N}}$ in this space is continuous $K \to E$.

Proof. The last conclusion of the theorem, *i.e.* that any adherence value v of $(v_m)_{m \in \mathbb{N}}$ in $\mathcal{F}(K, E)$ is continuous, is obtained by passing to the limit along this subsequence in (C.14), showing that the modulus of continuity of v is bounded above by ω.
The proof of the compactness result is an easy generalisation of the proof of the classical Ascoli–Arzelà compactness result. We start by taking a countable

dense subset $\{s_l : l \in \mathbb{N}\}$ in K (the existence of this set is ensured since K is compact metric). Since each set $\{v_m(s_l) : m \in \mathbb{N}\}$ is relatively compact in E, by diagonal extraction we can select a subsequence of $(v_m)_{m\in\mathbb{N}}$, denoted the same way, such that for any $l \in \mathbb{N}$, $(v_m(s_l))_{m\in\mathbb{N}}$ converges in E. We then proceed in showing that $(v_m)_{m\in\mathbb{N}}$ is a Cauchy sequence in $(\mathcal{F}(K,E), d_{\mathcal{F}})$. Since this space is complete, this will show that this sequence converges in this space and will therefore complete the proof.
Let $\varepsilon > 0$ and, using (C.13), take $\delta > 0$ and $M \in \mathbb{N}$ such that $\omega(s,s') \le \varepsilon$ whenever $d_K(s,s') \le \delta$ and $\tau_m \le \varepsilon$ whenever $m \ge M$. Select a finite set $\{s_{l_1}, \ldots, s_{l_N}\}$ such that any $s \in K$ is within distance δ of a s_{l_i}. Then, for any $m, m' \ge M$, by (C.14),

$$\begin{aligned} d_E(v_m(s), v_{m'}(s)) &\le d_E(v_m(s), v_m(s_{l_i})) + d_E(v_m(s_{l_i}), v_{m'}(s_{l_i})) \\ &\quad + d_E(v_{m'}(s_{l_i}), v_{m'}(s)) \\ &\le \omega(s, s_{l_i}) + \tau_m + d_E(v_m(s_{l_i}), v_{m'}(s_{l_i})) + \omega(s, s_{l_i}) + \tau_{m'} \\ &\le 4\varepsilon + d_E(v_m(s_{l_i}), v_{m'}(s_{l_i})). \end{aligned} \tag{C.15}$$

Let $i \in \{1, \ldots, N\}$. The sequence $(v_m(s_{l_i}))_{m\in\mathbb{N}}$ converges in E, and is therefore a Cauchy sequence in this space. We can thus find $M_i \in \mathbb{N}$ such that

$$\forall m, m' \ge M_i\,, \quad d_E(v_m(s_{l_i}), v_{m'}(s_{l_i})) \le \varepsilon. \tag{C.16}$$

Take $M' = \max(M, M_1, \ldots, M_N)$. Estimates (C.16) and (C.15) show that, for all $m, m' \ge M$ and all $s \in K$, $d_E(v_m(s), v_{m'}(s)) \le 5\varepsilon$. This concludes the proof that $(v_m)_{m\in\mathbb{N}}$ is a Cauchy sequence in $(\mathcal{F}(K,E), d_{\mathcal{F}})$. ∎

Corollary C.12 (Uniform-in-time compactness from estimates on discrete derivatives). *Let $T > 0$, $\theta \in [0,1]$, B be a Banach space, and $(X_m, \|\cdot\|_{X_m})_{m\in\mathbb{N}}$ be a sequence of Banach spaces included in B. For any $m \in \mathbb{N}$, we take*

- $N_m \in \mathbb{N}^*$,
- $0 = t_m^{(0)} < t_m^{(1)} < \ldots < t_m^{(N_m)} = T$, *and*
- $u_m = (u_m^{(n)})_{n=0,\ldots,N_m} \in X_m^{N_m+1}$.

Let $(u_m)_\theta : [0,T] \to X_m$ be the piecewise constant function in time defined by

$$\begin{aligned} &(u_m)_\theta(0) = u_m^{(0)} \text{ and, for all } n = 0, \ldots, N_m - 1 \text{ and } t \in (t^{(n)}, t^{(n+1)}], \\ &(u_m)_\theta(t) = \theta u_m^{(n+1)} + (1-\theta) u_m^{(n)}. \end{aligned} \tag{C.17}$$

Set $\delta t_m^{(n+\frac{1}{2})} = t_m^{(n+1)} - t_m^{(n)}$ for $n = 0, \ldots, N_m - 1$, and define the discrete derivative $\delta_m u_m$ by:

$$\forall n = 0, \ldots, N_m - 1\,, \quad \text{for a.e. } t \in (t_m^{(n)}, t_m^{(n+1)})\,,\ \delta_m u_m(t) = \frac{u_m^{(n+1)} - u_m^{(n)}}{\delta t_m^{(n+\frac{1}{2})}}.$$

We assume that

(h1) The sequence $(X_m)_{m\in\mathbb{N}}$ is compactly embedded in B (see Definition C.4).
(h2) The sequence $(\|(u_m)_\theta\|_{L^\infty(0,T;X_m)})_{m\in\mathbb{N}}$ is bounded.
(h3) The sequence $(\|\delta_m u_m\|_{L^q(0,T;B)})_{m\in\mathbb{N}}$ is bounded for some $q>1$.
(h4) Setting $\eth_m = \max_{n=0,\dots,N_m-1} \eth_m^{(n+\frac12)}$, it holds $\lim_{m\to\infty}\eth_m = 0$.

Then, there exists $u\in C([0,T];B)$ such that, up to a subsequence,

$$\lim_{m\to\infty}\sup_{t\in[0,T]}\|(u_m)_\theta(t)-u(t)\|_B = 0. \tag{C.18}$$

Proof. Let

$$u_m^{(n+\theta)} = \theta u_m^{(n+1)} + (1-\theta)u_m^{(n)} \quad \text{and} \quad \delta_m^{(n+\frac12)}u_m := \frac{u_m^{(n+1)}-u_m^{(n)}}{\eth_m^{(n+\frac12)}}.$$

Take $n_2\ge n_1$ in $\{0,\dots,N_m-1\}$, $s_1\in(t^{(n_1)},t^{(n_1+1)}]$ and $s_2\in(t^{(n_2)},t^{(n_2+1)}]$. By writing a telescopic sum, we get

$$\begin{aligned}
&(u_m)_\theta(s_2)-(u_m)_\theta(s_1)\\
&\quad= u_m^{(n_2+\theta)}-u_m^{(n_1+\theta)}\\
&\quad= \sum_{n=n_1+1}^{n_2}\Big[u_m^{(n+\theta)}-u_m^{(n-1+\theta)}\Big]\\
&\quad= \sum_{n=n_1+1}^{n_2}\Big[\theta(u_m^{(n+1)}-u_m^{(n)})+(1-\theta)(u_m^{(n)}-u_m^{(n-1)})\Big]\\
&\quad= \sum_{n=n_1+1}^{n_2}\Big[\theta\eth_m^{(n+\frac12)}\delta_m^{(n+\frac12)}u_m+(1-\theta)\eth_m^{(n-\frac12)}\delta_m^{(n-\frac12)}u_m\Big].
\end{aligned} \tag{C.19}$$

It can easily be checked that this relation extends to the case $s_1=0$, $n_1=-1$ and $n_2\in\{0,\dots,N_m-1\}$ by defining $\eth_m^{(-\frac12)}=0$ and $\delta_m^{(-\frac12)}u_m=0$; consider for example $n_2=0$ and notice that

$$u_m^{(\theta)}-u_m^{(0)} = \theta(u_m^{(1)}-u_m^{(0)}) = \theta\eth_m^{(\frac12)}\delta_m^{(\frac12)}u_m.$$

By the discrete Hölder inequality (D.3) with $\omega_i=\eth_m^{(i\pm\frac12)}$, $b_i=1$ and $a_i=\|\delta_m^{(i\pm\frac12)}u_m\|_B$, since $\frac{q}{q'}=q-1$,

$$\begin{aligned}
&\Big(\sum_{n=n_1+1}^{n_2}\eth_m^{(n\pm\frac12)}\Big\|\delta_m^{(n+\frac12)}u_m\Big\|_B\Big)^q\\
&\quad\le\Big(\sum_{n=n_1+1}^{n_2}\eth_m^{(n\pm\frac12)}\Big)^{q-1}\Big(\sum_{n=n_1+1}^{n_2}\eth_m^{(n\pm\frac12)}\Big\|\delta_m^{(n\pm\frac12)}u_m\Big\|_B^q\Big)\\
&\quad\le\Big[t_m^{(n_2+\frac12\pm\frac12)}-t_m^{(n_1+\frac12\pm\frac12)}\Big]^{q-1}C^q,
\end{aligned} \tag{C.20}$$

where C is a bound of $(\|\delta_m u_m\|_{L^q(0,T;B)})_{m\in\mathbb{N}}$. Take the norm in B of (C.19), use the triangle inequality, then take the power q and use the convexity of $s \to s^q$. Invoking finally the estimate (C.20) yields

$$\begin{aligned}&\|(u_m)_\theta(s_2) - (u_m)_\theta(s_1)\|_B^q \\ &\qquad \le \theta C^q \left[t_m^{(n_2+1)} - t_m^{(n_1+1)}\right]^{q-1} + (1-\theta)C^q \left[t_m^{(n_2)} - t_m^{(n_1)}\right]^{q-1},\end{aligned}$$

where we set $t_m^{(-1)} = 0$. This gives C_1, that depends only on C and q, such that, for all $s_1, s_2 \in [0,T]$ and all $m \in \mathbb{N}$,

$$\begin{aligned}\|(u_m)_\theta(s_1) - (u_m)_\theta(s_2)\|_B &\le C_1(|s_2 - s_1| + \eth t_m)^{\frac{q-1}{q}} \\ &\le C_1|s_2 - s_1|^{\frac{q-1}{q}} + C_1 \eth t_m^{\frac{q-1}{q}}. \qquad \text{(C.21)}\end{aligned}$$

In the last line, we used the power-of-sums inequality (D.13).
This relation and (h4) show that $v_m = (u_m)_\theta$ satisfies Assumptions (C.13)–(C.14) in the discontinuous Ascoli–Arzelà theorem (Theorem C.11), with $K = [0,T]$ and $E = B$. The proof of Corollary C.12 is therefore complete if we can establish that, for all $s \in [0,T]$,

$$\{(u_m)_\theta(s) \,:\, m \in \mathbb{N}\} \text{ is relatively compact in } B. \qquad \text{(C.22)}$$

Assume first that $s > 0$. Since $(u_m)_\theta$ is piecewise constant on $(0,T]$, the $L^\infty(0,T;X_m)$ norm of $(u_m)_\theta$ is actually a supremum norm on $(0,T]$. Hence, $\|(u_m)_\theta(s)\|_{X_m} \le \|(u_m)_\theta\|_{L^\infty(0,T;X_m)}$ and Hypotheses (h1) and (h2) show that $((u_m)_\theta(s))_{m\in\mathbb{N}}$ is indeed relatively compact in B.
Let us now consider the case $s = 0$. Since (C.22) holds for any $s > 0$, by diagonal extraction we can find a subsequence, still denoted by $(u_m)_{m\in\mathbb{N}}$, such that, for any $k \in \mathbb{N}$ satisfying $k^{-1} \in (0,T]$, the sequence $((u_m)_\theta(k^{-1}))_{m\in\mathbb{N}}$ converges in B. We now prove that, along the same subsequence, $((u_m)_\theta(0))_{m\in\mathbb{N}}$ is a Cauchy sequence in B. This will conclude the proof that (C.22) holds for any $s = 0$.
Owing to (C.21) we have, for $(m,m') \in \mathbb{N}^2$ and $k \in \mathbb{N}$ such that $k^{-1} \le T$,

$$\begin{aligned}&\|(u_m)_\theta(0) - (u_{m'})_\theta(0)\|_B \\ &\quad \le \left\|(u_m)_\theta(0) - (u_m)_\theta(k^{-1})\right\|_B + \left\|(u_m)_\theta(k^{-1}) - (u_{m'})_\theta(k^{-1})\right\|_B \\ &\qquad + \left\|(u_{m'})_\theta(k^{-1}) - (u_{m'})_\theta(0)\right\|_B \\ &\quad \le 2C_1 k^{-\frac{q-1}{q}} + C_1 \eth t_m^{\frac{q-1}{q}} + C_1 \eth t_{m'}^{\frac{q-1}{q}} + \left\|(u_m)_\theta(k^{-1}) - (u_{m'})_\theta(k^{-1})\right\|_B.\end{aligned}$$

Given $\varepsilon > 0$, fix k such that $2C_1 k^{-\frac{q-1}{q}} < \varepsilon/4$. Using (h4) and the convergence of $((u_m)_\theta(k^{-1}))_{m\in\mathbb{N}}$, we can then find $m_0 = m_0(k) \in \mathbb{N}$ such that, if $m, m' \ge m_0$,

$$C_1 \eth t_m^{\frac{q-1}{q}} \le \frac{\varepsilon}{4}, \; C_1 \eth t_{m'}^{\frac{q-1}{q}} \le \frac{\varepsilon}{4} \text{ and } \left\|(u_m)_\theta(k^{-1}) - (u_{m'})_\theta(k^{-1})\right\|_B \le \frac{\varepsilon}{4}.$$

This shows that $\|(u_m)_\theta(0) - (u_{m'})_\theta(0)\|_B \leq \varepsilon$ whenever $m, m' \geq m_0$. The sequence $((u_m)_\theta(0))_{m\in\mathbb{N}}$ is therefore Cauchy in B, and the proof is complete. ∎

The following lemma states an equivalent condition for the uniform convergence of functions, which proves extremely useful to establish uniform-in-time convergence of numerical schemes for parabolic equations when no smoothness is assumed on the data.

Lemma C.13. *Let (K, d_K) be a compact metric space, (E, d_E) be a metric space and $(\mathcal{F}(K,E), d_\mathcal{F})$ be as in Definition C.9. Let $(v_m)_{m\in\mathbb{N}}$ be a sequence in $\mathcal{F}(K,E)$, and let $v \in \mathcal{F}(K,E)$. The following properties are equivalent.*

1. *$v \in C(K,E)$ and $v_m \to v$ for $d_\mathcal{F}$,*
2. *for any $s \in K$ and for any sequence $(s_m)_{m\in\mathbb{N}} \subset K$ converging to s for d_K, we have $v_m(s_m) \to v(s)$ for d_E.*

Proof.
Step 1: Property 1 implies Property 2.
For any sequence $(s_m)_{m\in\mathbb{N}}$ converging to s,

$$\begin{aligned} d_E(v_m(s_m), v(s)) &\leq d_E(v_m(s_m), v(s_m)) + d_E(v(s_m), v(s)) \\ &\leq d_\mathcal{F}(v_m, v) + d_E(v(s_m), v(s)). \end{aligned}$$

The right-hand side tends to 0 by definition of $v_m \to v$ for $d_\mathcal{F}$, and by continuity of v.

Step 2: Property 2 implies Property 1.
Let us first prove that $v \in C(K,E)$. Let $(s_m)_{m\in\mathbb{N}} \subset K$ be a sequence converging to s for d_K. Since for any $t \in K$ the sequence $(v_n(t))_{n\in\mathbb{N}}$ converges to $v(t)$, we can find $\varphi(0) \in \mathbb{N}$ such that $d_E(v_{\varphi(0)}(s_0), v(s_0)) < 1$. Assuming that, for $n \in \mathbb{N}^\star$, $\varphi(n-1) \in \mathbb{N}$ is given, we can also find $\varphi(n) \in \mathbb{N}$ such that $\varphi(n) > \varphi(n-1)$ and $d_E(v_{\varphi(n)}(s_n), v(s_n)) < 1/(n+1)$.
We define the sequence $(\hat{s}_m)_{m\in\mathbb{N}}$ by $\hat{s}_m = s_n$ if $m = \varphi(n)$ for some $n \in \mathbb{N}$, and $\hat{s}_m = s$ if $m \notin \varphi(\mathbb{N})$. The sequence $(\hat{s}_m)_{m\in\mathbb{N}}$ is constructed by interlacing the sequence $(s_m)_{m\in\mathbb{N}}$ and the constant sequence equal to s. Hence, $\hat{s}_m \to s$ as $m \to \infty$ and, by assumption, $(v_m(\hat{s}_m))_{m\in\mathbb{N}}$ converges to $v(s)$. The sequence $(v_{\varphi(n)}(s_n))_{n\in\mathbb{N}}$ is a subsequence of $(v_m(\hat{s}_m))_{m\in\mathbb{N}}$, and it therefore also converges to $v(s)$. A triangle inequality then gives

$$\begin{aligned} d_E(v(s_n), v(s)) &\leq d_E(v(s_n), v_{\varphi(n)}(s_n)) + d_E(v_{\varphi(n)}(s_n), v(s)) \\ &\leq \frac{1}{n+1} + d_E(v_{\varphi(n)}(s_n), v(s)), \end{aligned}$$

which shows that $v(s_n) \to v(s)$. This completes the proof that $v \in C(K,E)$.

We now prove by way of contradiction that $v_m \to v$ for $d_\mathcal{F}$. If $(v_m)_{m\in\mathbb{N}}$ does not converge to v for $d_\mathcal{F}$, then there exists $\varepsilon > 0$ and a subsequence $(v_{m_k})_{k\in\mathbb{N}}$,

such that, for any $k \in \mathbb{N}$, $\sup_{s\in K} d_E(v_{m_k}(s), v(s)) \geq \varepsilon$. We can then find a sequence $(r_k)_{k\in\mathbb{N}} \subset K$ such that, for any $k \in \mathbb{N}$,

$$d_E(v_{m_k}(r_k), v(r_k)) \geq \varepsilon/2. \tag{C.23}$$

K being compact, up to another subsequence, denoted the same way, we can assume that $r_k \to s$ in K as $k \to \infty$. As before, we then construct a sequence $(s_m)_{m\in\mathbb{N}}$ converging to s, such that $s_{m_k} = r_k$ for all $k \in \mathbb{N}$ and $s_m = s$ if $m \notin \{r_k : k \in \mathbb{N}\}$. By assumption, $v_m(s_m) \to v(s)$ in E and, by continuity of v, $v(s_m) \to v(s)$ in E. A triangle inequality then shows that $d_E(v_m(s_m), v(s_m)) \to 0$, which contradicts (C.23) and concludes the proof.

■

Uniform-in-time convergence of numerical solutions to schemes for parabolic equations often starts with a weak convergence with respect to the time variable. This weak convergence is then used to prove a stronger convergence. The following definition and proposition recall standard notions related to the weak topology on $L^2(\Omega)$. The inner product in $L^2(\Omega)$ is denoted by $\langle\cdot,\cdot\rangle_{L^2(\Omega)}$.

Definition C.14 (Uniform-in-time $L^2(\Omega)$-weak convergence). *Let $(u_m)_{m\in\mathbb{N}}$ and u be functions $[0,T] \to L^2(\Omega)$. We say that $(u_m)_{m\in\mathbb{N}}$ converges weakly in $L^2(\Omega)$ uniformly on $[0,T]$ to u if, for all $\varphi \in L^2(\Omega)$, as $m \to \infty$ the sequence of functions $t \in [0,T] \to \langle u_m(t), \varphi\rangle_{L^2(\Omega)}$ converges uniformly on $[0,T]$ to the function $t \in [0,T] \to \langle u(t), \varphi\rangle_{L^2(\Omega)}$.*

Proposition C.15. *Let E be a closed bounded ball in $L^2(\Omega)$ and let $\{\varphi_l : l \in \mathbb{N}\}$ be a dense set in $L^2(\Omega)$. Then, on E, the weak topology of $L^2(\Omega)$ is given by the metric*

$$d_E(v,w) = \sum_{l\in\mathbb{N}} \frac{\min(1, |\langle v-w, \varphi_l\rangle_{L^2(\Omega)}|)}{2^l}. \tag{C.24}$$

Moreover, a sequence of functions $u_m : [0,T] \to E$ converges uniformly to $u : [0,T] \to E$ for the weak topology of $L^2(\Omega)$ if and only if, as $m \to \infty$, the sequence of functions $d_E(u_m, u) : [0,T] \to [0,\infty)$ converges uniformly to 0.

Proof. The sets $E_{\varphi,\varepsilon} = \{v \in E : |\langle v, \varphi\rangle_{L^2(\Omega)}| < \varepsilon\}$, for $\varphi \in L^2(\Omega)$ and $\varepsilon > 0$, define a neighbourhood basis of 0 for the $L^2(\Omega)$-weak topology on E. A neighbourhood basis of any other points is obtained by translation of this particular basis. If R is the radius of the ball E then, for any $\varphi \in L^2(\Omega)$, $l \in \mathbb{N}$ and $v \in E$,

$$|\langle v, \varphi\rangle_{L^2(\Omega)}| \leq R\,\|\varphi - \varphi_l\|_{L^2(\Omega)} + |\langle v, \varphi_l\rangle_{L^2(\Omega)}|.$$

By density of $\{\varphi_l : l \in \mathbb{N}\}$ we can select $l \in \mathbb{N}$ such that $\|\varphi - \varphi_l\|_{L^2(\Omega)} \leq \varepsilon/(2R)$, and we then see that $E_{\varphi_l,\varepsilon/2} \subset E_{\varphi,\varepsilon}$. Hence, a neighbourhood basis of 0 in E for the $L^2(\Omega)$-weak topology is also given by $(E_{\varphi_l,\varepsilon})_{l\in\mathbb{N},\,\varepsilon>0}$.

From the definition of d_E we see that, for any $l \in \mathbb{N}$, $\min(1, |\langle v, \varphi_l \rangle_{L^2(\Omega)}|) \leq 2^l d_E(0, v)$. If $d_E(0, v) < 2^{-l}$ this shows that $|\langle v, \varphi_l \rangle_{L^2(\Omega)}| \leq 2^l d_E(0, v)$ and therefore that

$$B_{d_E}(0, \min(2^{-l}, \varepsilon 2^{-l})) \subset E_{\varphi_l, \varepsilon}.$$

Hence, any neighbourhood of 0 in E for the $L^2(\Omega)$-weak topology is a neighbourhood of 0 for d_E. Conversely, for any $\varepsilon > 0$, selecting $N \in \mathbb{N}$ such that $\sum_{l \geq N+1} 2^{-l} < \varepsilon/2$ gives, from the definition (C.24) of d_E,

$$\bigcap_{l=1}^{N} E_{\varphi_l, \varepsilon/4} \subset B_{d_E}(0, \varepsilon).$$

Hence, any ball for d_E centred at 0 is a neighbourhood of 0 for the $L^2(\Omega)$-weak topology. Since d_E and $L^2(\Omega)$-weak neighbourhoods are invariant by translation, this concludes the proof that this weak topology is identical to the topology generated by d_E.

The conclusion on weak uniform convergence of sequences of functions follows from the preceding result, and more precisely by noticing that all previous inclusions are, when applied to $u_m(t) - u(t)$, uniform with respect to $t \in [0, T]$.

■

D

Technical results

D.1 Standard notations, inequalities and relations

We gather here a few notations and standard inequalities that are used throughout the book, sometimes implicitly.

D.1.1 $\mathbb{R}^d$ and measures

For $\boldsymbol{\xi}$ and $\boldsymbol{\eta}$ vectors in $\mathbb{R}^d$, $\boldsymbol{\xi} \cdot \boldsymbol{\eta}$ is the Euclidean (dot) product of $\boldsymbol{\xi}$ and $\boldsymbol{\eta}$, and $|\boldsymbol{\xi}|$ denotes the Euclidean norm of $\boldsymbol{\xi}$. If M is a $d \times d$ matrix, we also denote by $|M|$ the norm of M induced by the Euclidean norm on $\mathbb{R}^d$, that is,

$$|M| = \sup_{\boldsymbol{\xi} \in \mathbb{R}^d \setminus \{0\}} \frac{|M\boldsymbol{\xi}|}{|\boldsymbol{\xi}|}.$$

The Lebesgue measure of a measurable subset A of $\mathbb{R}^d$ is written $|A|$. The integral of a function $f : A \to \mathbb{R}$ with respect to this measure is written

$$\int_A f(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}.$$

If B is a measurable subset of an hyperplane of $\mathbb{R}^d$, then $|B|$ denotes the $(d-1)$-dimensional Lebesgue measure of B in that hyperplane. The integral of a function $g : B \to \mathbb{R}$ with respect to this measure is written

$$\int_B g(\boldsymbol{x}) \mathrm{d}\gamma(\boldsymbol{x}).$$

D.1.2 Lebesgue and Sobolev spaces

For $q \in [1, +\infty]$ and A a measurable subset of $\mathbb{R}^d$, $L^q(A)$ is the Lebesgue space with exponent q, that is the set of (class of) measurable functions from

A to $\mathbb{R}$ such that $\|u\|_{L^q(A)} < +\infty$, where $\|\cdot\|_{L^q(A)}$ is the usual norm defined by

$$\|u\|_{L^q(A)} = \begin{cases} \left(\int_A |u(\boldsymbol{x})|^q \mathrm{d}\boldsymbol{x}\right)^{1/q} & \text{if } q < +\infty, \\ \inf\{M \geq 0 \,:\, |u(\boldsymbol{x})| \leq M \text{ for a.e. } \boldsymbol{x} \in A\} & \text{if } q = +\infty. \end{cases}$$

For a vector-valued function $u : A \to \mathbb{R}^d$, we sometimes write $\|u\|_{L^p(A)^d}$ for $\| \, |u| \, \|_{L^p(A)}$.
If $A = \Omega$ is an open set of $\mathbb{R}^d$ and $k \in \mathbb{N}$, the standard Sobolev space is denoted as usual by

$$W^{k,q}(\Omega) := \Big\{ u \in L^q(\Omega) \,:\, \partial^\alpha u \in L^q(\Omega) \text{ for all } \alpha = (\alpha_1, \ldots, \alpha_d) \in \mathbb{N}^d \\ \text{such that } |\alpha| := \textstyle\sum_{i=1}^n \alpha_i \leq k \Big\},$$

and is endowed with the norm

$$\|u\|_{W^{k,q}(\Omega)} = \begin{cases} \left(\displaystyle\sum_{\alpha \in \mathbb{N}^d,\, |\alpha| \leq k} \|\partial^\alpha u\|^q_{L^q(\Omega)} \right)^{1/q} & \text{if } q < +\infty, \\ \max_{\alpha \in \mathbb{N}^d,\, |\alpha| \leq k} \|\partial^\alpha u\|_{L^\infty(\Omega)} & \text{if } q = +\infty. \end{cases}$$

The space $W^{k,q}_0(\Omega)$ is the closure in $W^{k,q}(\Omega)$ of the space $C^\infty_c(\Omega)$ of infinitely differentiable functions with compact support in Ω.
The space of vector-valued functions whose divergence (but not necessarily all derivatives) belongs to $L^q(\Omega)$ is

$$W^q_{\mathrm{div}}(\Omega) = \{\boldsymbol{\varphi} \in L^q(\Omega)^d \,:\, \mathrm{div}\boldsymbol{\varphi} \in L^q(\Omega)\}.$$

It is endowed with the norm

$$\|\boldsymbol{\varphi}\|_{W^q_{\mathrm{div}}(\Omega)} = \|\boldsymbol{\varphi}\|_{L^q(\Omega)^d} + \|\mathrm{div}\boldsymbol{\varphi}\|_{L^q(\Omega)}.$$

In the particular case $q = 2$, we use the standard notations $H^k(\Omega) := W^{k,2}(\Omega)$, $H^k_0(\Omega) := W^{k,2}_0(\Omega)$ and $H_{\mathrm{div}}(\Omega) := W^2_{\mathrm{div}}(\Omega)$.

D.1.3 Hölder inequalities

Let $(a_i)_{i\in I}$ and $(b_i)_{i\in I}$ be finite families of real numbers, and let $(p, p') \in (1, \infty)^2$ be such that $\frac{1}{p} + \frac{1}{p'} = 1$ (p and p' are *conjugate* exponents). Then the Hölder inequality for sums is

$$\sum_{i\in I} |a_i b_i| \leq \left(\sum_{i\in I} |a_i|^p \right)^{\frac{1}{p}} \left(\sum_{i\in I} |b_i|^{p'} \right)^{\frac{1}{p'}}. \tag{D.1}$$

It is frequently used after the introduction of some non-zero real numbers $(d_i)_{i\in I}$ in the product a_ib_i. More precisely, writing $a_ib_i = (a_id_i)(\frac{b_i}{d_i})$ and applying (D.1) to this new product, we have

$$\sum_{i\in I} |a_ib_i| \leq \left(\sum_{i\in I} |a_i|^p |d_i|^p\right)^{\frac{1}{p}} \left(\sum_{i\in I} \frac{|b_i|^{p'}}{|d_i|^{p'}}\right)^{\frac{1}{p'}}. \tag{D.2}$$

Another frequent use is to evenly split an existing weight. If $(w_i)_{i\in I}$ are non-negative numbers, writing $w_i|a_ib_i| = (w_i^{1/p}|a_i|)(w_i^{1/p'}|b_i|)$ and using (D.1) leads to

$$\sum_{i\in I} w_i|a_ib_i| \leq \left(\sum_{i\in I} w_i|a_i|^p\right)^{\frac{1}{p}} \left(\sum_{i\in I} w_i|b_i|^{p'}\right)^{\frac{1}{p'}}. \tag{D.3}$$

Using both weights and the introduction of non-zero numbers, we also have

$$\sum_{i\in I} w_i|a_ib_i| \leq \left(\sum_{i\in I} w_i|a_i|^p|d_i|^p\right)^{\frac{1}{p}} \left(\sum_{i\in I} w_i\frac{|b_i|^{p'}}{|d_i|^{p'}}\right)^{\frac{1}{p'}}. \tag{D.4}$$

The Hölder inequalities are also valid in Lebesgue spaces over a measurable set (X, μ). For example, the equivalent of (D.1) for integrals is: if $f, g : X \to \mathbb{R}$ are measurable functions, then

$$\int_X |fg|\mathrm{d}\mu \leq \left(\int_X |f|^p\mathrm{d}\mu\right)^{\frac{1}{p}} \left(\int_X |g|^{p'}\mathrm{d}\mu\right)^{\frac{1}{p'}}. \tag{D.5}$$

In other words, $\|fg\|_{L^1(X)} \leq \|f\|_{L^p(X)} \|g\|_{L^{p'}(X)}$. If X has a finite measure, this is sometimes used with $g \equiv 1$ to give

$$\int_X |f|\mathrm{d}\mu \leq \left(\int_X |f|^p\mathrm{d}\mu\right)^{\frac{1}{p}} \mu(X)^{\frac{1}{p'}} = \left(\int_X |f|^p\mathrm{d}\mu\right)^{\frac{1}{p}} \mu(X)^{1-\frac{1}{p}}. \tag{D.6}$$

A variant consists in taking $q > r > 1$ and in applying this to $|f|^r$, instead of f, with the exponent $p = q/r$. This leads to

$$\|f\|_{L^r(X)} \leq \mu(X)^{\frac{1}{r}-\frac{1}{q}} \|f\|_{L^q(X)}. \tag{D.7}$$

D.1.4 Young inequality

For $a, b \geq 0$ and (p, p') conjugate exponents, the Young inequality reads

$$ab \leq \frac{1}{p}a^p + \frac{1}{p'}b^{p'}. \tag{D.8}$$

As in the Hölder inequality, it is standard to introduce a (usually small) parameter when applying Young's inequality. Taking $\varepsilon > 0$ and writing $ab = (\varepsilon^{1/p}a)(\varepsilon^{-1/p}b)$, we obtain

$$ab \leq \frac{\varepsilon}{p}a^p + \frac{1}{p'\varepsilon^{p'/p}}b^{p'}. \tag{D.9}$$

D.1.5 Jensen inequality

Let A be a measurable subset of $\mathbb{R}^d$ with non-zero measure, and $\Psi : \mathbb{R} \to \mathbb{R}$ be a convex function. If f is integrable on A, then the Jensen inequality states that

$$\Psi\left(\frac{1}{|A|}\int_A f(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\right) \leq \frac{1}{|A|}\int_A \Psi(f(\boldsymbol{x}))\mathrm{d}\boldsymbol{x}. \tag{D.10}$$

Although mostly used for integrals over subsets of $\mathbb{R}^d$, Jensen's inequality is of course also valid for sums. If $w_i \geq 0$ are such that $W = \sum_{i\in I} w_i > 0$, then

$$\Psi\left(\frac{1}{W}\sum_{i\in I} w_i a_i\right) \leq \frac{1}{W}\sum_{i\in I} w_i \Psi(a_i). \tag{D.11}$$

D.1.6 Power of sums inequality

The last inequality we want to mention is a simple one for powers of a sum. If $\alpha \geq 0$ and $a, b \geq 0$, a basic estimate is

$$(a+b)^\alpha \leq 2^\alpha a^\alpha + 2^\alpha b^\alpha.$$

This generic inequality can be improved by looking separately at the cases $\alpha \leq 1$ and $\alpha \geq 1$. Using the convexity of $s \mapsto s^\alpha$ if $\alpha \geq 1$, we actually have $(\frac{a+b}{2})^\alpha \leq \frac{1}{2}a^\alpha + \frac{1}{2}b^\alpha$, that is

$$\forall \alpha \geq 1\,,\ (a+b)^\alpha \leq 2^{\alpha-1}a^\alpha + 2^{\alpha-1}b^\alpha. \tag{D.12}$$

If $\alpha \leq 1$, the mapping $s \to (1+s)^\alpha - s^\alpha$ is non-increasing and takes value 1 at $s = 0$. Hence, $(1+s)^\alpha \leq 1 + s^\alpha$. Applied to $s = b/a$, this gives

$$\forall \alpha \leq 1\,,\ (a+b)^\alpha \leq a^\alpha + b^\alpha. \tag{D.13}$$

This inequality is often applied with $\alpha = 1/2$.

An easy generalisation of the above inequalities can be obtained for sums of more than two terms. For example, if $\alpha \geq 1$ and $(a_i)_{i=1,\ldots,\ell}$ are non-negative numbers,

$$\left(\sum_{i=1}^{\ell} a_i\right)^\alpha \leq \ell^{\alpha-1}\sum_{i=1}^{\ell} a_i^\alpha. \tag{D.14}$$

D.1.7 Discrete integration-by-parts (summation-by-parts)

Let $(a_n)_{n=0,\ldots,N}$ and $(b_n)_{n=0,\ldots,N}$ be two families of real numbers. Splitting the sum and re-indexing the first term (with $j = n+1$), we have

$$\sum_{n=0}^{N-1}(a_{n+1} - a_n)b_n = \sum_{n=0}^{N-1} a_{n+1}b_n - \sum_{n=0}^{N-1} a_n b_n$$

$$
\begin{aligned}
&= \sum_{n=0}^{N-1} a_{n+1} b_n - \left(a_0 b_0 + \sum_{n=0}^{N-1} a_{n+1} b_{n+1} - a_N b_N \right) \\
&= \sum_{n=0}^{N-1} a_{n+1}(b_n - b_{n+1}) + a_N b_N - a_0 b_0.
\end{aligned}
$$

To summarise,

$$
\sum_{n=0}^{N-1} (a_{n+1} - a_n) b_n = - \sum_{n=0}^{N-1} a_{n+1}(b_{n+1} - b_n) + a_N b_N - a_0 b_0. \tag{D.15}
$$

The quantities $a_{n+1} - a_n$ and $b_{n+1} - b_n$ can be seen as discrete derivatives of $(a_n)_{n=0,\dots,N}$ and $(b_n)_{n=0,\dots,N}$. Relation (D.15) is therefore a form of discrete integration-by-parts, with $a_N b_N$ and $a_0 b_0$ playing the role of the boundary (integrated) terms.

Set, for example, $b_{N+1} = 0$ and let $\widetilde{b}_n = b_{n+1}$ for $n = 0, \dots, N$. Applying (D.15) to $(\widetilde{b}_n)_{n=0,\dots,N}$ instead of $(b_n)_{n=0,\dots,N}$ gives

$$
\begin{aligned}
&\sum_{n=0}^{N-1} (a_{n+1} - a_n) b_{n+1} = \\
&\qquad = - \sum_{n=0}^{N-1} a_{n+1}(b_{n+2} - b_{n+1}) - a_0 b_1 \\
&\qquad = - \sum_{n=1}^{N} a_n (b_{n+1} - b_n) - a_0 b_1 \\
&\qquad = - \sum_{n=0}^{N-1} a_n (b_{n+1} - b_n) + a_0(b_1 - b_0) - a_N(b_{N+1} - b_N) - a_0 b_1.
\end{aligned}
$$

In other words,

$$
\sum_{n=0}^{N-1} (a_{n+1} - a_n) b_{n+1} = - \sum_{n=0}^{N-1} a_n (b_{n+1} - b_n) + a_N b_N - a_0 b_0. \tag{D.16}
$$

This is the equivalent of (D.15) with an offset of the second family $(b_n)_{n=0,\dots,N}$.

By creating a convex combination of (D.15) and (D.16) we arrive at a formula that is instrumental when dealing with time terms in θ-schemes. If $(x_n)_{n=0,\dots,N}$ is a family of numbers and $\nu \in [0,1]$, for all $n = 0, \dots, N-1$ we set $x_{n+\nu} = \nu x_{n+1} + (1-\nu) x_n$. Adding up $\nu \times$ (D.16) and $(1-\nu) \times$ (D.15) yields

$$
\sum_{n=0}^{N-1} (a_{n+1} - a_n) b_{n+\nu} = - \sum_{n=0}^{N-1} (\nu a_n + (1-\nu) a_{n+1})(b_{n+1} - b_n) + a_N b_N - a_0 b_0.
$$

In other words,

$$\sum_{n=0}^{N-1}(a_{n+1}-a_n)b_{n+\nu} = -\sum_{n=0}^{N-1} a_{n+(1-\nu)}(b_{n+1}-b_n) + a_N b_N - a_0 b_0. \quad \text{(D.17)}$$

D.2 Topological degree

The following theorem, which can be found in [88], is a consequence of the theory of the topological degree [56].

Theorem D.1 (Application of the topological degree, finite dimensional case). *Let V be a finite dimensional vector space on $\mathbb{R}$ and $\Phi : V \to V$ be a continuous function. Assume that there exists a continuous function $\Psi : V \times [0,1] \to V$ satisfying:*

1. *$\Psi(\cdot, 1) = \Phi$.*
2. *There exists $R > 0$ such that, for any $(v, \rho) \in V \times [0,1]$, if $\Psi(v, \rho) = 0$ then $\|v\|_V \neq R$.*
3. *$\Psi(\cdot, 0)$ is affine and the equation $\Psi(v, 0) = 0$ has a solution $v \in V$ such that $\|v\|_V < R$.*

Then, there exists at least one $v \in V$ such that $\Phi(v) = 0$ and $\|v\|_V < R$.

As an easy consequence of this, we have the Brouwer fixed point theorem.

Theorem D.2 (Brouwer fixed point). *Let V be a finite dimensional vector space on $\mathbb{R}$, B a closed ball in V and $F : B \to B$ be continuous. Then F has a fixed point,* i.e. *there exists $v \in B$ such that $F(v) = v$.*

Proof. Without loss of generality, we can assume that B is centred at 0 and has radius $r > 0$. Let θ_r be the retraction of V on B, that is $\theta_r(v) = v$ if $v \in B$ and $\theta_r(v) = rv/\|v\|_V$ if $v \notin B$. Set $\Phi(v) = v - F(\theta_r(v))$ and $\Psi(v,t) = v - tF(\theta_r(v))$. Then $\Phi : V \to V$ is continuous, $\Phi = \Psi(\cdot, 1)$, $\Psi(\cdot, 0)$ is affine and the equation $\Psi(v, 0) = 0$ has the unique solution $v = 0 \in B$.
Moreover, if $\Psi(v,t) = 0$ then $v = tF(\theta_R(v)) \in tB \subset B$, and thus $\|v\|_V \leq r < r + 1 =: R$. Theorem D.1 then shows that $\Phi(v) = 0$ has a solution in V, that is that there exists $v \in V$ such that $v = F(\theta_r(v))$. Since F takes values in B, $v \in B$ and thus $v = F(v)$. ∎

D.3 Derivation and convergence in the sense of distributions

This section gives the generalisation of the notion of derivative that is used in this book. We refer to, e.g., [107] for more details on this subject.

Given an open set Ω of $\mathbb{R}^d$, the following lemma allows to merge the (class of) function(s) $f \in L^1_{\text{loc}}(\Omega)$ with the linear mapping $T_f : C_c^\infty(\Omega) \to \mathbb{R}$ defined by

$$T_f(\varphi) =: \int_\Omega f(\boldsymbol{x})\varphi(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \text{ for any } \varphi \in C_c^\infty(\Omega), \tag{D.18}$$

Recall that $f \in L^1_{loc}(\Omega)$ means that for any compact subset K of Ω, the restriction $f|_K$ of f to K belongs to $L^1(K)$.

Lemma D.3 (Almost everywhere equality). *Let Ω be an open subset of $\mathbb{R}^d$, $d \geq 1$, and let f and $g \in L^1_{loc}(\Omega)$. Then:*

$$\left[\forall \varphi \in C_c^\infty(\Omega), \int_\Omega f(\boldsymbol{x})\varphi(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = \int_\Omega g(\boldsymbol{x})\varphi(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\right] \Longleftrightarrow [f = g \text{ a.e. on } \Omega.]$$

Let T be a linear mapping from $C_c^\infty(\Omega)$ to $\mathbb{R}$ and $\varphi \in C_c^\infty(\Omega)$, then the real number $T(\varphi)$ is called the action of T on φ. Lemma D.3 allows the definition of a weak derivative of a L^1_{loc} function in the following way:

Definition D.4 (Derivatives in the sense of distributions, weak derivative). *Let Ω be an open subset of $\mathbb{R}^d$, $d \geq 1$ and $1 \leq i \leq d$. Let T be a linear form on $C_c^\infty(\Omega)$; its i-th derivative D_iT in the sense of distributions is defined by:*

$$D_iT(\varphi) =: -T(\partial_i\varphi),\ \forall \varphi \in C_c^\infty(\Omega), \tag{D.19}$$

where $\partial_i\varphi$ is the classical partial derivative of φ with respect to its i-th variable. Let $f \in L^1_{loc}(\Omega)$, and T_f is the related linear form on $C_c^\infty(\Omega)$ defined by (D.18); then identifying f and T_f, the i-th derivative $D_if := D_iT_f$ in the sense of distributions is given by:

$$D_if(\varphi) =: -\int_\Omega f(\boldsymbol{x})\partial_i\varphi(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \tag{D.20}$$

Note that if $f \in C^1(\Omega)$, then D_if is nothing but ∂_if, merging ∂_if and T_{∂_if} (which is the linear form on $C_c^\infty(\Omega)$ induced by ∂_if). The derivative in the sense of distributions is a generalisation of the notion of derivative. If the linear form D_if can be identified as a locally integrable function in the sense of Lemma D.3, then f is said to admit a weak derivative.

Definition D.5 (Convergence in the sense of distributions). *Let Ω be an open subset of $\mathbb{R}^d$, $d \geq 1$, $(T_n)_{n\in\mathbb{N}}$ be a sequence linear forms on $C_c^\infty(\Omega)$ and T a linear form on $C_c^\infty(\Omega)$. Then T_n converges to T pointwise in the set of mappings from $C_c^\infty(\Omega)$ to $\mathbb{R}$, as $n \to +\infty$, if*

$$T_n(\varphi) \to T(\varphi) \text{ for any } \varphi \in C_c^\infty(\Omega). \tag{D.21}$$

Such a converging sequence is said to converge in the sense of distributions.

Remark D.6 (Distribution theory)
In the framework of the distribution theory, the space $C_c^\infty(\Omega)$ is equipped with a (rather complicated) topology and usually denoted $\mathcal{D}(\Omega)$. This topology is actually not needed for most applications in PDEs. Even though the distribution theory involves a smaller space consisting of the *continuous* linear mappings from $C_c^\infty(\Omega)$ to $\mathbb{R}$ (usually denoted by $\mathcal{D}'(\Omega)$), the notion of convergence is still given by (D.21). Similarly, when $C_c^\infty(\Omega)$ is equipped with its topology, the notion of derivative in the sense of distribution coincides with that given by (D.19).

Lemma D.7 (Weak convergence and convergence of the derivatives). *Let Ω be an open subset of $\mathbb{R}^d$, $d \geq 1$, $p \in (1, +\infty)$, $(f_n)_{n\in\mathbb{N}} \subset L^p(\Omega)$, and $f \in L^p(\Omega)$, such that $f_n \to f$ weakly in $L^p(\Omega)$ as $n \to +\infty$, that is:*

$$\int_\Omega f_n(\boldsymbol{x})g(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \to \int_\Omega f(\boldsymbol{x})g(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \text{ as } n \to +\infty\,, \text{ for any } g \in L^{p'}(\Omega),$$

with $1/p + 1/p' = 1$. Then, identifying f_n (resp. f) with a linear form T_{f_n} (resp. T_f) on $C_c^\infty(\Omega)$, T_{f_n} tends to T_f in the sense of distributions and $D_iT_{f_n}$ tends to D_iT_f in the sense of distributions. Hence, identifying T_{f_n} with f_n and T_f with f,

$$D_if_n \to D_if \text{ in the sense of distributions as } n \to +\infty.$$

D.4 Weak and strong convergence results

Lemma D.8 (Weak-strong convergence). *Let $p \in [1, \infty)$ and $p' = \frac{p}{1-p}$ be the conjugate exponent of p. Let (X, μ) be a measured space. If $f_n \to f$ strongly in $L^p(X)^d$ and $g_n \to g$ weakly in $L^{p'}(X)^d$, then*

$$\int_X f_n \cdot g_n \mathrm{d}\mu \to \int_X f \cdot g \mathrm{d}\mu.$$

Proof. By Banach–Steinhaus theorem, $(g_n)_{n\in\mathbb{N}}$ is bounded, say by C, in $L^{p'}(X)^d$. We therefore write, using Hölder's inequality,

$$\begin{aligned}
&\left|\int_X f_n \cdot g_n \mathrm{d}\mu - \int_X f \cdot g \mathrm{d}\mu\right| \\
&\qquad = \left|\int_X (f_n - f) \cdot g_n \mathrm{d}\mu + \int_X f \cdot (g_n - g)\mathrm{d}\mu\right| \\
&\qquad \leq \|f_n - f\|_{L^p(X)^d} \|g_n\|_{L^{p'}(X)^d} + \left|\int_X f \cdot (g_n - g)\mathrm{d}\mu\right| \\
&\qquad \leq C \|f_n - f\|_{L^p(X)^d} + \left|\int_X f \cdot (g_n - g)\mathrm{d}\mu\right|.
\end{aligned}$$

The first term converges to 0 by strong convergence of $(f_n)_{n\in\mathbb{N}}$, and the second term tends to 0 by weak convergence of $(g_n)_{n\in\mathbb{N}}$. ∎

We now state a lemma that is particularly useful to pass to the limit in terms involving solution-dependent diffusion tensors.

Lemma D.9 (Non-linear strong convergence). *Let (X,μ) be a measure space and $\Lambda : X \times \mathbb{R} \to \mathcal{M}_d(\mathbb{R})$ be a Caratheodory function (*i.e. *$\Lambda(x,\cdot)$ is continuous for a.e. $x \in X$, and $\Lambda(\cdot,s)$ is measurable for all $s \in \mathbb{R}$), that is bounded over $X \times \mathbb{R}$. Assume that, as $n \to \infty$, $u_n \to u$ in $L^1(X)$ and that $H_n \to H$ in $L^p(X)^d$, for some $p \in [1,\infty)$. Then, $\Lambda(\cdot,u_n)H_n \to \Lambda(\cdot,u)H$ in $L^p(X)^d$.*

Proof. Up to a subsequence, we can assume that $u_n \to u$ a.e. on X. Then, by continuity of Λ with respect to its second argument, $\Lambda(\cdot,u_n) \to \Lambda(\cdot,u)$ a.e. on X. Still extracting a subsequence, we have $H_n \to H$ a.e. on X, and $|H_n| \le g$ a.e. on X for some fixed $g \in L^p(X)$.
Then, $\Lambda(\cdot,u_n)H_n \to \Lambda(\cdot,u)H$ a.e. on X and, denoting by C an upper bound of Λ, $|\Lambda(\cdot,u_n)H_n| \le C|H_n| \le Cg \in L^p(X)$. The dominated convergence theorem therefore gives $\Lambda(\cdot,u_n)H_n \to \Lambda(\cdot,u)H$ in $L^p(X)^d$.
This convergence is established up to a subsequence, but since the reasoning can be made starting from any subsequence of $(\Lambda(\cdot,u_n)H_n)_{n\in\mathbb{N}}$ and since the limit is unique, this shows that the whole sequence converges. ∎

D.5 Minty trick and convexity inequality

The next lemma, whose proof is based on a technique called in the literature as the Minty trick [131], is used to identify limits of non-linear functions of weakly convergent sequences.

Lemma D.10 (Minty trick). *Let $\beta,\zeta \in C^0(\mathbb{R})$ be two non-decreasing functions such that $\beta(0)=\zeta(0)=0$, $\beta+\zeta$ is strictly increasing, and $\lim_{s\to\pm\infty}(\beta+\zeta)(s)=\pm\infty$. Let (X,μ) be a measurable set and let $(w_n)_{n\in\mathbb{N}} \subset L^2(X)$ be such that*

(i) $(\beta(w_n))_{n\in\mathbb{N}} \subset L^2(X)$ and there exists $\overline{\beta} \in L^2(X)$ such that $\beta(w_n) \to \overline{\beta}$ weakly in $L^2(X)$ as $n \to \infty$;
(ii) $(\zeta(w_n))_{n\in\mathbb{N}} \subset L^2(X)$ and there exists $\overline{\zeta} \in L^2(X)$ such that $\zeta(w_n) \to \overline{\zeta}$ weakly in $L^2(X)$ as $n \to \infty$;
(iii) there holds:

$$\liminf_{n\to\infty} \int_X \beta(w_n)\zeta(w_n)\mathrm{d}\mu \le \int_X \overline{\beta}\,\overline{\zeta}\mathrm{d}\mu. \tag{D.22}$$

Then,

$$\overline{\beta} = \beta(w) \text{ and } \overline{\zeta} = \zeta(w) \text{ a.e. in } X, \tag{D.23}$$

where

$$w = \left(\frac{\beta+\zeta}{2}\right)^{-1}\left(\frac{\overline{\beta}+\overline{\zeta}}{2}\right).$$

Proof. Notice first that the assumptions on β and ζ ensure that $\frac{\beta+\zeta}{2} : \mathbb{R} \to \mathbb{R}$ is an homeomorphism. Hence, w is well defined. Since $\beta(0) = \zeta(0) = 0$, the two functions $\beta \circ (\frac{\beta+\zeta}{2})^{-1}$ and $\zeta \circ (\frac{\beta+\zeta}{2})^{-1}$ have the same sign (positive on $\mathbb{R}^+$, negative on $\mathbb{R}^-$) and their sum is equal to 2Id. The absolute value of each one of them is therefore bounded above by $2|\mathrm{Id}|$, and the property $\frac{\overline{\beta}+\overline{\zeta}}{2} \in L^2(X)$ shows that

$$\beta(w) = \left[\beta \circ \left(\frac{\beta+\zeta}{2}\right)^{-1}\right]\left(\frac{\overline{\beta}+\overline{\zeta}}{2}\right)$$

and

$$\zeta(w) = \left[\zeta \circ \left(\frac{\beta+\zeta}{2}\right)^{-1}\right]\left(\frac{\overline{\beta}+\overline{\zeta}}{2}\right)$$

both belong to $L^2(X)$. By monotony of β and ζ,

$$\int_X [\beta(w_m) - \beta(w)]\,[\zeta(w_m) - \zeta(w)]\,\mathrm{d}\mu \geq 0.$$

Develop this relation and use (D.22) and the weak convergences of $\beta(w_m)$ and $\zeta(w_m)$ to take the inferior limit as $m \to \infty$. This gives

$$\int_X \left[\overline{\beta} - \beta(w)\right]\left[\overline{\zeta} - \zeta(w)\right]\mathrm{d}\mu \geq 0. \tag{D.24}$$

With w defined as in the lemma,

$$\frac{\overline{\beta}+\overline{\zeta}}{2} = \frac{\beta(w)+\zeta(w)}{2}. \tag{D.25}$$

Hence, $\beta(w) = \frac{\overline{\beta}+\overline{\zeta}}{2} + \left(\frac{\beta-\zeta}{2}\right)(w)$ and $\zeta(w) = \frac{\overline{\beta}+\overline{\zeta}}{2} - \left(\frac{\beta-\zeta}{2}\right)(w)$. Used in (D.24), this leads to

$$-\int_X \left(\frac{\overline{\beta}-\overline{\zeta}}{2} - \left(\frac{\beta-\zeta}{2}\right)(w)\right)^2 \mathrm{d}\mu \geq 0.$$

Therefore, $\frac{\overline{\beta}-\overline{\zeta}}{2} = \frac{\beta(w)-\zeta(w)}{2}$ a.e. in X and (D.23) follows from this latter relation and (D.25). ∎

The proof of this lemma is classical, and only given for the convenience of the reader.

Lemma D.11 (Weak Fatou for convex functions). *Let I be an interval of $\mathbb{R}$ and $H : I \to [0, +\infty]$ be a convex lower semi-continuous function. Denote by $L^2(\Omega; I)$ the convex set of functions in $L^2(\Omega)$ with values in I. Let $v \in$*

$L^2(\Omega; I)$ and $(v_m)_{m\in\mathbb{N}}$ be a sequence of functions in $L^2(\Omega; I)$ which converges weakly to v in $L^2(\Omega)$. Then,

$$\int_\Omega H(v(\boldsymbol{x}))\mathrm{d}\boldsymbol{x} \leq \liminf_{m\to\infty} \int_\Omega H(v_m(\boldsymbol{x}))\mathrm{d}\boldsymbol{x}.$$

Proof.
Let $\Phi : L^2(\Omega; I) \to [0, \infty]$ be defined by $\Phi(w) = \int_\Omega H(w(\boldsymbol{x}))\mathrm{d}\boldsymbol{x}$. If $(w_k)_{k\in\mathbb{N}}$ converges strongly to w in $L^2(\Omega; I)$ then, up to a subsequence, $w_k \to w$ a.e. on Ω. H being lower semi-continuous, $H(w) \leq \liminf_{k\to\infty} H(w_k)$ a.e. on Ω. Since $H \geq 0$, Fatou's lemma then show that $\Phi(w) \leq \liminf_{k\to\infty} \Phi(w_k)$.
Hence, Φ is lower semi-continuous for the strong topology of $L^2(\Omega; I)$. Since Φ (as H) is convex, we deduce that this lower semi-continuity property is also valid for the weak topology of $L^2(\Omega; I)$, see [82]. The result of the lemma is just the translation of this weak lower semi-continuity of Φ. ∎

E

Some numerical examples

The numerical examples presented here illustrate theoretical convergence results proved in other chapters. Section E.1 is focused on gradient schemes for a 3D linear elliptic equation of the form (2.100); two different GDs are considered: the HMM GD in its SUSHI version (Chapter 13), and the VAG scheme (Section 8.5). The results show that these schemes yield a very good approximation of a quite singular solution on complex meshes. In Section E.2, the ADGGD scheme (see Section 11.3) is applied to the p-Laplace problem (Section 2.1.5). Section E.3 contains numerical results based on the VAG scheme for a degenerate parabolic problem as in Chapter 6.

E.1 A 3D elliptic problem

This numerical test is part of the 3D benchmark [104] and features an elliptic linear problem with non-homogeneous Dirichlet boundary conditions arising for instance in the exploitation of fluids in porous media through the use of a slanted well. The goal is to approximate the solution $\overline{u} \in H^1(\Omega)$ solution of the problem

$$-\operatorname{div}\Lambda\nabla\overline{u} = 0, \tag{E.1}$$

in the domain $\Omega = P\backslash W$, where P is the parallelepiped $(-15,15)\times(-15,15)\times(-7.5,7.5)$ and W is a slanted circular cylinder with radius $r_w = 0.1$. The axis of this well is a straight line located in the $x0z$ plane, passing through the origin at an angle $\theta = \frac{70\pi}{180}$ with the x axis, as shown in Figure E.1. The permeability tensor Λ is constant and anisotropic in the third coordinate direction:

$$\Lambda = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & \tau \end{pmatrix},$$

with $\tau = 0.2$. The solution $\overline{u}$, inspired from [7], is equal to 0 on the well boundary $\partial W \cap \partial\Omega$ and is strictly positive inside Ω. It is defined by $\bar{u}(x,y,z)$

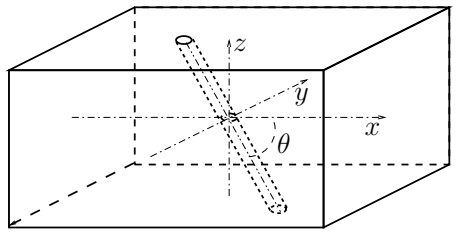


Fig. E.1. The circular slanted well.

$= v(Y(x,y,z),\, Z(x,y,z))$, where the linear functions Y and Z are defined by

$$Y(x,y,z) = y \text{ and } Z(x,y,z) = (\sin\beta)x + \frac{\cos\beta}{\sqrt{\tau}}z,$$

with $\beta = \arctan(\frac{\tan\theta}{\sqrt{\tau}})$. We then define $\alpha = \sqrt{(\frac{\sin\beta}{\sin\theta})^2 - 1}$, and we let $a = \alpha r_w$ and $\mu_0 = \log\left(\frac{1}{\alpha} + \sqrt{\frac{1}{\alpha^2} + 1}\right)$. The function v is given by

$$v(Y,Z) = \log(\sqrt{S} + \sqrt{S+1}) - \mu_0,$$

with $S > 0$ such that $a^2S^2 + (a^2 - Y^2 - Z^2)S - Y^2 = 0$.

The simulation uses 3D meshes (`mesh1` – `mesh7`) created for the 3D benchmark of [104]. These meshes are refined around the well as can be seen in Figure E.2. The first step of the meshing process is to create a radial mesh that is exponentially refined down to the well boundary. This radial local refinement implies a matching mesh between the radial grid and the reservoir Corner Point Geometry grid using hexahedral cells.

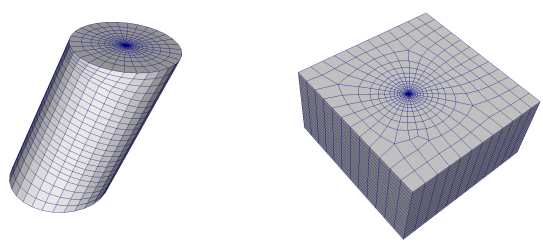


Fig. E.2. Radial mesh without (left) and with transition zone (right).

We present in Table E.1 the results obtained on these grids using on one hand the SUSHI scheme presented in Chapter 13 (see Remark 13.2 for the particular choice of the discrete gradient), and on the other hand the VAG scheme presented in Chapter 8, Section 8.5. The orders of convergence are

computed with respect to the number of unknowns at the power 1/3, which cannot put in evidence the effects of local refinement. The main observation is that these two gradient discretisations are adapted to complex polytopal meshes issued from a quite realistic situation (see Figure E.3 for a visualisation of the approximate solution).

	SUSHI		VAG	
	u	∇u	u	∇u
`mesh1`	3.79E-03	9.69E-02	6.22E-03	5.73E-02
order	0.69	2.03	3.24	3.03
`mesh2`	3.07E-03	5.21E-02	2.60E-03	2.53E-02
order	2.42	2.29	3.46	2.88E+00
`mesh3`	1.60E-03	2.81E-02	1.10E-03	1.24E-02
order	1.38	1.08	1.72	1.69
`mesh4`	1.10E-03	2.10E-02	7.13E-04	8.10E-03
order	1.45	1.19	2.64	1.05
`mesh5`	7.77E-04	1.57E-02	3.85E-04	6.35E-03
order	2.39	1.89	3.60	1.46
`mesh6`	4.78E-04	1.07E-02	1.90E-04	4.77E-03
order	0.26	0.37	-0.16	-0.06
`mesh7`	4.56E-04	9.98E-03	1.95E-04	4.82E-03

Table E.1. L^2 error for the solution and its gradient in the case of the 3D slanted well, using schemes SUSHI and VAG. "Order" represents the rate of convergence from the line above to the line below.

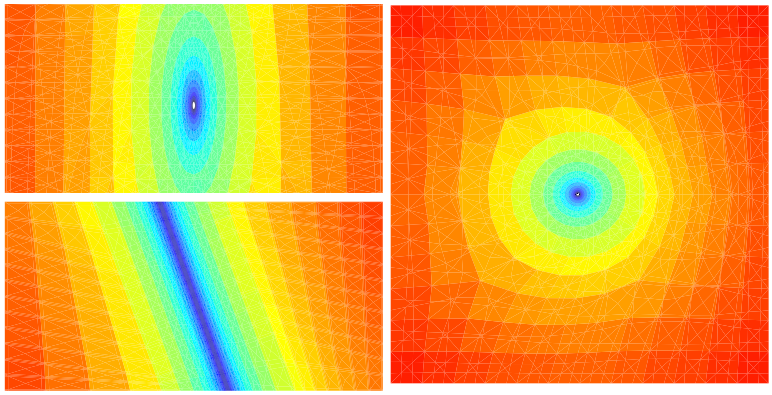


Fig. E.3. Approximate solution using SUSHI on `mesh3`. Top left: slice in the plane $x = 0.2$; bottom left: slice in the plane $y = 0.2$; right: slice in the plane $z = 0.4$.

E.2 ADGGD for the $p-$Laplace problem

We consider the p-Laplace problem (2.61a) with Dirichlet boundary conditions (2.61b). The aim of this section is to assess the accuracy of the error estimate provided by Theorem 2.38.

E.2.1 The one-dimensional case

We consider the case where $d = 1$, $\Omega = (0, 1)$ and $f(x) = 1$. The analytical solution is then given by

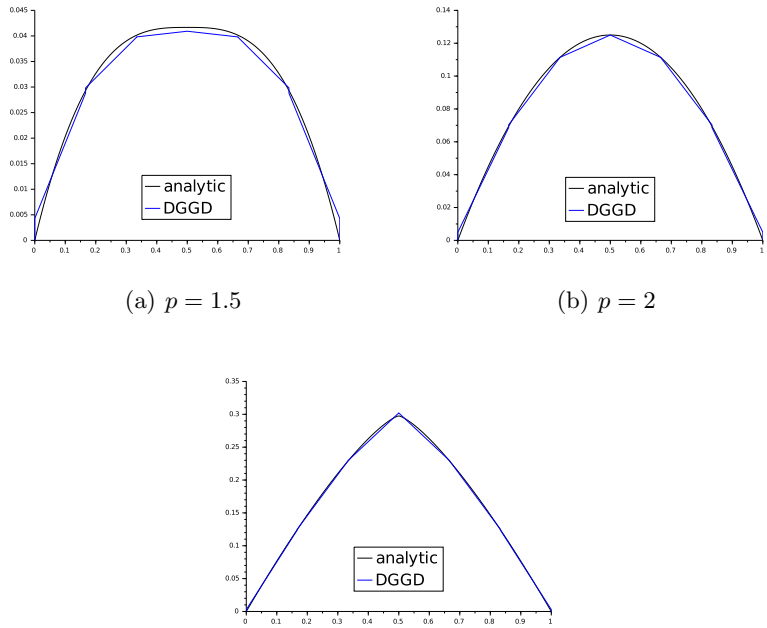
$$\overline{u}(x) = \frac{p-1}{p}\left[\left(\frac{1}{2}\right)^{p/(p-1)} - \left|x - \frac{1}{2}\right|^{p/(p-1)}\right]. \tag{E.2}$$

We consider a mesh with constant space step $h = 1/N$ (with $N \in \mathbb{N}^\star$) and we use Scheme (2.64) together with the Discontinuous Galerkin gradient discretisation given by Definition 11.1, with $k = 1$ and $\beta = 1/2$ (note that in the one-dimensional case, the two definitions (11.4) and (11.40) for the discrete gradient are identical, so the DGGD is identical to the ADGGD).
We see in Figure E.4 that the approximate solution matches quite well the analytical solution for $N = 6$, considering the three cases $p = 1.5$, $p = 2$ and $p = 4$.

	$p = 1.5$		$p = 2$		$p = 4$	
	u	∇u	u	∇u	u	∇u
$N = 10$	5.51E-04	6.34E-03	9.65E-04	9.40E-03	1.48E-03	8.11E-03
order	1.85	1.55	1.96	1.50	1.62	1.33
$N = 20$	1.53E-04	2.17E-03	2.48E-04	3.32E-03	4.80E-04	3.23E-03
order	1.92	1.61	1.98	1.50	1.60	1.29
$N = 40$	4.02E-05	7.11E-04	6.29E-05	1.18E-03	1.58E-04	1.33E-03
order	1.96	1.64	1.99	1.50	1.59	1.59
$N = 80$	1.03E-05	2.28E-04	1.58E-05	4.15E-04	5.24E-05	5.51E-04
order	1.98	1.65	2.00	1.50	1.59	1.26
$N = 160$	2.62E-06	7.26E-05	3.97E-06	3.97E-06	1.74E-05	2.30E-04

Table E.2. Errors and rates of convergences, on the functions and the gradient, for the ADGGD GS applied to the p-Laplace equation in dimension 1. "Order" represents the rate of convergence from the line above to the line below.

Combining Remark 2.39 and Lemmas 11.14 and 11.15 (for $\ell = k = 1$) shows that, if the solution is smooth enough, the expected rates of convergence in L^p norms on both the function and gradient are $\mathcal{O}(h^{p-1})$ if $p \leq 2$ and $\mathcal{O}(h^{1/(p-1)})$ if $p \geq 2$. Hence, for $p = 1.5$ (resp. 2, resp. 4), the expected order would be $\mathcal{O}(h^{0.5})$ (resp. $\mathcal{O}(h)$, resp. $\mathcal{O}(h^{1/3})$). As seen in Table E.2, these



(a) $p = 1.5$ (b) $p = 2$

(c) $p = 4$

Fig. E.4. Exact and DGGD approximate solutions for the p-Laplace equation ($k = 1$, $\beta = 0.5$, $N = 6$).

orders are pessimistic as, even for the non-smooth function given by (E.2), the theoretical rates are beat by at least half an order. This was expected for the L^p norm of the function, as estimates that are common to the function and the gradient, as in Theorem 2.38, are known to be sub-optimal for the function (but, at least for linear problems, better estimates can be established in the GDM framework [80]). This was less obvious for the gradient.

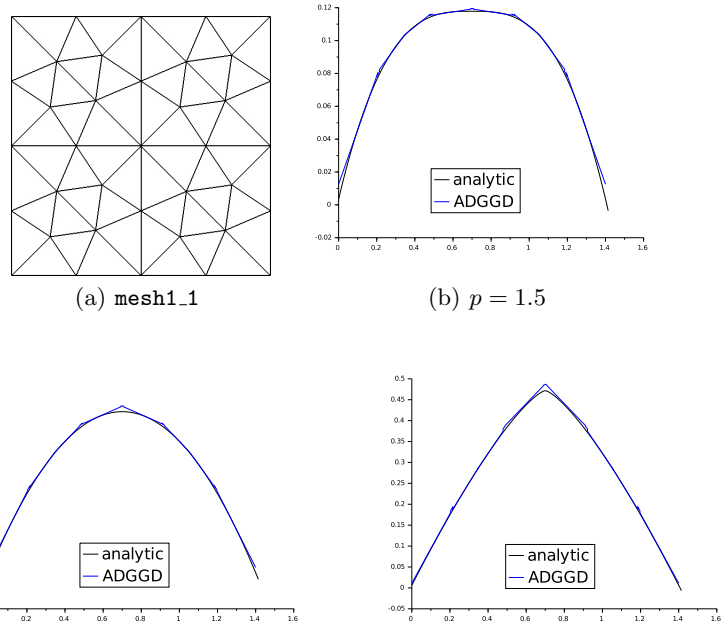
E.2.2 The two-dimensional case

We take here $d = 2$, $\Omega = (0,1) \times (0,1)$ and $f(\boldsymbol{x}) = 2$ for all $\boldsymbol{x} \in \Omega$. Set $\boldsymbol{x}_\Omega = (1/2, 1/2)$ and fix non-homogeneous Dirichlet boundary conditions in agreement with the analytical solution

$$\overline{u}(\boldsymbol{x}) = \frac{p-1}{p}\left[\left(\frac{1}{\sqrt{2}}\right)^{p/(p-1)} - |\boldsymbol{x} - \boldsymbol{x}_\Omega|^{p/(p-1)}\right]. \tag{E.3}$$

We apply the GS (2.64) together with the Average Discontinuous Galerkin Gradient Discretisation given by Definition 11.1 and definition (11.40) for the

discrete gradient, letting $k = 1$ and $\beta = 4/5$. Note that the discrete gradient is piecewise constant, which leads to simple computations, in particular for the $p-$Laplace problem. The triangular meshes from the family `mesh1` of [116] are used for the numerical tests.



(a) `mesh1_1` (b) $p = 1.5$

(c) $p = 2$ (d) $p = 4$

Fig. E.5. Mesh `mesh1_1` and exact and ADGGD approximate profiles along the line $x_2 = x_1 + 0.01$ for the p-Laplace equation ($k = 1$, $\beta = 0.8$, using `mesh1_1`).

Figure E.5 presents the profile of the approximate solution along the line $x_2 = x_1 + 0.01$, for the three cases $p = 1.5$, $p = 2$ and $p = 4$, on the coarsest triangular mesh. We notice a rather good match of approximate solution on this line.

Table E.3 shows that the practical rates of convergence are better than the theoretical ones from Theorem 2.38; however, the rates for the gradient are degraded with respect to the similar test case in dimension $d = 1$.

	$p = 1.5$		$p = 2$		$p = 4$	
	u	∇u	u	∇u	u	∇u
`mesh1_1`	0.944E-03	0.314E-02	0.120E-02	0.423E-02	0.138E-02	0.432E-02
order	1.96	1.48	1.95	1.42	1.32	1.41
`mesh1_2`	0.243E-03	0.113E-02	0.308E-03	0.158E-02	0.555E-03	0.162E-02
order	1.97	1.48	1.98	1.38	1.57	1.16
`mesh1_3`	0.621E-04	0.405E-03	0.783E-04	0.608E-03	0.187E-03	0.727E-03
order	1.98	1.40	1.99	1.31	1.67	0.93
`mesh1_4`	0.157E-04	0.154E-03	0.197E-04	0.245E-03	0.587E-04	0.381E-03
order	1.99	1.29	1.99	1.23	1.73	0.85
`mesh1_5`	0.396E-05	0.630E-04	0.495E-05	0.105E-03	0.177E-04	0.211E-03

Table E.3. Errors and rates of convergences, on the functions and the gradient, for the ADGGD GS applied to the p-Laplace equation in dimension 2. "Order" represents the rate of convergence from the line above to the line below.

E.3 An example of the application of the GDM to a degenerate parabolic problem

We consider the evolution problem (6.1) in 2D, letting $\beta(s) = s$ and $\Lambda = \mathrm{I}_d$, which means that we approximate the Stefan problem. The scheme used here is the VAG scheme described in Section 8.5. The domain is $\Omega = (0,1)^2$, and we use the following definition of $\zeta(\overline{u})$,

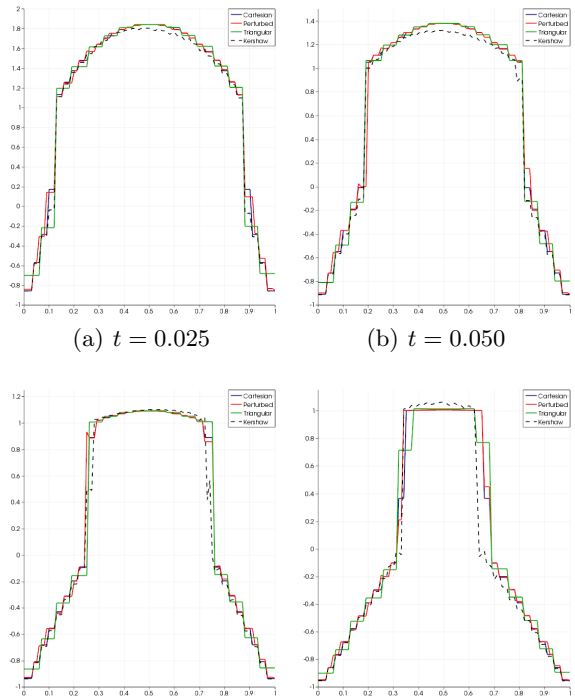
$$\zeta(\overline{u}) = \begin{cases} \overline{u} & \text{if } \overline{u} < 0, \\ \overline{u} - 1 & \text{if } \overline{u} > 1, \\ 0 & \text{otherwise.} \end{cases}$$

Dirichlet boundary conditions are given by $\overline{u} = -1$ on $\partial\Omega$ and the initial condition is $\overline{u}(\boldsymbol{x}, 0) = 2$. Four grids are used for the computations: a Cartesian grid with $32^2 = 1024$ cells, the same grid randomly perturbed, a triangular grids with 896 cells, and a "Kershaw mesh" with 1089 cells as illustrated in Figure E.8 (such meshes are standard in the framework of underground engineering). The final time is 0.1 and the simulation is ran with a constant time step of 0.001.

Figures E.8, E.9, E.10 and E.11 represent the discrete solution $u(\cdot, t)$ on all grids for $t = 0.025, 0.05, 0.075$ and 0.1. For a better comparison we have also plotted the interpolation of u along two lines of the mesh. The first line is horizontal and joins the two points $(0, 0.5)$ and $(1, 0.5)$. The second line is diagonal and joins points $(0, 0)$ and $(1, 1)$. The results for these slices are shown in Figures E.6 and E.7.

The numerical outputs are weakly dependent on the grid, and the interface between the regions $u < 0$ and $u > 1$ are located at the same place for all grids. It is worth noticing that this remains true even for the very irregular Kershaw mesh (which presents high regularity factors $\theta_{\mathfrak{T}}$ – see (7.8), that is

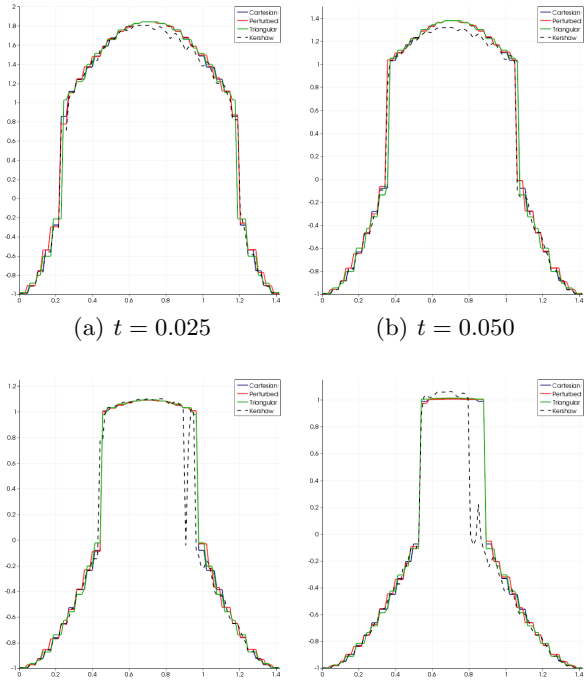
high ratios for some cells between the radii of inscribed balls and the diameter of the cell).



(a) $t = 0.025$ (b) $t = 0.050$

(c) $t = 0.075$ (d) $t = 0.1$

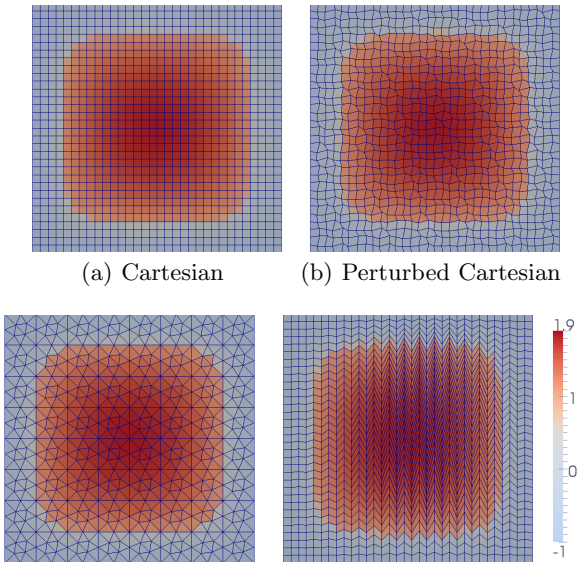
Fig. E.6. Interpolation of u along the line $x_2 = 0.5$ of the mesh for each grids: Cartesian in blue, perturbed Cartesian in red, triangular in green, and Kershaw in black dashed.



(a) $t = 0.025$ (b) $t = 0.050$

(c) $t = 0.075$ (d) $t = 0.1$

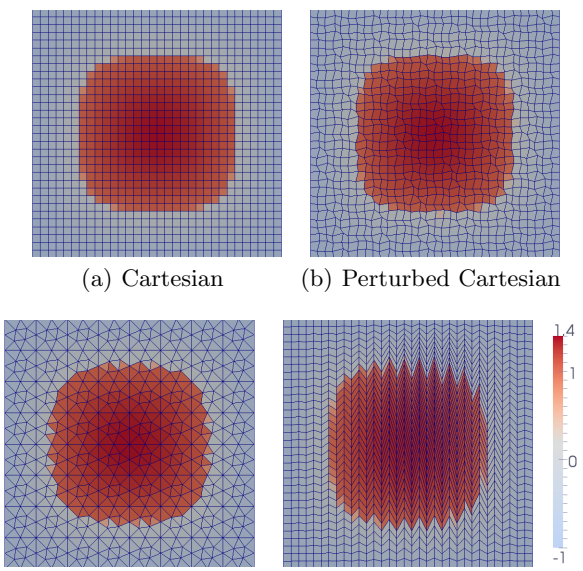
Fig. E.7. Interpolation of u along a diagonal axe of the mesh for each grids: Cartesian in blue, perturbed Cartesian in red, triangular in green, and Kershaw in black dashed.



(a) Cartesian (b) Perturbed Cartesian

(c) Triangular (d) Kershaw

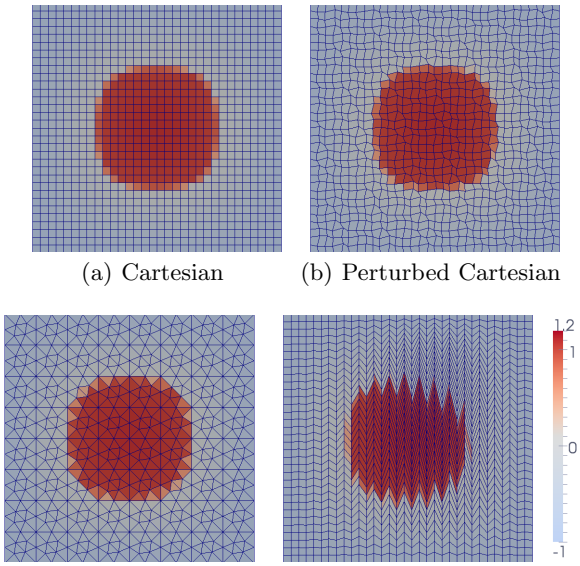
Fig. E.8. Discrete solution u on all grids at $t = 0.025$.



(a) Cartesian (b) Perturbed Cartesian

(c) Triangular (d) Kershaw

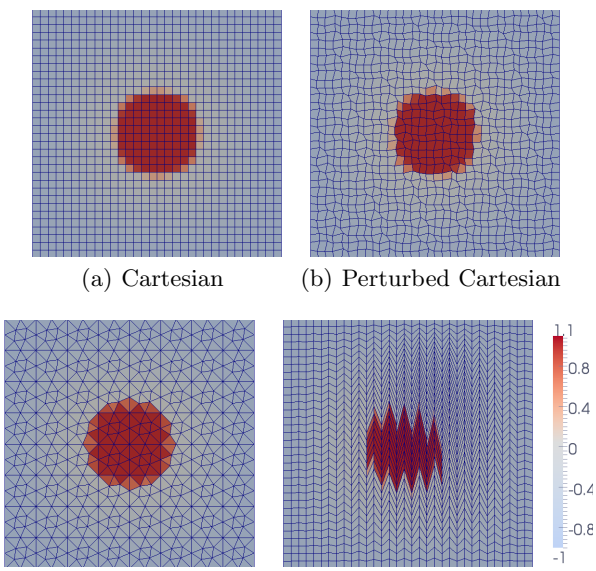
Fig. E.9. Discrete solution u on all grids at $t = 0.050$.



(a) Cartesian (b) Perturbed Cartesian

(c) Triangular (d) Kershaw

Fig. E.10. Discrete solution u on all grids at $t = 0.075$.



(a) Cartesian (b) Perturbed Cartesian

(c) Triangular (d) Kershaw

Fig. E.11. Discrete solution u on all grids at $t = 0.1$.

References

1. I. Aavatsmark. An introduction to multipoint flux approximations for quadrilateral grids. *Comput. Geosci.*, 6(3-4):405–432, 2002. Locally conservative numerical methods for flow in porous media.
2. I. Aavatsmark, T. Barkve, O. Boe, and T. Mannseth. Discretization on non-orthogonal, quadrilateral grids for inhomogeneous, anisotropic media. *J. Comput. Phys.*, 127(1):2–14, 1996.
3. I. Aavatsmark, T. Barkve, O. Boe, and T. Mannseth. Discretization on unstructured grids for inhomogeneous, anisotropic media. II. Discussion and numerical results. *SIAM J. Sci. Comput.*, 19(5):1717–1736 (electronic), 1998.
4. I. Aavatsmark, T. Barkve, O. Boe, and T. Mannseth. Discretization on unstructured grids for inhomogeneous, anisotropic media. part i: Derivation of the methods. *SIAM Journal on Sc. Comp.*, 19:1700–1716, 1998.
5. I. Aavatsmark, G. T. Eigestad, and R. A. Klausen. Numerical convergence of the MPFA O-method for general quadrilateral grids in two and three dimensions. In *Compatible spatial discretizations*, volume 142 of *IMA Vol. Math. Appl.*, pages 1–21. Springer, New York, 2006.
6. I. Aavatsmark, G. T. Eigestad, R. A. Klausen, M. F. Wheeler, and I. Yotov. Convergence of a symmetric MPFA method on quadrilateral grids. *Comput. Geosci.*, 11(4):333–345, 2007.
7. I. Aavatsmark and R. Klausen. Well index in reservoir simulation for slanted and slightly curved wells in 3D grids. *SPE Journal*, 8:41–48, 2003.
8. L. Agélas, D. A. Di Pietro, and J. Droniou. The G method for heterogeneous anisotropic diffusion on general meshes. *M2AN Math. Model. Numer. Anal.*, 44(4):597–625, 2010.
9. L. Agélas and R. Masson. Convergence of the finite volume MPFA O scheme for heterogeneous anisotropic diffusion problems on general meshes. *C. R. Math. Acad. Sci. Paris*, 346(17-18):1007–1012, 2008.
10. Y. Alnashri and J. Droniou. Gradient schemes for an obstacle problem. In J. Fuhrmann, M. Ohlberger, and C. Rohde, editors, *Finite Volumes for Complex Applications VII-Methods and Theoretical Aspects*, volume 77, pages 67–75. Springer International Publishing, 2014.
11. Y. Alnashri and J. Droniou. Gradient schemes for the Signorini and the obstacle problems, and application to hybrid mimetic mixed methods. *Computers and Mathematics with Applications*, 72:2788–2807, 2016.

12. H. Amann. Compact embeddings of vector-valued Sobolev and Besov spaces. *Glas. Mat. Ser. III*, 35(55)(1):161–177, 2000. Dedicated to the memory of Branko Najman.
13. B. Andreianov, M. Bendahmane, and F. Hubert. On 3D DDFV discretization of gradient and divergence operators: discrete functional analysis tools and applications to degenerate parabolic problems. *Comput. Methods Appl. Math.*, 13(4):369–410, 2013.
14. B. Andreianov, M. Bendahmane, F. Hubert, and S. Krell. On 3D DDFV discretization of gradient and divergence operators. I. Meshing, operators and discrete duality. *IMA J. Numer. Anal.*, 32(4):1574–1603, 2012.
15. B. Andreianov, F. Boyer, and F. Hubert. Discrete duality finite volume schemes for Leray-Lions-type elliptic problems on general 2D meshes. *Numer. Methods Partial Differential Equations*, 23(1):145–195, 2007.
16. D. N. Arnold and F. Brezzi. Mixed and conforming finite element methods; implementation, postprocessing and error estimates. *Modélisation mathématique et analyse numérique*, 19(1):7–32, 1985.
17. D. N. Arnold, F. Brezzi, B. Cockburn, and L. D. Marini. Unified analysis of discontinuous Galerkin methods for elliptic problems. *SIAM J. Numer. Anal.*, 39(5):1749–1779, 2001/02.
18. H. Attouch, G. Buttazzo, and G. Michaille. *Variational analysis in Sobolev and BV spaces*, volume 6 of *MPS/SIAM Series on Optimization*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA; Mathematical Programming Society (MPS), Philadelphia, PA, 2006.
19. B. Ayuso de Dios, K. Lipnikov, and G. Manzini. The nonconforming virtual element method. *ESAIM: Math. Model Numer. Anal.*, 50(3):879–904, 2016.
20. J. W. Barrett and W. B. Liu. Quasi-norm error bounds for the finite element approximation of a non-Newtonian flow. *Numer. Math.*, 68(4):437–456, 1994.
21. L. Beirão da Veiga, F. Brezzi, A. Cangiani, G. Manzini, L. D. Marini, and A. Russo. Basic principles of virtual element methods. *Math. Models Methods Appl. Sci. (M3AS)*, 199(23):199–214, 2013.
22. L. Beirão da Veiga, J. Droniou, and G. Manzini. A unified approach for handling convection terms in finite volumes and mimetic discretization methods for elliptic problems. *IMA J. Numer. Anal.*, 31(4):1357–1401, 2011.
23. M. Bertsch, P. De Mottoni, and L. Peletier. The Stefan problem with heating: appearance and disappearance of a mushy region. *Trans. Amer. Math. Soc*, 293:677–691, 1986.
24. M. Bertsch, R. Kersner, and L. A. Peletier. Positivity versus localization in degenerate diffusion equations. *Nonlinear Anal.*, 9(9):987–1008, 1985.
25. M. Bessemoulin-Chatard, C. Chainais-Hillairet, and F. Filbet. On discrete functional inequalities for some finite volume schemes. *IMA J. Numer. Anal.*, 35(3):1125–1149, 2015.
26. J. Bonelle. *Compatible Discrete Operator schemes on polyhedral meshes for elliptic and Stokes equations*. PhD thesis, University of Paris-Est, 2014.
27. J. Bonelle and A. Ern. Analysis of compatible discrete operator schemes for elliptic problems on polyhedral meshes. *ESAIM Math. Model. Numer. Anal.*, 48(2):553–581, 2014.
28. F. Boyer and F. Hubert. Finite volume method for 2D linear and nonlinear elliptic problems with discontinuities. *SIAM J. Numer. Anal.*, 46(6):3032–3070, 2008.

29. F. Boyer and F. Hubert. Finite volume method for 2D linear and nonlinear elliptic problems with discontinuities. *SIAM Journal on Numerical Analysis*, 46(6):3032–3070, 2008.
30. D. Braess. *Finite elements*. Cambridge University Press, Cambridge, third edition, 2007. Theory, fast solvers, and applications in elasticity theory, Translated from the German by Larry L. Schumaker.
31. S. C. Brenner and L. Owens. A weakly over-penalized non-symmetric interior penalty method. *JNAIAM J. Numer. Anal. Ind. Appl. Math.*, 2(1-2):35–48, 2007.
32. S. C. Brenner and L. R. Scott. *The mathematical theory of finite element methods*, volume 15 of *Texts in Applied Mathematics*. Springer, New York, third edition, 2008.
33. H. Brezis. *Functional analysis, Sobolev spaces and partial differential equations*. Universitext. Springer, New York, 2011.
34. F. Brezzi, A. Buffa, and K. Lipnikov. Mimetic finite differences for elliptic problems. *M2AN Math. Model. Numer. Anal.*, 43(2):277–295, 2009.
35. F. Brezzi, K. Lipnikov, and M. Shashkov. Convergence of the mimetic finite difference method for diffusion problems on polyhedral meshes. *SIAM J. Numer. Anal.*, 43(5):1872–1896, 2005.
36. F. Brezzi, K. Lipnikov, and V. Simoncini. A family of mimetic finite difference methods on polygonal and polyhedral meshes. *Math. Models Methods Appl. Sci.*, 15(10):1533–1551, 2005.
37. A. Cangiani, G. Manzini, and A. Russo. Convergence analysis of a mimetic finite difference method for general second-order elliptic problems. *SIAM J. Numer. Anal.*, 47(4):2612–2637, 2009.
38. A. Cangiani, G. Manzini, and O. J. Sutton. Conforming and nonconforming virtual element methods for elliptic problems. *IMA J. Numer. Anal.*, 37(3):1317–1354, 2017.
39. P. Cantin, J. Bonelle, E. Burman, and A. Ern. A vertex-based scheme on polyhedral meshes for advection-reaction equations with sub-mesh stabilization. *Computers & Mathematics with Applications*, 72(9):2057–2071, 2016.
40. P. Cantin and A. Ern. Vertex-based compatible discrete operator schemes on polyhedral meshes for advection-diffusion equations. *Comput. Methods Appl. Math.*, 16(2):187–212, 2016.
41. J. Carrillo. Entropy solutions for nonlinear degenerate problems. *Arch. Ration. Mech. Anal.*, 147(4):269–361, 1999.
42. F. Catté, P.-L. Lions, J.-M. Morel, and T. Coll. Image selective smoothing and edge detection by nonlinear diffusion. *SIAM J. Numer. Anal.*, 29(1):182–193, 1992.
43. C. Chainais-Hillairet and J. Droniou. Convergence analysis of a mixed finite volume scheme for an elliptic-parabolic system modeling miscible fluid flows in porous media. *SIAM J. Numer. Anal.*, 45(5):2228–2258, 2007.
44. C. Chainais-Hillairet, S. Krell, and A. Mouton. Convergence analysis of a DDFV scheme for a system describing miscible fluid flows in porous media. *Numer. Methods Partial Differential Equations*, 31(3):723–760, 2015.
45. X. Chen, A. Jüngel, and J.-G. Liu. A note on Aubin-Lions-Dubinskiĭ lemmas. *Acta Appl. Math.*, 133:33–43, 2014.
46. Y. Chen, B. Vemuri, and L. Wang. Image denoising and segmentation via nonlinear diffusion. *Computers and Mathematics with Applications*, 39:131–149, 2000.

47. H. M. Cheng, J. Droniou, and K.-N. Le. Convergence analysis of a family of ELLAM schemes for a fully coupled model of miscible displacement in porous media. Submitted. `https://arxiv.org/abs/1710.01897`, 2017.
48. P. Ciarlet. The finite element method. In P. G. Ciarlet and J.-L. Lions, editors, *Part I*, Handbook of Numerical Analysis, III. North-Holland, Amsterdam, 1991.
49. P. G. Ciarlet. The finite element method for elliptic problems. In *Studies in Mathematics and its Applications, Vol. 4*, pages xix+530. North-Holland Publishing Co., Amsterdam-New York-Oxford, 1978.
50. P. Clément. Approximation by finite element functions using local regularization. *RAIRO Anal. Numér.*, 9:77–84, 1975.
51. Y. Coudière, T. Gallouët, and R. Herbin. Discrete Sobolev Inequalities and L^p Error Estimates for Approximate Finite Volume Solutions of Convection Diffusion Equations. *Mathematical Modelling and Numerical Analysis*, 35:767–778, 1998.
52. Y. Coudière and F. Hubert. A 3d discrete duality finite volume method for nonlinear elliptic equations. *SIAM Journal on Scientific Computing*, 33(4):1739–1764, 2011.
53. Y. Coudière, F. Hubert, and G. Manzini. A CeVeFE DDFV scheme for discontinuous anisotropic permeability tensors. In *Finite volumes for complex applications VI*, volume 4 of *Springer Proc. Math.*, pages 283–291. Springer, Heidelberg, 2011.
54. Y. Coudière, J.-P. Vila, and P. Villedieu. Convergence rate of a finite volume scheme for a two-dimensional convection-diffusion problem. *M2AN Math. Model. Numer. Anal.*, 33(3):493–516, 1999.
55. M. Crouzeix and P.-A. Raviart. Conforming and nonconforming finite element methods for solving the stationary Stokes equations. I. *Rev. Française Automat. Informat. Recherche Opérationnelle Sér. Rouge*, 7(R-3):33–75, 1973.
56. K. Deimling. *Nonlinear functional analysis.* Springer-Verlag, Berlin, 1985.
57. D. A. Di Pietro and J. Droniou. A Hybrid High-Order method for Leray–Lions elliptic equations on general meshes. *Math. Comp.*, 86(307):2159–2191, 2017.
58. D. A. Di Pietro and J. Droniou. $W^{s,p}$-approximation properties of elliptic projectors on polynomial spaces, with application to the error analysis of a Hybrid High-Order discretisation of Leray–Lions problems. *Math. Models Methods Appl. Sci.*, 27(5):879–908, 2017.
59. D. A. Di Pietro, J. Droniou, and G. Manzini. Discontinuous skeletal gradient discretisation methods on polytopal meshes. *J. Comput. Phys.*, 2017. To appear. DOI: 10.1016/j.jcp.2017.11.018.
60. D. A. Di Pietro and A. Ern. Discrete functional analysis tools for discontinuous Galerkin methods with application to the incompressible Navier-Stokes equations. *Math. Comp.*, 79(271):1303–1330, 2010.
61. D. A. Di Pietro and A. Ern. *Mathematical aspects of discontinuous Galerkin methods*, volume 69 of *Mathématiques & Applications (Berlin) [Mathematics & Applications]*. Springer, Heidelberg, 2012.
62. D. A. Di Pietro, A. Ern, and S. Lemaire. An arbitrary-order and compact-stencil discretization of diffusion on general meshes based on local reconstruction operators. *Comput. Meth. Appl. Math.*, 14(4):461–472, 2014.
63. J. Diestel and J. J. Uhl, Jr. *Vector measures.* American Mathematical Society, Providence, R.I., 1977. With a foreword by B. J. Pettis, Mathematical Surveys, No. 15.

64. K. Domelevo and P. Omnes. A finite volume method for the Laplace equation on almost arbitrary two-dimensional grids. *M2AN Math. Model. Numer. Anal.*, 39(6):1203–1249, 2005.
65. O. Drblíková and K. Mikula. Convergence analysis of finite volume scheme for nonlinear tensor anisotropic diffusion in image processing. *SIAM J. Numer. Anal.*, 46(1):37–60, 2007/08.
66. M. Dreher and A. Jüngel. Compact families of piecewise constant functions in $L^p(0,T;B)$. *Nonlinear Anal.*, 75(6):3072–3077, 2012.
67. J. Droniou. Intégration et espaces de sobolev à valeurs vectorielles. Polycopiés de l'Ecole Doctorale de Maths-Info de Marseille. `https://hal.archives-ouvertes.fr/hal-01382368`, 2001.
68. J. Droniou. Finite volume schemes for fully non-linear elliptic equations in divergence form. *ESAIM: Mathematical Modelling and Numerical Analysis*, 40(6):1069–1100, 2006.
69. J. Droniou. Finite volume schemes for diffusion equations: introduction to and review of modern methods. *Math. Models Methods Appl. Sci. (M3AS)*, 24(8):1575–1619, 2014. Special issue on Recent Techniques for PDE Discretizations on Polyhedral Meshes.
70. J. Droniou and R. Eymard. A mixed finite volume scheme for anisotropic diffusion problems on any grid. *Numer. Math.*, 105(1):35–71, 2006.
71. J. Droniou and R. Eymard. Study of the mixed finite volume method for Stokes and Navier-Stokes equations. *Numerical methods for partial differential equations*, 25(1):137–171, 2009.
72. J. Droniou and R. Eymard. Uniform-in-time convergence of numerical methods for non-linear degenerate parabolic equations. *Numer. Math.*, 132(4):721–766, 2016.
73. J. Droniou and R. Eymard. The asymmetric gradient discretisation method. In *Finite volumes for complex applications. VIII. (Lille, 2017)*, Springer Proc. Math. Stat. Springer, Cham, 2017.
74. J. Droniou, R. Eymard, and P. Feron. Gradient Schemes for Stokes problem. *IMA J. Numer. Anal.*, 36(4):1636–1669, 2016.
75. J. Droniou, R. Eymard, T. Gallouët, and R. Herbin. A unified approach to mimetic finite difference, hybrid finite volume and mixed finite volume methods. *Math. Models Methods Appl. Sci.*, 20(2):265–295, 2010.
76. J. Droniou, R. Eymard, T. Gallouët, and R. Herbin. Gradient schemes: a generic framework for the discretisation of linear, nonlinear and nonlocal elliptic and parabolic equations. *Math. Models Methods Appl. Sci. (M3AS)*, 23(13):2395–2432, 2013.
77. J. Droniou, R. Eymard, and R. Herbin. Gradient schemes: generic tools for the numerical analysis of diffusion equations. *M2AN Math. Model. Numer. Anal.*, 50(3):749–781, 2016. Special issue – Polyhedral discretization for PDE.
78. J. Droniou, R. Eymard, and K. S. Talbot. Convergence in $C([0;T];L^2(\Omega))$ of weak solutions to perturbed doubly degenerate parabolic equations. *J. Differential Equations*, 260(11):7821–7860, 2016.
79. J. Droniou and B. P. Lamichhane. Gradient schemes for linear and non-linear elasticity equations. *Numer. Math.*, 129(2):251–277, 2015.
80. J. Droniou and N. Nataraj. Improved L^2 estimate for gradient schemes and super-convergence of the tpfa finite volume scheme. *IMA J. Numer. Anal.*, page 40p, 2017. To appear, DOI: 10.1093/imanum/drx028.

81. R. Durán. Mixed finite elements. In D. Boffi and L. Gastaldi, editors, *Mixed finite elements, compatibility conditions, and applications: lectures given at the CIME Summer School held in Cetraro, Italy, June 26-July 1, 2006*, volume 1939 of *Lecture Notes in Mathematics*, pages 1–44. Springer, 2008.
82. I. Ekeland and R. Témam. *Convex analysis and variational problems*, volume 28 of *Classics in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, english edition, 1999. Translated from the French.
83. Y. Epshteyn and B. Rivière. Estimation of penalty parameters for symmetric interior penalty Galerkin methods. *J. Comput. Appl. Math.*, 206(2):843–872, 2007.
84. A. Ern and J.-L. Guermond. *Theory and practice of finite elements*, volume 159. Springer, 2004.
85. A. Ern and J.-L. Guermond. Mollification in strongly lipschitz domains with application to continuous and discrete de rham complexes. *Comput. Meth. in Appl. Math.*, 16(1):51–75, 2016.
86. A. Ern and J.-L. Guermond. Finite element quasi-interpolation and best approximation. *ESAIM: M2AN*, 51(4):1367–1385, 2017.
87. A. Ern and J.-L. Guermond. *Finite elements methods, updated version*. Springer, to be published in 2017.
88. L. C. Evans. *Partial differential equations*, volume 19 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, second edition, 2010.
89. L. C. Evans and J. Spruck. Motion of level sets by mean curvature I. *J. Differential Geometry*, 33:635–681, 1991.
90. R. Eymard, P. Feron, and C. Guichard. Family of convergent numerical schemes for the incompressible Navier–Stokes equations. *Mathematics and Computers in Simulation*, 144(Supplement C):196–218, 2018.
91. R. Eymard, T. Gallouët, and R. Herbin. Finite volume methods. In P. G. Ciarlet and J.-L. Lions, editors, *Techniques of Scientific Computing, Part III*, Handbook of Numerical Analysis, VII, pages 713–1020. North-Holland, Amsterdam, 2000.
92. R. Eymard, T. Gallouët, and R. Herbin. A finite volume scheme for anisotropic diffusion problems. *C. R. Math. Acad. Sci. Paris*, 339(4):299–302, 2004.
93. R. Eymard, T. Gallouët, and R. Herbin. A cell-centered finite-volume approximation for anisotropic diffusion operators on unstructured meshes in any space dimension. *IMA J. Numer. Anal.*, 26(2):326–353, 2006.
94. R. Eymard, T. Gallouët, and R. Herbin. A new finite volume scheme for anisotropic diffusion problems on general grids: convergence analysis. *C. R., Math., Acad. Sci. Paris*, 344(6):403–406, 2007.
95. R. Eymard, T. Gallouët, and R. Herbin. Cell centred discretization of fully non linear elliptic problems on general multidimensional polyhedral grids. *J. Numer. Math.*, 17(3):173–193, 2009.
96. R. Eymard, T. Gallouët, and R. Herbin. Discretization of heterogeneous and anisotropic diffusion problems on general nonconforming meshes SUSHI: a scheme using stabilization and hybrid interfaces. *IMA J. Numer. Anal.*, 30(4):1009–1043, 2010.
97. R. Eymard and C. Guichard. Discontinuous Galerkin gradient discretisations for the approximation of second-order differential operators in divergence form. working paper or preprint, Nov. 2017.

98. R. Eymard, C. Guichard, and R. Herbin. Benchmark 3D: the VAG scheme. In *Finite volumes for complex applications. VI. Problems & perspectives. Volume 1, 2*, volume 4 of *Springer Proc. Math.*, pages 1013–1022. Springer, Heidelberg, 2011.
99. R. Eymard, C. Guichard, and R. Herbin. Small-stencil 3d schemes for diffusive flows in porous media. *M2AN*, 46:265–290, 2012.
100. R. Eymard, C. Guichard, R. Herbin, and R. Masson. Vertex centred discretization of two-phase Darcy flows on general meshes. In *Congrès National de Mathématiques Appliquées et Industrielles*, volume 35 of *ESAIM Proc.*, pages 59–78. EDP Sci., Les Ulis, 2011.
101. R. Eymard, C. Guichard, R. Herbin, and R. Masson. Vertex-centred discretization of multiphase compositional darcy flows on general meshes. *Computational Geosciences*, pages 1–19, 2012.
102. R. Eymard, C. Guichard, R. Herbin, and R. Masson. Gradient schemes for two-phase flow in heterogeneous porous media and Richards equation. *ZAMM Z. Angew. Math. Mech.*, 94(7-8):560–585, 2014.
103. R. Eymard, A. Handlovičová, R. Herbin, K. Mikula, and O. Stašová. Applications of approximate gradient schemes for nonlinear parabolic equations. *Appl. Math.*, 60(2):135–156, 2015.
104. R. Eymard, G. Henry, R. Herbin, F. Hubert, R. Klofkorn, and G. Manzini. 3d benchmark on discretization schemes for anisotropic diffusion problems on general grids. In *Proceedings of Finite Volumes for Complex Applications VI*, pages 895–930, Praha, 2011. Springer, Springer.
105. I. Faille. A control volume method to solve an elliptic equation on a two-dimensional irregular mesh. *Comput. Methods Appl. Mech. Engrg.*, 100(2):275–290, 1992.
106. F. Filbet. A finite volume scheme for the Patlak-Keller-Segel chemotaxis model. *Numer. Math.*, 104(4):457–488, 2006.
107. T. Gallouët and R. Herbin. Partial differential equations. Master course, `https://hal.archives-ouvertes.fr/cel-01196782`, Sept. 2015.
108. T. Gallouët, R. Herbin, J.-C. Latché, and K. Mallem. Convergence of the MAC scheme for the incompressible Navier–Stokes equations. *Foundations of Computational Mathematics*, pages 1–41, 2016.
109. T. Gallouët and J. Latché. Compactness of discrete approximate solutions to parabolic pdes – application to a turbulence model. *Commun. Pure Appl. Anal*, 12(6):2371–2391, 2012.
110. A. Glitzky and J. A. Griepentrog. Discrete Sobolev-Poincaré inequalities for Voronoi finite volume approximations. *SIAM J. Numer. Anal.*, 48(1):372–391, 2010.
111. A. Glitzky and J. A. Griepentrog. On discrete Sobolev-Poincaré inequalities for Voronoi finite volume approximations. In *Finite volumes for complex applications VI. Problems & perspectives. Volume 1, 2*, volume 4 of *Springer Proc. Math.*, pages 533–541. Springer, Heidelberg, 2011.
112. M. Guedda, D. Hilhorst, and M. A. Peletier. Disappearing interfaces in nonlinear diffusion. *Adv. Math. Sci. Appl.*, 7(2):695–710, 1997.
113. J. Hadamard. *Lectures on Cauchy's Problem in Linear Partial Differential Equations.* Dover Phoenix, New York, NY, USA, 1923.
114. P. Hansbo and M. G. Larson. Discontinuous Galerkin and the Crouzeix-Raviart element: application to elasticity. *M2AN Math. Model. Numer. Anal.*, 37(1):63–72, 2003.

115. R. Herbin. An error estimate for a finite volume scheme for a diffusion-convection problem on a triangular mesh. *Numer. Methods Partial Differential Equations*, 11(2):165–173, 1995.
116. R. Herbin and F. Hubert. Benchmark on discretization schemes for anisotropic diffusion problems on general grids. In *Finite volumes for complex applications V*, pages 659–692. ISTE, London, 2008.
117. F. Hermeline. Une méthode de volumes finis pour les équations elliptiques du second ordre. *C. R. Acad. Sci. Paris Sér. I Math.*, 326(12):1433–1436, 1998.
118. F. Hermeline. A finite volume method for the approximation of diffusion operators on distorted meshes. *J. Comput. Phys.*, 160(2):481–499, 2000.
119. F. Hermeline. Approximation of diffusion operators with discontinuous tensor coefficients on distorted meshes. *Comput. Methods Appl. Mech. Engrg.*, 192(16-18):1939–1959, 2003.
120. F. Hermeline. Approximation of 2-d and 3-d diffusion operators with variable full tensor coefficients on arbitrary meshes. *Computer methods in applied mechanics and engineering*, 196(21):2497–2526, 2007.
121. F. Hermeline. Numerical experiments with the DDFV method. In *Finite volumes for complex applications V*, pages 851–863. ISTE, London, 2008.
122. F. Hermeline. A finite volume method for approximating 3d diffusion operators on general meshes. *Journal of computational Physics*, 228(16):5763–5786, 2009.
123. J. Jaffré. Décentrage et éléments finis mixtes pour les équations de diffusion-convection. *Calcolo*, 21:171–197, 1984.
124. T. Kato. *Introduction to the theory of operators in Banach spaces*, pages 126–188. Springer Berlin Heidelberg, Berlin, Heidelberg, 1995.
125. S. Lemaire. *Discrétisations non-conformes d'un modèle poromécanique sur maillages généraux*. PhD thesis, University of Paris-Est Marne-la-Vallée, 2012. [oai:tel.archives-ouvertes.fr:tel-00957292] – http://tel.archives-ouvertes.fr/tel-00957292.
126. J. Leray and J. Lions. Quelques résultats de Višik sur les problèmes elliptiques non linéaires par les méthodes de Minty-Browder. *Bull. Soc. Math. France*, 93:97–107, 1965.
127. K. Lipnikov and G. Manzini. A high-order mimetic method on unstructured polyhedral meshes for the diffusion equation. *J. Comput. Phys.*, 272:360–385, 2014.
128. K. Lipnikov, G. Manzini, and M. Shashkov. Mimetic finite difference method. *J. Comput. Phys.*, 257-Part B:1163–1227, 2014.
129. K. Lipnikov, M. Shashkov, and I. Yotov. Local flux mimetic finite difference methods. *Numer. Math.*, 112(1):115–152, 2009.
130. R. Liska, M. Shashkov, and V. Ganzha. Analysis and optimization of inner products for mimetic finite difference methods on triangular grid. *Math. Comput. Simu.*, 67:55–66, 2004.
131. G. Minty. On a monotonicity method for the solution of non-linear equations in Banach spaces. *Proceedings of the National Academy of Sciences of the United States of America*, 50(6):1038, 1963.
132. G. Strang. Variational crimes in the finite element method. In *The mathematical foundations of the finite element method with applications to partial differential equations (Proc. Sympos., Univ. Maryland, Baltimore, Md., 1972)*, pages 689–710. Academic Press, New York, 1972.
133. G. Strang and G. Fix. *An analysis of the finite element method*. Wellesley-Cambridge Press, Wellesley, MA, second edition, 2008.

134. R. Temam. *Navier–Stokes equations theory and numerical analysis*. North-Holland, 1977.
135. R. Temam. *Navier–Stokes equations*, volume 2 of *Studies in Mathematics and its Applications*. North-Holland Publishing Co., Amsterdam, third edition, 1984. Theory and numerical analysis, With an appendix by F. Thomasset.
136. V. Thomée. *Galerkin Finite Element Methods for Parabolic Problems (Springer Series in Computational Mathematics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.
137. A. Veeser and R. Verfürth. Poincaré constants for finite element stars. *IMA Journal of Numerical Analysis*, 32(1):30–47, 2012.
138. J. Weickert. Coherence-enhancing diffusion filtering. *International Journal of Computer Vision*, 31(2/3):111–127, 1999.

Abbreviations

BC	Boundary Condition
CeVeFE	Cell–Vertex–Face/Edge
CDO	Compatible Discrete Operator
DDFV	Discrete Duality Finite Volume
DGGD	Discontinuous Galerkin gradient discretization
GD	Gradient Discretisation
GDM	Gradient Discretisation Method
GS	Gradient Scheme
HFV	Hybrid Finite Volume
hMFD	hybrid Mimetic Finite Difference
HMM	Hybrid Mimetic Mixed
LLE	Local Linearly Exact
MFD	Mimetic Finite Difference
MFE	Mixed Finite Element
MPFA	Multi-Point Flux Approximation
MPFA-O	Multi-Point Flux Approximation O-scheme
nMFD	nodal Mimetic Finite Difference
PDE	Partial Differential Equation
SIPG	Symmetric Interior Penalty Galerkin
SUSHI	Scheme Using Stabilization and Hybrid Interfaces
TPFA	Two-Point Flux Approximation
TPFA-CG	Two-Point Flux Approximation for Cartesian Grids
VAG	Vertex Approximate Gradient

Notations

Functional spaces
- $C_c^\infty(\Omega)$, 458
- $H_{\mathrm{div}}(\Omega)$, 20, 458
- $H^k(\Omega)$, 458
- $H_0^k(\Omega)$, 458
- $H_\star^1(\Omega)$, 77
- $L^p(\Omega)$, 457
- $\mathbb{P}_\ell$, 225
- $W_{\mathrm{div}}^q(\Omega)$, 20, 458
- $W_{\mathrm{div},\partial}^{p'}(\Omega)$, 71
- $W_{\mathrm{div},\Gamma_n}^{p'}(\Omega)$, 90
- $W^{k,q}(\Omega)$, 458
- $W_0^{k,q}(\Omega)$, 458
- $W_\bullet^{1,p}(\Omega)$, 97

GD operators and indicators
- $C_{\mathcal{D}}$, 18
- $\delta_{\mathcal{D}}$, 98
- $\nabla_{\mathcal{D}}^{(\theta)}$, 98
- $\mathcal{I}_{\mathcal{D}}$, 98
- $\Pi_{\mathcal{D}}^{ML}$, 234
- $\Pi_{\mathcal{D}}^{(\theta)}$, 98
- $C_{\mathcal{D}}$, 66, 70, 90, 398
- $S_{\mathcal{D}}$, 19, 59, 66, 86, 90, 401
- $W_{\mathcal{D}}$, 20, 67, 71, 91, 398
- $\mathbb{T}_{\mathcal{D}}^{(\theta)}$, 98

GDM notions
- $\mathcal{D}^{\mathrm{BA}}$, 229
- I^{BA}, 229
- $\mathcal{D}_T$, 98
- $\mathrm{reg}_{\mathrm{BA}}(\mathcal{D}^{\mathrm{BA}})$, 229
- $\mathrm{reg}_{\mathrm{LLE}}(\mathcal{D})$, 223
- $X_{\mathcal{D},\bullet}$, 97

Miscellaneous
- $\Delta-$admissible mesh, x
- v_θ, 98

Norms and semi-norms
- $\|w\|_{BV(\mathbb{R}^d)}$, 418
- $\|\mathcal{G}_K\|_p$, 219
- $\|u\|_{L^p(\Omega)}$, 458
- $\|\pi_K\|_p$, 217
- $\|\boldsymbol{\varphi}\|_{W_{\mathrm{div}}^q(\Omega)}$, 458
- $|\cdot|_{\mathfrak{T},p}$, 206

Polytopal tools
- ∂K, 203
- $D_{K,\sigma}$, 204
- $d_{K,\sigma}$, 204
- D_σ, 204
- $\mathcal{F}$, 203
- $\mathcal{F}_K$, 203
- $\eta_{\mathfrak{T}}$, 206
- $\nabla_{\mathfrak{T}}$, 206
- $h_{\mathcal{M}}$, 204
- $\kappa_{\mathfrak{T}}$, 207
- $\mathcal{M}$, 203
- $|K|$, 203
- $|\sigma|$, 203
- $\mathcal{M}_\sigma$, 203
- $\mathcal{N}_K$, 203
- $\boldsymbol{n}_{K,\sigma}$, 203
- $\omega^\nabla(\mathcal{D},\mathfrak{T},\boldsymbol{\Phi})$, 208
- $\omega^\Pi(\mathcal{D},\mathfrak{T},\boldsymbol{\Phi})$, 208
- $\mathcal{P}$, 203
- $\Pi_{\mathfrak{T}}$, 206
- $\mathfrak{T}$, 205
- $\theta_{\mathfrak{T}}$, 206
- $\mathbb{T}_{\mathfrak{T}}$, 206
- $\boldsymbol{x}_K$, 203
- $X_{\mathfrak{T}}$, 205
- $X_{\mathfrak{T},0}$, 205
- $\boldsymbol{x}_\sigma$, 204

Index

approximation space, 3–5
Arzelà–Ascoli theorem, 115, 440, 449
  discontinuous, 449
Aubin–Simon theorem, 107, 446

barycentric condensation, 202, 228–230, 233, 240, 241, 243, 256, 273, 274, 348, 365
barycentric dual mesh, 271
broken
  gradient, 5, 11, 207, 282
  space, 251, 282, 416
Brouwer fixed point theorem, 43, 82, 84, 88, 462

Céa's lemma, 4
Caratheodory function, 41, 63, 81, 87, 118, 133, 465
CeVeFE, *see* finite volume
coercivity
  definition, 18, 66, 70, 90, 398
  of specific schemes, 304, 313, 330, 343, 358
compactly embedded sequence, 444
compactly–continuously embedded sequence, 445
compactness, 24
  definition, 22, 67, 71, 91, 402
  implies coercivity, 23, 68, 72, 402
  of specific schemes, 313, 330, 343, 358
compatible discrete operator (CDO), xii, 347
condensation
  barycentric, 202, 228–230, 233, 240, 241, 243, 256, 273, 274, 348, 365
  static, 230, 275, 353
conforming $\mathbb{P}_k$ LLE gradient discretisation, 261
conforming method, 4, 259
conformity defect, 12, 20, 33, 34
  estimate, 209, 213, 216, 267, 278, 286, 288, 312, 331, 343, 358, 384
consistency, 19, 38, 39
  see also GD-consistency
consistency defect, 33, 34
  estimate, 249, 251–254, 267, 278, 286, 288, 294, 312, 330, 343, 359, 384
control of a GD by a polytopal toolbox
  definition, 208, 213, 216
  of specific schemes, 209, 309, 328, 342, 357, 382
convergence
  in the sense of distributions, 463
  non-linear strong, 465
  space–time weak-strong, 111
  uniform in time strong, 174
  weak-strong, 464
Crank-Nicolson scheme, 97

DDFV, *see* finite volume
Δ-admissible mesh, x, 364
density of smooth functions, 75, 118, 148, 174, 217, 248, 422
DGGD, *see* discontinuous Galerkin
Dirichlet boundary condition
  homogeneous, ix, 18, 22, 25, 26, 33, 41
discontinuous Galerkin, 319, 320
  SIPG, 319, 322
discrete compactness in $H^{-1}(\Omega)$, 109
discrete integration-by-parts, 460
discrete Poincaré inequality, 12, 19, 421
discrete Stokes formula, 416

dual norm for space–time GD, 106

elliptic problem, 32
equicontinuous, 440, 449
error estimate, 34, 267, 285
  Dirichlet BCs, 34, 62
  mixed BCs, 94
  Neumann BCs, 79
  on the trace, Fourier BCs, 79
  parabolic problem, 120
  p-Laplace problem, 47
Estimates on the translates, 422
Euler scheme, 97

finite difference, 38
  see also mimetic finite difference
finite element
  conforming
    $\mathbb{P}_k$, 261
  mixed , 301, 302
  mixed –
    Raviart-Thomas, 305, 314
  non-conforming, 5, 285, 286, 291
finite volume, x, 7, 199, 339, 362
  discrete duality (DDFV), xi, xii, 384
  discrete duality –
    CeVeFE, 199, 386
  hybrid, 347, 349
  mixed , xii, 347, 362
  MPFA, x, xii, 224, 337, 338, 340, 344
  TPFA, 7–9, 361
fixed point method, 42
fully implicit scheme, 135

Galerkin discontinuous , see discontinuous Galerkin319
Galerkin method, 3
  conforming, 4, 259–261
GD-consistency, 19, 24, 25, 27, 30, 37, 59, 66, 86, 90, 98, 401, 402
  of LLE GDs, 222, 239, 241, 242
  of specific schemes, 223, 304, 312, 313, 330, 343, 358, 367
gradient discretisation
  abstract setting, 397
  Dirichlet BCs, 18, 58
  Fourier BCs, 85
  mixed BCs, 89
  Neumann BCs, 65, 69
  time-dependent problem, 98
gradient scheme, ix, 17, 201, 322, 469
  degenerate parabolic problem, 165
  error estimate, 47, 73, 94
  Leray-Lions problem, 52, 54
  linear elliptic problem, 32, 61, 62, 93
  p-Laplace problem, 45, 46
  quasi-linear elliptic problem, 41, 80, 81, 84, 87, 88
  quasi-linear parabolic problem, 119, 148
  quasi-linear problem, 63

Hölder inequalities, 458
heterogeneous medium, x, xi, 41, 161, 202, 256, 273, 364
higher order, 30, 35, 45, 225, 312
hybrid finite volume (HFV), *see* finite volume
hybrid high order (HHO), xii
hybrid mimetic finite difference (hMFD), *see* mimetic finite difference
hybrid mimetic mixed (HMM), xii, 347, 348, 364, 412

interpolant, 19, 39, 60, 85, 218, 219, 225, 226, 257, 268, 297, 410, 413, 414, 435
interpolation error, 39
interpolation of space–time functions, 102

Jensen inequality, 460

Lagrange interpolation, 262, 263, 275
Lebesgue spaces, 457
Leray–Lions type problem, 54, 149

limit-conformity, 20, 21, 24, 37, 67, 68, 71, 72, 91, 398
  of specific schemes, 304, 313, 330, 343, 358
linear spatial interpolator, 120
local linearly exact (LLE) GD, 202, 217, 222, 243, 249

mass lumping, 202, 234, 235, 240, 241, 243, 272, 273
  conforming $\mathbb{P}_1$, 271
  non-conforming $\mathbb{P}_1$, 299
maximal monotone operator, 163
mimetic finite difference (MFD), xi, 35
  hybrid, xii, 347
  mixed-hybrid, 315
  nodal, xii, 371, 372
Minty trick, 465
mixed finite element (MFE), *see* finite element
mixed finite volume (MFV), *see* finite volume
MPFA, 337, 338
MPFA-O, 344

Newton method, 42

parabolic problem, 118
  degenerate, 161
  non-conservative, 133
piecewise constant reconstruction, 23
p-Laplace problem, 45, 472
Poincaré inequality, *see* discrete Poincaré inequality
polytopal
  gradient, 349, 415
  mesh, 203
  subset, 203
  toolbox, 207, 208, 212, 215, 307
polytope, 202, 203
power of sums inequality, 460

quasi-linear problem, 41, 118

regularity of the limit
  abstract setting, 401
  Dirichlet BCs, 25, 61
  mixed BCs, 92
  Neumann BCs, 72
Rellich theorem, 422
resolvent, 163
Richards problem, x, xiii, 161, 176, 187, 199, 271

second Strang lemma, 6, 33, 285
segment condition, 68
semi-implicit scheme, 142
Sobolev
  embedding, 418, 431, 435, 437
  space, 457, 458
    broken, 251, 306, 416
space size of a GD
  definition, 29
  of specific schemes, 267, 288, 295, 312, 332, 344, 359, 384
space–time GD, 98
stability, 18, 38, 40
  finite element, 40
  interpolant, 413
  polytopal gradient, 415
static condensation, 230, 275, 353
Stefan problem, x, xiii, 161, 176, 187, 199, 271, 475
super-admissible, 10
super-convergence, 35
SUSHI, xi, xii, 347, 349, 364

θ-scheme, 97
topological degree, 462
trace operator, 58, 71, 74

uniform-in-time
  $L^2(\Omega)$-weak convergence, 454
  compactness, 450
unisolvence, 263

vertex approximate gradient (VAG), 273–277, 470
virtual element methods (VEM), xii

weak
  derivative, 462, 463
weak-strong
  convergence, 464
  space–time convergence, 111

Young inequality, 459