The gradient discretisation method

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

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

The gradient discretisation method

A framework for the discretization of linear and nonlinear elliptic and parabolic problems

October 16, 2016

Preface

This monograph is dedicated to the presentation of the Gradient Discretisation Method (GDM) and of 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 discretization schemes for diffusion problems of different kinds: linear or non linear, steady-state or time-dependent. The schemes may be conforming or non conforming and may rely on very general poygonal or polyhedral 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, linear or non-linear, for which the GDM is particularly adapted. As a result, for these models, any scheme entering the GDM framework can then be known to converge.
Appropriate tools are then developed so as to easily check whether a given scheme satisfies the expected properties of a GDM. Thanks to these tools a number of methods can be shown to enter the GDM framework: some of these methods are classical, such as the conforming Finite Elements, the Raviart-Thomas Mixed Finite Elements, or the $\mathbb{P}_1$ non-conforming Finite Elements. Others are more recent, such as the Hybrid Mixed Mimetic or Nodal Mimetic methods, some Discrete Duality Finite Volume schemes, and some Multi-Point Flux Approximation schemes.

Marseille, Melbourne, Paris
the authors, 2016

Contents

Part I Elliptic problems

1 Motivation and basic ideas 7
1.1 Some well-known approximations of linear elliptic problems . . . 7
1.1.1 Galerkin approximations 7
1.1.2 Non-conforming $\mathbb{P}_1$ finite elements 8
1.1.3 Two-point flux finite volume on Cartesian meshes 10
1.2 Towards Gradient Schemes 14
1.3 Generalization to non-linear problems 16

2 The gradient discretisation method 21
2.1 Dirichlet boundary conditions 22
2.1.1 Homogeneous Dirichlet conditions 22
2.1.2 Non-homogeneous Dirichlet conditions 30
2.2 Neumann boundary conditions 33
2.2.1 Homogeneous Neumann conditions 33
2.2.2 Non-homogeneous Neumann conditions 37
2.2.3 Complements on trace operators 42
2.3 Non-homogeneous Fourier boundary conditions 44
2.4 Mixed boundary conditions 46

3 Elliptic problems 51
3.1 The linear case 51
3.1.1 Homogeneous Dirichlet boundary conditions 51
3.1.2 Non-homogeneous Dirichlet boundary conditions 58
3.1.3 Neumann boundary conditions 59
3.1.4 Mixed boundary conditions 62
3.2 Unknown-dependent diffusion problems 63
3.2.1 Homogeneous Dirichlet boundary conditions 64
3.2.2 Non-homogeneous Dirichlet boundary conditions 68
3.2.3 Homogeneous Neumann boundary conditions 69

3.2.4 Non-homogeneous Neumann boundary conditions 71
3.2.5 Non-homogeneous Fourier boundary conditions 73
3.3 p-Laplacian type problems: $p \in (1, +\infty)$ 75
3.3.1 An error estimate for the p-Laplace problem 75
3.3.2 Convergence of gradient schemes for fully nonlinear Leray–Lions problems 82

Part II Parabolic problems

4 Time-dependent problems: GDM and DFA 91
4.1 Space–time gradient discretisation 91
4.2 Averaged-in-time compactness 99
4.2.1 Abstract setting 99
4.2.2 Application to space–time gradient discretisations 109
4.3 Uniform-in-time compactness 115
4.3.1 Definitions and abstract results 115
4.3.2 Application to space–time gradient discretisations 121

5 Parabolic problems 127
5.1 The gradient discretisation method for a quasilinear parabolic problem 127
5.1.1 The continuous problem 128
5.1.2 The gradient scheme 129
5.1.3 Error estimate in the linear case 130
5.1.4 Convergence analysis in the non-linear case 134
5.2 Non-conservative problems 143
5.2.1 The continuous problem 143
5.2.2 Fully implicit scheme 145
5.2.3 Semi-implicit scheme 152
5.3 Non-linear time-dependent Leray–Lions problems 157
5.3.1 Model 157
5.3.2 Gradient scheme and main results 159
5.3.3 *A priori* estimates 160
5.3.4 Proof of the convergence results 162

6 Degenerate parabolic problems 171
6.1 The continuous problem 172
6.1.1 Hypotheses and notion of solution 172
6.1.2 A maximal monotone operator viewpoint 173
6.2 Gradient scheme 175
6.3 Estimates on the approximate solution 175
6.4 A first convergence theorem 180
6.5 Uniform-in-time, strong L^2 convergence results 184
6.6 Auxiliary results 190

6.7 Proof of the uniqueness of the solution to the model 197
6.8 Numerical example 205

Part III Review of gradient discretisation methods

7 Meshes and discrete tools 215
7.1 Polytopal meshes 216
7.1.1 Definition and notations 216
7.1.2 Operators, norm and regularity factors associated with a polytopal mesh 218
7.2 Polytopal toolboxes 220
7.2.1 Dirichlet boundary conditions 220
7.2.2 Non-homogeneous Dirichlet boundary conditions 225
7.2.3 Neumann and Fourier boundary conditions 226
7.2.4 Mixed boundary conditions 229
7.3 Local linearly exact GDs 231
7.3.1 $\mathbb{P}_0$-exact and $\mathbb{P}_1$-exact reconstructions 231
7.3.2 Definition and consistency of local linearly exact GDs for Dirichlet boundary conditions 235
7.3.3 From local to global basis functions, and matrix assembly 239
7.3.4 Barycentric elimination of degrees of freedom 241
7.3.5 Mass lumping 246
7.3.6 Non-homogeneous Dirichlet, Neumann and Fourier boundary conditions 250

8 Conforming methods and derived methods 257
8.1 Conforming Galerkin methods 257
8.1.1 Homogeneous Dirichlet boundary conditions 257
8.1.2 Non-homogeneous Neumann boundary conditions 258
8.2 $\mathbb{P}_k$ finite elements for homogeneous Dirichlet boundary conditions 259
8.2.1 Definition of $\mathbb{P}_k$ gradient discretisations 259
8.2.2 Properties of $\mathbb{P}_k$ gradient discretisations 263
8.3 $\mathbb{P}_k$ FE for non-homogeneous Dirichlet, Neumann and Fourier BCs 265
8.3.1 Non-homogeneous Dirichlet conditions 266
8.3.2 Neumann boundary conditions 267
8.3.3 Fourier conditions 268
8.4 Mass-lumped $\mathbb{P}_1$ finite elements 268
8.5 Vertex approximate gradient (VAG) methods 270

9 Non-conforming finite element methods and derived methods 279
9.1 Non-conforming $\mathbb{P}_1$ FE for homogeneous Dirichlet BCs 279
9.1.1 Definition of the non-conforming $\mathbb{P}_1$ gradient discretisation 279
9.1.2 Preliminary lemmas 280
9.1.3 Properties of the non-conforming $\mathbb{P}_1$ finite element method 283
9.2 Non-conforming $\mathbb{P}_1$ methods for Neumann and Fourier BCs . . . 284
9.2.1 Neumann boundary conditions 284
9.2.2 Fourier boundary conditions 286
9.3 Non-conforming $\mathbb{P}_1$ FE for non-homogeneous Dirichlet BCs . . . 286
9.4 Mass-lumped non-conforming $\mathbb{P}_1$ reconstruction 288

10 Mixed finite element $\mathbb{RT}_k$ schemes 291
10.1 The $\mathbb{RT}_k$ mixed finite element scheme for linear elliptic problems 291
10.2 Gradient discretisation from primal mixed finite element 293
10.3 Gradient discretisation from dual mixed finite element formulation 298

11 The multi-point flux approximation MPFA-O scheme 303
11.1 MPFA methods for Dirichlet boundary conditions 303
11.1.1 Definition of the MPFA gradient discretisation 303
11.1.2 Preliminary lemmas 306
11.1.3 Properties of the MPFA-O gradient discretisation 309
11.2 MPFA-O methods for Neumann and Fourier boundary conditions 310
11.2.1 Neumann boundary conditions 310
11.2.2 Fourier boundary conditions 311

12 Hybrid mimetic mixed schemes 313
12.1 HMM methods for Dirichlet boundary conditions 314
12.1.1 Definition of HMM gradient discretisations 314
12.1.2 Preliminary lemmas 320
12.1.3 Properties of HMM gradient discretisations 323
12.2 HMM methods for Neumann and Fourier boundary conditions 326
12.2.1 Neumann boundary conditions 326
12.2.2 Fourier boundary conditions 327
12.3 HMM fluxes, and link with the two-point finite volume method 327
12.4 A cell-centered variant of HMM schemes on Δ-admissible meshes 329
12.5 The SUSHI scheme for homogeneous Dirichlet conditions 330
12.5.1 Harmonic interpolation coefficients 330

13 Nodal mimetic finite difference methods 335
13.1 Definition and properties of nMFD gradient discretisations . . . 335
13.1.1 Preliminary lemmas 343
13.1.2 Properties of nMFD gradient discretisations 347
13.2 Link with discrete duality finite volume methods 348

Part IV Appendix

A Complements on LLE GDs 355
A.1 $W^{2,p}$ estimates for $S_{\mathcal{D}}$ 355
A.2 LLE GDs with generalised degrees of freedom 367
A.3 Non-linearly exact barycentric combinations 368

B Discrete functional analysis 371
B.1 Preliminary results 371
B.1.1 Geometrical properties of cells 371
B.1.2 Approximation properties 376
B.2 Discrete functional analysis for Dirichlet boundary conditions . 381
B.2.1 Discrete Sobolev embeddings 381
B.2.2 Compactness in $L^p(\Omega)$ 385
B.3 Discrete functional analysis for Neumann and Fourier BCs 386
B.3.1 Estimates involving the reconstructed trace 386
B.3.2 Discrete Sobolev embeddings 392
B.3.3 Compactness in $L^p(\Omega)$ 395
B.4 Discrete functional analysis for mixed boundary condition 396
B.4.1 Discrete Sobolev embeddings 396
B.4.2 Compactness in $L^p(\Omega)$ 398

C Technical results 399
C.1 Standard notations, inequalities and relations 399
C.1.1 Notations 399
C.1.2 Hölder inequalities 399
C.1.3 Young inequality 400
C.1.4 Jensen inequality 400
C.1.5 Power of sums 401
C.1.6 Discrete integration-by-parts (summation-by-parts) 401
C.2 Topological degree 403
C.3 Weak and strong convergences in integrals 403
C.4 Minty trick and convexity inequality 404

References 407

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 diffusion problems. A Gradient Discretisation Method is based on the choice of a set of discrete spaces and operators, referred to as a "Gradient Discretisation" (GD). Using the discrete elements of a particular GD in lieu of continuous space and operators in the weak formulation of a diffusion problem then yields a numerical scheme called a Gradient Scheme (GS) for this problem.

Considering here only the case of homogeneous Dirichlet boundary conditions, the stationary linear and non-linear diffusion problems that we shall consider 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 a regular boundary denoted by $\partial\Omega = \overline{\Omega} \setminus \Omega$,
- $\boldsymbol{a}$ 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 $\boldsymbol{a}(\boldsymbol{x}, u, \boldsymbol{\xi}) = \Lambda(\boldsymbol{x})\boldsymbol{\xi}$, which yields a linear diffusion problem. Another possible choice for $\boldsymbol{a}$ is a Leray-Lions operator such as the p-Laplacian $\boldsymbol{a}(\boldsymbol{x}, u, \boldsymbol{\xi}) = |\boldsymbol{\xi}|^{p-2}\boldsymbol{\xi}$ with $p > 1$, which yields a non linear diffusion problem. We also consider the related evolution problem:

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

Degenerate transient problems are also treated, such as the following:

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

where the functions ζ and β will mainly be assumed to be Lipschitz-continuous and non-decreasing. This model includes both the Stefan problem arising from a melting material model, and the Richards problem which models a two phase flow in a porous media problem 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.), or in image processing. In the case of underground engineering, numerical simulations have to be performed on meshes adapted to the geological layers, which include complex geometrical features such as faults, vanishing layers, inclined wells, highly heterogeneous permeability fields, local non conforming refinement. A large number of discretisation methods have recently been developed for the numerical approximation of these equations; we show in this book that several of these recent methods, and many of the more classical ones, are GDMs, in particular:

1. the conforming or non-conforming Finite Element methods, including mass lumped versions,
2. the Raviart-Thomas Mixed Finite Element methods,
3. the Multi-Point Flux Approximation (MPFA) schemes and the Discrete Duality Finite Volume (DDFV) schemes on particular grids,
4. the Hybrid Mimetic Mixed (HMM) family which includes the hybrid Mimetic Finite Difference schemes, the SUSHI scheme and the Mixed Finite Volume scheme,
5. the nodal Mimetic Finite Difference scheme.

This book is written assuming that the reader is familiar with Sobolev spaces and weak formulations of elliptic and parabolic partial differential equations. We refer to [13] 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 partial differential equations (PDEs) : for example the knowledge of one of the aforementioned methods (such as the conforming $\mathbb{P}_1$ Finite Elements on triangles).

This book is organised as follows. In Part I, we first motivate in Chapter 1 the basic concepts used to define a GDM. This method is then formally introduced in Chapter 2. Chapter 3 shows how the GDM is applied to elliptic problems. An error estimate is first obtained in the linear case. The convergence of the GDM is then proved for a generalised Leray–Lions model, which involves a non-local dependency of the operator.
Part II is devoted to the study of the GDM for linear and non-linear parabolic problems. In Chapter 4, we present the definitions and main compactness results which are used to analyse the GDM for nonlinear parabolic problems. In

Chapter 5, we start with the classical parabolic heat equation, and then study the convergence of gradient schemes for transient Leray-Lions problem. Chapter 6 covers the study of GSs for degenerate parabolic problems, including the Stefan problem and the Richards problem. We also note that, even though numerous kinds of equations are covered here, they do not form an exhaustive list of the models for which gradient schemes have been developed and analysed. In particular, the following models are not covered in this monograph: linear and non-linear elasticity equations [39], the poro-elasticity equations [61], the Stokes and Navier–Stokes equations [34, 55], and the obstacle and Signorini problems [3, 4].
Part III lists some important examples of GDMs. We first show that the standard Finite Elements Methods, the Vertex Approximate Gradient scheme, the non-conforming $\mathbb{P}_1$ Finite Elements Method and the Mixed Finite Elements Methods are GDMs. We then analyse, in the framework of GDM, the HMM family, some particular Finite Volume methods (MPFA, DDFV...), and the nodal Mimetic Finite Difference method.
In the appendix, we provide tools to establish that particular gradient discretisations satisfy the required properties for the convergence analysis of Parts I and II to hold. In particular, the following notions are introduced in Chapter A:

- local linearly exact gradient reconstruction, which rigorously describes the basic idea of numerous schemes for diffusion equations,
- barycentric elimination, which enables a reduction of the number of degrees of freedom of a method;
- mass-lumping, particularly useful to deal with time-dependent or some non-linear models.

The second chapter of the appendix, Chapter B, describes further tools useful to analyse GDs. These *discrete functional analysis* tools are gathered into the notion of polytopal toolbox, which is relevant to a wide range of gradient discretisations.

Remark 0.1 (Shaded remarks)
Shaded remarks such as this one contains notions, comments or results that can be somewhat technical, and can be skipped in a first reading.

Part I

Elliptic problems

1

Motivation and basic ideas

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 Ω is a polygonal subset of $\mathbb{R}^d$ and $f \in L^2(\Omega)$. The weak formulation of (1.1) is:

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

1.1.1 Galerkin approximations

A classical family of numerical methods to approximate this problem is given by conforming Galerkin methods: the main idea is to seek the approximate solution in 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 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. [22] for more on Finite Element approximations.
Once such a finite dimensional subspace V_h has been chosen, the Galerkin approximation of (1.2) is

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

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 Cauchy-Schwarz' inequality in the right-hand side (see Section C.1) and recalling that $\|\varphi\|^2_{H_0^1(\Omega)} = \int_\Omega |\nabla\varphi(\boldsymbol{x})|^2 \mathrm{d}\boldsymbol{x}$, we infer that

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

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 may be generalised to the case of a bilinear form a and is known as Céa's lemma [21].
If a family of subspaces $(V_h)_{h>0}$ is such that V_h becomes "ultimately dense" in $H_0^1(\Omega)$ as $h \to 0$, *i.e.* for all $\varphi \in H_0^1(\Omega)$, $\min_{w_h \in V_h} \|w_h - \varphi\|_{H_0^1(\Omega)} \to 0$ 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 is its simplicity. It is however limited to methods for which the approximation space V_h is included in the space V in which the continuous solution lives. These methods are refered to as "conforming". Numerous numerical schemes for elliptic equations are "not conforming" in the sense that the provided approximate solutions are not in $H_0^1(\Omega)$. This happens for instance in the case of the non-conforming $\mathbb{P}_1$ finite element, which yields a piecewise affine approximation and in the case of the cell centered finite volume scheme, which yields a piecewise constant approximation.

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

We again consider a simplicial mesh $\mathcal{M}$ of a domain Ω. Let $\mathcal{F}$ be the finite set of the faces of the mesh (edges in 2D), $\mathcal{F}_{\text{ext}}$ be the set of all $\sigma \in \mathcal{F}$ such that $\sigma \subset \partial\Omega$, and $\mathcal{F}_{\text{int}} = \mathcal{F} \setminus \mathcal{F}_{\text{ext}}$ the set of interior faces. For any $\sigma \in \mathcal{F}$, $\overline{\boldsymbol{x}}_\sigma$ is the barycenter (center 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 simplices of the mesh; 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

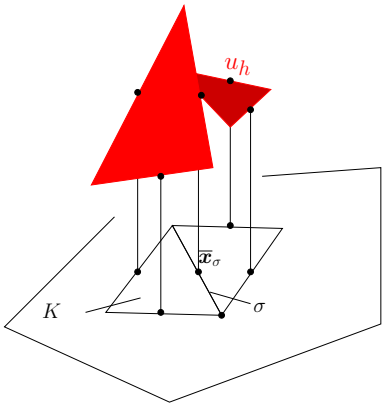


Fig. 1.1. Non-conforming $\mathbb{P}_1$ finite element, two-dimensional case

function such that $\varphi_\sigma(\overline{\boldsymbol{x}}_\sigma) = 1$ and $\varphi_\sigma(\overline{\boldsymbol{x}}_{\sigma'}) = 0$ for all $\sigma' \in \mathcal{F}\backslash\{\sigma\}$. The space V_h is clearly not a subspace of $H_0^1(\Omega)$; however, the restriction of a function of V_h is piecewise affine so that its gradient is well defined and constant. For $K \in \mathcal{M}$, let us denote by $\nabla_K \varphi_\sigma$ 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 interface 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

$$\begin{aligned} &\text{For any } u_h = \sum_{\sigma\in\mathcal{F}_{\rm int}} u_\sigma \varphi_\sigma \,, \\ &\forall K \in \mathcal{M}\,, \forall \boldsymbol{x} \in K\,,\ \nabla_{\mathcal{M}} u_h(\boldsymbol{x}) = \sum_{\sigma\in\mathcal{F}_{\rm int}} u_\sigma \nabla_K \varphi_\sigma \end{aligned} \tag{1.6}$$

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

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

If $u_h \in V_h$, then we have $\|u_h\|_h = \|\nabla_{\mathcal{M}} u_h\|_{L^2(\Omega)}$; if $u_h \in H_0^1(\Omega)$, then we have $\|u_h\|_h = \|\nabla u_h\|_{L^2(\Omega)}$. In order to approximate Problem (1.1), we define the bilinear form $a_h : (H_0^1(\Omega) + V_h)^2 \to \mathbb{R}$ by

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

We see that a_h is elliptic with an ellipticity constant $\alpha = 1$, since

$$\forall v_h \in V_h\,,\ a_h(v_h, v_h) = \|v_h\|_h^2\,.$$

The approximate solution of Problem (1.1) is defined by

$$\text{Find } u_h \in V_h \text{ such that, } \forall v_h \in V_h,\ a_h(u_h, v_h) = \int_\Omega f(\boldsymbol{x}) v_h(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}. \quad (1.7)$$

There exists one and only one solution to (1.7), and there holds the following error estimate [21, Theorem 4.2.2], based on the second Strang Lemma [67]: there exists $C > 0$, depending only on the regularity of $\mathcal{M}$ but not on h, such that

$$\begin{aligned} &\|\overline{u} - u_h\|_h \\ &\leq C \left(\inf_{v_h \in V_h} \|\overline{u} - v_h\|_h + \sup_{w_h \in V_h \setminus \{0\}} \frac{\left| a_h(\overline{u}, w_h) - \int_\Omega f(\boldsymbol{x}) w_h(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \right|}{\|w_h\|_h} \right). \end{aligned} \quad (1.8)$$

This estimate can be written in terms solely involving $\overline{u}$:

$$\|\overline{u} - u_h\|_h \leq C(S_{\mathcal{M}}(\overline{u}) + W_{\mathcal{M}}(\nabla \overline{u})), \quad (1.9)$$

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

$$S_{\mathcal{M}}(\varphi) = \inf_{v_h \in V_h} \|\varphi - v_h\|_h, \quad (1.10)$$

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

$$W_{\mathcal{M}}(\boldsymbol{\varphi}) = \sup_{w_h \in V_h \setminus \{0\}} \frac{\int_\Omega (\boldsymbol{\varphi}(\boldsymbol{x}) \cdot \nabla_{\mathcal{M}} w_h(\boldsymbol{x}) + \mathrm{div}\boldsymbol{\varphi}(\boldsymbol{x}) w_h(\boldsymbol{x})) \mathrm{d}\boldsymbol{x}}{\|w_h\|_h}. \quad (1.11)$$

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, see e.g. [22, 43].

1.1.3 Two-point flux finite volume on Cartesian meshes

A second example of a non-conforming scheme is given by the "Two-Point Flux Approximation" (TPFA) finite volume scheme [47]. The TPFA scheme on Cartesian grids, which we shall denote by TPFA-CG 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. Let us consider a rectangular mesh of a rectangle Ω ($d = 2$). In addition to the notations K, σ and $\overline{\boldsymbol{x}}_\sigma$ introduced above, we introduce the following (see Figure 1.2):

- We denote by $\boldsymbol{x}_K$ the center of mass of $K \in \mathcal{M}$, by $\mathcal{V}$ the set of vertices of the mesh, by $\mathcal{V}_K$ the set of vertices of K, and by $\mathcal{F}_K$ the set of the edges of K.

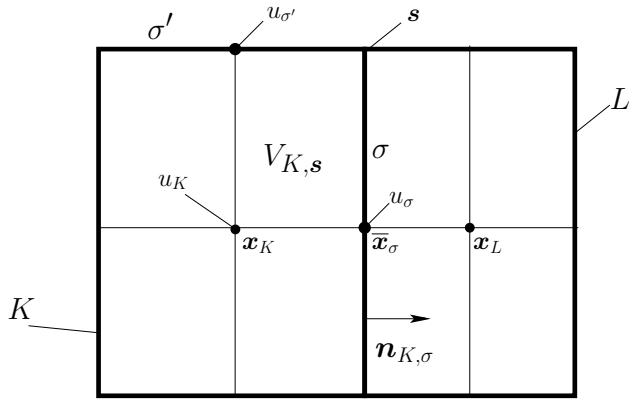


Fig. 1.2. Notation for a rectangular mesh.

- 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 $-\int_\sigma \nabla u \cdot \boldsymbol{n}_{K,\sigma} \mathrm{d}s(\boldsymbol{x})$ ($\boldsymbol{n}_{K,\sigma}$ is the normal to σ outward K, $\mathrm{d}s$ is the measure on the edges), and in writing the following discrete flux balance in each cell

$$\forall K \in \mathcal{M}\,, \ \sum_{\sigma \in \mathcal{F}_K} F_{K,\sigma} = \int_K f(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}, \tag{1.12}$$

and flux conservativity across each interior edge:

$$\forall \sigma \in \mathcal{F}_{\mathrm{int}} \text{ common face of } K \text{ and } L\,, \ F_{K,\sigma} + F_{L,\sigma} = 0. \tag{1.13}$$

Relation (1.12) simply mimicks the Stokes formula applied to the continuous problem (1.1):

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

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

$$F_{K,\sigma} = -|\sigma| \frac{u_\sigma - u_K}{\mathrm{dist}(\overline{\boldsymbol{x}}_\sigma, \boldsymbol{x}_K)}. \tag{1.14}$$

The boundary condition is imposed by setting

$$u_\sigma = 0 \text{ if } \sigma \subset \partial\Omega, \tag{1.15}$$

There is no clear way to see the TPFA-CG scheme method as a nonconforming finite element method. However, it can be recast into a variational formulation. We first 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 (1.12) by v_K and summing on $K \in \mathcal{M}$, we get

$$\sum_{K\in\mathcal{M}}\sum_{\sigma\in\mathcal{F}_K} v_K F_{K,\sigma} = \sum_{K\in\mathcal{M}} v_K \int_K f(\boldsymbol{x})\mathrm{d}\boldsymbol{x}. \tag{1.16}$$

We then notice that

$$\sum_{K\in\mathcal{M}}\sum_{\sigma\in\mathcal{F}_K} v_K F_{K,\sigma} = \sum_{K\in\mathcal{M}}\sum_{\sigma\in\mathcal{F}_K} (v_K - v_\sigma) F_{K,\sigma}. \tag{1.17}$$

Indeed, if $\sigma \subset \partial\Omega$, then $(v_K - v_\sigma)F_{K,\sigma} = v_K F_{K,\sigma}$. If σ is the common face between two control volumes K and L, then v_σ is multiplied in the above sum by $F_{K,\sigma} + F_{L,\sigma}$, which vanishes thanks to (1.13). Thus, using (1.17) into (1.16) and invoking (1.14), we get

$$\sum_{K\in\mathcal{M}}\sum_{\sigma\in\mathcal{F}_K} \frac{|\sigma|}{\mathrm{dist}(\overline{\boldsymbol{x}}_\sigma, \boldsymbol{x}_K)}(v_\sigma - v_K)(u_\sigma - u_K) = \sum_{K\in\mathcal{M}} v_K \int_K f(\boldsymbol{x})\mathrm{d}\boldsymbol{x}. \tag{1.18}$$

Conversely, it is possible, assuming Expression (1.14) for the fluxes, to deduce (1.12), (1.13) and (1.15) from (1.18) with convenient choices for 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 ones equal to 0). Moreover, Relation (1.18) can be expressed in terms of reconstructed functions and gradients, using the discrete values defined on K and σ.

- We define $X_{\mathcal{D},0}$ as the space of all real families $u_\mathcal{D} = ((u_K)_{K\in\mathcal{M}},(u_\sigma)_{\sigma\in\mathcal{F}})$ satisfying the boundary conditions(1.15).
- For $u_\mathcal{D} \in X_{\mathcal{D},0}$, we 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 faces of K sharing the vertex $\boldsymbol{s}$, we define the reconstructed gradient $\nabla_{K,\boldsymbol{s}} u_\mathcal{D} = \nabla^{(\sigma)}_{K,\boldsymbol{s}} u_\mathcal{D}\ \boldsymbol{n}_{K,\sigma} + \nabla^{(\sigma')}_{K,\boldsymbol{s}} u_\mathcal{D}\ \boldsymbol{n}_{K,\sigma'}$ with

$$\nabla^{(\sigma)}_{K,\boldsymbol{s}} u_\mathcal{D} = \frac{u_\sigma - u_K}{\mathrm{dist}(\overline{\boldsymbol{x}}_\sigma, \boldsymbol{x}_K)} \quad \text{and} \quad \nabla^{(\sigma')}_{K,\boldsymbol{s}} u_\mathcal{D} = \frac{u_{\sigma'} - u_K}{\mathrm{dist}(\overline{\boldsymbol{x}}_{\sigma'}, \boldsymbol{x}_K)}.$$

  We then denote 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

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

As a result, we can rewrite (1.18) in the form of a discrete variational problem:

$$\begin{cases}\text{Find } u_{\mathcal{D}}\in X_{\mathcal{D},0} \text{ such that, for all } v_{\mathcal{D}}\in X_{\mathcal{D},0},\\ \displaystyle\int_\Omega\nabla_{\mathcal{D}}u_{\mathcal{D}}(\boldsymbol{x})\cdot\nabla_{\mathcal{D}}v_{\mathcal{D}}(\boldsymbol{x})\mathrm{d}\boldsymbol{x}=\int_\Omega f(\boldsymbol{x})\Pi_{\mathcal{D}}v_{\mathcal{D}}(\boldsymbol{x})\mathrm{d}\boldsymbol{x}.\end{cases}\tag{1.19}$$

The study of the TPFA-CG scheme has been done in [47] using finite volume techniques, and the following results are proven: if the size of the mesh tends to 0, 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 (Unstructured meshes). The analysis of the TPFA scheme of [47] also holds in the case of unstructured meshes, provided an orthogonality condition holds (see [47, Definition 9.1]). However, in the unstructured case, it does not seem to be possible to write the scheme under the form (1.19), except in the so-called "super-admissible" case, *i.e.* when the center of mass is also the intersection of the orthogonal bisectors; therefore, the general TPFA scheme is not included in the framework of the gradient schemes studied in this book.

The question arises to know whether the TPFA-CG scheme could be studied using the non-conforming techniques of Section 1.1.2. There are a series of objections to this approach:

1. Comparing the right-hand sides of (1.7) and (1.19) we see that 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 degrees of freedom $(v_\sigma)_{\sigma\in\mathcal{F}}$ of $v_{\mathcal{D}}\in X_{\mathcal{D},0}$, and there is thus 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 $V_h+H^1_0(\Omega)$, which gives back $\int_\Omega\nabla_{\mathcal{D}}u_{\mathcal{D}}(\boldsymbol{x})\cdot\nabla_{\mathcal{D}}v_{\mathcal{D}}(\boldsymbol{x})\mathrm{d}\boldsymbol{x}$ for elements of V_h, and $\int_\Omega\nabla\overline{u}(\boldsymbol{x})\cdot\nabla\boldsymbol{v}(\boldsymbol{x})\mathrm{d}\boldsymbol{x}$ for elements of $H^1_0(\Omega)$.
3. The same problem arises for the definition of the norm $\|\cdot\|_h$.

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

1.2 Towards Gradient Schemes

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 degrees of freedom describing the finite dimensional space in which the approximate solution is sought. We already called $X_{\mathcal{D},0}$ this finite dimensional space ("$\mathcal{D}$" 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 degrees of freedom, 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

$$\begin{cases} \text{Find } u_{\mathcal{D}} \in X_{\mathcal{D},0} \text{ such that, for all } v_{\mathcal{D}} \in X_{\mathcal{D},0}, \\ \displaystyle\int_\Omega \nabla_{\mathcal{D}} u_{\mathcal{D}}(\boldsymbol{x}) \cdot \nabla_{\mathcal{D}} v_{\mathcal{D}}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} = \int_\Omega f(\boldsymbol{x}) \Pi_{\mathcal{D}} v_{\mathcal{D}}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \end{cases} \tag{1.20}$$

for suitable choices of $(X_{\mathcal{D},0}, \Pi_{\mathcal{D}}, \nabla_{\mathcal{D}})$. 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, $\Pi_{\mathcal{D}} v_{\mathcal{D}} \in C(\Omega)$ is the piecewise linear function on the mesh which 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 barycenters of the edges, $\Pi_{\mathcal{D}} v_{\mathcal{D}}$ is the piecewise linear function on the mesh which takes these values at these barycenters, and $\nabla_{\mathcal{D}} v_{\mathcal{D}} = \nabla_{\mathcal{M}}(\Pi_{\mathcal{D}} v_{\mathcal{D}})$ is the broken gradient defined in (1.6).
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 the previous section.

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 (1.20) (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 (1.20) 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 nonconforming finite element method; however, we no longer assume that a bilinear form a_h or a norm $\|\cdot\|_h$ can be defined with argument $\Pi_{\mathcal{D}} u$; as mentioned above, this is mandatory if we want to include the TPFA-CG scheme in this framework.
By noticing that (1.2) implies that $-\Delta\overline{u} = f$ in the sense of distributions, we get from (1.20) that, for any $v_{\mathcal{D}} \in X_{\mathcal{D},0}$,

$$\int_\Omega \nabla_{\mathcal{D}} u_{\mathcal{D}}(\boldsymbol{x}) \cdot \nabla_{\mathcal{D}} v_{\mathcal{D}}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} = \int_\Omega -\Delta\overline{u}(\boldsymbol{x}) \Pi_{\mathcal{D}} v_{\mathcal{D}}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}. \tag{1.21}$$

If $\Pi_{\mathcal{D}} v_{\mathcal{D}}$ were a classical regular function, 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 (1.11), by a function $W_{\mathcal{D}}(\boldsymbol{\varphi})$ such that, for any sufficiently regular function $\boldsymbol{\varphi} : \Omega \mapsto \mathbb{R}^d$, and for all $v_{\mathcal{D}} \in X_{\mathcal{D},0}$,

$$W_{\mathcal{D}}(\boldsymbol{\varphi}) = \sup_{v_{\mathcal{D}}\in X_{\mathcal{D},0}} \frac{\displaystyle\int_\Omega (\boldsymbol{\varphi}(\boldsymbol{x})\cdot\nabla_{\mathcal{D}} v_{\mathcal{D}}(\boldsymbol{x}) + \mathrm{div}\boldsymbol{\varphi}(x)\Pi_{\mathcal{D}} v_{\mathcal{D}}(\boldsymbol{x})\mathrm{d}\boldsymbol{x})\mathrm{d}\boldsymbol{x}}{\|\nabla_{\mathcal{D}} v_{\mathcal{D}}\|_{L^2(\Omega)^d}}. \quad (1.22)$$

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 – we come back to this further down. 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 (1.22) and using (1.21) to compute $\int_\Omega \Delta\overline{u}(\boldsymbol{x})\Pi_{\mathcal{D}} v_{\mathcal{D}}(\boldsymbol{x})\mathrm{d}\boldsymbol{x}$, we obtain an approximate version of (1.4):

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

We now take 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

$$\|\nabla_{\mathcal{D}} u_{\mathcal{D}} - \nabla_{\mathcal{D}} w_{\mathcal{D}}\|_{L^2(\Omega)^d} \leq \|\nabla\overline{u} - \nabla_{\mathcal{D}} w_{\mathcal{D}}\|_{L^2(\Omega)^d} + W_{\mathcal{D}}(\nabla\overline{u}). \quad (1.23)$$

We now define the "best interpolation error" (in the spirit of (1.10)) by

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

We pick $w_{\mathcal{D}}$ which 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 (1.23) gives

$$\|\nabla_{\mathcal{D}} u_{\mathcal{D}} - \nabla \overline{u}\|_{L^2(\Omega)^d} \leq 2S_{\mathcal{D}}(\overline{u}) + W_{\mathcal{D}}(\nabla \overline{u}). \tag{1.24}$$

Let us now study how $\Pi_{\mathcal{D}} u_{\mathcal{D}}$ approximates $\overline{u}$. To this aim, we assume a discrete Poincaré inequality:

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

This enables us to write

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

Estimate (1.23) then shows that

$$\|\Pi_{\mathcal{D}} u_{\mathcal{D}} - \overline{u}\|_{L^2(\Omega)} \leq (C_{\mathcal{D}} + 1) S_{\mathcal{D}}(\overline{u}) + C_{\mathcal{D}} W_{\mathcal{D}}(\nabla \overline{u}). \tag{1.25}$$

Equations (1.24) and (1.25) 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 we take a sequence $(X_{\mathcal{D}_m,0}, \Pi_{\mathcal{D}_m}, \nabla_{\mathcal{D}_m})_{m\in\mathbb{N}}$ such that

(P1) $(C_{\mathcal{D}_m})_{m\in\mathbb{N}}$ is bounded,
(P2) $S_{\mathcal{D}_m}(\overline{u}) \to 0$ as $m \to \infty$,
(P3) $W_{\mathcal{D}_m}(\nabla \overline{u}) \to 0$ as $m \to \infty$,

then (1.24) and (1.25) 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 $L^2(\Omega)^d$.

The previous reasoning highlights 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 (1.20). 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 (1.20). Property (P2) states that $\Pi_{\mathcal{D}}$ and $\nabla_{\mathcal{D}}$ are *consistent* reconstructions of functions and their gradient; it allows us to approximate $\overline{u}$ 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 must ultimately behave as in the conforming case and should, in the limit, satisfy the exact Stokes formula.

1.3 Generalization to non-linear problems

The framework of the convergence analysis must be able to handle non-linear equations. Assume we now wish to approximate the following problem:

$$\begin{cases} \beta(\overline{u}) - \Delta\overline{u} = f \text{ in } \Omega, \\ \overline{u} = 0 \text{ on } \partial\Omega, \end{cases} \tag{1.26}$$

with the same notations as in Section 1.1, and where the function $\beta : \mathbb{R} \to \mathbb{R}$ is continuous. The weak formulation of (1.26) is:

$$\begin{cases} \text{Find } \overline{u} \in H_0^1(\Omega) \text{ such that, for all } v \in H_0^1(\Omega), \\ \displaystyle\int_\Omega (\beta(\overline{u}(\boldsymbol{x}))v(\boldsymbol{x}) + \nabla\overline{u}(\boldsymbol{x}) \cdot \nabla v(\boldsymbol{x}))\mathrm{d}\boldsymbol{x} = \int_\Omega f(\boldsymbol{x})v(\boldsymbol{x})\mathrm{d}\boldsymbol{x}. \end{cases} \tag{1.27}$$

It can then be shown that there exists at least one solution to (1.27). Let us assume that we use a conforming Galerkin method, say the $\mathbb{P}_1$ finite element method. Denoting by V_h the space of continuous piecewise linear functions on a triangular mesh of Ω, the $\mathbb{P}_1$ finite element approximation of (1.27) could be

$$\begin{cases} \text{Find } u_h \in V_h \text{ such that, for all } v_h \in V_h, \\ \displaystyle\int_\Omega (\beta(u_h(\boldsymbol{x}))v_h(\boldsymbol{x}) + \nabla u_h(\boldsymbol{x}) \cdot \nabla v_h(\boldsymbol{x}))\mathrm{d}\boldsymbol{x} = \int_\Omega f(\boldsymbol{x})v_h(\boldsymbol{x})\mathrm{d}\boldsymbol{x}. \end{cases} \tag{1.28}$$

Although this approximate problem has at least one solution, its analysis presents three major difficulties. The first one is the difficulty to compute, if $u_h = \sum_{\boldsymbol{s}' \in \mathcal{V}_{\text{int}}} u_{\boldsymbol{s}'}\varphi_{\boldsymbol{s}'}$, the quantity

$$\int_\Omega \beta\left(\sum_{\boldsymbol{s}' \in \mathcal{V}_{\text{int}}} u_{\boldsymbol{s}'}\varphi_{\boldsymbol{s}'}(\boldsymbol{x})\right)\varphi_{\boldsymbol{s}}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \tag{1.29}$$

for a given interior vertex $\boldsymbol{s}$ of the mesh; due to the non-linearity β, the integrand may not be a piecewise polynomial and thus exact quadrature rules cannot be used. The second one is to define an algorithm to approximatie the solution of the non-linear system of equations provided by (1.28). The third one 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 (1.28) with $v_h = \varphi_{\boldsymbol{s}}$, the term (1.29) with $\omega_{\boldsymbol{s}}\beta(u_{\boldsymbol{s}})$ – where $\omega_{\boldsymbol{s}}$ is some weight to be defined. The gradient scheme 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). We simply let $X_{\mathcal{D},0} = \mathbb{R}^{\mathcal{V}_{\text{int}}}$ as before and we define, for $u \in X_{\mathcal{D},0}$,

$$\begin{aligned} \Pi_{\mathcal{D}}u &= \sum_{\boldsymbol{s} \in \mathcal{V}_{\text{int}}} u_{\boldsymbol{s}}\mathbf{1}_{K_{\boldsymbol{s}}} \qquad \text{(piecewise constant reconstruction)}, \\ \nabla_{\mathcal{D}}u &= \sum_{\boldsymbol{s} \in \mathcal{V}_{\text{int}}} u_{\boldsymbol{s}}\nabla\varphi_{\boldsymbol{s}}, \end{aligned} \tag{1.30}$$

where $\mathbf{1}_{K_{\boldsymbol{s}}}$ is the characteristic function of $K_{\boldsymbol{s}}$. Then the scheme (1.28) is replaced with

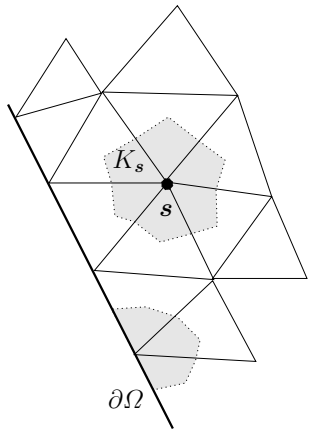


Fig. 1.3. Definition of $K_{\boldsymbol{s}}$

$$\begin{cases} \text{Find } u \in X_{\mathcal{D},0} \text{ such that, for all } v \in X_{\mathcal{D},0}, \\ \displaystyle\int_\Omega (\beta(\Pi_{\mathcal{D}} u(\boldsymbol{x}))\Pi_{\mathcal{D}} v(\boldsymbol{x}) + \nabla_{\mathcal{D}} u(\boldsymbol{x}) \cdot \nabla_{\mathcal{D}} v(\boldsymbol{x}))\mathrm{d}\boldsymbol{x} \\ \qquad = \displaystyle\int_\Omega f(\boldsymbol{x})\Pi_{\mathcal{D}} v(\boldsymbol{x})\mathrm{d}\boldsymbol{x}. \end{cases} \tag{1.31}$$

We first notice that, since $\Pi_{\mathcal{D}} u$ and $\Pi_{\mathcal{D}} v$ are piecewise constant, this scheme leads to a very simple computation of both the first term in the left-hand side and the term in the right-hand side. 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 is of crucial importance for the obtention of the estimates (see section 6.3). It is also important from a numerical point of view since it facilitates the implementation of the scheme. If $(X_{\mathcal{D}_m,0}, \Pi_{\mathcal{D}_m}, \nabla_{\mathcal{D}_m})_{m\in\mathbb{N}}$ is a sequence of mass-lumped $\mathbb{P}_1$ finite elements it is possible to show that properties (P1), (P2) and (P3) as defined at the end of Section 1.2 hold, provided that they hold for the underlying $\mathbb{P}_1$ finite elements. Then, only using those properties, it is possible to show that:

1. The scheme (1.31) has at least one solution, which we denote 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 $\overline{u} \in H^1_0(\Omega)$, $\nabla_{\mathcal{D}_m} u_m$ 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 necessary to prove that $\overline{\beta} = \beta(\overline{u})$, and thus complete the convergence proof of the scheme:

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

This property can be established for gradient schemes defined by 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), which provide the foundations to build gradient schemes, in addition with the piecewise constant reconstruction property in the case of some nonlinear problems (see (1.30)).

2

The gradient discretisation method

A gradient discretization method (GDM) is a numerical technique used to find approximate solutions to boundary value problems for elliptic and parabolic partial differential equations. As mentioned in 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** $X_{\mathcal{D}}$ is a, e.g. the values at the nodes of the mesh, as in the conforming P1 finite element method, at a particular point of the cell, as in the TPFA-CG scheme, or at a particular point of the faces, as in the non-conforming P1 finite element method),
- a **function reconstruction** operator $\Pi_{\mathcal{D}}$ which transforms an element of $X_{\mathcal{D}}$ into a function defined a.e. on the physical domain Ω.
- $\nabla_{\mathcal{D}}$ is a an **approximate gradient reconstruction**, which builds a "discrete gradient" (vector-valued function) defined a.e. on Ω from the discrete unknowns.

In the present chapter, we define the concept of gradient discretisation and list the properties of the spaces and mappings that are important for the convergence analysis of the GDM. This convergence analysis is performed in Chapter 3 for linear and non-linear elliptic problems, in Chapter 5 for linear and non-degenerate non-linear parabolic problems, and in Chapter 6 for some degenerate parabolic problems.

The idea of the GDM is then to mimick the weak formulation or the problem to be solved, by building, thanks to the function and gradient reconstructions, a discrete weak formulation which will be the gradient scheme.

The convergence analysis of the GDM depends, of course, on the nature of the PDE to be solved. The definition the GDM, on the other hand, only depends to a large extent only on the boundary conditions. Therefore, the construction of a GD varies accordingly to the nature of these conditions.
For simplicity, we start with homegeneous and non-homogeneous Dirichlet boundary conditions in Section 2.1, and then address the case of the home-

geneous and non-homogeneous Neumann boundary conditions in Section 2.2. The gradient discretisations defined for these cases may then be easily generalized to the case of Robin conditions (Section 2.3) and mixed conditions (Section 2.4).

Throughout this book, Ω is a connected open bounded subset of $\mathbb{R}^d$ which is the physical domain over which the p.d.e. is studied, $d \in \mathbb{N}^\star$ is the space dimension, and $p \in (1, +\infty)$ denotes a regularity index of the sought solution. In some abstract theorems, p might be allowed to take the value 1.

2.1 Dirichlet boundary conditions

2.1.1 Homogeneous Dirichlet conditions

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. This gradient reconstruction must be chosen such that $\|\cdot\|_{\mathcal{D}} := \|\nabla_{\mathcal{D}} \cdot\|_{L^p(\Omega)^d}$ is a norm on $X_{\mathcal{D},0}$.*

In the following chapters, we shall construct some gradient schemes for several problems, starting from a gradient discretisation. In order to show the convergence of the scheme, we use some properties of consistency and stability. As in the framework of the Finite Element method, stability is obtained through some uniform coercivity of the discrete operator which relies on a discrete Poincaré inequality.

Definition 2.2 (Coercivity, Dirichlet conditions)

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 $\|\Pi_{\mathcal{D}} v\|_{L^p(\Omega)} \leq C_{\mathcal{D}} \|\nabla_{\mathcal{D}} v\|_{L^p(\Omega)^d}$ 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

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

Remark 2.5 (Definition of the interpolant $P_{\mathcal{D}}$). 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 $P_{\mathcal{D}}\varphi \in X_{\mathcal{D},0}$ that realises the minimum in $S_{\mathcal{D}}(\varphi)$, that is, such that

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

We will write

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

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

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

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 $P_{\mathcal{D}}^{(2)}$ and of its approximation properties).

A wellknown 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 $H_{\mathrm{div}}(\Omega) = \{\boldsymbol{\varphi} \in L^2(\Omega)^d \ : \ \mathrm{div}\boldsymbol{\varphi} \in L^2(\Omega)\}$. When dealing with non-conforming method, this property is no longer exact at the discrete level. The concept of limit-conformity which we now introduce states that the discrete gradient and divergence operator satisfy this property asymptotically. Since we shall be dealing with non linear problems, we introduce, or 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 $W_{\mathrm{div}}^2(\Omega) = H_{\mathrm{div}}(\Omega)$.

Definition 2.6 (Limit-conformity, Dirichlet conditions)

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

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

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 gradients satisfy the divergence (Stokes) formula (2.4). If the method is *conforming* in the sense that $X_{\mathcal{D},0}$ is a subspace of $W_0^{1,p}(\Omega)$, $\nabla_{\mathcal{D}} u = \nabla u$ and $\Pi_{\mathcal{D}} u = u$ for all $u \in X_{\mathcal{D},0} \subset W_0^{1,p}$, the set $\{\Pi_{\mathcal{D}} v, v \in X_{\mathcal{D},0}\}$ is a subspace of $W_0^{1,p}(\Omega)$ and for all $v \in X_{\mathcal{D},0}$, there exists $u \in W_0^{1,p}(\Omega)$ such that $\Pi_{\mathcal{D}} v = u$ and $\nabla_{\mathcal{D}} v = \nabla u$. 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 equivalent condition for the limit-conformity property facilitates the proof of the regularity of a possible limit (Lemma 2.12 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*

$$\begin{aligned} &\forall (\boldsymbol{\varphi}, u) \in W^{p'}_{\mathrm{div}}(\Omega) \times X_{\mathcal{D},0}\,, \\ &\widetilde{W}_{\mathcal{D}}(\boldsymbol{\varphi}, u) = \int_\Omega (\nabla_{\mathcal{D}} u(\boldsymbol{x}) \cdot \boldsymbol{\varphi}(\boldsymbol{x}) + \Pi_{\mathcal{D}} u(\boldsymbol{x}) \mathrm{div} \boldsymbol{\varphi}(\boldsymbol{x}))\, \mathrm{d}\boldsymbol{x}. \end{aligned} \tag{2.8}$$

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,

$$\forall \boldsymbol{\varphi} \in W^{p'}_{\mathrm{div}}(\Omega),\ \lim_{m\to\infty} \widetilde{W}_{\mathcal{D}_m}(\boldsymbol{\varphi}, u_m) = 0. \tag{2.9}$$

Proof. Let us remark that $W_{\mathcal{D}}(\boldsymbol{\varphi}) = \sup_{u\in X_{\mathcal{D},0}\setminus\{0\}} |\widetilde{W}_{\mathcal{D}}(\boldsymbol{\varphi}, u)|/\|u\|_{\mathcal{D}}$. The proof that (2.7) implies (2.9) 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 (2.9). ∎

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

Definition 2.8 (Compactness, Dirichlet conditions)

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 $L^p(\Omega)$.

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 3.5 and Remark 3.6).

Lemma 2.9 (Compactness implies coercivity). *Let $(\mathcal{D}_m)_{m\in\mathbb{N}}$ be a compact sequence of gradient discretisations in the sense of Definition 2.8. Then it is coercive in the sense of Definition 2.2.*

Proof. Let us assume that the sequence 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. ∎

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 will be a key argument for non linear problems.

Definition 2.10 (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) degrees of freedom 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 degrees of freedom are involved in the definition of the reconstructed gradients, but in several examples some degrees of freedom 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 2.11. 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 degrees of freedom I in $X_{\mathcal{D},0}$. We should therefore denote $g^B(u)$ or $(uv)^B$ to emphasize

the dependency on B but, in practice, we will remove this exponent B as the degrees of freedom are usually canonically chosen and fixed throughout the whole study of a gradient scheme.

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 framework of gradient schemes, this means proving that $\|u_{\mathcal{D}}\|_{\mathcal{D}}$ remains bounded. Lemma 2.12 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 2.12 (Regularity of the limit, homogeneous Dirichlet BCs). *Let $(\mathcal{D}_m)_{m\in\mathbb{N}}$ be a sequence of gradient discretisations in the sense of Definition 2.1 which is coercive (Definition 2.2) and limit-conforming (Definition 2.6). Let $u_m \in X_{\mathcal{D}_m,0}$ be such that $(\|u_m\|_{\mathcal{D}_m})_{m\in\mathbb{N}}$ remains bounded. Then there exist a subsequence of $(\mathcal{D}_m, u_m)_{m\in\mathbb{N}}$, denoted in the same way, and $u \in W_0^{1,p}(\Omega)$ 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. Thanks to the coercivity, the sequence $(\Pi_{\mathcal{D}_m} u_m)_{m\in\mathbb{N}}$ remains 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 $\boldsymbol{v}$. We extend $\Pi_{\mathcal{D}_m} u_m$, u, $\nabla_{\mathcal{D}_m} u_m$ and $\boldsymbol{v}$ by 0 outside Ω, and the same 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 (2.9) gives

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

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 space-consistency, limit-conformity and compactness of a sequence of gradient discretisations.

Lemma 2.13 (On 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 $\mathcal{R}$ in $W_0^{1,p}(\Omega)$ such that*

$$\forall \psi \in \mathcal{R}\,,\ \lim_{m\to\infty} S_{\mathcal{D}_m}(\psi) = 0. \tag{2.10}$$

Proof. Let us assume that (2.10) holds and let us prove (2.3) (the converse is straightforward, take $\mathcal{R} = W_0^{1,p}(\Omega)$). Let $\varphi \in W_0^{1,p}(\Omega)$ and $\varepsilon > 0$. Take $\psi \in \mathcal{R}$ such that $\|\varphi - \psi\|_{W_0^{1,p}(\Omega)} \leq \varepsilon$. For $v \in X_{\mathcal{D},0}$, the triangle inequality yields

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

Hence, $S_{\mathcal{D}_m}(\varphi) \leq S_{\mathcal{D}_m}(\psi) + ||\varphi - \psi||_{W_0^{1,p}(\Omega)} \leq S_{\mathcal{D}_m}(\psi) + \varepsilon$, we get from (2.10) that $\limsup_{m\to\infty} S_{\mathcal{D}_m}(\varphi) \leq \varepsilon$. The proof is then completed by letting $\varepsilon \to 0$. ∎

Lemma 2.14 (Equivalent condition for limit-conformity, Dirichlet BCs).
Let $(\mathcal{D}_m)_{m\in\mathbb{N}}$ be a coercive sequence of gradient discretisations in the sense of Definition 2.2. Then $(\mathcal{D}_m)_{m\in\mathbb{N}}$ is limit conforming in the sense of Definition 2.6 if and only if there exists a dense subset $\mathcal{S}$ in $W_{\rm div}^{p'}(\Omega)$ (endowed with the norm $\|\boldsymbol{\varphi}\|_{W_{\rm div}^{p'}(\Omega)} = \|\boldsymbol{\varphi}\|_{L^{p'}(\Omega)^d} + \|{\rm div}\boldsymbol{\varphi}\|_{L^{p'}(\Omega)}$) such that

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

Remark 2.15. If Ω is a locally star-shaped open set in $\mathbb{R}^d$ (which is in particular the case if Ω is polyhedral as in Section 7.1), then $\mathcal{S} = C_c^\infty(\mathbb{R}^d)^d$ is dense in $W_{\rm div}^{p'}(\Omega)$ and can therefore be used in Lemma 2.14.

Proof. Let $C_P \in \mathbb{R}_+$ be such that $C_{\mathcal{D}_m} \leq C_P$. Let us assume that (2.11) holds and let us prove (2.7). Let $\boldsymbol{\varphi} \in W_{\rm div}^{p'}(\Omega)$. Take $\varepsilon > 0$ and $\boldsymbol{\psi} \in \mathcal{S}$ such that $\|\boldsymbol{\varphi} - \boldsymbol{\psi}\|_{W_{\rm div}^{p'}(\Omega)} \leq \varepsilon$, which means that $\|\boldsymbol{\varphi} - \boldsymbol{\psi}\|_{L^{p'}(\Omega)^d} \leq \varepsilon$ and $\|{\rm div}\boldsymbol{\varphi} - {\rm div}\boldsymbol{\psi}\|_{L^{p'}(\Omega)} \leq \varepsilon$. We have, using the coercivity assumption,

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

Using (2.11) we deduce that $\limsup_{m\to\infty} W_{\mathcal{D}_m}(\boldsymbol{\varphi}) \leq (1 + C_P)\varepsilon$ and the proof is completed by letting $\varepsilon \to 0$. ∎

Lemma 2.16 (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

$$\forall \boldsymbol{\xi} \in \mathbb{R}^d,\quad T_{\mathcal{D}}(\boldsymbol{\xi}) = \max_{v\in X_{\mathcal{D},0}\setminus\{0\}} \frac{\|\Pi_{\mathcal{D}} v(\cdot + \boldsymbol{\xi}) - \Pi_{\mathcal{D}} v\|_{L^p(\mathbb{R}^d)}}{\|v\|_{\mathcal{D}}}, \tag{2.12}$$

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 $||u_{m(k)}||_{\mathcal{D}_{m(k)}} = 1$. 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

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

and therefore $\tau_A(\boldsymbol{\xi} + \boldsymbol{\xi}') \leq \tau_A(\boldsymbol{\xi}) + \tau_A(\boldsymbol{\xi}')$. Hence, if $\lim_{|\boldsymbol{\xi}|\to 0} \tau_A(\boldsymbol{\xi}) = 0$, then τ_A is finite on a neighborhood 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 $|\boldsymbol{\xi}_0| > \text{diam}(\Omega)$, for all $w \in A$ we see that $\widetilde{w}(\cdot + \boldsymbol{\xi})$ and $\widetilde{w}$ have disjoint supports and therefore

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

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

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

(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. ∎

The following lemma is an immediate consequence of the previous ones and facilitates, in many practical situations, the verification of the core properties of gradient discretisations.

Lemma 2.17 (Sufficient conditions, homogeneous Dirichlet BCs). *Let $\mathcal{F}$ be a family of gradient discretisations in the sense of Definition 2.1. Assume that there exist $C, \nu \in (0,\infty)$ and, for all $\mathcal{D} \in \mathcal{F}$, a real value $h_{\mathcal{D}} \in (0,+\infty)$ such that:*

$$S_{\mathcal{D}}(\varphi) \le C h_{\mathcal{D}} \|\varphi\|_{W^{2,\infty}(\Omega)}, \text{ for all } \varphi \in C_c^\infty(\Omega), \tag{2.13a}$$

$$W_{\mathcal{D}}(\boldsymbol{\varphi}) \le C h_{\mathcal{D}} \|\boldsymbol{\varphi}\|_{(W^{1,\infty}(\mathbb{R}^d))^d}, \text{ for all } \boldsymbol{\varphi} \in C_c^\infty(\mathbb{R}^d)^d, \tag{2.13b}$$

$$\max_{v\in X_{\mathcal{D},0}\setminus\{0\}} \frac{\|\Pi_{\mathcal{D}} v(\cdot+\boldsymbol{\xi}) - \Pi_{\mathcal{D}} v\|_{L^p(\mathbb{R}^d)}}{\|v\|_{\mathcal{D}}} \le C|\boldsymbol{\xi}|^\nu, \text{ for all } \boldsymbol{\xi} \in \mathbb{R}^d. \tag{2.13c}$$

Then any sequence $(\mathcal{D}_m)_{m\in\mathbb{N}} \subset \mathcal{F}$ such that $h_{\mathcal{D}_m} \to 0$ as $m \to \infty$ is GD-consistent, limit-conforming and compact (and therefore coercive).

Note that for several gradient schemes, the parameter $h_{\mathcal{D}}$ in the above lemma can simply be chosen as the mesh size.

2.1.2 Non-homogeneous Dirichlet conditions

We present here the framework of gradient discretisations for diffusion problems with 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 $p \in (1,+\infty)$. This concept requires more regularity on Ω than in Section 2.1.1 and we therefore assume here that Ω has a Lipschitz boundary. We recall in Section 2.2.3 some facts about the trace operator γ.

Definition 2.18 (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 degrees of freedom and to the boundary degrees of freedom,*
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. It must be chosen such that $\|\cdot\|_\mathcal{D} := \|\nabla_\mathcal{D} \cdot\|_{L^p(\Omega)^d}$ is a norm on $X_{\mathcal{D},0}$.*

Remark 2.19 (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 (3.22b)), 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 12.2.
In that case, the GD-consistency (Definition 2.20) 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, 2.6, 2.8 and 2.10 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 2.20 (GD-consistency, non-homogeneous Dirichlet conditions)

If $\mathcal{D}$ is a gradient discretisation in the sense of Definition 2.18, define $S_\mathcal{D} : W^{1,p}(\Omega) \to [0, +\infty)$ by

$$\forall \varphi \in W^{1,p}(\Omega),\ S_\mathcal{D}(\varphi) = \min\{\|\Pi_\mathcal{D} v - \varphi\|_{L^p(\Omega)} + \|\nabla_\mathcal{D} v - \nabla\varphi\|_{L^p(\Omega)^d},\ v - \mathcal{I}_{\mathcal{D},\partial}\gamma\varphi \in X_{\mathcal{D},0}\}. \quad (2.14)$$

A sequence $(\mathcal{D}_m)_{m\in\mathbb{N}}$ of gradient discretisations in the sense of Definition 2.18 is **GD-consistent** if

$$\forall \varphi \in W^{1,p}(\Omega),\ \lim_{m\to\infty} S_{\mathcal{D}_m}(\varphi) = 0. \quad (2.15)$$

Since coercivity, limit-conformity and compactness are the same as for homogeneous Dirichlet conditions, the characterisation Lemmas 2.7, 2.14 and 2.16

may also be used in the context of non-homogeneous Dirichlet conditions. 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 controled by this trace.

Lemma 2.21 (Equivalent condition for GD-consistency, non-homogeneous Dirichlet BCs). *Let $(\mathcal{D}_m)_{m\in\mathbb{N}}$ be a sequence of gradient discretisations in the sense of Definition 2.18. We assume that there exists C_1 such that, for any $m \in \mathbb{N}$ and any $\varphi \in W^{1,p}(\Omega)$,*

$$\begin{aligned}\min\{||\Pi_{\mathcal{D}_m} v||_{L^p(\Omega)} + ||\nabla_{\mathcal{D}_m} v||_{L^p(\Omega)^d}\,,\, v - \mathcal{I}_{\mathcal{D}_m,\partial}\gamma\varphi \in X_{\mathcal{D},0}\} \\ \leq C_1||\varphi||_{W^{1,p}(\Omega)}.\end{aligned} \tag{2.16}$$

Then $(\mathcal{D}_m)_{m\in\mathbb{N}}$ is GD-consistent in the sense of Definition 2.20 if and only if there exists a dense subset $\mathcal{R}$ in $W^{1,p}(\Omega)$ such that

$$\forall \psi \in \mathcal{R}\,,\ \lim_{m\to\infty} S_{\mathcal{D}_m}(\psi) = 0. \tag{2.17}$$

Remark 2.22. Note that (2.16) is almost a requirement to GD-consistency in the sense of Definition 2.20. 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 2.20, estimate (2.16) is asymptotically true as $m \to \infty$ since $S_{\mathcal{D}_m}(\varphi) \to 0$.

Proof. The proof is very similar as the proof of Lemma 2.13 and we obviously only have to prove the "if" direction (the "only if" holds with $\mathcal{R} = W^{1,p}(\Omega)$). Let $\varphi \in W^{1,p}(\Omega)$ and $\varepsilon > 0$. Take $\psi \in \mathcal{R}$ 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 (2.16). Then $v + w \in X_{\mathcal{D}_m,0} + \mathcal{I}_{\mathcal{D}_m,\partial}\gamma\varphi$ and, therefore,

$$\begin{aligned}S_{\mathcal{D}_m}(\varphi) &\leq ||\Pi_{\mathcal{D}_m}(v+w) - \varphi||_{L^p(\Omega)} + ||\nabla_{\mathcal{D}_m}(v+w) - \nabla\varphi||_{L^p(\Omega)^d} \\ &\leq ||\Pi_{\mathcal{D}_m} v - \psi||_{L^p(\Omega)} + ||\nabla_{\mathcal{D}_m} v - \nabla\psi||_{L^p(\Omega)^d} \\ &\quad + ||\Pi_{\mathcal{D}_m} w||_{L^p(\Omega)} + ||\nabla_{\mathcal{D}_m} w||_{L^p(\Omega)^d} \\ &\quad + ||\varphi - \psi||_{L^p(\Omega)} + ||\nabla\varphi - \nabla\psi||_{L^p(\Omega)^d} \\ &\leq S_{\mathcal{D}_m}(\psi) + (C_1 + 1)||\varphi - \psi||_{W^{1,p}(\Omega)} \\ &\leq S_{\mathcal{D}_m}(\psi) + (C_1 + 1)\epsilon.\end{aligned}$$

The conclusion follows as in the proof of Lemma 2.13. ∎

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 (3.26) 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 2.23 (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 2.18 which is coercive (Definition 2.2), limit-conforming (Definition 2.6) and GD-consistent (Definition 2.20). 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. 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 2.12 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 Neumann boundary conditions

2.2.1 Homogeneous Neumann conditions

We take here Ω a connected open bounded subset of $\mathbb{R}^d$ with Lipschitz boundary and $p \in (1, +\infty)$.

Definition 2.24 (GD, homogeneous Neumann BCs).
A gradient discretisation $\mathcal{D}$ for homogeneous Neumann conditions is defined by $\mathcal{D} = (X_{\mathcal{D}}, \Pi_{\mathcal{D}}, \nabla_{\mathcal{D}})$ 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.*

The operators $\nabla_{\mathcal{D}}$ *and* $\Pi_{\mathcal{D}}$ *must be chosen such that*

$$\|v\|_{\mathcal{D}} := \left(\|\nabla_{\mathcal{D}} v\|_{L^p(\Omega)^d}^p + \left| \int_\Omega \Pi_{\mathcal{D}} v(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \right|^p \right)^{1/p} \tag{2.18}$$

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

Remark 2.25. 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 the scheme for Neumann problem is written and the *a priori* estimates that can be established on the approximate solution, cf Section 3.2.3.

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

Definition 2.26 (Coercivity, homogeneous Neumann conditions)

If $\mathcal{D}$ is a gradient discretisation in the sense of Definition 2.24, define

$$C_{\mathcal{D}} = \max_{v \in X_{\mathcal{D}} \setminus \{0\}} \frac{\|\Pi_{\mathcal{D}} v\|_{L^p(\Omega)}}{\|v\|_{\mathcal{D}}}. \tag{2.19}$$

A sequence $(\mathcal{D}_m)_{m \in \mathbb{N}}$ of gradient discretisations in the sense of Definition 2.24 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 2.27 (GD-consistency, Neumann conditions)

If $\mathcal{D}$ is a gradient discretisation in the sense of Definition 2.24, define $S_{\mathcal{D}} : W^{1,p}(\Omega) \to [0, +\infty)$ by

$$\begin{aligned} &\forall \varphi \in W^{1,p}(\Omega), \\ &S_{\mathcal{D}}(\varphi) = \min_{v \in X_{\mathcal{D}}} \left(\|\Pi_{\mathcal{D}} v - \varphi\|_{L^p(\Omega)} + \|\nabla_{\mathcal{D}} v - \nabla \varphi\|_{L^p(\Omega)^d} \right). \end{aligned} \tag{2.20}$$

A sequence $(\mathcal{D}_m)_{m \in \mathbb{N}}$ of gradient discretisations in the sense of Definition 2.24 is **GD-consistent** if

$$\forall \varphi \in W^{1,p}(\Omega), \lim_{m \to \infty} S_{\mathcal{D}_m}(\varphi) = 0. \tag{2.21}$$

Definition 2.28 (Limit-conformity, homogeneous Neumann conditions)

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 2.2.3). If $\mathcal{D}$ be a gradient discretisation in the sense of Definition 2.24, define $W_{\mathcal{D}}$: $W^{p'}_{\mathrm{div},0}(\Omega) \to [0, +\infty)$ by

$$\begin{aligned} &\forall \boldsymbol{\varphi} \in W^{p'}_{\mathrm{div},0}(\Omega)\,, \\ &W_{\mathcal{D}}(\boldsymbol{\varphi}) = \max_{v \in X_{\mathcal{D}} \setminus \{0\}} \frac{1}{\|v\|_{\mathcal{D}}} \left| \int_\Omega \nabla_{\mathcal{D}} v(\boldsymbol{x}) \cdot \boldsymbol{\varphi}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \right. \\ &\qquad\qquad \left. + \int_\Omega \Pi_{\mathcal{D}} v(\boldsymbol{x}) \mathrm{div}\boldsymbol{\varphi}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \right|. \end{aligned} \tag{2.22}$$

A sequence $(\mathcal{D}_m)_{m\in\mathbb{N}}$ of gradient discretisations in the sense of Definition 2.24 is **limit-conforming** if

$$\forall \boldsymbol{\varphi} \in W^{p'}_{\mathrm{div},0}(\Omega),\ \lim_{m\to\infty} W_{\mathcal{D}_m}(\boldsymbol{\varphi}) = 0. \tag{2.23}$$

Definition 2.29 (Compactness, homogeneous Neumann conditions)

A sequence $(\mathcal{D}_m)_{m\in\mathbb{N}}$ of gradient discretisations in the sense of Definition 2.24 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 2.10, replacing the space $X_{\mathcal{D},0}$ by $X_{\mathcal{D}}$.

As in the case of Dirichlet boundary conditions (see Lemma 2.13), the GD-consistency (resp. the limit-conformity) of sequences of gradient discretisations in the case of homogeneous Neumann conditions needs only be checked on a dense subset of $W^{1,p}(\Omega)$ (resp. $W_0^{1,p'}(\Omega)$). The proof of this result which is stated in the following lemma is identical to the proof of Lemma 2.13, replacing $W_0^{1,p}(\Omega)$ by $W^{1,p}(\Omega)$.

Lemma 2.30 (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 2.27 if and only if there exists a dense subset $\mathcal{R}$ in $W^{1,p}(\Omega)$ such that

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

We stated in Lemma 2.16 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 prolong the functions by 0 outside of Ω, and therefore the criterion can only involve the "interior" translations

Lemma 2.31 (A criterion for compactness). *Let Ω be an open set 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

$$\begin{array}{l} \lim_{|\boldsymbol{\xi}|\to 0} \sup_{m\in\mathbb{N}} \|u_m(\cdot+\boldsymbol{\xi}) - u_m\|_{L^p(\Omega_{\boldsymbol{\xi}})} = 0 \\ \textit{(where } \Omega_{\boldsymbol{\xi}} = \{\boldsymbol{x}\in\Omega, [\boldsymbol{x},\boldsymbol{x}+\boldsymbol{\xi}]\subset\Omega\}\textit{).} \end{array} \tag{2.24}$$

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 $u \in L^p_{\rm loc}(\Omega)$ and a subsequence, still denoted by $(u_m)_{m\in\mathbb{N}}$, such that $u_m \to u$ in $L^p_{\rm loc}(\Omega)$. 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 (2.24). 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 neighborhood 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

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

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 $U = \cup_{i=1}^k U_i$, neighborhood 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 $\varepsilon \to 0$. ∎

2.2.2 Non-homogeneous Neumann conditions

We present here the framework of gradient discretisations for diffusion problems with non-homogeneous Neumann boundary conditions. We again take Ω a connected open bounded subset of $\mathbb{R}^d$ with Lipschitz boundary and $p \in (1, +\infty)$.

Definition 2.32 (GD, non-homogeneous Neumann BCs). *A gradient discretisation $\mathcal{D}$ for non-homogeneous Neumann 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; 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.*

The operators $\nabla_{\mathcal{D}}$ and $\Pi_{\mathcal{D}}$ must be chosen such that

$$\|v\|_{\mathcal{D}} := \left(\|\nabla_{\mathcal{D}} v\|^p_{L^p(\Omega)^d} + \left| \int_\Omega \Pi_{\mathcal{D}} v(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \right|^p \right)^{1/p} \tag{2.25}$$

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 2.27 and 2.10 (replacing $X_{\mathcal{D},0}$ with $X_{\mathcal{D}}$ in the latter definition).

Definition 2.33 (Coercivity, non-homogeneous Neumann conditions)

If $\mathcal{D}$ is a gradient discretisation in the sense of Definition 2.32, define

$$C_{\mathcal{D}} = \max_{v \in X_{\mathcal{D}} \setminus \{0\}} \left(\max \left\{ \frac{\|\Pi_{\mathcal{D}} v\|_{L^p(\Omega)}}{\|v\|_{\mathcal{D}}}, \frac{\|\mathbb{T}_{\mathcal{D}} v\|_{L^p(\partial\Omega)}}{\|v\|_{\mathcal{D}}} \right\} \right). \tag{2.26}$$

A sequence $(\mathcal{D}_m)_{m\in\mathbb{N}}$ of gradient discretisations in the sense of Definition 2.32 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 2.34 (Limit-conformity, non-homogeneous Neumann conditions)

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

$$W^{p'}_{\mathrm{div},\partial}(\Omega) = \{ \boldsymbol{\varphi} \in L^{p'}(\Omega)^d \,:\, \mathrm{div}\boldsymbol{\varphi} \in L^{p'}(\Omega)\,,\ \gamma_{\mathbf{n}}(\boldsymbol{\varphi}) \in L^{p'}(\partial\Omega) \}, \tag{2.27}$$

where $\gamma_{\mathbf{n}}(\boldsymbol{\varphi})$ is the normal trace of $\boldsymbol{\varphi}$ on $\partial\Omega$ (see Section 2.2.3). If $\mathcal{D}$ is a gradient discretisation in the sense of Definition 2.32, define $W_{\mathcal{D}}$: $W^{p'}_{\mathrm{div},\partial}(\Omega) \to [0, +\infty)$ by

$$\begin{aligned} &\forall \boldsymbol{\varphi} \in W^{p'}_{\mathrm{div},\partial}(\Omega)\,, \\ &W_{\mathcal{D}}(\boldsymbol{\varphi}) = \max_{v \in X_{\mathcal{D}} \setminus \{0\}} \frac{1}{\|v\|_{\mathcal{D}}} \left| \int_\Omega \nabla_{\mathcal{D}} v(\boldsymbol{x}) \cdot \boldsymbol{\varphi}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \right. \\ &\qquad \left. + \int_\Omega \Pi_{\mathcal{D}} v(\boldsymbol{x}) \mathrm{div}\boldsymbol{\varphi}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} - \int_{\partial\Omega} \mathbb{T}_{\mathcal{D}} v(\boldsymbol{x}) \gamma_{\mathbf{n}}(\boldsymbol{\varphi})(\boldsymbol{x}) \mathrm{d}s(\boldsymbol{x}) \right|. \end{aligned} \tag{2.28}$$

A sequence $(\mathcal{D}_m)_{m\in\mathbb{N}}$ of gradient discretisations in the sense of Definition 2.32 is **limit-conforming** if

$$\forall \boldsymbol{\varphi} \in W^{p'}_{\mathrm{div},\partial}(\Omega),\ \lim_{m\to\infty} W_{\mathcal{D}_m}(\boldsymbol{\varphi}) = 0. \tag{2.29}$$

Remark 2.35. This definition of limit-conformity ensure both that the dual operator to $\nabla_{\mathcal{D}_m}$ approximates the continuous divergence operator, and that $\mathbb{T}_{\mathcal{D}_m}$ approximates the continuous trace operator (see also Lemma 2.40).

Definition 2.36 (Compactness, non-homogeneous Neumann conditions)

A sequence $(\mathcal{D}_m)_{m\in\mathbb{N}}$ of gradient discretisations in the sense of Definition 2.32 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 $L^p(\Omega)$ and the sequence $(\mathbb{T}_{\mathcal{D}_m} u_m)_{m\in\mathbb{N}}$ is weakly relatively compact in $L^p(\partial\Omega)$.

Note that the weak relative compactness of the sequence of discrete traces is an immediate consequence of the coercivity of the sequence of gradient discretisations. Moreover, as in the case of Dirichlet boundary conditions, the compactness of a sequence of gradient discretisations implies its coercivity.

Remark 2.37. As for Dirichlet problems, compactness of the gradient discretisation is only useful to deal with non-linearities in the equation. If these non-linearity involve the trace of the solution, then the compactness property should be modified and include also the relative strong compactness of $(\mathbb{T}_{\mathcal{D}_m} u_m)_{m\in\mathbb{N}}$ in $L^p(\partial\Omega)$.

Lemma 2.38 (Equivalent condition for limit-conformity, non-homogeneous Neumann BCs). *Let $(\mathcal{D}_m)$ be a sequence of coercive sequence of gradient discretisations in the sense of Definition 2.33. Then $(\mathcal{D}_m)_{m\in\mathbb{N}}$ is limit-conforming in the sense of Definition 2.34 if and only if there exists a dense subset $\mathcal{S}$ 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*

$$\forall \boldsymbol{\varphi} \in \mathcal{S},\ \lim_{m\to\infty} W_{\mathcal{D}_m}(\boldsymbol{\varphi}) = 0. \tag{2.30}$$

Remark 2.39. Lemma 2.46 shows that the set $\mathcal{S} = C^\infty(\mathbb{R}^d)^d$ is dense in $W^{\mathrm{div},p,\partial}(\Omega)$.

Proof. Let $C_P \in \mathbb{R}_+$ be such that $C_{\mathcal{D}_m} \le C_P$. Assuming that (2.30) holds, we take $\boldsymbol{\varphi} \in W^{p'}_{\mathrm{div},\partial}(\Omega)$ and, for $\varepsilon > 0$, select $\boldsymbol{\varphi} \in \mathcal{S}$ such that $\|\boldsymbol{\varphi}-\boldsymbol{\varphi}\|_{W^{p'}_{\mathrm{div},\partial}(\Omega)} \le \varepsilon$. Then, using the definition of C_P,

$$\begin{aligned}|W_{\mathcal{D}_m}(\boldsymbol{\varphi})| &\le |W_{\mathcal{D}_m}(\boldsymbol{\varphi})| + ||\boldsymbol{\varphi}-\boldsymbol{\varphi}||_{L^p(\Omega)^d} \\ &\quad +(||\mathrm{div}\boldsymbol{\varphi} - \mathrm{div}\boldsymbol{\varphi}||_{L^p(\Omega)} + ||\gamma_{\mathbf{n}}(\mathbf{U}) - \gamma_{\mathbf{n}}(\boldsymbol{\varphi})||_{L^{p'}(\partial\Omega)})C_P \\ &\le |W_{\mathcal{D}_m}(\boldsymbol{\varphi})| + (1+2C_P)\varepsilon.\end{aligned}$$

From (2.30) with $\boldsymbol{\varphi}$ instead of $\boldsymbol{\varphi}$ we then deduce $\limsup_{m\to\infty} |W_{\mathcal{D}_m}(\boldsymbol{\varphi})| \leq (1+2C_P)\varepsilon$ and the proof is complete. ∎

The following lemma is the equivalent of Lemma 2.12 for non-homogeneous Neumann conditions.

Lemma 2.40 (Regularity of the limit, non-homogeneous Neumann BCs). *Let $(\mathcal{D}_m)_{m\in\mathbb{N}}$ be a coercive and limit-conforming sequence of gradient discretisations in the sense of Definitions 2.33 and 2.34. 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,*

$$\Pi_{\mathcal{D}_m} u_m \to u \text{ weakly in } L^p(\Omega), \tag{2.31}$$

$$\mathbb{T}_{\mathcal{D}_m} u_m \to \gamma u \text{ weakly in } L^p(\partial\Omega), \tag{2.32}$$

$$\nabla_{\mathcal{D}_m} u_m \to \nabla u \text{ weakly in } L^p(\Omega)^d. \tag{2.33}$$

Remark 2.41. This lemma shows in particular that 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 2.42. In the case of gradient discretisations for homogeneous Neumann conditions, which do not involve a trace reconstruction, Lemma 2.40 is still valid with (2.32) removed.

Proof. By coercivity, the bound on $\|u_m\|_{\mathcal{D}_m}$ shows that $\|\Pi_{\mathcal{D}_m} u_m\|_{L^p(\Omega)}$, $\|\mathbb{T}_{\mathcal{D}_m} u_m\|_{L^p(\partial\Omega)}$ and $\|\nabla_{\mathcal{D}_m} u_m\|_{L^p(\Omega)^d}$ are bounded. There exists therefore $u \in L^p(\Omega)$, $w \in L^p(\partial\Omega)$ and $\boldsymbol{v} \in L^p(\Omega)^d$ such that, up to a subsequence, $\Pi_{\mathcal{D}_m} u_m \to u$ weakly in $L^p(\Omega)$, $\mathbb{T}_{\mathcal{D}_m} u_m \to w$ weakly in $L^p(\partial\Omega)$ and $\nabla_{\mathcal{D}_m} u_m \to \boldsymbol{v}$ weakly in $L^p(\Omega)^d$.
Let $\boldsymbol{\varphi} \in W^{p'}_{\mathrm{div},\partial}(\Omega)$. By Definition (2.28) of $W_{\mathcal{D}}$,

$$||u_m||_{\mathcal{D}_m} W_{\mathcal{D}_m}(\boldsymbol{\varphi}) \geq \left| \int_\Omega (\nabla_{\mathcal{D}_m} u_m(\boldsymbol{x}) \cdot \boldsymbol{\varphi}(\boldsymbol{x}) + \Pi_{\mathcal{D}_m} u(\boldsymbol{x}) \mathrm{div}\boldsymbol{\varphi}(\boldsymbol{x})) \mathrm{d}\boldsymbol{x} \right.$$
$$\left. - \int_{\partial\Omega} \mathbb{T}_{\mathcal{D}_m} u_m(\boldsymbol{x}) \gamma_{\mathbf{n}}(\boldsymbol{\varphi})(\boldsymbol{x}) \mathrm{d}s(\boldsymbol{x}) \right|. \tag{2.34}$$

By limit-conformity of $(\mathcal{D}_m)_{m\in\mathbb{N}}$ and the bound on $||u_m||_{\mathcal{D}_m}$, the left-hand side of this expression tends to 0 as $m \to \infty$. Hence, passing to the limit we get

$$\int_\Omega (\boldsymbol{v}(\boldsymbol{x}) \cdot \boldsymbol{\varphi}(\boldsymbol{x}) + u(\boldsymbol{x}) \mathrm{div}\boldsymbol{\varphi}(\boldsymbol{x})) \mathrm{d}\boldsymbol{x} - \int_{\partial\Omega} w(\boldsymbol{x}) \gamma_{\mathbf{n}}(\boldsymbol{\varphi})(\boldsymbol{x}) \mathrm{d}s(\boldsymbol{x}) = 0. \tag{2.35}$$

Applied to $\boldsymbol{\varphi} \in C_c^\infty(\Omega)^d$, this relation shows that $\boldsymbol{v} = \nabla u$ and therefore that $u \in W^{1,p}(\Omega)$ and that (2.31) and (2.33) hold.

For any $l \in L^{p'}(\Omega) \subset (W^{1-\frac{1}{p},p}(\partial\Omega))'$, take $\boldsymbol{\varphi} \in W^{p'}_{\mathrm{div}}(\Omega)$ such that $\gamma_{\mathbf{n}}(\boldsymbol{\varphi}) = l$ (see Lemma 2.45). Then $\boldsymbol{\varphi} \in W^{p'}_{\mathrm{div},\partial}(\Omega)$ and therefore, (2.35) and the definition (2.39) of $\gamma_{\mathbf{n}}(\boldsymbol{\varphi}) = l$ (with u instead of $\mathcal{L}_\partial \gamma u$, see the comments after (2.39)) give

$$\int_{\partial\Omega} l(\boldsymbol{x})\gamma u(\boldsymbol{x})\mathrm{d}s(\boldsymbol{x}) = \int_{\partial\Omega} w(\boldsymbol{x})l(\boldsymbol{x})\mathrm{d}s(\boldsymbol{x}),$$

which proves that $w = \gamma u$. Hence (2.32) is verified and the proof is complete. ∎

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 3.12).

Proposition 2.43 (Approximation property of $\mathbb{T}_{\mathcal{D}}$ – Neumann BCs). *Let $\mathcal{D}$ be a gradient discretisation in the sense of Definition 2.32. We define, for $\varphi \in W^{1,p}(\Omega)$,*

$$\overline{S}_{\mathcal{D}}(\varphi) = \min_{w\in X_{\mathcal{D}}} \Big(\|\Pi_{\mathcal{D}} w - \varphi\|_{L^p(\Omega)} + \|\mathbb{T}_{\mathcal{D}} w - \gamma\varphi\|_{L^p(\partial\Omega)} + \|\nabla_{\mathcal{D}} w - \nabla\varphi\|_{L^p(\Omega)^d} \Big). \quad (2.36)$$

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 2.44. The quantity $\overline{S}_{\mathcal{D}}$ in (2.36) is actually the measure of the GD-consistency for Fourier boundary conditions (see (2.49)).

Proof. We introduce

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

and we notice that

$$\|\Pi_{\mathcal{D}} P_{\mathcal{D}}\varphi - \varphi\|_{L^p(\Omega)} + \|\mathbb{T}_{\mathcal{D}} P_{\mathcal{D}}\varphi - \gamma\varphi\|_{L^p(\partial\Omega)} + \|\nabla_{\mathcal{D}} P_{\mathcal{D}}\varphi - \nabla\varphi\|_{L^p(\Omega)^d} \le \overline{S}_{\mathcal{D}}(\varphi). \quad (2.37)$$

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

$$\|\mathbb{T}_{\mathcal{D}} w\|_{L^p(\partial\Omega)} \le C_{\mathcal{D}} \left(\|\nabla_{\mathcal{D}} w\|_{L^p(\Omega)^d} + |\Omega|^{\frac{1}{p'}} \|\Pi_{\mathcal{D}} w\|_{L^p(\Omega)} \right). \quad (2.38)$$

A triangular inequality therefore provides

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

We then use the triangular 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 (2.37) to conclude the proof. ∎

2.2.3 Complements on trace operators

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 $g \in W^{1-\frac{1}{p},p}(\partial\Omega)$. We 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 $\boldsymbol{\varphi} \in W^{p'}_{\rm div}(\Omega)$ (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 $g \in W^{1-\frac{1}{p},p}(\partial\Omega)$,

$$\langle \gamma_{\mathbf{n}}(\boldsymbol{\varphi}), g\rangle_\partial = \int_\Omega (\boldsymbol{\varphi}(\boldsymbol{x}) \cdot \nabla\mathcal{L}_\partial g(\boldsymbol{x}) + \mathcal{L}_\partial g(\boldsymbol{x}) {\rm div}\boldsymbol{\varphi}(\boldsymbol{x}))\, {\rm d}\boldsymbol{x}. \tag{2.39}$$

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,

$$\int_\Omega (\nabla(\mathcal{L}_\partial g - \varphi)(\boldsymbol{x}) \cdot \boldsymbol{\varphi}(\boldsymbol{x}) + (\mathcal{L}_\partial g - \varphi)(\boldsymbol{x}) {\rm div}\boldsymbol{\varphi}(\boldsymbol{x}))\, {\rm d}\boldsymbol{x} = 0.$$

This shows that (2.39) 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 2.45 (Surjectivity of the normal trace). *The normal trace $\gamma_{\mathbf{n}} : W^{p'}_{\rm div}(\Omega) \to (W^{1-\frac{1}{p},p}(\partial\Omega))'$ 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 $\boldsymbol{\varphi} \in W^{p'}_{\rm div}(\Omega)$ satisfying $\gamma_{\mathbf{n}}(\boldsymbol{\varphi}) = l$ and*

$$||\boldsymbol{\varphi}||_{W^{p'}_{\rm div}(\Omega)} \le C_\partial ||l||_{(W^{1-\frac{1}{p},p}(\partial\Omega))'}, \tag{2.40}$$

where we recall that $||\boldsymbol{\varphi}||_{W^{p'}_{\rm div}(\Omega)} = ||\boldsymbol{\varphi}||_{L^{p'}(\Omega)^d} + ||{\rm div}\boldsymbol{\varphi}||_{L^{p'}(\Omega)}$.

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 $\varphi \in W^{1,p}(\Omega)$,

$$\begin{aligned} \langle l, \gamma\varphi\rangle_\partial &= \langle \gamma^* l, \varphi\rangle_{(W^{1,p}(\Omega))', W^{1,p}(\Omega)} \\ &= \int_\Omega (\boldsymbol{\varphi}(\boldsymbol{x}) \cdot \nabla\varphi(\boldsymbol{x}) + h(\boldsymbol{x})\varphi(\boldsymbol{x}))\, \mathrm{d}\boldsymbol{x} \end{aligned} \tag{2.41}$$

and

$$||\boldsymbol{\varphi}||_{L^{p'}(\Omega)^d} + ||h||_{L^{p'}(\Omega)} \leq ||\gamma^* l||_{(W^{1,p}(\Omega))'} \leq C_\partial ||l||_{(W^{1-\frac{1}{p},p}(\partial\Omega))'} \tag{2.42}$$

where C_∂ is the norm of γ (it is also the norm of γ^*). Testing (2.41) with functions φ in $C_c^\infty(\Omega)$ 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 (2.41) with $\varphi = \mathcal{L}_\partial g$ gives $\gamma_{\mathbf{n}}(\boldsymbol{\varphi}) = l$ and Estimate (2.42) gives (2.40). ∎

Lemma 2.46 (Density of smooth functions in $W^{\mathrm{div},p,\partial}(\Omega)$). *Let Ω be a polytopal open set (see Definition 7.2) and let $p \in (1, +\infty)$. The space $W^{\mathrm{div},p,\partial}(\Omega)$ is defined by* (2.27) *and endowed with the norm $\|\boldsymbol{\varphi}\|_{W^{\mathrm{div},p,\partial}(\Omega)} = \|\boldsymbol{\varphi}\|_{L^p(\Omega)^d} + \|\mathrm{div}\boldsymbol{\varphi}\|_{L^p(\Omega)} + \|\gamma_{\mathbf{n}}(\boldsymbol{\varphi})\|_{L^p(\partial\Omega)}$. Then*

1. *$C_c^\infty(\Omega)^d$ is dense in $\{\boldsymbol{\varphi} \in W^{\mathrm{div},p,\partial}(\Omega) : \gamma_{\mathbf{n}}(\boldsymbol{\varphi}) = 0\}$,*
2. *$C_c^\infty(\mathbb{R}^d)^d$ is dense in $W^{\mathrm{div},p,\partial}(\Omega)$.*

Remark 2.47. We only stated the lemma for polytopal open sets Ω, but the proof shows that the result is more general than this (in particular, it holds for 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).

Proof.
Item 1: using the localisation techniques of [68, 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 $\boldsymbol{\varphi} \in W^{\mathrm{div},p,\partial}(\Omega)$ 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 $\mathrm{div}(\widetilde{\boldsymbol{\varphi}}_\lambda * \rho_\epsilon) = \mathrm{div}(\widetilde{\boldsymbol{\varphi}}_\lambda) * \rho_\epsilon \to \mathrm{div}(\widetilde{\boldsymbol{\varphi}}_\lambda)$ 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 $W^{\mathrm{div},p,\partial}(\Omega)$ of $\boldsymbol{\varphi}$ by functions in $C_c^\infty(\Omega)^d$, and the proof of Item 1 is complete.

ITEM 2: let $\boldsymbol{\varphi} \in W^{\mathrm{div},p,\partial}(\Omega)$ 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 neighborhood of S, and such that $\|\gamma_{\mathbf{n}}(\boldsymbol{\varphi}) - \psi_\varepsilon\|_{L^p(\partial\Omega)} \leq \varepsilon$. 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 $\boldsymbol{\varphi} - \boldsymbol{\psi}_\varepsilon \in W^{\mathrm{div},p,\partial}(\Omega)$. We have

$$\|\gamma_{\mathbf{n}}(\boldsymbol{\varphi} - \boldsymbol{\psi}_\varepsilon)\|_{L^p(\partial\Omega)} = \|\gamma_{\mathbf{n}}(\boldsymbol{\varphi}) - \psi_\varepsilon\|_{L^p(\partial\Omega)} \leq \varepsilon. \tag{2.43}$$

By Lemma 2.45 (applied with p and p' swapped) 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

$$\gamma_{\mathbf{n}}(\boldsymbol{\zeta}_\varepsilon) = \gamma_{\mathbf{n}}(\boldsymbol{\varphi} - \boldsymbol{\psi}_\varepsilon) \tag{2.44}$$

and

$$\|\boldsymbol{\zeta}_\varepsilon\|_{W^p_{\mathrm{div}}(\Omega)} \leq C \|\gamma_{\mathbf{n}}(\boldsymbol{\varphi} - \boldsymbol{\psi}_\varepsilon)\|_{L^p(\partial\Omega)} \leq C\varepsilon. \tag{2.45}$$

Property (2.44) shows that $\boldsymbol{\zeta}_\varepsilon \in W^{p'}_{\mathrm{div},\partial}(\Omega)$ and, combined with (2.43) and (2.45), that

$$\|\boldsymbol{\zeta}_\varepsilon\|_{W^{\mathrm{div},p,\partial}(\Omega)} = \|\boldsymbol{\zeta}_\varepsilon\|_{W^p_{\mathrm{div}}(\Omega)} + \|\gamma_{\mathbf{n}}(\boldsymbol{\zeta}_\varepsilon)\|_{L^p(\partial\Omega)} \leq C\varepsilon. \tag{2.46}$$

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

$$\|(\boldsymbol{\varphi} - \boldsymbol{\psi}_\varepsilon - \boldsymbol{\zeta}_\varepsilon) - \boldsymbol{\xi}_\varepsilon\|_{W^{\mathrm{div},p,\partial}(\Omega)} \leq \varepsilon. \tag{2.47}$$

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 (2.46) and (2.47), $\|\boldsymbol{\varphi} - \boldsymbol{\varphi}_\varepsilon\|_{W^{\mathrm{div},p,\partial}(\Omega)} \leq C\varepsilon$. ∎

2.3 Non-homogeneous Fourier boundary conditions

Although we shall present few convergence results for PDE's with Fourier (also known as Robin) boundary conditions, for the sake of completeness we give here the gradient discretisation framework for these conditions. Here Ω

is again a connected open bounded subset of $\mathbb{R}^d$ with Lipschitz boundary and $p \in (1, +\infty)$.
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 2.48 (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.*

The operators $\nabla_{\mathcal{D}}$ and $\mathbb{T}_{\mathcal{D}}$ must be chosen such that

$$\|v\|_{\mathcal{D}} := \left(\|\nabla_{\mathcal{D}} v\|^p_{L^p(\Omega)^d} + \|\mathbb{T}_{\mathcal{D}} v\|^p_{L^p(\partial\Omega)} \right)^{1/p} \tag{2.48}$$

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 2.33, 2.34, 2.36 and 2.10 (with $X_{\mathcal{D},0}$ replaced by $X_{\mathcal{D}}$ in the latter, and using the norm (2.48) in all these definitions). The GD-consistency, however, must take into account the trace reconstruction.

Definition 2.49 (GD-consistency, non-homogeneous Fourier conditions)

If $\mathcal{D}$ is a gradient discretisation in the sense of Definition 2.48, define $S_{\mathcal{D}} : W^{1,p}(\Omega) \to [0, +\infty)$ by

$$\begin{aligned} &\forall \varphi \in W^{1,p}(\Omega), \\ &S_{\mathcal{D}}(\varphi) = \min_{v \in X_{\mathcal{D}}} \big(\|\Pi_{\mathcal{D}} v - \varphi\|_{L^p(\Omega)} + \|\mathbb{T}_{\mathcal{D}} v - \gamma\varphi\|_{L^p(\partial\Omega)} \\ &\qquad\qquad + \|\nabla_{\mathcal{D}} v - \nabla\varphi\|_{L^p(\Omega)^d} \big). \end{aligned} \tag{2.49}$$

A sequence $(\mathcal{D}_m)_{m \in \mathbb{N}}$ of gradient discretisations in the sense of Definition 2.48 is **GD-consistent** if

$$\forall \varphi \in W^{1,p}(\Omega), \lim_{m \to \infty} S_{\mathcal{D}_m}(\varphi) = 0. \tag{2.50}$$

We notice that Lemma 2.30 (characterisation of GD-consistency using a dense set of $W^{1,p}(\Omega)$), Lemma 2.38 (characterisation of limit-conformity using a dense set of $W^{p'}_{\mathrm{div},\partial}(\Omega)$) and Lemma 2.40 (regularity of the limit) still hold in the framework of gradient schemes for non-homogeneous 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 2.43. 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 (2.51). In practice, for many gradient discretisation this assumption is easy to check by using Lemma B.16 and the notion of control by a polytopal toolbox (cf. Section 7.2.3).
The proof of this proposition is identical to the proof of Proposition 2.43, the assumption (2.51) playing the role of (2.38).

Proposition 2.50 (Approximation property of $\mathbb{T}_{\mathcal{D}}$ – Fourier BCs). *Let $\mathcal{D}$ be a gradient discretisation in the sense of Definition 2.48. We assume that there exists $\theta > 0$ such that*

$$\forall v \in X_{\mathcal{D}} \,:\, \|\mathbb{T}_{\mathcal{D}} v\|_{L^p(\partial\Omega)} \le \theta \left(\|\Pi_{\mathcal{D}} v\|_{L^p(\Omega)} + \|\nabla_{\mathcal{D}} v\|_{L^p(\Omega)^d} \right). \tag{2.51}$$

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

$$\|\mathbb{T}_{\mathcal{D}} v - \gamma\varphi\|_{L^p(\partial\Omega)} \le \theta \left(\|\Pi_{\mathcal{D}} v - \varphi\|_{L^p(\Omega)} + \|\nabla_{\mathcal{D}} v - \nabla\varphi\|_{L^p(\Omega)^d} \right) + \max(1,\theta) S_{\mathcal{D}}(\varphi).$$

2.4 Mixed boundary conditions

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

$$\begin{array}{l} \Gamma_d, \Gamma_n \text{ are two disjoint relatively open subsets of } \partial\Omega \\ \text{such that } |\partial\Omega \backslash (\Gamma_d \cup \Gamma_n)| = 0 \text{ and } |\Gamma_d| > 0 \end{array} \tag{2.52}$$

($|\cdot|$ denotes here the $(d-1)$-dimensional measure).

Definition 2.51 (GD, mixed BCs). *Under Assumption* (2.52)*, 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 degrees of freedom in* Ω *and on* Γ_n *and to the degrees of freedom 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.*

The operator $\nabla_{\mathcal{D}}$ *must be chosen such that*

$$\|v\|_{\mathcal{D}} := \|\nabla_{\mathcal{D}} v\|_{L^p(\Omega)^d}$$

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

Definition 2.52 (Coercivity, mixed boundary conditions)

Under Assumption (2.52), if $\mathcal{D}$ is a gradient discretisation in the sense of Definition 2.51, define

$$C_{\mathcal{D}} = \max_{v \in X_{\mathcal{D},\Omega,\Gamma_n} \setminus \{0\}} \left(\max \left\{ \frac{\|\Pi_{\mathcal{D}} v\|_{L^p(\Omega)}}{\|v\|_{\mathcal{D}}}, \frac{\|\mathbb{T}_{\mathcal{D},\Gamma_n} v\|_{L^p(\Gamma_n)}}{\|v\|_{\mathcal{D}}} \right\} \right). \quad (2.53)$$

A sequence $(\mathcal{D}_m)_{m\in\mathbb{N}}$ of gradient discretisations in the sense of Definition 2.51 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 2.53 (GD-consistency, mixed boundary conditions)

Under Assumption (2.52), if $\mathcal{D}$ is a gradient discretisation in the sense of Definition 2.51, define $S_{\mathcal{D}} : W^{1,p}(\Omega) \to [0, +\infty)$ by

$$\forall \varphi \in W^{1,p}(\Omega),\ S_{\mathcal{D}}(\varphi) = \min\{\|\Pi_{\mathcal{D}} v - \varphi\|_{L^p(\Omega)} + \|\nabla_{\mathcal{D}} v - \nabla\varphi\|_{L^p(\Omega)^d}, \\ v - \mathcal{I}_{\mathcal{D},\Gamma_d}\gamma\varphi \in X_{\mathcal{D},\Omega,\Gamma_n}\}. \quad (2.54)$$

A sequence $(\mathcal{D}_m)_{m\in\mathbb{N}}$ of gradient discretisations in the sense of Definition 2.51 is **GD-consistent** if

$$\forall \varphi \in W^{1,p}(\Omega), \lim_{m\to\infty} S_{\mathcal{D}_m}(\varphi) = 0. \quad (2.55)$$

Definition 2.54 (Limit-conformity, mixed boundary conditions)

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

$$W^{\mathrm{div},p',\Gamma_n}(\Omega) = \{\boldsymbol{\varphi} \in L^{p'}(\Omega)^d \,:\, \mathrm{div}\boldsymbol{\varphi} \in L^{p'}(\Omega)\,,\ \gamma_{\mathbf{n}}(\boldsymbol{\varphi}) \in L^{p'}(\Gamma_n)\}, \tag{2.56}$$

where $\gamma_{\mathbf{n}}(\boldsymbol{\varphi})$ is the normal trace of $\boldsymbol{\varphi}$ on $\partial\Omega$. Under Assumption (2.52), if $\mathcal{D}$ is a gradient discretisation in the sense of Definition 2.51, define $W_{\mathcal{D}}$: $W^{\mathrm{div},p',\Gamma_n}(\Omega) \to [0,+\infty)$ by

$$\begin{aligned}&\forall \boldsymbol{\varphi} \in W^{\mathrm{div},p',\Gamma_n}(\Omega),\\ &W_{\mathcal{D}}(\boldsymbol{\varphi}) = \max\left\{ \frac{1}{\|v\|_{\mathcal{D}}} \left| \int_\Omega (\nabla_{\mathcal{D}} v(\boldsymbol{x}) \cdot \boldsymbol{\varphi}(\boldsymbol{x}) + \Pi_{\mathcal{D}} v(\boldsymbol{x}) \mathrm{div}\boldsymbol{\varphi}(\boldsymbol{x}))\,\mathrm{d}\boldsymbol{x} \right.\right.\\ &\qquad \left.\left. - \int_{\Gamma_n} \mathbb{T}_{\mathcal{D},\Gamma_n} v(\boldsymbol{x}) \gamma_{\mathbf{n}}(\boldsymbol{\varphi})(\boldsymbol{x}) \mathrm{d}s(\boldsymbol{x}) \right|,\ v \in X_{\mathcal{D},\Omega,\Gamma_n} \setminus \{0\} \right\}. \end{aligned} \tag{2.57}$$

A sequence $(\mathcal{D}_m)_{m\in\mathbb{N}}$ of gradient discretisations in the sense of Definition 2.51 is **limit-conforming** if

$$\forall \boldsymbol{\varphi} \in W^{\mathrm{div},p',\Gamma_n}(\Omega),\ \lim_{m\to\infty} W_{\mathcal{D}_m}(\boldsymbol{\varphi}) = 0. \tag{2.58}$$

Remark 2.55. 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,

$$\langle \gamma_{\mathbf{n}}(\boldsymbol{\varphi}), w\rangle_{(W^{1-\frac{1}{p},p}(\partial\Omega))',W^{1-\frac{1}{p},p}(\partial\Omega)} = \int_{\Gamma_n} g(\boldsymbol{x}) w(\boldsymbol{x}) \mathrm{d}s(\boldsymbol{x}).$$

Definition 2.56 (Compactness, mixed boundary conditions)

Under Assumption (2.52), a sequence $(\mathcal{D}_m)_{m\in\mathbb{N}}$ of gradient discretisations in the sense of Definition 2.51 is said to be **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 $L^p(\Omega)$ and the sequence $(\mathbb{T}_{\mathcal{D}_m,\Gamma_n} u_m)_{m\in\mathbb{N}}$ is weakly relatively compact in $L^p(\Gamma_n)$.

The definition of **piecewise constant reconstruction** for a gradient discretisation for mixed boundary conditions is the same as Definition 2.10, replacing the space $X_{\mathcal{D},0}$ by $X_{\mathcal{D}}$.

As in the non-homogeneous Neumann case, the weak compactness of the reconstructed traces is an immediate consequence of the coercivity of the sequence of gradient discretisations.

The equivalent of Lemmas 2.12, 2.23 and 2.40 is the following lemma.

Lemma 2.57 (Regularity of the limit, mixed BCs). *Under Assumption (2.52), 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 2.52, 2.53 and 2.54. Let $g \in W^{1-\frac{1}{p},p}(\partial\Omega)$ and, for $m \in \mathbb{N}$, let $u_m \in X_{\mathcal{D}_m}$ 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 $m \to \infty$,*

$$\begin{aligned} &\Pi_{\mathcal{D}_m} u_m \to u \text{ weakly in } L^p(\Omega), \\ &\nabla_{\mathcal{D}_m} u_m \to \nabla u \text{ weakly in } L^p(\Omega)^d. \end{aligned} \tag{2.59}$$

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 $m \to \infty$,

$$\begin{aligned} \min\{&\|\Pi_{\mathcal{D}_m} v - \varphi_g\|_{L^p(\Omega)} + \|\mathbb{T}_{\mathcal{D}_m,\Gamma_n} v - \gamma\varphi_g\|_{L^p(\Gamma_n)} \\ &+ \|\nabla_{\mathcal{D}_m} v - \nabla\varphi_g\|_{L^p(\Omega)^d},\ v - \mathcal{I}_{\mathcal{D}_m,\Gamma_d}\gamma\varphi \in X_{\mathcal{D}_m,\Omega,\Gamma_n}\} \to 0, \end{aligned} \tag{2.60}$$

then we also have

$$\mathbb{T}_{\mathcal{D}_m,\Gamma_n} u_m \to (\gamma u)_{|\Gamma_n} \text{ weakly in } L^p(\Gamma_n). \tag{2.61}$$

Remark 2.58. Assumption (2.60), 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 $u_m = u_m - \mathcal{I}_{\mathcal{D}_m,\Gamma_d} g \in X_{\mathcal{D}_m,\Omega,\Gamma_n}$, we proceed with u_m as in the proof of Lemma 2.40, replacing $W^{p'}_{\mathrm{div},\partial}(\Omega)$ with $W^{\mathrm{div},p',\Gamma_n}(\Omega)$ and, in (2.34), the integrals over $\partial\Omega$ with integrals over Γ_n. This gives $u \in L^p(\Omega)$, $\mathbf{v} \in L^p(\Omega)^d$ and $w \in L^p(\Gamma_n)$ such that , up to a subsequence, $\Pi_{\mathcal{D}_m} u_m \to u$ weakly in $L^p(\Omega)$, $\nabla_{\mathcal{D}_m} u_m \to \mathbf{v}$ weakly in $L^p(\Omega)^d$, $\mathbb{T}_{\mathcal{D}_m,\Gamma_n} u_m \to w$ weakly in $L^p(\Gamma_n)$, and such that the following version of (2.35) holds for all $\boldsymbol{\varphi} \in W^{\mathrm{div},p',\Gamma_n}(\Omega)$:

$$\int_\Omega (\mathbf{v}(\boldsymbol{x}) \cdot \boldsymbol{\varphi}(\boldsymbol{x}) + u(\boldsymbol{x})\mathrm{div}\boldsymbol{\varphi}(\boldsymbol{x}))\mathrm{d}\boldsymbol{x} - \int_{\Gamma_n} w(\boldsymbol{x})\gamma_{\boldsymbol{n}}(\boldsymbol{\varphi})(\boldsymbol{x})\mathrm{d}s(\boldsymbol{x}) = 0.$$

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

$$\int_{\partial\Omega} \gamma_{\mathbf{n}}(\boldsymbol{\varphi})(\boldsymbol{x})\gamma u(\boldsymbol{x})\mathrm{d}s(\boldsymbol{x}) = \int_{\Gamma_n} w(\boldsymbol{x})\gamma_{\mathbf{n}}(\boldsymbol{\varphi})(\boldsymbol{x})\mathrm{d}s(\boldsymbol{x}).$$

This shows that $\gamma u = w$ on Γ_n and that $\gamma u = 0$ on Γ_d, which concludes the proof of (2.59) and (2.61) 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 2.23, 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,

$$\begin{aligned} &\Pi_{\mathcal{D}_m}(u_m - v_m) \to U \text{ weakly in } L^p(\Omega)\,, \\ &\nabla_{\mathcal{D}_m}(u_m - v_m) \to \nabla U \text{ weakly in } L^p(\Omega)^d\,, \\ &\mathbb{T}_{\mathcal{D}_m,\Gamma_n}(u_m - v_m) \to \gamma U \text{ weakly in } L^p(\Gamma_n). \end{aligned} \tag{2.62}$$

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 (2.62) then show that (2.59) holds.

Step 3: we consider a general $g \in W^{1-\frac{1}{p},p}(\partial\Omega)$, and we assume that (2.60) 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 (2.62)) ensures that (2.61) holds. ∎

3

Elliptic problems

The ingredients necessary to implement the gradient discretization method (GDM) for a diffusion problem were introduced in Chapter 2 for various boundary conditions (BCs). They can now be applied to write a gradient scheme (GS) for elliptic and parabolic PDEs; The present chapter is devoted to the study of gradient schemes for the approximation of linear and non-linear elliptic problems. Error estimates are provided in the linear case. In the non-linear case, the convergence is proved thanks to compactness arguments.

3.1 The linear case

In this section, linear problems are considered, so that $p = 2$ is chosen in all the definitions of Chapter 2.

3.1.1 Homogeneous Dirichlet boundary conditions

We consider here the following problem:

$$-\mathrm{div}(\Lambda(\boldsymbol{x})\nabla\overline{u}) = f + \mathrm{div}(\boldsymbol{F}) \text{ in } \Omega, \tag{3.1a}$$

with boundary conditions

$$\overline{u} = 0 \text{ on } \partial\Omega, \tag{3.1b}$$

under the following assumptions:

- Ω is an open bounded connected subset of $\mathbb{R}^d$ ($d \in \mathbb{N}^\star$), (3.2a)
- 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 $[\underline{\lambda}, \overline{\lambda}]$, (3.2b)
- $f \in L^2(\Omega)$, $\boldsymbol{F} \in L^2(\Omega)^d$. (3.2c)

Note that the assumptions on the right hand side include both the case of a right hand side 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 (3.1a), but it is commonly satisfied in applications. Under these hypotheses, the weak solution of (3.1) is the unique function $\overline{u}$ satisfying:

$$\begin{aligned}&\overline{u} \in H^1_0(\Omega),\ \forall v \in H^1_0(\Omega),\\ &\int_\Omega \Lambda(\boldsymbol{x})\nabla\overline{u}(\boldsymbol{x})\cdot\nabla v(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = \int_\Omega f(\boldsymbol{x})v(\boldsymbol{x})\mathrm{d}\boldsymbol{x} - \int_\Omega \boldsymbol{F}(\boldsymbol{x})\cdot\nabla v(\boldsymbol{x})\mathrm{d}\boldsymbol{x}.\end{aligned} \tag{3.3}$$

Let us now introduce an approximation of Problem (3.3) by the GDM.

Definition 3.1 (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 *for Problem* (3.3) *is defined by*

Find $u \in X_{\mathcal{D},0}$ such that for any $v \in X_{\mathcal{D},0}$,

$$\begin{aligned}&\int_\Omega \Lambda(\boldsymbol{x})\nabla_{\mathcal{D}}u(\boldsymbol{x})\cdot\nabla_{\mathcal{D}}v(\boldsymbol{x})\mathrm{d}\boldsymbol{x} =\\ &\qquad\qquad \int_\Omega f(\boldsymbol{x})\Pi_{\mathcal{D}}v(\boldsymbol{x})\mathrm{d}\boldsymbol{x} - \int_\Omega \boldsymbol{F}(\boldsymbol{x})\cdot\nabla_{\mathcal{D}}v(\boldsymbol{x})\mathrm{d}\boldsymbol{x}.\end{aligned} \tag{3.4}$$

Note that, considering a basis $(\xi^{(i)})_{i=1,\dots,N}$ of the space $X_{\mathcal{D},0}$, Scheme (3.4) is equivalent to solving the linear square system $AU = B$, letting

$$\begin{aligned}u &= \sum_{j=1}^N U_j\xi^{(j)},\\ A_{ij} &= \int_\Omega \Lambda(\boldsymbol{x})\nabla_{\mathcal{D}}\xi^{(j)}(\boldsymbol{x})\cdot\nabla_{\mathcal{D}}\xi^{(i)}(\boldsymbol{x})\mathrm{d}\boldsymbol{x},\\ B_i &= \int_\Omega f(\boldsymbol{x})\Pi_{\mathcal{D}}\xi^{(i)}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} - \int_\Omega \boldsymbol{F}(\boldsymbol{x})\cdot\nabla_{\mathcal{D}}\xi^{(i)}(\boldsymbol{x})\mathrm{d}\boldsymbol{x}.\end{aligned} \tag{3.5}$$

The following theorem, first proved in [50] in the case $\boldsymbol{F} = 0$, is in the spirit of the second Strang lemma [67].

Theorem 3.2 (Control of the approximation error, homogeneous Dirichlet BCs). *Under Assumptions* (3.2)*, let $\overline{u} \in H^1_0(\Omega)$ be the solution of Problem* (3.3) *(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* (3.4)*; this solution satisfies the following inequalities:*

$$\|\nabla\overline{u} - \nabla_{\mathcal{D}}u_{\mathcal{D}}\|_{L^2(\Omega)^d} \le \frac{1}{\underline{\lambda}}\left[W_{\mathcal{D}}(\Lambda\nabla\overline{u} + \boldsymbol{F}) + (\overline{\lambda} + \underline{\lambda})S_{\mathcal{D}}(\overline{u})\right], \tag{3.6}$$

$$\|\overline{u} - \Pi_{\mathcal{D}} u_{\mathcal{D}}\|_{L^2(\Omega)} \leq \frac{1}{\underline{\lambda}} \left[C_{\mathcal{D}} W_{\mathcal{D}} (\Lambda \nabla \overline{u} + \boldsymbol{F}) + (C_{\mathcal{D}} \overline{\lambda} + \underline{\lambda}) S_{\mathcal{D}}(\overline{u}) \right], \qquad (3.7)$$

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

Remark 3.3 (Mesh-based GSs). Gradient schemes are often mesh-based. Under usual non-degeneracy assumptions on the meshes, it is proved in many cases that there exists $C \in \mathbb{R}_+$, not depending on $\mathcal{D}$, such that

$$\forall \varphi \in H^2(\Omega) \cap H^1_0(\Omega)\,,\ S_{\mathcal{D}}(\varphi) \leq C h_{\mathcal{D}} ||\varphi||_{H^2(\Omega)}$$

and

$$\forall \boldsymbol{\varphi} \in H^1(\Omega)^d\,,\ W_{\mathcal{D}}(\boldsymbol{\varphi}) \leq C h_{\mathcal{D}} ||\boldsymbol{\varphi}||_{H^1(\Omega)^d},$$

where $h_{\mathcal{D}}$ measures the mesh size. The proofs of these inequalities are done, for some important examples of gradient schemes, in Part III. Then, under the assumptions that the coefficients of Λ are Lipschitz-continuous, that $\boldsymbol{F} \in H^1(\Omega)$ and that $\overline{u} \in H^2(\Omega) \cap H^1_0(\Omega)$, Theorem 3.2 gives in fact $\mathcal{O}(h_{\mathcal{D}})$ error estimates.

Remark 3.4 (Super-convergence)
As noticed in Remark 3.3 above, the L^2 estimate in Theorem 3.2 only provides an $\mathcal{O}(h_{\mathcal{D}})$ 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. This phenomenon is known as super-convergence. It is possible to establish, in the framework of gradient schemes, an improved L^2 estimate that provides such super-convergence results for various schemes, including some for which super-convergence was previously not proved. See [40].

Proof. Let us first prove that, if (3.6) holds for any solution $u_{\mathcal{D}} \in X_{\mathcal{D},0}$ to Scheme (3.4), then the solution to Scheme (3.4) exists and is unique. Indeed, let us prove that, assuming (3.6), the matrix denoted by A of the linear system (3.5) 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 (3.3) is equal to zero a.e. Then from (3.6), 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 (3.5) (as well as (3.4)) 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 (3.4) satisfies (3.6) and (3.7). As noticed in the statement of the theorem, we can take $\boldsymbol{\varphi} = \Lambda \nabla \overline{u} + \boldsymbol{F} \in H_{\rm div}(\Omega)$ in the definition (2.6) of $W_{\mathcal{D}}$. We then obtain, for a given $v \in X_{\mathcal{D},0}$,

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

$$\leq \|v\|_{\mathcal{D}}\ W_{\mathcal{D}}(\Lambda\nabla\overline{u}+\boldsymbol{F}),$$

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

$$\left|\int_\Omega \nabla_{\mathcal{D}}v(\boldsymbol{x})\cdot(\Lambda(\boldsymbol{x})\nabla\overline{u}(\boldsymbol{x})+\boldsymbol{F}(\boldsymbol{x}))-\Pi_{\mathcal{D}}v(\boldsymbol{x})f(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\right| \leq \|v\|_{\mathcal{D}}\ W_{\mathcal{D}}(\Lambda\nabla\overline{u}+\boldsymbol{F}). \quad (3.8)$$

Since $u_{\mathcal{D}}$ is a solution to (3.4), we get

$$\left|\int_\Omega \Lambda\nabla_{\mathcal{D}}v(\boldsymbol{x})\cdot(\nabla\overline{u}(\boldsymbol{x})-\nabla_{\mathcal{D}}u_{\mathcal{D}}(\boldsymbol{x}))\mathrm{d}\boldsymbol{x}\right| \leq \|v\|_{\mathcal{D}}\ W_{\mathcal{D}}(\Lambda\nabla\overline{u}+\boldsymbol{F}). \quad (3.9)$$

We define

$$P_{\mathcal{D}}\overline{u} = \operatorname*{argmin}_{w\in X_{\mathcal{D},0}} \left(\|\Pi_{\mathcal{D}}w-\overline{u}\|_{L^2(\Omega)}+\|\nabla_{\mathcal{D}}w-\nabla\overline{u}\|_{L^2(\Omega)^d}\right) \quad (3.10)$$

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

$$\|\Pi_{\mathcal{D}}\overline{u}-\overline{u}\|_{L^2(\Omega)}+\|\nabla_{\mathcal{D}}P_{\mathcal{D}}\overline{u}-\nabla\overline{u}\|_{L^2(\Omega)^d}=S_{\mathcal{D}}(\overline{u}). \quad (3.11)$$

Recalling the definition of $\|\cdot\|_{\mathcal{D}}$ in Definition 2.1, by (3.9) we get

$$\begin{aligned}
&\left|\int_\Omega \Lambda\nabla_{\mathcal{D}}v(\boldsymbol{x})\cdot(\nabla_{\mathcal{D}}P_{\mathcal{D}}\overline{u}(\boldsymbol{x})-\nabla_{\mathcal{D}}u_{\mathcal{D}}(\boldsymbol{x}))\mathrm{d}\boldsymbol{x}\right| \\
&\quad\leq \|\nabla_{\mathcal{D}}v\|_{L^2(\Omega)^d}\ W_{\mathcal{D}}(\Lambda\nabla\overline{u}+\boldsymbol{F}) \\
&\qquad+\left|\int_\Omega \Lambda(\boldsymbol{x})\nabla_{\mathcal{D}}v(\boldsymbol{x})\cdot(\nabla_{\mathcal{D}}P_{\mathcal{D}}\overline{u}(\boldsymbol{x})-\nabla\overline{u}(\boldsymbol{x}))\mathrm{d}\boldsymbol{x}\right| \\
&\quad\leq \|\nabla_{\mathcal{D}}v\|_{L^2(\Omega)^d}\left(W_{\mathcal{D}}(\Lambda\nabla\overline{u}+\boldsymbol{F})+\overline{\lambda}\,\|\nabla_{\mathcal{D}}P_{\mathcal{D}}\overline{u}-\nabla\overline{u}\|_{L^2(\Omega)^d}\right) \\
&\quad\leq \|\nabla_{\mathcal{D}}v\|_{L^2(\Omega)^d}\left(W_{\mathcal{D}}(\Lambda\nabla\overline{u}+\boldsymbol{F})+\overline{\lambda}S_{\mathcal{D}}(\overline{u})\right).
\end{aligned} \quad (3.12)$$

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

$$\underline{\lambda}\|\nabla_{\mathcal{D}}(P_{\mathcal{D}}\overline{u}-u_{\mathcal{D}})\|_{L^2(\Omega)^d}\leq W_{\mathcal{D}}(\Lambda\nabla\overline{u}+\boldsymbol{F})+\overline{\lambda}S_{\mathcal{D}}(\overline{u}) \quad (3.13)$$

and (3.6) follows by writing

$$\begin{aligned}
&\|\nabla\overline{u}-\nabla_{\mathcal{D}}u_{\mathcal{D}}\|_{L^2(\Omega)^d} \\
&\qquad\leq \|\nabla\overline{u}-\nabla_{\mathcal{D}}P_{\mathcal{D}}\overline{u}\|_{L^2(\Omega)^d}+\|\nabla_{\mathcal{D}}(P_{\mathcal{D}}\overline{u}-u_{\mathcal{D}})\|_{L^2(\Omega)^d} \\
&\qquad\leq S_{\mathcal{D}}(\overline{u})+\frac{1}{\underline{\lambda}}\left(W_{\mathcal{D}}(\Lambda\nabla\overline{u}+\boldsymbol{F})+\overline{\lambda}S_{\mathcal{D}}(\overline{u})\right).
\end{aligned} \quad (3.14)$$

Using (2.1) and (3.13), we get

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

which yields (3.7) by using, as in (3.14), a triangular inequality and the estimate $\|\overline{u} - \Pi_{\mathcal{D}} P_{\mathcal{D}} \overline{u}\|_{L^2(\Omega)^d} \leq S_{\mathcal{D}}(\overline{u})$. ∎

Theorem 3.2 gives a control of the approximation error thanks to the space-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 space-consistent, limit-conforming and coercive, as stated in Corollary 3.5 below. Can this be re-stated as the usual

$$\textit{Consistency} \text{ and } \textit{Stability} \Longrightarrow \textit{Convergence} \tag{3.15}$$

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 discretized 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 $(g(\boldsymbol{x}_i))_{i=1,\ldots,N}$ of g at the discretisation points $(\boldsymbol{x}_i)_{i=1,\ldots,N}$. 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 (3.15) 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.* $\max_{i=1,\ldots,N} |\Pi_h(u) - u_h|(\boldsymbol{x}_i) \to 0$ as $h \to 0$).
In the finite element context, (3.15) no longer holds under these terms, although the spirit remains the same. The reason for the failure of (3.15) 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:

$$u \in V, \tag{3.16}$$
$$a(u, v) = (f, v)\,, \quad \forall v \in V, \tag{3.17}$$

where $V = H_0^1(\Omega)$, $f \in L^2(\Omega)$, and a is a continuous coercive bilinar form on V. Consider a finite element scheme for the discretisation of Problem (3.17), which reads

$$u_h \in V_h, \tag{3.18}$$
$$a_h(u_h, v) = (f, v)\,, \quad \forall v \in V_h, \tag{3.19}$$

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

$$\kappa_h = \max_{v \in V_h \setminus \{0\}} \frac{|a_h(\Pi_h u, v) - (f, v)|}{\|v\|_V}, \tag{3.20}$$

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. [21, Chapter 4] or [44, Chapter 20]. Hence in the FEM context, (3.15) still holds provided

$$\textit{Consistency} = \textit{FEM-consistency} \text{ and } \textit{interpolation error control}. \tag{3.21}$$

For the stability issue in the FEM context, we also refer to [44].
Let us now view a finite element method as a GDM. In this context, the space-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 (3.21). Indeed, in the context of the GDM, the equivalent of the term κ_h defined by (3.20) reads (for $\boldsymbol{F} = 0$)

$$\kappa_{\mathcal{D}} = \underbrace{\frac{\displaystyle\int_\Omega \Lambda(\boldsymbol{x}) \nabla_{\mathcal{D}}(P_{\mathcal{D}} \overline{u})(\boldsymbol{x}) \cdot \nabla_{\mathcal{D}} v(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} - \int_\Omega \Lambda(\boldsymbol{x}) \nabla \overline{u}(\boldsymbol{x}) \cdot \nabla_{\mathcal{D}} v(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}}{\|\nabla_{\mathcal{D}} v\|_{L^2(\Omega)^d}}}_{\text{Controlled by } S_{\mathcal{D}}(\bar{u}), \text{ see (3.12)}}$$
$$+ \underbrace{\frac{\displaystyle\int_\Omega \Lambda(\boldsymbol{x}) \nabla \overline{u}(\boldsymbol{x}) \cdot \nabla_{\mathcal{D}} v(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} - \int_\Omega f(\boldsymbol{x}) \Pi_{\mathcal{D}} v(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}}{\|\nabla_{\mathcal{D}} v\|_{L^2(\Omega)^d}}}_{\text{Controlled by } W_{\mathcal{D}}(\lambda \nabla \bar{u}), \text{ see (3.8)}}.$$

Note that space-consistency and stability (or coercivity) are not sufficient to prove the convergence of a general GDM. The limit-conformity (which is inherent to all conforming finite element methods) 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

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

provided

Consistency = Space-consistency and *Limit-conformity.*

Let us conclude this section by stating the convergence of the GS, which follows easily from Theorem 3.2.

Corollary 3.5 (Convergence). *Under Hypotheses* (3.2)*, let* $(\mathcal{D}_m)_{m\in\mathbb{N}}$ *be a sequence of GDs in the sense of Definition 2.1, which is coercive, space-consistent and limit-conforming in the sense of definitions 2.2, 2.4 and 2.6. Then, for any* $m \in \mathbb{N}$*, there exists a unique solution* $u_m \in X_{\mathcal{D}_m,0}$ *to the gradient scheme* (3.4) *and* $\Pi_{\mathcal{D}_m} u_m$ *converges in* $L^2(\Omega)$ *to the solution* $\overline{u}$ *of* (3.3) *and* $\nabla_{\mathcal{D}_m} u_m$ *converges in* $L^2(\Omega)^d$ *to* $\nabla\overline{u}$ *as* $m \to \infty$*.*

Remark 3.6 (On the compactness assumption). Note that, in the present linear case, the compactness of the sequence of GDs is not required to obtain the convergence ; this compactness assumption will be needed in general for non linear problems (see Remark 3.36 for non linear problems that do not require it).

Remark 3.7 (Space-consistency and limit conformity are necessary conditions)
We state here a kind of reciprocal property to Corollary 3.5. Let us assume that, under Assumptions (3.2a)-(3.2b), 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 (3.4) 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 (3.3) (resp. in $L^2(\Omega)^d$ to $\nabla\overline{u}$). Then $(\mathcal{D}_m)_{m\in\mathbb{N}}$ is space-consistent and limit-conforming in the sense of definitions 2.4 and 2.6.
Indeed, for $\varphi \in H_0^1(\Omega)$, let us consider $f = 0$ and $\mathbf{F} = -\Lambda\nabla\varphi$ in (3.3). 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 (3.3) (resp. converges in $L^2(\Omega)^d$ to $\nabla\varphi$) suffices to prove 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 space-consistent.
For $\boldsymbol{\varphi} \in H_{\mathrm{div}}(\Omega)$, let us set $f = \mathrm{div}\boldsymbol{\varphi}$ and $\boldsymbol{F} = -\boldsymbol{\varphi}$ in (3.3). In this case, the solution $\overline{u}$ is equal to 0 a.e., since the right-hand side of (3.3) vanishes for any $v \in H_0^1(\Omega)$. Since $u_m \in X_{\mathcal{D}_m,0}$ is a solution to the GS (3.4), we get for all $v \in X_{\mathcal{D}_m,0}$,

$$\int_\Omega (\nabla_{\mathcal{D}_m} v(\boldsymbol{x}) \cdot \boldsymbol{\varphi}(\boldsymbol{x}) + \Pi_{\mathcal{D}_m} v(\boldsymbol{x})\mathrm{div}\boldsymbol{\varphi}(\boldsymbol{x}))\,\mathrm{d}\boldsymbol{x} \leq \overline{\lambda}\|\nabla_{\mathcal{D}_m} u_m\|_{L^2(\Omega)^d}\|\nabla_{\mathcal{D}_m} v\|_{L^2(\Omega)^d}.$$

Using that $\nabla_{\mathcal{D}_m} u_m$ converges in $L^2(\Omega)^d$ to 0, we get that

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

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, these hypotheses are sufficient to prove that $\nabla_{\mathcal{D}_m} u_m$ converges 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 (3.3) and $v = u_m$ in (3.4), 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.

3.1.2 Non-homogeneous Dirichlet boundary conditions

As already mentioned in Section 2.1.2, the case of non-homogeneous Dirichlet boundary conditions requires the concept of trace of functions in $H^1(\Omega)$ which demands the Lipschitz regularity of the boundary $\partial\Omega$. We refer to Section 2.2.3 for the properties of the trace operator in this context. We consider the linear problem defined in its strong form by:

$$-\mathrm{div}(\Lambda(\boldsymbol{x})\nabla \overline{u}) = f + \mathrm{div}(\boldsymbol{F}) \text{ in } \Omega, \tag{3.22a}$$

with boundary conditions

$$\overline{u} = g \text{ on } \partial\Omega, \tag{3.22b}$$

under Assumptions (3.2) and

$$\begin{array}{l} \Omega \text{ is an open bounded connected subset of } \mathbb{R}^d \ (d \in \mathbb{N}^\star) \\ \text{ with Lipschitz boundary,} \end{array} \tag{3.23}$$

$$g \in H^{1/2}(\partial\Omega). \tag{3.24}$$

Under these hypotheses, the weak solution of (3.22a) is the unique function $\overline{u}$ satisfying:

$$\begin{array}{l} \overline{u} \in \{w \in H^1(\Omega), \gamma(w) = g\}, \ \forall v \in H^1_0(\Omega), \\ \displaystyle\int_\Omega \Lambda(\boldsymbol{x}) \nabla \overline{u}(\boldsymbol{x}) \cdot \nabla v(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} = \int_\Omega f(\boldsymbol{x}) v(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} - \int_\Omega \boldsymbol{F}(\boldsymbol{x}) \cdot \nabla v(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}. \end{array} \tag{3.25}$$

Under Assumptions (3.2)-(3.24), the GDM applied to Problem (3.25) yields the following gradient scheme.

Definition 3.8 (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 2.18, then we define the related* gradient scheme *for* (3.25) *by*

$$\begin{aligned}&\text{Find } u \in \mathcal{I}_{\mathcal{D},\partial} g + X_{\mathcal{D},0} \text{ such that, for any } v \in X_{\mathcal{D},0},\\ &\int_\Omega \Lambda(\boldsymbol{x}) \nabla_{\mathcal{D}} u(\boldsymbol{x}) \cdot \nabla_{\mathcal{D}} v(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}\\ &= \int_\Omega f(\boldsymbol{x}) \Pi_{\mathcal{D}} v(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} - \int_\Omega \boldsymbol{F}(\boldsymbol{x}) \cdot \nabla_{\mathcal{D}} v(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}.\end{aligned} \tag{3.26}$$

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 3.5.

Theorem 3.9 (Control of the approximation error, non-hom. Dirichlet BCs). *Under Hypotheses* (3.2), (3.23), (3.24), *let* $\overline{u} \in H^1(\Omega)$ *be the solution of* (3.25) *(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 2.18. Then there exists one and only one* $u_{\mathcal{D}} \in X_{\mathcal{D}}$ *solution to the GS* (3.26)*, and it satisfies the following inequalities:*

$$\|\nabla \overline{u} - \nabla_{\mathcal{D}} u_{\mathcal{D}}\|_{L^2(\Omega)^d} \leq \frac{1}{\underline{\lambda}} \left[W_{\mathcal{D}}(\Lambda \nabla \overline{u} + \boldsymbol{F}) + (\overline{\lambda} + \underline{\lambda}) S_{\mathcal{D}}(\overline{u}) \right], \tag{3.27}$$

$$\|\overline{u} - \Pi_{\mathcal{D}} u_{\mathcal{D}}\|_{L^2(\Omega)} \leq \frac{1}{\underline{\lambda}} \left[C_{\mathcal{D}} W_{\mathcal{D}}(\Lambda \nabla \overline{u} + \boldsymbol{F}) + (C_{\mathcal{D}} \overline{\lambda} + \underline{\lambda}) S_{\mathcal{D}}(\overline{u}) \right], \tag{3.28}$$

where $C_{\mathcal{D}}$*,* $S_{\mathcal{D}}$ *and* $W_{\mathcal{D}}$ *are defined by Definitions 2.2, 2.20 and 2.6.*

Proof. Reasoning as in the proof of Theorem 3.2, we arrive at (3.9) for any $v \in X_{\mathcal{D},0}$. We then define

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

and we notice that, by definition (2.14) of $S_{\mathcal{D}}$, (3.11) is still valid. Moreover, the vector $v = P_{\mathcal{D}} \overline{u} - u_{\mathcal{D}}$ 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 (3.9). The rest of the proof is then exactly as in the proof of Theorem 3.2. ∎

3.1.3 Neumann boundary conditions

We consider here a linear elliptic problem with non-homogeneous Neumann boundary conditions

$$\begin{aligned}&-\mathrm{div}(\Lambda(\boldsymbol{x}) \nabla \overline{u}) = f + \mathrm{div}(\boldsymbol{F}) \text{ in } \Omega,\\ &\Lambda \nabla \overline{u} \cdot \mathbf{n} + \boldsymbol{F} \cdot \mathbf{n} = h \text{ on } \partial\Omega,\end{aligned} \tag{3.29}$$

where $\mathbf{n}$ is the unit normal outward Ω to $\partial\Omega$, assumed to be Lipschitz, under Assumptions (3.2), (3.23) and

$$h \in L^2(\partial\Omega) \text{ and } \int_\Omega f(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} + \int_{\partial\Omega} h(\boldsymbol{x}) \mathrm{d}s(\boldsymbol{x}) = 0. \tag{3.30}$$

Under these hypotheses and defining

$$H^1_\star(\Omega) = \{\varphi \in H^1(\Omega)\,,\ \int_\Omega \varphi(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = 0\},$$

the weak formulation of (3.29) is

$$\begin{aligned}
&\overline{u} \in H^1_\star(\Omega),\ \forall v \in H^1_\star(\Omega),\\
&\int_\Omega \Lambda(\boldsymbol{x})\nabla\overline{u}(\boldsymbol{x})\cdot\nabla v(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\\
&= \int_\Omega f(\boldsymbol{x})v(\boldsymbol{x})\mathrm{d}\boldsymbol{x} - \int_\Omega \boldsymbol{F}(\boldsymbol{x})\cdot\nabla v(\boldsymbol{x})\mathrm{d}\boldsymbol{x} + \int_{\partial\Omega} h(\boldsymbol{x})\gamma(v)(\boldsymbol{x})\mathrm{d}s(\boldsymbol{x}).
\end{aligned} \tag{3.31}$$

We recall that, owing to Hypothesis (3.30), Problem (3.31) is equivalent to

$$\begin{aligned}
&\overline{u} \in H^1(\Omega),\ \forall v \in H^1(\Omega),\\
&\int_\Omega \Lambda(\boldsymbol{x})\nabla\overline{u}(\boldsymbol{x})\cdot\nabla v(\boldsymbol{x})\mathrm{d}\boldsymbol{x} + \int_\Omega \overline{u}(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\int_\Omega v(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\\
&= \int_\Omega f(\boldsymbol{x})v(\boldsymbol{x})\mathrm{d}\boldsymbol{x} - \int_\Omega \boldsymbol{F}(\boldsymbol{x})\cdot\nabla v(\boldsymbol{x})\mathrm{d}\boldsymbol{x} + \int_{\partial\Omega} h(\boldsymbol{x})\gamma(v)(\boldsymbol{x})\mathrm{d}s(\boldsymbol{x}),
\end{aligned} \tag{3.32}$$

since 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 in the next definition.

Definition 3.10 (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 2.32, then we define the related* gradient scheme *for* (3.31) *by*

$$\begin{aligned}
&\text{Find } u \in X_\mathcal{D} \text{ such that, for any } v \in X_\mathcal{D},\\
&\int_\Omega \Lambda(\boldsymbol{x})\nabla_\mathcal{D}u(\boldsymbol{x})\cdot\nabla_\mathcal{D}v(\boldsymbol{x})\mathrm{d}\boldsymbol{x} + \int_\Omega \Pi_\mathcal{D}u(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\ \int_\Omega \Pi_\mathcal{D}v(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\\
&= \int_\Omega f(\boldsymbol{x})\Pi_\mathcal{D}v(\boldsymbol{x})\mathrm{d}\boldsymbol{x} - \int_\Omega \boldsymbol{F}(\boldsymbol{x})\cdot\nabla_\mathcal{D}v(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\\
&\quad + \int_{\partial\Omega} h(\boldsymbol{x})\mathbb{T}_\mathcal{D}v(\boldsymbol{x})\mathrm{d}s(\boldsymbol{x}).
\end{aligned} \tag{3.33}$$

The error estimates for Neumann boundary conditions are stated in Theorem 3.11. We do not explicitly the convergence result, similar to Corollary 3.5, that stems from these error estimates.

Theorem 3.11 (Control of the approximation error, Neumann BCs). *Under Hypotheses* (3.2), (3.23) *and* (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 $\Lambda\nabla\overline{u} + \boldsymbol{F} \in W^{\mathrm{div},2,\partial}(\Omega)$, see* (2.27)*).*
Let $\mathcal{D}$ be a GD for a Neumann problem in the sense of Definition 2.32.
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:*

$$\|\nabla\overline{u} - \nabla_{\mathcal{D}} u_{\mathcal{D}}\|_{L^2(\Omega)^d} \leq \mathrm{Err} + S_{\mathcal{D}}(\overline{u}), \tag{3.34}$$

$$\|\overline{u} - \Pi_{\mathcal{D}} u_{\mathcal{D}}\|_{L^2(\Omega)} \leq C_{\mathcal{D}}\mathrm{Err} + S_{\mathcal{D}}(\overline{u}), \tag{3.35}$$

$$\text{with } \mathrm{Err} := \frac{1}{\min(\underline{\lambda}, 1)} \left[W_{\mathcal{D}}(\Lambda\nabla\overline{u} + \boldsymbol{F}) + (\overline{\lambda} + |\Omega|^{1/2} C_{\mathcal{D}}) S_{\mathcal{D}}(\overline{u}) \right],$$

where $C_{\mathcal{D}}$, $S_{\mathcal{D}}$ and $W_{\mathcal{D}}$ are defined by (2.26), (2.20) *and* (2.28).

Remark 3.12 (Error estimate on the traces). If we let $\overline{S}_{\mathcal{D}}$ be the measure of space-consistency for Fourier boundary conditions (*i.e.* (2.36)), then Proposition 2.43 and Theorem 3.11 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 only depends on $\underline{\lambda}$, $\overline{\lambda}$, $|\Omega|$ and an upper bound of $C_{\mathcal{D}}$.

Proof. Recall that in Definition 2.32, it is assumed that

$$\|v\|_{\mathcal{D}} := \left(\|\nabla_{\mathcal{D}} v\|^2_{L^2(\Omega)^d} + \left| \int_\Omega \Pi_{\mathcal{D}} v(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \right|^2 \right)^{1/2}$$

defines a norm on $X_{\mathcal{D}}$. Therefore, proving (3.34) for any solution $u_{\mathcal{D}} \in X_{\mathcal{D}}$ to Scheme (3.33) is enough to prove the existence and uniqueness of this solution (because this estimate shows that, whenever $f = 0$, $h = 0$ and $\boldsymbol{F} = 0$, the only possible solution to the scheme is $u_{\mathcal{D}} = 0$).
To prove the estimates, we take $\boldsymbol{\varphi} = \Lambda\nabla\overline{u} + \boldsymbol{F} \in W^{\mathrm{div},2,\partial}(\Omega)$ in the definition (2.28) of $W_{\mathcal{D}}$ and using that $\overline{u}$ is the solution to (3.31). We then have, for any $v \in X_{\mathcal{D}}$,

$$\left| \int_\Omega \left[\nabla_{\mathcal{D}} v(\boldsymbol{x}) \cdot (\Lambda(\boldsymbol{x})\nabla\overline{u}(\boldsymbol{x}) + \boldsymbol{F}(\boldsymbol{x})) - \Pi_{\mathcal{D}} v(\boldsymbol{x}) f(\boldsymbol{x}) \right] \mathrm{d}\boldsymbol{x} \right.$$
$$\left. - \int_{\partial\Omega} h(\boldsymbol{x}) \mathbb{T}_{\mathcal{D}} v(\boldsymbol{x}) \mathrm{d}s(\boldsymbol{x}) \right| \leq \|v\|_{\mathcal{D}}\ W_{\mathcal{D}}(\Lambda\nabla\overline{u} + \boldsymbol{F}).$$

Therefore, since $u_{\mathcal{D}}$ is a solution to (3.33), we get

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

We then introduce

$$P_{\mathcal{D}}\overline{u} = \operatorname*{argmin}_{w \in X_{\mathcal{D}}} \left(\|\Pi_{\mathcal{D}} w - \overline{u}\|_{L^2(\Omega)} + \|\nabla_{\mathcal{D}} w - \nabla\overline{u}\|_{L^2(\Omega)^d} \right) \tag{3.36}$$

in the above inequality. It leads to

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

Observing that

$$\left| \int_\Omega \Pi_{\mathcal{D}} P_{\mathcal{D}} \overline{u}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \right| = \left| \int_\Omega (\Pi_{\mathcal{D}} P_{\mathcal{D}} \overline{u}(\boldsymbol{x}) - \overline{u}(\boldsymbol{x})) \mathrm{d}\boldsymbol{x} \right|$$
$$\leq |\Omega|^{1/2} \left\| \Pi_{\mathcal{D}} P_{\mathcal{D}} \overline{u} - \overline{u} \right\|_{L^2(\Omega)},$$

we can write, using the definition (2.26) of $C_{\mathcal{D}}$,

$$\left| \int_\Omega \Lambda \nabla_{\mathcal{D}} v(\boldsymbol{x}) \cdot (\nabla_{\mathcal{D}} P_{\mathcal{D}} \overline{u}(\boldsymbol{x}) - \nabla_{\mathcal{D}} u_{\mathcal{D}}(\boldsymbol{x})) \mathrm{d}\boldsymbol{x} \right.$$
$$\left. + \int_\Omega (\Pi_{\mathcal{D}} P_{\mathcal{D}} \overline{u}(\boldsymbol{x}) - \Pi_{\mathcal{D}} u_{\mathcal{D}}(\boldsymbol{x})) \mathrm{d}\boldsymbol{x} \int_\Omega \Pi_{\mathcal{D}} v(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \right|$$
$$\leq \|v\|_{\mathcal{D}} \left[W_{\mathcal{D}}(\Lambda \nabla \overline{u} + \boldsymbol{F}) + (\overline{\lambda} + |\Omega|^{1/2} C_{\mathcal{D}}) S_{\mathcal{D}}(\overline{u}) \right].$$

We now let $v = P_{\mathcal{D}} \overline{u} - u_{\mathcal{D}}$. Recalling the definition (2.25) of $\|\cdot\|_{\mathcal{D}}$, we obtain

$$\|P_{\mathcal{D}} \overline{u} - u_{\mathcal{D}}\|_{\mathcal{D}} \leq \frac{W_{\mathcal{D}}(\Lambda \nabla \overline{u} + \boldsymbol{F}) + (\overline{\lambda} + |\Omega|^{1/2} C_{\mathcal{D}}) S_{\mathcal{D}}(\overline{u})}{\min(\underline{\lambda}, 1)}.$$

The conclusion follows as in the proof of Theorem 3.2. ∎

3.1.4 Mixed boundary conditions

To conclude the case of linear problem, we consider here a linear elliptic problem with mixed boundary conditions

$$\begin{aligned} &-\mathrm{div}(\Lambda(\boldsymbol{x}) \nabla \overline{u}) = f + \mathrm{div}(\boldsymbol{F}) \text{ in } \Omega, \\ &\overline{u} = g \text{ on } \Gamma_d, \\ &\Lambda \nabla \overline{u} \cdot \mathbf{n} + \boldsymbol{F} \cdot \mathbf{n} = h \text{ on } \Gamma_n, \end{aligned} \tag{3.37}$$

under Assumptions (3.2), (3.23), (2.52) and

$$g \in H^{1/2}(\partial\Omega)\,,\ h \in L^2(\Gamma_n). \tag{3.38}$$

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 (3.37) is

$$\begin{aligned} &\overline{u} \in \{w \in H^1(\Omega) \,:\, \gamma(w) = g \text{ on } \Gamma_d\},\ \forall v \in H^1_{\Gamma_d}(\Omega),\\ &\int_\Omega \Lambda(\boldsymbol{x})\nabla\overline{u}(\boldsymbol{x})\cdot\nabla v(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = \int_\Omega f(\boldsymbol{x})v(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\\ &\qquad -\int_\Omega \boldsymbol{F}(\boldsymbol{x})\cdot\nabla v(\boldsymbol{x})\mathrm{d}\boldsymbol{x} + \int_{\Gamma_n} h(\boldsymbol{x})\gamma(v)(\boldsymbol{x})\mathrm{d}s(\boldsymbol{x}). \end{aligned} \tag{3.39}$$

The GDM applied to this mixed problem yields the following scheme.

Definition 3.13 (GS, mixed linear problem).
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 2.51, then the related gradient scheme *for* (3.39) *is defined by:*

$$\begin{aligned} &\text{Find } u \in \mathcal{I}_{\mathcal{D},\Gamma_d} g + X_{\mathcal{D},\Omega,\Gamma_n} \text{ such that, for any } v \in X_{\mathcal{D},\Omega,\Gamma_n},\\ &\int_\Omega \Lambda(\boldsymbol{x})\nabla_\mathcal{D} u(\boldsymbol{x})\cdot\nabla_\mathcal{D} v(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = \int_\Omega f(\boldsymbol{x})\Pi_\mathcal{D} v(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\\ &\qquad -\int_\Omega \boldsymbol{F}(\boldsymbol{x})\cdot\nabla_\mathcal{D} v(\boldsymbol{x})\mathrm{d}\boldsymbol{x} + \int_{\Gamma_n} h(\boldsymbol{x})\mathbb{T}_{\mathcal{D},\Gamma_n} v(\boldsymbol{x})\mathrm{d}s(\boldsymbol{x}). \end{aligned} \tag{3.40}$$

The proof of the following error estimates for mixed boundary conditions is similar to the proofs made in the case of other boundary conditions. Likewise, a convergence result similar to Corollary 3.5 follows from these error estimates.

Theorem 3.14 (Control of the approximation error). *Under Hypotheses* (3.2), (3.23), (2.52) *and* (3.38), *let $\overline{u} \in H^1(\Omega)$ be the solution of* (3.39) *(remark that since $f \in L^2(\Omega)$ and $h \in L^2(\Gamma_n)$, we have $\Lambda\nabla\overline{u}+\boldsymbol{F} \in W^{\mathrm{div},2,\Gamma_n}(\Omega)$, see* (2.56)*).*
Let $\mathcal{D}$ be a GD for mixed boundary conditions in the sense of Definition 2.51. Then there exists one and only one $u_\mathcal{D} \in X_\mathcal{D}$ solution to the gradient scheme (3.40), *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 (2.53), (2.54) *and* (2.57).

3.2 Unknown-dependent diffusion problems

We shall consider here 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

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

operator, we remain in the functional framework of the linear case, considering again that $p = 2$.

3.2.1 Homogeneous Dirichlet boundary conditions

We consider the following problem:

$$-\mathrm{div}(\Lambda(\boldsymbol{x}, \overline{u}(\boldsymbol{x}))\nabla\overline{u}) = f \text{ in } \Omega, \tag{3.41a}$$

with boundary conditions

$$\overline{u} = 0 \text{ on } \partial\Omega, \tag{3.41b}$$

under the following assumptions:

- Ω is an open bounded connected subset of $\mathbb{R}^d$, $d \in \mathbb{N}^\star$, (3.42a)
- Λ is a Caratheodory function from $\Omega \times \mathbb{R}$ to $\mathcal{M}_d(\mathbb{R})$,
  $\Lambda(\boldsymbol{x}, s)$ 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 $[\underline{\lambda}, \overline{\lambda}]$, (3.42b)
- $f \in L^2(\Omega)$. (3.42c)

Under these hypotheses, a weak solution of (3.41a) is a function $\overline{u}$ (not necessarily unique) satisfying:

$$\begin{aligned}&\overline{u} \in H^1_0(\Omega),\ \forall v \in H^1_0(\Omega),\\ &\int_\Omega \Lambda(\boldsymbol{x}, \overline{u}(\boldsymbol{x}))\nabla\overline{u}(\boldsymbol{x}) \cdot \nabla v(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = \int_\Omega f(\boldsymbol{x})v(\boldsymbol{x})\mathrm{d}\boldsymbol{x}.\end{aligned} \tag{3.43}$$

Then Problem (3.43) is approximated under Assumptions (3.42) by the following gradient scheme.

Definition 3.15 (Gradient scheme, unknown-dependent diffusion). *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* (3.3) *by*

$$\begin{aligned}&\text{Find } u \in X_{\mathcal{D},0} \text{ such that, for any } v \in X_{\mathcal{D},0},\\ &\int_\Omega \Lambda(\boldsymbol{x}, \Pi_\mathcal{D}u(\boldsymbol{x}))\nabla_\mathcal{D}u(\boldsymbol{x}) \cdot \nabla_\mathcal{D}v(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = \int_\Omega f(\boldsymbol{x})\Pi_\mathcal{D}v(\boldsymbol{x})\mathrm{d}\boldsymbol{x}.\end{aligned} \tag{3.44}$$

Note that, considering a basis $(\xi^{(i)})_{i=1,\dots,N}$ of the space $X_{\mathcal{D},0}$, Scheme (3.44) is equivalent to solving the system of N non-linear equations with N unknowns $A(u)U = B$ with

$$u = \sum_{j=1}^N U_j \xi^{(j)},$$

$$A_{ij}(u) = \int_\Omega \Lambda(\boldsymbol{x}, \Pi_\mathcal{D} u(\boldsymbol{x})) \nabla_\mathcal{D} \xi^{(j)}(\boldsymbol{x}) \cdot \nabla_\mathcal{D} \xi^{(i)}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}, \tag{3.45}$$
$$B_i = \int_\Omega f(\boldsymbol{x}) \Pi_\mathcal{D} \xi^{(i)}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}.$$

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-Raphson method.

Let us now state a convergence result (note that an error estimate between an approximate solution and a weak solution to (3.44) cannot be stated, since the uniqueness of the solution to neither (3.45) nor (3.44) is known in the general case).

Theorem 3.16 (Convergence, unknown-dependent diffusion). *Under assumptions* (3.42)*, take a sequence* $(\mathcal{D}_m)_{m\in\mathbb{N}}$ *of GDs in the sense of Definition 2.1, which is space-consistent, limit-conforming and compact in the sense of Definitions 2.4, 2.6 and 2.8 (it is then coercive in the sense of Definition 2.2, see Lemma 2.9).*
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* (3.44) *and, up to a subsequence,* $\Pi_{\mathcal{D}_m} u_m$ *converges strongly in* $L^2(\Omega)$ *to a solution* $\overline{u}$ *of* (3.43) *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* (3.43) *is unique, then the whole sequence converges to* $\overline{u}$ *as* $m \to \infty$ *in the senses above.*

Proof.
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

$$\begin{aligned}&\text{Find } u \in X_{\mathcal{D},0} \text{ such that, } \forall v \in X_{\mathcal{D},0},\\ &\int_\Omega \Lambda(\boldsymbol{x}, \Pi_\mathcal{D} w(\boldsymbol{x})) \nabla_\mathcal{D} u(\boldsymbol{x}) \cdot \nabla_\mathcal{D} v(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} = \int_\Omega f(\boldsymbol{x}) \Pi_\mathcal{D} v(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}.\end{aligned} \tag{3.46}$$

Therefore, u is solution to the square linear system $A(w)U = B$, where $A(w)$ and B are defined in (3.45). Let us prove that the matrix $A(w)$ is invertible. Letting $v = u$ in (3.46), and applying the Cauchy-Schwarz inequality and Hypothesis (3.42b), 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

$$\|\nabla_\mathcal{D} u\|_{L^2(\Omega)^d} \le \frac{C_\mathcal{D}}{\underline{\lambda}} \|f\|_{L^2(\Omega)}. \tag{3.47}$$

This completes the proof that $A(w)$ is invertible, since (3.47) 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 (3.47) 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 (Theorem C.2) proves that the equation $F(U) = U$ has at least one solution. This proves the existence of at least one discrete solution to (3.44).
We note that the previous estimates easily show that any solution to this scheme satisfies (3.47).

Step 2: convergence of $\Pi_{\mathcal{D}_m} u_m$ and $\nabla_{\mathcal{D}_m} u_m$.
Thanks to the coercivity hypothesis and (3.47), we have

$$\|\nabla_{\mathcal{D}_m} u_m\|_{L^2(\Omega)^d} \leq \frac{C_P}{\underline{\lambda}} \|f\|_{L^2(\Omega)}. \tag{3.48}$$

We may then apply Lemma 2.12, 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 $\overline{u} \in H^1_0(\Omega)$ 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, there exists again a subsequence of the preceding 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 (3.43).
This proof is done by passing to the limit in the gradient scheme (3.44), considering as test function the following interpolation of a given function $\varphi \in H^1_0(\Omega)$.
Let us define, for a given GD $\mathcal{D}$, $P_{\mathcal{D}} : H^1_0(\Omega) \to X_{\mathcal{D},0}$ by

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

We have

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

and therefore, by space-consistency of the sequence $(\mathcal{D}_m)_{m\in\mathbb{N}}$, $\Pi_{\mathcal{D}_m}(P_{\mathcal{D}_m}\varphi) \to \varphi$ strongly in $L^2(\Omega)$ and $\nabla_{\mathcal{D}_m}(P_{\mathcal{D}_m}\varphi) \to \nabla\varphi$ strongly in $L^2(\Omega)^d$.
Using Lemma C.4 page 404 (non-linear strong convergence), we infer that $\Lambda(\cdot, \Pi_{\mathcal{D}_m} u_m)\nabla_{\mathcal{D}_m}(P_{\mathcal{D}_m}\varphi) \to \Lambda(\cdot,\overline{u})\nabla\varphi$ strongly in $L^2(\Omega)^d$. By symmetry of Λ and the weak-strong convergence property (Lemma C.3), this shows that

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

$$= \int_\Omega \Lambda(\boldsymbol{x}, \overline{u}(\boldsymbol{x}))\nabla\overline{u}(\boldsymbol{x}) \cdot \nabla\varphi(\boldsymbol{x})\mathrm{d}\boldsymbol{x}. \tag{3.49}$$

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

$$\int_\Omega f(\boldsymbol{x})\Pi_{\mathcal{D}_m}(P_{\mathcal{D}_m}\varphi)(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \to \int_\Omega f(\boldsymbol{x})\varphi(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \quad \text{as } m \to \infty. \tag{3.50}$$

Letting $v = P_{\mathcal{D}_m}\varphi$ in (3.44), we can use (3.49) and (3.50) to pass to the limit and see that $\overline{u}$ is a solution to (3.43).

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 (3.44) and we pass to the limit in the right-hand side. Since $\overline{u}$ is a solution to (3.43), we obtain

$$\begin{aligned}\lim_{m\to\infty} \int_\Omega \Lambda(\boldsymbol{x},&\Pi_{\mathcal{D}_m} u_m(\boldsymbol{x}))\nabla_{\mathcal{D}_m} u_m(\boldsymbol{x}) \cdot \nabla_{\mathcal{D}_m} u_m(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \\ &= \int_\Omega f(\boldsymbol{x})\overline{u}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = \int_\Omega \Lambda(\boldsymbol{x}, \overline{u}(\boldsymbol{x}))\nabla\overline{u}(\boldsymbol{x}) \cdot \nabla\overline{u}(\boldsymbol{x})\mathrm{d}\boldsymbol{x}.\end{aligned} \tag{3.51}$$

We have

$$\begin{aligned}\int_\Omega \Lambda(\boldsymbol{x},&\Pi_{\mathcal{D}_m} u_m(\boldsymbol{x}))(\nabla_{\mathcal{D}_m} u_m(\boldsymbol{x}) - \nabla\overline{u}(\boldsymbol{x})) \cdot (\nabla_{\mathcal{D}_m} u_m(\boldsymbol{x}) - \nabla\overline{u}(\boldsymbol{x}))\mathrm{d}\boldsymbol{x} \\ &= \int_\Omega \Lambda(\boldsymbol{x}, \Pi_{\mathcal{D}_m} u_m(\boldsymbol{x}))\nabla_{\mathcal{D}_m} u_m(\boldsymbol{x}) \cdot \nabla_{\mathcal{D}_m} u_m(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \\ &\quad - \int_\Omega \Lambda(\boldsymbol{x}, \Pi_{\mathcal{D}_m} u_m(\boldsymbol{x}))\nabla_{\mathcal{D}_m} u_m(\boldsymbol{x}) \cdot \nabla\overline{u}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \\ &\quad - \int_\Omega \Lambda(\boldsymbol{x}, \Pi_{\mathcal{D}_m} u_m(\boldsymbol{x}))\nabla\overline{u}(\boldsymbol{x}) \cdot (\nabla_{\mathcal{D}_m} u_m(\boldsymbol{x}) - \nabla\overline{u}(\boldsymbol{x}))\mathrm{d}\boldsymbol{x}.\end{aligned} \tag{3.52}$$

By (3.51), 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 C.4 page 404), and the weak-strong convergence lemma (Lemma C.3), we infer that

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

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. ∎

3.2.2 Non-homogeneous Dirichlet boundary conditions

We again refer to Section 2.2.3 for the properties of the trace operator in the case of domain Ω with Lipschitz boundary, and we consider Problem (3.41a), replacing the homogeneous Dirichlet boundary condition by

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

under Assumptions (3.42) and

$$g \in H^{1/2}(\partial\Omega). \tag{3.53}$$

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

$$\begin{aligned} &\overline{u} \in \{w \in H^1(\Omega), \gamma(w) = g\},\ \forall v \in H^1_0(\Omega), \\ &\int_\Omega \Lambda(\boldsymbol{x}, \overline{u}(\boldsymbol{x}))\nabla\overline{u}(\boldsymbol{x}) \cdot \nabla v(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = \int_\Omega f(\boldsymbol{x})v(\boldsymbol{x})\mathrm{d}\boldsymbol{x}, \end{aligned} \tag{3.54}$$

and it is approximated by the following gradient scheme.

Definition 3.17 (Non-linear problem, 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 2.18, then we define the related gradient scheme *for* (3.54) *by*

$$\begin{aligned} &\text{Find } u \in \mathcal{I}_{\mathcal{D},\partial}g + X_{\mathcal{D},0} \text{ such that, for any } v \in X_{\mathcal{D},0}, \\ &\int_\Omega \Lambda(\boldsymbol{x}, \Pi_\mathcal{D}u(\boldsymbol{x}))\nabla_\mathcal{D}u(\boldsymbol{x}) \cdot \nabla_\mathcal{D}v(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = \int_\Omega f(\boldsymbol{x})\Pi_\mathcal{D}v(\boldsymbol{x})\mathrm{d}\boldsymbol{x}. \end{aligned} \tag{3.55}$$

This scheme is again leading to a nonlinear system of equations under the form $A(u)U = B$, similar to (3.45). We then have the following convergence result.

Theorem 3.18 (Convergence).
Under assumptions (3.42)-(3.53)*, let $(\mathcal{D}_m)_{m\in\mathbb{N}}$ be a sequence of GDs in the sense of Definition 2.18, which is space-consistent, limit-conforming and compact in the sense of Definitions 2.20, 2.6 and 2.8 (it is then coercive in the sense of Definition 2.2, see Lemma 2.9).*
Then, for any $m \in \mathbb{N}$, there exists at least one $u_m \in X_{\mathcal{D}_m}$ solution to the gradient scheme (3.55) *and, up to a subsequence, $\Pi_{\mathcal{D}_m}u_m$ converges strongly in $L^2(\Omega)$ to a solution $\overline{u}$ of* (3.54) *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 (3.54) *is unique, then the whole sequence converges to $\overline{u}$ as $m \to \infty$ in the senses above.*

Proof. Let us first consider any GD $\mathcal{D}$ in the sense of Definition 2.18. We consider any lifting $\bar{g} \in H^1(\Omega)$ such that $\gamma\bar{g} = g$, and we define

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

Thanks to Definition 2.20, we get that $\Pi_{\mathcal{D}_m}P^{\partial}_{\mathcal{D}_m}\bar{g}$ converges strongly in $L^2(\Omega)$ to $\bar{g}$ and $\nabla_{\mathcal{D}_m}P^{\partial}_{\mathcal{D}_m}\bar{g}$ converges strongly in $L^2(\Omega)^d$ to $\nabla\bar{g}$. Then, for any solution u to (3.55), writing $w = u - P^{\partial}_{\mathcal{D}}\bar{g} \in X_{\mathcal{D},0}$, we have

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

The remaining of the proof is then similar to that of Theorem 3.16, reasoning on w instead of u. ∎

3.2.3 Homogeneous Neumann boundary conditions

We consider Problem (3.41a), replacing the homogeneous Dirichlet boundary condition by

$$\Lambda(\cdot,\overline{u})\nabla\overline{u}\cdot\boldsymbol{n} = 0 \text{ on } \partial\Omega,$$

where $\mathbf{n}$ is the unit normal outward Ω to $\partial\Omega$, assumed to be Lipschitz, under Assumptions (3.42) and

$$\int_\Omega f(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = 0. \tag{3.56}$$

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 of this problem is a function $\overline{u}$ (not necessarily unique) satisfying:

$$\begin{aligned}
&\overline{u}\in H^1_\star(\Omega),\ \forall v\in H^1_\star(\Omega),\\
&\int_\Omega \Lambda(\boldsymbol{x},\overline{u}(\boldsymbol{x}))\nabla\overline{u}(\boldsymbol{x})\cdot\nabla v(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = \int_\Omega f(\boldsymbol{x})v(\boldsymbol{x})\mathrm{d}\boldsymbol{x},
\end{aligned} \tag{3.57}$$

and it is approximated by the following gradient scheme.

Definition 3.19 (Non-linear problem, homogeneous Neumann case). *If $\mathcal{D} = (X_{\mathcal{D}},\Pi_{\mathcal{D}},\nabla_{\mathcal{D}})$ is a GD in the sense of Definition 2.24, then we define the related* gradient scheme *for* (3.57) *by*

$$\begin{aligned}
&\textit{Find } u\in X_{\mathcal{D}} \textit{ such that for any } v\in X_{\mathcal{D}},\\
&\int_\Omega \Lambda(\boldsymbol{x},\Pi_{\mathcal{D}}u(\boldsymbol{x}))\nabla_{\mathcal{D}}u(\boldsymbol{x})\cdot\nabla_{\mathcal{D}}v(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\\
&+\int_\Omega \Pi_{\mathcal{D}}u(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\ \int_\Omega \Pi_{\mathcal{D}}v(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = \int_\Omega f(\boldsymbol{x})\Pi_{\mathcal{D}}v(\boldsymbol{x})\mathrm{d}\boldsymbol{x}.
\end{aligned} \tag{3.58}$$

This scheme is again leading to a nonlinear system of equations under the form $A(u)U = B$, similar to (3.45). But the matrix such obtained is in general full, and equivalent algebraic methods leading to sparse matrices must be used. We then have the following convergence result.

Theorem 3.20 (Convergence). *Assume* (3.42)-(3.56)*, and let* $(\mathcal{D}_m)_{m\in\mathbb{N}}$ *be a sequence of GDs in the sense of Definition 2.24, which is space-consistent, limit-conforming and compact in the sense of Definitions 2.27, 2.28 and 2.29 (it is then coercive in the sense of Definition 2.26).*
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) *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* (3.57) *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 3.16.
For any GD $\mathcal{D}$ in the sense of Definition 2.32, let $w \in X_{\mathcal{D}}$ be given, and let $u \in X_{\mathcal{D}}$ be such that

$$\begin{aligned}&\int_\Omega \Lambda(\boldsymbol{x}, \Pi_{\mathcal{D}} w(\boldsymbol{x}))\nabla_{\mathcal{D}} u(\boldsymbol{x}) \cdot \nabla_{\mathcal{D}} v(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \\ &+ \int_\Omega \Pi_{\mathcal{D}} u(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \int_\Omega \Pi_{\mathcal{D}} v(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = \int_\Omega f(\boldsymbol{x})\Pi_{\mathcal{D}} v(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\,, \quad \forall v \in X_{\mathcal{D}}.\end{aligned} \tag{3.59}$$

Then, letting $v = u$ in (3.59), and applying the Cauchy-Schwarz inequality and the coercivity property in the sense of Definition 2.26, 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

$$\|u\|_{\mathcal{D}} \le \frac{C_{\mathcal{D}}}{\min(\underline{\lambda}, 1)}\|f\|_{L^2(\Omega)}. \tag{3.60}$$

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. Applying the Brouwer theorem (Theorem C.2), we see that this mapping $w \to u$ has at least one fixed point. This shows the existence of at least one discrete solution to (3.58). It is clear that any solution to this scheme satisfies (3.60).
We denote by $u_m \in X_{\mathcal{D}_m}$ such a solution for $\mathcal{D} = \mathcal{D}_m$. The estimate (3.60) 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 2.42 and the compactness of the sequence of GDs.
We define $P_{\mathcal{D}} : H^1(\Omega) \to X_{\mathcal{D}}$ by

$$P_{\mathcal{D}}\varphi = \operatorname*{argmin}_{v\in X_{\mathcal{D}}} \left(\|\Pi_{\mathcal{D}} v - \varphi\|_{L^2(\Omega)} + \|\nabla_{\mathcal{D}} v - \nabla\varphi\|_{L^2(\Omega)^d} \right). \tag{3.61}$$

By space-consistency of the sequence of GDs, for any $\varphi \in H^1(\Omega)$ we have $\Pi_{\mathcal{D}_m}(P_{\mathcal{D}_m}\varphi) \to \varphi$ strongly in $L^2(\Omega)$ and $\nabla_{\mathcal{D}_m}(P_{\mathcal{D}_m}\varphi) \to \nabla\varphi$ strongly in $L^2(\Omega)^d$.
Since $\mathbf{1}_\Omega$ (the characteristic function of Ω) belongs to $H^1(\Omega)$, we can take $v = P_{\mathcal{D}_m}\mathbf{1}_\Omega$ in (3.58) and pass to the limit. We get, thanks to Hypothesis (3.56), that

$$\begin{aligned} 0 &= \lim_{m\to\infty} \left(\int_\Omega \Pi_{\mathcal{D}_m} u_m(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \right) \left(\int_\Omega \Pi_{\mathcal{D}_m}(P_{\mathcal{D}_m}\varphi)(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \right) \\ &= \left(\int_\Omega \overline{u}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \right) |\Omega|. \end{aligned}$$

This shows that $\overline{u} \in H^1_\star(\Omega)$ and that

$$\lim_{m\to\infty} \int_\Omega \Pi_{\mathcal{D}_m} u_m(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = 0. \tag{3.62}$$

Let $\varphi \in H^1_\star(\Omega)$ be given. Using the non-linear strong convergence property of Lemma C.4 page 404, $\Lambda(\cdot, \Pi_{\mathcal{D}_m} u_m)\nabla_{\mathcal{D}_m}(P_{\mathcal{D}_m}\varphi) \to \Lambda(\cdot, u)\nabla\varphi$ strongly in $L^2(\Omega)^d$. Lemma C.3 (weak-strong convergence property) enables us to pass to the limit in (3.58) with $v = P_{\mathcal{D}_m}\varphi$, which proves that $\overline{u}$ is a solution to (3.57).
By passing to the limit in the left-hand side of (3.58) with $v = u_m$ and using (3.62), 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 3.16. ∎

3.2.4 Non-homogeneous Neumann boundary conditions

We again refer to Section 2.2.3 for the properties of the trace operator in the case of domain Ω with Lipschitz boundary. We consider Problem (3.41a), replacing the homogeneous Dirichlet boundary condition by

$$\Lambda(\cdot, \overline{u})\nabla\overline{u} \cdot \boldsymbol{n} = h \text{ on } \partial\Omega,$$

where $\mathbf{n}$ is the unit normal outward Ω to $\partial\Omega$, assumed to be Lipschitz, under Assumptions (3.42) and

$$h \in L^2(\partial\Omega),\ \int_\Omega f(\boldsymbol{x})\mathrm{d}\boldsymbol{x} + \int_{\partial\Omega} h(\boldsymbol{x})\mathrm{d}s(\boldsymbol{x}) = 0. \tag{3.63}$$

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 of this problem is a function $\overline{u}$ (not necessarily unique) satisfying:

$$\begin{aligned}&\overline{u} \in H^1_\star(\Omega),\ \forall v \in H^1_\star(\Omega),\\ &\int_\Omega \Lambda(\boldsymbol{x},\overline{u}(\boldsymbol{x}))\nabla\overline{u}(\boldsymbol{x})\cdot\nabla v(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = \int_\Omega f(\boldsymbol{x})v(\boldsymbol{x})\mathrm{d}\boldsymbol{x} + \int_{\partial\Omega} h(\boldsymbol{x})\gamma(v)(\boldsymbol{x})\mathrm{d}s(\boldsymbol{x}).\end{aligned} \tag{3.64}$$

We again recall that, owing to Hypothesis (3.63), Problem (3.64) is equivalent to

$$\begin{aligned}&\overline{u} \in H^1(\Omega),\ \forall v \in H^1(\Omega),\\ &\int_\Omega \Lambda(\boldsymbol{x},\overline{u}(\boldsymbol{x}))\nabla\overline{u}(\boldsymbol{x})\cdot\nabla v(\boldsymbol{x})\mathrm{d}\boldsymbol{x} + \int_\Omega \overline{u}(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\int_\Omega v(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\\ &= \int_\Omega f(\boldsymbol{x})v(\boldsymbol{x})\mathrm{d}\boldsymbol{x} + \int_{\partial\Omega} h(\boldsymbol{x})\gamma(v)(\boldsymbol{x})\mathrm{d}s(\boldsymbol{x}),\end{aligned} \tag{3.65}$$

since letting $v \equiv 1$ in (3.65) implies that $\int_\Omega \overline{u}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = 0$.
This problem is therefore approximated by the following gradient scheme.

Definition 3.21 (Non-linear problem, Neumann case).
If $\mathcal{D} = (X_\mathcal{D}, \Pi_\mathcal{D}, \mathbb{T}_\mathcal{D}, \nabla_\mathcal{D})$ is a GD in the sense of Definition 2.32, then we define the related gradient scheme *for* (3.64) *by*

$$\begin{aligned}&\text{Find } u \in X_\mathcal{D} \text{ such that, for any } v \in X_\mathcal{D},\\ &\int_\Omega \Lambda(\boldsymbol{x},\Pi_\mathcal{D}u(\boldsymbol{x}))\nabla_\mathcal{D}u(\boldsymbol{x})\cdot\nabla_\mathcal{D}v(\boldsymbol{x})\mathrm{d}\boldsymbol{x} + \int_\Omega \Pi_\mathcal{D}u(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\ \int_\Omega \Pi_\mathcal{D}v(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\\ &= \int_\Omega f(\boldsymbol{x})\Pi_\mathcal{D}v(\boldsymbol{x})\mathrm{d}\boldsymbol{x} + \int_{\partial\Omega} h(\boldsymbol{x})\mathbb{T}_\mathcal{D}(v)(\boldsymbol{x})\mathrm{d}s(\boldsymbol{x}).\end{aligned} \tag{3.66}$$

We then have the following convergence result.

Theorem 3.22 (Convergence).
Under assumptions (3.42)-(3.63)*, let $(\mathcal{D}_m)_{m\in\mathbb{N}}$ be a sequence of GDs in the sense of Definition 2.32, which is space-consistent, limit-conforming and compact in the sense of Definitions 2.27, 2.34 and 2.36 (it is then coercive in the sense of Definition 2.33).*
Then, for any $m \in \mathbb{N}$, there exists at least one $u_m \in X_{\mathcal{D}_m}$ solution to the gradient scheme (3.66) *and, up to a subsequence, $\Pi_{\mathcal{D}_m}u_m$ converges strongly in $L^2(\Omega)$ to a solution $\overline{u}$ of* (3.64) *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 (3.64) *is unique, then the whole sequence converges to $\overline{u}$ as $m \to \infty$ in the senses above.*

Proof. We follow the same line as that of the proof of Theorem 3.20, in addition to the use of Remark 2.41 for the weak convergence of the discrete trace.

We first get, thanks again to Brouwer's fix-point theorem, the existence of at least one discrete solution $u_m \in X_{\mathcal{D}_m}$ to (3.66). Thanks to the coercivity hypothesis (Definition 2.33) which involves the discrete trace, we then get that

$$\|u_m\|_{\mathcal{D}_m} \leq \frac{C_P}{\min(\underline{\lambda}, 1)}(\|f\|_{L^2(\Omega)} + \|h\|_{L^2(\partial\Omega)}). \tag{3.67}$$

Then the same arguments as those used in the proof of Theorem 3.20 show that there exists $\overline{u} \in H^1_\star(\Omega)$, and a subsequence 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 $\overline{u}$. Moreover, for the interpolation $v_m = P_{\mathcal{D}_m}\varphi$ defined by (3.61), for any $\varphi \in H^1_\star(\Omega)$, we get using Remark 2.41, that $\mathbb{T}_{\mathcal{D}_m} v_m \to \gamma\varphi$ weakly in $L^2(\partial\Omega)$. The remaining of the proof is then similar to that of the proof of Theorem 3.20. ∎

3.2.5 Non-homogeneous Fourier boundary conditions

We consider the same hypotheses on Ω as in the preceding section (in particular, it is assumed to have a Lipschitz boundary), and we then consider Problem (3.41a) with Fourier boundary conditions:

$$\Lambda(\cdot, \overline{u})\nabla\overline{u} \cdot \boldsymbol{n} + b\overline{u} = h \text{ on } \partial\Omega,$$

under Assumptions (3.42) and

$$\begin{aligned} &h \in L^2(\partial\Omega)\,,\ b \in L^\infty(\partial\Omega) \text{ and} \\ &\text{there exists } \underline{b} > 0 \text{ such that } b(\boldsymbol{x}) \geq \underline{b} \text{ for a.e. } \boldsymbol{x} \in \partial\Omega. \end{aligned} \tag{3.68}$$

A weak solution of this problem is:

$$\begin{aligned} &\overline{u} \in H^1(\Omega),\ \forall v \in H^1(\Omega), \\ &\int_\Omega \Lambda(\boldsymbol{x}, \overline{u}(\boldsymbol{x}))\nabla\overline{u}(\boldsymbol{x}) \cdot \nabla v(\boldsymbol{x})\mathrm{d}\boldsymbol{x} + \int_{\partial\Omega} b(\boldsymbol{x})\gamma(\overline{u})(\boldsymbol{x})\gamma(v)(\boldsymbol{x})\mathrm{d}s(\boldsymbol{x}) \\ &\quad = \int_\Omega f(\boldsymbol{x})v(\boldsymbol{x})\mathrm{d}\boldsymbol{x} + \int_{\partial\Omega} h(\boldsymbol{x})\gamma(v)(\boldsymbol{x})\mathrm{d}s(\boldsymbol{x}). \end{aligned} \tag{3.69}$$

This problem is then approximated by the following gradient scheme.

Definition 3.23 (GS for the non-linear problem, non-homogeneous Fourier case). *If $\mathcal{D} = (X_\mathcal{D}, \Pi_\mathcal{D}, \mathbb{T}_\mathcal{D}, \nabla_\mathcal{D})$ is a GD in the sense of Definition 2.48, then we define the related* gradient scheme *for* (3.69) *by*

$$\begin{aligned} &\text{Find } u \in X_\mathcal{D} \text{ such that, for any } v \in X_\mathcal{D}, \\ &\int_\Omega \Lambda(\boldsymbol{x}, \Pi_\mathcal{D} u(\boldsymbol{x}))\nabla_\mathcal{D} u(\boldsymbol{x}) \cdot \nabla_\mathcal{D} v(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \\ &\quad + \int_{\partial\Omega} b(\boldsymbol{x})\mathbb{T}_\mathcal{D} u(\boldsymbol{x})\mathbb{T}_\mathcal{D} v(\boldsymbol{x})\mathrm{d}s(\boldsymbol{x}) \\ &\qquad = \int_\Omega f(\boldsymbol{x})\Pi_\mathcal{D} v(\boldsymbol{x})\mathrm{d}\boldsymbol{x} + \int_{\partial\Omega} h(\boldsymbol{x})\mathbb{T}_\mathcal{D}(v)(\boldsymbol{x})\mathrm{d}s(\boldsymbol{x}). \end{aligned} \tag{3.70}$$

The convergence result is similar to the previous ones.

Theorem 3.24 (Convergence, Fourier BCs). *Under assumptions* (3.42)-(3.69)*, let* $(\mathcal{D}_m)_{m\in\mathbb{N}}$ *be a sequence of GDs in the sense of Definition 2.48, which is space-consistent, limit-conforming and compact in the sense of Definitions 2.49, 2.34 and 2.36 (it is then coercive in the sense of Definition 2.33).*
Then, for any $m \in \mathbb{N}$*, there exists at least one* $u_m \in X_{\mathcal{D}_m}$ *solution to the gradient scheme* (3.70) *and, up to a subsequence,* $\Pi_{\mathcal{D}_m} u_m$ *converges strongly in* $L^2(\Omega)$ *to a solution* $\overline{u}$ *of* (3.69)*,* $\nabla_{\mathcal{D}_m} u_m$ *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* (3.69) *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 proof of Theorems 3.16 and 3.20, we only indicate here the elements which differ.
Letting $u = v$ in (3.70), by assumption on Λ and b and Definition 2.48 of $||\cdot||_{\mathcal{D}}$ we obtain

$$\begin{aligned}
\min(\underline{\lambda}, \underline{b})||u||_{\mathcal{D}}^2 &\leq \int_\Omega \Lambda(\boldsymbol{x}, \Pi_{\mathcal{D}} u(\boldsymbol{x}))\nabla_{\mathcal{D}} u(\boldsymbol{x}) \cdot \nabla_{\mathcal{D}} u(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \\
&\quad + \int_{\partial\Omega} b(\boldsymbol{x})\mathbb{T}_{\mathcal{D}} u(\boldsymbol{x})\mathbb{T}_{\mathcal{D}} u(\boldsymbol{x})\mathrm{d}s(\boldsymbol{x}) \\
&= \int_\Omega f(\boldsymbol{x})\Pi_{\mathcal{D}} u(\boldsymbol{x})\mathrm{d}\boldsymbol{x} + \int_{\partial\Omega} h(\boldsymbol{x})\mathbb{T}_{\mathcal{D}}(u)(\boldsymbol{x})\mathrm{d}s(\boldsymbol{x}) \\
&\leq \|f\|_{L^2(\Omega)} \|\Pi_{\mathcal{D}} u\|_{L^2(\Omega)} + \|h\|_{L^2(\partial\Omega)} \|\mathbb{T}_{\mathcal{D}}(u)\|_{L^2(\partial\Omega)} \\
&\leq C_{\mathcal{D}}(\|f\|_{L^2(\Omega)} + \|h\|_{L^2(\partial\Omega)}) \|u\|_{\mathcal{D}} .
\end{aligned}$$

This gives an estimate on $\|u\|_{\mathcal{D}}$ which allows us, as in the proof of Theorem 3.16, to use Brouwer's fixed point theorem to prove the existence of a solution to (3.70).
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 2.40 and the compactness of the GDs, that, for some $\overline{u} \in H^1(\Omega)$,

$$\begin{aligned}
&\Pi_{\mathcal{D}_m} u_m \to \overline{u} \text{ strongly in } L^2(\Omega) \text{ and a.e.,} \\
&\mathbb{T}_{\mathcal{D}_m} u_m \to \gamma\overline{u} \text{ weakly in } L^2(\partial\Omega) \text{ and} \\
&\nabla_{\mathcal{D}_m} u_m \to \nabla\overline{u} \text{ weakly in } L^2(\Omega)^d.
\end{aligned} \tag{3.71}$$

Defining then $P_{\mathcal{D}} : H^1(\Omega) \to X_{\mathcal{D}}$ by

$$\begin{aligned}
P_{\mathcal{D}}\varphi = \operatorname*{argmin}_{v\in X_{\mathcal{D}}} \big(&\|\Pi_{\mathcal{D}} w - \varphi\|_{L^2(\Omega)} + \|\nabla_{\mathcal{D}} v - \nabla\varphi\|_{L^2(\Omega)^d} \\
&+ \|\mathbb{T}_{\mathcal{D}} v - \gamma\varphi\|_{L^2(\Omega)^d}\big)
\end{aligned}$$

the space-consistency of the sequence of GDs shows that, for any $\varphi \in H^1(\Omega)$, $\Pi_{\mathcal{D}_m}(P_{\mathcal{D}_m}\varphi) \to \varphi$ strongly in $L^2(\Omega)$, $\nabla_{\mathcal{D}_m}(P_{\mathcal{D}_m}\varphi) \to \nabla\varphi$ strongly in $L^2(\Omega)^d$ and $\mathbb{T}_{\mathcal{D}_m}(P_{\mathcal{D}_m}\varphi) \to \gamma\varphi$ strongly in $L^2(\partial\Omega)$.

We can then, as in the proof of Theorem 3.16, use $v = P_{\mathcal{D}_m}\varphi$ in (3.70) and pass to the limit, thanks to these strong convergences and to (3.71), to see that $\overline{u}$ is a solution to (3.69).
We then take $v = u_m$ in (3.70) and pass to the limit to obtain

$$\begin{aligned}
&\lim_{m\to\infty} \left(\int_\Omega \Lambda(\boldsymbol{x}, \Pi_{\mathcal{D}_m} u_m(\boldsymbol{x})) \nabla_{\mathcal{D}_m} u_m(\boldsymbol{x}) \cdot \nabla_{\mathcal{D}_m} u_m(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \right. \\
&\qquad \left. + \int_{\partial\Omega} b(\boldsymbol{x}) \mathbb{T}_{\mathcal{D}_m} u_m(\boldsymbol{x})^2 \mathrm{d}s(\boldsymbol{x}) \right) \\
&= \int_\Omega f(\boldsymbol{x}) \overline{u}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} + \int_{\partial\Omega} h(\boldsymbol{x}) \gamma\overline{u}(\boldsymbol{x}) \mathrm{d}s(\boldsymbol{x}) \\
&= \int_\Omega \Lambda(\boldsymbol{x}, \overline{u}(\boldsymbol{x})) \nabla\overline{u}(\boldsymbol{x}) \cdot \nabla\overline{u}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} + \int_{\partial\Omega} b(\boldsymbol{x}) \gamma\overline{u}(\boldsymbol{x})^2 \mathrm{d}s(\boldsymbol{x}).
\end{aligned}$$

This limit and (3.71) allows us to see that

$$\begin{aligned}
&\int_\Omega \Lambda(\boldsymbol{x}, \Pi_{\mathcal{D}_m} u_m(\boldsymbol{x})) (\nabla_{\mathcal{D}_m} u_m(\boldsymbol{x}) - \nabla\overline{u}(\boldsymbol{x})) \cdot (\nabla_{\mathcal{D}_m} u_m(\boldsymbol{x}) - \nabla\overline{u}(\boldsymbol{x})) \mathrm{d}\boldsymbol{x} \\
&\qquad + \int_{\partial\Omega} b(\boldsymbol{x}) (\mathbb{T}_{\mathcal{D}_m} u_m(\boldsymbol{x}) - \gamma\overline{u}(\boldsymbol{x}))^2 \mathrm{d}s(\boldsymbol{x}) \to 0.
\end{aligned}$$

By Assumptions (3.42b) on Λ and (3.68) on b, the left-hand side of this limit is greater than $\min(\underline{\lambda}, \underline{b})(\|\nabla_{\mathcal{D}_m} u_m - \nabla\overline{u}\|^2_{L^2(\Omega)^d} + \|\mathbb{T}_{\mathcal{D}_m} u_m - \gamma\overline{u}\|^2_{L^2(\partial\Omega)^d})$. The strong convergences of the reconstructed gradient and trace therefore follow. ∎

Remark 3.25. In the linear case (Λ independent of u), it is very easy to obtain error estimates for (3.70) similar to the ones in Theorem 3.2 but with an additional error estimate on the traces.

3.3 p-Laplacian type problems: $p \in (1, +\infty)$

After the study of the approximation of quasilinear elliptic problem in Section 3.2, we now turn to the study of the approximation of problems involving a nonlinear expression with respect to the gradient of the unknown function, using gradient schemes. The first case which is studied is that of the p-Laplace problem, which enables an error estimate, in terms of $W_{\mathcal{D}}$ and $S_{\mathcal{D}}$ in the same way as error estimates are provided in Section 3.1. In the more general case which we consider in Section 3.3.2, only convergence results are given.

3.3.1 An error estimate for the p-Laplace problem

We consider in this section a particular case of a non-linear Leray-Lions problem, the so-called p-Laplace equation:

$$-\mathrm{div}(|\nabla \overline{u}|^{p-2}\nabla \overline{u}) = f + \mathrm{div}(\boldsymbol{F}) \text{ in } \Omega, \tag{3.72a}$$

with boundary conditions

$$\overline{u} = 0 \text{ on } \partial\Omega, \tag{3.72b}$$

under the following assumptions:

- Ω is an open bounded connected subset of $\mathbb{R}^d$ ($d \in \mathbb{N}^\star$), (3.73a)
- $p \in (1, +\infty)$ (3.73b)
- $f \in L^{p'}(\Omega)\,,\ \boldsymbol{F} \in L^{p'}(\Omega)^d$. (3.73c)

Under these hypotheses, the weak solution of (3.72) is the unique function $\overline{u}$ satisfying:

$$\begin{aligned} &\overline{u} \in W_0^{1,p}(\Omega) \text{ and, for all } v \in W_0^{1,p}(\Omega),\\ &\int_\Omega |\nabla \overline{u}|^{p-2}\nabla \overline{u}(\boldsymbol{x}) \cdot \nabla v(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}\\ &\qquad = \int_\Omega f(\boldsymbol{x}) v(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} - \int_\Omega \boldsymbol{F}(\boldsymbol{x}) \cdot \nabla v(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}. \end{aligned} \tag{3.74}$$

Definition 3.26 (Gradient scheme for the 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* (3.74) *is defined by*

$$\begin{aligned} &\text{Find } u \in X_{\mathcal{D},0} \text{ such that, for any } v \in X_{\mathcal{D},0},\\ &\int_\Omega |\nabla_{\mathcal{D}} u(\boldsymbol{x})|^{p-2}\nabla_{\mathcal{D}} u(\boldsymbol{x}) \cdot \nabla_{\mathcal{D}} v(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}\\ &\qquad = \int_\Omega f(\boldsymbol{x}) \Pi_{\mathcal{D}} v(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} - \int_\Omega \boldsymbol{F}(\boldsymbol{x}) \cdot \nabla_{\mathcal{D}} v(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}. \end{aligned} \tag{3.75}$$

The following lemma establishes the existence and uniqueness of the solutions to (3.74) and (3.75), as well as estimates on these solutions.

Lemma 3.27. *Under Hypotheses* (3.73)*, there exists one and only one solution to each of the problems* (3.74) *and* (3.75)*. These solutions moreover satisfy*

$$\|\nabla \overline{u}\|_{L^p(\Omega)^d} \le (C_{P,p}\|f\|_{L^{p'}(\Omega)} + \|\boldsymbol{F}\|_{L^{p'}(\Omega)^d})^{\frac{1}{p-1}} \tag{3.76}$$

and

$$\|\nabla_{\mathcal{D}} u_{\mathcal{D}}\|_{L^p(\Omega)^d} \le (C_{\mathcal{D}}\|f\|_{L^{p'}(\Omega)} + \|\boldsymbol{F}\|_{L^{p'}(\Omega)^d})^{\frac{1}{p-1}}, \tag{3.77}$$

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 (3.74) and (3.75) are respectively equivalent to the minimisation problems

$$\overline{u} \in \operatorname*{argmin}_{v \in W_0^{1,p}(\Omega)} \left(\frac{1}{p}\int_\Omega |\nabla v|^p \mathrm{d}\boldsymbol{x} - \int_\Omega f(\boldsymbol{x}) v(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} + \int_\Omega \boldsymbol{F}(\boldsymbol{x}) \cdot \nabla v(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \right) \tag{3.78}$$

and

$$u_{\mathcal{D}} \in \operatorname*{argmin}_{v \in X_{\mathcal{D},0}} \Big(\frac{1}{p} \int_\Omega |\nabla_{\mathcal{D}} v(\boldsymbol{x})|^p \mathrm{d}\boldsymbol{x} - \int_\Omega f(\boldsymbol{x}) \Pi_{\mathcal{D}} v(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} + \int_\Omega \boldsymbol{F}(\boldsymbol{x}) \cdot \nabla_{\mathcal{D}} v(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \Big). \quad (3.79)$$

This equivalence is a consequence of the inequality

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

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 (3.78) and (3.79) are classical consequence of standard convex minimisation theorems, see e.g. [6].
Then inequalities (3.76) and (3.77) follow by taking, in each corresponding problem, the solution itself as a test function. ∎

Theorem 3.28 (Control of the approximation error). *Under Hypotheses* (3.73)*, let $\overline{u} \in W_0^{1,p}(\Omega)$ be the solution of Problem* (3.74)*, 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* (3.75)*. Then there exists $C_2 > 0$, depending only on p such that:*

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

$$\|\nabla \overline{u} - \nabla_{\mathcal{D}} u_{\mathcal{D}}\|_{L^p(\Omega)^d} \leq S_{\mathcal{D}}(\overline{u}) + C_2 \big[W_{\mathcal{D}}(|\nabla \overline{u}|^{p-2} \nabla \overline{u} + \boldsymbol{F}) + S_{\mathcal{D}}(\overline{u})^{p-1} \big] \times \Big[S_{\mathcal{D}}(\overline{u})^p + \big[(C_{\mathcal{D}} + C_{P,p}) \|f\|_{L^{p'}(\Omega)} + \|\boldsymbol{F}\|_{L^{p'}(\Omega)^d} \big]^{\frac{p}{p-1}} \Big]^{\frac{2-p}{2}}. \quad (3.80)$$

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

$$\|\nabla \overline{u} - \nabla_{\mathcal{D}} u_{\mathcal{D}}\|_{L^p(\Omega)^d} \leq S_{\mathcal{D}}(\overline{u}) + C_2 \Big[W_{\mathcal{D}}(|\nabla \overline{u}|^{p-2} \nabla \overline{u} + \boldsymbol{F}) + S_{\mathcal{D}}(\overline{u}) \big[(C_{P,p} \|f\|_{L^{p'}(\Omega)} + \|\boldsymbol{F}\|_{L^{p'}(\Omega)^d})^{\frac{1}{p-1}} + S_{\mathcal{D}}(\overline{u}) \big]^{p-2} \Big]^{\frac{1}{p-1}}. \quad (3.81)$$

As a consequence of (3.80)–(3.81)*, we have the following error estimate:*

$$\|\overline{u} - \Pi_{\mathcal{D}} u_{\mathcal{D}}\|_{L^p(\Omega)} \leq S_{\mathcal{D}}(\overline{u}) + C_{\mathcal{D}} (S_{\mathcal{D}}(\overline{u}) + \|\nabla \overline{u} - \nabla_{\mathcal{D}} u_{\mathcal{D}}\|_{L^p(\Omega)^d}). \quad (3.82)$$

Remark 3.29 (Mesh-based gradient schemes). As in Remark 3.3 (for the case $p = 2$ and $d \leq 3$), under non-degeneracy assumptions on the meshes, for many mesh-based gradient schemes it can be proved that there exists $C \in \mathbb{R}_+$, not depending on $\mathcal{D}$, such that

$$\forall \varphi \in W^{2,p}(\Omega) \cap W_0^{1,p}(\Omega),\ S_{\mathcal{D}}(\varphi) \leq C h_{\mathcal{D}} ||\varphi||_{W^{2,p}(\Omega)}$$

and

$$\forall \boldsymbol{\varphi} \in W^{1,p'}(\Omega)^d\,,\ W_{\mathcal{D}}(\boldsymbol{\varphi}) \le C h_{\mathcal{D}} ||\boldsymbol{\varphi}||_{W^{1,p'}(\Omega)^d},$$

where $h_{\mathcal{D}}$ measures the mesh size. Under the condition $p > d/2$, the proofs of these inequalities are done, for some important examples of gradient schemes, in Part III. Therefore, in the case where $\overline{u} \in W^{2,p}(\Omega)$ and $|\nabla \overline{u}|^{p-2}\nabla \overline{u} + \boldsymbol{F} \in W^{1,p'}(\Omega)^d$, the preceding theorem gives an error estimate of the form $\mathcal{O}(h_{\mathcal{D}}^{p-1})$ if $p \in (1,2]$ and $\mathcal{O}(h_{\mathcal{D}}^{1/(p-1)})$ if $p \ge 2$.

Proof. We notice that, since $f \in L^{p'}(\Omega)$, the equation (3.72a) in the sense of distributions (*i.e.* taking $v \in C_c^\infty(\Omega)$ in (3.74)) 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 $\mathrm{div}\boldsymbol{\varphi} = -f$. We can therefore take $\boldsymbol{\varphi}$ in the definition (2.6) of $W_{\mathcal{D}}$ and we obtain, for any $v \in X_{\mathcal{D},0}$, denoting by $\overline{W} = W_{\mathcal{D}}(|\nabla \overline{u}|^{p-2}\nabla \overline{u} + \boldsymbol{F})$ and $\overline{S} = S_{\mathcal{D}}(\overline{u})$,

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

We use the fact that $u_{\mathcal{D}}$ satisfies (3.75) to replace the term $\Pi_{\mathcal{D}} v\, f$ and we get

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

We set

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

and we obtain

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

Case $p \in (1,2]$.
Thanks to (3.86) in Lemma 3.30 below, we get the existence of C_3 depending only on p such that

$$\| \, |\nabla_{\mathcal{D}} P_{\mathcal{D}} \overline{u}|^{p-2}\nabla_{\mathcal{D}} P_{\mathcal{D}} \overline{u} - |\nabla \overline{u}|^{p-2}\nabla \overline{u}\|_{L^{p'}(\Omega)^d}^{p'} \le C_3 \|\nabla_{\mathcal{D}} P_{\mathcal{D}} \overline{u} - \nabla \overline{u}\|_{L^p(\Omega)^d}^p,$$

which leads, recalling the definition (2.2) of $S_{\mathcal{D}}$, to

$$A(v) \le \|\nabla_{\mathcal{D}} v\|_{L^p(\Omega)^d} \Big[\overline{W} + C_3^{p-1} \overline{S}^{p-1} \Big]. \tag{3.83}$$

We then apply (3.88) in Lemma 3.30 with $\xi = \nabla_{\mathcal{D}} P_{\mathcal{D}} \overline{u}$ and $\chi = \nabla_{\mathcal{D}} u_{\mathcal{D}}$, and use Hölder's inequality with exponents $2/p$ and $2/(2-p)$. Taking $v = P_{\mathcal{D}} \overline{u} - u_{\mathcal{D}}$, we get C_4 depending only on p such that

$$\begin{aligned} &\|\nabla_{\mathcal{D}} P_{\mathcal{D}} \overline{u} - \nabla_{\mathcal{D}} u_{\mathcal{D}}\|^p_{L^p(\Omega)^d} \\ &\qquad \le C_4 A(P_{\mathcal{D}} \overline{u} - u_{\mathcal{D}})^{\frac{p}{2}} (\|\nabla_{\mathcal{D}} P_{\mathcal{D}} \overline{u}\|^p_{L^p(\Omega)^d} + \|\nabla_{\mathcal{D}} u_{\mathcal{D}}\|^p_{L^p(\Omega)^d})^{\frac{2-p}{2}}, \end{aligned}$$

and thus

$$\begin{aligned} &\|\nabla_{\mathcal{D}} P_{\mathcal{D}} \overline{u} - \nabla_{\mathcal{D}} u_{\mathcal{D}}\|^2_{L^p(\Omega)^d} \\ &\qquad \le C_4^{\frac{2}{p}} A(P_{\mathcal{D}} \overline{u} - u_{\mathcal{D}}) (\|\nabla_{\mathcal{D}} P_{\mathcal{D}} \overline{u}\|^p_{L^p(\Omega)^d} + \|\nabla_{\mathcal{D}} u_{\mathcal{D}}\|^p_{L^p(\Omega)^d})^{\frac{2-p}{p}}. \end{aligned}$$

Plugging (3.83) into this estimate gives C_5 depending only on p such that

$$\begin{aligned} \|\nabla_{\mathcal{D}} P_{\mathcal{D}} \overline{u} - \nabla_{\mathcal{D}} u_{\mathcal{D}}\|_{L^p(\Omega)^d} &\le C_5 \big[\overline{W} + \overline{S}^{p-1} \big] \\ &\quad \times \big[\|\nabla_{\mathcal{D}} P_{\mathcal{D}} \overline{u}\|^p_{L^p(\Omega)^d} + \|\nabla_{\mathcal{D}} u_{\mathcal{D}}\|^p_{L^p(\Omega)^d} \big]^{\frac{2-p}{2}}. \end{aligned}$$

We have $\|\nabla \overline{u} - \nabla_{\mathcal{D}} u_{\mathcal{D}}\|_{L^p(\Omega)^d} \le \overline{S} + \|\nabla_{\mathcal{D}} P_{\mathcal{D}} \overline{u} - \nabla_{\mathcal{D}} u_{\mathcal{D}}\|_{L^p(\Omega)^d}$, and Estimate (3.80) therefore follows from (3.76) and (3.77).

Case $p \in (2, +\infty)$.
We use (3.85) in Lemma 3.30, Hölder's inequality with exponents $p/p' = p-1$ and $\frac{p-1}{p-2}$, and $(a+b)^\theta \le 2^\theta (a^\theta + b^\theta)$ with $\theta = \frac{p}{p-2}$, $a = |\nabla_{\mathcal{D}} P_{\mathcal{D}} \overline{u}|^{p-2}$ and $b = |\nabla \overline{u}|^{p-2}$. This gives C_6 depending only on p such that

$$\begin{aligned} &\| \, |\nabla_{\mathcal{D}} P_{\mathcal{D}} \overline{u}|^{p-2} \nabla_{\mathcal{D}} P_{\mathcal{D}} \overline{u} - |\nabla \overline{u}|^{p-2} \nabla \overline{u}\|_{L^{p'}(\Omega)^d} \\ &\qquad \le C_6 \|\nabla_{\mathcal{D}} P_{\mathcal{D}} \overline{u} - \nabla \overline{u}\|_{L^p(\Omega)^d} (\|\nabla_{\mathcal{D}} P_{\mathcal{D}} \overline{u}\|_{L^p(\Omega)^d} + \|\nabla \overline{u}\|_{L^p(\Omega)^d})^{p-2}. \end{aligned}$$

This leads to

$$\begin{aligned} &A(v) \le \\ &\qquad \|\nabla_{\mathcal{D}} v\|_{L^p(\Omega)^d} \Big[\overline{W} + C_6 \overline{S} \big[\|\nabla_{\mathcal{D}} P_{\mathcal{D}} \overline{u}\|_{L^p(\Omega)^d} + \|\nabla \overline{u}\|_{L^p(\Omega)^d} \big]^{p-2} \Big]. \end{aligned} \tag{3.84}$$

As before, we take $v = P_{\mathcal{D}} \overline{u} - u_{\mathcal{D}}$. Thanks to (3.89) in Lemma 3.30, we get the existence of C_7 depending only on p such that

$$\|\nabla_{\mathcal{D}} P_{\mathcal{D}} \overline{u} - \nabla_{\mathcal{D}} u_{\mathcal{D}}\|^p_{L^p(\Omega)^d} \le C_7 A(P_{\mathcal{D}} \overline{u} - u_{\mathcal{D}}).$$

Using (3.84) we infer

$$\|\nabla_{\mathcal{D}} P_{\mathcal{D}} \overline{u} - \nabla_{\mathcal{D}} u_{\mathcal{D}}\|_{L^p(\Omega)^d}^{p-1} \le C_7 \Big[\overline{W} + C_6 \overline{S} \big[\|\nabla \overline{u}\|_{L^p(\Omega)^d} + \overline{S} \big]^{p-2} \Big],$$

and the proof of (3.81) is complete by invoking (3.76). ∎

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

Lemma 3.30. *Let $p \in (1, +\infty)$ and $d \in \mathbb{N}^\star$. Then*

$$\forall \xi, \chi \in \mathbb{R}^d, \quad \big| |\xi|^{p-2}\xi - |\chi|^{p-2}\chi \big| \le \max(1, p-1) |\xi - \chi| \ (|\xi|^{p-2} + |\chi|^{p-2}), \tag{3.85}$$

which implies

$$\forall p \in (1, 2], \ \forall \xi, \chi \in \mathbb{R}^d, \ \big| |\xi|^{p-2}\xi - |\chi|^{p-2}\chi \big| \le 5 |\xi - \chi|^{p-1}. \tag{3.86}$$

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

$$\forall \xi, \chi \in \mathbb{R}^d, \quad C_0(p) (|\xi|^{p-2}\xi - |\chi|^{p-2}\chi) \cdot (\xi - \chi) \ge |\xi - \chi|^2 \ (|\xi| + |\chi|)^{p-2}, \tag{3.87}$$

which implies

$$\forall p \in (1, 2], \ \forall \xi, \chi \in \mathbb{R}^d, \quad |\xi - \chi|^p \le \left(\frac{2}{p-1} (|\xi|^{p-2}\xi - |\chi|^{p-2}\chi) \cdot (\xi - \chi) \right)^{\frac{p}{2}} \times (2^{p-1} (|\xi|^p + |\chi|^p))^{\frac{2-p}{2}}, \tag{3.88}$$

and

$$\forall p \ge 2, \ \forall \xi, \chi \in \mathbb{R}^d, \ |\xi - \chi|^p \le 2^{p-1} (|\xi|^{p-2}\xi - |\chi|^{p-2}\chi) \cdot (\xi - \chi). \tag{3.89}$$

Proof. Estimates (3.85) and (3.86) originally appeared in [7]. Let $H(\xi) = |\xi|^{p-2}\xi$. If $p \ge 2$ then $H \in C^1(\mathbb{R}^d)^d$ and $|DH(\xi)| \le (p-1)|\xi|^{p-2}$ (where DH is the differential of H and $|DH|$ the norm induced by the Euclidean norm). Hence, for all $\xi, \chi \in \mathbb{R}^d$,

$$|H(\xi) - H(\chi)| \le |\xi - \chi| (p-1) \max_{\zeta \in [\xi, \chi]} |\zeta|^{p-2}. \tag{3.90}$$

The proof of (3.85) is complete in the case $p \ge 2$ since the mapping $s \mapsto s^{p-2}$ is non-decreasing, and thus $\max_{\zeta \in [\xi, \chi]} |\zeta|^{p-2} = \max(|\xi|^{p-2}, |\chi|^{p-2}) \le (|\xi|^{p-2} + |\chi|^{p-2})$.

If $p < 2$, (3.90) remains valid but does directly lead to (3.85). Without loss of generality, we can assume that $0 < |\chi| \le |\xi|$ (the case where $\chi = 0$ is trivial since the right-hand side of (3.85) 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 center 0 and radius $|\chi|$ (see Figure 3.1). Since $|\widetilde{\xi}| = |\chi|$ we have $|H(\widetilde{\xi}) - H(\chi)| = |\chi|^{p-2}|\widetilde{\xi} - \chi|$. Hence, by the triangular inequality and (3.90) between ξ and $\widetilde{\xi}$,

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

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

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

The proof of (3.85) in the case $p < 2$ is complete by noticing that $|\xi - \widetilde{\xi}| + |\widetilde{\xi} - \chi| = |\xi - \chi|$.

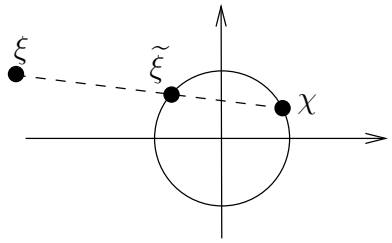


Fig. 3.1. Illustration of the proof of (3.85) in the case $p \in (1, 2)$.

Let us now prove (3.86). Let $\eta > 0$. If $|\xi|$ and $|\chi|$ belong to $[\eta, +\infty)$, by (3.85) we have

$$\left| |\xi|^{p-2}\xi - |\chi|^{p-2}\chi \right| \le 2\eta^{p-2}|\xi - \chi|. \tag{3.91}$$

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

$$\left| |\xi|^{p-2}\xi - |\chi|^{p-2}\chi \right| \le |\xi|^{p-1} + \eta^{p-1} \le (|\xi - \chi| + \eta)^{p-1} + \eta^{p-1}. \tag{3.92}$$

Combining (3.91) and (3.92) we see that, for all $\xi, \chi \in \mathbb{R}^d$ and all $\eta > 0$,

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

Estimate (3.86) follows by choosing $\eta = |\xi - \chi|$.
We now turn to the proof of (3.87). Set $A = (|\xi|^{p-2}\xi - |\chi|^{p-2}\chi) \cdot (\xi - \chi)$. By developing both sides we see that

$$A = (|\xi|^{p-1} - |\chi|^{p-1})(|\xi| - |\chi|) + (|\xi|^{p-2} + |\chi|^{p-2})(|\xi|\,|\chi| - \xi \cdot \chi). \tag{3.93}$$

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 $1 < p \leq 2$, we write $f'(x) \geq (p-1)x^{p-2} - \frac{2}{C_0(p)}x^{p-2} = 0$ since $C_0(p) = \frac{2}{p-1}$.
- If $p > 2$, $f'(x) \geq (p-1)x^{p-2} - \frac{2}{C_0(p)}(p-2)(x+y)^{p-3}(x+y) - \frac{2}{C_0(p)}(x+y)^{p-2}$, and therefore, since $x \geq y$, $f'(x) \geq (p-1)(x^{p-2} - \frac{2}{C_0(p)}2^{p-2}x^{p-2}) = 0$ since $C_0(p) = 2^{p-1}$.

Since $f(y) = 0$, this shows that, if $x \geq y \geq 0$,

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

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

$$(|\xi|^{p-1} - |\chi|^{p-1})(|\xi| - |\chi|) \geq \frac{1}{C_0(p)}(|\xi| + |\chi|)^{p-2}(|\xi| - |\chi|)^2. \tag{3.94}$$

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 (3.93) and using (3.94) 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 (3.87) 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 (3.88) is obtained by raising (3.87) to the power $p/2$ and by using $(|\xi| + |\chi|)^p \leq 2^{p-1}(|\xi|^p + |\chi|^p)$. Estimate (3.89) follows by writing $|\xi - \chi|^p = |\xi - \chi|^2|\xi - \chi|^{p-2} \leq |\xi - \chi|^2(|\xi| + |\chi|)^{p-2}$ and by using (3.87). ∎

3.3.2 Convergence of gradient schemes for fully nonlinear Leray–Lions problems

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

$$\begin{array}{l} -\text{div } \boldsymbol{a}(\boldsymbol{x}, \overline{u}, \nabla \overline{u}) = f \text{ in } \Omega, \\ \overline{u} = 0 \text{ on } \partial\Omega, \end{array} \tag{3.95}$$

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 (3.96a)

(*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. $\boldsymbol{x} \in \Omega$, $\forall u \in L^p(\Omega),\ \forall \boldsymbol{\xi} \in \mathbb{R}^d$, (3.96b)
- $(\boldsymbol{a}(\boldsymbol{x}, u, \boldsymbol{\xi}) - \boldsymbol{a}(\boldsymbol{x}, u, \boldsymbol{\chi})) \cdot (\boldsymbol{\xi} - \boldsymbol{\chi}) \geq 0$ for a.e. $\boldsymbol{x} \in \Omega$, $\forall u \in L^p(\Omega),\ \forall \boldsymbol{\xi}, \boldsymbol{\chi} \in \mathbb{R}^d$, (3.96c)
- $\exists \overline{a} \in L^{p'}(\Omega)\,, \exists \mu \in (0, +\infty)$ such that $|\boldsymbol{a}(\boldsymbol{x}, u, \boldsymbol{\xi})| \leq \overline{a}(\boldsymbol{x}) + \mu |\boldsymbol{\xi}|^{p-1}$ for a.e. $\boldsymbol{x} \in \Omega,\ \forall u \in L^p(\Omega),\ \forall \boldsymbol{\xi} \in \mathbb{R}^d$, (3.96d)
- $f \in L^{p'}(\Omega)\,,$ where $p' = \dfrac{p}{p-1}$. (3.96e)

Remark 3.31. 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 [18, 28, 69].
These assumptions (in particular (3.96a)) 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 [62]. However, the adaptation of the following results to this case is quite easy and more classical, see e.g. [30] for an adaptation of the original Leray-Lions method to a numerical scheme (based on the Mixed Finite Volume method) for local non-monotone operators.

If a function $\boldsymbol{a}$ satisfies (3.96), then the mapping $u \mapsto -\text{div} \boldsymbol{a}(\cdot, u, \nabla u)$ is called a generalised Leray-Lions operator. A classical example is the p-Laplacian operator, obtained by setting $\boldsymbol{a}(\boldsymbol{x}, u, \boldsymbol{\xi}) = |\boldsymbol{\xi}|^{p-2} \boldsymbol{\xi}$. Note that the existence of at least one solution to (3.95) is shown in [62] under hypotheses (3.96) 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 (3.95) if:

$$\begin{array}{l} \overline{u} \in W_0^{1,p}(\Omega),\ \forall \overline{v} \in W_0^{1,p}(\Omega), \\ \displaystyle\int_\Omega \boldsymbol{a}(\boldsymbol{x}, \overline{u}, \nabla \overline{u}(\boldsymbol{x})) \cdot \nabla \overline{v}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} = \int_\Omega f(\boldsymbol{x}) \overline{v}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}. \end{array} \tag{3.97}$$

Remark 3.32. Note that, even if $\boldsymbol{a}$ does not depend on $u \in L^p(\Omega)$, the solution to (3.97) 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 (3.96b) is satisfied with $\underline{a} = \frac{1}{2}$, (3.96c) is satisfied since $\boldsymbol{a}$ is non decreasing with respect to $\boldsymbol{\xi}$ and (3.96d) 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 (3.97) for any value $\alpha\in[1,2]$.

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

$$\begin{aligned}&(\boldsymbol{a}(\boldsymbol{x},u,\boldsymbol{\xi}) - \boldsymbol{a}(\boldsymbol{x},u,\boldsymbol{\chi}))\cdot(\boldsymbol{\xi}-\boldsymbol{\chi}) > 0,\\ &\text{for a.e. } \boldsymbol{x}\in\Omega,\ \forall u\in L^p(\Omega),\ \forall \boldsymbol{\xi},\boldsymbol{\chi}\in\mathbb{R}^d \text{ with } \boldsymbol{\xi}\neq\boldsymbol{\chi},\end{aligned} \tag{3.98}$$

is only used to prove the strong convergence of the approximate gradient (see theorem below). We now define the gradient scheme for Problem (3.95).

Definition 3.33 (GS for fully non-linear Leray–Lions problems, homogeneous Dirichlet BCs). *If $\mathcal{D} = (X_{\mathcal{D},0}, \Pi_{\mathcal{D}}, \nabla_{\mathcal{D}})$ is a GD, then we define the related* gradient scheme *for* (3.95) *by*

$$\begin{aligned}&\text{Find } u\in X_{\mathcal{D},0} \text{ such that, } \forall v\in X_{\mathcal{D},0},\\ &\int_\Omega \boldsymbol{a}(\boldsymbol{x},\Pi_{\mathcal{D}}u,\nabla_{\mathcal{D}}u(\boldsymbol{x}))\cdot\nabla_{\mathcal{D}}v(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = \int_\Omega f(\boldsymbol{x})\Pi_{\mathcal{D}}v(\boldsymbol{x})\mathrm{d}\boldsymbol{x}.\end{aligned} \tag{3.99}$$

Theorem 3.34 (Convergence). *Under assumptions* (3.96)-(3.96e)*, take a sequence $(\mathcal{D}_m)_{m\in\mathbb{N}}$ of GDs in the sense of Definition 2.1, which is space-consistent, limit-conforming and compact in the sense of Definitions 2.4, 2.6 and 2.8 (it is then coercive in the sense of Definition 2.2).*
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 (3.99) *and, up to a subsequence, $\Pi_{\mathcal{D}_m}u_{\mathcal{D}_m}$ converges strongly in $L^p(\Omega)$ to a solution $\overline{u}$ of* (3.97) *and $\nabla_{\mathcal{D}_m}u_{\mathcal{D}_m}$ converges weakly in $L^p(\Omega)^d$ to $\nabla\overline{u}$ as $m\to\infty$. Moreover, if we assume that the Leray-Lions operator $\boldsymbol{a}$ is strictly monotone in the sense of* (3.98)*, 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 (3.97) *is unique, then the whole sequence converges to $\overline{u}$ as $m\to\infty$ in the above senses.*

Remark 3.35. As a by-product, this theorem also gives the existence of a solution $\bar{u}$ to (3.97). Indeed, under the assumptions of the theorem, the proof shows that the sequence $u_{\mathcal{D}_m}$ has a converging subsequence and that the limit $\bar{u}$ of this subsequence is in fact a solution to the continuous problem. Since there exists at least one (in fact there exist several, see Part III) gradient scheme which satisfies the assumptions of this theorem, this gives the existence of a solution to (3.97).

Remark 3.36 (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 (3.97) does not require the compactness of the sequence of GDs. In this case the strong convergence results from (3.98) (which gives the strong convergence of the approximate gradient) and from the coercivity and the space-consistency of the sequence $(\mathcal{D}_m)_{m\in\mathbb{N}}$.

Proof.
This proof follows the same ideas as in [30, 48].

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 a 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 $\langle F(u), u\rangle \geq C_8 |u|^p$ with C_8 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 [62] or [27, Theorem 3.3, page 19]). Note that we could as well use Theorem C.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 (3.99).

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

Letting $v = u_{\mathcal{D}_m}$ in (3.99) with $\mathcal{D} = \mathcal{D}_m$ and using (2.1) and Hypothesis (3.96b), 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 2.12 then gives $\overline{u} \in W^{1,p}_0(\Omega)$ 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 (3.96d), 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 susbsequence, to some $\boldsymbol{A}$ weakly in $L^{p'}(\Omega)^d$, as $m \to \infty$.
Let us now show that $\overline{u}$ is solution to (3.97), using the well-known Minty trick [64]. 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

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

as a test function in (3.99). By the space-consistency of $(\mathcal{D}_m)_{m\in\mathbb{N}}$, letting $m \to \infty$ we get

$$\int_\Omega \boldsymbol{A}(\boldsymbol{x}) \cdot \nabla\varphi(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} = \int_\Omega f(\boldsymbol{x})\varphi(\boldsymbol{x})\mathrm{d}\boldsymbol{x}, \quad \forall \varphi \in W_0^{1,p}(\Omega). \tag{3.100}$$

On the other hand, we may let $m \to \infty$ in (3.99) with $u_{\mathcal{D}_m}$ as a test function. Using (3.100) with $\varphi = \overline{u}$, this leads to

$$\begin{aligned} &\lim_{m\to\infty} \int_\Omega \boldsymbol{a}(\boldsymbol{x}, \Pi_{\mathcal{D}_m} u_{\mathcal{D}_m}, \nabla_{\mathcal{D}_m} u_{\mathcal{D}_m}(\boldsymbol{x})) \cdot \nabla_{\mathcal{D}_m} u_{\mathcal{D}_m}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \\ &= \int_\Omega f(\boldsymbol{x})\overline{u}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = \int_\Omega \boldsymbol{A}(\boldsymbol{x}) \cdot \nabla\overline{u}(\boldsymbol{x})\mathrm{d}\boldsymbol{x}. \end{aligned} \tag{3.101}$$

Hypothesis (3.96c) 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 C.3 for the weak-strong convergences and (3.101) 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

$$\boldsymbol{A}(\boldsymbol{x}) = \boldsymbol{a}(\boldsymbol{x}, \overline{u}, \nabla \overline{u}(\boldsymbol{x})), \text{ for a.e. } \boldsymbol{x} \in \Omega. \tag{3.102}$$

In addition to (3.100), this shows that $\overline{u}$ is a solution to (3.97). 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 (3.98), strong convergence of the approximate gradient.

We follow here the ideas of [62]. Thanks to (3.101) and (3.102), 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

$$[\boldsymbol{a}(\cdot, \Pi_{\mathcal{D}_m} u_{\mathcal{D}_m}, \nabla_{\mathcal{D}_m} u_{\mathcal{D}_m}) - \boldsymbol{a}(\cdot, \Pi_{\mathcal{D}_m} u_{\mathcal{D}_m}, \nabla \overline{u})] \cdot [\nabla_{\mathcal{D}_m} u_{\mathcal{D}_m} - \nabla \overline{u}] \to 0 \text{ in } L^1(\Omega), \tag{3.103}$$

and therefore a.e. for a sub-sequence. Then, thanks to the strict monotonicity assumption (3.98), we may use Lemma 3.37 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 (3.101) and (3.102),

$$\lim_{m \to \infty} \int_{\Omega} \boldsymbol{a}(\boldsymbol{x}, \Pi_{\mathcal{D}_m} u_{\mathcal{D}_m}, \nabla_{\mathcal{D}_m} u_{\mathcal{D}_m}(\boldsymbol{x})) \cdot \nabla_{\mathcal{D}_m} u_{\mathcal{D}_m}(\boldsymbol{x}) d\boldsymbol{x} = \int_{\Omega} \boldsymbol{a}(\boldsymbol{x}, \overline{u}, \nabla \overline{u}(\boldsymbol{x})) \cdot \nabla \overline{u}(\boldsymbol{x}) d\boldsymbol{x}. \tag{3.104}$$

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 3.38 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 (3.96b), 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 3.37. *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*

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

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

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

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

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

$$\begin{aligned} &(\boldsymbol{b}(u_m, \beta_m) - \boldsymbol{b}(u_m, \beta)) \cdot \frac{\beta_m - \beta}{s_m} \\ &= h_{\delta_m,m}(s_m) \geq h_{\delta_m,m}(\varepsilon) \\ &= (\boldsymbol{b}(u_m, \beta + \varepsilon\delta_m) - \boldsymbol{b}(u_m, \beta)) \cdot \delta_m. \end{aligned}$$

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

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

which is impossible. ∎

The following result is classical (see [62]). Its proof is given for the sake of completeness.

Lemma 3.38. *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 $m \to \infty$. ∎

Part II

Parabolic problems

4

Time-dependent problems: gradient discretisation method and discrete functional analysis

In this chapter, we first give the definition of gradient discretisations (GDs) for time-dependent problems. We then present compactness results for the analysis of such problems. These results include discrete Ascoli–Arzelà and Aubin–Simon theorems, and are presented first in an general setting, before their consequences for gradient discretisations are discussed.

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"), and $\theta = \frac{1}{2}$ provides the Crank-Nicolson scheme. Implicit schemes correspond to $\theta \in [\frac{1}{2}, 1]$, and are the most frequently considered in this book due to the parabolic nature of the equations under study.
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 $W_\bullet^{1,p}(\Omega)$ for $W_0^{1,p}(\Omega)$ in the case of homogeneous Dirichlet boundary conditions, and $W^{1,p}(\Omega)$ in the case of other boundary conditions.

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 $\mathcal{D}_T = (\mathcal{D}, \mathcal{I}_\mathcal{D}, (t^{(n)})_{n=0,\ldots,N})$ is a space–time gradient discretisation if*

- $\mathcal{D} = (X_{\mathcal{D},\bullet}, \Pi_{\mathcal{D}}, \nabla_{\mathcal{D}}, \ldots)$ *is a GD in the sense of Definition 2.1 (resp. Definition 2.18, 2.24, 2.32, 2.48 or 2.51 – 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,*
- $t^{(0)} = 0 < t^{(1)} \ldots < t^{(N)} = T$.

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* $n = 0, \ldots, N-1$, *and* $\eth t_{\mathcal{D}} = \max_{n=0,\ldots,N-1} \eth t^{(n+\frac{1}{2})}$. *To a family* $v = (v^{(n)})_{n=0,\ldots,N} \in X_{\mathcal{D},\bullet}^{N+1}$ *we associate the 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*

$$\begin{aligned} &\forall n = 0, \ldots, N-1, \text{ for all } t \in (t^{(n)}, t^{(n+1)}], \\ &\quad v_\theta(t) = v^{(n+\theta)} := \theta v^{(n+1)} + (1-\theta) v^{(n)} \text{ and, for a.e. } \boldsymbol{x} \in \Omega, \\ &\quad \Pi_{\mathcal{D}}^{(\theta)} v(\boldsymbol{x}, t) = \Pi_{\mathcal{D}}[v_\theta(t)](\boldsymbol{x}), \ \nabla_{\mathcal{D}}^{(\theta)} v(\boldsymbol{x}, t) = \nabla_{\mathcal{D}}[v_\theta(t)](\boldsymbol{x}) \text{ and} \\ &\quad \mathbb{T}_{\mathcal{D}}^{(\theta)} v(\boldsymbol{x}, t) = \mathbb{T}_{\mathcal{D}}[v_\theta(t)](\boldsymbol{x}). \end{aligned} \tag{4.2}$$

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*

$$\begin{aligned} &\forall n = 0, \ldots, N-1, \text{ for a.e. } t \in (t^{(n)}, t^{(n+1)}), \\ &\quad \delta_{\mathcal{D}} v(t) = \delta_{\mathcal{D}}^{(n+\frac{1}{2})} v := \frac{\Pi_{\mathcal{D}} v^{(n+1)} - \Pi_{\mathcal{D}} v^{(n)}}{\eth t^{(n+\frac{1}{2})}}. \end{aligned} \tag{4.4}$$

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), (2.14), (2.20), (2.49) or (2.54) (depending on the considered boundary conditions), where $W_\bullet^{1,p}(\Omega)$ has been replaced with $W_\bullet^{1,p}(\Omega) \cap L^2(\Omega)$ and $\|\Pi_{\mathcal{D}} v - \varphi\|_{L^p(\Omega)}$ has been 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

Definition 4.3 (Space–time-consistency for space–time GD) (cont)

1. It holds
$$\forall \varphi \in W^{1,p}_{\bullet}(\Omega) \cap L^2(\Omega)\,, \ \lim_{m\to\infty} \widehat{S}_{\mathcal{D}_m}(\varphi) = 0.$$
2. It holds
$$\forall u \in L^2(\Omega), \ \lim_{m\to\infty} \|u - \Pi_{\mathcal{D}_m}\mathcal{I}_{\mathcal{D}_m} u\|_{L^2(\Omega)} = 0. \tag{4.5}$$
3. $\delta t_{\mathcal{D}_m} \to 0$ as $m \to \infty$.

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

$$\begin{aligned} &\forall u \in L^2(\Omega)\,, \\ &\mathcal{I}_{\mathcal{D}} u = \operatorname{argmin}\{\|v\|_{\mathcal{D}} : v \in X_{\mathcal{D},\bullet}\,,\ \Pi_{\mathcal{D}} v = \operatorname{Pr}_{\Pi_{\mathcal{D}}(X_{\mathcal{D},\bullet})} u\}, \end{aligned} \tag{4.6}$$

where $\operatorname{Pr}_{\Pi_{\mathcal{D}}(X_{\mathcal{D},\bullet})} : L^2(\Omega) \to \Pi_{\mathcal{D}}(X_{\mathcal{D},\bullet})$ 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 Chapter 2, 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 $\Pi_{\mathcal{D}_m}\mathcal{I}_{\mathcal{D}_m} = \operatorname{Pr}_{\Pi_{\mathcal{D}_m}(X_{\mathcal{D}_m,\bullet})}$, and thus, by the properties of the orthogonal projector,

$$\begin{aligned} \|u - \Pi_{\mathcal{D}_m}\mathcal{I}_{\mathcal{D}_m} u\|_{L^2(\Omega)} &= \left\|u - \operatorname{Pr}_{\Pi_{\mathcal{D}_m}(X_{\mathcal{D}_m,\bullet})} u\right\|_{L^2(\Omega)} \\ &\le \|u - \Pi_{\mathcal{D}_m} u_m\|_{L^2(\Omega)} \to 0 \text{ as } m \to \infty. \end{aligned}$$

Hence, (4.5) holds for $u \in W^{1,p}_{\bullet}(\Omega) \cap L^2(\Omega)$. Since the mapping $\Pi_{\mathcal{D}_m}\mathcal{I}_{\mathcal{D}_m} = \operatorname{Pr}_{\Pi_{\mathcal{D}_m}(X_{\mathcal{D}_m,\bullet})} : L^2(\Omega) \to L^2(\Omega)$ 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 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.

The following lemma is the counterpart of Lemma 2.12 and Lemma 2.40. 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 2.23 or Lemma 2.57).

Lemma 4.7 (Regularity of the limit, space–time problems). *Let $p \in (1,\infty)$ and $((\mathcal{D}_T)_m)_{m\in\mathbb{N}}$ be a coercive and 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 $u \in L^q(0,T;W_\bullet^{1,p}(\Omega))$ such that, up to a subsequence, $\Pi_{\mathcal{D}_m}^{(\theta)} u_m \to u$ weakly in $L^q(0,T;L^p(\Omega))$ and $\nabla_{\mathcal{D}_m}^{(\theta)} u_m \to \nabla u$ 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, $\mathbb{T}_{\mathcal{D}_m}^{(\theta)} u_m \to \gamma u$ 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 4.8. 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 (2.18) 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

$$\left(\int_0^T \|(u_m)_\theta(t)\|_{\mathcal{D}_m}^q \,\mathrm{d}t\right)^{1/q}.$$

Proof. We only consider the case of homogeneous Dirichlet boundary conditions, the other case being handled similarly by following the proof of Lemma 2.40.
By coercivity of $((\mathcal{D}_T)_m)_{m\in\mathbb{N}}$,

$$\begin{aligned}\|(u_m)_\theta\|_{L^q(0,T;X_{\mathcal{D}_m,\bullet})} &= \left\|\nabla^{(\theta)}_{\mathcal{D}_m} u_m\right\|_{L^q(0,T;L^p(\Omega))^d} \\ &\geq \frac{1}{C_P}\left\|\Pi^{(\theta)}_{\mathcal{D}_m} u_m\right\|_{L^q(0,T;L^p(\Omega))}.\end{aligned}$$

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 completed by showing that $\boldsymbol{v}=\nabla u$ in the sense of distributions on $\mathbb{R}^d\times(0,T)$.
Let $\boldsymbol{\varphi}\in C^\infty_c(\mathbb{R}^d)^d$ and $\psi\in C^\infty_c(0,T)$. We drop the indices m for legibility. We have, for $t\in(0,T)$, by definition (2.6) of $W_{\mathcal{D}}$,

$$\int_\Omega \Big[\nabla_{\mathcal{D}}[u_\theta(t)](\boldsymbol{x})\cdot\boldsymbol{\varphi}(\boldsymbol{x}) + \Pi_{\mathcal{D}}[u_\theta(t)](\boldsymbol{x})\mathrm{div}\boldsymbol{\varphi}(\boldsymbol{x})\Big]\mathrm{d}\boldsymbol{x} \leq \|u_\theta(t)\|_{X_{\mathcal{D},0}} W_{\mathcal{D}}(\varphi).$$

Multiply this by $\psi(t)$, integrate over $t\in(0,T)$ and use Hölder's inequality:

$$\begin{aligned}&\int_0^T\int_\Omega \Big[\nabla^{(\theta)}_{\mathcal{D}} u(\boldsymbol{x},t)\cdot(\psi(t)\boldsymbol{\varphi}(\boldsymbol{x})) + \Pi^{(\theta)}_{\mathcal{D}} u(\boldsymbol{x},t)\mathrm{div}(\psi(t)\boldsymbol{\varphi})(\boldsymbol{x})\Big]\mathrm{d}\boldsymbol{x}\mathrm{d}t \\ &\leq W_{\mathcal{D}}(\varphi)\int_0^T \|u_\theta(t)\|_{X_{\mathcal{D},0}}\psi(t)\mathrm{d}t \leq W_{\mathcal{D}}(\varphi)\,\|u_\theta\|_{L^q(0,T;X_{\mathcal{D},0})}\,\|\psi\|_{L^{q'}(0,T)}.\end{aligned}$$

Changing ψ into $-\psi$ shows that the same inequality is satisfied if the left-hand side is replaced with its absolute value. Since $(\|u_\theta\|_{L^q(0,T;X_{\mathcal{D},0})})_{m\in\mathbb{N}}$ is bounded and, by limit-conformity, $W_{\mathcal{D}}(\varphi)\to 0$ as $m\to\infty$, we can 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 $(\boldsymbol{x},t)\to\psi(t)\varphi(\boldsymbol{x})$, 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 $C^\infty_c(\mathbb{R}^d\times(0,T))^d$, see [29, Appendix D]. This shows that, for all $\boldsymbol{\Phi}\in C^\infty_c(\mathbb{R}^d\times(0,T))^d$,

$$\int_0^T\int_\Omega \Big[\boldsymbol{v}(\boldsymbol{x},t)\cdot\boldsymbol{\Phi}(\boldsymbol{x},t) + u(\boldsymbol{x},t)\mathrm{div}(\boldsymbol{\Phi}(\cdot,t))(\boldsymbol{x})\Big]\mathrm{d}\boldsymbol{x}\mathrm{d}t = 0.$$

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 their gradient, with reconstructed functions and gradients built from space–time-consistent GDs.

Lemma 4.9 (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, which is space–time-consistent in the sense of Definition 4.3. Let $\overline{v} \in L^p(0,T;W_\bullet^{1,p}(\Omega))$. Then:*

1. *There exists a sequence $(v_m)_{m\in\mathbb{N}}$ such that $v_m = (v_m^{(n)})_{n=0,\dots,N_m} \in X_{\mathcal{D}_m,\bullet}^{N_m+1}$ for all $m \in \mathbb{N}$, and, as $m \to \infty$,*

$$\Pi_{\mathcal{D}_m}^{(\theta)} v_m \to \overline{v} \text{ strongly in } L^p(\Omega\times(0,T)), \tag{4.7a}$$

$$\nabla_{\mathcal{D}_m}^{(\theta)} v_m \to \nabla\overline{v} \text{ strongly in } L^p(\Omega\times(0,T))^d. \tag{4.7b}$$

2. *In the case of non-homogeneous Neumann boundary conditions, if the sequence of underlying spatial discretisations is coercive and limit–conforming in the sense of Definitions 2.33 and 2.34, then the sequence $(v_m)_{m\in\mathbb{N}}$ in Item 1 also satisfies*

$$\mathbb{T}_{\mathcal{D}_m}^{(\theta)} v_m \to \gamma\overline{v} \text{ weakly in } L^p(\partial\Omega\times(0,T)) \text{ as } m\to\infty. \tag{4.8}$$

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*

$$\mathbb{T}_{\mathcal{D}_m}^{(\theta)} v_m \to \gamma\overline{v} \text{ strongly in } L^p(\partial\Omega\times(0,T)) \text{ as } m\to\infty. \tag{4.9}$$

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 [29, 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 2.13).

Given the structure (4.11) of functions in $\mathcal{T}(0,T;W^{1,p}_\bullet(\Omega))$, 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 $w \in W^{1,p}_\bullet(\Omega)$. Let $v_m \in X^{N_m+1}_{\mathcal{D}_m,\bullet}$ be defined by $v_m^{(n)} = \varphi(t^{(n)})P_{\mathcal{D}_m}w$ for $n = 0,\dots,N_m$, where

$$P_{\mathcal{D}_m}w = \operatorname*{argmin}_{z\in X_{\mathcal{D}_m,\bullet}} \left(\|\Pi_{\mathcal{D}_m}z - w\|_{L^{\max(p,2)}(\Omega)} + \|\nabla_{\mathcal{D}_m}z - \nabla w\|_{L^p(\Omega)^d} \right). \tag{4.12}$$

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 $\Pi^{(\theta)}_{\mathcal{D}_m}$, for all $t \in (0,T)$ and a.e. $\boldsymbol{x} \in \Omega$,

$$\begin{aligned}\overline{v}(\boldsymbol{x},t) - \Pi^{(\theta)}_{\mathcal{D}_m}v_m(\boldsymbol{x},t) &= \varphi(t)w(\boldsymbol{x}) - \Phi_m(t)\Pi_{\mathcal{D}_m}(P_{\mathcal{D}_m}w)(\boldsymbol{x}) \\ &= [\varphi(t) - \Phi_m(t)]w(\boldsymbol{x}) + \Phi_m(t)[w(\boldsymbol{x}) - \Pi_{\mathcal{D}_m}(P_{\mathcal{D}_m}w)(\boldsymbol{x})].\end{aligned} \tag{4.13}$$

Using the definitions of $\widehat{S}_{\mathcal{D}_m}$ and $P_{\mathcal{D}_m}$, we infer that

$$\begin{aligned}&\|\overline{v} - \Pi^{(\theta)}_{\mathcal{D}_m}v_m\|_{L^p(\Omega\times(0,T))} \\ &\le \|\varphi - \Phi_m\|_{L^p(0,T)}\|w\|_{L^p(\Omega)} + \|\Phi_m\|_{L^p(0,T)}\|w - \Pi_{\mathcal{D}_m}(P_{\mathcal{D}_m}w)\|_{L^p(\Omega)} \\ &\le \|\varphi - \Phi_m\|_{L^p(0,T)}\|w\|_{L^p(\Omega)} + \|\Phi_m\|_{L^p(0,T)}\widehat{S}_{\mathcal{D}_m}(w).\end{aligned} \tag{4.14}$$

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 4.7 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

$$\begin{aligned}P_{\mathcal{D}_m}w = \operatorname{argmin}_{z\in X_{\mathcal{D}_m}} \Big(&\|\Pi_{\mathcal{D}_m}z - w\|_{L^{\max(p,2)}(\Omega)} + \|\nabla_{\mathcal{D}_m}z - \nabla w\|_{L^p(\Omega)^d} \\ &+ \|\mathbb{T}_{\mathcal{D}_m}z - \gamma w\|_{L^p(\partial\Omega)} \Big)\end{aligned}$$

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 $\|\mathbb{T}_{\mathcal{D}_m}(P_{\mathcal{D}_m}w) - \gamma w\|_{L^p(\partial\Omega)} \le \widehat{S}_{\mathcal{D}_m}(w)$. 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 [29, 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 $n \to \infty$,

$$\overline{v}_n \to \overline{v} \text{ in } L^p(0,T;W^{1,p}_\bullet(\Omega)) \cap C([0,T];L^2(\Omega)), \text{ and}$$
$$\partial_t \overline{v}_n \to \partial_t \overline{v} \text{ in } L^2(0,T;L^2(\Omega)).$$

The proof of [29, 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 [29, 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 $i = \ell_\tau - 3, \dots, \ell_\tau$ 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 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, $v_m^{(n)} = \varphi(t^{(n)})P_{\mathcal{D}_m} w$ for $n = 0,\dots,N_m$. 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 $\mathit{\Pi}^{(\theta)}_{\mathcal{D}_m} v_m$ at $t = 0$,

$$\begin{aligned} u(\boldsymbol{x},0) - \mathit{\Pi}^{(\theta)}_{\mathcal{D}_m} v_m(\boldsymbol{x},0) &= \varphi(0)w(\boldsymbol{x}) - \varphi(t^{(0)})\mathit{\Pi}_{\mathcal{D}_m}(P_{\mathcal{D}_m} w)(\boldsymbol{x}) \\ &= \varphi(0)[w(\boldsymbol{x}) - \mathit{\Pi}_{\mathcal{D}_m}(P_{\mathcal{D}_m} w)(\boldsymbol{x})]. \end{aligned}$$

Then, by definition of $P_{\mathcal{D}_m}$ and $\widehat{S}_{\mathcal{D}_m}$,

$$\left\|\overline{v}(0) - \mathit{\Pi}^{(\theta)}_{\mathcal{D}_m} v_m(0)\right\|_{L^2(\Omega)} = |\varphi(0)|\,\|w - \mathit{\Pi}_{\mathcal{D}_m}(P_{\mathcal{D}_m} w)\|_{L^2(\Omega)} \leq |\varphi(0)|\widehat{S}_{\mathcal{D}_m}(w)$$

and (4.10b) follows from the space–time-consistency of $((\mathcal{D}_T)_m)_{m\in\mathbb{N}}$.
To establish (4.10c), we define $\mathit{\Psi}_m : (0,T] \to \mathbb{R}$ 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 $n = 0,\dots,N_m-1$.
Then, by definition (4.4) of $\delta_{\mathcal{D}_m} v_m$, for all $t \in (0,T)$ and a.e. $\boldsymbol{x} \in \Omega$,

$$\partial_t \overline{v}(\boldsymbol{x},t) - \delta_{\mathcal{D}_m} v_m(\boldsymbol{x},t) = \varphi'(t)w(\boldsymbol{x}) - \mathit{\Psi}_m(t)\mathit{\Pi}_{\mathcal{D}_m}(P_{\mathcal{D}_m} w)(\boldsymbol{x}).$$

By smoothness of φ we have $\mathit{\Psi}_m \to \varphi'$ 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 $(\overline{v}, \mathit{\Pi}_{\mathcal{D}_m} v_m, \varphi, \mathit{\Phi}_m)$ with $(\partial_t\overline{v}, \delta_{\mathcal{D}_m} v_m, \varphi', \mathit{\Psi}_m)$, and the L^p norms with L^2 norms (note that $\widehat{S}_{\mathcal{D}_m}(w) \geq \|w - \mathit{\Pi}_{\mathcal{D}_m}(P_{\mathcal{D}_m} w)\|_{L^2(\Omega)}$ by definition of $\widehat{S}_{\mathcal{D}}$). ∎

4.2 Averaged-in-time compactness

4.2.1 Abstract setting

In this section, we gather a few notions and compactness results used in the convergence analysis of numerical methods for time-dependent problems.
The first two theorems are generalisations to vector-valued Lebesgue spaces of the classical Kolmogorov compactness theorem for L^p spaces [13]. 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. [29, 45] for a definition and some properties of these spaces.

Theorem 4.10 (Kolmogorov (1)). *Let B be a Banach space, $1 \leq p < +\infty$, $T > 0$ 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 $\|Pf\|_{L^p(\mathbb{R};B)} \leq C$, 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 4.11 (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 $(\rho_m)_{m\in\mathbb{N}}$ be a sequence of mollifiers constructed by scaling a given smooth function ρ, that is:

$$\begin{aligned} &\rho \in C_c^\infty(-1,1),\ \int_{\mathbb{R}} \rho \mathrm{d}t = 1,\ \rho \geq 0,\ \rho(-t) = \rho(t) \text{ for all } t \in \mathbb{R} \\ &\text{and, for all } m \in \mathbb{N} \text{ and } t \in \mathbb{R},\ \rho_m(t) = m\rho(mt). \end{aligned} \tag{4.16}$$

We set $K = [0,T]$ and, for $m \in \mathbb{N}$, $A_m = \{(Pf \star \rho_m)_{|K},\ f \in A\}$ where $\star$ denotes the convolution product in $\mathbb{R}$.
The proof is divided in two steps. In Step 1 we prove, using Ascoli's theorem and Assumption 2, that, for $m \in \mathbb{N}$, 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 $m \in \mathbb{N}$. In order to prove that A_m is relatively compact in $C(K;B)$, we use Ascoli–Arzelà's theorem. 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,

$$\begin{aligned}\|Pf\star\rho_m(t_2)-Pf\star\rho_m(t_1)\|_B &\\ &\leq \int_{\mathbb{R}} \|Pf(s)\|_B \,|\rho_m(t_2-s)-\rho_m(t_1-s)|\,\mathrm{d}s\\ &\leq \|Pf\|_{L^p(\mathbb{R};B)}\,\|\rho_m(t_2-\cdot)-\rho_m(t_1-\cdot)\|_{L^{p'}(\mathbb{R})}\,.\end{aligned}$$

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 $\overline{s}=ms$,

$$\begin{aligned}Pf\star\rho_m(t)-Pf(t) &= \int_{\mathbb{R}}\left[Pf(t-s)-Pf(t)\right]\rho_m(s)\mathrm{d}s\\ &= \int_{-1}^{1}\left[Pf\left(t-\frac{\overline{s}}{m}\right)-Pf(t)\right]\rho(\overline{s})\mathrm{d}\overline{s}.\end{aligned}$$

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

$$\begin{aligned}\|Pf\star\rho_m-Pf\|_{L^p(0,T;B)}^p &\\ &\leq \|\rho\|_{L^{p'}}^p\int_{-1}^{1}\left\|Pf\left(\cdot-\frac{\overline{s}}{m}\right)-Pf\right\|_{L^p(0,T;B)}^p\mathrm{d}\overline{s}\\ &\leq 2\,\|\rho\|_{L^{p'}}^p\sup\left\{\|Pf(\cdot+h)-Pf\|_{L^p(\mathbb{R};B)}^p\,:\,|h|\leq\frac{1}{m}\right\}.\end{aligned}$$

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

$$\left\|Pf \star \rho_{m(\varepsilon)} - Pf\right\|_{L^p(0,T;B)} \leq \varepsilon/2 \quad \text{for all } f \in A. \tag{4.17}$$

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 (4.17) 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 4.12 (Kolmogorov (2)). *Let B be a Banach space, $1 \leq p < +\infty$, $T > 0$ 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 $f \in A$,*

$$\int_0^{T-h} \|f(t+h) - f(t)\|_B^p \,\mathrm{d}t \leq \eta(h).$$

Proof.
The proof uses Theorem 4.10 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 4.12, Items 1 and 2 of Theorem 4.10 are clearly satisfied.
We prove now prove, in two steps, Item 3 of Theorem 4.10. 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 $\tau + h \leq T$. For all $t \in (0,\tau)$ one has $\|f(t)\|_B \leq \|f(t+h)\|_B + \|f(t+h) - f(t)\|_B$ and thus, by the power-of-sums inequality (C.12),

$$\|f(t)\|_B^p \leq 2^{p-1} \|f(t+h)\|_B^p + 2^{p-1} \|f(t+h) - f(t)\|_B^p .$$

Integrating this inequality for $t \in (0,\tau)$ gives

$$\int_0^\tau \|f(t)\|_B^p \,\mathrm{d}t \leq 2^{p-1} \int_0^\tau \|f(t+h)\|_B^p \,\mathrm{d}t + 2^{p-1} \int_0^\tau \|f(t+h) - f(t)\|_B^p \,\mathrm{d}t. \tag{4.18}$$

Now let $h_0 \in (0,T)$ et $\tau \in (0, T-h_0)$. For all $h \in (0,h_0)$, Inequality (4.18) gives, using $\eta(h) \leq \eta(h_0)$,

$$\int_0^\tau \|f(t)\|_B^p \,\mathrm{d}t \le 2^{p-1} \int_0^\tau \|f(t+h)\|_B^p \,\mathrm{d}t + 2^{p-1}\eta(h_0).$$

Integrating this inequality over $h \in (0, h_0)$ leads to

$$h_0 \int_0^\tau \|f(t)\|_B^p \,\mathrm{d}t \le 2^{p-1} \int_0^{h_0} \left(\int_0^\tau \|f(t+h)\|_B^p \,\mathrm{d}t \right) dh + 2^{p-1} h_0 \eta(h_0). \quad (4.19)$$

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 (4.19),

$$\int_0^\tau \|f(t)\|_B^p \,\mathrm{d}t \le \frac{\tau 2^{p-1}}{h_0} \|f\|_{L^p(0,T;B)}^p + 2^{p-1}\eta(h_0).$$

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 4.10 is satisfied, and thus conclude the proof of Theorem 4.12.
Recall that $Pf(t) = 0$ if $t \notin [0, T]$ so that, for all $h \in (0, T)$ and $f \in A$,

$$\begin{aligned} &\int_{\mathbb{R}} \|Pf(t+h) - Pf(t)\|_B^p \mathrm{d}t \\ &\le \int_{-h}^0 \|f(t+h)\|_B^p \,\mathrm{d}t + \int_0^{T-h} \|f(t+h) - f(t)\|_B^p \,\mathrm{d}t + \int_{T-h}^T \|f(t)\|_B^p \,\mathrm{d}t \\ &\le \int_0^h \|f(t)\|_B^p \,\mathrm{d}t + \eta(h) + \int_{T-h}^T \|f(t)\|_B^p \,\mathrm{d}t. \end{aligned} \quad (4.20)$$

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 (4.20), for all $f \in A$ and $h \le \min(h_1, h_2)$,

$$\int_{\mathbb{R}} \|Pf(t+h) - Pf(t)\|_B^p \,\mathrm{d}t \le 3\varepsilon.$$

This concludes the proof that Assumption 3 in Theorem 4.10 is satisfied. ∎

We now turn to compactness theorems involving sequences of spaces as co-domains 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 space in a fixed Banach space.

Definition 4.13 (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. We say that 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 results for sequences of subspaces is a straightforward translation in that setting of the second Kolmogorov theorem above.

Proposition 4.14 (Time compactness with a sequence of subspaces). *Let $1 \le p < +\infty$, $T > 0$, B be a Banach space, and $(X_m)_{m\in\mathbb{N}}$ be compactly embedded in B (see Definition 4.13). 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 $m \in \mathbb{N}$,*

$$\int_0^{T-h} \|f_m(t+h) - f_m(t)\|_B^p \,\mathrm{d}t \le \eta(h).$$

Then, the sequence $(f_m)_{m\in\mathbb{N}}$ is relatively compact in $L^p(0,T;B)$.

Proof. We aim at applying Theorem 4.12 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[58] and generalised in [57], see also [56].
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 4.15 (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 4.13).*
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 4.16. *Let B be a Banach space and $(X_m, Y_m)_{m\in\mathbb{N}}$ be compactly–continuously embedded in B (see Definition 4.15). 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 that there exists $\varepsilon > 0$ such that, for any $m_0 \in \mathbb{N}$, we can find $m = \varphi(m_0) \geq m_0$ and $w \in X_{\varphi(m_0)}$ such that

$$\|w\|_B > \varepsilon \|w\|_{X_{\varphi(m_0)}} + m_0 \|w\|_{Y_{\varphi(m_0)}} .$$

There is 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

$w \neq 0$, we can then set $u_{\varphi(m_0)} = \frac{w}{\|w\|_B} \in X_{\varphi(m_0)}$ (there is no ambiguity in the definition of $u_{\varphi(m_0)}$ since φ is one-to-one). We then have, for any $m \in \varphi(\mathbb{N})$,

$$1 = \|u_m\|_B \geq \varepsilon \|u_m\|_{X_m} + \psi(m) \|u_m\|_{Y_m}, \tag{4.21}$$

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 (4.21) 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 (4.21), $(\|u_m\|_{X_m})_{m\in\mathbb{N}}$ 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 (4.21) again, $\|u_m\|_{Y_m} \leq 1/\psi(m) \to 0$ as $m \to +\infty$, and thus, by Definition 4.15, the limit of $(u_m)_{m\in\mathbb{N}}$ in B must be 0. This contradicts (4.21) 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 4.17 (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 4.15). 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 $\delta t^{(n+\frac{1}{2})} = t^{(n+1)} - t^{(n)}$,

$$\delta_m f_m(t) = \frac{v^{(n+1)} - v^{(n)}}{\delta t^{(n+\frac{1}{2})}} \text{ for } n \in \{0, \ldots, N-1\} \text{ and } t \in (t^{(n)}, t^{(n+1)}).$$

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 4.14. 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 "unformly 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

$$\forall m \geq m_0,\, \forall h \in (0,h_0)\,,\ \int_0^{T-h} \|f_m(\cdot+h) - f_m\|_B^p \,\mathrm{d}t \leq \eta. \tag{4.22}$$

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 4.16 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 (C.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

$$\begin{aligned}
\int_0^{T-h} \|f_m(t+h) - f_m(t)\|_B^p \,\mathrm{d}t \leq\ & 2 \times 3^{p-1}\varepsilon^p \|f_m\|_{L^p(0,T;X_m)}^p \\
&+ 3^{p-1}C_\varepsilon^p \int_0^{T-h} \|f_m(t+h) - f_m(t)\|_{Y_m}^p \,\mathrm{d}t.
\end{aligned} \tag{4.23}$$

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

$$\begin{aligned}
f_m(t+h) - f_m(t) &= \sum_{n:\, t^{(n)} \in (t,t+h)} (f_m)_{|(t^{(n)}, t^{(n+1)})} - (f_m)_{|(t^{(n-1)}, t^{(n)})} \\
&= \sum_{n:\, t^{(n)} \in (t,t+h)} (\theta v^{(n+1)} + (1-\theta)v^{(n)}) - (\theta v^{(n)} + (1-\theta)v^{(n-1)}) \\
&= \sum_{n:\, t^{(n)} \in (t,t+h)} \Big[\theta(v^{(n+1)} - v^{(n)}) + (1-\theta)(v^{(n)} - v^{(n-1)})\Big] \\
&= \sum_{n=1}^{N-1} \Big[\theta(v^{(n+1)} - v^{(n)}) + (1-\theta)(v^{(n)} - v^{(n-1)})\Big]\, \mathbf{1}_{(t,t+h)}(t^{(n)})
\end{aligned}$$

$$= \theta \sum_{n=1}^{N-1} \frac{v^{(n+1)} - v^{(n)}}{\delta t^{(n+\frac{1}{2})}} \delta t^{(n+\frac{1}{2})} \mathbf{1}_{(t,t+h)}(t^{(n)}) + (1-\theta) \sum_{n=1}^{N-1} \frac{v^{(n)} - v^{(n-1)}}{\delta t^{(n-\frac{1}{2})}} \delta t^{(n-\frac{1}{2})} \mathbf{1}_{(t,t+h)}(t^{(n)}), \tag{4.24}$$

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 $m \in \mathbb{N}$,

$$\sum_{n=0}^{N-1} \left\| \frac{v^{(n+1)} - v^{(n)}}{\delta t^{(n+\frac{1}{2})}} \right\|_{Y_m} \delta t^{(n+\frac{1}{2})} \leq M.$$

Taking the Y_m-norm of (4.24), then the power p, and using the convexity of $s \to s^p$ gives

$$\begin{aligned}
&\|f_m(t+h) - f_m(t)\|_{Y_m}^p \\
&\leq \theta \left(\sum_{n=1}^{N-1} \left\| \frac{v^{(n+1)} - v^{(n)}}{\delta t^{(n+\frac{1}{2})}} \right\|_{Y_m} \delta t^{(n+\frac{1}{2})} \mathbf{1}_{(t,t+h)}(t^{(n)}) \right)^p \\
&\quad + (1-\theta) \left(\sum_{n=1}^{N-1} \left\| \frac{v^{(n)} - v^{(n-1)}}{\delta t^{(n-\frac{1}{2})}} \right\|_{Y_m} \delta t^{(n-\frac{1}{2})} \mathbf{1}_{(t,t+h)}(t^{(n)}) \right)^p \\
&\leq \theta M^{p-1} \left(\sum_{n=1}^{N-1} \left\| \frac{v^{(n+1)} - v^{(n)}}{\delta t^{(n+\frac{1}{2})}} \right\|_{Y_m} \delta t^{(n+\frac{1}{2})} \mathbf{1}_{(t,t+h)}(t^{(n)}) \right) \\
&\quad + (1-\theta) M^{p-1} \left(\sum_{n=1}^{N-1} \left\| \frac{v^{(n)} - v^{(n-1)}}{\delta t^{(n-\frac{1}{2})}} \right\|_{Y_m} \delta t^{(n-\frac{1}{2})} \mathbf{1}_{(t,t+h)}(t^{(n)}) \right).
\end{aligned} \tag{4.25}$$

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

$$\int_0^{T-h} \|f_m(t+h) - f_m(t)\|_{Y_m}^p \, \mathrm{d}t \leq M^p h. \tag{4.26}$$

Plugging this inequality into (4.23), we obtain

$$\int_0^{T-h} \|f_m(t+h) - f_m(t)\|_B^p \, \mathrm{d}t \leq 2 \times 3^{p-1} \varepsilon^p \|f_m\|_{L^p(0,T;X_m)}^p + 3^{p-1} C_\varepsilon^p M^p h. \tag{4.27}$$

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 (4.27) then shows that (4.22) holds, which proves the theorem. ∎

4.2.2 Application to space–time gradient discretisations

In this section, we apply the previous abstract compactness results to the framework of space–time GD.

Aubin–Simon theorem

The first step is to defined a dual norm $\|w\|_{\star,\mathcal{D}}$ on $\Pi_{\mathcal{D}}(X_{\mathcal{D},\bullet})$, which will enable us to define the spaces Y_m in the above theorems.

Definition 4.18 (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:*

$$\begin{aligned} &\forall w \in \Pi_{\mathcal{D}}(X_{\mathcal{D},\bullet}), \\ &\|w\|_{\star,\mathcal{D}} = \sup\left\{ \int_\Omega w(\boldsymbol{x}) \Pi_{\mathcal{D}} v(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \,:\, v \in X_{\mathcal{D},\bullet}\,,\ \|v\|_{\mathcal{D}} = 1 \right\}. \end{aligned} \tag{4.28}$$

A straightforward consequence of this definition is

$$\forall w \in \Pi_{\mathcal{D}}(X_{\mathcal{D},\bullet})\,,\ \forall v \in X_{\mathcal{D},\bullet}\,,\ \left| \int_\Omega w(\boldsymbol{x}) \Pi_{\mathcal{D}} v(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \right| \le \|w\|_{\star,\mathcal{D}} \|v\|_{\mathcal{D}}\,. \tag{4.29}$$

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 4.19 (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 2.24).

Remark 4.20 ($\|\cdot\|_{\star,\mathcal{D}}$ 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, (4.28) therefore reads: for $w \in L^2(\Omega)$,

$$\|w\|_\star := \sup\left\{ \int_\Omega w(\boldsymbol{x}) v(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \,:\, v \in H_0^1(\Omega)\,,\ \|v\|_{H_0^1(\Omega)} = 1 \right\}. \tag{4.30}$$

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, (4.30) 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 4.17).

Theorem 4.21 (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 sequence of space–time-consistent and compact space–time GDs in the sense of Definitions 4.3 and 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*

$$\forall m \in \mathbb{N},\ \int_0^T \|(v_m)_\theta(t)\|_{\mathcal{D}_m}^p \,\mathrm{d}t \le C \tag{4.31}$$

and

$$\forall m \in \mathbb{N},\ \int_0^T \|\delta_{\mathcal{D}_m} v_m(t)\|_{\star,\mathcal{D}_m} \,\mathrm{d}t \le C. \tag{4.32}$$

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 4.17, let $B = L^p(\Omega)$, $X_m = \Pi_{\mathcal{D}_m}(X_{\mathcal{D}_m,\bullet})$, and define the norm on X_m by

$$\|u\|_{X_m} = \min\{\|w\|_{\mathcal{D}_m} : w \in X_{\mathcal{D}_m,\bullet} \text{ such that } \Pi_{\mathcal{D}_m} w = u\}. \tag{4.33}$$

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 4.15. First, the compactness hypothesis on $(\mathcal{D}_m)_{m\in\mathbb{N}}$ is exactly stating that $(X_m)_{m\in\mathbb{N}}$ is compactly embedded in B, in the sense of Definition 4.13. 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 (4.29) 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 $L^p(\Omega)$ of $(u_m)_{m\in\mathbb{N}}$ must be 0. The proof $(X_m, Y_m)_{m\in\mathbb{N}}$ is compactly–continuously embedded in $B = L^p(\Omega)$ is complete.

The relative compactness of $(\Pi_{\mathcal{D}_m}^{(\theta)} v_m)_{m\in\mathbb{N}}$ in $L^p(0,T;L^p(\Omega))$ follows from Theorem 4.17 with $f_m = \Pi_{\mathcal{D}_m}^{(\theta)} v_m$ if we can check the four assumptions stated in this theorem.
The first of this assumption is obviously satisfied by the definition of $\Pi_{\mathcal{D}_m}^{(\theta)}$ in (4.2).
Since the sequence of underlying spatial discretisations is compact, it is also coercive (see, e.g., Lemma 2.9 for homogeneous Dirichlet boundary conditions). The definition of $C_{\mathcal{D}_m}$ combined with (4.31) and the definition of $\Pi_{\mathcal{D}_m}^{(\theta)}$ then shows that $(\Pi_{\mathcal{D}_m}^{(\theta)} v_m)_{m\in\mathbb{N}}$ is bounded in $L^p(0,T;L^p(\Omega))$. This takes care of the second assumption in Theorem 4.17.
The third assumption follows immediately from (4.31) and the fact that

$$\left\|\Pi_{\mathcal{D}_m}^{(\theta)} v_m(t)\right\|_{X_m} = \|\Pi_{\mathcal{D}_m}((v_m)_\theta(t))\|_{X_m} \le \|(v_m)_\theta(t)\|_{\mathcal{D}_m}\,.$$

To prove the fourth assumption in Theorem 4.17, we notice that

$$\left\|\delta_m \Pi_{\mathcal{D}_m}^{(\theta)} v_m(t)\right\|_{Y_m} = \|\delta_{\mathcal{D}_m} v_m(t)\|_{Y_m} = \|\delta_{\mathcal{D}_m} v_m(t)\|_{\star,\mathcal{D}_m}$$

and we use (4.32). ∎

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 4.24, 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 4.24 for $p \neq 2$ requires a different path; see [33, Theorem 5.4] for details.

We first define the inverses of the discrete and continuous Laplace operators with homogeneous Dirichlet boundary conditions.

Definition 4.22 (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 $\Delta_{\mathcal{D}}^i : X_{\mathcal{D},0} \to X_{\mathcal{D},0}$ such that, for all $v \in X_{\mathcal{D},0}$,*

$$\forall w \in X_{\mathcal{D},0}\,,\ \int_\Omega \nabla_{\mathcal{D}}(\Delta_{\mathcal{D}}^i v)(\boldsymbol{x})\cdot\nabla_{\mathcal{D}} w(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = \int_\Omega \Pi_{\mathcal{D}} v(\boldsymbol{x})\Pi_{\mathcal{D}} w(\boldsymbol{x})\mathrm{d}\boldsymbol{x}. \quad (4.34)$$

We also define $\Delta^i : L^2(\Omega) \to H_0^1(\Omega)$ such that, for all $v \in L^2(\Omega)$,

$$\forall w \in H_0^1(\Omega)\,,\ \int_\Omega \nabla(\Delta^i v)(\boldsymbol{x})\cdot\nabla w(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = \int_\Omega v(\boldsymbol{x})w(\boldsymbol{x})\mathrm{d}\boldsymbol{x}. \quad (4.35)$$

Theorem 4.23 (Compactness of Δ^i). *Let $p = 2$, $T > 0$, $\theta \in [0,1]$ and $((\mathcal{D}_T)_m)_{m\in\mathbb{N}} = (\mathcal{D}_m, \mathcal{I}_{\mathcal{D}_m}, (t_m^{(n)})_{n=0,\dots,N_m})_{m\in\mathbb{N}}$ 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 $v_m \in X_{\mathcal{D}_m,0}^{N_m+1}$ be such that there exists $C > 0$ and $q \geq 1$ satisfying*

$$\forall m \in \mathbb{N},\ \int_0^T \left\| \Pi_{\mathcal{D}_m}^{(\theta)} v_m(t) \right\|_{L^2(\Omega)}^2 \mathrm{d}t \leq C \tag{4.36}$$

and

$$\forall m \in \mathbb{N},\ \int_0^T \|\delta_{\mathcal{D}_m} v_m(t)\|_{\star,\mathcal{D}_m}^q \,\mathrm{d}t \leq C. \tag{4.37}$$

We also assume that $\Pi_{\mathcal{D}_m}^{(\theta)} v_m$ 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 $m \to \infty$,

$$\begin{aligned} &\Pi_{\mathcal{D}_m}^{(\theta)}(\Delta_{\mathcal{D}_m}^i v_m) \to \Delta^i v \text{ in } L^2(0,T;L^2(\Omega)), \text{ and} \\ &\nabla_{\mathcal{D}_m}^{(\theta)}(\Delta_{\mathcal{D}_m}^i v_m) \to \nabla(\Delta^i \overline{v}) \text{ in } L^2(0,T;L^2(\Omega))^d. \end{aligned} \tag{4.38}$$

Moreover, if $q > 1$ then

$$\delta_{\mathcal{D}_m}(\Delta_{\mathcal{D}_m}^i v_m) \to \partial_t(\Delta^i \overline{v}) \text{ weakly in } L^q(0,T;L^2(\Omega)) \text{ as } m \to \infty. \tag{4.39}$$

Proof.
Step 1: we prove, using Proposition 4.14 with $p = 2$, that $(\Pi_{\mathcal{D}_m}^{(\theta)}(\Delta_{\mathcal{D}_m}^i v_m))_{m\in\mathbb{N}}$ is relatively compact in $L^2(0,T;L^2(\Omega))$.
Let $u_m = \Delta_{\mathcal{D}_m}^i v_m$. Since the sequence of underlying spatial discretisations is compact, it is coercive (Lemma 2.9). Denote by C_P a coercivity constant of this sequence. Using the definition (4.34) of $\Delta_{\mathcal{D}_m}^i$ 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

$$\int_0^T \|(u_m)_\theta(t)\|_{\mathcal{D}_m}^2 \leq C_P^2 \int_0^T \left\| \Pi_{\mathcal{D}_m}^{(\theta)} v_m(t) \right\|_{L^2(\Omega)}^2 \leq C_P^2 C. \tag{4.40}$$

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 4.13. Estimate (4.40) and the coercivity of $(\mathcal{D}_m)_{m\in\mathbb{N}}$ prove that items 1 and 2 of the hypotheses of Proposition 4.14 hold for $f_m = \Pi_{\mathcal{D}_m}^{(\theta)} u_m$. Let us now observe that, for all $z \in X_{\mathcal{D}_m,0}$,

$\|\Pi_{\mathcal{D}_m} z\|_{\star,\mathcal{D}_m}$

$$= \sup\left\{\int_\Omega \nabla_{\mathcal{D}_m}(\Delta^i_{\mathcal{D}_m} z)(\boldsymbol{x})\cdot\nabla_{\mathcal{D}_m} w(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \,:\, w\in X_{\mathcal{D}_m},\ \|w\|_{\mathcal{D}_m}=1\right\}$$
$$= \left\|\Delta^i_{\mathcal{D}_m} z\right\|_{\mathcal{D}_m}. \tag{4.41}$$

Therefore, since $\delta_{\mathcal{D}_m} u_m(t) = \frac{\Pi_{\mathcal{D}} u_m^{(n+1)} - \Pi_{\mathcal{D}} u_m^{(n)}}{\delta t^{(n+\frac12)}}$, Hypothesis (4.37) and the use of coercivity with constant C_P imply that

$$\forall m\in\mathbb{N},\ \int_0^T \|\delta_{\mathcal{D}_m} u_m(t)\|^q_{L^2(\Omega)}\,\mathrm{d}t \le (C_P)^q C. \tag{4.42}$$

Apply the same computation as in (4.24), followed by (4.25) with Y_m replaced by $L^2(\Omega)$. Using (4.42), this proves that (4.26) holds (still with $L^2(\Omega)$ instead of Y_m). Hence, item 3 of the hypotheses of Proposition 4.14 holds with $\eta(h) = h$. Note that, contrary to the proof of Theorem 4.17, we do not use an inequality similar to (4.23), which is the discrete equivalent of the Lions lemma.
Therefore, Proposition 4.14 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 4.7, $\overline{u}$ 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 4.9 for $\overline{w}$. Writing (4.34) 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

$$\int_0^T\int_\Omega \nabla\overline{u}(\boldsymbol{x},t)\cdot\nabla\overline{w}(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t = \int_0^T\int_\Omega \overline{v}(\boldsymbol{x},t)\overline{w}(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t. \tag{4.43}$$

This precisely shows that $\overline{u} = \Delta^i\overline{v}$.

Step 3: proof of (4.38).
Write now (4.34) with $v = (v_m)_\theta(t)$ and $w = (u_m)_\theta(t)$, and integrate over $t\in(0,T)$. By 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 (4.43) 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 show 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 (4.38) is thus complete.

Step 4: assuming that $q>1$, proof of (4.39).

We proved in Step 1 that $(\|\delta_{\mathcal{D}_m}(\Delta^i_{\mathcal{D}_m} v_m)\|_{L^q(0,T;Y_m)})_{m\in\mathbb{N}}$ 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 (C.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 wearkly compact in time and strongly in space (estimate on the spatial derivatives).

Theorem 4.24 (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

$$\lim_{m\to\infty}\int_0^T\int_\Omega \Pi^{(\theta)}_{\mathcal{D}_m}\beta_m(\boldsymbol{x},t)\Pi^{(\theta)}_{\mathcal{D}_m}\zeta_m(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t = \int_0^T\int_\Omega \overline{\beta}(\boldsymbol{x},t)\,\overline{\zeta}(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t. \quad (4.44)$$

Proof. The sequence $(\beta_m)_{m\in\mathbb{N}}$ satisfies the hypotheses of Theorem 4.23. 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

$$\int_0^T\int_\Omega \Pi^{(\theta)}_{\mathcal{D}_m}\beta_m(\boldsymbol{x},t)\Pi^{(\theta)}_{\mathcal{D}_m}\zeta_m(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t = \int_0^T\int_\Omega \nabla^{(\theta)}_{\mathcal{D}_m}(\Delta^i_{\mathcal{D}_m}\beta_m)(\boldsymbol{x},t)\cdot\nabla^{(\theta)}_{\mathcal{D}_m}\zeta_m(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t. \quad (4.45)$$

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 4.7, $\nabla^{(\theta)}_{\mathcal{D}_m}\zeta_m \to \nabla\overline{\zeta}$ weakly in $L^2(0,T;L^2(\Omega)^d)$. Passing to the limit in the right-hand side of (4.45), 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 (4.44). ∎

4.3 Uniform-in-time compactness

Most of the results and techniques developed here come from [33].

4.3.1 Definitions and abstract results

Solutions of numerical schemes for parabolic equations are usually piecewise constant in time, and therefore not continous. 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 are based on a generalisation to non-continuous functions of the Ascoli–Arzelà theorem.

Definition 4.25. *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 4.26 (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 4.25. 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*

$$\lim_{d_K(s,s')\to 0}\omega(s,s')=0\,,\quad \lim_{m\to\infty}\tau_m=0\,, \tag{4.46}$$

$$\forall (s,s')\in K^2\,,\ \forall m\in\mathbb{N}\,,\ d_E(v_m(s),v_m(s'))\le\omega(s,s')+\tau_m. \tag{4.47}$$

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 trivially obtained by passing to the limit along this subsequence in (4.47), showing that the modulus of continuity of v is bounded above by ω.
The proof of the compactness result 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 (4.46), 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 $\{s_{l_1}, \dots, s_{l_N}\}$ such that any $s \in K$ is within distance δ of a s_{l_i}. Then, for any $m, m' \ge M$, by (4.47),

$$\begin{aligned} d_E(v_m(s), v_{m'}(s)) &\le d_E(v_m(s), v_m(s_{l_i})) + d_E(v_m(s_{l_i}), v_{m'}(s_{l_i})) \\ &\quad + d_E(v_{m'}(s_{l_i}), v_{m'}(s)) \\ &\le \omega(s, s_{l_i}) + \tau_m + d_E(v_m(s_{l_i}), v_{m'}(s_{l_i})) + \omega(s, s_{l_i}) + \tau_{m'} \\ &\le 4\varepsilon + d_E(v_m(s_{l_i}), v_{m'}(s_{l_i})). \end{aligned} \tag{4.48}$$

Let $i \in \{1, \dots, N\}$. 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

$$\forall m, m' \ge M_i\,, \quad d_E(v_m(s_{l_i}), v_{m'}(s_{l_i})) \le \varepsilon. \tag{4.49}$$

Take $M' = \max(M, M_1, \dots, M_N)$. Estimates (4.49) and (4.48) 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 4.27 (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*
- $u_m = (u_m^{(n)})_{n=0,\dots,N_m} \in X_m^{N_m+1}$.

Let $(u_m)_\theta : [0,T] \to X_m$ be the piecewise-constant function in time defined by

$$\begin{aligned} &(u_m)_\theta(0) = u_m^{(0)} \text{ and, for all } n = 0, \dots, N_m - 1 \text{ and } t \in (t^{(n)}, t^{(n+1)}], \\ &(u_m)_\theta(t) = \theta u_m^{(n+1)} + (1-\theta) u_m^{(n)}. \end{aligned} \tag{4.50}$$

Set $\eth_m^{(n+\frac{1}{2})} = t_m^{(n+1)} - t_m^{(n)}$ *for* $n = 0, \dots, N_m - 1$, *and define the discrete derivative* $\delta_m u_m$ *by:*

$$\forall n = 0, \dots, N_m - 1\,, \quad \text{for a.e. } t \in (t_m^{(n)}, t_m^{(n+1)})\,,\ \delta_m u_m(t) = \frac{u_m^{(n+1)} - u_m^{(n)}}{\eth_m^{(n+\frac{1}{2})}}.$$

We assume that

(h1) The sequence $(X_m)_{m\in\mathbb{N}}$ *is compactly embedded in* B *(see Definition 4.13).*
(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 $\eth_m = \max_{n=0,\dots,N_m-1} \eth_m^{(n+\frac{1}{2})}$, *it holds* $\lim_{m\to\infty} \eth_m = 0$.

Then, there exists $u \in C([0,T]; B)$ *such that, up to a subsequence,*

$$\lim_{m\to\infty} \sup_{t\in[0,T]} \|(u_m)_\theta(t) - u(t)\|_B = 0. \tag{4.51}$$

Proof. Let

$$u_m^{(n+\theta)} = \theta u_m^{(n+1)} + (1-\theta) u_m^{(n)} \quad \text{and} \quad \delta_m^{(n+\frac{1}{2})} u_m := \frac{u_m^{(n+1)} - u_m^{(n)}}{\eth_m^{(n+\frac{1}{2})}}.$$

Take $n_2 \geq n_1$ 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

$$\begin{aligned}
&(u_m)_\theta(s_2) - (u_m)_\theta(s_1) \\
&\quad = u_m^{(n_2+\theta)} - u_m^{(n_1+\theta)} \\
&\quad = \sum_{n=n_1+1}^{n_2} \left[u_m^{(n+\theta)} - u_m^{(n-1+\theta)} \right] \\
&\quad = \sum_{n=n_1+1}^{n_2} \left[\theta(u_m^{(n+1)} - u_m^{(n)}) + (1-\theta)(u_m^{(n)} - u_m^{(n-1)}) \right] \\
&\quad = \sum_{n=n_1+1}^{n_2} \left[\theta \eth_m^{(n+\frac{1}{2})} \delta_m^{(n+\frac{1}{2})} u_m + (1-\theta) \eth_m^{(n-\frac{1}{2})} \delta_m^{(n-\frac{1}{2})} u_m \right].
\end{aligned} \tag{4.52}$$

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 $\eth_m^{(-\frac{1}{2})} = 0$ and $\delta_m^{(-\frac{1}{2})} u_m = 0$; consider for example $n_2 = 0$ and notice that

$$u_m^{(\theta)} - u_m^{(0)} = \theta(u_m^{(1)} - u_m^{(0)}) = \theta \eth_m^{(\frac{1}{2})} \delta_m^{(\frac{1}{2})} u_m.$$

By the discrete Hölder inequality (C.3) with $\omega_i = \eth_m^{(i\pm\frac{1}{2})}$, $b_i = 1$ and $a_i = \|\delta_m^{(i\pm\frac{1}{2})} u_m\|_B$, since $\frac{q}{q'} = q - 1$,

$$\Big(\sum_{n=n_1+1}^{n_2} \delta t_m^{(n\pm\frac{1}{2})} \left\| \delta_m^{(n+\frac{1}{2})} u_m \right\|_B \Big)^q$$
$$\leq \left(\sum_{n=n_1+1}^{n_2} \delta t_m^{(n\pm\frac{1}{2})} \right)^{q-1} \left(\sum_{n=n_1+1}^{n_2} \delta t_m^{(n\pm\frac{1}{2})} \left\| \delta_m^{(n\pm\frac{1}{2})} u_m \right\|_B^q \right)$$
$$\leq \left[t_m^{(n_2+\frac{1}{2}\pm\frac{1}{2})} - t_m^{(n_1+\frac{1}{2}\pm\frac{1}{2})} \right]^{q-1} C^q, \tag{4.53}$$

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 (4.52), use the triangle inequality, then take the power q and use the convexity of $s \to s^q$. Invoking finally the estimate (4.53) yields

$$\|(u_m)_\theta(s_2) - (u_m)_\theta(s_1)\|_B^q$$
$$\leq \theta C^q \left[t_m^{(n_2+1)} - t_m^{(n_1+1)} \right]^{q-1} + (1-\theta) C^q \left[t_m^{(n_2)} - t_m^{(n_1)} \right]^{q-1},$$

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 $m \in \mathbb{N}$,

$$\|(u_m)_\theta(s_1) - (u_m)_\theta(s_2)\|_B \leq C_1 (|s_2 - s_1| + \delta t_m)^{\frac{q-1}{q}}$$
$$\leq C_1 |s_2 - s_1|^{\frac{q-1}{q}} + C_1 \delta t_m^{\frac{q-1}{q}}. \tag{4.54}$$

In the last line, we used the power-of-sums inequality (C.13).
This relation and (h4) show that $v_m = (u_m)_\theta$ satisfies Assumptions (4.46)–(4.47) in the discontinuous Ascoli–Arzelà theorem (Theorem 4.26), with $K = [0,T]$ and $E = B$. The proof of Corollary 4.27 is therefore complete if we can establish that, for all $s \in [0,T]$,

$$\{(u_m)_\theta(s) \,:\, m \in \mathbb{N}\} \text{ is relatively compact in } B. \tag{4.55}$$

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, $\|(u_m)_\theta(s)\|_{X_m} \leq \|(u_m)_\theta\|_{L^\infty(0,T;X_m)}$ 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 (4.55) 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 (4.55) holds for any $s = 0$.
Owing to (4.54) we have, for $(m,m') \in \mathbb{N}^2$ and $k \in \mathbb{N}$ such that $k^{-1} \leq T$,

$$\|(u_m)_\theta(0) - (u_{m'})_\theta(0)\|_B$$
$$\leq \left\|(u_m)_\theta(0) - (u_m)_\theta(k^{-1})\right\|_B + \left\|(u_m)_\theta(k^{-1}) - (u_{m'})_\theta(k^{-1})\right\|_B$$

$$
\begin{aligned}
&+ \left\|(u_{m'})_\theta(k^{-1}) - (u_{m'})_\theta(0)\right\|_B \\
&\leq 2C_1 k^{-\frac{q-1}{q}} + C_1 \delta t_m^{\frac{q-1}{q}} + C_1 \delta t_{m'}^{\frac{q-1}{q}} + \left\|(u_m)_\theta(k^{-1}) - (u_{m'})_\theta(k^{-1})\right\|_B .
\end{aligned}
$$

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 $m, m' \geq m_0$,

$$
C_1 \delta t_m^{\frac{q-1}{q}} \leq \frac{\varepsilon}{4}\,,\ C_1 \delta t_{m'}^{\frac{q-1}{q}} \leq \frac{\varepsilon}{4} \text{ and } \left\|(u_m)_\theta(k^{-1}) - (u_{m'})_\theta(k^{-1})\right\|_B \leq \frac{\varepsilon}{4}.
$$

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 4.28. *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 4.25. 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 $k \in \mathbb{N}$,

$$d_E(v_{m_k}(r_k), v(r_k)) \geq \varepsilon/2. \tag{4.56}$$

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 (4.56) 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 4.29 (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 4.30. *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*

$$d_E(v, w) = \sum_{l\in\mathbb{N}} \frac{\min(1, |\langle v - w, \varphi_l\rangle_{L^2(\Omega)}|)}{2^l}. \tag{4.57}$$

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 neighborhood basis of 0 for the $L^2(\Omega)$-weak topology on E. A neighborhood 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 $v \in E$,

$$|\langle v, \varphi\rangle_{L^2(\Omega)}| \le R \|\varphi - \varphi_l\|_{L^2(\Omega)} + |\langle v, \varphi_l\rangle_{L^2(\Omega)}|.$$

By density of $\{\varphi_l : l \in \mathbb{N}\}$ we can select $l \in \mathbb{N}$ such that $\|\varphi - \varphi_l\|_{L^2(\Omega)} \le \varepsilon/(2R)$, and we then see that $E_{\varphi_l,\varepsilon/2} \subset E_{\varphi,\varepsilon}$. Hence, a neighborhood 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 $l \in \mathbb{N}$, $\min(1, |\langle v, \varphi_l\rangle_{L^2(\Omega)}|) \le 2^l d_E(0, v)$. If $d_E(0, v) < 2^{-l}$ this shows that $|\langle v, \varphi_l\rangle_{L^2(\Omega)}| \le 2^l d_E(0, v)$ and therefore that

$$B_{d_E}(0, \min(2^{-l}, \varepsilon 2^{-l})) \subset E_{\varphi_l,\varepsilon}.$$

Hence, any neighborhood of 0 in E for the $L^2(\Omega)$-weak topology is a neighborhood of 0 for d_E. Conversely, for any $\varepsilon > 0$, selecting $N \in \mathbb{N}$ such that $\sum_{l\ge N+1} 2^{-l} < \varepsilon/2$ gives, from the definition (4.57) of d_E,

$$\bigcap_{l=1}^{N} E_{\varphi_l,\varepsilon/4} \subset B_{d_E}(0, \varepsilon).$$

Hence, any ball for d_E centered at 0 is a neighborhood of 0 for the $L^2(\Omega)$-weak topology. Since d_E and $L^2(\Omega)$-weak neighborhoods 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 $t \in [0, T]$. ∎

4.3.2 Application to space–time gradient discretisations

We now consider applications of the previous results to the framework of space–time gradient discretisations for generic boundary conditions, as described in Section 4.1.
The following theorem is a consequence of Corollary 4.27.

Theorem 4.31 ($L^\infty(0,T;L^p(\Omega))$ 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 $v_m \in X_{\mathcal{D}_m,\bullet}^{N_m+1}$. Assume that there exist $C > 0$ and $q > 1$ satisfying*

$$\forall m \in \mathbb{N},\ \|(v_m)_\theta\|_{L^\infty(0,T;X_{\mathcal{D}_m,\bullet})} \le C, \tag{4.58}$$

and

$$\forall m \in \mathbb{N}\,,\ \|\delta_{\mathcal{D}_m} v_m\|_{L^q(0,T;L^p(\Omega))} \le C. \tag{4.59}$$

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*

$$\lim_{m\to\infty} \sup_{t\in[0,T]} \left\| \Pi^{(\theta)}_{\mathcal{D}_m} v_m(t) - u(t) \right\|_{L^p(\Omega)} = 0. \tag{4.60}$$

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 4.27 with $B = L^p(\Omega)$, $X_m = \Pi_{\mathcal{D}_m}(X_{\mathcal{D}_m,\bullet})$ endowed with the norm (4.33), 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 4.13, which yields Hypothesis (h1) in Corollary 4.27. Hypothesis (h2) is satisfied owing to (4.58) and

$$\|(u_m)_\theta(t)\|_{X_m} = \|\Pi_{\mathcal{D}_m}[(v_m)_\theta(t)]\|_{X_m} \le \|(v_m)_\theta(t)\|_{\mathcal{D}_m}\,.$$

Hypothesis (h3) of Corollary 4.27 is obtained by (4.59) 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 4.27, we obtain $u \in C([0,T];L^p(\Omega))$ such that, up to a subsequence, (4.60) holds. The fact that u belongs to $L^\infty(0,T;W^{1,p}_\bullet(\Omega))$ follows by Lemma 4.7.

It remains to prove the convergence of the discrete time derivative. By (4.59) 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 $\psi \in C^\infty_c(\Omega\times(0,T))$ and write, by definition (4.4) of $\delta_{\mathcal{D}_m} v_m$,

$$\begin{aligned}
&\int_0^T \int_\Omega \delta_{\mathcal{D}_m} v_m(\boldsymbol{x},t)\psi(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t \\
&= \sum_{n=1}^{N_m-1} \frac{1}{\delta t^{(n+\frac12)}} \int_{t^{(n)}}^{t^{(n+1)}} \int_\Omega \left(\Pi_{\mathcal{D}_m} v_m^{(n+1)}(\boldsymbol{x}) - \Pi_{\mathcal{D}_m} v_m^{(n)}(\boldsymbol{x}) \right) \psi(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t.
\end{aligned} \tag{4.61}$$

Set $\psi_n(\boldsymbol{x}) = \frac{1}{\delta t^{(n+\frac12)}} \int_{t^{(n)}}^{t^{(n+1)}} \psi(\boldsymbol{x},t)\mathrm{d}t$ and, for $\nu = 1-\theta$, $\psi_{n+\nu} = \nu\psi_{n+1} + (1-\nu)\psi_n$. Since ψ is smooth, $|\psi_n(\boldsymbol{x}) - \psi_{n+\nu}(\boldsymbol{x})| \le C_\psi \delta t_{\mathcal{D}_m}$ for some C_ψ not depending on $\boldsymbol{x}$ or n. Using the discrete integration-by-parts formula (C.17), (4.61) 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),

$$\int_0^T \int_\Omega \delta_{\mathcal{D}_m} v_m(\boldsymbol{x},t)\psi(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t \tag{4.62}$$

$$= \sum_{n=1}^{N_m-1} \int_\Omega \left(\Pi_{\mathcal{D}_m} v_m^{(n+1)}(\boldsymbol{x}) - \Pi_{\mathcal{D}_m} v_m^{(n)}(\boldsymbol{x})\right)\psi_n(\boldsymbol{x})\mathrm{d}\boldsymbol{x}$$

$$= \sum_{n=1}^{N_m-1} \int_\Omega \left(\Pi_{\mathcal{D}_m} v_m^{(n+1)}(\boldsymbol{x}) - \Pi_{\mathcal{D}_m} v_m^{(n)}(\boldsymbol{x})\right)\psi_{n+\nu}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} + R_m$$

$$= -\sum_{n=1}^{N_m-1} \int_\Omega \Pi_{\mathcal{D}_m} v_m^{(n+\theta)}(\boldsymbol{x})\left(\psi_{n+1}(\boldsymbol{x}) - \psi_n(\boldsymbol{x})\right)\mathrm{d}\boldsymbol{x} + R_m$$

$$= -\sum_{n=1}^{N_m-1} \int_{t^{(n)}}^{t^{(n+1)}} \int_\Omega \Pi_{\mathcal{D}_m} v_m^{(n+\theta)}(\boldsymbol{x}) \frac{\psi_{n+1}(\boldsymbol{x}) - \psi_n(\boldsymbol{x})}{\delta t^{(n+\frac{1}{2})}}\mathrm{d}\boldsymbol{x} + R_m \tag{4.63}$$

where, owing to (4.59),

$$|R_m| \le C_\psi \delta t_{\mathcal{D}_m} \|\delta_{\mathcal{D}_m} v_m\|_{L^1(\Omega\times(0,T))} \to 0 \text{ as } m \to \infty.$$

The term (4.62) converges to

$$\int_0^T \int_\Omega U(\boldsymbol{x},t)\psi(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t \tag{4.64}$$

and, owing to the smoothness of ψ and the convergence of $\Pi_{\mathcal{D}_m}^{(\theta)} v_m$, the term (4.63) converges to

$$-\int_0^T \int_\Omega u(\boldsymbol{x},t)\partial_t\psi(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t. \tag{4.65}$$

The proof that $U = \partial_t u$ is complete by equating (4.64) and (4.65). ∎

The uniform-in-time weak-in-space compactness result provided by the next theorem is the initial step to proving a uniform-in-time strong-in-space convergence result for gradient scheme approximations of parabolic equations.

Theorem 4.32 (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 $m \in \mathbb{N}$,*

$$\sup_{t\in[0,T]} \left\|\Pi_{\mathcal{D}_m}^{(\theta)} v_m(t)\right\|_{L^2(\Omega)} \le C \quad \text{and} \quad \int_0^T \|\delta_{\mathcal{D}_m} v_m(t)\|_{\star,\mathcal{D}_m}^q \,\mathrm{d}t \le C \tag{4.66}$$

(see Definition 4.18 of $\|\cdot\|_{\star,\mathcal{D}}$).

Then, the sequence $(\Pi_{\mathcal{D}_m}^{(\theta)} v_m)_{m\in\mathbb{N}}$ is relatively compact weakly in $L^2(\Omega)$ uniformly on $[0,T]$, that is, it has a subsequence which converges accordingly to Definition 4.29.

Moreover, any limit of such a subsequence is continuous $[0,T] \to L^2(\Omega)$ for the weak topology.

Proof. Theorem 4.32 is a consequence of the discontinuous Ascoli–Arzelà theorem (Theorem 4.26), 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 (4.57) from these φ_l. By Proposition 4.30, this metric defines the weak $L^2(\Omega)$ topology.

The set E is metric compact and therefore complete, and the functions $\Pi_{\mathcal{D}_m}^{(\theta)} v_m$ have values in E. It remains to estimate $d_E(\Pi_{\mathcal{D}_m}^{(\theta)} v_m(s), \Pi_{\mathcal{D}_m}^{(\theta)} v_m(s'))$. 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 (4.52), we write

$$\begin{aligned}
&\Pi_{\mathcal{D}}^{(\theta)} v_m(s') - \Pi_{\mathcal{D}}^{(\theta)} v_m(s) \\
&\quad = \theta \sum_{n=n_1+1}^{n_2} \eth^{(n+\frac{1}{2})} \delta_{\mathcal{D}}^{(n+\frac{1}{2})} v_m + (1-\theta) \sum_{n=n_1+1}^{n_2} \eth^{(n-\frac{1}{2})} \delta_{\mathcal{D}}^{(n-\frac{1}{2})} v_m,
\end{aligned}$$

where $\eth^{(-\frac{1}{2})} = 0$ and $\delta_{\mathcal{D}}^{(-\frac{1}{2})} v_m = 0$. Take $P_{\mathcal{D}} \varphi_l \in X_{\mathcal{D},\bullet}$ that realises the minimum defining $S_{\mathcal{D}}(\varphi_l)$, multiply the previous relation by by $\Pi_{\mathcal{D}} P_{\mathcal{D}} \varphi_l$ and integrate over Ω. Estimates (4.29), (4.66) and the Hölder inequality (C.3) (used as in (4.53)) yield

$$\begin{aligned}
&\left| \int_\Omega \left(\Pi_{\mathcal{D}}^{(\theta)} v_m(\boldsymbol{x}, s') - \Pi_{\mathcal{D}}^{(\theta)} v_m(\boldsymbol{x}, s) \right) \Pi_{\mathcal{D}} P_{\mathcal{D}} \varphi_l(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \right| \\
&\le \left| \theta \sum_{n=n_1+1}^{n_2} \eth^{(n+\frac{1}{2})} \int_\Omega \delta_{\mathcal{D}}^{(n+\frac{1}{2})} v_m(\boldsymbol{x}) \Pi_{\mathcal{D}} P_{\mathcal{D}} \varphi_l(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \right| \\
&\quad + \left| (1-\theta) \sum_{n=n_1+1}^{n_2} \eth^{(n-\frac{1}{2})} \int_\Omega \delta_{\mathcal{D}}^{(n-\frac{1}{2})} v_m(\boldsymbol{x}) \Pi_{\mathcal{D}} P_{\mathcal{D}} \varphi_l(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \right| \\
&\le \theta \left\| P_{\mathcal{D}} \varphi_l \right\|_{\mathcal{D}} \sum_{n=n_1+1}^{n_2} \eth^{(n+\frac{1}{2})} \left\| \delta_{\mathcal{D}}^{(n+\frac{1}{2})} v_m \right\|_{\star,\mathcal{D}} \\
&\quad + (1-\theta) \left\| P_{\mathcal{D}} \varphi_l \right\|_{\mathcal{D}} \sum_{n=n_1+1}^{n_2} \eth^{(n-\frac{1}{2})} \left\| \delta_{\mathcal{D}}^{(n-\frac{1}{2})} v_m \right\|_{\star,\mathcal{D}} \\
&\le C^{1/q} \Big[\theta (t^{(n_2+1)} - t^{(n_1+1)})^{1/q'} \\
&\qquad\qquad + (1-\theta)(t^{(n_2)} - t^{(n_1)})^{1/q'} \Big] \left\| P_{\mathcal{D}} \varphi_l \right\|_{\mathcal{D}}. \qquad (4.67)
\end{aligned}$$

By definition of $P_{\mathcal{D}}$ and of $\|\cdot\|_{\mathcal{D}}$ (depending on the specific boundary conditions), we have

$$\left\| \Pi_{\mathcal{D}} P_{\mathcal{D}} \varphi_l - \varphi_l \right\|_{L^2(\Omega)} \le \widehat{S}_{\mathcal{D}}(\varphi_l)$$

and, using a triangle inequality,

$$\|P_{\mathcal{D}}\varphi_l\|_{\mathcal{D}} \le \widehat{S}_{\mathcal{D}}(\varphi_l) + D_{\varphi_l} \le C_{\varphi_l}$$

where D_{φ_l} and C_{φ_l} do not depend on $\mathcal{D}$ (and therefore on m). Since $t^{(n_2+1)} - t^{(n_1+1)} \le |s'-s| + \delta t$ and $t^{(n_2)} - t^{(n_1)} \le |s'-s| + \delta t$, the estimate on $\Pi_{\mathcal{D}_m}^{(\theta)} v_m$ in (4.66) gives, owing to (4.67),

$$\begin{aligned}
&\left| \int_\Omega \left(\Pi_{\mathcal{D}}^{(\theta)} v_m(\boldsymbol{x}, s') - \Pi_{\mathcal{D}}^{(\theta)} v_m(\boldsymbol{x}, s) \right) \varphi_l(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \right| \\
&\le \left| \int_\Omega \left(\Pi_{\mathcal{D}}^{(\theta)} v_m(\boldsymbol{x}, s') - \Pi_{\mathcal{D}}^{(\theta)} v_m(\boldsymbol{x}, s) \right) \Pi_{\mathcal{D}} P_{\mathcal{D}} \varphi_l(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \right| + 2C \widehat{S}_{\mathcal{D}}(\varphi_l) \\
&\le C^{1/q} C_{\varphi_l} |s'-s|^{1/q'} + C^{1/q} C_{\varphi_l} \delta t^{1/q'} + 2C \widehat{S}_{\mathcal{D}}(\varphi_l). \qquad (4.68)
\end{aligned}$$

Plugged into the definition (4.57) of the distance in E, this yields

$$\begin{aligned}
&d_E\left(\Pi_{\mathcal{D}}^{(\theta)} v_m(s'), \Pi_{\mathcal{D}}^{(\theta)} v_m(s) \right) \\
&\quad \le \sum_{l \in \mathbb{N}} \frac{\min(1, C^{1/q'} C_{\varphi_l} |s'-s|^{1/q'})}{2^l} \\
&\qquad + \sum_{l \in \mathbb{N}} \frac{\min(1, 2C \widehat{S}_{\mathcal{D}_m}(\varphi_l) + C^{1/q'} C_{\varphi_l} \delta t_m^{1/q'})}{2^l} =: \omega(s, s') + \tau_m.
\end{aligned}$$

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 4.26 are satisfied and the proof is complete. ∎

The following easy lemma is used in the proofs of uniform-in-time convergence and strong convergence of the gradient for gradient schemes approximations of parabolic equations.

Lemma 4.33. *Let $(a_m)_{m\in\mathbb{N}}$ and $(b_m)_{m\in\mathbb{N}}$ be two sequences of real numbers, and $a, b \in \mathbb{R}$. We assume that $a \le \liminf_{m\to\infty} a_m$, $b \le \liminf_{m\to\infty} b_m$ and $\limsup_{m\to\infty}(a_m + b_m) \le a + b$. Then $a_m \to a$ and $b_m \to b$ as $m \to \infty$.*

Proof. We have

$$\begin{aligned}
a + b &\le \liminf_{m\to\infty} a_m + \liminf_{m\to\infty} b_m \\
&\le \liminf_{m\to\infty} (a_m + b_m) \le \limsup_{m\to\infty} (a_m + b_m) \le a + b.
\end{aligned}$$

Hence, all inequalities involved in this sequence are equalities and, in particular, $\liminf_{m\to\infty}(a_m + b_m) = \limsup_{m\to\infty}(a_m + b_m) = \lim_{m\to\infty}(a_m + b_m) = a + b$, and $a = \liminf_{m\to\infty} a_m$. We then write

$$\limsup_{m\to\infty} b_m = \limsup_{m\to\infty} (b_m + a_m - a_m)$$

$$\begin{aligned} &\leq \limsup_{m\to\infty}(b_m + a_m) + \limsup_{m\to\infty}(-a_m) \\ &= \lim_{m\to\infty}(a_m + b_m) - \liminf_{m\to\infty} a_m = a + b - a = b. \end{aligned}$$

Combined with $b \leq \liminf_{m\to\infty} b_m$, this proves that $b_m \to b$ as $m \to \infty$. We then have $a_m = a_m + b_m - b_m \to a + b - b = a$ 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 Chapter 3. We first prove an error estimate for the GDM approximation of the linear version of this problem, under additional regularity hypotheses. For the complete non-linear problem, the mathematical arguments used in the convergence analysis of the GDM come from Chapter 4. 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 indludes 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 generalised (non-local) fully non-linear Leray–Lions parabolic problems with Neumann boundary conditions. These problems arise in particular from image processing models. Several convergence results for the GDM are obtained, including a uniform-in-time strong-in-space convergence result (based on the tools developed in Section 4.3). 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 quasilinear parabolic problem

In the 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)
- $\Lambda : \Omega \times (0,T) \to \mathcal{M}_d(\mathbb{R})$ 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

$$\left\{\begin{array}{l} \overline{u} \in L^2(0,T; H^1_0(\Omega)) \text{ and, for all } \overline{v} \in L^2(0,T; H^1_0(\Omega)) \\ \text{ such that } \partial_t \overline{v} \in L^2(\Omega \times (0,T)) \text{ and } \overline{v}(\cdot, T) = 0, \\ -\displaystyle\int_0^T \int_\Omega \overline{u}(\boldsymbol{x},t)\partial_t \overline{v}(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t - \int_\Omega u_{\text{ini}}(\boldsymbol{x})\overline{v}(\boldsymbol{x},0)\mathrm{d}\boldsymbol{x} \\ \qquad + \displaystyle\int_0^T \int_\Omega \Lambda(\boldsymbol{x}, \overline{u}(\boldsymbol{x},t))\nabla \overline{u}(\boldsymbol{x},t) \cdot \nabla \overline{v}(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t \\ \qquad\quad = \displaystyle\int_0^T \int_\Omega (f(\boldsymbol{x},t)\overline{v}(\boldsymbol{x},t) - \boldsymbol{F}(\boldsymbol{x},t) \cdot \nabla \overline{v}(\boldsymbol{x},t))\mathrm{d}\boldsymbol{x}\mathrm{d}t. \end{array}\right. \tag{5.3}$$

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 $C_c^\infty([0,T]; H^1_0(\Omega))$ in $L^2(0,T; H^1_0(\Omega))$ (see [29, 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 5.6).
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 $\mathcal{D}_T = (X_{\mathcal{D},0}, \Pi_\mathcal{D}, \nabla_\mathcal{D}, \mathcal{I}_\mathcal{D}, (t^{(n)})_{n=0,\dots,N})$ 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 $(u^{(n)})_{n=0,\dots,N} \in X_{\mathcal{D},0}^{N+1}$ such that, recalling the notations (4.2) and (4.4),

$$\begin{cases} u^{(0)} = \mathcal{I}_\mathcal{D} u_{\rm ini} \text{ and, for all } n = 0,\dots,N-1,\ u^{(n+1)} \text{ satisfies} \\ \displaystyle\int_\Omega \delta_\mathcal{D}^{(n+\frac{1}{2})} u(\boldsymbol{x}) \Pi_\mathcal{D} v(\boldsymbol{x}) {\rm d}\boldsymbol{x} \\ \qquad + \displaystyle\int_\Omega \Lambda(\boldsymbol{x},\Pi_\mathcal{D} u^{(n+\theta)}(\boldsymbol{x}))\nabla_\mathcal{D} u^{(n+\theta)}(\boldsymbol{x})\cdot\nabla_\mathcal{D} v(\boldsymbol{x}){\rm d}\boldsymbol{x} \\ \qquad = \dfrac{1}{\delta t^{(n+\frac{1}{2})}} \displaystyle\int_{t^{(n)}}^{t^{(n+1)}} \int_\Omega (f(\boldsymbol{x},t)\Pi_\mathcal{D} v(\boldsymbol{x}) - \boldsymbol{F}(\boldsymbol{x},t)\cdot\nabla_\mathcal{D} v(\boldsymbol{x})){\rm d}\boldsymbol{x}{\rm d}t, \\ \qquad\qquad \forall v \in X_{\mathcal{D},0}. \end{cases} \tag{5.5}$$

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 $n = 0,\dots,N-1$, 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 (3.44) for the steady quasilinear 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 we consider is therefore $\partial_t \overline{u} - \Delta \overline{u} = f$, with homogeneous Dirichlet boundary conditions.

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. We assume the existence of $h_{\mathcal{D}} > 0$ such that*

$$\forall \varphi \in W^{2,\infty}(\Omega) \cap H_0^1(\Omega),\ S_{\mathcal{D}}(\varphi) \le h_{\mathcal{D}} \|\varphi\|_{W^{2,\infty}(\Omega)}, \tag{5.7a}$$

$$\forall \boldsymbol{\varphi} \in W^{1,\infty}(\Omega)^d,\ W_{\mathcal{D}}(\boldsymbol{\varphi}) \le h_{\mathcal{D}} \|\boldsymbol{\varphi}\|_{W^{1,\infty}(\Omega)^d}, \tag{5.7b}$$

$$\forall \varphi \in W^{1,\infty}(\Omega) \cap H_0^1(\Omega),\ \|\Pi_{\mathcal{D}} \mathcal{I}_{\mathcal{D}} \varphi - \varphi\|_{L^2(\Omega)} \le h_{\mathcal{D}} \|\varphi\|_{W^{1,\infty}(\Omega)}. \tag{5.7c}$$

Assume that the solution $\overline{u}$ to (5.4) *belongs to $W^{1,\infty}(0,T; W^{2,\infty}(\Omega))$, and let u be the solution to the GS* (5.5) *with $\theta = 1$. Then there exists $C > 0$, depending only on $\overline{u}$, Ω, T and (in a non-decreasing way) of $C_{\mathcal{D}}$, such that*

$$\max_{t\in[0,T]} \left\| \Pi_{\mathcal{D}}^{(1)} u(\cdot,t) - \overline{u}(\cdot,t) \right\|_{L^2(\Omega)} \le C(\delta t_{\mathcal{D}} + h_{\mathcal{D}})$$

and

$$\left\| \nabla_{\mathcal{D}}^{(1)} u - \nabla \overline{u} \right\|_{L^2(\Omega\times(0,T))^d} \le C(\delta t_{\mathcal{D}} + h_{\mathcal{D}}).$$

Remark 5.4 (Existence of $h_{\mathcal{D}}$)
If $\mathcal{I}_{\mathcal{D}}$ is linear and continuous (such as, e.g., in Remark 4.4), there always exists $h_{\mathcal{D}}$ satisfying (5.7). Indeed, defining $P_{\mathcal{D}}^{(2)} : H_0^1(\Omega) \to X_{\mathcal{D},0}$ as in (5.8) below, the property (5.7) holds with $h_{\mathcal{D}}$ upper bound of the norms of the following linear or bilinear operators:

$$W^{2,\infty}(\Omega) \cap H_0^1(\Omega) \to L^2(\Omega) \times L^2(\Omega)^d,$$
$$u \mapsto (u - \Pi_{\mathcal{D}} P_{\mathcal{D}}^{(2)} u, \nabla u - \nabla_{\mathcal{D}} P_{\mathcal{D}}^{(2)} u),$$

$$W^{1,\infty}(\Omega)^d \times X_{\mathcal{D},0} \to \mathbb{R},$$
$$(\boldsymbol{\varphi}, v) \mapsto \int_\Omega (\mathrm{div}\boldsymbol{\varphi}(\boldsymbol{x}) \Pi_{\mathcal{D}} v(\boldsymbol{x}) + \boldsymbol{\varphi}(\boldsymbol{x}) \cdot \nabla_{\mathcal{D}} v(\boldsymbol{x})) \mathrm{d}\boldsymbol{x},$$

and

$$W^{1,\infty}(\Omega) \cap H_0^1(\Omega) \to L^2(\Omega),$$
$$u \mapsto u - \Pi_{\mathcal{D}} \mathcal{I}_{\mathcal{D}} u.$$

Proof. In the following proof, we denote by C_i various quantities having the same dependencies as C in the theorem.
For the sake of brevity, if $n \in \{0, \dots, N-1\}$ 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 $P_{\mathcal{D}}$ defined by (3.10) enables us to "plug" the exact solution into the scheme, which is an essential process in establishing error estimates. However, this $P_{\mathcal{D}}$ is not necessarily linear, which becomes a problem for parabolic equations. We therefore need a slightly modified version of this interpolator. We define $P_{\mathcal{D}}^{(2)} : H_0^1(\Omega) \to X_{\mathcal{D},0}$ by: for $\varphi \in H_0^1(\Omega)$,

$$P_{\mathcal{D}}^{(2)}\varphi = \operatorname*{argmin}_{w \in X_{\mathcal{D},0}} \left(\|\Pi_{\mathcal{D}} w - \varphi\|_{L^2(\Omega)}^2 + \|\nabla_{\mathcal{D}} w - \nabla\varphi\|_{L^2(\Omega)^d}^2 \right). \tag{5.8}$$

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 $P_{\mathcal{D}}^{(2)}\varphi = \mathcal{R} \circ \mathcal{P}(\varphi, \nabla\varphi)$ for all $\varphi \in H_0^1(\Omega)$. Hence, $P_{\mathcal{D}}^{(2)}\varphi$ is uniquely defined and $P_{\mathcal{D}}^{(2)}$ is linear continuous. The characterisation of the orthogonal projection $\mathcal{P}$ also shows that, for all $\varphi \in H_0^1(\Omega)$ and $w \in X_{\mathcal{D},0}$,

$$\begin{aligned} \int_\Omega \Pi_{\mathcal{D}} P_{\mathcal{D}}^{(2)}\varphi(\boldsymbol{x}) \Pi_{\mathcal{D}} w(\boldsymbol{x}) + \nabla_{\mathcal{D}} P_{\mathcal{D}}^{(2)}\varphi(\boldsymbol{x}) \cdot \nabla_{\mathcal{D}} w(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \\ = \int_\Omega \varphi(\boldsymbol{x}) \Pi_{\mathcal{D}} w(\boldsymbol{x}) + \nabla\varphi(\boldsymbol{x}) \cdot \nabla_{\mathcal{D}} w(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}. \end{aligned}$$

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 $P_{\mathcal{D}}^{(2)}$ shows that

$$\begin{aligned} &\left(\left\| \Pi_{\mathcal{D}} P_{\mathcal{D}}^{(2)}\varphi - \varphi \right\|_{L^2(\Omega)}^2 + \left\| \nabla_{\mathcal{D}} P_{\mathcal{D}}^{(2)}\varphi - \nabla\varphi \right\|_{L^2(\Omega)^d}^2 \right)^{1/2} \\ &\qquad \le \left(\|\Pi_{\mathcal{D}} v - \varphi\|_{L^2(\Omega)}^2 + \|\nabla_{\mathcal{D}} v - \nabla\varphi\|_{L^2(\Omega)^d}^2 \right)^{1/2} \le \sqrt{2} S_{\mathcal{D}}(\varphi). \end{aligned} \tag{5.9}$$

We have $\overline{u} \in C([0,T]; W^{2,\infty}(\Omega))$ and $\nabla\overline{u} \in W^{1,\infty}(0,T; L^2(\Omega)^d)$, which implies that $\nabla\overline{u} : [0,T] \to L^2(\Omega)$ is Lipschitz-continuous. Hence, (5.9) with $\varphi = \overline{u}(t^{(n+1)})$ and Assumption (5.7a) yield

$$\left\| \nabla\overline{u}^{(n+1)} - \nabla_{\mathcal{D}} P_{\mathcal{D}}^{(2)} \overline{u}(t^{(n+1)}) \right\|_{L^2(\Omega)^d} \le \left\| \nabla\overline{u}^{(n+1)} - \nabla\overline{u}(t^{(n+1)}) \right\|_{L^2(\Omega)^d}$$

$$
\begin{aligned}
&\quad + S_{\mathcal{D}}(\overline{u}(t^{(n+1)})) \\
&\le C_1(\delta t_{\mathcal{D}} + h_{\mathcal{D}}).
\end{aligned} \tag{5.10}
$$

Since $\partial_t \overline{u} \in L^\infty(0,T;W^{2,\infty}(\Omega))$, the quantity $\left\|\partial_t \overline{u}^{(n+1)}\right\|_{W^{2,\infty}(\Omega)}$ 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 $P_{\mathcal{D}}^{(2)}$ and invoking (5.7a), we obtain

$$
\left\| \frac{\Pi_{\mathcal{D}} P_{\mathcal{D}}^{(2)} \overline{u}(t^{(n+1)}) - \Pi_{\mathcal{D}} P_{\mathcal{D}}^{(2)} \overline{u}(t^{(n)})}{\delta t^{(n+\frac{1}{2})}} - \partial_t \overline{u}^{(n+1)} \right\|_{L^2(\Omega)} \le C_2 h_{\mathcal{D}}. \tag{5.11}
$$

Step 2: proof of the error estimates.
Since $\nabla \overline{u}^{(n+1)} \in H_{\mathrm{div}}(\Omega)$ we can write, for all $v \in X_{\mathcal{D},0}$,

$$
\begin{aligned}
&\int_\Omega \Big(\Pi_{\mathcal{D}} v(\boldsymbol{x}) \mathrm{div}(\nabla \overline{u}^{(n+1)})(\boldsymbol{x}) + \nabla \overline{u}^{(n+1)}(\boldsymbol{x}) \cdot \nabla_{\mathcal{D}} v(\boldsymbol{x}) \Big) \mathrm{d}\boldsymbol{x} \\
&\qquad \le W_{\mathcal{D}}(\nabla \overline{u}^{(n+1)}) \left\| v \right\|_{\mathcal{D}}.
\end{aligned}
$$

Owing to the regularity of $\overline{u}$, the equation $\partial_t \overline{u} - f = \mathrm{div}(\nabla \overline{u})$ is satisfied a.e. in space and time. Averaging over time in $(t^{(n)}, t^{(n+1)})$ gives $\partial_t \overline{u}^{(n+1)} - f^{(n+1)} = \mathrm{div}(\nabla \overline{u}^{(n+1)})$ a.e. in space, and thus

$$
\begin{aligned}
&\int_\Omega \Big(\Pi_{\mathcal{D}} v(\boldsymbol{x}) \left(\partial_t \overline{u}^{(n+1)}(\boldsymbol{x}) - f^{(n+1)}(\boldsymbol{x}) \right) + \nabla \overline{u}^{(n+1)}(\boldsymbol{x}) \cdot \nabla_{\mathcal{D}} v(\boldsymbol{x}) \Big) \mathrm{d}\boldsymbol{x} \\
&\qquad \le W_{\mathcal{D}}(\nabla \overline{u}^{(n+1)}) \left\| v \right\|_{\mathcal{D}}.
\end{aligned}
$$

Use the GS (5.5) to replace the term $f^{(n+1)}$ in the left-hand side. Since $\nabla \overline{u} \in L^\infty(0,T;W^{1,\infty}(\Omega)^d)$, the quantity $\|\nabla \overline{u}^{(n+1)}\|_{W^{1,\infty}(\Omega)^d}$ is bounded independently on n and Assumption (5.7b) yields

$$
\begin{aligned}
&\int_\Omega \Pi_{\mathcal{D}} v(\boldsymbol{x}) \left(\partial_t \overline{u}^{(n+1)}(\boldsymbol{x}) - \delta_{\mathcal{D}}^{(n+\frac{1}{2})} u(\boldsymbol{x}) \right) \mathrm{d}\boldsymbol{x} \\
&\qquad + \int_\Omega \left(\nabla \overline{u}^{(n+1)}(\boldsymbol{x}) - \nabla_{\mathcal{D}} u^{(n+1)} \right) \cdot \nabla_{\mathcal{D}} v(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \le C_3 h_{\mathcal{D}} \left\| v \right\|_{\mathcal{D}}.
\end{aligned} \tag{5.12}
$$

For $k = 0, \ldots, N$, set $e^{(k)} = P_{\mathcal{D}}^{(2)} \overline{u}(t^{(k)}) - u^{(k)}$. We have

$$
\begin{aligned}
\delta_{\mathcal{D}}^{(n+\frac{1}{2})} e = &\left[\frac{\Pi_{\mathcal{D}} P_{\mathcal{D}}^{(2)} \overline{u}(t^{(n+1)}) - \Pi_{\mathcal{D}} P_{\mathcal{D}}^{(2)} \overline{u}(t^{(n)})}{\delta t^{(n+\frac{1}{2})}} - \partial_t \overline{u}^{(n+1)} \right] \\
&+ \left[\partial_t \overline{u}^{(n+1)} - \delta_{\mathcal{D}}^{(n+\frac{1}{2})} u \right].
\end{aligned}
$$

and

$$
\nabla_{\mathcal{D}} e^{(n+1)} = \left[\nabla_{\mathcal{D}} P_{\mathcal{D}}^{(2)} \overline{u}(t^{(n+1)}) - \nabla \overline{u}^{(n+1)} \right] + \left[\nabla \overline{u}^{(n+1)} - \nabla_{\mathcal{D}} u^{(n+1)} \right]
$$

Then (5.12), (5.11), (5.10) and the definition of $C_{\mathcal{D}}$ give

$$\int_\Omega \Pi_{\mathcal{D}} v(\boldsymbol{x}) \delta_{\mathcal{D}}^{(n+\frac{1}{2})} e(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} + \int_\Omega \nabla_{\mathcal{D}} e^{(n+1)}(\boldsymbol{x}) \cdot \nabla_{\mathcal{D}} v(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \\ \leq C_4(\eth t_{\mathcal{D}} + h_{\mathcal{D}}) \|v\|_{\mathcal{D}}.$$

Take $v = \eth t^{(n+\frac{1}{2})} e^{(n+1)}$, and sum on $n = 0, \ldots, m-1$ for some $m \in \{1, \ldots, N\}$. Recalling the definition of $\|\cdot\|_{\mathcal{D}}$,

$$\sum_{n=0}^{m-1} \int_\Omega \Pi_{\mathcal{D}} e^{(n+1)}(\boldsymbol{x}) \left[\Pi_{\mathcal{D}} e^{(n+1)}(\boldsymbol{x}) - \Pi_{\mathcal{D}} e^{(n)}(\boldsymbol{x}) \right] \mathrm{d}\boldsymbol{x} \\ + \sum_{n=0}^{m-1} \eth t^{(n+\frac{1}{2})} \int_\Omega |\nabla_{\mathcal{D}} e^{(n+1)}(\boldsymbol{x})|^2 \mathrm{d}\boldsymbol{x} \\ \leq \sum_{n=0}^{N-1} C_4(\eth t_{\mathcal{D}} + h_{\mathcal{D}}) \eth t^{(n+\frac{1}{2})} \left(\int_\Omega |\nabla_{\mathcal{D}} e^{(n+1)}(\boldsymbol{x})|^2 \mathrm{d}\boldsymbol{x} \right)^{1/2}. \quad (5.13)$$

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 (C.8) with $p = p' = 2$ in the right-hand side of (5.13), this leads to

$$\frac{1}{2} \int_\Omega (\Pi_{\mathcal{D}} e^{(m)}(\boldsymbol{x}))^2 \mathrm{d}\boldsymbol{x} + \sum_{n=0}^{N-1} \eth t^{(n+\frac{1}{2})} \int_\Omega |\nabla_{\mathcal{D}} e^{(n+1)}(\boldsymbol{x})|^2 \mathrm{d}\boldsymbol{x} \\ \leq \frac{1}{2} \int_\Omega (\Pi_{\mathcal{D}} e^{(0)}(\boldsymbol{x}))^2 \mathrm{d}\boldsymbol{x} + \frac{1}{2} \sum_{n=0}^{m-1} \eth t^{(n+\frac{1}{2})} \int_\Omega |\nabla_{\mathcal{D}} e^{(n+1)}(\boldsymbol{x})|^2 \mathrm{d}\boldsymbol{x} \\ + \frac{1}{2} \sum_{n=0}^{N-1} C_4^2 (\eth t_{\mathcal{D}} + h_{\mathcal{D}})^2 \eth t^{(n+\frac{1}{2})}. \quad (5.15)$$

By Assumption (5.7c) and Estimate (5.9), since $u^{(0)} = \mathcal{I}_{\mathcal{D}} u_{\mathrm{ini}} = \mathcal{I}_{\mathcal{D}} \overline{u}(0)$,

$$\left\| \Pi_{\mathcal{D}} e^{(0)} \right\|_{L^2(\Omega)} \leq \left\| \Pi_{\mathcal{D}} P_{\mathcal{D}}^{(2)} \overline{u}(0) - \overline{u}(0) \right\|_{L^2(\Omega)} + \| \overline{u}(0) - \Pi_{\mathcal{D}} \mathcal{I}_{\mathcal{D}} \overline{u}(0) \|_{L^2(\Omega)} \\ \leq C_5 h_{\mathcal{D}}.$$

Hence, recalling the definition of $\nabla^{(1)}$ and using $\sum_{n=0}^{N-1} \eth t^{(n+\frac{1}{2})} = T$, Equation (5.15) yields

$$\frac{1}{2} \int_\Omega (\Pi_{\mathcal{D}} e^{(m)}(\boldsymbol{x}))^2 \mathrm{d}\boldsymbol{x} + \frac{1}{2} \int_0^{t^{(m)}} \int_\Omega |\nabla_{\mathcal{D}}^{(1)} e(\boldsymbol{x}, t)|^2 \mathrm{d}\boldsymbol{x} \mathrm{d}t \leq C_6 (\eth t_{\mathcal{D}} + h_{\mathcal{D}})^2. \quad (5.16)$$

Using a triangle inequality, (5.9), and the power-of-sums inequality (C.13) with $\alpha = 1/2$, Equation (5.16) leads on one hand to

$$\forall m = 1, \dots, N\,, \ \left\| I\!I_{\mathcal{D}} u^{(m)} - \overline{u}(t^{(m)}) \right\|_{L^2(\Omega)} \leq C_7(\delta t_{\mathcal{D}} + h_{\mathcal{D}}) + \sqrt{2} S_{\mathcal{D}}(\overline{u}(t^{(m)})) \leq C_8(\delta t_{\mathcal{D}} + h_{\mathcal{D}}). \tag{5.17}$$

On the other hand, using again (5.9) and a triangle inequality, Equations (5.16) with $m = N-1$ and the power-of-sums inequality (C.12) with $\alpha = 2$ lead to

$$\sum_{n=0}^{N-1} \delta t^{(n+\frac{1}{2})} \left\| \nabla_{\mathcal{D}} u^{(n+1)} - \nabla \overline{u}(t^{(n+1)}) \right\|_{L^2(\Omega)}^2 \leq 4C_6(\delta t_{\mathcal{D}} + h_{\mathcal{D}})^2 + 4 \sum_{n=0}^{N-1} \delta t^{(n+\frac{1}{2})} S_{\mathcal{D}}(\overline{u}(t^{(n+1)}))^2 \leq C_9^2(\delta t_{\mathcal{D}} + h_{\mathcal{D}})^2. \tag{5.18}$$

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 quasilinear model (5.1). The convergence result we intend on proving is the following.

Theorem 5.5 (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* $m \to \infty$*,*

$$\sup_{t \in [0,T]} \left\| I\!I^{(\theta)}_{\mathcal{D}_m} u_m(t) - \overline{u}(t) \right\|_{L^2(\Omega)} \to 0 \tag{5.19a}$$

$$\nabla^{(\theta)}_{\mathcal{D}_m} u_m \to \nabla \overline{u} \quad \text{in } L^2(\Omega \times (0,T))^d, \tag{5.19b}$$

where $\overline{u}$ *is a solution to* (5.3) *(and thus also* (5.4)*).*

Remark 5.6. 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 first allow us to prove that such a solution exist, and are then useful to invoke the compactness results of Chapter 4.

Lemma 5.7 ($L^\infty(0,T;L^2(\Omega))$ 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 $k = 0,\dots,N$,*

$$\begin{aligned}
&\frac{1}{2}\int_\Omega (\Pi_\mathcal{D} u^{(k)}(\boldsymbol{x}))^2 \mathrm{d}\boldsymbol{x} \\
&+ \int_0^{t^{(k)}} \int_\Omega \Lambda(\boldsymbol{x}, \Pi_\mathcal{D}^{(\theta)} u(\boldsymbol{x},t)) \nabla_\mathcal{D}^{(\theta)} u(\boldsymbol{x},t) \cdot \nabla_\mathcal{D}^{(\theta)} u(\boldsymbol{x},t) \mathrm{d}\boldsymbol{x}\mathrm{d}t \\
&\qquad \le \frac{1}{2}\int_\Omega (\Pi_\mathcal{D}\mathcal{I}_\mathcal{D} u_{\mathrm{ini}}(\boldsymbol{x}))^2 \mathrm{d}\boldsymbol{x} \\
&\qquad\quad + \int_0^{t^{(k)}} \int_\Omega (f(\boldsymbol{x},t)\Pi_\mathcal{D}^{(\theta)} u(\boldsymbol{x},t) - \boldsymbol{F}(\boldsymbol{x},t)\cdot \nabla_\mathcal{D}^{(\theta)} u(\boldsymbol{x},t)) \mathrm{d}\boldsymbol{x}\mathrm{d}t.
\end{aligned} \tag{5.20}$$

Consequently, there exists $C_{10} > 0$ depending only on $C_P \ge C_\mathcal{D}$ (see Definition 2.2), $C_{\mathrm{ini}} \ge \|\Pi_\mathcal{D}\mathcal{I}_\mathcal{D} u_{\mathrm{ini}}\|_{L^2(\Omega)}$, f, $\boldsymbol{F}$, and $\underline{\lambda}$ such that

$$\sup_{t\in[0,T]} \left\|\Pi_\mathcal{D}^{(\theta)} u(t)\right\|_{L^2(\Omega)} \le C_{10} \quad \text{and} \quad \left\|\nabla_\mathcal{D}^{(\theta)} u\right\|_{L^2(\Omega\times(0,T))^d} \le C_{10}. \tag{5.21}$$

Moreover, there exists at least one solution u to the GS (5.5)*.*

Proof. Relation (5.14) is generalised to the following: for all $a, b \in \mathbb{R}$,

$$\begin{aligned}
(a-b)(\theta a + (1-\theta)b) &= (a-b)\left[\left(\theta - \frac{1}{2}\right)a + \left(\frac{1}{2}-\theta\right)b\right] + \frac{1}{2}(a-b)(a+b) \\
&= \left(\theta - \frac{1}{2}\right)(a-b)^2 + \frac{1}{2}(a^2-b^2) \ge \frac{1}{2}(a^2-b^2).
\end{aligned}$$

Let $n \in \{0,\dots,N-1\}$. Applying the above relation to $a = \Pi_\mathcal{D} u^{(n+1)}$ and $b = \Pi_\mathcal{D} u^{(n)}$ yields

$$\eth^{(n+\frac{1}{2})} \delta_\mathcal{D}^{(n+\frac{1}{2})} u \; \Pi_\mathcal{D} u^{(n+\theta)} \ge \frac{1}{2}\left((\Pi_\mathcal{D} u^{(n+1)})^2 - (\Pi_\mathcal{D} u^{(n)})^2\right). \tag{5.22}$$

Setting $v = \eth^{(n+\frac{1}{2})} u^{(n+\theta)}$ in (5.5) and summing over $n = 0,\dots,k-1$ (we assume here that $k \ge 1$, the case $k = 0$ in (5.20) is trivial) therefore leads to

$$\begin{aligned}
&\frac{1}{2}\sum_{n=0}^{k-1}\left(\int_\Omega (\Pi_\mathcal{D} u^{(n+1)}(\boldsymbol{x}))^2 \mathrm{d}\boldsymbol{x} - \int_\Omega (\Pi_\mathcal{D} u^{(n)}(\boldsymbol{x}))^2 \mathrm{d}\boldsymbol{x}\right) \\
&+ \sum_{n=0}^{k-1} \eth^{(n+\frac{1}{2})} \int_\Omega \Lambda(\boldsymbol{x}, \Pi_\mathcal{D} u^{(n+\theta)}(\boldsymbol{x})) \nabla_\mathcal{D} u^{(n+\theta)}(\boldsymbol{x}) \cdot \nabla_\mathcal{D} u^{(n+\theta)}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \\
&\le \sum_{n=0}^{k-1} \int_{t^{(n)}}^{t^{(n+1)}} \int_\Omega (f(\boldsymbol{x},t)\Pi_\mathcal{D} u^{(n+\theta)}(\boldsymbol{x}) - \boldsymbol{F}(\boldsymbol{x},t)\cdot \nabla_\mathcal{D} u^{(n+\theta)}(\boldsymbol{x})) \mathrm{d}\boldsymbol{x}\mathrm{d}t.
\end{aligned} \tag{5.23}$$

The first sum is telescopic and reduces to

$$\int_\Omega (\Pi_{\mathcal{D}} u^{(k)}(\boldsymbol{x}))^2 \mathrm{d}\boldsymbol{x} - \int_\Omega (\Pi_{\mathcal{D}} u^{(0)}(\boldsymbol{x}))^2 \mathrm{d}\boldsymbol{x} = \int_\Omega (\Pi_{\mathcal{D}} u^{(k)}(\boldsymbol{x}))^2 \mathrm{d}\boldsymbol{x} - \int_\Omega (\Pi_{\mathcal{D}} \mathcal{I}_{\mathcal{D}} u_{\mathrm{ini}}(\boldsymbol{x}))^2 \mathrm{d}\boldsymbol{x}.$$

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.* (C.5) with $p = p' = 2$), the Young inequality (C.9) and the definition (2.1) of $C_{\mathcal{D}}$, we write

$$\begin{aligned}
\int_0^{t^{(k)}} \int_\Omega & (f(\boldsymbol{x},t)\Pi_{\mathcal{D}}^{(\theta)} u(\boldsymbol{x},t) - \boldsymbol{F}(\boldsymbol{x},t)\cdot\nabla_{\mathcal{D}}^{(\theta)} u(\boldsymbol{x},t))\mathrm{d}\boldsymbol{x}\mathrm{d}t \\
&\le \|f\|_{L^2(\Omega\times(0,t^{(k)}))} \left\|\Pi_{\mathcal{D}}^{(\theta)} u\right\|_{L^2(\Omega\times(0,t^{(k)}))} \\
&\quad + \|\boldsymbol{F}\|_{L^2(\Omega\times(0,t^{(k)}))^d} \left\|\nabla_{\mathcal{D}}^{(\theta)} u\right\|_{L^2(\Omega\times(0,t^{(k)}))^d} \\
&\le \frac{C_{\mathcal{D}}^2}{\underline{\lambda}} \|f\|^2_{L^2(\Omega\times(0,t^{(k)}))} + \frac{\underline{\lambda}}{4C_{\mathcal{D}}^2} \left\|\Pi_{\mathcal{D}}^{(\theta)} u\right\|^2_{L^2(\Omega\times(0,t^{(k)}))} \\
&\quad + \frac{1}{\underline{\lambda}} \|\boldsymbol{F}\|^2_{L^2(\Omega\times(0,t^{(k)}))^d} + \frac{\underline{\lambda}}{4} \left\|\nabla_{\mathcal{D}}^{(\theta)} u\right\|^2_{L^2(\Omega\times(0,t^{(k)}))^d} \\
&\le \frac{C_{\mathcal{D}}^2}{\underline{\lambda}} \|f\|^2_{L^2(\Omega\times(0,t^{(k)}))} + \frac{1}{\underline{\lambda}} \|\boldsymbol{F}\|^2_{L^2(\Omega\times(0,t^{(k)}))^d} \\
&\quad + \frac{\underline{\lambda}}{2} \left\|\nabla_{\mathcal{D}}^{(\theta)} u\right\|^2_{L^2(\Omega\times(0,t^{(k)}))^d}.
\end{aligned} \tag{5.24}$$

Plugged into (5.20) and using the coercivity of Λ, this gives

$$\begin{aligned}
\frac{1}{2}\int_\Omega (\Pi_{\mathcal{D}} u^{(k)}(\boldsymbol{x}))^2 \mathrm{d}\boldsymbol{x} + \underline{\lambda} \left\|\nabla_{\mathcal{D}}^{(\theta)} u\right\|^2_{L^2(\Omega\times(0,t^{(k)}))^d} & \\
\le \frac{1}{2} C_{\mathrm{ini}}^2 + \frac{C_{\mathcal{D}}^2}{\underline{\lambda}} \|f\|^2_{L^2(\Omega\times(0,t^{(k)}))} + \frac{1}{\underline{\lambda}} \|\boldsymbol{F}\|^2_{L^2(\Omega\times(0,t^{(k)}))^d} & \\
+ \frac{\underline{\lambda}}{2} \left\|\nabla_{\mathcal{D}}^{(\theta)} u\right\|^2_{L^2(\Omega\times(0,t^{(k)}))^d}. &
\end{aligned}$$

Hence,

$$\begin{aligned}
\frac{1}{2} \max_{k=0,\dots,N} \left\|\Pi_{\mathcal{D}} u^{(k)}\right\|^2_{L^2(\Omega)} + \frac{\underline{\lambda}}{2} \left\|\nabla_{\mathcal{D}}^{(\theta)} u\right\|^2_{L^2(\Omega\times(0,T)^d} & \\
\le \frac{1}{2} C_{\mathrm{ini}}^2 + \frac{C_{\mathcal{D}}^2}{\underline{\lambda}} \|f\|^2_{L^2(\Omega\times(0,T))} + \frac{1}{\underline{\lambda}} \|\boldsymbol{F}\|^2_{L^2(\Omega\times(0,T))^d}. &
\end{aligned}$$

The estimates in (5.21) follow from this inequality and from the fact that, by definition (4.2) of $\Pi_{\mathcal{D}}^{(\theta)}$,

$$\|\Pi_{\mathcal{D}} u^{(n+\theta)}\|_{L^2(\Omega)} \le \theta \|\Pi_{\mathcal{D}} u^{(n+1)}\|_{L^2(\Omega)} + (1-\theta)\|\Pi_{\mathcal{D}} u^{(n)}\|_{L^2(\Omega)}.$$

Following the same arguments as in the proof of Theorem 3.16, it is easy to establish by induction that, for each $n = 0, \dots, N-1$, 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 5.8 (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* $C_P \ge C_{\mathcal{D}}$*,* $C_{\rm ini} \ge \|\Pi_{\mathcal{D}} I_{\mathcal{D}} u_{\rm ini}\|_{L^2(\Omega)}$*,* 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* (4.28)*.*

Proof. In (5.5), choose $v \in X_{\mathcal{D},0}$ which realises the supremum in the definition (4.28) 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

$$\begin{aligned}
\left\|\delta_{\mathcal{D}}^{(n+\frac{1}{2})} u\right\|_{\star,\mathcal{D}} &\le \overline{\lambda} \left\|\nabla_{\mathcal{D}} u^{(n+\theta)}\right\|_{L^2(\Omega)} \\
&\quad + \frac{1}{\delta t^{(n+\frac{1}{2})}} \int_{t^{(n)}}^{t^{(n+1)}} (C_{\mathcal{D}} \|f(\cdot,t)\|_{L^2(\Omega)} + \|\boldsymbol{F}(\cdot,t)\|_{L^2(\Omega)^d}) \mathrm{d}t \\
&= \frac{1}{\delta t^{(n+\frac{1}{2})}} \int_{t^{(n)}}^{t^{(n+1)}} \Big[\overline{\lambda} \left\|\nabla_{\mathcal{D}}^{(\theta)} u(t)\right\|_{L^2(\Omega)} + C_{\mathcal{D}} \|f(\cdot,t)\|_{L^2(\Omega)} \\
&\qquad\qquad + \|\boldsymbol{F}(\cdot,t)\|_{L^2(\Omega)^d}\Big] \mathrm{d}t
\end{aligned}$$

Square, use the Jensen inequality (C.10), multiply by $\delta t^{(n+\frac{1}{2})}$ and apply the power-of-sums inequality (C.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 $n = 0, \dots, N-1$ and by invoking Estimates (5.21). ∎

We are now ready to prove the convergence of the GS (5.5).
Proof of Theorem 5.5.

We note that since $((\mathcal{D}_T)_m)_{m\in\mathbb{N}}$ is compact, it is also coercive (see Lemma 2.9).

Step 1: Application of compactness results.

By Estimates (5.21), Lemma 4.7 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 4.21 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 4.9 (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 $n=0,\dots,N-1$. 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 (C.17), with $\nu = 1-\theta$, to $T_1^{(m)}$ and using the fact that $v^{(N)}=0$, we write

$$\begin{aligned}
T_1^{(m)} =& -\sum_{n=0}^{N-1} \int_\Omega \Pi_{\mathcal{D}} u^{(n+\theta)}(\boldsymbol{x}) \left[\Pi_{\mathcal{D}} v^{(n+1)}(\boldsymbol{x}) - \Pi_{\mathcal{D}} v^{(n)}(\boldsymbol{x})\right]\mathrm{d}\boldsymbol{x} \\
&- \int_\Omega \Pi_{\mathcal{D}} u^{(0)}(\boldsymbol{x})\Pi_{\mathcal{D}} v^{(0)}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \\
=& -\sum_{n=0}^{N-1} \int_{t^{(n)}}^{t^{(n+1)}} \int_\Omega \Pi_{\mathcal{D}}^{(\theta)} u(\boldsymbol{x},t)\delta_{\mathcal{D}} v(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t \\
&- \int_\Omega \Pi_{\mathcal{D}} u^{(0)}(\boldsymbol{x})\Pi_{\mathcal{D}} v^{(0)}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \\
=& -\int_0^T \int_\Omega \Pi_{\mathcal{D}}^{(\theta)} u(\boldsymbol{x},t)\delta_{\mathcal{D}} v(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t \\
&- \int_\Omega \Pi_{\mathcal{D}}\mathcal{I}_{\mathcal{D}} u_{\mathrm{ini}}(\boldsymbol{x})\Pi_{\mathcal{D}}^{(1-\theta)} v(\boldsymbol{x},0)\mathrm{d}\boldsymbol{x}.
\end{aligned}$$

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 $\Pi_{\mathcal{D}} \mathcal{I}_{\mathcal{D}} u_{\rm ini} \to u_{\rm ini}$ 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

$$\lim_{m\to\infty} T_1^{(m)} = -\int_0^T \int_\Omega \overline{u}(\boldsymbol{x},t)\partial_t \overline{v}(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t - \int_\Omega u_{\rm ini}(\boldsymbol{x})\overline{v}(\boldsymbol{x},0)\mathrm{d}\boldsymbol{x}. \quad (5.26)$$

Since $\Pi_{\mathcal{D}}^{(\theta)} u \to \overline{u}$ in $L^2(\Omega\times(0,T))$, Lemma C.4 (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 C.3,

$$\begin{aligned}\lim_{m\to\infty} T_2^{(m)} &= \lim_{m\to\infty} \int_0^T \int_\Omega \nabla_{\mathcal{D}}^{(\theta)} u(\boldsymbol{x},t)\cdot \Lambda(\boldsymbol{x}, \Pi_{\mathcal{D}}^{(\theta)} u(\boldsymbol{x},t))\nabla_{\mathcal{D}}^{(1-\theta)} v(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t \\ &= \int_0^T \int_\Omega \Lambda(\boldsymbol{x},\overline{u}(\boldsymbol{x},t))\nabla\overline{u}(\boldsymbol{x},t)\cdot\nabla\overline{v}(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t. \quad (5.27)\end{aligned}$$

The convergences of $\Pi_{\mathcal{D}}^{(1-\theta)} v$ and $\nabla_{\mathcal{D}}^{(1-\theta)} v$ readily give

$$\lim_{m\to\infty} T_3^{(m)} = \int_0^T \int_\Omega \left(f(\boldsymbol{x},t)\overline{v}(\boldsymbol{x},t) - \boldsymbol{F}(\boldsymbol{x},t)\cdot\nabla\overline{v}(\boldsymbol{x},t)\right)\mathrm{d}\boldsymbol{x}\mathrm{d}t. \quad (5.28)$$

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 $k(m)\in\{0,\dots,N_m-1\}$ 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 $\Pi_{\mathcal{D}}^{(\theta)}$,

$$\begin{aligned}(\Pi_{\mathcal{D}_m}^{(\theta)} u_m(\boldsymbol{x},s_m))^2 &= \left(\theta \Pi_{\mathcal{D}_m} u_m^{(k(m)+1)} + (1-\theta)\Pi_{\mathcal{D}_m} u_m^{(k(m))}\right)^2 \\ &\leq \theta(\Pi_{\mathcal{D}_m} u_m^{(k(m)+1)})^2 + (1-\theta)(\Pi_{\mathcal{D}_m} u_m^{(k(m))})^2. \quad (5.29)\end{aligned}$$

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

$$
\begin{aligned}
&\frac{1}{2}\int_\Omega (\Pi_{\mathcal{D}_m}^{(\theta)} u(\boldsymbol{x}, s_m))^2 \mathrm{d}\boldsymbol{x} \\
&+ \int_0^{s_m^{(-)}} \int_\Omega \Lambda(\boldsymbol{x}, \Pi_{\mathcal{D}_m}^{(\theta)} u(\boldsymbol{x},t)) \nabla_{\mathcal{D}_m}^{(\theta)} u(\boldsymbol{x},t) \cdot \nabla_{\mathcal{D}_m}^{(\theta)} u(\boldsymbol{x},t) \mathrm{d}\boldsymbol{x}\mathrm{d}t \\
&+ \left[\theta \int_{s_m^{(-)}}^{s_m^{(+)}} \int_\Omega \Lambda(\boldsymbol{x}, \Pi_{\mathcal{D}_m}^{(\theta)} u(\boldsymbol{x},t)) \nabla_{\mathcal{D}_m}^{(\theta)} u(\boldsymbol{x},t) \cdot \nabla_{\mathcal{D}_m}^{(\theta)} u(\boldsymbol{x},t) \mathrm{d}\boldsymbol{x}\mathrm{d}t\right] \\
&\quad \le \frac{1}{2}\int_\Omega (\Pi_{\mathcal{D}_m} \mathcal{I}_{\mathcal{D}_m} u_{\mathrm{ini}}(\boldsymbol{x}))^2 \mathrm{d}\boldsymbol{x} \\
&\qquad + \int_0^{s_m^{(-)}} \int_\Omega (f(\boldsymbol{x},t) \Pi_{\mathcal{D}_m}^{(\theta)} u(\boldsymbol{x},t) - \boldsymbol{F}(\boldsymbol{x},t) \cdot \nabla_{\mathcal{D}_m}^{(\theta)} u(\boldsymbol{x},t)) \mathrm{d}\boldsymbol{x}\mathrm{d}t \\
&\qquad + \theta \int_{s_m^{(-)}}^{s_m^{(+)}} \int_\Omega (f(\boldsymbol{x},t) \Pi_{\mathcal{D}_m}^{(\theta)} u(\boldsymbol{x},t) - \boldsymbol{F}(\boldsymbol{x},t) \cdot \nabla_{\mathcal{D}_m}^{(\theta)} u(\boldsymbol{x},t)) \mathrm{d}\boldsymbol{x}\mathrm{d}t.
\end{aligned}
\tag{5.30}
$$

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 C.4 page 404 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 C.3 (weak-strong convergence property) and to the weak convergence in $L^2(\Omega\times(0,T))^d$ of $\nabla^{(\theta)}_{\mathcal{D}_m}u_m$ 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

$$\begin{aligned}&\int_0^s\int_\Omega \Lambda(\boldsymbol{x},\overline{u}(\boldsymbol{x},t))\nabla\overline{u}(\boldsymbol{x},t)\cdot\nabla\overline{u}(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t\\&\le\liminf_{m\to\infty}\int_0^{s_m^{(-)}}\int_\Omega \Lambda(\boldsymbol{x},\Pi^{(\theta)}_{\mathcal{D}_m}u_m(\boldsymbol{x},t))\nabla^{(\theta)}_{\mathcal{D}_m}u_m(\boldsymbol{x},t)\cdot\nabla^{(\theta)}_{\mathcal{D}_m}u_m(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t.\end{aligned}\tag{5.32}$$

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 $\Pi^{(\theta)}_{\mathcal{D}_m}u_m$ and $\nabla^{(\theta)}_{\mathcal{D}_m}u_m$ thus give, as $m\to\infty$,

$$\begin{aligned}&\int_0^{s_m^{(-)}}\int_\Omega(f(\boldsymbol{x},t)\Pi^{(\theta)}_{\mathcal{D}_m}u(\boldsymbol{x},t)-\boldsymbol{F}(\boldsymbol{x},t)\cdot\nabla^{(\theta)}_{\mathcal{D}_m}u(\boldsymbol{x},t))\mathrm{d}\boldsymbol{x}\mathrm{d}t\\&\qquad\to\int_0^s\int_\Omega(f(\boldsymbol{x},t)\overline{u}(\boldsymbol{x},t)-\boldsymbol{F}(\boldsymbol{x},t)\cdot\nabla\overline{u}(\boldsymbol{x},t))\mathrm{d}\boldsymbol{x}\mathrm{d}t.\end{aligned}\tag{5.34}$$

Finally, since $\mathbf{1}_{[s_m^{(-)},s_m^{(+)}]}f\to0$ in $L^2(\Omega\times(0,T))$ and $\mathbf{1}_{[s_m^{(-)},s_m^{(+)}]}\boldsymbol{F}\to0$ in $L^2(\Omega\times(0,T))^d$,

$$\int_{s_m^{(-)}}^{s_m^{(+)}}\int_\Omega(f(\boldsymbol{x},t)\Pi^{(\theta)}_{\mathcal{D}_m}u(\boldsymbol{x},t)-\boldsymbol{F}(\boldsymbol{x},t)\cdot\nabla^{(\theta)}_{\mathcal{D}_m}u(\boldsymbol{x},t))\mathrm{d}\boldsymbol{x}\mathrm{d}t\to0.\tag{5.35}$$

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

$$\begin{aligned}&\limsup_{m\to\infty}\frac12\int_\Omega(\Pi^{(\theta)}_{\mathcal{D}_m}u_m(\boldsymbol{x},s_m))^2\mathrm{d}\boldsymbol{x}\\&\quad\le\frac12\int_\Omega u_{\mathrm{ini}}(\boldsymbol{x})^2\mathrm{d}\boldsymbol{x}+\int_0^s\int_\Omega(f(\boldsymbol{x},t)\overline{u}(\boldsymbol{x},t)-\boldsymbol{F}(\boldsymbol{x},t)\cdot\nabla\overline{u}(\boldsymbol{x},t))\mathrm{d}\boldsymbol{x}\mathrm{d}t\\&\qquad-\liminf_{m\to\infty}\int_0^{s_m^{(-)}}\int_\Omega\Lambda(\boldsymbol{x},\Pi^{(\theta)}_{\mathcal{D}_m}u_m(\boldsymbol{x},t))\nabla^{(\theta)}_{\mathcal{D}_m}u_m(\boldsymbol{x},t)\cdot\nabla^{(\theta)}_{\mathcal{D}_m}u_m(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t\\&\quad\le\frac12\int_\Omega u_{\mathrm{ini}}(\boldsymbol{x})^2\mathrm{d}\boldsymbol{x}+\int_0^s\int_\Omega(f(\boldsymbol{x},t)\overline{u}(\boldsymbol{x},t)-\boldsymbol{F}(\boldsymbol{x},t)\cdot\nabla\overline{u}(\boldsymbol{x},t))\mathrm{d}\boldsymbol{x}\mathrm{d}t\end{aligned}$$

$$- \int_0^s \int_\Omega \Lambda(\boldsymbol{x}, \overline{u}(\boldsymbol{x},t))\nabla\overline{u}(\boldsymbol{x},t)\cdot\nabla\overline{u}(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t. \tag{5.36}$$

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 [29, 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

$$\begin{aligned}&\frac{1}{2}\int_\Omega \overline{u}(\boldsymbol{x},s)^2\mathrm{d}\boldsymbol{x} + \int_0^s\int_\Omega \Lambda(\boldsymbol{x},\overline{u}(\boldsymbol{x},t))\nabla\overline{u}(\boldsymbol{x},t)\cdot\nabla\overline{u}(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t\\ &\quad= \frac{1}{2}\int_\Omega u_{\mathrm{ini}}(\boldsymbol{x})^2\mathrm{d}\boldsymbol{x} + \int_0^s\int_\Omega (f(\boldsymbol{x},t)\overline{u}(\boldsymbol{x},t) - \boldsymbol{F}(\boldsymbol{x},t)\cdot\nabla\overline{u}(\boldsymbol{x},t))\mathrm{d}\boldsymbol{x}\mathrm{d}t.\end{aligned} \tag{5.37}$$

Used in (5.36), this relation gives

$$\limsup_{m\to\infty}\int_\Omega (\Pi_{\mathcal{D}_m}^{(\theta)}u_m(\boldsymbol{x},s_m))^2\mathrm{d}\boldsymbol{x} \le \int_\Omega \overline{u}(\boldsymbol{x},s)^2\mathrm{d}\boldsymbol{x}, \tag{5.38}$$

Owing to Theorem 4.32 and to Estimates (5.21) and (5.25), $(\Pi_{\mathcal{D}_m}^{(\theta)}u_m)_{m\in\mathbb{N}}$ converges to $\overline{u}$ weakly in $L^2(\Omega)$ uniformly in $[0,T]$ (in the sense of Definition 4.29). Hence, $\Pi_{\mathcal{D}_m}^{(\theta)}u_m(\cdot,s_m) \to \overline{u}(\cdot,s_m)$ 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 4.28, we infer that (5.19a) holds.

Step 4: Strong convergence of $\nabla_{\mathcal{D}_m}^{(\theta)}u_m$.

Note that, by (5.19a), $\Pi_{\mathcal{D}_m}^{(\theta)}u_m(\cdot,T) \to \overline{u}(\cdot,T)$ 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 $h_m : [0,T] \to \mathbb{R}$ be the function such that $h_m = 1$ on $[0, s_m^{(-)}]$ and $h_m = \theta$ on $(s_m^{(-)}, T]$. Using (5.37) with $s = T$, we obtain

$$\begin{aligned}&\limsup_{m\to\infty}\int_0^T\int_\Omega h_m(t)\Lambda(\boldsymbol{x},\Pi_{\mathcal{D}_m}^{(\theta)}u_m(\boldsymbol{x},t))\nabla_{\mathcal{D}_m}^{(\theta)}u_m(\boldsymbol{x},t)\cdot\nabla_{\mathcal{D}_m}^{(\theta)}u_m(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t\\ &\quad\le \frac{1}{2}\int_\Omega u_{\mathrm{ini}}(\boldsymbol{x})^2\mathrm{d}\boldsymbol{x} + \int_0^T\int_\Omega (f(\boldsymbol{x},t)\overline{u}(\boldsymbol{x},t) - \boldsymbol{F}(\boldsymbol{x},t)\cdot\nabla\overline{u}(\boldsymbol{x},t))\mathrm{d}\boldsymbol{x}\mathrm{d}t\\ &\qquad - \frac{1}{2}\int_\Omega \overline{u}(\boldsymbol{x},T)^2\mathrm{d}\boldsymbol{x}\\ &\quad= \int_0^T\int_\Omega \Lambda(\boldsymbol{x},\overline{u}(\boldsymbol{x},t))\nabla\overline{u}(\boldsymbol{x},t)\cdot\nabla\overline{u}(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t.\end{aligned}$$

Using this estimate, the strong convergence in $L^2(\Omega\times(0,T))$ of $\Pi_{\mathcal{D}_m}^{(\theta)}u_m$ to $\overline{u}$, Lemma C.4, the strong convergence in $L^2(\Omega\times(0,T))^d$ of $h_m\nabla\overline{u}$ to $\nabla\overline{u}$, the weak convergence in $L^2(\Omega\times(0,T))^d$ of $\nabla_{\mathcal{D}_m}^{(\theta)}u_m$ to $\nabla\overline{u}$, and developing the following expression in a similar fashion as (3.52), we infer that

$$\limsup_{m\to\infty}\int_0^T\int_\Omega h_m(t)\Lambda(\boldsymbol{x},\Pi_{\mathcal{D}_m}^{(\theta)}u_m(\boldsymbol{x},t))(\nabla_{\mathcal{D}_m}^{(\theta)}u_m(\boldsymbol{x},t)-\nabla\overline{u}(\boldsymbol{x},t))\\ \cdot(\nabla_{\mathcal{D}_m}^{(\theta)}u_m(\boldsymbol{x},t)-\nabla\overline{u}(\boldsymbol{x},t))\mathrm{d}\boldsymbol{x}\mathrm{d}t\le 0.$$

By coercivity of Λ and since $h_m\ge\theta\ge\frac12$, this shows that, as $m\to\infty$,

$$\int_0^T\int_\Omega\left|\nabla_{\mathcal{D}_m}^{(\theta)}u_m(\boldsymbol{x},t)-\nabla\overline{u}(\boldsymbol{x},t)\right|^2\mathrm{d}\boldsymbol{x}\mathrm{d}t\to 0.$$

This concludes the proof that $\nabla_{\mathcal{D}_m}^{(\theta)}u_m\to\nabla\overline{u}$ strongly in $L^2(\Omega\times(0,T))^d$ as $m\to\infty$. ∎

Remark 5.9 (About the discrete IBP formula (C.17)*)*
The usage in Step 2 of the "ν"-discrete integration by parts formula (C.17) is non-standard. A usual way of proceeding, see e.g. [36] or the proof of Theorem 5.20, is to analyse in this proof the convergence of $\Pi_{\mathcal{D}_m}^{(1)}u_m$ 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 (C.15) can then be applied.
Thanks to (C.17), we can however fully analyse in the proof of Theorem 5.5 the convergence of $\Pi_{\mathcal{D}_m}^{(\theta)}u_m$ without having to analyse at the same time $\Pi_{\mathcal{D}_m}^{(1)}u_m$, 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:

$$\nu(\boldsymbol{x},t,u(\boldsymbol{x},t),\nabla u(\boldsymbol{x},t))\partial_t u(\boldsymbol{x},t)-\mathrm{div}(\mu(|\nabla u(\boldsymbol{x},t)|)\nabla u(\boldsymbol{x},t))\\ =f(\boldsymbol{x},t),\ \text{for a.e. } (\boldsymbol{x},t)\in\Omega\times(0,T) \tag{5.39a}$$

with the initial condition

$$u(\boldsymbol{x},0)=u_{\mathrm{ini}}(\boldsymbol{x}),\ \text{for a.e. } \boldsymbol{x}\in\Omega, \tag{5.39b}$$

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)

(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)

On 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 to the regularised level set equation [46]. 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 5.10 (Weak solution of (5.39)). *Under Hypotheses (5.40), we say that u is a weak solution of (5.39) if*

1. *$u\in L^2(0,T;H^1_0(\Omega))$ 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*

$$\int_0^T \int_\Omega \nu(\boldsymbol{x},t,u(\boldsymbol{x},t),\nabla u(\boldsymbol{x},t))\partial_t u(\boldsymbol{x},t)v(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t + \int_0^T \int_\Omega \mu(|\nabla u(\boldsymbol{x},t)|)\nabla u(\boldsymbol{x},t)\cdot\nabla v(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t = \int_0^T \int_\Omega f(\boldsymbol{x},t)v(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t\,, \quad \forall v \in L^2(0,T;H_0^1(\Omega)). \tag{5.41}$$

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, $\mathrm{div}(\mu(|\nabla u|)\nabla u) \in L^2(\Omega\times(0,T))$.

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 model, 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

$$\forall s\in\mathbb{R}_+,\ F(s) = \int_0^s z\mu(z)\mathrm{d}z \in \left[\mu_{\min}\frac{s^2}{2}, \mu_{\max}\frac{s^2}{2}\right], \tag{5.42}$$

any sufficiently regular function u satisfies

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_\Omega F(|\nabla u(\boldsymbol{x},t)|)\mathrm{d}\boldsymbol{x} = \int_\Omega \mu(|\nabla u(\boldsymbol{x},t)|)\nabla u(\boldsymbol{x},t)\cdot\nabla\partial_t u(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t. \tag{5.43}$$

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

$$\int_0^T \int_\Omega \nu(u,\nabla u)\partial_t u(\boldsymbol{x},t)^2\mathrm{d}\boldsymbol{x}\mathrm{d}t + \int_\Omega F(|\nabla u(\boldsymbol{x},t)|)\mathrm{d}\boldsymbol{x} = \int_\Omega F(|\nabla u_{\mathrm{ini}}(\boldsymbol{x})|)\mathrm{d}\boldsymbol{x}. \tag{5.44}$$

The discrete equivalent of this essential estimate is established in Lemma 5.11 for the fully-implicit scheme (using that $x\mapsto x\mu(\boldsymbol{x})$ is strictly increasing), and in Lemma 5.15 for the semi-implicit scheme (using that μ is decreasing). The hypothesis that $x\mapsto x\mu(\boldsymbol{x})$ is instrumental, for both schemes, to prove 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

$$\begin{cases} u^{(0)} = \mathcal{I}_{\mathcal{D}} u_{\mathrm{ini}} \text{ and, for } n = 0, \dots, N-1,\ u^{(n+1)} \text{ satisfies} \\ \displaystyle\int_{t^{(n)}}^{t^{(n+1)}} \int_\Omega \nu(\boldsymbol{x}, t, \Pi_{\mathcal{D}} u^{(n+1)}, \nabla_{\mathcal{D}} u^{(n+1)}) \delta_{\mathcal{D}}^{(n+\frac{1}{2})} u(\boldsymbol{x}) \Pi_{\mathcal{D}} v(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}\mathrm{d}t \\ \quad + \delta t^{(n+\frac{1}{2})} \displaystyle\int_\Omega \mu(|\nabla_{\mathcal{D}} u^{(n+1)}(\boldsymbol{x})|) \nabla_{\mathcal{D}} u^{(n+1)}(\boldsymbol{x}) \cdot \nabla_{\mathcal{D}} v(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \\ \quad = \displaystyle\int_{t^{(n)}}^{t^{(n+1)}} \int_\Omega f(\boldsymbol{x}, t) \Pi_{\mathcal{D}} v(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}\mathrm{d}t\,, \quad \forall v \in X_{\mathcal{D},0}. \end{cases} \tag{5.45}$$

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 5.11 ($L^2(\Omega \times (0,T))$ 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 $m = 1, \dots, N$,*

$$\nu_{\min} \int_0^{t^{(m)}} \int_\Omega \delta_{\mathcal{D}} u(\boldsymbol{x}, t)^2 \mathrm{d}\boldsymbol{x}\mathrm{d}t + \mu_{\min} \left\| \nabla_{\mathcal{D}} u^{(m)} \right\|_{L^2(\Omega)^d}^2 \\ \leq \mu_{\max} \left\| \nabla_{\mathcal{D}} \mathcal{I}_{\mathcal{D}} u_{\mathrm{ini}} \right\|_{L^2(\Omega)^d}^2 + \frac{1}{\nu_{\min}} \left\| f \right\|_{L^2(\Omega \times (0,T))}^2. \tag{5.46}$$

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 $n = 0, \dots, m-1$ leads to

$$\nu_{\min} \int_0^{t^{(m)}} \int_\Omega \delta_{\mathcal{D}} u(\boldsymbol{x}, t)^2 \mathrm{d}\boldsymbol{x}\mathrm{d}t \\ + \sum_{n=0}^{m-1} \int_\Omega \mu(|\nabla_{\mathcal{D}} u^{(n+1)}(\boldsymbol{x})|) \nabla_{\mathcal{D}} u^{(n+1)}(\boldsymbol{x}) \cdot \left[\nabla_{\mathcal{D}} u^{(n+1)}(\boldsymbol{x}) - \nabla_{\mathcal{D}} u^{(n)}(\boldsymbol{x}) \right] \mathrm{d}\boldsymbol{x} \\ \leq \int_0^{t^{(m)}} \int_\Omega f(\boldsymbol{x}, t) \delta_{\mathcal{D}} u(\boldsymbol{x}, t) \mathrm{d}\boldsymbol{x}\mathrm{d}t. \tag{5.47}$$

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

$$|\nabla_{\mathcal{D}} u^{(n+1)}(\boldsymbol{x})| \left[|\nabla_{\mathcal{D}} u^{(n+1)}(\boldsymbol{x})| - |\nabla_{\mathcal{D}} u^{(n)}(\boldsymbol{x})|\right]$$
$$\leq \nabla_{\mathcal{D}} u^{(n+1)}(\boldsymbol{x}) \cdot \left[\nabla_{\mathcal{D}} u^{(n+1)}(\boldsymbol{x}) - \nabla_{\mathcal{D}} u^{(n)}(\boldsymbol{x})\right]. \quad (5.49)$$

Combining (5.48) and (5.49) and plugging the result into (5.47) yields

$$\begin{aligned}
&\nu_{\min} \int_0^{t^{(m)}} \int_\Omega \delta_{\mathcal{D}} u(\boldsymbol{x},t)^2 \mathrm{d}\boldsymbol{x}\mathrm{d}t \\
&\quad + \sum_{n=0}^{m-1} \int_\Omega \left[F(|\nabla_{\mathcal{D}} u^{(n+1)}(\boldsymbol{x})|) - F(|\nabla_{\mathcal{D}} u^{(n)}(\boldsymbol{x})|)\right] \mathrm{d}\boldsymbol{x} \\
&\quad \leq \int_0^{t^{(m)}} \int_\Omega f(\boldsymbol{x},t)\delta_{\mathcal{D}} u(\boldsymbol{x},t) \mathrm{d}\boldsymbol{x}\mathrm{d}t. \qquad (5.50)
\end{aligned}$$

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 (C.9). Since the range of F is in $[\mu_{\min} s^2/2, \mu_{\max} s^2/2]$, this gives

$$\begin{aligned}
&\nu_{\min} \int_0^{t^{(m)}} \int_\Omega \delta_{\mathcal{D}} u(\boldsymbol{x},t)^2 \mathrm{d}\boldsymbol{x}\mathrm{d}t + \frac{\mu_{\min}}{2} \int_\Omega |\nabla_{\mathcal{D}} u^{(m)}(\boldsymbol{x})|^2 \mathrm{d}\boldsymbol{x} \\
&\leq \frac{\mu_{\max}}{2} \int_\Omega |\nabla_{\mathcal{D}} u^{(0)}(\boldsymbol{x})|^2 \mathrm{d}\boldsymbol{x} + \|f\|_{L^2(\Omega\times(0,T))} \|\delta_{\mathcal{D}} u\|_{L^2(\Omega\times(0,t^{(m)}))} \\
&\leq \frac{\mu_{\max}}{2} \int_\Omega |\nabla_{\mathcal{D}} \mathcal{I}_{\mathcal{D}} u_{\mathrm{ini}}(\boldsymbol{x})|^2 \mathrm{d}\boldsymbol{x} + \frac{1}{2\nu_{\min}} \|f\|^2_{L^2(\Omega\times(0,T))} \\
&\quad + \frac{\nu_{\min}}{2} \|\delta_{\mathcal{D}} u\|^2_{L^2(\Omega\times(0,t^{(m)}))}.
\end{aligned}$$

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 $(v_i)_{i=1,\dots,M}$:

$$\begin{aligned}\Phi(w,\lambda)_i = \int_{t^{(n)}}^{t^{(n+1)}} \int_\Omega \nu_\lambda(\boldsymbol{x},t,\Pi_{\mathcal{D}} w(\boldsymbol{x}),\nabla_{\mathcal{D}} w(\boldsymbol{x})) \\ \times \frac{\Pi_{\mathcal{D}} w(\boldsymbol{x}) - \Pi_{\mathcal{D}} u^{(n)}(\boldsymbol{x})}{\delta t^{(n+\frac{1}{2})}} \Pi_{\mathcal{D}} v_i(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}\mathrm{d}t \\ + \delta t^{(n+\frac{1}{2})} \int_\Omega \mu_\lambda(|\nabla_{\mathcal{D}} w(\boldsymbol{x})|) \nabla_{\mathcal{D}} w(\boldsymbol{x}) \cdot \nabla_{\mathcal{D}} v_i(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \\ - \int_{t^{(n)}}^{t^{(n+1)}} \int_\Omega f(\boldsymbol{x},t) \Pi_{\mathcal{D}} v_i(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}\mathrm{d}t.\end{aligned}$$

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 C.1 page 403) 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 implicity GS for (5.39).

Lemma 5.12 (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* $(\nabla_{\mathcal{D}_m}\mathcal{I}_{\mathcal{D}_m}u_{\rm ini})_{m\in\mathbb{N}}$ *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*

$$\begin{aligned}&\overline{u} \in L^\infty(0,T;H_0^1(\Omega)) \cap C([0,T];L^2(\Omega)) \text{ with } \partial_t\overline{u} \in L^2(\Omega\times(0,T))\\ &\qquad \text{and } u(\cdot,0) = u_{\rm ini}\,,\\ &\overline{G} \in L^2(\Omega\times(0,T))^d\,,\ \text{and}\\ &\overline{w} \in L^2(\Omega\times(0,T))\end{aligned}$$

such that, along a subsequence as $m \to \infty$*,*

- $\sup_{t\in[0,T]} \|\Pi^{(1)}_{\mathcal{D}_m}u_m(t) - \overline{u}(t)\|_{L^2(\Omega)} \to 0$*,*
- $\nabla^{(1)}_{\mathcal{D}_m}u_m$ *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*

$$\begin{aligned}&\int_0^T\int_\Omega G_{\mathcal{D}_m}(\boldsymbol{x},t)\cdot\nabla^{(1)}_{\mathcal{D}_m}u_m(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t\\ &\qquad\longrightarrow \int_0^T\int_\Omega \overline{G}(\boldsymbol{x},t)\cdot\nabla\overline{u}(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t.\end{aligned} \tag{5.54}$$

Remark 5.13. Note that, 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 5.14, 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 5.11, 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 4.31 thus provides $\overline{u} \in L^\infty(0,T;H_0^1(\Omega)) \cap C([0,T];L^2(\Omega))$ 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 4.7.
The definition (4.3) gives $\Pi^{(1)}_{\mathcal{D}_m}u_m(0) = \Pi_{\mathcal{D}_m}u_m^{(0)} = \Pi_{\mathcal{D}_m}\mathcal{I}_{\mathcal{D}_m}u_{\rm ini}$. The space–time-consistency of $((\mathcal{D}_T)_m)_{m\in\mathbb{N}}$ then yields $\Pi^{(1)}_{\mathcal{D}_m}u_m(0) \to u_{\rm ini}$ 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 $\overline{u}(\cdot,0) = u_{\rm ini}$.
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 4.9 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 C.3 page 403) 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

$$\begin{aligned}\lim_{m\to\infty}\int_0^T\int_\Omega G_{\mathcal{D}_m}(\boldsymbol{x},t)\cdot\nabla^{(1)}_{\mathcal{D}_m}u_m(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t &= \int_0^T\int_\Omega \overline{w}(\boldsymbol{x},t)\overline{u}(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t\\ &= \int_0^T\int_\Omega \overline{G}(\boldsymbol{x},t)\cdot\nabla\overline{u}(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t\end{aligned}$$

and the proof of (5.54) is complete. ∎

We now state and prove the convergence of the fully implicit GS for (5.39).

Theorem 5.14.
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 $(\nabla_{\mathcal{D}_m}\mathcal{I}_{\mathcal{D}_m}u_{\rm ini})_{m\in\mathbb{N}}$ *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 5.10 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 5.12. 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 $V, W \in L^2(\Omega\times(0,T))^d$,

$$\begin{aligned}\int_0^T\int_\Omega [\mu(|W|)W - \mu(|V|)V]\cdot[W-V]\,\mathrm{d}\boldsymbol{x}\mathrm{d}t&\\ \geq \alpha\,\||W|-|V|\|^2_{L^2(\Omega\times(0,T))^d}.&\end{aligned}\tag{5.56}$$

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 C.4 page 404, 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 C.3 page 403 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 4.9 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 5.10 are ascertained in Lemma 5.12, 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

$$\begin{cases} u^{(0)} = \mathcal{I}_{\mathcal{D}} u_{\rm ini} \text{ and, for } n = 0, \dots, N-1,\ u^{(n+1)} \text{ satisfies} \\ \displaystyle\int_{t^{(n)}}^{t^{(n+1)}} \int_\Omega \nu(\boldsymbol{x}, t, \Pi_{\mathcal{D}} u^{(n)}, \nabla_{\mathcal{D}} u^{(n)}) \delta_{\mathcal{D}}^{(n+\frac{1}{2})} u(\boldsymbol{x}) \Pi_{\mathcal{D}} v(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}\mathrm{d}t \\ \quad + \delta t^{(n+\frac{1}{2})} \displaystyle\int_\Omega \mu(|\nabla_{\mathcal{D}} u^{(n)}|) \nabla_{\mathcal{D}} u^{(n+1)} \cdot \nabla_{\mathcal{D}} v(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \\ \quad = \displaystyle\int_{t^{(n)}}^{t^{(n+1)}} \int_\Omega f(\boldsymbol{x}, t) \Pi_{\mathcal{D}} v(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}\mathrm{d}t\,, \qquad \forall v \in X_{\mathcal{D},0}. \end{cases} \tag{5.60}$$

Quite naturally, the analysis of this semi-implicit implicit scheme uses both $\Pi_{\mathcal{D}}^{(1)}$, $\nabla_{\mathcal{D}}^{(1)}$ and $\Pi_{\mathcal{D}}^{(0)}$, $\nabla_{\mathcal{D}}^{(0)}$. Recall that the definition of these latter operators (see (4.2)):

$$\Pi_{\mathcal{D}}^{(0)} u(\boldsymbol{x}, t) = \Pi_{\mathcal{D}} u^{(n)}(\boldsymbol{x}) \text{ and } \nabla_{\mathcal{D}}^{(0)} u(\boldsymbol{x}, t) = \nabla_{\mathcal{D}} u^{(n)}(\boldsymbol{x}),$$
$$\text{for a.e. } (\boldsymbol{x}, t) \in \Omega \times (t^{(n)}, t^{(n+1)}),\ \forall n = 0, \dots, N-1.$$

Estimates and existence of a solution to the semi-implicit scheme

Lemma 5.15 ($L^2(\Omega \times (0,T))$ 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 $m = 1, \dots, N$,*

$$\begin{aligned} &\nu_{\min} \int_0^{t^{(m)}} \int_\Omega \delta_{\mathcal{D}} u(\boldsymbol{x}, t)^2 \mathrm{d}\boldsymbol{x}\mathrm{d}t + \mu_{\min} \left\| \nabla_{\mathcal{D}} u^{(m)} \right\|_{L^2(\Omega)^d}^2 \\ &+ \mu_{\min} \sum_{n=0}^{m-1} \int_\Omega \left| \nabla_{\mathcal{D}} u^{(n+1)}(\boldsymbol{x}) - \nabla_{\mathcal{D}} u^{(n)}(\boldsymbol{x}) \right|^2 \mathrm{d}\boldsymbol{x} \\ &\qquad \le \mu_{\max} \left\| \nabla_{\mathcal{D}} \mathcal{I}_{\mathcal{D}} u_{\rm ini} \right\|_{L^2(\Omega)^d}^2 + \frac{1}{\nu_{\min}} \left\| f \right\|_{L^2(\Omega \times (0,T))}^2. \end{aligned} \tag{5.61}$$

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

$$\begin{aligned} &F(|\nabla_{\mathcal{D}} u^{(n+1)}(\boldsymbol{x})|) - F(|\nabla_{\mathcal{D}} u^{(n)}(\boldsymbol{x})|) \\ &\qquad + \frac{\mu_{\min}}{2} \left| \nabla_{\mathcal{D}} u^{(n+1)}(\boldsymbol{x}) - \nabla_{\mathcal{D}} u^{(n)}(\boldsymbol{x}) \right|^2 \\ &\le \mu(|\nabla_{\mathcal{D}} u^{(n)}(\boldsymbol{x})|) \nabla_{\mathcal{D}} u^{(n+1)}(\boldsymbol{x}) \cdot \left[\nabla_{\mathcal{D}} u^{(n+1)}(\boldsymbol{x}) - \nabla_{\mathcal{D}} u^{(n)}(\boldsymbol{x}) \right]. \end{aligned} \tag{5.63}$$

Estimate (5.61) is then established as the proof of Lemma 5.11, by plugging $v = \frac{u^{(n+1)}-u^{(n)}}{\delta t^{(n+\frac{1}{2})}}$ in (5.60), summing over $n = 0, \ldots, m-1$, 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 5.12.

Lemma 5.16 (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 $m\to\infty$,

- $\sup_{t\in[0,T]} \|\mathit{\Pi}^{(1)}_{\mathcal{D}_m} u_m(t) - \overline{u}(t)\|_{L^2(\Omega)} \to 0$,
- $\nabla^{(0)}_{\mathcal{D}_m} u_m$ *and* $\nabla^{(1)}_{\mathcal{D}_m} u_m$ *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*

$$\int_0^T \int_\Omega (\widetilde{G}_{\mathcal{D}_m}(\boldsymbol{x},t) - \widehat{G}_{\mathcal{D}_m}(\boldsymbol{x},t)) \cdot \nabla^{(0)}_{\mathcal{D}} u_m(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t \to 0, \tag{5.68}$$

- $\widetilde{w}_{\mathcal{D}_m}$ *converges weakly to* $\overline{w}$ *in* $L^2(\Omega\times(0,T))$,
- *it holds*

$$\begin{aligned}
&\int_0^T \int_\Omega \widehat{G}_{\mathcal{D}_m}(\boldsymbol{x},t) \cdot \nabla^{(0)}_{\mathcal{D}_m} u_m(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t \\
&\qquad\qquad \longrightarrow \int_0^T \int_\Omega \overline{G}(\boldsymbol{x},t)\cdot\nabla\overline{u}(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t.
\end{aligned} \tag{5.69}$$

Proof. The proof is similar to the proof of Lemma 5.12. The *a priori* estimate (5.61) provide the existence of $\overline{u}$ such that $(\mathit{\Pi}^{(1)}_{\mathcal{D}_m} u_m)_{m\in\mathbb{N}}$, $(\nabla^{(1)}_{\mathcal{D}_m} u_m)_{m\in\mathbb{N}}$ 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

$$\begin{aligned}
&\left\|\nabla^{(1)}_{\mathcal{D}_m} u_m - \nabla^{(0)}_{\mathcal{D}_m} u_m\right\|^2_{L^2(\Omega\times(0,T))^d} \\
&\qquad = \sum_{n=0}^{N-1} \eth^{(n+\frac12)} \int_\Omega \left|\nabla_{\mathcal{D}_m} u_m^{(n+1)}(\boldsymbol{x}) - \nabla_{\mathcal{D}_m} u_m^{(n)}(\boldsymbol{x})\right|^2 \mathrm{d}\boldsymbol{x} \\
&\qquad \le \eth_{\mathcal{D}_m} \sum_{n=0}^{N-1} \int_\Omega \left|\nabla_{\mathcal{D}_m} u_m^{(n+1)}(\boldsymbol{x}) - \nabla_{\mathcal{D}_m} u_m^{(n)}(\boldsymbol{x})\right|^2 \mathrm{d}\boldsymbol{x}\,,
\end{aligned} \tag{5.70}$$

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 5.12. The GS (5.67) applied to $\mathcal{D} = \mathcal{D}_m$ and $v = u_m$ and the strong convergence of $\mathit{\Pi}^{(1)}_{\mathcal{D}_m} u_m$ then show that

$$\begin{aligned}
\int_0^T\int_\Omega \widetilde{G}_{\mathcal{D}_m}(\boldsymbol{x},t)\cdot\nabla^{(1)}_{\mathcal{D}_m} u_m(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t &= \int_0^T\int_\Omega \widetilde{w}_{\mathcal{D}_m}(\boldsymbol{x},t)\mathit{\Pi}^{(1)}_{\mathcal{D}_m} u_m(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t \\
&\to \int_0^T\int_\Omega \overline{w}(\boldsymbol{x},t)\overline{u}(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t \\
&= \int_0^T\int_\Omega \overline{G}(\boldsymbol{x},t)\cdot\nabla\overline{u}(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t.
\end{aligned}$$

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 5.14, 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 5.17.
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 $(\nabla_{\mathcal{D}_m}\mathcal{I}_{\mathcal{D}_m}u_{\rm ini})_{m\in\mathbb{N}}$ *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 5.10 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 problems

5.3.1 Model

We consider here an evolution problem based on a Leray–Lions operator, with non-homogeneous Neumann boundary conditions and non-local dependency on the lower order terms. The model reads

$$\begin{aligned}
&\partial_t\overline{u}-\operatorname{div}(\boldsymbol{a}(\boldsymbol{x},\overline{u},\nabla\overline{u}))=f && \text{in } \Omega\times(0,T),\\
&\overline{u}(\boldsymbol{x},0)=u_{\rm ini}(\boldsymbol{x}) && \text{in } \Omega,\\
&\boldsymbol{a}(\boldsymbol{x},\overline{u},\nabla\overline{u})\cdot\boldsymbol{n}=g && \text{on } \partial\Omega\times(0,T),
\end{aligned}\tag{5.73}$$

where $\boldsymbol{a}$ satisfies (3.96a)–(3.96d) and

$$\begin{aligned}
&\bullet\ T\in(0,+\infty),\\
&\bullet\ u_{\rm ini}\in L^2(\Omega),\\
&\bullet\ f\in L^{p'}(\Omega\times(0,T)) \text{ and } g\in L^{p'}(\partial\Omega\times(0,T))\,, \text{ where } p'=\tfrac{p}{p-1}.
\end{aligned}\tag{5.74}$$

The non-linear equation (5.73) covers a number of models, including semi-linear ones appearing in image processing [18, 20]. The analysis of the GDM applied to (5.73) with homogeneous Dirichlet boundary conditions is done in [36].
The precise notion of solution to (5.73) is the following:

$$\left\{\begin{aligned}
&\overline{u}\in L^p(0,T;W^{1,p}(\Omega))\cap C([0,T];L^2(\Omega))\,,\ \overline{u}(\cdot,0)=u_{\rm ini},\\
&\partial_t\overline{u}\in L^{p'}(0,T;(W^{1,p}(\Omega))') \text{ and}\\
&\int_0^T\langle\partial_t\overline{u}(\cdot,t),\overline{v}(\cdot,t)\rangle_{(W^{1,p}(\Omega))',W^{1,p}(\Omega)}\mathrm{d}t\\
&+\int_0^T\int_\Omega\boldsymbol{a}(\boldsymbol{x},\overline{u}(\cdot,t),\nabla\overline{u}(\boldsymbol{x},t))\cdot\nabla\overline{v}(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t\\
&\quad=\int_0^T\int_\Omega f(\boldsymbol{x},t)\overline{v}(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t+\int_0^T\int_{\partial\Omega}g(\boldsymbol{x},t)\gamma\overline{v}(\boldsymbol{x},t)\mathrm{d}s(\boldsymbol{x})\mathrm{d}t\,,\\
&\quad\forall\overline{v}\in L^p(0;T;W^{1,p}(\Omega)).
\end{aligned}\right.\tag{5.75}$$

Remark 5.18. 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 [29, 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 5.19. Using regularisation and integration-by-parts techniques [29, Section 2.5.2], it is possible to see that any solution $\overline{u}$ to (5.75) also satisfies, for any $s \in [0,T]$,

$$\begin{aligned} \frac{1}{2}\left\|\overline{u}(s)\right\|_{L^2(\Omega)}^2 + \int_0^s \int_\Omega \boldsymbol{a}(\boldsymbol{x},\overline{u}(\cdot,\tau),\nabla\overline{u}(\boldsymbol{x},\tau))\mathrm{d}\boldsymbol{x}\mathrm{d}\tau \\ = \frac{1}{2}\left\|u_{\mathrm{ini}}\right\|_{L^2(\Omega)}^2 + \int_0^s \int_\Omega f(\boldsymbol{x},\tau)\overline{u}(\boldsymbol{x},\tau)\mathrm{d}\boldsymbol{x}\mathrm{d}\tau \\ + \int_0^s \int_{\partial\Omega} g(\boldsymbol{x},\tau)\gamma\overline{u}(\boldsymbol{x},\tau)\mathrm{d}s(\boldsymbol{x})\mathrm{d}\tau. \end{aligned} \tag{5.77}$$

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:

$$\left\{\begin{array}{l} \overline{u} \in L^p(0,T;W^{1,p}(\Omega)) \cap L^\infty(0,T;L^2(\Omega)) \text{ and,} \\ \text{for all } \overline{v} \in C^1([0,T];W^{1,p}(\Omega) \cap L^2(\Omega)) \text{ such that } \overline{v}(\cdot,T) = 0, \\ -\displaystyle\int_0^T \int_\Omega \overline{u}(\boldsymbol{x},t)\partial_t\overline{v}(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t - \int_\Omega u_{\mathrm{ini}}(\boldsymbol{x})\overline{v}(\boldsymbol{x},0)\mathrm{d}\boldsymbol{x} \\ + \displaystyle\int_0^T \int_\Omega \boldsymbol{a}(\boldsymbol{x},\overline{u}(\cdot,t),\nabla\overline{u}(\boldsymbol{x},t)) \cdot \nabla\overline{v}(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t \\ \quad = \displaystyle\int_0^T \int_\Omega f(\boldsymbol{x},t)\overline{v}(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t + \int_0^T \int_{\partial\Omega} g(\boldsymbol{x},t)\gamma\overline{v}(\boldsymbol{x},t)\mathrm{d}s(\boldsymbol{x})\mathrm{d}t. \end{array}\right. \tag{5.78}$$

To prove this equivalence, we use [29, 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 [29, 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

$$\begin{cases} u^{(0)} = \mathcal{I}_\mathcal{D} u_{\rm ini} \in X_\mathcal{D} \text{ and, for all } n = 0, \dots, N-1,\ u^{(n+1)} \text{ satisfies} \\ \displaystyle\int_\Omega \delta_\mathcal{D}^{(n+\frac{1}{2})} u(\boldsymbol{x}) \Pi_\mathcal{D} v(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \\ \displaystyle + \int_\Omega \boldsymbol{a}\left(\boldsymbol{x}, \Pi_\mathcal{D} u^{(n+\theta)}, \nabla_\mathcal{D} u^{(n+\theta)}(\boldsymbol{x})\right) \cdot \nabla_\mathcal{D} v(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \\ \qquad \displaystyle = \frac{1}{\delta t^{(n+\frac{1}{2})}} \int_{t^{(n)}}^{t^{(n+1)}} \int_\Omega f(\boldsymbol{x}, t) \Pi_\mathcal{D} v(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \mathrm{d}t \\ \qquad\quad \displaystyle + \frac{1}{\delta t^{(n+\frac{1}{2})}} \int_{t^{(n)}}^{t^{(n+1)}} \int_{\partial\Omega} g(\boldsymbol{x}, t) \mathbb{T}_\mathcal{D} v(\boldsymbol{x}) \mathrm{d}s(\boldsymbol{x}) \mathrm{d}t\,, \quad \forall v \in X_\mathcal{D}. \end{cases} \tag{5.79}$$

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-Nicholson time discretisation.
Recalling the notations in (4.2), we now state our initial convergence results for this GS.

Theorem 5.20 (Convergence of the GS for transient Leray–Lions). *Under Assumptions* (3.96a)–(3.96d) *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 4.29),*
- $\nabla_{\mathcal{D}_m}^{(\theta)} u_m$ *converges to* $\nabla\overline{u}$ *weakly in* $L^p(\Omega \times (0,T))^d$*,*

where $\overline{u}$ *is a solution to* (5.75)*.*

Remark 5.21. As for the stationary problem (see Remark 3.35), 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 5.22 (Uniform-in-time convergence of the GS). *Under the assumptions and notations of Theorem 5.20, 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 operator $\boldsymbol{a}$ is strictly monotone, then a strong convergence result can also be stated on the gradients.

Theorem 5.23 (Strong convergence of the gradients in the strictly monotone case). *Let us assume the hypotheses of Theorem 5.20, and that $\boldsymbol{a}$ is strictly monotone in the sense of* (3.98)*. Then, with the same notations and along the same subsequence as in Theorem 5.20, $\nabla^{(\theta)}_{\mathcal{D}_m} u_m$ 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 5.24 ($L^\infty(0,T;L^2(\Omega))$ estimate, discrete $L^p(0,T;W^{1,p}(\Omega))$ estimate, and existence of a solution to the GS). *Under Hypotheses* (3.96a)–(3.96d) *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, $C_P \ge C_{\mathcal{D}}$, $C_{\rm ini} \ge \|\Pi_{\mathcal{D}}\mathcal{I}_{\mathcal{D}}u_{\rm ini}\|_{L^2(\Omega)}$, f, g and $\underline{a}$ such that, for any solution u to this scheme,*

$$\begin{aligned}&\sup_{t\in[0,T]} \left\|\Pi^{(1)}_{\mathcal{D}} u(t)\right\|_{L^2(\Omega)} \le C_{12}\,, \quad \sup_{t\in[0,T]} \left\|\Pi^{(\theta)}_{\mathcal{D}} u(t)\right\|_{L^2(\Omega)} \le C_{12}\\ &\text{and } \left\|\nabla^{(\theta)}_{\mathcal{D}} u\right\|_{L^p(\Omega\times(0,T))^d} \le C_{12}.\end{aligned} \tag{5.80}$$

Proof. Let us first prove the estimates. Recall (5.22), that is

$$\delta t^{(n+\frac12)}\delta^{(n+\frac12)}_{\mathcal{D}} u\ \Pi_{\mathcal{D}}u^{(n+\theta)} \ge \frac12\left((\Pi_{\mathcal{D}}u^{(n+1)})^2 - (\Pi_{\mathcal{D}}u^{(n)})^2\right),$$

choose $v = \delta t^{(n+\frac12)}u^{(n+\theta)}$ in (5.79), and sum on $n=0,\dots,k-1$ for a given $k\in\{1,\dots,N\}$. This yields

$$\begin{aligned}&\frac12\left\|\Pi_{\mathcal{D}}u^{(k)}\right\|^2_{L^2(\Omega)} + \int_0^{t^{(k)}}\int_\Omega \boldsymbol{a}\left(\boldsymbol{x},\Pi^{(\theta)}_{\mathcal{D}}u(\cdot,t),\nabla^{(\theta)}_{\mathcal{D}}u(\boldsymbol{x},t)\right)\cdot\nabla^{(\theta)}_{\mathcal{D}}u(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t\\ &\quad\le \frac12\left\|\Pi_{\mathcal{D}}u^{(0)}\right\|^2_{L^2(\Omega)} + \int_0^{t^{(k)}}\int_\Omega f(\boldsymbol{x},t)\Pi^{(\theta)}_{\mathcal{D}}u(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t\end{aligned}$$

$$+ \int_0^{t^{(k)}} \int_{\partial\Omega} g(\boldsymbol{x},t)\mathbb{T}_{\mathcal{D}}^{(\theta)} u(\boldsymbol{x},t)\mathrm{d}s(\boldsymbol{x})\mathrm{d}t. \tag{5.81}$$

In particular, owing to the coercivity property (3.96b) of $\boldsymbol{a}$, and using Hölder's inequality and Young's inequality (C.9) (the latter with $\varepsilon = \frac{p\underline{a}}{4C_{\mathcal{D}}^p}$),

$$\begin{aligned}
&\frac{1}{2}\left\|\Pi_{\mathcal{D}} u^{(k)}\right\|_{L^2(\Omega)}^2 + \underline{a}\int_0^{t^{(k)}} \left\|\nabla_{\mathcal{D}}^{(\theta)} u(\cdot,t)\right\|_{L^p(\Omega)^d}^p \mathrm{d}t \\
&\le \frac{1}{2}\left\|\Pi_{\mathcal{D}} u^{(0)}\right\|_{L^2(\Omega)}^2 + \frac{4^{1/(p-1)}C_{\mathcal{D}}^{p'}}{(p\underline{a})^{1/(p-1)}\ p'}\left\|f\right\|_{L^{p'}(\Omega\times(0,t^{(k)}))}^{p'} \\
&\quad + \frac{\underline{a}}{4C_{\mathcal{D}}^p}\left\|\Pi_{\mathcal{D}}^{(\theta)} u\right\|_{L^p(\Omega\times(0,t^{(k)}))}^p + \frac{4^{1/(p-1)}C_{\mathcal{D}}^{p'}}{(p\underline{a})^{1/(p-1)}\ p'}\left\|g\right\|_{L^{p'}(\partial\Omega\times(0,t^{(k)}))}^{p'} \\
&\quad + \frac{\underline{a}}{4C_{\mathcal{D}}^p}\left\|\mathbb{T}_{\mathcal{D}}^{(\theta)} u\right\|_{L^p(\partial\Omega\times(0,t^{(k)}))}^p.
\end{aligned}$$

Apply the definition (2.26) of $C_{\mathcal{D}}$ and recall that $u^{(0)} = \mathcal{I}_{\mathcal{D}} u_{\mathrm{ini}}$ to deduce

$$\begin{aligned}
&\frac{1}{2}\left\|\Pi_{\mathcal{D}} u^{(k)}\right\|_{L^2(\Omega)}^2 + \frac{\underline{a}}{2}\left\|\nabla_{\mathcal{D}}^{(\theta)} u\right\|_{L^p(\Omega\times(0,t^{(k)}))^d}^p \\
&\qquad \le \frac{1}{2}\left\|\Pi_{\mathcal{D}}\mathcal{I}_{\mathcal{D}} u_{\mathrm{ini}}\right\|_{L^2(\Omega)}^2 + \frac{4^{1/(p-1)}C_{\mathcal{D}}^{p'}}{(p\underline{a})^{1/(p-1)}\ p'}\left\|f\right\|_{L^{p'}(\Omega\times(0,t^{(k)}))}^{p'} \\
&\qquad\quad + \frac{4^{1/(p-1)}C_{\mathcal{D}}^{p'}}{(p\underline{a})^{1/(p-1)}\ p'}\left\|g\right\|_{L^{p'}(\partial\Omega\times(0,t^{(k)}))}^{p'}.
\end{aligned}$$

This establishes the estimates on $\Pi_{\mathcal{D}}^{(1)} u$ and $\nabla_{\mathcal{D}}^{(\theta)} u$. The estimate on $\Pi_{\mathcal{D}}^{(\theta)} u$ 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 3.34, 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 4.21).

Lemma 5.25 (Estimate on the dual norm of the discrete time derivative). *Under Hypotheses* (3.96a)–(3.96d) *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*

$$\int_0^T \|\delta_{\mathcal{D}} u(t)\|_{\star,\mathcal{D}}^{p'}\,\mathrm{d}t \le C_{13}, \tag{5.82}$$

where the dual norm $\|\cdot\|_{\star,\mathcal{D}}$ is given by Definition 4.18.

Proof. Let us take a generic $v \in X_{\mathcal{D}}$ as a test function in (5.79). We have, thanks to Assumption (3.96d) on $\boldsymbol{a}$,

$$\begin{aligned}\int_\Omega \delta_{\mathcal{D}}^{(n+\frac12)} u(\boldsymbol{x}) \Pi_{\mathcal{D}} v(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} &\\ &\le \int_\Omega \left(\overline{a}(\boldsymbol{x}) + \mu |\nabla_{\mathcal{D}} u^{(n+\theta)}(\boldsymbol{x})|^{p-1}\right) |\nabla_{\mathcal{D}} v(\boldsymbol{x})| \mathrm{d}\boldsymbol{x} \\ &\quad + \frac{1}{\delta t^{(n+\frac12)}} \int_{t^{(n)}}^{t^{(n+1)}} \int_\Omega f(\boldsymbol{x},t) \Pi_{\mathcal{D}} v(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}\mathrm{d}t \\ &\quad + \frac{1}{\delta t^{(n+\frac12)}} \int_{t^{(n)}}^{t^{(n+1)}} \int_{\partial\Omega} g(\boldsymbol{x},t) \mathbb{T}_{\mathcal{D}} v(\boldsymbol{x}) \mathrm{d}s(\boldsymbol{x}) \mathrm{d}t.\end{aligned}$$

This leads, by definition (2.26) of $C_{\mathcal{D}}$, to the existence of $C_{14} > 0$ depending only on p, μ such that

$$\begin{aligned}\int_\Omega \delta_{\mathcal{D}}^{(n+\frac12)} u(\boldsymbol{x}) \Pi_{\mathcal{D}} v(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} &\\ &\le C_{14} \left(\begin{array}{l} \|\overline{a}\|_{L^{p'}(\Omega)} + \left\| \nabla_{\mathcal{D}} u^{(n+\theta)} \right\|_{L^p(\Omega)^d}^{p-1} \\ + \dfrac{C_{\mathcal{D}}}{\delta t^{(n+\frac12)}} \displaystyle\int_{t^{(n)}}^{t^{(n+1)}} \|f(\cdot,t)\|_{L^{p'}(\Omega)} \mathrm{d}t \\ + \dfrac{C_{\mathcal{D}}}{\delta t^{(n+\frac12)}} \displaystyle\int_{t^{(n)}}^{t^{(n+1)}} \|g(\cdot,t)\|_{L^{p'}(\partial\Omega)} \mathrm{d}t \end{array} \right) \|v\|_{\mathcal{D}}.\end{aligned}$$

Taking the supremum on $v \in X_{\mathcal{D}}$ such that $\|v\|_{\mathcal{D}} = 1$ gives

$$\begin{aligned}\left\| \delta_{\mathcal{D}}^{(n+\frac12)} u \right\|_{\star,\mathcal{D}} &\le C_{14} \|\overline{a}\|_{L^{p'}(\Omega)} + C_{14} \left\| \nabla_{\mathcal{D}} u^{(n+\theta)} \right\|_{L^p(\Omega)^d}^{p-1} \\ &\quad + \frac{C_{14} C_{\mathcal{D}}}{\delta t^{(n+\frac12)}} \int_{t^{(n)}}^{t^{(n+1)}} \|f(\cdot,t)\|_{L^{p'}(\Omega)} \mathrm{d}t + \frac{C_{14} C_{\mathcal{D}}}{\delta t^{(n+\frac12)}} \int_{t^{(n)}}^{t^{(n+1)}} \|g(\cdot,t)\|_{L^{p'}(\partial\Omega)} \mathrm{d}t.\end{aligned}$$

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 (C.14), using Jensen's inequality for the integral terms, multiplying by $\delta t^{(n+\frac12)}$, summing on n and invoking Lemma 5.24 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 model (5.73).

Proof of Theorem 5.20.

Step 1: Application of compactness results.
The definition (2.25) of $\|\cdot\|_{\mathcal{D}_m}$ and Estimates (5.80) and (5.82) show that the hypotheses of Lemma 4.7 (regularity of the limit) and Theorem 4.21 (Aubin–Simon for GD) are satisfied by $(u_m)_{m\in\mathbb{N}}$. There exists therefore $\overline{u} \in L^p(0,T;W^{1,p}(\Omega))$ 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 4.32 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 4.29).
Estimates (5.80) and Assumption (3.96d) 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

$$P_{\mathcal{D}_m}\varphi(t) = \operatorname*{argmin}_{w\in X_{\mathcal{D}_m}} \Big(\|\Pi_{\mathcal{D}_m} w - \varphi(t)\|_{L^{\max(p,2)}(\Omega)} + \|\nabla_{\mathcal{D}_m} w - \nabla\varphi(t)\|_{L^p(\Omega)^d} \Big).$$

Since $0 \in X_{\mathcal{D}_m}$, a triangle inequality shows that

$$\|\Pi_{\mathcal{D}_m} P_{\mathcal{D}_m}\varphi(t)\|_{L^{\max(p,2)}(\Omega)} + \|\nabla_{\mathcal{D}_m} P_{\mathcal{D}_m}\varphi(t)\|_{L^p(\Omega)^d} \le 2\|\varphi(t)\|_{L^{\max(p,2)}(\Omega)} + 2\|\nabla\varphi(t)\|_{L^p(\Omega)^d}. \quad (5.83)$$

In particular, by definition of $\|\cdot\|_{\mathcal{D}_m}$ and smoothness of φ, $\|P_{\mathcal{D}_m}\varphi\|_{L^p(0,T;X_{\mathcal{D}_m})}$ remains bounded. Moreover, the space–time-consistency of $((\mathcal{D}_T)_m)_{m\in\mathbb{N}}$ ensures that, for all $t\in(0,T)$, $\Pi_{\mathcal{D}_m} P_{\mathcal{D}_m}\varphi(t) \to \varphi(t)$ in $L^2(\Omega)$ as $m\to\infty$. Combined with the dominated convergence theorem and (5.83), this yields $\Pi_{\mathcal{D}_m} P_{\mathcal{D}_m}\varphi \to \varphi$ in $L^2(\Omega\times(0,T))$ as $m\to\infty$.
For $n\in\{0,\dots,N-1\}$ and $t\in(t^{(n)},t^{(n+1)}]$,

$$\begin{aligned}\Pi^{(1)}_{\mathcal{D}_m} u_m(t) - \Pi^{(\theta)}_{\mathcal{D}_m} u_m(t) &= \Pi_{\mathcal{D}_m} u_m^{(n+1)} - \Pi_{\mathcal{D}_m} u_m^{(n+\theta)} \\ &= (1-\theta)(\Pi_{\mathcal{D}_m} u_m^{(n+1)} - \Pi_{\mathcal{D}_m} u_m^{(n)})\end{aligned}$$

$$= (1-\theta)\delta t^{(n+\frac{1}{2})}\delta_{\mathcal{D}_m} u_m(t) \tag{5.84}$$

and thus, by (4.29),

$$\left|\int_0^T \int_\Omega \left(\Pi^{(1)}_{\mathcal{D}_m} u_m(\boldsymbol{x},t) - \Pi^{(\theta)}_{\mathcal{D}_m} u_m(\boldsymbol{x},t)\right) \Pi_{\mathcal{D}_m}[P_{\mathcal{D}_m}\varphi(t)](\boldsymbol{x})\mathrm{d}\boldsymbol{x}\mathrm{d}t\right|$$
$$\leq (1-\theta)\delta t_{\mathcal{D}_m} \int_0^T \|\delta_{\mathcal{D}_m} u(t)\|_{\star,\mathcal{D}_m} \|P_{\mathcal{D}_m}\varphi(t)\|_{\mathcal{D}_m} \,\mathrm{d}t.$$

Use Lemma 5.25 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

$$\begin{aligned}
&\int_0^T \int_\Omega (\widetilde{u}(\boldsymbol{x},t) - \overline{u}(\boldsymbol{x},t))\varphi(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t \\
&= \lim_{m\to\infty} \int_0^T \int_\Omega (\Pi^{(1)}_{\mathcal{D}_m} u_m(\boldsymbol{x},t) - \Pi^{(\theta)}_{\mathcal{D}_m} u_m(\boldsymbol{x},t)) \Pi_{\mathcal{D}_m}[P_{\mathcal{D}_m}\varphi(t)](\boldsymbol{x})\mathrm{d}\boldsymbol{x}\mathrm{d}t \\
&= 0.
\end{aligned} \tag{5.85}$$

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 4.9. Properties (4.7), (4.8) and (4.10) therefore hold, with $\theta = 0$.
We drop some indices m for legibility. Using $\delta t^{(n+\frac{1}{2})} v^{(n)}$ 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

$$T_4^{(m)} = \int_0^T \int_{\partial\Omega} g(\boldsymbol{x},t)\mathbb{T}^{(0)}_{\mathcal{D}} v(\boldsymbol{x},t)\mathrm{d}s(\boldsymbol{x})\mathrm{d}t.$$

Accounting for $v^{(N)} = 0$, the discrete integrate-by-parts formula (C.15) gives

$$\begin{aligned} T_1^{(m)} = & -\sum_{n=0}^{N-1} \int_\Omega \Pi_{\mathcal{D}} u^{(n+1)}(\boldsymbol{x}) \big[\Pi_{\mathcal{D}} v^{(n+1)}(\boldsymbol{x}) - \Pi_{\mathcal{D}} v^{(n)}(\boldsymbol{x})\big] \mathrm{d}\boldsymbol{x} \\ & - \int_\Omega \Pi_{\mathcal{D}} u^{(0)}(\boldsymbol{x}) \Pi_{\mathcal{D}} v^{(0)}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \\ = & -\int_0^T \int_\Omega \Pi_{\mathcal{D}}^{(1)} u(\boldsymbol{x},t) \delta_{\mathcal{D}} v(\boldsymbol{x},t) \mathrm{d}\boldsymbol{x}\mathrm{d}t - \int_\Omega \Pi_{\mathcal{D}} \mathcal{I}_{\mathcal{D}} u_{\mathrm{ini}}(\boldsymbol{x}) \Pi_{\mathcal{D}}^{(0)} v(\boldsymbol{x},0) \mathrm{d}\boldsymbol{x}. \end{aligned}$$

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 $m \to \infty$,

$$\begin{aligned} T_3^{(m)} &\to \int_0^T \int_\Omega f(\boldsymbol{x},t) \overline{v}(\boldsymbol{x},t) \mathrm{d}\boldsymbol{x}\mathrm{d}t, \quad \text{and} \\ T_4^{(m)} &\to \int_0^T \int_{\partial\Omega} g(\boldsymbol{x},t) \gamma \overline{v}(\boldsymbol{x},t) \mathrm{d}s(\boldsymbol{x})\mathrm{d}t. \end{aligned} \tag{5.89}$$

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

$$\begin{aligned} & -\int_0^T \int_\Omega \overline{u}(\boldsymbol{x},t) \partial_t \overline{v}(\boldsymbol{x},t) \mathrm{d}\boldsymbol{x}\mathrm{d}t - \int_\Omega u_{\mathrm{ini}}(\boldsymbol{x}) \overline{v}(\boldsymbol{x},0) \mathrm{d}\boldsymbol{x} \\ & + \int_0^T \int_\Omega \boldsymbol{A}(\boldsymbol{x},t) \cdot \nabla \overline{v}(\boldsymbol{x},t) \mathrm{d}\boldsymbol{x}\mathrm{d}t \\ & \quad = \int_0^T \int_\Omega f(\boldsymbol{x},t) \overline{v}(\boldsymbol{x},t) \mathrm{d}\boldsymbol{x}\mathrm{d}t + \int_0^T \int_{\partial\Omega} g(\boldsymbol{x},t) \gamma \overline{v}(\boldsymbol{x},t) \mathrm{d}s(\boldsymbol{x})\mathrm{d}t. \end{aligned}$$

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 $\overline{v} \in L^p(0,T; W^{1,p}(\Omega))$,

$$\int_0^T \langle \partial_t \overline{u}(\cdot,t), \overline{v}(\cdot,t)\rangle_{(W^{1,p}(\Omega))',W^{1,p}(\Omega)} \mathrm{d}t + \int_0^T \int_\Omega \boldsymbol{A}(\boldsymbol{x},t)\cdot\nabla\overline{v}(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t$$
$$= \int_0^T \int_\Omega f(\boldsymbol{x},t)\overline{v}(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t + \int_0^T \int_{\partial\Omega} g(\boldsymbol{x},t)\gamma\overline{v}(\boldsymbol{x},t)\mathrm{d}s(\boldsymbol{x})\mathrm{d}t. \quad (5.90)$$

It remains to prove that

$$\boldsymbol{A}(\boldsymbol{x},t) = \boldsymbol{a}(\boldsymbol{x},\overline{u}(\cdot,t),\nabla\overline{u}(\boldsymbol{x},t)), \text{ for a.e. } (\boldsymbol{x},t)\in\Omega\times(0,T). \quad (5.91)$$

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 [29, 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

$$\frac{1}{2}\int_\Omega \overline{u}(\boldsymbol{x},T)^2\mathrm{d}\boldsymbol{x} - \frac{1}{2}\int_\Omega u_{\mathrm{ini}}(\boldsymbol{x})^2\mathrm{d}\boldsymbol{x} + \int_0^T\int_\Omega \boldsymbol{A}(\boldsymbol{x},t)\cdot\nabla\overline{u}(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t$$
$$= \int_0^T\int_\Omega f(\boldsymbol{x},t)\overline{u}(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t + \int_0^T\int_{\partial\Omega} g(\boldsymbol{x},t)\gamma\overline{u}(\boldsymbol{x},t)\mathrm{d}s(\boldsymbol{x})\mathrm{d}t. \quad (5.92)$$

Relation (5.81) with $k=N$ yields

$$\begin{aligned}
&\frac{1}{2}\int_\Omega (\Pi_{\mathcal{D}}^{(1)}u(\boldsymbol{x},T))^2\mathrm{d}\boldsymbol{x}\mathrm{d}t\\
&+\int_0^T\int_\Omega \boldsymbol{a}\left(\boldsymbol{x},\Pi_{\mathcal{D}}^{(\theta)}u(\cdot,t),\nabla_{\mathcal{D}}^{(\theta)}u(\boldsymbol{x},t)\right)\cdot\nabla_{\mathcal{D}}^{(\theta)}u(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t\\
&\le \frac{1}{2}\int_\Omega (\Pi_{\mathcal{D}}\mathcal{I}_{\mathcal{D}}u_{\mathrm{ini}}(\boldsymbol{x}))^2\mathrm{d}\boldsymbol{x} + \int_0^T\int_\Omega f(\boldsymbol{x},t)\Pi_{\mathcal{D}}^{(\theta)}u(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t\\
&\quad + \int_0^T\int_{\partial\Omega} g(\boldsymbol{x},t)\mathbb{T}_{\mathcal{D}}^{(\theta)}u(\boldsymbol{x},t)\mathrm{d}s(\boldsymbol{x})\mathrm{d}t.
\end{aligned} \quad (5.93)$$

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

$$\limsup_{m\to\infty}\int_0^T\int_\Omega \boldsymbol{a}\left(\boldsymbol{x},\Pi_{\mathcal{D}}^{(\theta)}u(\cdot,t),\nabla_{\mathcal{D}}^{(\theta)}u(\boldsymbol{x},t)\right)\cdot\nabla_{\mathcal{D}}^{(\theta)}u(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t$$
$$\le -\liminf_{m\to\infty}\frac{1}{2}\int_\Omega (\Pi_{\mathcal{D}}^{(1)}u(\boldsymbol{x},T))^2\mathrm{d}\boldsymbol{x}\mathrm{d}t + \frac{1}{2}\int_\Omega u_{\mathrm{ini}}(\boldsymbol{x})^2\mathrm{d}\boldsymbol{x}$$

$$
\begin{aligned}
&+ \int_0^T \int_\Omega f(\boldsymbol{x},t)\overline{u}(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t + \int_0^T \int_{\partial\Omega} g(\boldsymbol{x},t)\gamma\overline{u}(\boldsymbol{x},t)\mathrm{d}s(\boldsymbol{x})\mathrm{d}t \\
\leq\ & -\frac{1}{2}\int_\Omega \overline{u}(\boldsymbol{x},T)^2\mathrm{d}\boldsymbol{x}\mathrm{d}t + \frac{1}{2}\int_\Omega u_{\mathrm{ini}}(\boldsymbol{x})^2\mathrm{d}\boldsymbol{x} \\
&+ \int_0^T \int_\Omega f(\boldsymbol{x},t)\overline{u}(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t + \int_0^T \int_{\partial\Omega} g(\boldsymbol{x},t)\gamma\overline{u}(\boldsymbol{x},t)\mathrm{d}s(\boldsymbol{x})\mathrm{d}t.
\end{aligned}
$$

Relation (5.92) then yields

$$
\begin{aligned}
\limsup_{m\to\infty} \int_0^T \int_\Omega \boldsymbol{a}\left(\boldsymbol{x}, \mathit{\Pi}_{\mathcal{D}_m}^{(\theta)} u_m(\cdot,t), \nabla_{\mathcal{D}_m}^{(\theta)} u_m(\boldsymbol{x},t)\right) \cdot \nabla_{\mathcal{D}_m}^{(\theta)} u_m(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t \\
\leq \int_0^T \int_\Omega \boldsymbol{A}(\boldsymbol{x},t)\cdot\nabla\overline{u}(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t. \quad (5.94)
\end{aligned}
$$

It is now possible to apply Minty's trick. Consider, for $\boldsymbol{G} \in L^p(0,T;L^p(\Omega))^d$, the quantity

$$
\begin{aligned}
\int_0^T \int_\Omega \left[\boldsymbol{a}\left(\boldsymbol{x}, \mathit{\Pi}_{\mathcal{D}}^{(\theta)} u(\cdot,t), \nabla_{\mathcal{D}}^{(\theta)} u(\boldsymbol{x},t)\right) - \boldsymbol{a}\left(\boldsymbol{x}, \mathit{\Pi}_{\mathcal{D}}^{(\theta)} u(\cdot,t), \boldsymbol{G}(\boldsymbol{x},t)\right)\right] \\
\cdot \left[\nabla_{\mathcal{D}}^{(\theta)} u(\boldsymbol{x},t) - \boldsymbol{G}(\boldsymbol{x},t)\right]\mathrm{d}\boldsymbol{x}\mathrm{d}t \geq 0. \quad (5.95)
\end{aligned}
$$

Since $\mathit{\Pi}_{\mathcal{D}_m}^{(\theta)} u_m \to \overline{u}$ strongly in $L^p(0,T;L^p(\Omega))$, up to a subsequence we can assume that $\mathit{\Pi}_{\mathcal{D}_m}^{(\theta)} u_m(t) \to \overline{u}(t)$ strongly in $L^p(\Omega)$ for a.e. $t \in (0,T)$. Assumptions (3.96a) and (3.96d) and the dominated convergence theorem then show that $\boldsymbol{a}(\cdot, \mathit{\Pi}_{\mathcal{D}}^{(\theta)} u, \boldsymbol{G}) \to \boldsymbol{a}(\cdot, \overline{u}, \boldsymbol{G})$ strongly in $L^{p'}(\Omega\times(0,T))^d$. Developping (5.95), all the terms except one pass to the limit by "weak-strong" convergence (cf. Lemma C.3). 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 3.34, 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 5.22.
Step 1: a preliminary result.

Take $(s_m)_{m\in\mathbb{N}} \subset [0,T]$ that converges to some $s \in [0,T]$. Since $\Pi^{(\theta)}_{\mathcal{D}_m} u_m \to \overline{u}$ strongly in $L^p(\Omega\times(0,T))$, as in Step 3 of the proof of Theorem 5.20, Assumptions (3.96a) and (3.96d) show that $(\mathbf{1}_{[0,s_m]}\boldsymbol{a}(\boldsymbol{x}, \Pi^{(\theta)}_{\mathcal{D}_m} u_m, \nabla\overline{u}))_{m\in\mathbb{N}}$ converges strongly in $L^{p'}(\Omega\times(0,T))^d$. The weak convergence of $(\nabla^{(\theta)}_{\mathcal{D}_m} u_m)_{m\in\mathbb{N}}$ 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

$$\begin{aligned}\frac{1}{2}\left\|\Pi^{(1)}_{\mathcal{D}_m} u_m(s)\right\|^2_{L^2(\Omega)} &+ \int_0^s\int_\Omega \boldsymbol{a}(\boldsymbol{x}, \Pi^{(\theta)}_{\mathcal{D}_m} u_m(\cdot,t), \nabla^{(\theta)}_{\mathcal{D}_m} u_m(\boldsymbol{x},t))\cdot\nabla^{(\theta)}_{\mathcal{D}_m} u_m(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t \\ &\leq \frac{1}{2}\left\|\Pi_{\mathcal{D}_m}\mathcal{I}_{\mathcal{D}_m} u_{\mathrm{ini}}\right\|^2_{L^2(\Omega)} + \int_0^s\int_\Omega f(\boldsymbol{x},t)\Pi_{\mathcal{D}_m} u_m(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t \\ &\quad + \int_0^s\int_{\partial\Omega} g(\boldsymbol{x},t)\mathbb{T}_{\mathcal{D}_m} u_m(\boldsymbol{x},t)\mathrm{d}s(\boldsymbol{x})\mathrm{d}t + \rho(\delta t_{\mathcal{D}_m}) \quad (5.98)\end{aligned}$$

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 4.28. 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

$I\!\!\Pi^{(\theta)}_{\mathcal{D}_m} u_m$ (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

$$\begin{aligned}\limsup_{m\to\infty} \frac{1}{2} \left\| I\!\!\Pi^{(1)}_{\mathcal{D}_m} u_m(s_m) \right\|^2_{L^2(\Omega)} \\ \leq -\int_0^s \int_\Omega \boldsymbol{a}(\boldsymbol{x}, \overline{u}(\cdot,t), \nabla\overline{u}(\boldsymbol{x},t)) \cdot \nabla\overline{u}(\boldsymbol{x},t) \mathrm{d}\boldsymbol{x}\mathrm{d}t \\ + \frac{1}{2}\|u_{\mathrm{ini}}\|^2_{L^2(\Omega)} + \int_0^s \int_\Omega f(\boldsymbol{x},t)\overline{u}(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t \\ + \int_0^s \int_{\partial\Omega} g(\boldsymbol{x},t)\gamma\overline{u}(\boldsymbol{x},t)\mathrm{d}s(\boldsymbol{x})\mathrm{d}t = \frac{1}{2}\|\overline{u}(s)\|^2_{L^2(\Omega)}\,. \end{aligned} \tag{5.99}$$

The uniform-in-time weak $L^2(\Omega)$ convergence of $(I\!\!\Pi^{(1)}_{\mathcal{D}_m} u_m)_{m\in\mathbb{N}}$ towards $\overline{u}$ and Lemma 4.28 show that $I\!\!\Pi^{(1)}_{\mathcal{D}_m} u_m(s_m) \to \overline{u}(s)$ in $L^2(\Omega)$ weak. Owing to (5.99), this convergence is actually strong in $L^2(\Omega)$. Invoke again Lemma 4.28 to conclude that $\sup_{t\in[0,T]} \|I\!\!\Pi^{(1)}_{\mathcal{D}_m} u_m(t) - \overline{u}(t)\|_{L^2(\Omega)} \to 0$.

The strong convergence of $(I\!\!\Pi^{(\theta)}_{\mathcal{D}_m} u_m)_{m\in\mathbb{N}}$ in the same sense follows immediately from the definition of these functions, the strong convergence of $(I\!\!\Pi^{(1)}_{\mathcal{D}_m} u_m)_{m\in\mathbb{N}}$, and the continuity of $\overline{u} : [0,T] \to L^2(\Omega)$. Indeed, $I\!\!\Pi^{(\theta)}_{\mathcal{D}_m} u_m(\cdot,t)$ is a convex combination of values of $I\!\!\Pi^{(1)}_{\mathcal{D}_m} u_m$ at two times within distance $\delta t_{\mathcal{D}_m}$ of t. ∎

Proof of Theorem 5.23.
Using (5.94) and (5.96) with $s_m = T$,

$$\begin{aligned}\limsup_{m\to\infty} \int_0^T \int_\Omega \left[\boldsymbol{a}\left(\boldsymbol{x}, I\!\!\Pi^{(\theta)}_{\mathcal{D}_m} u_m, \nabla^{(\theta)}_{\mathcal{D}_m} u_m\right) - \boldsymbol{a}\left(\boldsymbol{x}, I\!\!\Pi^{(\theta)}_{\mathcal{D}_m} u_m, \nabla\overline{u}\right) \right] \\ \cdot \left[\nabla^{(\theta)}_{\mathcal{D}_m} u_m - \nabla\overline{u}\right] \mathrm{d}\boldsymbol{x}\mathrm{d}t \leq 0.\end{aligned}$$

This relation and the strict monotonicity of $\boldsymbol{a}$ enable us to conclude, as in Step 3 of the proof of Theorem 3.34, that $\nabla^{(\theta)}_{\mathcal{D}_m} u_m \to \nabla\overline{u}$ a.e. on $\Omega\times(0,T)$. From (5.94) and (5.91) we also infer

$$\begin{aligned}\limsup_{m\to\infty} \int_0^T \int_\Omega \boldsymbol{a}\left(\boldsymbol{x}, I\!\!\Pi^{(\theta)}_{\mathcal{D}_m} u_m, \nabla^{(\theta)}_{\mathcal{D}_m} u_m\right) \cdot \nabla^{(\theta)}_{\mathcal{D}_m} u_m \mathrm{d}\boldsymbol{x}\mathrm{d}t \\ \leq \int_0^T \int_\Omega \boldsymbol{a}\left(\boldsymbol{x}, \overline{u}, \nabla\overline{u}\right) \cdot \nabla\overline{u}\mathrm{d}\boldsymbol{x}\mathrm{d}t.\end{aligned}$$

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 3.34 then yields the strong convergence of $\nabla_{\mathcal{D}_m} u_m$ to $\nabla\overline{u}$ in $L^p(\Omega\times(0,T))^d$. ∎

6

Degenerate parabolic problems

In this chapter, we study the following generic nonlinear parabolic model

$$\begin{aligned} &\partial_t \beta(\overline{u}) - \operatorname{div}\left(\Lambda(\boldsymbol{x})\nabla\zeta(\overline{u})\right) = f && \text{in } \Omega\times(0,T),\\ &\beta(\overline{u})(\boldsymbol{x},0) = \beta(u_{\mathrm{ini}})(\boldsymbol{x}) && \text{in } \Omega,\\ &\zeta(\overline{u}) = 0 && \text{on } \partial\Omega\times(0,T), \end{aligned} \tag{6.1}$$

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 [8], 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).

Remark 6.1. The techniques developed in this chapter also apply to the following more general non-linear PDE, which mixes (6.1) and Leray–Lions 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 $\nu' = \beta'\zeta'$. We refer to [33] for the analysis of gradient schemes for (6.2).

The chapter is organised as follows. In Section 6.1, we present the assumptions and the notion of weak solution for (6.1), and we show that this problem

can be 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. Numerical examples, presented in Section 6.8 and based on the VAG scheme (see Section 8.5), complete this chapter.

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 polytopal 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)
- $u_{\rm ini} \in L^2(\Omega)$, $f \in L^2(\Omega \times (0,T))$. (6.3f)

Remark 6.2 (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:

$$\begin{cases} \zeta(\overline{u}) \in L^2(0,T;H_0^1(\Omega))\,, \\ -\displaystyle\int_0^T \int_\Omega \beta(\overline{u})(\boldsymbol{x},t)\partial_t\overline{v}(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t - \int_\Omega \beta(u_{\mathrm{ini}}(\boldsymbol{x}))\overline{v}(\boldsymbol{x},0)\mathrm{d}\boldsymbol{x} \\ \qquad + \displaystyle\int_0^T \int_\Omega \Lambda(\boldsymbol{x})\nabla\zeta(\overline{u})(\boldsymbol{x},t)\cdot\nabla\overline{v}(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t \\ \qquad = \displaystyle\int_0^T \int_\Omega f(\boldsymbol{x},t)\overline{v}(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t, \\ \qquad\qquad \forall \overline{v} \in L^2(0;T;H_0^1(\Omega)) \text{ such that } \partial_t\overline{v} \in L^2((0,T)\times\Omega) \\ \qquad\qquad \text{and } \overline{v}(\cdot,T) = 0. \end{cases} \tag{6.4}$$

Remark 6.3 (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 [29]), 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 6.15). The uniqueness of this solution is proved in Section 6.7.

Theorem 6.4 (Existence and uniqueness of the weak solution). *Under Hypotheses* (6.3)*, there exists a unique solution to* (6.4)*.*

Remark 6.5. We will see in Corollary 6.17 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 [38], we show here that (6.1) can be recast in a maximal monotone operator framework.

Lemma 6.6 (Maximal monotone operator). *Let $\mathcal{T} : \mathbb{R} \to \mathbb{R}$ be a multivalued operator, that is a function from $\mathbb{R}$ to the set $\mathcal{P}(\mathbb{R})$ of all subsets of $\mathbb{R}$. The following properties are equivalent:*

1. *$\mathcal{T}$ is a maximal monotone operator with domain $\mathbb{R}$, $0 \in \mathcal{T}(0)$ 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 nondecreasing. We now have to prove that $\mathcal{T}$ is maximal, that is, if x, y satisfy $(\zeta(s)-x)(\beta(s)-y) \geq 0$ for all $s \in \mathbb{R}$ then $(x,y) \in \mathrm{Gr}(\mathcal{T})$. 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

$$\beta(s) + \zeta(s) = x + y. \tag{6.5}$$

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 (6.5), $\zeta(s) = x$. Hence $(x, y) \in \mathrm{Gr}(\mathcal{T})$. The sub-linearity of T follows from $|\beta(s)| \leq L_\beta |s| \leq L_\beta(|\zeta(s)| + M_2)/M_1$.

(1)⇒(2). Recall that the resolvent $\mathcal{R}(\mathcal{T}) = (\mathrm{Id} + \mathcal{T})^{-1}$ of the maximal monotone operator $\mathcal{T}$ is a single-valued function $\mathbb{R} \to \mathbb{R}$ that is nondecreasing and Lipschitz continuous with Lipschitz constant 1. Set $\zeta = \mathcal{R}(\mathcal{T})$ and $\beta = \mathrm{Id} - \zeta$. These functions are nondecreasing 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

$$\begin{cases} \partial_t \mathcal{T}(z) - \mathrm{div}\,(\Lambda(\boldsymbol{x})\nabla z) = f & \text{in } \Omega \times (0,T), \\ \mathcal{T}(z)(\cdot, 0) = b_{\mathrm{ini}} & \text{in } \Omega, \\ z = 0 & \text{on } \partial\Omega \times (0,T) \end{cases} \tag{6.6}$$

where $b_{\mathrm{ini}} = \beta(u_{\mathrm{ini}}) \in L^2(\Omega)$. Hypotheses (6.3c) and (6.3b) are translated into:

$$\begin{aligned} &\mathcal{T} : \mathbb{R} \to \mathcal{P}(\mathbb{R}) \text{ is a maximal monotone operator, } 0 \in \mathcal{T}(0) \\ &\text{and } \mathcal{T} \text{ is sublinear: } \exists T_1, T_2 \geq 0 \text{ such that, for all } x \in \mathbb{R} \\ &\text{all } y \in \mathcal{T}(x),\ |y| \leq T_1|x| + T_2. \end{aligned} \tag{6.7}$$

Definition 6.7. *Let us assume* (6.3a), (6.3e), (6.3f) *and* (6.7). *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* (6.6) *is a pair of functions* (z, b) *satisfying*

$$\begin{cases} z \in L^2(0,T; H_0^1(\Omega)), \\ b(\boldsymbol{x},t) \in \mathcal{T}(z(\boldsymbol{x},t)) \text{ for a.e. } (\boldsymbol{x},t) \in \Omega \times (0,T), \\ -\displaystyle\int_0^T \int_\Omega b(\boldsymbol{x},t)\partial_t \overline{v}(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t - \int_\Omega b_{\mathrm{ini}}(\boldsymbol{x})\overline{v}(\boldsymbol{x},0)\mathrm{d}\boldsymbol{x} \\ \qquad + \displaystyle\int_0^T \int_\Omega \Lambda(\boldsymbol{x})\nabla z(\boldsymbol{x},t)\cdot\nabla\overline{v}(\boldsymbol{x},t)\mathrm{d}x\mathrm{d}t \\ \qquad = \displaystyle\int_0^T \int_\Omega f(\boldsymbol{x},t)\overline{v}(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t \\ \qquad \forall \overline{v} \in L^2(0;T;H_0^1(\Omega)) \text{ such that } \partial_t\overline{v} \in L^2((0,T)\times\Omega) \\ \qquad\quad \text{and } \overline{v}(\cdot,T) = 0. \end{cases} \tag{6.8}$$

Remark 6.8. 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 6.7 and fixing $\zeta = \mathcal{R}(\mathcal{T})$ and $\beta = {\rm Id} - \zeta$ (as in the proof of Item 2 of Lemma 6.6), 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 6.4) gives the existence and uniqueness of the solution to (6.8). 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 $\mathcal{D}_T = (\mathcal{D}, \mathcal{I}_\mathcal{D}, (t^{(n)})_{n=0,\dots,N})$ 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 2.10. 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 (6.25)). 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 $(u^{(n)})_{n=0,\dots,N} \subset X_{\mathcal{D},0}$ such that

$$\left\{\begin{array}{l} u^{(0)} = \mathcal{I}_\mathcal{D} u_{\rm ini} \text{ and, for all } v = (v^{(n)})_{n=1,\dots,N} \subset X_{\mathcal{D},0}, \\ \displaystyle\int_0^T \int_\Omega \Big[\delta_\mathcal{D}\beta(u)(\boldsymbol{x},t)\Pi_\mathcal{D}^{(1)} v(\boldsymbol{x},t) + \Lambda(\boldsymbol{x})\nabla_\mathcal{D}^{(1)}\zeta(u)(\boldsymbol{x},t)\cdot\nabla_\mathcal{D}^{(1)} v(\boldsymbol{x},t)\Big]{\rm d}\boldsymbol{x}{\rm d}t \\ \qquad = \displaystyle\int_0^T \int_\Omega f(\boldsymbol{x},t)\Pi_\mathcal{D}^{(1)} v(\boldsymbol{x},t){\rm d}\boldsymbol{x}{\rm d}t. \end{array}\right. \tag{6.9}$$

We recall the definition, in Remark 2.11, of $\zeta(u)$ and $\beta(u)$, which is coherent with $\Pi_\mathcal{D}$ since this reconstruction is piecewise constant.

Remark 6.9 (Crank–Nicolson and θ-scheme)
As in Section 5.3, we could as well consider, instead of a fully implicit time-stepping, a Crank–Nicolson scheme or any scheme in between those two. Such a scheme is defined by taking $\theta \in [\frac{1}{2}, 1]$ in (4.2). All the results we establish for (6.9) would hold for such a scheme.

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 6.10 ($L^\infty(0,T;L^2(\Omega))$ 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 2.10, and that u is a solution to the corresponding GS* (6.9)*. Let $\eta:\mathbb{R}\to\mathbb{R}$ be defined by*

$$\forall s\in\mathbb{R},\quad \eta(s)=\int_0^s \zeta(q)\beta'(q)\mathrm{d}q. \tag{6.10}$$

Let $T_0\in(0,T]$ and denote by $k=1,\dots,N$ the index such that $T_0\in(t^{(k-1)},t^{(k)}]$. Then

$$\begin{aligned}\int_\Omega \Pi_{\mathcal{D}}^{(1)}\eta(u)(\boldsymbol{x},T_0)\mathrm{d}\boldsymbol{x}+\int_0^{T_0}\int_\Omega \Lambda(\boldsymbol{x})\nabla_{\mathcal{D}}^{(1)}\zeta(u)(\boldsymbol{x},t)\cdot\nabla_{\mathcal{D}}^{(1)}\zeta(u)(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t\\ \le\int_\Omega \Pi_{\mathcal{D}}\eta(\mathcal{I}_{\mathcal{D}}u_{\rm ini})(\boldsymbol{x})\mathrm{d}\boldsymbol{x}+\int_0^{t^{(k)}}\int_\Omega f(\boldsymbol{x},t)\Pi_{\mathcal{D}}^{(1)}\zeta(u)(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t.\end{aligned} \tag{6.11}$$

Consequently, there exists $C_1>0$, depending only on L_β, L_ζ, $C_P\ge C_{\mathcal{D}}$ (see Definition 2.2), $C_{\rm ini}\ge\|\Pi_{\mathcal{D}}\mathcal{I}_{\mathcal{D}}u_{\rm ini}\|_{L^2(\Omega)}$, f and $\underline{\lambda}$ such that

$$\begin{aligned}&\sup_{t\in[0,T]}\left\|\Pi_{\mathcal{D}}^{(1)}\eta(u)(t)\right\|_{L^1(\Omega)}\le C_1\,,\ \left\|\nabla_{\mathcal{D}}^{(1)}\zeta(u)\right\|_{L^2(\Omega\times(0,T))^d}\le C_1\\ &\text{and}\ \sup_{t\in[0,T]}\left\|\Pi_{\mathcal{D}}^{(1)}\beta(u)(t)\right\|_{L^2(\Omega)}\le C_1.\end{aligned} \tag{6.12}$$

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 $\int_a^b\zeta'(q)(\beta(q)-\beta(a))\mathrm{d}q\ge0$ (as ζ and β are non-decreasing), we get

$$\eta(b)-\eta(a)\le\zeta(b)(\beta(b)-\beta(a)). \tag{6.13}$$

Using Remark 2.11 (consequence of the definition 2.10 of piecewise constant reconstruction) and (6.13), we infer that for any $n=0,\dots,N-1$, any $t\in(t^{(n)},t^{(n+1)}]$,

$$\begin{aligned}\delta_{\mathcal{D}}\beta(u)(t)\Pi_{\mathcal{D}}\zeta(u^{(n+1)})&=\frac{1}{\delta t^{(n+\frac12)}}\left(\beta(\Pi_{\mathcal{D}}u^{(n+1)})-\beta(\Pi_{\mathcal{D}}u^{(n)})\right)\zeta(\Pi_{\mathcal{D}}u^{(n+1)})\\ &\ge\frac{1}{\delta t^{(n+\frac12)}}\left(\eta(\Pi_{\mathcal{D}}u^{(n+1)})-\eta(\Pi_{\mathcal{D}}u^{(n)})\right).\end{aligned}$$

Hence, taking $v=(\zeta(u^{(0)}),\zeta(u^{(1)}),\dots,\zeta(u^{(k)}),0,\dots,0)\subset X_{\mathcal{D},0}$ in (6.9), we find

$$\int_\Omega \eta(\Pi_{\mathcal{D}}^{(1)} u(\boldsymbol{x}, t^{(k)}))\mathrm{d}\boldsymbol{x} + \int_0^{t^{(k)}} \int_\Omega \Lambda(\boldsymbol{x}) \nabla_{\mathcal{D}}^{(1)} \zeta(u)(\boldsymbol{x}, t) \cdot \nabla_{\mathcal{D}}^{(1)} \zeta(u)(\boldsymbol{x}, t)\mathrm{d}\boldsymbol{x}\mathrm{d}t$$
$$\leq \int_\Omega \eta(\Pi_{\mathcal{D}} u^{(0)}(\boldsymbol{x}))\mathrm{d}\boldsymbol{x} + \int_0^{t^{(k)}} \int_\Omega f(\boldsymbol{x}, t)\Pi_{\mathcal{D}}^{(1)} \zeta(u)(\boldsymbol{x}, t)\mathrm{d}\boldsymbol{x}\mathrm{d}t. \quad (6.14)$$

Equation (6.11) 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 nonnegative on $[T_0, t^{(k)}]$.
Using the Young inequality (C.9), we write

$$\int_0^{t^{(k)}} \int_\Omega f(\boldsymbol{x}, t)\Pi_{\mathcal{D}}^{(1)} \zeta(u)(\boldsymbol{x}, t)\mathrm{d}\boldsymbol{x}\mathrm{d}t$$
$$\leq \frac{C_{\mathcal{D}}^2}{2\underline{\lambda}} \|f\|_{L^2(\Omega\times(0,t^{(k)}))}^2 + \frac{\underline{\lambda}}{2C_{\mathcal{D}}^2} \|\Pi_{\mathcal{D}}^{(1)} \zeta(u)\|_{L^2(\Omega\times(0,t^{(k)}))}^2. \quad (6.15)$$

We also notice that

$$0 \leq \eta(s) \leq L_\beta L_\zeta \int_0^s q\mathrm{d}q = L_\beta L_\zeta \frac{s^2}{2}, \quad (6.16)$$

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

$$\left\| \eta(\Pi_{\mathcal{D}} u^{(0)}) \right\|_{L^1(\Omega)} = \|\eta(\Pi_{\mathcal{D}} \mathcal{I}_{\mathcal{D}} u_{\mathrm{ini}})\|_{L^1(\Omega)} \leq \frac{L_\beta L_\zeta}{2} \|\Pi_{\mathcal{D}} \mathcal{I}_{\mathcal{D}} u_{\mathrm{ini}}\|_{L^2(\Omega)}^2 .$$

The first two estimates in (6.12) therefore follow from (6.14), (6.15), 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,

$$\eta(s) = \int_0^s \zeta(q)\beta'(q)\mathrm{d}q \geq \frac{M_0}{L_\beta} \int_0^s \beta(q)\beta'(q)\mathrm{d}q - M_1 \int_0^s \beta'(q)\mathrm{d}q$$
$$= \frac{M_0}{2L_\beta}\beta(s)^2 - M_1\beta(s)$$
$$\geq \frac{M_0}{2L_\beta}\beta(s)^2 - \frac{M_0}{4L_\beta}\beta(s)^2 - \frac{L_\beta M_1^2}{M_0}.$$

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,

$$\forall s \in \mathbb{R}, \quad \frac{M_0}{4L_\beta}\beta(s)^2 - \frac{L_\beta M_1^2}{M_0} \le \eta(s). \tag{6.17}$$

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 (6.12). ∎

Corollary 6.11 (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 2.10. Then there exists at least a solution to the GS* (6.9)*.*

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 6.10 to see that there exists C_2, not depending on ρ, such that any solution u_ρ to (6.9) with $\beta = \beta_\rho$ and $\zeta = \zeta_\rho$ satisfies

$$\left\| \nabla_{\mathcal{D}}^{(1)} \zeta_\rho(u_\rho) \right\|_{L^2((0,T)\times\Omega)^d} \le C_2.$$

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 (6.9) 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 C.1) combined with the uniform bound on $(u_\rho)_{\rho\in[0,1]}$ then shows that the scheme corresponding to $\rho = 1$, that is (6.9), possesses at least one solution. ∎

Lemma 6.12 (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 2.10. Let $u, \widetilde{u}$ be solutions to the GS* (6.9)*. Then, for all $n = 0, \dots, N$, $\Pi_{\mathcal{D}} u^{(n)} = \Pi_{\mathcal{D}} \widetilde{u}^{(n)}$ 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 $u^{(0)} = \widetilde{u}^{(0)} = \mathcal{I}_{\mathcal{D}} u_{\rm ini}$. Let us now assume that, for some $n \le N-1$, $\Pi_{\mathcal{D}} u^{(n)}(\boldsymbol{x}) = \Pi_{\mathcal{D}} \widetilde{u}^{(n)}(\boldsymbol{x})$ 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

$$\int_\Omega \Big[\frac{\Pi_\mathcal{D}(\beta(u^{(n+1)}) - \beta(\widetilde{u}^{(n+1)}))(\boldsymbol{x})}{\delta t^{(n+\frac{1}{2})}} \Pi_\mathcal{D} v(\boldsymbol{x}) + \nabla_\mathcal{D}(\zeta(u^{(n+1)}) - \zeta(\widetilde{u}^{(n+1)}))(\boldsymbol{x}) \cdot \nabla_\mathcal{D} v(\boldsymbol{x}) \Big] \mathrm{d}\boldsymbol{x} = 0, \quad \forall v \in X_{\mathcal{D},0}. \quad (6.18)$$

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 (6.18),

$$\int_\Omega |\nabla_\mathcal{D}(\zeta(u^{(n+1)}) - \zeta(\widetilde{u}^{(n+1)}))(\boldsymbol{x})|^2 \mathrm{d}\boldsymbol{x} = 0.$$

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 (6.18), 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 6.13 (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 2.10. Let u be a solution to Scheme* (6.9)*. 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*

$$\int_0^T \|\delta_\mathcal{D}\beta(u)(t)\|_{\star,\mathcal{D}}^2 \,\mathrm{d}t \leq C_3, \quad (6.19)$$

where the dual norm $\|\cdot\|_{\star,\mathcal{D}}$ is given by Definition 4.18.

Proof. Let us take a generic $v = (v^{(n)})_{n=1,\dots,N} \subset X_{\mathcal{D},0}$ as test function in (6.9). 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 (6.12), this gives $C_4 > 0$ depending only on L_β, C_P, $C_{\rm ini}$, f, $\underline{\lambda}$ and $\overline{\lambda}$ such that

$$\int_0^T \int_\Omega \delta_{\mathcal{D}}\beta(u)(\boldsymbol{x},t)\Pi_{\mathcal{D}}^{(1)}v(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t \le C_4\|\nabla_{\mathcal{D}}^{(1)}v\|_{L^2(0,T;L^2(\Omega))^d}.$$

The proof of (6.19) is completed by selecting

$$v = \left(\left\|\delta_{\mathcal{D}}^{(n+\frac{1}{2})}\beta(u)\right\|_{\star,\mathcal{D}} z^{(n)}\right)_{n=0,\dots,N}$$

with $(z^{(n)})_{n=0,\dots,N} \subset X_{\mathcal{D},0}$ such that, for any $n = 0,\dots,N-1$, $z^{(n+1)}$ realises the supremum in (4.28) 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 6.14 (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 2.10). Let* $\nu : \mathbb{R}\to\mathbb{R}$ *be defined by*

$$\forall s\in\mathbb{R}, \quad \nu(s) = \int_0^s \zeta'(q)\beta'(q)\mathrm{d}q. \tag{6.20}$$

For any $m\in\mathbb{N}$*, let* u_m *be a solution to* (6.9) *with* $\mathcal{D} = \mathcal{D}_m$*. Then, as* $m\to\infty$*,*

$$\begin{aligned}
&\Pi_{\mathcal{D}_m}^{(1)}\beta(u_m)\to\beta(\overline{u}) && \text{weakly in } L^2(\Omega) \text{ uniformly on } [0,T]\\
&&& \text{(see Definition 4.29),}\\
&\Pi_{\mathcal{D}_m}^{(1)}\zeta(u_m)\to\zeta(\overline{u}) && \text{weakly in } L^2(\Omega\times(0,T)),\\
&\Pi_{\mathcal{D}_m}^{(1)}\nu(u_m)\to\nu(\overline{u}) && \text{in } L^2(\Omega\times(0,T)),\\
&\nabla_{\mathcal{D}_m}^{(1)}\zeta(u_m)\to\nabla\zeta(\overline{u}) && \text{weakly in } L^2(\Omega\times(0,T))^d,
\end{aligned} \tag{6.21}$$

where $\overline{u}$ *is the unique solution to* (6.4)*.*

Remark 6.15. 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 6.4.

Proof.
Note that, since $((\mathcal{D}_T)_m)_{m\in\mathbb{N}}$ is compact, it is also coercive.
Step 1: Application of compactness results.
Thanks to Theorem 4.32 and Estimates (6.12) and (6.19), 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 4.29) 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 (6.12) and applying Lemma 4.7, 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 (6.12) and (6.19) also show that $\beta_m = \beta(u_m)$ and $\zeta_m = \zeta(u_m)$ satisfy the assumptions of Theorem 4.24 (weak-strong time-space convergence of a product theorem). Hence,

$$\lim_{m\to\infty}\int_0^T\int_\Omega \beta\left(\Pi^{(1)}_{\mathcal{D}_m}u_m(\boldsymbol{x},t)\right)\zeta\left(\Pi^{(1)}_{\mathcal{D}_m}u_m(\boldsymbol{x},t)\right)\mathrm{d}\boldsymbol{x}\mathrm{d}t = \int_0^T\int_\Omega \overline{\beta}(\boldsymbol{x},t)\overline{\zeta}(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t. \quad (6.22)$$

Assumptions (6.3b)–(6.3d) allow us to apply Lemma C.5 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

$$\int_0^T\int_\Omega\left[\nu(\Pi^{(1)}_{\mathcal{D}_m}u_m(\boldsymbol{x},t)) - \nu(\overline{u}(\boldsymbol{x},t))\right]^2\mathrm{d}\boldsymbol{x}\mathrm{d}t$$

$$\leq L_\zeta L_\beta \int_0^T \int_\Omega \left[\beta(\Pi^{(1)}_{\mathcal{D}_m} u_m(\boldsymbol{x},t)) - \beta(\overline{u}(\boldsymbol{x},t))\right] \times \left[\zeta(\Pi^{(1)}_{\mathcal{D}_m} u_m(\boldsymbol{x},t)) - \zeta(\overline{u}(\boldsymbol{x},t))\right] \mathrm{d}\boldsymbol{x}\mathrm{d}t. \quad (6.23)$$

Developing the right-hand side of this inequality, using (6.22) 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 (6.23) 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 4.9 (for $\theta = 0$) and introduce $(0, v^{(0)}, \dots, v^{(N-1)})$ as test function in (6.9). 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 (C.15) in $T_1^{(m)}$:

$$T_1^{(m)} = -\int_0^T \int_\Omega \Pi^{(1)}_{\mathcal{D}} \beta(u)(\boldsymbol{x},t) \delta_{\mathcal{D}} v(\boldsymbol{x},t) \mathrm{d}\boldsymbol{x}\mathrm{d}t - \int_\Omega \beta(\Pi_{\mathcal{D}} \mathcal{I}_{\mathcal{D}} u_{\mathrm{ini}})(\boldsymbol{x}) \Pi_{\mathcal{D}} v^{(0)} \mathrm{d}\boldsymbol{x}.$$

Hence, by the convergence properties (4.10c) and (4.10b) of $(v_m)_{m\in\mathbb{N}}$, as $m \to \infty$ we have

$$T_1^{(m)} \to -\int_0^T \int_\Omega \beta(\overline{u})(\boldsymbol{x},t) \partial_t \overline{v}(\boldsymbol{x},t) \mathrm{d}\boldsymbol{x}\mathrm{d}t - \int_\Omega \beta(u_{\mathrm{ini}})(\boldsymbol{x}) \overline{v}(\boldsymbol{x},0) \mathrm{d}\boldsymbol{x}. \quad (6.24)$$

Using the convergence (4.7b) and (4.7a) of $(v_m)_{m\in\mathbb{N}}$, we have, as $m \to \infty$,

$$T_2^{(m)} \to \int_0^T \int_\Omega \Lambda(\boldsymbol{x}) \nabla \zeta(\overline{u})(\boldsymbol{x},t) \cdot \nabla \overline{v}(\boldsymbol{x},t) \mathrm{d}\boldsymbol{x}\mathrm{d}t,$$

$$T_3^{(m)} \to \int_0^T \int_\Omega f(\boldsymbol{x},t) \overline{v}(\boldsymbol{x},t) \mathrm{d}\boldsymbol{x}\mathrm{d}t.$$

Plugged alongside (6.24) in $T_1^{(m)} + T_2^{(m)} = T_3^{(m)}$, these convergences show that $\overline{u}$ satisfies (6.4). ∎

Remark 6.16 (Convergence of $\Pi_{\mathcal{D}}^{(1)} u_m$?)
We do not prove here that $\overline{u}$ is a weak limit of $\Pi_{\mathcal{D}_m}^{(1)} u_m$. Such a limit is not stated in (6.21) 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 6.4), an equivalent form of (6.4) can be stated.

Corollary 6.17 (Equivalent from of (6.4)). *Under Hypotheses* (6.3)*, Problem* (6.4) *is equivalent to*

$$\begin{cases} \overline{u} \in L^2(0,T;L^2(\Omega))\,,\ \zeta(\overline{u}) \in L^2(0,T;H_0^1(\Omega))\,, \\ \beta(\overline{u}) \in C([0,T],L^2(\Omega)\text{-w})\,,\ \partial_t \beta(\overline{u}) \in L^2(0,T;H^{-1}(\Omega))\,, \\ \beta(\overline{u})(\cdot,0) = \beta(u_{\text{ini}}) \text{ in } L^2(\Omega), \\ \displaystyle\int_0^T \langle \partial_t \beta(\overline{u})(\cdot,t), \overline{v}(\cdot,t)\rangle_{H^{-1},H_0^1} \mathrm{d}t \\ \qquad \displaystyle + \int_0^T \int_\Omega \Lambda(\boldsymbol{x}) \nabla \zeta(\overline{u})(\boldsymbol{x},t) \cdot \nabla \overline{v}(\boldsymbol{x},t) \mathrm{d}\boldsymbol{x}\mathrm{d}t \\ \qquad \displaystyle = \int_0^T \int_\Omega f(\boldsymbol{x},t)\overline{v}(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t, \quad \forall \overline{v} \in L^2(0;T;H_0^1(\Omega)), \end{cases} \tag{6.25}$$

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 (6.25). There is a unique solution to (6.4) (Theorem 6.4), so it must be the $\overline{u}$ constructed in the proof of Theorem 6.14. 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 $\beta(\overline{u})(0,\cdot) = \overline{\beta}(0,\cdot) = \beta(u_{\text{ini}})$. Using $C_c^\infty((0,T)\times\Omega)$ test functions in (6.4), we see that

$$\partial_t \beta(\overline{u}) = \operatorname{div}(\Lambda \nabla \zeta(\overline{u})) + f$$

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 $\overline{v} \in C_c^\infty((0,T)\times\Omega)$. By definition of the distribution derivative,

$$-\int_0^T \int_\Omega \beta(\overline{u})(\boldsymbol{x},t)\partial_t \overline{v}(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t = \int_0^T \langle \partial_t \beta(\overline{u})(t), \overline{v}(t)\rangle_{H^{-1},H_0^1}\mathrm{d}t$$

and thus (6.4) shows that the equation in (6.25) is satisfied for such smooth compactly supported $\overline{v}$. Since these functions are dense in $L^2(0,T;H_0^1(\Omega))$ (see [29]), we infer that (6.25) is fully satisfied.

Let us now assume that $\overline{u}$ satisfies (6.25). 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 $\overline{v} \in C_c^\infty((-\infty,T)\times\Omega)$. 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 $\beta(\overline{u})(\boldsymbol{x}, 0) = \beta(u_{\mathrm{ini}})$, (6.25) proves that (6.4) is satisfied for such $\overline{v}$. As discussed at the end of the proof of Theorem 6.14, this shows that (6.25) is satisfied for all required test functions. ∎

Remark 6.18 (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 in $L^2(\Omega)$ of such φ, 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

$$\forall s \in R_\beta\,,\ \beta^i(s) = \begin{cases} \inf\{t \in \mathbb{R} \,|\, \beta(t) = s\} & \text{if } s \geq 0, \\ \sup\{t \in \mathbb{R} \,|\, \beta(t) = s\} & \text{if } s < 0, \end{cases} \\ = \text{closest } t \text{ to } 0 \text{ such that } \beta(t) = s. \tag{6.26}$$

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

$$B(z) = \int_0^z \zeta(\beta^i(s))\, ds. \tag{6.27}$$

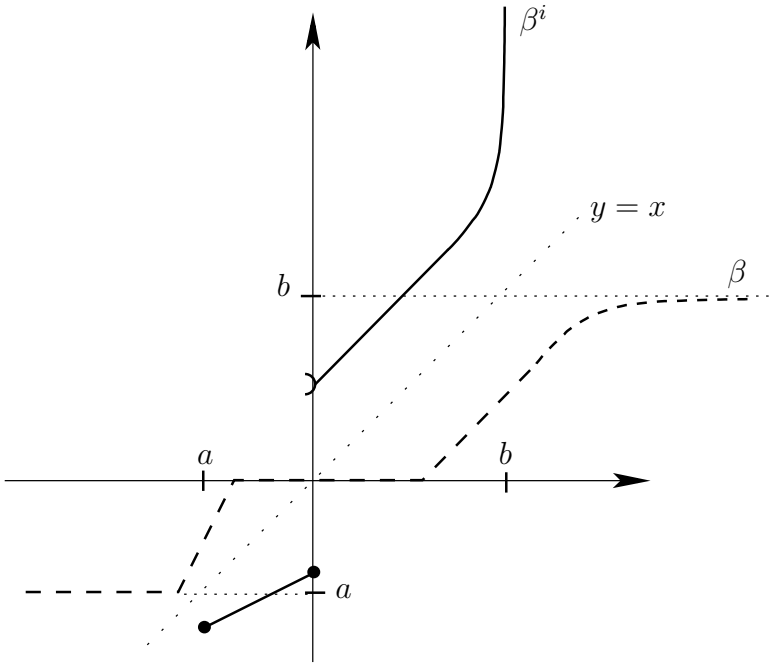


Fig. 6.1. An example of β (dashed line) and its pseudo-inverse function β^i (continuous line). Here, the range of β is $[a, b)$.

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 6.23 in Section 6.6 states a few useful properties of B.

Remark 6.19 (Range of $\beta(\overline{u})$)
The a.e. equality $\beta(\overline{u})(\boldsymbol{x}, t) = \beta(\overline{u}(\boldsymbol{x}, t))$ (see Remark 6.18) 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 that established in Theorem 6.14.

Theorem 6.20 (Uniform-in-time convergence of the GS). *Under the assumptions of Theorem 6.14, the solution u_m to the GS* (6.9) *with $\mathcal{D}_T = (\mathcal{D}_T)_m$ satisfies the following convergence results, as $m \to \infty$:*

$$\begin{aligned}
&\sup_{t\in[0,T]} \left\| \Pi^{(1)}_{\mathcal{D}_m} \nu(u_m)(t) - \nu(\overline{u})(t) \right\|_{L^2(\Omega)} \to 0,\\
&\Pi^{(1)}_{\mathcal{D}_m} \zeta(u_m) \to \zeta(\overline{u}) \ \textit{in}\ L^2(\Omega\times(0,T)),\\
&\nabla^{(1)}_{\mathcal{D}_m} \zeta(u_m) \to \nabla\zeta(\overline{u}) \ \textit{in}\ L^2(\Omega\times(0,T))^d,
\end{aligned} \tag{6.28}$$

where $\overline{u}$ is the unique solution to (6.4).

Remark 6.21. For the Stefan model, $\beta = \mathrm{Id}$ 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 (6.38) in Lemma 6.23, $\eta = B\circ\beta$. The energy estimate (6.11) can thus be written

$$\begin{aligned}
&\int_\Omega B(\beta(\Pi^{(1)}_{\mathcal{D}_m} u_m))(\boldsymbol{x}, T_0)\mathrm{d}\boldsymbol{x}\\
&\qquad + \int_0^{T_0}\int_\Omega \Lambda(\boldsymbol{x})\nabla^{(1)}_{\mathcal{D}_m}\zeta(u_m)(\boldsymbol{x},t)\cdot\nabla^{(1)}_{\mathcal{D}_m}\zeta(u_m)(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t\\
&\qquad\qquad \le \int_\Omega B(\beta(\Pi_{\mathcal{D}_m}\mathcal{I}_{\mathcal{D}_m} u_{\mathrm{ini}}))(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\\
&\qquad\qquad\qquad + \int_0^{t^{(k)}}\int_\Omega f(\boldsymbol{x},t)\Pi^{(1)}_{\mathcal{D}_m}\zeta(u_m)(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t.
\end{aligned} \tag{6.29}$$

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 C.3 page 403) 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

$$\liminf_{m\to\infty}\int_0^{T_m}\int_\Omega \Lambda(\boldsymbol{x})\nabla^{(1)}_{\mathcal{D}_m}\zeta(u_m)(\boldsymbol{x},t)\cdot\nabla^{(1)}_{\mathcal{D}_m}\zeta(u_m)(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t \geq \int_0^{T_0}\int_\Omega \Lambda(\boldsymbol{x})\nabla\zeta(\overline{u})(\boldsymbol{x},t)\cdot\nabla\zeta(\overline{u})(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t. \quad (6.30)$$

By space–time-consistency of $((\mathcal{D}_T)_m)_{m\in\mathbb{N}}$ (Definition 4.3) and quadratic growth (6.16) of $\eta = B\circ\beta$, it holds

$$B(\beta(\Pi_{\mathcal{D}_m}\mathcal{I}_{\mathcal{D}_m}u_{\mathrm{ini}})) \to B(\beta(u_{\mathrm{ini}}) \text{ in } L^1(\Omega) \text{ as } m\to\infty. \quad (6.31)$$

We then write (6.29) 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 (6.30) and the weak convergence of $\Pi^{(1)}_{\mathcal{D}_m}\zeta(u_m)$ to $\zeta(\overline{u})$,

$$\begin{aligned}
&\limsup_{m\to\infty}\int_\Omega B(\beta(\Pi^{(1)}_{\mathcal{D}_m}u_m(\boldsymbol{x},T_m)))\mathrm{d}\boldsymbol{x}\\
&\quad\leq \limsup_{m\to\infty}\left(\int_\Omega B(\beta(\Pi_{\mathcal{D}_m}\mathcal{I}_{\mathcal{D}_m}u_{\mathrm{ini}}))(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\right.\\
&\qquad\quad+\int_0^{t^{(k(m))}}\int_\Omega f(\boldsymbol{x},t)\Pi^{(1)}_{\mathcal{D}_m}\zeta(u_m)(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t\\
&\qquad\quad\left.-\int_0^{T_m}\int_\Omega \Lambda(\boldsymbol{x})\nabla^{(1)}_{\mathcal{D}_m}\zeta(u_m)(\boldsymbol{x},t)\cdot\nabla^{(1)}_{\mathcal{D}_m}\zeta(u_m)(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t\right)\\
&\quad\leq \int_\Omega B(\beta(u_{\mathrm{ini}}))(\boldsymbol{x})\mathrm{d}\boldsymbol{x}+\int_0^{T_0}\int_\Omega f(\boldsymbol{x},t)\zeta(\overline{u})(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t\\
&\qquad-\liminf_{m\to\infty}\int_0^{T_m}\int_\Omega \Lambda(\boldsymbol{x})\nabla_{\mathcal{D}_m}\zeta(u_m)(\boldsymbol{x},t)\cdot\nabla_{\mathcal{D}_m}\zeta(u_m)(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t\\
&\quad\leq \int_\Omega B(\beta(u_{\mathrm{ini}}))(\boldsymbol{x})\mathrm{d}\boldsymbol{x}+\int_0^{T_0}\int_\Omega f(\boldsymbol{x},t)\zeta(\overline{u})(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t\\
&\qquad-\int_0^{T_0}\int_\Omega \Lambda(\boldsymbol{x})\nabla\zeta(\overline{u})(\boldsymbol{x},t)\cdot\nabla\zeta(\overline{u})(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t.
\end{aligned}$$

Corollary 6.26 therefore gives

$$\limsup_{m\to\infty}\int_\Omega B(\beta(\Pi^{(1)}_{\mathcal{D}_m}u_m(\boldsymbol{x},T_m)))\mathrm{d}\boldsymbol{x}\leq\int_\Omega B(\beta(\overline{u})(\boldsymbol{x},T_0))\mathrm{d}\boldsymbol{x}. \quad (6.32)$$

By Lemma 4.28, 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 6.17), we have

$$\beta(\Pi^{(1)}_{\mathcal{D}_m} u_m)(T_m) \to \beta(\overline{u})(T_0) \text{ weakly in } L^2(\Omega) \text{ as } m \to \infty. \tag{6.33}$$

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 C.6 then gives, by convexity of B,

$$\int_\Omega B(\beta(\overline{u})(\boldsymbol{x}, T_0))\mathrm{d}\boldsymbol{x} \leq \liminf_{m\to\infty} \int_\Omega B\left(\frac{\beta(\Pi^{(1)}_{\mathcal{D}_m} u_m(\boldsymbol{x}, T_m)) + \beta(\overline{u})(\boldsymbol{x}, s_m)}{2}\right)\mathrm{d}\boldsymbol{x}. \tag{6.34}$$

Property (6.40) of B and the two inequalities (6.32) and (6.34) 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 6.18 and 6.27), 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 6.26, we have

$$\nu(\overline{u}(\cdot, s_m)) \to \nu(\overline{u})(\cdot, T_0) \text{ in } L^2(\Omega) \text{ as } m \to \infty. \tag{6.35}$$

Inequality (6.40) gives

$$\begin{aligned}
\left\|\nu(\Pi^{(1)}_{\mathcal{D}_m} u_m(\cdot, T_m)) - \nu(\overline{u})(\cdot, T_0)\right\|^2_{L^2(\Omega)} &\leq 2\left\|\nu(\Pi^{(1)}_{\mathcal{D}_m} u_m(\cdot, T_m)) - \nu(\overline{u}(\cdot, s_m))\right\|^2_{L^2(\Omega)} \\
&\quad + 2\left\|\nu(\overline{u}(\cdot, s_m)) - \nu(\overline{u})(\cdot, T_0))\right\|^2_{L^2(\Omega)} \\
&\leq 8L_\beta L_\zeta \int_\Omega \left[B(\beta(\Pi^{(1)}_{\mathcal{D}_m} u_m(\boldsymbol{x}, T_m))) + B(\beta(\overline{u}(\boldsymbol{x}, s_m)))\right]\mathrm{d}\boldsymbol{x} \\
&\quad - 16L_\beta L_\zeta \int_\Omega B\left(\frac{\beta(\Pi^{(1)}_{\mathcal{D}_m} u_m(\boldsymbol{x}, T_m)) + \beta(\overline{u}(\boldsymbol{x}, s_m))}{2}\right)\mathrm{d}\boldsymbol{x} \\
&\quad + 2\left\|\nu(\overline{u}(\cdot, s_m)) - \nu(\overline{u})(\cdot, T_0))\right\|^2_{L^2(\Omega)}.
\end{aligned}$$

We then take the lim sup as $m \to \infty$ of this expression. Thanks to (6.32) 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 6.26), 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 (6.32), (6.34), (6.35) and to the continuity of B (Corollary 6.26),

$$\limsup_{m\to\infty} \left\| \nu(\Pi^{(1)}_{\mathcal{D}_m} u_m(\cdot, T_m)) - \nu(\overline{u})(\cdot, T_0) \right\|^2_{L^2(\Omega)} \leq 0.$$

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 4.28, 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 (6.33) (with $T_m = T_0 = T$) and Lemma C.6 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 (6.29) with $T_0 = T$, taking the lim sup as $m \to \infty$, using (6.31) and the continuous integration-by-part formula (6.49), we therefore find

$$\begin{aligned}\limsup_{m\to\infty} \int_0^T \int_\Omega \Lambda(\boldsymbol{x}) \nabla^{(1)}_{\mathcal{D}_m} \zeta(u_m)(\boldsymbol{x}, t) \cdot \nabla^{(1)}_{\mathcal{D}_m} \zeta(u_m)(\boldsymbol{x}, t)\mathrm{d}\boldsymbol{x}\mathrm{d}t \\ \leq \int_0^{T_0} \int_\Omega \Lambda(\boldsymbol{x}) \nabla\zeta(\overline{u})(\boldsymbol{x}, t) \cdot \nabla\zeta(\overline{u})(\boldsymbol{x}, t)\mathrm{d}\boldsymbol{x}\mathrm{d}t.\end{aligned}$$

Combined with (6.30) with $T_m = T_0 = T$, this shows that

$$\begin{aligned}\lim_{m\to\infty} \int_0^T \int_\Omega \Lambda(\boldsymbol{x}) \nabla^{(1)}_{\mathcal{D}_m} \zeta(u_m)(\boldsymbol{x}, t) \cdot \nabla^{(1)}_{\mathcal{D}_m} \zeta(u_m)(\boldsymbol{x}, t)\mathrm{d}\boldsymbol{x}\mathrm{d}t \\ = \int_0^{T_0} \int_\Omega \Lambda(\boldsymbol{x}) \nabla\zeta(\overline{u})(\boldsymbol{x}, t) \cdot \nabla\zeta(\overline{u})(\boldsymbol{x}, t)\mathrm{d}\boldsymbol{x}\mathrm{d}t. \quad (6.36)\end{aligned}$$

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

$$\begin{aligned}\lim_{m\to\infty} \int_0^T \int_\Omega \Lambda(\boldsymbol{x}) \left[\nabla^{(1)}_{\mathcal{D}_m} \zeta(u_m)(\boldsymbol{x}, t) - \nabla\zeta(\overline{u})(\boldsymbol{x}, t)\right] \\ \cdot \left[\nabla^{(1)}_{\mathcal{D}_m} \zeta(u_m)(\boldsymbol{x}, t) - \nabla\zeta(\overline{u})(\boldsymbol{x}, t)\right] \mathrm{d}\boldsymbol{x}\mathrm{d}t = 0.\end{aligned}$$

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 4.9 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

$$\left\|\Pi^{(1)}_{\mathcal{D}_m}\zeta(u_m)-\Pi^{(1)}_{\mathcal{D}_m}v_m\right\|_{L^2(\Omega\times(0,T))} \le C_P\left\|\nabla^{(1)}_{\mathcal{D}_m}\zeta(u_m)-\nabla^{(1)}_{\mathcal{D}_m}v_m\right\|_{L^2(\Omega\times(0,T))^d}.$$

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 6.22 (Convergence of $B(\beta(\Pi^{(1)}_{\mathcal{D}_m}u_m(T_m)))$)
Let $T_m\to T_0$. The convergence property (6.33), the convexity of B and Lemma C.6 show that

$$\int_\Omega B(\beta(\overline{u})(\boldsymbol{x},T_0))\mathrm{d}\boldsymbol{x}\le\liminf_{m\to\infty}\int_\Omega B(\beta(\Pi^{(1)}_{\mathcal{D}_m}u_m)(\boldsymbol{x},T_m))\mathrm{d}\boldsymbol{x}.$$

Combined with (6.32), this gives

$$\lim_{m\to\infty}\int_\Omega B(\beta(\Pi^{(1)}_{\mathcal{D}_m}u_m(\boldsymbol{x},T_m)))\mathrm{d}\boldsymbol{x}=\int_\Omega B(\beta(\overline{u})(\boldsymbol{x},T_0))\mathrm{d}\boldsymbol{x}. \tag{6.37}$$

6.6 Auxiliary results

We state here a family of technical lemmas, starting with a few properties on ν and B.

Lemma 6.23. *Under Assumptions* (6.3b) *and* (6.3c)*, let ν be defined by* (6.20)*, B be defined by* (6.27)*, and η be defined by* (6.10)*. 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,*

$$\forall s\in\mathbb{R},\quad \eta(s)=B(\beta(s))=\int_0^s\zeta(q)\beta'(q)\mathrm{d}q, \tag{6.38}$$

$$\forall a\in\mathbb{R},\forall r\in\overline{R_\beta},\quad B(r)-B(\beta(a))\ge\zeta(a)(r-\beta(a)), \tag{6.39}$$

and

$$\forall s,s'\in\mathbb{R},(\nu(s)-\nu(s'))^2\le 4L_\beta L_\zeta\left[B(\beta(s))+B(\beta(s'))-2B\left(\frac{\beta(s)+\beta(s')}{2}\right)\right]. \tag{6.40}$$

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 (6.38), 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. 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 (6.38) is complete. The continuity of $B \circ \beta$ follows from this relation.
We now prove (6.39), which states that $\zeta(a)$ belongs to the convex sub-differential 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 (6.39) 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 (6.38),

$$B(\beta(s)) + B(\beta(s')) - 2B(\beta(\bar{s}))$$

$$
\begin{aligned}
&= \int_0^s \zeta(q)\beta'(q)\mathrm{d}q + \int_0^s \zeta(q)\beta'(q)\mathrm{d}q - 2\int_0^{\bar{s}} \zeta(q)\beta'(q)\mathrm{d}q \\
&= \int_{\bar{s}}^s \zeta(q)\beta'(q)\mathrm{d}q + \int_{\bar{s}}^{s'} \zeta(q)\beta'(q)\mathrm{d}q \\
&= \int_{\bar{s}}^s (\zeta(q)-\zeta(\bar{s}))\beta'(q)\mathrm{d}q + \int_{\bar{s}}^{s'} (\zeta(q)-\zeta(\bar{s}))\beta'(q)\mathrm{d}q. \qquad (6.41)
\end{aligned}
$$

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 (6.40) follows from (6.41). ∎

The following property states an expected integration-by-parts result, which can be formally obained by writing $(\partial_t\beta(v))\zeta(v) = \beta'(v)\zeta(v)\partial_t v = \partial_t B(\beta(v))$ (owing to (6.38)). 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 6.24. *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* $t_1, t_2 \in [0,T]$,

$$
\begin{aligned}
&\int_{t_1}^{t_2} \langle \partial_t\beta(v)(t), \zeta(v(t))\rangle_{H^{-1},H_0^1}\mathrm{d}t \\
&\qquad = \int_\Omega B(\beta(v)(\boldsymbol{x},t_2))\mathrm{d}\boldsymbol{x} - \int_\Omega B(\beta(v)(\boldsymbol{x},t_1))\mathrm{d}\boldsymbol{x}. \qquad (6.42)
\end{aligned}
$$

Remark 6.25 (Continuity of $\beta(v)$)
Since $\eta = B \circ \beta$ satisfies (6.17), 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 6.18, 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

$$t \in \mathbb{R} \to D_h \overline{\beta(v)} := \frac{1}{h} \int_t^{t+h} \partial_t \overline{\beta(v)}(s) \mathrm{d}s \tag{6.43}$$

$$= \frac{\overline{\beta(v)}(t+h) - \overline{\beta(v)}(t)}{h} \in H^{-1}(\Omega) \tag{6.44}$$

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 C.6 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. $t \in (t_1, t_2)$,

$$\langle \overline{\beta(v)}(\tau) - \beta(v)(t), \zeta(v(\cdot, t)) \rangle_{H^{-1}, H^1_0}$$
$$\le \int_\Omega B(\overline{\beta(v)}(\boldsymbol{x}, \tau)) - B(\beta(v)(\boldsymbol{x}, t)) \mathrm{d}\boldsymbol{x}. \tag{6.45}$$

Let t such that $\beta(v)(\cdot, t) = \beta(v(\cdot, t))$ a.e. on Ω. Almost every t satisfies this property. By Remark 6.19, 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 (6.39) with $r = \overline{\beta(v)}(\boldsymbol{x}, \tau)$ and $a = v(\boldsymbol{x}, t)$,

$$B(\overline{\beta(v)}(\boldsymbol{x}, \tau)) - B(\beta(v(\boldsymbol{x}, t))) \ge \zeta(v(\boldsymbol{x}, t))(\overline{\beta(v)}(\boldsymbol{x}, \tau) - \beta(v(\boldsymbol{x}, t))).$$

Integrating this relation over $\boldsymbol{x} \in \Omega$, Property (6.45) follows since the H^{-1}–H_0^1 duality product in (6.45) can be replaced with an L^2 inner product, as all terms in this product belong to $L^2(\Omega)$.

Step 4: proof of (6.42)
By convergence of $D_h\overline{\beta(v)}$ to $\partial_t\overline{\beta(v)}$ in $L^2(0,T;H^{-1}(\Omega))$ and since $\mathbf{1}_{(t_1,t_2)}\zeta(v) \in L^2(\mathbb{R};H_0^1(\Omega))$, we have

$$\begin{aligned}
&\int_{t_1}^{t_2} \langle \partial_t \beta(v)(t), \zeta(v(t))\rangle_{H^{-1},H_0^1} \mathrm{d}t \\
&\qquad = \int_{\mathbb{R}} \langle \partial_t \overline{\beta(v)}(t), \mathbf{1}_{(t_1,t_2)}(t)\zeta(v(\cdot,t))\rangle_{H^{-1},H_0^1} \mathrm{d}t \\
&\qquad = \lim_{h\to 0} \int_{\mathbb{R}} \langle D_h \overline{\beta(v)}(t), \mathbf{1}_{(t_1,t_2)}(t)\zeta(v(\cdot,t))\rangle_{H^{-1},H_0^1} \mathrm{d}t \\
&\qquad = \lim_{h\to 0} \frac{1}{h} \int_{t_1}^{t_2} \langle \overline{\beta(v)}(s+h) - \overline{\beta(v)}(t), \zeta(v(\cdot,t)\rangle_{H^{-1},H_0^1} \mathrm{d}t. \qquad (6.46)
\end{aligned}$$

We then use (6.45) for a.e. $t \in (t_1,t_2)$ to obtain, for h small enough such that $t_1 + h < t_2$,

$$\begin{aligned}
&\frac{1}{h}\int_{t_1}^{t_2} \langle \overline{\beta(v)}(t+h) - \overline{\beta(v)}(t), \zeta(v(\cdot,t))\rangle_{H^{-1},H_0^1} \mathrm{d}t \\
&\qquad \le \frac{1}{h}\int_{t_1}^{t_2}\int_\Omega B(\overline{\beta(v)}(\boldsymbol{x},t+h)) - B(\overline{\beta(v)}(\boldsymbol{x},t))\mathrm{d}\boldsymbol{x}\mathrm{d}t \\
&\qquad = \frac{1}{h}\int_{t_2}^{t_2+h}\int_\Omega B(\overline{\beta(v)}(\boldsymbol{x},t))\mathrm{d}\boldsymbol{x}\mathrm{d}t - \frac{1}{h}\int_{t_1}^{t_1+h}\int_\Omega B(\overline{\beta(v)}(\boldsymbol{x},t))\mathrm{d}\boldsymbol{x}\mathrm{d}t \\
&\qquad = \int_\Omega B(\beta(v)(\boldsymbol{x},t_2))\mathrm{d}\boldsymbol{x} - \frac{1}{h}\int_{t_1}^{t_1+h}\int_\Omega B(\beta(v)(\boldsymbol{x},t))\mathrm{d}\boldsymbol{x}\mathrm{d}t. \qquad (6.47)
\end{aligned}$$

In the last line, we used $\overline{\beta(v)}(t) = \beta(v)(t_2)$ for all $t \ge t_2$. We then take the superior limit of (6.47), 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 (6.46) we obtain

$$\begin{aligned}
&\int_{t_1}^{t_2} \langle \partial_t \beta(v)(t), \zeta(v(t))\rangle_{H^{-1},H_0^1} \mathrm{d}t \\
&\qquad \le \int_\Omega B(\beta(v)(\boldsymbol{x},t_2))\mathrm{d}\boldsymbol{x} - \liminf_{h\to 0} \frac{1}{h}\int_{t_1}^{t_1+h}\int_\Omega B(\beta(v)(\boldsymbol{x},t))\mathrm{d}\boldsymbol{x}\mathrm{d}t. \qquad (6.48)
\end{aligned}$$

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 C.6 and Jensen's inequality give

$$\int_\Omega B(\beta(v)(\boldsymbol{x},t_1))\mathrm{d}\boldsymbol{x} \le \liminf_{h\to 0} \int_\Omega B\left(\frac{1}{h}\int_{t_1}^{t_1+h} \beta(v)(\boldsymbol{x},t)\mathrm{d}t\right)\mathrm{d}\boldsymbol{x}$$

$$\leq \liminf_{h\to 0} \int_\Omega \frac{1}{h} \int_{t_1}^{t_1+h} B(\beta(v)(\boldsymbol{x},t))\mathrm{d}t\mathrm{d}\boldsymbol{x}.$$

Plugged into (6.48), this inequality shows that (6.42) holds with $\leq$ 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 (6.42) with "$\leq$" 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 (6.42) with "$\geq$" instead of "$=$" and the proof of (6.42) 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 (6.42), 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 6.26. *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* $T_0\in[0,T]$,

$$\int_\Omega B(\beta(\overline{u})(\boldsymbol{x},T_0))\mathrm{d}\boldsymbol{x} + \int_0^{T_0}\int_\Omega \Lambda(\boldsymbol{x})\nabla\zeta(\overline{u})(\boldsymbol{x},t)\cdot\nabla\zeta(\overline{u})(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t$$
$$= \int_\Omega B(\beta(u_{\mathrm{ini}}(\boldsymbol{x})))\mathrm{d}\boldsymbol{x} + \int_0^{T_0}\int_\Omega f(\boldsymbol{x},t)\zeta(\overline{u})(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t; \quad (6.49)$$

3. $\nu(\overline{u})$ *is continuous* $[0,T]\to L^2(\Omega)$.

Remark 6.27 (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 6.18), 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 6.17 was established using Theorems 6.4 and 6.14, which do not make use of Corollary 6.26. Hence, we invoke Corollary 6.17, which tells us that $\overline{u}$ is also a solution to (6.25). 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 (6.49) therefore follow from Lemma 6.24 applied to $v=\overline{u}$, by using $\overline{v}=\zeta(\overline{u})\mathbf{1}_{[0,T_0]}$ in (6.25).

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 (6.40),

$$\begin{aligned}
&\int_\Omega [\nu(\overline{u}(\boldsymbol{x},s_l)) - \nu(\overline{u}(\boldsymbol{x},t_k))]^2 \mathrm{d}\boldsymbol{x} \\
&\quad \le 4L_\beta L_\zeta \left(\int_\Omega B(\beta(\overline{u}(\boldsymbol{x},s_l)))\mathrm{d}\boldsymbol{x} + \int_\Omega B(\beta(\overline{u}(\boldsymbol{x},t_k)))\mathrm{d}\boldsymbol{x} \right) \\
&\qquad - 8L_\beta L_\zeta \int_\Omega B\left(\frac{\beta(\overline{u}(\boldsymbol{x},s_l)) + \beta(\overline{u}(\boldsymbol{x},t_k))}{2} \right) \mathrm{d}\boldsymbol{x} \\
&\quad = 4L_\beta L_\zeta \left(\int_\Omega B(\beta(\overline{u})(\boldsymbol{x},s_l))\mathrm{d}\boldsymbol{x} + \int_\Omega B[\beta(\overline{u})(\boldsymbol{x},t_k))\mathrm{d}\boldsymbol{x} \right) \\
&\qquad - 8L_\beta L_\zeta \int_\Omega B\left(\frac{\beta(\overline{u})(\boldsymbol{x},s_l) + \beta(\overline{u})(\boldsymbol{x},t_k)}{2} \right) \mathrm{d}\boldsymbol{x}.
\end{aligned} \tag{6.50}$$

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 C.6 and the convexity of B (Lemma 6.23) 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 (6.50) and using the continuity of $t\mapsto \int_\Omega B(\beta(\overline{u})(\boldsymbol{x},t))\mathrm{d}\boldsymbol{x}$ thus shows that

$$\|\nu(\overline{u}(\cdot,s_l)) - \nu(\overline{u}(\cdot,t_k))\|_{L^2(\Omega)} \to 0 \quad \text{ as } l,k\to\infty. \tag{6.51}$$

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 $t_k=s_k$, (6.51) 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 (6.51) 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

$$\begin{aligned}
\|\nu(\overline{u})(\cdot,\tau_r) - \nu(\overline{u}(\cdot,s_r))\|_{L^2(\Omega)} &\le \frac{1}{r}, \\
\|\nu(\overline{u})(\cdot,\tau) - \nu(\overline{u}(\cdot,t_r))\|_{L^2(\Omega)} &\le \frac{1}{r}.
\end{aligned}$$

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 (6.51) 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 (6.6)). 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 [17], 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 [60]. This idea consists in using the solution to an approximate dual problem. It was successfully applied to the one-dimensional Stefan problem in [9], and subsequently generalised to the higher dimensional case in [59].
The proof provided here was originally developed in [38] 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 $\psi \in L^2(0,T;H^1_0(\Omega))$ 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

$$\int_0^T \int_\Omega u_d(\boldsymbol{x},t)\Big((1-q(\boldsymbol{x},t))\partial_t\psi(\boldsymbol{x},t) + q(\boldsymbol{x},t)\mathrm{div}(\Lambda\nabla\psi)(\boldsymbol{x},t)\Big)\mathrm{d}\boldsymbol{x}\mathrm{d}t = 0. \tag{6.52}$$

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

$$\frac{(q_\varepsilon - q)^2}{q_\varepsilon} \leq \varepsilon \quad \text{and} \quad \frac{(q_\varepsilon - q)^2}{1-q_\varepsilon} \leq \varepsilon. \tag{6.53}$$

Let ψ_ε be given by Lemma 6.28 below, with $g = q_\varepsilon$ and some $w \in C_c^\infty(\Omega\times(0,T))$. Making $\psi = \psi_\varepsilon$ in (6.52) and using (6.56),

$$\left|\int_0^T\int_\Omega u_d(\boldsymbol{x},t)w(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t\right|$$
$$\leq\left|\int_0^T\int_\Omega u_d(\boldsymbol{x},t)(q_\varepsilon(\boldsymbol{x},t)-q(\boldsymbol{x},t))(\mathrm{div}(\Lambda\nabla\psi_\varepsilon)(\boldsymbol{x},t)-\partial_t\psi_\varepsilon(\boldsymbol{x},t))\mathrm{d}\boldsymbol{x}\mathrm{d}t\right|. \tag{6.54}$$

The Cauchy-Schwarz inequality, (6.57) and (6.53) imply

$$\begin{aligned}
&\left[\int_0^T\int_\Omega u_d(x,t)(q_\varepsilon(x,t)-q(x,t))(\mathrm{div}(\Lambda\nabla\psi_\varepsilon)(x,t)-\partial_t\psi_\varepsilon(x,t))\mathrm{d}x\mathrm{d}t\right]^2\\
&\quad\leq 2\left(\int_0^T\int_\Omega u_d(\boldsymbol{x},t)^2\frac{(q(\boldsymbol{x},t)-q_\varepsilon(\boldsymbol{x},t))^2}{q_\varepsilon(\boldsymbol{x},t)}\mathrm{d}\boldsymbol{x}\mathrm{d}t\right)\\
&\qquad\times\left(\int_0^T\int_\Omega q_\varepsilon(\boldsymbol{x},t)\Big(\mathrm{div}(\Lambda\nabla\psi_\varepsilon)(\boldsymbol{x},t)\Big)^2\mathrm{d}\boldsymbol{x}\mathrm{d}t\right)\\
&\quad+2\left(\int_0^T\int_\Omega u_d(\boldsymbol{x},t)^2\frac{(q(\boldsymbol{x},t)-q_\varepsilon(x,t))^2}{1-q_\varepsilon(\boldsymbol{x},t)}\mathrm{d}\boldsymbol{x}\mathrm{d}t\right)\\
&\qquad\times\left(\int_0^T\int_\Omega(1-q_\varepsilon(\boldsymbol{x},t))\Big(\partial_t\psi_\varepsilon(\boldsymbol{x},t)\Big)^2\mathrm{d}\boldsymbol{x}\mathrm{d}t\right)\\
&\quad\leq 2\varepsilon C_0\left\|u_d\right\|_{L^2(\Omega\times(0,T))}\\
&\qquad\times\Big(\|\nabla w\|^2_{L^2(\Omega\times(0,T))^d}+\|w\|^2_{L^2(\Omega\times(0,T))}+\|\partial_t w\|^2_{L^2(\Omega\times(0,T))}\Big).
\end{aligned}\tag{6.55}$$

Letting $\varepsilon\to 0$ and using (6.54) 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 6.28. *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. $(\boldsymbol{x},t)\in\Omega\times(0,T)$,

$$(1-g(\boldsymbol{x},t))\partial_t\psi(\boldsymbol{x},t)+g(\boldsymbol{x},t)\mathrm{div}(\Lambda\nabla\psi)(\boldsymbol{x},t)=w(\boldsymbol{x},t); \tag{6.56}$$

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*

$$\begin{aligned}&\int_0^T\int_\Omega\left((1-g(\boldsymbol{x},t))\Big(\partial_t\psi(\boldsymbol{x},t)\Big)^2+g(\boldsymbol{x},t)\Big(\mathrm{div}(\Lambda\nabla\psi)(\boldsymbol{x},t)\Big)^2\right)\mathrm{d}\boldsymbol{x}\mathrm{d}t\\&\leq C_0\left(\|\nabla w\|^2_{L^2(\Omega\times(0,T))^d}+\|w\|^2_{L^2(\Omega\times(0,T))}+\|\partial_t w\|^2_{L^2(\Omega\times(0,T))}\right).\end{aligned} \tag{6.57}$$

Proof.

Step 1: existence of ψ satisfying 1, 2 and 3.
After dividing through by g, observe that (6.56) is equivalent to

$$\text{For a.e. } (\boldsymbol{x},t)\in\Omega\times(0,T)\,,\qquad \Phi(\boldsymbol{x},t)\partial_t\psi(\boldsymbol{x},t)+\mathrm{div}(\Lambda(\boldsymbol{x})\nabla\psi(\boldsymbol{x},t))=f(\boldsymbol{x},t), \tag{6.58}$$

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 (6.58) 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 $H^1_0(\Omega)$. We look for $\psi_k:[0,T]\to V_k$ solution to the following Galerkin approximation of (6.58), with final condition:

$$\begin{aligned}&\psi_k(T)=0 \text{ and } \forall t\in[0,T]\,,\forall v\in V_k\,,\\&(\Phi(\cdot,t)\psi_k'(t),v)_{L^2}-(\Lambda\nabla\psi_k(t),\nabla v)_{(L^2)^d}=(f(\cdot,t),v)_{L^2}.\end{aligned} \tag{6.59}$$

Here, $(\cdot,\cdot)_{L^2}$ is the $L^2(\Omega)$ inner product. Choosing an orthonormal (for this inner product) basis $(\mathbf{e}_i)_{i=1,\dots,N_k}$ of V_k and writing $\psi_k(t)=\sum_{i=1}^{N_k}\theta_i(t)\mathbf{e}_i$, (6.59) can be re-cast as

$$\Theta(T)=0 \text{ and, for all } t\in[0,T],\ M(t)\Theta'(t)-S(t)\Theta(t)=F(t) \tag{6.60}$$

where $\Theta(t)=(\theta_i(t))_{i=1,\dots,N_k}$, $M(t)$ 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 $F(t)=((f(\cdot,t),\mathbf{e}_j))_{j=1,\dots,N_k}$. Since $\Phi\geq\varphi^*$ and $(\mathbf{e}_i)_{i=1,\dots,N_k}$ 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 (6.60) 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 (6.59), with $\psi_k\in W^{1,\infty}(0,T;V_k)\subset W^{1,\infty}(0,T;H^1_0(\Omega))$. 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 (6.59) 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 (C.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 Aubin–Simon's 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 (6.59) 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 $\rho(\boldsymbol{x},t) = \theta(t)v(\boldsymbol{x})$,

$$\begin{aligned}\int_0^T\int_\Omega \Phi(\boldsymbol{x},t)\partial_t\psi(\boldsymbol{x},t)\rho(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t - \int_0^T\int_\Omega \Lambda(\boldsymbol{x})\nabla\psi(\boldsymbol{x},t)\cdot\nabla\rho(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t \\ = \int_0^T\int_\Omega f(\boldsymbol{x},t)\rho(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t. \quad (6.61)\end{aligned}$$

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 [29]). Hence, (6.61) also holds for any $\rho\in L^2(0,T;H_0^1(\Omega))$. Considering smooth compactly supported functions ρ, (6.61) 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 (6.61), that (6.58) is satisfied.
Note that Lemma 6.29 below provides an additional regularity property and an integration-by-part formula on ψ.

Step 2: proof of (6.57).
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 (6.56) 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 (6.65) in Lemma 6.29, and recalling that $\psi(\cdot,T)=0$, we obtain

$$\begin{aligned}
&\frac{1}{2}\int_\Omega \Lambda(\boldsymbol{x})\nabla\psi(\boldsymbol{x},s)\cdot\nabla\psi(\boldsymbol{x},s)\mathrm{d}\boldsymbol{x} \\
&\quad + \int_s^T\int_\Omega \left((1-g(\boldsymbol{x},t))\Big(\partial_t\psi(\boldsymbol{x},t)\Big)^2 + g(\boldsymbol{x},t)\Big(\mathrm{div}(\Lambda\nabla\psi)(\boldsymbol{x},t)\Big)^2\right)\mathrm{d}\boldsymbol{x}\mathrm{d}t \\
&= -\int_s^T\int_\Omega \Lambda(\boldsymbol{x})\nabla w(\boldsymbol{x},t)\cdot\nabla\psi(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t - \int_\Omega w(\boldsymbol{x},s)\psi(\boldsymbol{x},s)\mathrm{d}\boldsymbol{x} \\
&\quad - \int_s^T\int_\Omega \psi(\boldsymbol{x},t)\partial_t w(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t. \qquad (6.62)
\end{aligned}$$

Integrating (6.62) with respect to $s\in(0,T)$ leads to

$$\begin{aligned}
&\frac{1}{2}\int_0^T\int_\Omega \Lambda(\boldsymbol{x})\nabla\psi(\boldsymbol{x},s)\cdot\nabla\psi(\boldsymbol{x},s)\mathrm{d}\boldsymbol{x}\mathrm{d}s \\
&\le T\int_0^T\int_\Omega |\Lambda(\boldsymbol{x})\nabla w(\boldsymbol{x},t)\cdot\nabla\psi(\boldsymbol{x},t)|\mathrm{d}\boldsymbol{x}\mathrm{d}t + \int_0^T\int_\Omega |w(\boldsymbol{x},s)\psi(\boldsymbol{x},s)|\mathrm{d}\boldsymbol{x}\mathrm{d}s \\
&\quad + T\int_0^T\int_\Omega |\psi(\boldsymbol{x},t)\partial_t w(\boldsymbol{x},t)|\mathrm{d}\boldsymbol{x}\mathrm{d}t. \qquad (6.63)
\end{aligned}$$

Apply the Cauchy-Schwarz and Poincaré inequalities to obtain

$$\begin{aligned}
\frac{\underline{\lambda}}{2}\|\nabla\psi\|_{L^2(\Omega\times(0,T))^d} &\le T\overline{\lambda}\|\nabla w\|_{L^2(\Omega\times(0,T))} \\
&\quad + \mathrm{diam}(\Omega)\left(\|w\|_{L^2(\Omega\times(0,T))} + T\|\partial_t w\|_{L^2(\Omega\times(0,T))}\right). \qquad (6.64)
\end{aligned}$$

Letting $s=0$ in (6.62), recalling that $w(\cdot,0)=0$, and using (6.64) 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 (6.64), this shows that (6.57) holds. ∎

Lemma 6.29. *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 $s,\tau\in[0,T]$,*

$$\int_s^\tau \int_\Omega \partial_t\psi(\boldsymbol{x},t)\mathrm{div}(\Lambda\nabla\psi)(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t = -\frac{1}{2}\int_\Omega \Lambda(\boldsymbol{x})\nabla\psi(\boldsymbol{x},\tau)\cdot\nabla\psi(\boldsymbol{x},\tau)\mathrm{d}\boldsymbol{x} + \frac{1}{2}\int_\Omega \Lambda(\boldsymbol{x})\nabla\psi(\boldsymbol{x},s)\cdot\nabla\psi(\boldsymbol{x},s)\mathrm{d}\boldsymbol{x}. \quad (6.65)$$

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 $\psi \in C([0,T];L^2(\Omega))$,

$$\begin{aligned}(\psi(t_n),\gamma)_{H_0^1} \to & -\int_\Omega \psi(\boldsymbol{x},t)\Delta\gamma(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\mathrm{d}t \\ &= \int_\Omega \nabla\psi(\boldsymbol{x},t)\cdot\nabla\gamma(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\mathrm{d}t = (\psi(t),\gamma)_{H_0^1}. \end{aligned} \quad (6.66)$$

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

$$\begin{aligned}&\left|(\psi(t_n),\gamma)_{H_0^1} - (\psi(t),\gamma)_{H_0^1}\right| \\ &\quad\le \left|(\psi(t_n),\gamma)_{H_0^1} - (\psi(t_n),\gamma_\varepsilon)_{H_0^1}\right| + \left|(\psi(t_n),\gamma_\varepsilon)_{H_0^1} - (\psi(t),\gamma_\varepsilon)_{H_0^1}\right| \\ &\qquad + \left|(\psi(t),\gamma_\varepsilon)_{H_0^1} - (\psi(t),\gamma)_{H_0^1}\right| \\ &\quad\le M\varepsilon + \left|(\psi(t_n),\gamma_\varepsilon)_{H_0^1} - (\psi(t),\gamma_\varepsilon)_{H_0^1}\right| + M\varepsilon.\end{aligned}$$

Taking the superior limit as $n \to \infty$ (using (6.66) 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 (6.65).
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

$$\overline{\psi}(t) = \begin{cases} \psi(s) & \text{if } t \le s, \\ \psi(t) & \text{if } t \in (s,\tau), \\ \psi(\tau) & \text{if } t \ge \tau. \end{cases}$$

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,

$$\begin{aligned}
&\int_s^\tau \langle \partial_t\overline{\psi}_n(t), \operatorname{div}(\Lambda\nabla\overline{\psi}_n)(t)\rangle_{H_0^1,H^{-1}}\mathrm{d}t \\
&\qquad = -\int_s^\tau\int_\Omega \partial_t\nabla\overline{\psi}_n(\boldsymbol{x},t)\cdot\Lambda(\boldsymbol{x})\nabla\overline{\psi}_n(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t \\
&\qquad = -\frac{1}{2}\int_s^\tau \frac{d}{dt}\int_\Omega \Lambda(\boldsymbol{x})\nabla\overline{\psi}_n(\boldsymbol{x},t)\cdot\nabla\overline{\psi}_n(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\mathrm{d}t \\
&\qquad = -\frac{1}{2}\int_\Omega \Lambda(\boldsymbol{x})\nabla\overline{\psi}_n(\boldsymbol{x},\tau)\cdot\nabla\overline{\psi}_n(\boldsymbol{x},\tau)\mathrm{d}\boldsymbol{x} \\
&\qquad\quad + \frac{1}{2}\int_\Omega \Lambda(\boldsymbol{x})\nabla\overline{\psi}_n(\boldsymbol{x},s)\cdot\nabla\overline{\psi}_n(\boldsymbol{x},s)\mathrm{d}\boldsymbol{x}. \qquad (6.67)
\end{aligned}$$

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 $n\in\mathbb{N}$,

$$\begin{aligned}
&\frac{1}{2}\int_\Omega \Lambda(\boldsymbol{x})\nabla\overline{\psi}_n(\boldsymbol{x},\tau)\cdot\nabla\overline{\psi}_n(\boldsymbol{x},\tau)\mathrm{d}\boldsymbol{x} \\
&\qquad\qquad = \frac{1}{2}\int_\Omega \Lambda(\boldsymbol{x})\nabla\psi(\boldsymbol{x},\tau)\cdot\nabla\psi(\boldsymbol{x},\tau)\mathrm{d}\boldsymbol{x}. \qquad (6.68)
\end{aligned}$$

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

$$\liminf_{n\to\infty} \frac{1}{2}\int_\Omega \Lambda(\boldsymbol{x})\nabla\overline{\psi}_n(\boldsymbol{x},s)\cdot\nabla\overline{\psi}_n(\boldsymbol{x},s)\mathrm{d}\boldsymbol{x} \\ \geq \frac{1}{2}\int_\Omega \Lambda(\boldsymbol{x})\nabla\psi(\boldsymbol{x},s)\cdot\nabla\psi(\boldsymbol{x},s)\mathrm{d}\boldsymbol{x}. \quad (6.69)$$

Dealing with the left-hand side of (6.67) 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 (6.67) can therefore be re-cast as

$$\begin{aligned}
&\int_s^\tau \langle \partial_t\overline{\psi}_n(t), \mathrm{div}(\Lambda\nabla\overline{\psi}_n)(t)\rangle_{H_0^1,H^{-1}}\mathrm{d}t \\
&\quad= \int_s^\tau\int_\Omega \partial_t\overline{\psi}_n(\boldsymbol{x},t)\left[\mathrm{div}(\Lambda\nabla\psi)(\boldsymbol{x},\cdot)\mathbf{1}_{(s,\tau)}\right] * \rho_n(t)\mathrm{d}\boldsymbol{x}\mathrm{d}t \\
&\qquad+ \int_s^\tau \langle \partial_t\overline{\psi}_n(t), \mathrm{div}(\Lambda\nabla\psi(\tau))\rangle_{H_0^1,H^{-1}}(\mathbf{1}_{[\tau,\infty)} * \rho_n)(t)\mathrm{d}t \\
&\quad= \int_s^\tau\int_\Omega \partial_t\overline{\psi}_n(\boldsymbol{x},t)\left[\mathrm{div}(\Lambda\nabla\psi)(\boldsymbol{x},\cdot)\mathbf{1}_{(s,\tau)}\right] * \rho_n(t)\mathrm{d}\boldsymbol{x}\mathrm{d}t + T_n, \quad (6.70)
\end{aligned}$$

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

$$T_n = F_n(\tau)(\mathbf{1}_{[\tau,\infty)} * \rho_n)(\tau) - F_n(s)(\mathbf{1}_{[\tau,\infty)} * \rho_n)(s) \\ - \int_s^\tau F_n(t)(\mathbf{1}_{[\tau,\infty)} * \rho_n)'(t)\mathrm{d}t.$$

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 $n \to \infty$,

$$T_n = F_n(\tau) - \int_s^\tau F_n(t)(\mathbf{1}_{[\tau,\infty)} * \rho_n)'(t)\mathrm{d}t \to F(\tau) - F(\tau) = 0. \quad (6.71)$$

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

$$\big[\mathrm{div}(\Lambda\nabla\psi)(\boldsymbol{x},\cdot)\mathbf{1}_{(s,\tau)}\big]*\rho_n \to \mathrm{div}(\Lambda\nabla\psi)(\boldsymbol{x},\cdot)\mathbf{1}_{(s,\tau)} \quad \text{ in } L^2(\Omega\times\mathbb{R}).$$

Using (6.71), we can therefore pass to the limit in (6.70) 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 (6.67), (6.68) and (6.69), this gives (6.65) 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 6.24.

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 (6.65) 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 $H_0^1(\Omega)$. ∎

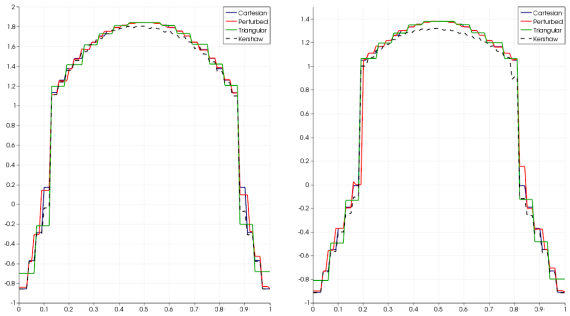
6.8 Numerical example

We consider a 2D test case with $\beta(s)=s$ 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 6.4 (such meshes are standard in the framework of underground

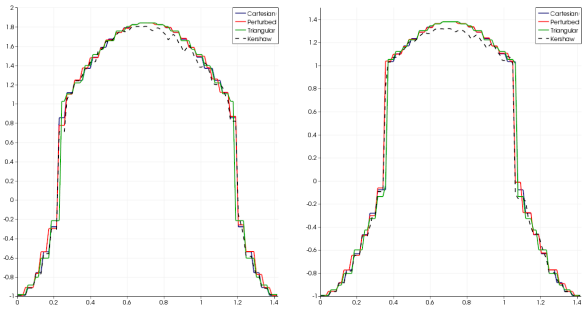
engineering). The time simulation is 0.1 for a constant given time step of 0.001.
Figures 6.4, 6.5, 6.6 and 6.7 represent the discrete solution $u(\cdot, t)$ on all grids for $t = .025, 0.05, 0.075$ and 0.1. For a better comparison we have also plotted the interpolation of u along two lines of the mesh. 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 6.2 and 6.3.
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 $\kappa_{\mathfrak{T}}$ – see (7.10), that is high ratios for some cells between the radii of inscribed balls and the diameter of the cell).



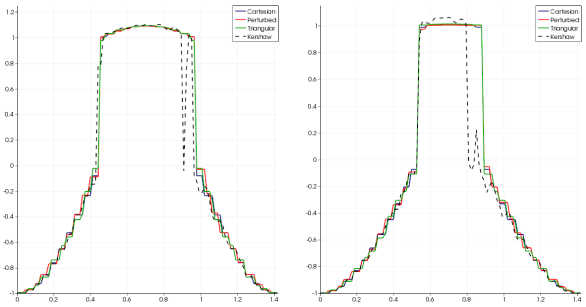
(a) $t = 0.025$ (b) $t = 0.050$

(c) $t = 0.075$ (d) $t = 0.1$

Fig. 6.2. 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.

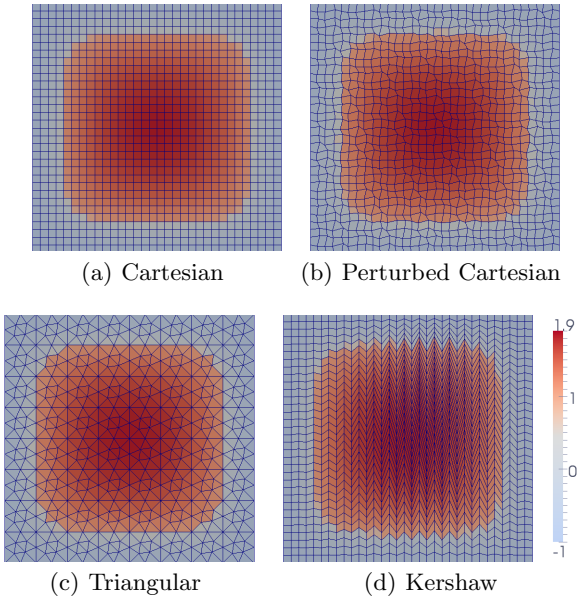


(a) $t = 0.025$ (b) $t = 0.050$



(c) $t = 0.075$ (d) $t = 0.1$

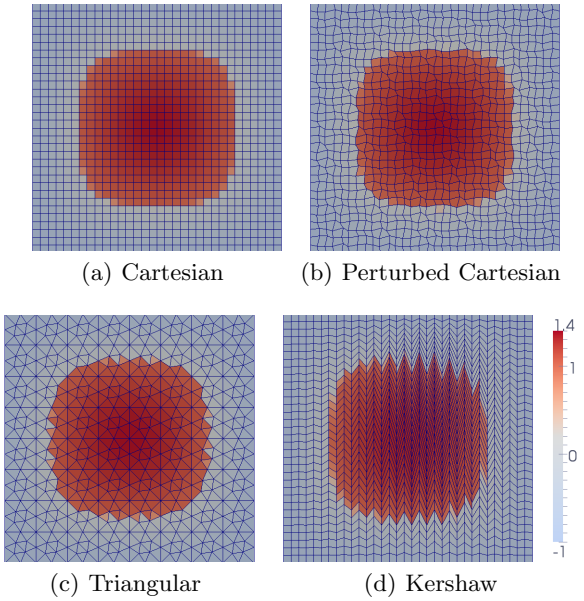
Fig. 6.3. Interpolation of u along a diagonal axe of the mesh for each grids: Cartesian in blue, perturbed Cartesien in red, triangular in green, and Kershaw in black dashed.



(a) Cartesian (b) Perturbed Cartesian

(c) Triangular (d) Kershaw

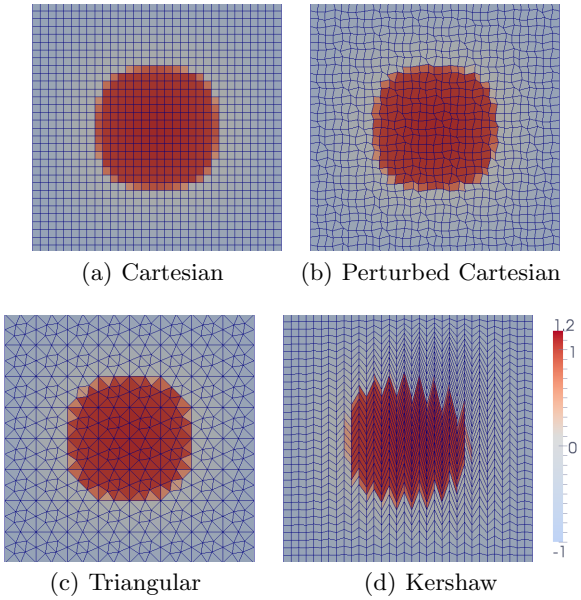
Fig. 6.4. Discrete solution u on all grids at $t = 0.025$.



(a) Cartesian (b) Perturbed Cartesian

(c) Triangular (d) Kershaw

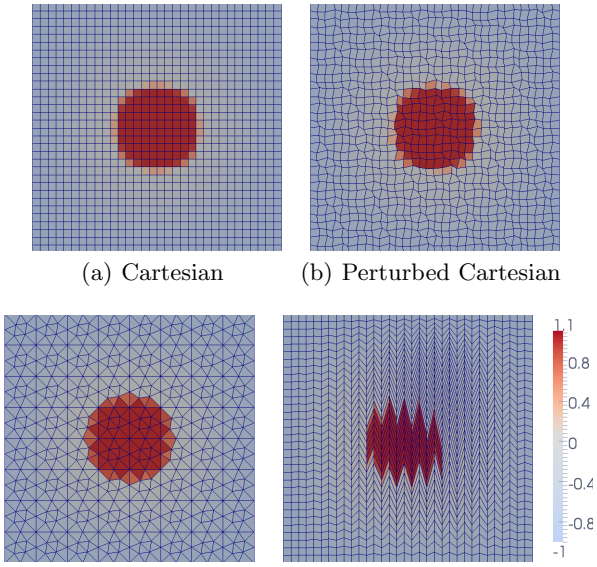
Fig. 6.5. Discrete solution u on all grids at $t = 0.050$.



(a) Cartesian (b) Perturbed Cartesian

(c) Triangular (d) Kershaw

Fig. 6.6. Discrete solution u on all grids at $t = 0.075$.



(a) Cartesian (b) Perturbed Cartesian

(c) Triangular (d) Kershaw

Fig. 6.7. Discrete solution u on all grids at $t = 0.1$.

Part III

Review of gradient discretisation methods

In this part several classical and popular numerical methods are shown to fit in the gradient discretisation method (GDM).
Most of the considered schemes are mesh-based numerical methods; hence Chapter 7 begins with meshes and discrete tools which are used to establish 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.
Each of the chapters 8 to 13 are devoted to a particular well known class of methods, namely: conforming Galerkin methods, non-conforming finite element methods and derived methods, mixed finite element $\mathbb{RT}_k$ schemes, the multi-point flux approximation (MPFA)-O scheme, hybrid mimetic mixed schemes, nodal mimetic finite difference methods (which is also compared with the CeVeFE-DDFV method). For each of these methods, a gradient discretisation is constructed in such a way that the corresponding gradient scheme (GS) (3.4) for the standard linear diffusion model (3.1) corresponds to the considered numerical method applied to this model.
The properties (defined in Part I) of the GDs thus constructed are then analysed. Once these known numerical methods are recasted 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.

As in the previous part, in all the following chapters 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 $\partial\Omega$ (except for Galerkin methods, Ω is in fact polytopal in this part).

7

Meshes and discrete tools

This chapter is devoted to the introduction of polytopal meshes, which are used in most of the examples of GDs reviewed in this part.
Section 7.1 presents the notion of *polytopal toolbox* which enables the "control" of numerous GDs by mapping them into polytopal toolboxes. Using the discrete functional analysis tools of Appendix B, this notion of "control" of a GD by a polytopal toolbox is shown to give the *coercivity*, *compactness* and *limit-conformity* of numerous mesh-based GDs. Precise estimates on $C_{\mathcal{D}}$ and $W_{\mathcal{D}}$ are also established.
The notion of local linearly exact (LLE) gradient discretisation is then introduced and analysed in Section 7.3. It is shown to ensure the *consistency* of sequences of GDs. Section A.1 in the appendix extends the analysis done here to prove explicit estimates on $S_{\mathcal{D}}$, in the case $p > d/2$.
The notion of LLE GD also enables us to describe a generic process to eliminate unknowns in gradient schemes, by replacing them with barycentric combinations of other unknowns. This process is standard in the construction of numerous schemes (e.g., SUSHI [49], VAG [50, 52]) but always on a case-by-case basis. The description of this barycentric elimination, given in the context of LLE GDs, is detailed, generic and ensures the preservation of 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 of performing mass-lumping is set up so that, under a single easily-checked assumption, the mass-lumped GDs enjoys the same properties as the initial GDs.

Example 7.1 (Illustration of the notions)
Boxes such as this 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 meshes

7.1.1 Definition and notations

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

$$\mathcal{S}((\boldsymbol{x}_i)_{i=1,\ldots,k+1}) = \left\{ \sum_{i=1}^{k+1} \alpha_i \boldsymbol{x}_i \,:\, \alpha_i \geq 0, i = 1, \ldots, k+1, \sum_{i=1}^{k+1} \alpha_i = 1 \right\}. \quad (7.1)$$

An open d-polytope Ω is defined as the interior of the union of a finite number of d-simplices $(S_j)_{j=1,\ldots,M}$, 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

$$\overline{\Omega} = \bigcup_{j=1,\ldots,M} S_j$$

and Ω is the interior of $\overline{\Omega}$. The boundary of Ω is then the union of the faces of the simplices $(S_j)_{j=1,\ldots,M}$ which are not common to two different simplices. $\partial\Omega$ is therefore the union of $d-1$-simplices.
In this chapter, we work with the following conditions and notations:

$$\begin{array}{l} d \in \mathbb{N} \setminus \{0\} \text{ denotes the space dimension,} \\ \Omega \text{ is a } d\text{-polytopal bounded connected open subset of } \mathbb{R}^d, \\ \text{with boundary } \partial\Omega. \end{array} \quad (7.2)$$

Definition 7.2 (Polytopal mesh). *Let $\Omega \subset \mathbb{R}^d$ satisfy Assumption (7.2); a polytopal mesh of Ω is given by $\mathfrak{T} = (\mathcal{M}, \mathcal{F}, \mathcal{P}, \mathcal{V})$, 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 K.*

2. *$\mathcal{F} = \mathcal{F}_{\rm int} \cup \mathcal{F}_{\rm ext}$ is a finite family of disjoint subsets of $\overline{\Omega}$ (the "faces" of the mesh – "edges" in 2D), such that, for all $\sigma \in \mathcal{F}_{\rm int}$, σ is a non empty open subset of a hyperplane of $\mathbb{R}^d$ included in Ω and, for all $\sigma \in \mathcal{F}_{\rm ext}$, σ is a non empty open subset of $\partial\Omega$; furthermore, the $(d-1)$-dimensional measure $|\sigma|$ of any $\sigma \in \mathcal{F}$ is stricly positive, and we denote by $\overline{\boldsymbol{x}}_\sigma$ its center of mass. We assume that, 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 then denote by $\mathcal{M}_\sigma = \{K \in \mathcal{M}, \sigma \in \mathcal{F}_K\}$. We then assume that, for all $\sigma \in \mathcal{F}$, either $\mathcal{M}_\sigma$ has exactly one element and then $\sigma \in \mathcal{F}_{\rm ext}$ or $\mathcal{M}_\sigma$ has exactly two elements and then $\sigma \in \mathcal{F}_{\rm int}$. For all $K \in \mathcal{M}$ and for any $\sigma \in \mathcal{F}_K$, we denote by $\boldsymbol{n}_{K,\sigma}$ the (constant) unit vector normal to σ outward to K.*
   For all $K \in \mathcal{M}$, we denote by $\mathcal{N}_K$ the set of the neighbours of K:

$$\mathcal{N}_K = \{L \in \mathcal{M} \setminus \{K\},\ \exists\sigma \in \mathcal{F}_{\rm int},\ \mathcal{M}_\sigma = \{K, L\}\}. \tag{7.3}$$

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:

$$h_{\mathcal{M}} = \sup\{h_K, K \in \mathcal{M}\}. \tag{7.6}$$

Remark 7.3. The above definition applies to a large variety of meshes. In particular, the cells are not assumed to be convex. Hence generalized "hexahedra" with non planar faces can be used (in fact, such sets have then 12 faces if each non planar face is shared in two triangles, but only 6 neighbouring cells).

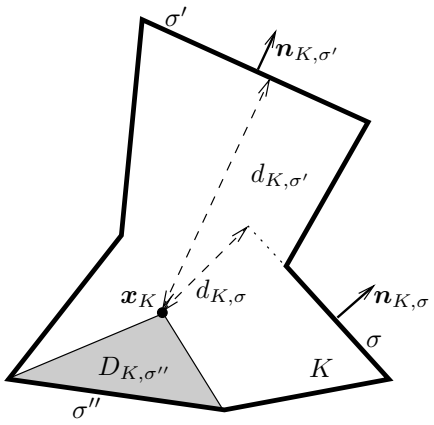


Fig. 7.1. A cell K of a polytopal mesh

Remark 7.4. The common boundary of two neighbouring cells can include more than one face.

A number of Finite Element methods require the notion of simplicial mesh.

Definition 7.5 (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 centers 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 $\mathit{\Pi}_{\mathfrak{T}} : X_{\mathfrak{T}} \to L^\infty(\Omega)$, 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

$$\forall v \in X_{\mathfrak{T}},\ \forall K \in \mathcal{M},\ \text{for a.e. } \boldsymbol{x} \in K,\ \Pi_{\mathfrak{T}} v(\boldsymbol{x}) = v_K, \tag{7.7c}$$

$$\forall v \in X_{\mathfrak{T}},\ \forall \sigma \in \mathcal{F}_{\rm ext},\ \text{for a.e. } \boldsymbol{x} \in \sigma,\ \mathbb{T}_{\mathfrak{T}} v(\boldsymbol{x}) = v_\sigma, \tag{7.7d}$$

$$\begin{aligned} &\forall v \in X_{\mathfrak{T}},\ \forall K \in \mathcal{M},\ \text{for a.e. } \boldsymbol{x} \in K,\\ &\qquad \overline{\nabla}_{\mathfrak{T}} v(\boldsymbol{x}) = \overline{\nabla}_K v := \frac{1}{|K|} \sum_{\sigma \in \mathcal{F}_K} |\sigma| (v_\sigma - v_K) \boldsymbol{n}_{K,\sigma}\\ &\qquad\qquad = \frac{1}{|K|} \sum_{\sigma \in \mathcal{F}_K} |\sigma| v_\sigma \boldsymbol{n}_{K,\sigma}. \end{aligned} \tag{7.7e}$$

We notice that the last equality in (7.7e) comes from 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 $p \in [1, +\infty)$ we define a discrete $W^{1,p}$ semi-norm on $X_{\mathfrak{T}}$ by

$$\forall v \in X_{\mathfrak{T}},\ |v|_{\mathfrak{T},p}^p = \sum_{K \in \mathcal{M}} \sum_{\sigma \in \mathcal{F}_K} |\sigma| d_{K,\sigma} \left| \frac{v_\sigma - v_K}{d_{K,\sigma}} \right|^p. \tag{7.7f}$$

We remark that $|\cdot|_{\mathfrak{T},p}$ is in fact a norm when restricted to $X_{\mathfrak{T},0}$.

Remark 7.6 (Cell-centred schemes)
For cell-centred schemes, whose unknowns are $v = (v_K)_{K \in \mathcal{M}}$, a more natural norm than (7.7f) would be

$$|v|_{\mathfrak{T},p,c}^p = \sum_{\sigma \in \mathcal{F}} |\sigma| d_{K,L} \left| \frac{v_L - v_K}{d_{K,L}} \right|^p$$

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 $\sigma \in \mathcal{F}_K \cap \mathcal{F}_{\rm ext}$. It is however very easy, for such schemes, to come back to a space and norm using cell and face unknowns as in (7.7a) and (7.7f). 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 which only has face unknowns, in order to use the results of this section) is also easy to do – see the analysis of non-conforming finite elements in Chapter 9.

Finally, for a given polytopal mesh $\mathfrak{T}$ we define two numbers that measure the regularity properties of the mesh:

$$\theta_{\mathfrak{T}} = \max_{K \in \mathcal{M}} \left(\max_{\sigma \in \mathcal{F}_K} \frac{h_K}{d_{K,\sigma}} + \mathrm{Card}(\mathcal{F}_K) \right), \tag{7.8}$$

$$\eta_{\mathfrak{T}} = \max_{\sigma \in \mathcal{F}_{\rm int},\ \mathcal{M}_\sigma = \{K,L\}} \left(\frac{d_{K,\sigma}}{d_{L,\sigma}} + \frac{d_{L,\sigma}}{d_{K,\sigma}} \right). \tag{7.9}$$

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 centered at the center of mass $\overline{\boldsymbol{x}}_K$ of K. It is proved in Lemma B.4 page 374 that, for simplicial meshes, this regularity factor controls the other two.

7.2 Polytopal toolboxes

Example 7.7 (GD for the non-conforming $\mathbb{P}_1$ finite elements) Since the non-conforming $\mathbb{P}_1$ finite element is used to illustrate notions introduced below, we need to briefly describe the corresponding gradient discretisation.
The non-conforming $\mathbb{P}_1$ finite element method is defined on a simplicial mesh $\mathfrak{T}$ (Definition 7.5). The DOFs of this method consist in face unknowns, gathered in the space

$$\begin{aligned} X_{\mathcal{D},0} = \{v = (v_\sigma)_{\sigma\in\mathcal{F}_K} \,:\, & v_\sigma \in \mathbb{R} \text{ for all } \sigma \in \mathcal{F}_{\rm int}\,, \\ & v_\sigma = 0 \text{ for all } \sigma \in \mathcal{F}_{\rm ext}\}. \end{aligned} \tag{7.11}$$

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 centers $(\overline{\boldsymbol{x}}_\sigma)_{\sigma\in\mathcal{F}}$, and takes the values $(v_\sigma)_{\sigma\in\mathcal{F}}$ at these centers. 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

A polytopal toolbox is nothing more than a polytopal mesh with associated reconstruction operators and norm.

Definition 7.8 (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 $(X_{\mathfrak{T},0}, \Pi_{\mathfrak{T}}, \overline{\nabla}_{\mathfrak{T}}, |\ |_{\mathfrak{T},p})$ 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)*:*

$$\forall v \in X_{\mathfrak{T},0},\ \forall K \in \mathcal{M},\ \text{for a.e. } \boldsymbol{x} \in K,$$
$$\overline{\nabla}_{\mathfrak{T}} v(\boldsymbol{x}) = \frac{1}{|K|} \sum_{\sigma \in \mathcal{F}_K} |\sigma|(v_\sigma - v_K)\boldsymbol{n}_{K,\sigma} = \frac{1}{|K|} \sum_{\sigma \in \mathcal{F}_K} |\sigma| v_\sigma \boldsymbol{n}_{K,\sigma}.$$

4. The space $X_{\mathfrak{T},0}$ *is endowed with the norm* (7.7f)*:*

$$\forall v \in X_{\mathfrak{T},0},\ |v|_{\mathfrak{T},p}^p = \sum_{K \in \mathcal{M}} \sum_{\sigma \in \mathcal{F}_K} |\sigma| d_{K,\sigma} \left| \frac{v_\sigma - v_K}{d_{K,\sigma}} \right|^p.$$

Remark 7.9. 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 $X_{\mathfrak{T},0}$.

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.), but the polytopal toolbox depends on these conditions as seen in Section 7.2.3. However, the context will always make clear which boundary conditions are considered, and thus which kind of polytopal toolbox should be used.
The notion of "control of a GD" by a polytopal toolbox consists in comparing the GD operators with those of a polytopal toolbox, through a linear mapping of the GD DOFs on the polytopal toolbox DOFs. Under some assumptions, this comparison enables us to establish the coercivity, limit-conformity and compactness of sequences of GDs.

Definition 7.10 (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 7.8. A control of* $\mathcal{D}$ *by* $\mathfrak{T}$ *is a linear mapping* $\boldsymbol{\Phi} : X_{\mathcal{D},0} \longrightarrow X_{\mathfrak{T},0}$*. 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 7.11 (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 DOFs 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 9.2 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 7.14 shows how such bounds are used.

Theorem 7.12 (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 7.8, and let $\boldsymbol{\Phi}$ be a control of $\mathcal{D}$ by $\mathfrak{T}$ in the sense of Definition 7.10. 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 triangular inequality, Lemma B.12 and the Hölder's inequality (C.7) we observe that, for any $v \in X_{\mathcal{D},0}$,

$$\begin{aligned}\|\Pi_{\mathcal{D}} v\|_{L^p(\Omega)} &\le \omega^{\Pi}(\mathcal{D},\mathfrak{T},\boldsymbol{\Phi})\,\|v\|_{\mathcal{D}} + \|\Pi_{\mathfrak{T}}\boldsymbol{\Phi}(v)\|_{L^p(\Omega)}\\ &\le \omega^{\Pi}(\mathcal{D},\mathfrak{T},\boldsymbol{\Phi})\,\|v\|_{\mathcal{D}} + C_{23}|\Omega|^{\frac{1}{p}-\frac{1}{q}}\,|\boldsymbol{\Phi}(v)|_{\mathfrak{T},p}\,,\end{aligned}$$

with q and C_{23} given in Lemma B.12. 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 triangular inequality, the definition of $\omega^{\Pi}(\mathcal{D},\mathfrak{T},\boldsymbol{\Phi})$ and (B.22) (notice that $\mathbb{T}_{\mathcal{D}}\boldsymbol{\Phi}(v) = 0$ since $\boldsymbol{\Phi}(v) \in X_{\mathfrak{T},0}$) to obtain

$$\begin{aligned}
&\left|\int_\Omega \Big(\nabla_{\mathcal{D}} v(\boldsymbol{x})\cdot\boldsymbol{\varphi}(\boldsymbol{x}) + \Pi_{\mathcal{D}} v(\boldsymbol{x})\mathrm{div}\boldsymbol{\varphi}(\boldsymbol{x})\Big)\mathrm{d}\boldsymbol{x}\right| \\
&\quad\le \left|\int_\Omega [\nabla_{\mathcal{D}} v(\boldsymbol{x}) - \overline{\nabla}_{\mathfrak{T}}\boldsymbol{\Phi}(v)(\boldsymbol{x})]\cdot\boldsymbol{\varphi}(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\right| + \|\mathrm{div}\boldsymbol{\varphi}\|_{L^{p'}(\Omega)}\,\omega^{\Pi}(\mathcal{D},\mathfrak{T},\boldsymbol{\Phi})\,\|v\|_{\mathcal{D}} \\
&\qquad + \left|\int_\Omega \big(\overline{\nabla}_{\mathfrak{T}}\boldsymbol{\Phi}(v)(\boldsymbol{x})\cdot\boldsymbol{\varphi}(\boldsymbol{x}) + \Pi_{\mathfrak{T}}\boldsymbol{\Phi}(v)(\boldsymbol{x})\mathrm{div}\boldsymbol{\varphi}(\boldsymbol{x})\big)\,\mathrm{d}\boldsymbol{x}\right| \\
&\quad\le \left|\int_\Omega [\nabla_{\mathcal{D}} v(\boldsymbol{x}) - \overline{\nabla}_{\mathfrak{T}}\boldsymbol{\Phi}(v)(\boldsymbol{x})]\cdot\boldsymbol{\varphi}(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\right| + \|\mathrm{div}\boldsymbol{\varphi}\|_{L^{p'}(\Omega)}\,\omega^{\Pi}(\mathcal{D},\mathfrak{T},\boldsymbol{\Phi})\,\|v\|_{\mathcal{D}} \\
&\qquad + C_{20}\,\|\,|\nabla\boldsymbol{\varphi}|\,\|_{L^{p'}(\Omega)^d}\,|\boldsymbol{\Phi}(v)|_{\mathfrak{T},p}\,h_{\mathcal{M}}. \qquad (7.15)
\end{aligned}$$

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.7 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,

$$\begin{aligned}
&\left|\int_\Omega [\nabla_{\mathcal{D}} v(\boldsymbol{x}) - \overline{\nabla}_{\mathfrak{T}}\boldsymbol{\Phi}(v)(\boldsymbol{x})]\cdot\boldsymbol{\varphi}(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\right| \\
&\quad= \left|\sum_{K\in\mathcal{M}}\int_K [\nabla_{\mathcal{D}} v(\boldsymbol{x}) - \overline{\nabla}_{\mathfrak{T}}\boldsymbol{\Phi}(v)(\boldsymbol{x})]\cdot\boldsymbol{\varphi}(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\right| \\
&\quad= \left|\sum_{K\in\mathcal{M}}\Big(\int_K \nabla_{\mathcal{D}} v(\boldsymbol{x})\cdot[\boldsymbol{\varphi}(\boldsymbol{x}) - \boldsymbol{\varphi}_K]\mathrm{d}\boldsymbol{x}\right. \\
&\qquad\qquad \left.+ \boldsymbol{\varphi}_K\cdot\int_K [\nabla_{\mathcal{D}} v(\boldsymbol{x}) - \overline{\nabla}_{\mathfrak{T}}\boldsymbol{\Phi}(v)(\boldsymbol{x})]\mathrm{d}\boldsymbol{x}\Big)\right| \\
&\quad\le C_2 h_{\mathcal{M}}\,\|\,|\nabla\boldsymbol{\varphi}|\,\|_{L^{p'}(\Omega)}\,\|\nabla_{\mathcal{D}} v\|_{L^p(\Omega)^d} \\
&\qquad + \sum_{K\in\mathcal{M}}|\boldsymbol{\varphi}_K|\left|\int_K [\nabla_{\mathcal{D}} v(\boldsymbol{x}) - \overline{\nabla}_{\mathfrak{T}}\boldsymbol{\Phi}(v)(\boldsymbol{x})]\mathrm{d}\boldsymbol{x}\right|
\end{aligned}$$

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

$$\begin{aligned}
&\left|\int_\Omega [\nabla_{\mathcal{D}} v(\boldsymbol{x}) - \overline{\nabla}_{\mathfrak{T}}\boldsymbol{\Phi}(v)(\boldsymbol{x})]\cdot\boldsymbol{\varphi}(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\right| \\
&\quad\le C_2 h_{\mathcal{M}}\,\|\,|\nabla\boldsymbol{\varphi}|\,\|_{L^{p'}(\Omega)}\,\|\nabla_{\mathcal{D}} v\|_{L^p(\Omega)^d}
\end{aligned}$$

$$+ \|\boldsymbol{\varphi}\|_{L^{p'}(\Omega)^d} \left(\sum_{K\in\mathcal{M}} |K|^{1-p} \left| \int_K \left[\nabla_{\mathcal{D}} v(\boldsymbol{x}) - \overline{\nabla}_{\mathfrak{T}} \boldsymbol{\Phi}(v)(\boldsymbol{x}) \right] \mathrm{d}\boldsymbol{x} \right|^p \right)^{1/p}$$
$$\leq \left(C_2 h_{\mathcal{M}} + \omega^{\nabla}(\mathcal{D},\mathfrak{T},\boldsymbol{\Phi}) \right) \|\boldsymbol{\varphi}\|_{W^{1,p'}(\Omega)^d} \|\nabla_{\mathcal{D}} v\|_{L^p(\Omega)^d}. \tag{7.16}$$

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 $p' = +\infty$), we 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. We can then use, in the previous calculations, the estimate $|\varphi(\boldsymbol{x}) - \varphi_K| \leq C_3 h_K \| \, |\nabla \boldsymbol{\varphi}| \, \|_{L^\infty(\Omega)}$ for any $\boldsymbol{x} \in K$. ∎

An immediate consequence of Theorem 7.12 is the following corollary.

Corollary 7.13 (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 7.8. 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 7.10, 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 2.6, and compact in the sense of Definition 2.8.

Example 7.14 (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 7.11, 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 2.14 and the fact that $W^{1,p'}(\Omega)^d$ is dense in $W^{\mathrm{div},p'}(\Omega)$ – see Remark 2.15).
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.15, 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

$$\|\Pi_{\mathcal{D}_m} u_m - \Pi_{\mathfrak{T}_m} \boldsymbol{\Phi}_m(u_m)\|_{L^p(\Omega)} \le \omega^{\Pi}(\mathcal{D}_m, \mathfrak{T}_m, \boldsymbol{\Phi}_m) \|u_m\|_{\mathcal{D}_m} \to 0$$

as $m \to \infty$, we deduce that $\Pi_{\mathcal{D}_m} u_m \to u$ in $L^p(\Omega)$ and the proof is complete. ∎

7.2.2 Non-homogeneous Dirichlet boundary conditions

The definition of coercivity, limit-conformity and compactness of GDs for non-homogeneous Dirichlet boundary conditions are identical to the same definitions for homogeneous Dirichlet conditions. Hence, all the previous results (and in particular Corollary 7.13) can be used to in the context of non-homogeneous Dirichlet boundary conditions.

Although polytopal meshes/toolboxes are not directly useful to establish the consistency of GDs, the structures provided by meshes can sometimes be used to construct interpolants of functions and prove that $S_{\mathcal{D}_m}(\varphi) \to 0$ for all smooth φ. In the context of non-homogenous Dirichlet boundary conditions, using Lemma 2.21 requires to check the condition (2.16), which can be facilitated by the following proposition.

Proposition 7.15 (Estimate of the discrete norm of an interpolate). *Let $\mathfrak{T}$ be a polytopal mesh of Ω in the sense of Definition 7.2, and let $\theta \ge \theta_{\mathfrak{T}}$ (see (7.8)). Let $\varphi \in W^{1,p}(\Omega)$ and define $v \in X_{\mathfrak{T}}$ by*

$$\begin{aligned} &\forall K \in \mathcal{M}, \quad v_K = \frac{1}{|K|} \int_K \varphi(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \\ &\forall \sigma \in \mathcal{F}, \quad v_\sigma = \frac{1}{|\sigma|} \int_\sigma \varphi(\boldsymbol{x}) \mathrm{d}s(\boldsymbol{x}). \end{aligned} \tag{7.17}$$

Then, there exists C_4 depending only on d, p and θ such that

$$\|\Pi_{\mathfrak{T}} v\|_{L^p(\Omega)} \le \|\varphi\|_{L^p(\Omega)} \quad \text{and} \quad |v|_{\mathfrak{T},p} \le C_4 \|\nabla \varphi\|_{L^p(\Omega)^d}. \tag{7.18}$$

Proof. By Jensen's inequality

$$|v_K|^p \le \frac{1}{|K|} \int_K |\varphi(\boldsymbol{x})|^p \mathrm{d}\boldsymbol{x}.$$

Multiplying this inequality by $|K|$ and summing on $K \in \mathcal{M}$ gives

$$\|\Pi_{\mathfrak{T}} v\|_{L^p(\Omega)} \le \|\varphi\|_{L^p(\Omega)}.$$

To estimate $|v|_{\mathfrak{T},p}$, we apply (B.11) in Lemma B.7 to find C_5 depending only on d, p and θ such that

$$|v_\sigma - v_K|^p \le \frac{C_5 h_K^{p-1}}{|\sigma|} \int_K |\nabla \varphi(\boldsymbol{x})|^p \mathrm{d}\boldsymbol{x}.$$

Multiplying this inequality by $|\sigma| d_{K,\sigma}^{1-p}$, summing on $\sigma \in \mathcal{F}_K$ and $K \in \mathcal{M}$, and using the definition of θ we deduce that $|v|_{\mathfrak{T},p}^p \le C_5 \theta^p \| \, |\nabla \varphi| \, \|_{L^p(\Omega)}^p$.

∎

7.2.3 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 7.20 and 7.21 we indicate the minor modifications that needs to be made to the following definitions and results for homogeneous Neumann and Fourier boundary conditions.

Definition 7.16 (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)*:*

$$\forall v \in X_{\mathfrak{T}},\ \forall K \in \mathcal{M},\ \text{for a.e. } \boldsymbol{x} \in K,\ \Pi_{\mathfrak{T}} v(\boldsymbol{x}) = v_K.$$

3. *The trace reconstruction $\mathbb{T}_{\mathfrak{T}} : X_{\mathfrak{T}} \to L^\infty(\partial\Omega)$ is defined by* (7.7d)*:*

$$\forall v \in X_{\mathfrak{T}},\ \forall \sigma \in \mathcal{F}_{\rm ext},\ \text{for a.e. } \boldsymbol{x} \in \sigma,\ \mathbb{T}_{\mathfrak{T}} v(\boldsymbol{x}) = v_\sigma.$$

4. *The gradient reconstruction $\overline{\nabla}_{\mathfrak{T}} : X_{\mathfrak{T}} \to L^\infty(\Omega)^d$ is defined by* (7.7e)*:*

$$\forall v \in X_{\mathfrak{T}},\ \forall K \in \mathcal{M},\ \text{for a.e. } \boldsymbol{x} \in K,$$
$$\overline{\nabla}_{\mathfrak{T}} v(\boldsymbol{x}) = \frac{1}{|K|} \sum_{\sigma\in\mathcal{F}_K} |\sigma|(v_\sigma - v_K)\boldsymbol{n}_{K,\sigma} = \frac{1}{|K|} \sum_{\sigma\in\mathcal{F}_K} |\sigma| v_\sigma \boldsymbol{n}_{K,\sigma}.$$

5. *Recalling the definition* (7.7f) *of the semi-norm $|\ |_{\mathfrak{T},p}$, the space $X_{\mathfrak{T}}$ is endowed with the norm*

$$\|v\|_{\mathfrak{T},p}^p = |v|_{\mathfrak{T},p}^p + \left| \int_\Omega \Pi_{\mathfrak{T}} v(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \right|^p. \tag{7.19}$$

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 7.17 (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 2.32, and let $\mathfrak{T}$ be a polytopal toolbox in the sense of Definition 7.16. A control of $\mathcal{D}$ by $\mathfrak{T}$ is a linear mapping $\boldsymbol{\Phi} : X_{\mathcal{D}} \longrightarrow X_{\mathfrak{T}}$. 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 7.18 (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 2.32, and let $\mathfrak{T}$ be a polytopal toolbox in the sense of Definition 7.16. We take $\boldsymbol{\Phi}$ a control of $\mathcal{D}$ by $\mathfrak{T}$ in the sense of Definition 7.17, and $\varrho \geq \theta_{\mathfrak{T}} + \eta_{\mathfrak{T}}$ (see* (7.8) *and* (7.9)*).*
Then, there exists C_6 depending only on Ω, p and ϱ such that

$$C_{\mathcal{D}} \leq \max\left(\omega^{\Pi}(\mathcal{D},\mathfrak{T},\boldsymbol{\Phi}),\omega^{\mathbb{T}}(\mathcal{D},\mathfrak{T},\boldsymbol{\Phi})\right) + C_6\,\|\boldsymbol{\Phi}\|_{\mathcal{D},\mathfrak{T}} \tag{7.20}$$

and, for all $\boldsymbol{\varphi}\in W^{1,p'}(\Omega)^d$,

$$W_{\mathcal{D}}(\boldsymbol{\varphi}) \leq \|\boldsymbol{\varphi}\|_{W^{1,p'}(\Omega)^d}\Big[C_6 h_{\mathcal{M}}(1+\|\boldsymbol{\Phi}\|_{\mathcal{D},\mathfrak{T}}) + \omega^{\Pi}(\mathcal{D},\mathfrak{T},\boldsymbol{\Phi}) + \omega^{\nabla}(\mathcal{D},\mathfrak{T},\boldsymbol{\Phi}) + \omega^{\mathbb{T}}(\mathcal{D},\mathfrak{T},\boldsymbol{\Phi})\Big]. \tag{7.21}$$

Here, $C_{\mathcal{D}}$ and $W_{\mathcal{D}}$ are the coercivity constant and limit-conformity measure defined by (2.26) *and* (2.28)*.*

Proof. The proof is similar to the case of Dirichlet boundary conditions (Theorem 7.12). By Lemma B.16 and B.20 we have $\|\mathbb{T}_{\mathfrak{T}}\boldsymbol{\Phi}(v)\|_{L^p(\partial\Omega)} \leq C_7\,\|\boldsymbol{\Phi}(v)\|_{\mathfrak{T},p}$ for some C_7 depending only on Ω, p and ϱ. Hence, using the triangular inequality,

$$\begin{aligned}\|\mathbb{T}_{\mathcal{D}}v\|_{L^p(\partial\Omega)} &\leq \omega^{\mathbb{T}}(\mathcal{D},\mathfrak{T},\boldsymbol{\Phi})\,\|v\|_{\mathcal{D}} + \|\mathbb{T}_{\mathfrak{T}}\boldsymbol{\Phi}(v)\|_{L^p(\partial\Omega)}\\ &\leq \omega^{\mathbb{T}}(\mathcal{D},\mathfrak{T},\boldsymbol{\Phi})\,\|v\|_{\mathcal{D}} + C_7\,\|\boldsymbol{\Phi}(v)\|_{\mathfrak{T},p}\\ &\leq \omega^{\mathbb{T}}(\mathcal{D},\mathfrak{T},\boldsymbol{\Phi})\,\|v\|_{\mathcal{D}} + C_7\,\|\boldsymbol{\Phi}\|_{\mathcal{D},\mathfrak{T}}\,\|v\|_{\mathcal{D}}.\end{aligned}$$

The proof of (7.20) is concluded by reproducing the same steps starting from $\|\Pi_{\mathcal{D}}v\|_{L^p(\partial\Omega)}$ and using Lemma B.20 to control $\|\Pi_{\mathfrak{T}}\boldsymbol{\Phi}(v)\|_{L^p(\Omega)}$ by $\|\boldsymbol{\Phi}(v)\|_{\mathfrak{T},p}$.
We turn to (7.21). Let $\boldsymbol{\varphi}\in W^{1,p'}(\Omega)^d$ and use the triangular inequality, the definition of $\omega^{\Pi}(\mathcal{D},\mathfrak{T},\boldsymbol{\Phi})$ and $\omega^{\mathbb{T}}(\mathcal{D},\mathfrak{T},\boldsymbol{\Phi})$ and (B.22) to obtain

$$\left|\int_\Omega\Big(\nabla_{\mathcal{D}}v(\boldsymbol{x})\cdot\boldsymbol{\varphi}(\boldsymbol{x}) + \Pi_{\mathcal{D}}v(\boldsymbol{x})\mathrm{div}\boldsymbol{\varphi}(\boldsymbol{x})\Big)\mathrm{d}\boldsymbol{x} - \int_{\partial\Omega}\mathbb{T}_{\mathcal{D}}v(\boldsymbol{x})\gamma_{\boldsymbol{n}}(\boldsymbol{\varphi})(\boldsymbol{x})\mathrm{d}s(\boldsymbol{x})\right|$$

$$
\begin{aligned}
&\le \left| \int_\Omega [\nabla_{\mathcal{D}} v(\boldsymbol{x}) - \overline{\nabla}_{\mathfrak{T}} \boldsymbol{\Phi}(v)(\boldsymbol{x})] \cdot \boldsymbol{\varphi}(\boldsymbol{x}) \right| + \|\mathrm{div}\boldsymbol{\varphi}\|_{L^{p'}(\Omega)} \omega^{\Pi}(\mathcal{D}, \mathfrak{T}, \boldsymbol{\Phi}) \|v\|_{\mathcal{D}} \\
&\quad + \|\gamma_{\boldsymbol{n}}(\boldsymbol{\varphi})\|_{L^{p'}(\partial\Omega)} \omega^{\mathbb{T}}(\mathcal{D}, \mathfrak{T}, \boldsymbol{\Phi}) \|v\|_{\mathcal{D}} \\
&\quad + \left| \int_\Omega \left(\overline{\nabla}_{\mathfrak{T}} \boldsymbol{\Phi}(v)(\boldsymbol{x}) \cdot \boldsymbol{\varphi}(\boldsymbol{x}) + \Pi_{\mathfrak{T}} \boldsymbol{\Phi}(v)(\boldsymbol{x}) \mathrm{div}\boldsymbol{\varphi}(\boldsymbol{x}) \right) \mathrm{d}\boldsymbol{x} \right. \\
&\quad \left. - \int_{\partial\Omega} \mathbb{T}_{\mathfrak{T}} v(\boldsymbol{x}) \gamma_{\boldsymbol{n}}(\boldsymbol{\varphi})(\boldsymbol{x}) \mathrm{d}s(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \right| \\
&\le \left| \int_\Omega [\nabla_{\mathcal{D}} v(\boldsymbol{x}) - \overline{\nabla}_{\mathfrak{T}} \boldsymbol{\Phi}(v)(\boldsymbol{x})] \cdot \boldsymbol{\varphi}(\boldsymbol{x}) \right| + \|\mathrm{div}\boldsymbol{\varphi}\|_{L^{p'}(\Omega)} \omega^{\Pi}(\mathcal{D}, \mathfrak{T}, \boldsymbol{\Phi}) \|v\|_{\mathcal{D}} \\
&\quad + \|\gamma_{\boldsymbol{n}}(\boldsymbol{\varphi})\|_{L^{p'}(\partial\Omega)} \omega^{\mathbb{T}}(\mathcal{D}, \mathfrak{T}, \boldsymbol{\Phi}) \|v\|_{\mathcal{D}} + C_8 \| \, |\nabla\boldsymbol{\varphi}| \, \|_{L^{p'}(\Omega)^d} |\boldsymbol{\Phi}(v)|_{\mathfrak{T},p} h_{\mathcal{M}},
\end{aligned}
$$

where C_8 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 7.19 (Properties of GDs controlled by polytopal toolboxes – Neumann 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 2.32, and let $(\mathfrak{T}_m)_{m\in\mathbb{N}}$ be a sequence of polytopal toolboxes in the sense of Definition 7.16. 1ssume 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 7.17, 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 2.33, limit-conforming in the sense of Definition 2.34, and compact in the sense of Definition 2.36.

Proof. The coercivity and limit-conformity follow from Estimates (7.20) and (7.21), from Lemma 2.38, and from the fact that $W^{1,p'}(\Omega)^d$ is dense in $W^{\mathrm{div},p',\partial}(\Omega)$ (see Lemma 2.46).
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

$$
\|\boldsymbol{\Phi}_m(v_m)\|_{\mathfrak{T}_m} \le \|\boldsymbol{\Phi}_m\|_{\mathcal{D}_m,\mathfrak{T}_m} \|v_m\|_{\mathcal{D}_m}.
$$

Hence, by Lemma B.22 and the definition of $\|\boldsymbol{\Phi}_m(v_m)\|_{\mathfrak{T}_m,p}$, the sequence $(\Pi_{\mathfrak{T}_m} \boldsymbol{\Phi}_m(v_m))_{m\in\mathbb{N}}$ converges up to a subsequence in $L^p(\Omega)$. Using Lemma

B.16 we also see that, up to a subsequence, $(\mathbb{T}_{\mathfrak{T}_m}\boldsymbol{\Phi}_m(v_m))_{m\in\mathbb{N}}$ converges weakly in $L^p(\partial\Omega)$ (recall that $p \in (1,\infty)$ throughout this chapter). The convergences of $\omega^{\Pi}(\mathcal{D}_m, \mathfrak{T}_m, \boldsymbol{\Phi}_m)$ and $\omega^{\mathbb{T}}(\mathcal{D}_m, \mathfrak{T}_m, \boldsymbol{\Phi}_m)$ then ensure that, along the same subsequence, $(\Pi_{\mathcal{D}_m} v_m)_{m\in\mathbb{N}}$ converges in $L^p(\Omega)$ and $(\mathbb{T}_{\mathcal{D}_m} v_m)_{m\in\mathbb{N}}$ converges weakly in $L^p(\partial\Omega)$, which completes the proof. ∎

Remark 7.20 (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 7.16, $\omega^{\mathbb{T}}$ in Definition 7.17 and Corollary 7.19, and Theorem 7.18 holds with $\omega^{\mathbb{T}}$ replaced with 0.

Remark 7.21 (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 (7.19) 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 7.18 in the case of Fourier boundary conditions is straightforward thanks to Lemma B.17.

7.2.4 Mixed boundary conditions

Finally, we cite the definition of a polytopal toolbox for mixed boundary conditions, as well as associated results without proofs (they can be established exactly as for Dirichlet and Neumann boundary conditions, using Lemma B.27 and B.28).

Definition 7.22 (Polytopal toolbox for mixed BCs). *Under Assumptions* (7.2) *and* (2.52)*, 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.65).
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 restruction 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 7.23 (Control of a GD by a polytopal toolbox – mixed BCs). *Under Assumptions* (7.2) *and* (2.52)*, let $\mathcal{D}$ be a GD in the sense of*

Definition 2.51, and let $\mathfrak{T}$ be a polytopal toolbox in the sense of Definition 7.22. A control of $\mathcal{D}$ by $\mathfrak{T}$ is a linear mapping $\boldsymbol{\Phi} : X_{\mathcal{D},\Omega,\Gamma_n} \longrightarrow X_{\mathfrak{T},\Omega,\Gamma_n}$. 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 7.24 (Estimates for a GD controlled by polytopal toolboxes – mixed BCs). *Under Assumptions (7.2) and (2.52), let $\mathcal{D}$ be a GD in the sense of Definition 2.51, and let $\mathfrak{T}$ be a polytopal toolbox in the sense of Definition 7.22. We take $\boldsymbol{\Phi}$ a control of $\mathcal{D}$ by $\mathfrak{T}$ in the sense of Definition 7.23, and $\varrho \geq \theta_{\mathfrak{T}} + \eta_{\mathfrak{T}}$ (see (7.8) and (7.9)).*
Then, there exists C_9 depending only on Ω, p and ϱ such that

$$C_{\mathcal{D}} \leq \max\left(\omega^{\Pi}(\mathcal{D},\mathfrak{T},\boldsymbol{\Phi}), \omega^{\mathbb{T}}(\mathcal{D},\mathfrak{T},\boldsymbol{\Phi})\right) + C_6 \|\boldsymbol{\Phi}\|_{\mathcal{D},\mathfrak{T}}$$

and, for all $\boldsymbol{\varphi} \in W^{1,p'}(\Omega)^d$,

$$W_{\mathcal{D}}(\boldsymbol{\varphi}) \leq \|\boldsymbol{\varphi}\|_{W^{1,p'}(\Omega)^d} \Big[C_6 h_{\mathcal{M}}(1 + \|\boldsymbol{\Phi}\|_{\mathcal{D},\mathfrak{T}}) + \omega^{\Pi}(\mathcal{D},\mathfrak{T},\boldsymbol{\Phi}) + \omega^{\nabla}(\mathcal{D},\mathfrak{T},\boldsymbol{\Phi}) + \omega^{\mathbb{T}}(\mathcal{D},\mathfrak{T},\boldsymbol{\Phi})\Big].$$

Here, $C_{\mathcal{D}}$ and $W_{\mathcal{D}}$ are the coercivity constant and limit-conformity measure defined by (2.53) and (2.57).

Corollary 7.25 (Properties of GDs controlled by polytopal toolboxes – mixed BCs). *Under Assumptions (7.2) and (2.52), let $(\mathcal{D}_m)_{m\in\mathbb{N}}$ be a sequence of GDs in the sense of Definition 2.51, and let $(\mathfrak{T}_m)_{m\in\mathbb{N}}$ be a sequence of polytopal toolboxes in the sense of Definition 7.22. 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 7.23, 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 2.52, limit-conforming in the sense of Definition 2.54, and compact in the sense of Definition 2.56.

The only (slightly) non-trivial adaptation of the preceding proofs to establish this corollary is the density of smooth functions in $W^{\mathrm{div},p',\Gamma_n}(\Omega)$, 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 2.46, 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, explicitly or implicitly, on reconstructions of functions – or gradients – from discrete degrees of freedoms. These reconstructions are designed to match certain simple functions (e.g. constant, or affine) – or their gradients – when the degrees of freedom 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 7.26 ($\mathbb{P}_0$-exact function reconstruction). *Let I be a finite set, K be a bounded set 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

$$\text{for a.e. } \boldsymbol{x} \in K\,,\ \sum_{i\in I} \pi_K^i(\boldsymbol{x}) = 1. \tag{7.22}$$

The norm of π_K is defined by (setting $|K|^{-\frac{1}{p}} = 1$ if $p = +\infty$)

$$\|\pi_K\|_p = |K|^{-\frac{1}{p}} \left\| \sum_{i\in I} |\pi_K^i| \right\|_{L^p(K)}. \tag{7.23}$$

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 (7.22) 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 7.27 (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 7.28 (Interpolation estimate for $\mathbb{P}_0$-exact function reconstruction). *Let I be a finite set, K be a bounded set 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 $v = (\varphi(\boldsymbol{x}_i))_{i\in I}$,

$$\|\pi_K v - \varphi\|_{L^p(K)} \leq \left(1 + \max_{i\in I} \frac{\operatorname{dist}(\boldsymbol{x}_i, K)}{\operatorname{diam}(K)}\right) \|\pi_K\|_p \, |K|^{\frac{1}{p}} \operatorname{diam}(K) \, \|\varphi\|_{W^{1,\infty}(\mathbb{R}^d)}.$$

Proof. For a.e. $\boldsymbol{x} \in K$, using (7.22) 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 $\boldsymbol{x} \in K$,

$$\begin{aligned} |\varphi(\boldsymbol{x}_i) - \varphi(\boldsymbol{x})| &\leq |\boldsymbol{x}_i - \boldsymbol{x}| \, \|\varphi\|_{W^{1,\infty}(\mathbb{R}^d)} \\ &\leq (\operatorname{diam}(K) + \operatorname{dist}(\boldsymbol{x}_i, K)) \, \|\varphi\|_{W^{1,\infty}(\mathbb{R}^d)}. \end{aligned}$$

Hence, for a.e. $\boldsymbol{x} \in K$,

$$\begin{aligned} |\pi_K v(\boldsymbol{x}) - \varphi(\boldsymbol{x})| &= \left|\sum_{i\in I} \pi_K^i(\boldsymbol{x})(v_i - \varphi(\boldsymbol{x}))\right| \qquad (7.24) \\ &\leq \max_{i\in I} |\varphi(\boldsymbol{x}_i) - \varphi(\boldsymbol{x})| \sum_{i\in I} |\pi_K^i(\boldsymbol{x})| \\ &\leq \left(1 + \max_{i\in I} \frac{\operatorname{dist}(\boldsymbol{x}_i, K)}{\operatorname{diam}(K)}\right) \operatorname{diam}(K) \, \|\varphi\|_{W^{1,\infty}(\mathbb{R}^d)} \sum_{i\in I} |\pi_K^i(\boldsymbol{x})|. \end{aligned}$$

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 7.29 ($\mathbb{P}_1$-exact gradient reconstruction). *Let I be a finite set, K be a bounded set 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:

$$\text{for any affine } A : \mathbb{R}^d \to \mathbb{R} \text{ and a.e. } \boldsymbol{x} \in K\,,\ \sum_{i\in I} A(\boldsymbol{x}_i)\mathcal{G}_K^i(\boldsymbol{x}) = \nabla A. \quad (7.25)$$

The norm of $\mathcal{G}_K$ is defined by (setting $|K|^{-\frac{1}{p}} = 1$ if $p = +\infty$)

$$\|\mathcal{G}_K\|_p = \operatorname{diam}(K)|K|^{-\frac{1}{p}} \left\| \sum_{i\in I} |\mathcal{G}_K^i| \right\|_{L^p(K)}. \quad (7.26)$$

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 (7.25) that

$$\text{for all affine function } A, \text{ if } v = (A(\boldsymbol{x}_i))_{i\in I} \text{ then } \mathcal{G}_K v = \nabla A. \quad (7.27)$$

This is the $\mathbb{P}_1$-exactness of the gradient reconstruction $\mathcal{G}_K$.

Example 7.30 ($\mathcal{G}_K$ 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 7.27, we let $\mathcal{G}_K^\sigma = \nabla\pi_K^\sigma \in L^p(K)^d$. As proved in Lemma 9.1, 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 $S = (\overline{\boldsymbol{x}}_\sigma)_{\sigma\in\mathcal{F}_K}$.
This property that $\mathcal{G}_K^i$ is the gradient of π_K^i, which also holds for conforming finite elements, is a very specific one. It is not satisfied by a number of other schemes such as mixed finite elements, hybrid mimetic mixed methods, etc. (see Chapters 10, 11, 12 and 13), or if performing mass-lumping of conforming and non-conforming finite elements (see Example 7.44). Hence, for many methods, a full description cannot be given by just describing the elementary functions π_K^i, but also requires a separate definition of the local gradients $\mathcal{G}_K^i$.

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 set 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 $v = (\varphi(\boldsymbol{x}_i))_{i\in I}$,

$$\|\mathcal{G}_K v - \nabla\varphi\|_{L^p(K)^d} \le \left(1 + \frac{1}{2}\|\mathcal{G}_K\|_p \left[1 + \max_{i\in I} \frac{\operatorname{dist}(\boldsymbol{x}_i, K)}{\operatorname{diam}(K)}\right]^2\right) \times |K|^{\frac{1}{p}} \operatorname{diam}(K) \|\varphi\|_{W^{2,\infty}(\mathbb{R}^d)}. \tag{7.28}$$

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 (7.27) 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 $\|\varphi\|_{W^{2,\infty}(\mathbb{R}^d)}$, we write

$$\begin{aligned}\|\mathcal{G}_K\xi - \nabla\varphi\|_{L^p(K)^d} &= \|\nabla\varphi(\boldsymbol{x}_K) - \nabla\varphi\|_{L^p(K)^d} \\ &\le |K|^{\frac{1}{p}} \|\nabla\varphi(\boldsymbol{x}_K) - \nabla\varphi\|_{L^\infty(K)^d} \\ &\le |K|^{\frac{1}{p}} \operatorname{diam}(K) \|\varphi\|_{W^{2,\infty}(\mathbb{R}^d)}. \end{aligned} \tag{7.29}$$

For any $i \in I$ we have, by Taylor's expansion,

$$\begin{aligned} v_i - \xi_i &= \varphi(\boldsymbol{x}_i) - A(\boldsymbol{x}_i) \\ &= \varphi(\boldsymbol{x}_i) - \varphi(\boldsymbol{x}_K) - \nabla\varphi(\boldsymbol{x}_K)\cdot(\boldsymbol{x}_i - \boldsymbol{x}_K) \\ &= \int_0^1 (1-s) D^2\varphi(\boldsymbol{x}_K + s(\boldsymbol{x}_i - \boldsymbol{x}_K))(\boldsymbol{x}_i - \boldsymbol{x}_K)\cdot(\boldsymbol{x}_i - \boldsymbol{x}_K)\mathrm{d}s. \end{aligned} \tag{7.30}$$

Using $|\boldsymbol{x}_i - \boldsymbol{x}_K| \le \operatorname{diam}(K) + \operatorname{dist}(\boldsymbol{x}_i, K)$ yields

$$|v_i - \xi_i| \le \frac{1}{2}[\operatorname{diam}(K) + \operatorname{dist}(\boldsymbol{x}_i, K)]^2 \|\varphi\|_{W^{2,\infty}(\mathbb{R}^d)}. \tag{7.31}$$

Hence, for a.e. $\boldsymbol{x} \in K$,

$$\begin{aligned} |\mathcal{G}_K v(\boldsymbol{x}) - \mathcal{G}_K\xi(\boldsymbol{x})| &= \left|\sum_{i\in I} (v_i - \xi_i)\mathcal{G}_K^i(\boldsymbol{x})\right| \\ &\le \frac{1}{2}\left[\operatorname{diam}(K) + \max_{i\in I}\operatorname{dist}(\boldsymbol{x}_i, K)\right]^2 \|\varphi\|_{W^{2,\infty}(\mathbb{R}^d)} \sum_{i\in I} |\mathcal{G}_K^i(\boldsymbol{x})|. \end{aligned}$$

Taking the $L^p(K)$ norm over $\boldsymbol{x}$ and recalling the definition of $\|\mathcal{G}_K\|_p$ leads to

$$\begin{aligned}
&\|\mathcal{G}_K v - \mathcal{G}\xi\|_{L^p(K)^d} \\
&\qquad \le \frac{|K|^{\frac{1}{p}}\|\mathcal{G}_K\|_p}{\operatorname{diam}(K)}\frac{1}{2}\left[\operatorname{diam}(K) + \max_{i\in I}\operatorname{dist}(\boldsymbol{x}_i, K)\right]^2 \|\varphi\|_{W^{2,\infty}(\mathbb{R}^d)} \\
&\qquad = \frac{1}{2}\|\mathcal{G}_K\|_p\left[1 + \max_{i\in I}\frac{\operatorname{dist}(\boldsymbol{x}_i, K)}{\operatorname{diam}(K)}\right]^2 |K|^{\frac{1}{p}}\operatorname{diam}(K)\,\|\varphi\|_{W^{2,\infty}(\mathbb{R}^d)}\,.
\end{aligned}$$

Combined with (7.29) and a triangle inequality, this completes the proof of the lemma if $\varphi \in C_b^2(\mathbb{R}^d)$. Since any function $\varphi \in W^{2,\infty}(\mathbb{R}^d)$ can be approximated (using convolution) by functions $\varphi_n \in C_b^2(\mathbb{R}^d)$ such that $\varphi_n \to \varphi$ and $\nabla\varphi_n \to \nabla\varphi$ uniformly on compact sets, and $\|\varphi_n\|_{W^{2,\infty}(\mathbb{R}^d)} \le \|\varphi\|_{W^{2,\infty}(\mathbb{R}^d)}$, the proof follows by passing to the limit $n \to \infty$ in (7.28) written for φ_n. ∎

Remark 7.32. For all functions and gradient reconstructions considered in Chapters 8–13, 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 8–13, 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 $\pi_K^i \ge 0$ 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 8–13. 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 degrees of freedom (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 degrees of freedom (DOFs), where I is partitioned into I_Ω (interior geometrical entities attached to the DOFs) and I_∂ (boundary geometrical entities attached to the DOFs), such that*

$$X_{\mathcal{D},0} = \{v = (v_i)_{i\in I} \,:\, v_i \in \mathbb{R} \text{ for all } i \in I\,,\ v_i = 0 \text{ if } i \in I_\partial\}. \tag{7.32}$$

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 7.26) such that

$$\begin{aligned} &\forall v \in X_{\mathcal{D},0}\,, \quad \text{for a.e. } x \in K\,, \\ &\Pi_{\mathcal{D}} v(\boldsymbol{x}) = \pi_K[(v_i)_{i\in I_K}](\boldsymbol{x}) = \sum_{i\in I_K} v_i \pi_K^i(\boldsymbol{x}), \end{aligned} \tag{7.33}$$

   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 7.29) such that

$$\begin{aligned} &\forall v \in X_{\mathcal{D},0}\,, \quad \text{for a.e. } x \in K\,, \\ &\nabla_{\mathcal{D}} v(\boldsymbol{x}) = \mathcal{G}_K[(v_i)_{i\in I_K}](\boldsymbol{x}) = \sum_{i\in I_K} v_i \mathcal{G}_K^i(\boldsymbol{x}). \end{aligned} \tag{7.34}$$

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 $h_{\mathcal{M}} = \max_{K\in\mathcal{M}} \mathrm{diam}(K)$, and the LLE regularity of $\mathcal{D}$ is defined by

$$\mathrm{reg}_{\mathrm{LLE}}(\mathcal{D}) = \max_{K\in\mathcal{M}} \left(\|\pi_K\|_p + \|\mathcal{G}_K\|_p + \max_{i\in I_K} \frac{\mathrm{dist}(\boldsymbol{x}_i, K)}{\mathrm{diam}(K)} \right). \tag{7.35}$$

Example 7.34 (LLE GD interpretation of the non-conforming $\mathbb{P}_1$ gradient discretisation)
Example 7.7 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 DOFs are the faces of the mesh (that is, $I = \mathcal{F}$), the elementary basis functions are the π_K^σ described in Remark 7.27, and the local gradients $\mathcal{G}_K^\sigma$ are given in Remark 7.30.
It is proved in Lemma 9.1 that, under standard regularity assumptions on $\mathfrak{T}$, $\mathrm{reg}_{\mathrm{LLE}}(\mathcal{D})$ is bounded. Used in Proposition 7.36 below, this bound yields the consistency of non-conforming $\mathbb{P}_1$ gradient discretisations.

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 8–13, 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 DOFs may be located at the same point $\boldsymbol{x}_i = \boldsymbol{x}_j$. This happens for instance in the case of the MPFA-0 scheme, see Remark 11.1 in p.305.

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 2.10), and in particular mass-lumped GDs (see Section 7.3.5 below).

Remark 7.35 (Generalisation of $\mathrm{reg}_{\mathrm{LLE}}$*)*
The term $\mathrm{diam}(K)$ in $\mathrm{reg}_{\mathrm{LLE}}(\mathcal{D})$ could be replaced with any quantity $\omega_K > 0$, the requirement to prove Proposition 7.36 below being that $\max_{K\in\mathcal{M}_m} \omega_K \to 0$ as $m \to \infty$.

Proposition 7.36 (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 2.13 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 7.28 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 (7.33) of $\Pi_{\mathcal{D}_m}$, Lemma 7.28 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

$$\begin{aligned}&\|\Pi_{\mathcal{D}_m} v^m - \varphi\|_{L^p(\Omega)} \\ &\qquad \le (1 + \mathrm{reg}_{\mathrm{LLE}}(\mathcal{D}_m)) \, \mathrm{reg}_{\mathrm{LLE}}(\mathcal{D}_m) |\Omega|^{\frac{1}{p}} h_{\mathcal{M}_m} \, \|\varphi\|_{W^{1,\infty}(\mathbb{R}^d)} . \quad (7.36)\end{aligned}$$

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 (7.34) of $\nabla_{\mathcal{D}_m}$, to Lemma 7.31 and to the definition of $\mathrm{reg}_{\mathrm{LLE}}(\mathcal{D}_m)$,

$$\begin{aligned}\|\nabla_{\mathcal{D}_m} v^m - \nabla\varphi\|_{L^p(K)^d} &= \left\|\mathcal{G}_{K,m}[(v_i^m)_{i\in I_K^m}] - \nabla\varphi\right\|_{L^p(K)^d} \\ &\le \left(1 + \frac{1}{2}\,\mathrm{reg}_{\mathrm{LLE}}(\mathcal{D}_m)[1 + \mathrm{reg}_{\mathrm{LLE}}(\mathcal{D}_m)]^2\right) |K|^{\frac{1}{p}} \mathrm{diam}(K) \, \|\varphi\|_{W^{2,\infty}(\mathbb{R}^d)} .\end{aligned}$$

Again, raising to the power p, sum over $K \in \mathcal{M}_m$ and take the power $1/p$ to obtain

$$\begin{aligned}&\|\nabla_{\mathcal{D}_m} v^m - \nabla\varphi\|_{L^p(\Omega)^d} \\ &\quad \le \left(1 + \frac{1}{2}\,\mathrm{reg}_{\mathrm{LLE}}(\mathcal{D}_m)[1 + \mathrm{reg}_{\mathrm{LLE}}(\mathcal{D}_m)]^2\right) |\Omega|^{\frac{1}{p}} h_{\mathcal{M}_m} \, \|\varphi\|_{W^{2,\infty}(\mathbb{R}^d)} . \quad (7.37)\end{aligned}$$

Since $(\mathrm{reg}_{\mathrm{LLE}}(\mathcal{D}_m))_{m\in\mathbb{N}}$ is bounded and $S_{\mathcal{D}_m}(\varphi) \le \|\Pi_{\mathcal{D}_m} v^m - \varphi\|_{L^p(\Omega)} + \|\nabla_{\mathcal{D}_m} v^m - \nabla\varphi\|_{L^p(\Omega)^d}$, we infer from (7.36) and (7.37) the existence of C_{10} not depending on m or φ such that

$$S_{\mathcal{D}_m}(\varphi) \le C_{10} h_{\mathcal{M}_m} \, \|\varphi\|_{W^{2,\infty}(\mathbb{R}^d)} . \quad (7.38)$$

Thus, $S_{\mathcal{D}_m}(\varphi) \to 0$ as $m \to \infty$ and the proof is complete. ∎

Remark 7.37 (Order of approximation)
The order 1 approximation (7.38) is expected for reconstructions that are only linearly exact.
Assume that, for each $K \in \mathcal{M}$, the family $(\pi_K^i)_{i\in I_K}$ is a $\mathbb{P}_{(k-1)}$-exact function reconstruction, that is

$$\forall \boldsymbol{j} = (j_1, \ldots, j_d) \in \mathbb{N}^d \text{ such that } j_1 + \ldots + j_d \le k-1, \ \forall \boldsymbol{x} \in K ,$$

$$\sum_{i\in I_K} \boldsymbol{x}_i^{\boldsymbol{j}} \pi_K^i(\boldsymbol{x}) = \boldsymbol{x}^{\boldsymbol{j}}.$$

Here, $\boldsymbol{x}^{\boldsymbol{j}} = x_1^{j_1}\cdots x_d^{j_d}$ if $\boldsymbol{x} = (x_1,\ldots,x_d)$. Similarly, suppose that the family $(\mathcal{G}_K^i)_{i\in I_K}$ is a $\mathbb{P}_k$-exact gradient reconstruction, that is:

$$\forall \boldsymbol{j} = (j_1,\ldots,j_d) \in \mathbb{N}^d \text{ such that } j_1+\ldots+j_d \le k,\ \forall \boldsymbol{x}\in K\,,$$
$$\sum_{i\in I_K} \boldsymbol{x}_i^{\boldsymbol{j}} \mathcal{G}_K^i(\boldsymbol{x}) = \nabla \boldsymbol{x}^{\boldsymbol{j}}.$$

Then, in (7.24) in the proof of Lemma 7.28, express $v_i - \varphi(\boldsymbol{x}) = \varphi(\boldsymbol{x}_i) - \varphi(\boldsymbol{x})$ using the Taylor expansion of φ of order $k-1$ to see that, under boundedness assumption on $\mathrm{reg}_{\mathrm{LLE}}(\mathcal{D})$,

$$\|\Pi_{\mathcal{D}} v - \varphi\|_{L^p(K)} = \mathcal{O}(h_{\mathcal{M}}^k |K|^{\frac{1}{p}} \|\varphi\|_{W^{k,\infty}(\mathbb{R}^d)}). \tag{7.39}$$

Similarly, using in the proof of Lemma 7.31 the k-th order (instead of the first order) Taylor expansion of φ shows that (7.31) becomes $|v_i - \xi_i| = \mathcal{O}(h_{\mathcal{M}}^{k+1} \|\varphi\|_{W^{k+1,\infty}(\mathbb{R}^d)})$, and thus that the final estimate in Lemma 7.31 is

$$\|\mathcal{G}_K v - \nabla\varphi\|_{L^p(K)^d} = \mathcal{O}(h_{\mathcal{M}}^k |K|^{\frac{1}{p}} \|\varphi\|_{W^{k+1,\infty}(\mathbb{R}^d)}). \tag{7.40}$$

Raising (7.39) and (7.40) to the power p and summing over $K\in\mathcal{M}$ yields

$$\forall \varphi \in W_0^{1,p}(\Omega)\cap W^{k+1,\infty}(\mathbb{R}^d)\,,\ S_{\mathcal{D}}(\varphi) \le C h_{\mathcal{M}}^k \|\varphi\|_{W^{k+1,\infty}(\mathbb{R}^d)}\,, \tag{7.41}$$

where C depends only on an upper bound of $\mathrm{reg}_{\mathrm{LLE}}(\mathcal{D})$. This latter estimate is particularly useful for problems in which $S_{\mathcal{D}}$ participates in the error estimates established for the GDM (cf. Theorems 3.2 and 3.28 for example).
The relation (7.41) is not optimal in the sense that the regularity assumptions on φ can often be relaxed. In particular, we only need the $W^{k+1,\infty}$ regularity around each $K\in\mathcal{M}$, and this regularity can often be further relaxed to H^{k+1}. See Section A.1 in Appendix A for examples of more efficient estimates.

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 DOF 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:

$$\forall K\in\mathcal{M} \text{ such that } i\in I_K\,,\ (\pi^i)_{|K} = \pi_K^i \text{ and } (\mathcal{G}^i)_{|K} = \mathcal{G}_K^i\,,$$
$$\forall K\in\mathcal{M} \text{ such that } i\notin I_K\,,\ (\pi^i)_{|K} = 0 \text{ and } (\mathcal{G}^i)_{|K} = 0.$$

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 DOF v_i of a generic $v \in X_{\mathcal{D},0}$.

This construction is illustrated in Figure 7.2, 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., $\boldsymbol{s} \notin \mathcal{V}_{K'}$).

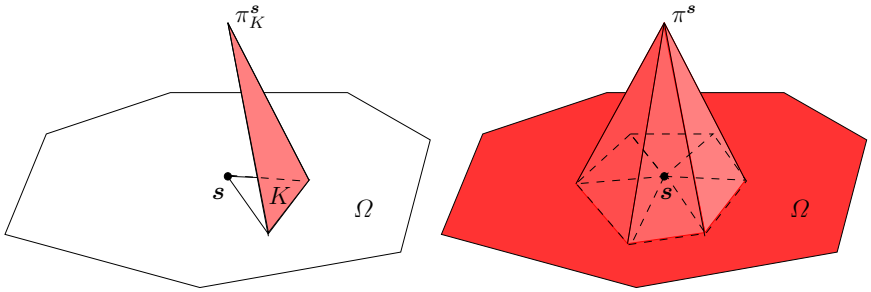


Fig. 7.2. Elementary basis function (left) and global basis functions (right) for $\mathbb{P}_1$ finite elements

Let us now consider the problem (3.3) and its GS approximation (3.4). As seen in Section 3.1.1, this scheme can be re-cast as a linear system $AU = B$ with

$$u = \sum_{i \in I_\Omega} U_i v^{(i)},$$
$$A_{ij} = \int_\Omega \Lambda(\boldsymbol{x}) \nabla_{\mathcal{D}} v^{(j)}(\boldsymbol{x}) \cdot \nabla_{\mathcal{D}} v^{(i)}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x},$$
$$B_i = \int_\Omega f(\boldsymbol{x}) \Pi_{\mathcal{D}} v^{(i)}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} - \int_\Omega \boldsymbol{F}(\boldsymbol{x}) \cdot \nabla_{\mathcal{D}} v^{(i)}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}.$$

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

$$\forall i \in I_K,\ \forall j \in I_K,$$
$$A_{ij}^K = \int_K \Lambda(\boldsymbol{x}) \mathcal{G}_K^j(\boldsymbol{x}) \cdot \mathcal{G}_K^i(\boldsymbol{x}) \mathrm{d}\boldsymbol{x},$$

$$B_i^K = \int_K f(\boldsymbol{x})\pi_K^i(\boldsymbol{x})\mathrm{d}\boldsymbol{x} - \int_K \boldsymbol{F}(\boldsymbol{x})\cdot\mathcal{G}_K^i(\boldsymbol{x})\mathrm{d}\boldsymbol{x}.$$

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 elimination of degrees of freedom

The construction of a numerical scheme often requires several interpolation points, the approximation points S of an LLE GD, corresponding to as many DOFs of the scheme. The higher the number of DOFs, 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 DOFs through barycentric combinations. This consists in replacing (and thus, eliminating) certain DOFs by averages of other DOFs.
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 DOFs of an LLE GD, and all other degrees of freedom (associated with $I\backslash I^{\mathrm{BA}}$) are eliminated from the GD space by being expressed as local barycentric combinations of the degrees of freedom corresponding to I^{BA}.

Definition 7.38 (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 $\mathcal{D}^{\mathrm{BA}}$ is a barycentric condensation of $\mathcal{D}$ if there exists a strict subset $I^{\mathrm{BA}} \subset I$ and, for all $i \in I\backslash I^{\mathrm{BA}}$, a set $H_i \subset I^{\mathrm{BA}}$ and real numbers $(\beta_j^i)_{j\in H_i}$ satisfying*

$$\sum_{j\in H_i} \beta_j^i = 1 \quad \text{and} \quad \sum_{j\in H_i} \beta_j^i \boldsymbol{x}_j = \boldsymbol{x}_i, \tag{7.42}$$

such that

- $I_\partial \subset I^{\mathrm{BA}}$,
- $X_{\mathcal{D}^{\mathrm{BA}},0} = \{u = (u_i)_{i\in I^{\mathrm{BA}}} : u_i \in \mathbb{R} \text{ for all } i \in I^{\mathrm{BA}},\ u_i = 0 \text{ for all } i \in I_\partial\}$,
- *The function and gradient reconstructions $\Pi_{\mathcal{D}^{\mathrm{BA}}}$ and $\nabla_{\mathcal{D}^{\mathrm{BA}}}$ are given by:*

$$\forall v \in X_{\mathcal{D}^{\mathrm{BA}},0}\,, \quad \Pi_{\mathcal{D}^{\mathrm{BA}}} v = \Pi_{\mathcal{D}} \widetilde{v} \quad \text{and} \quad \nabla_{\mathcal{D}^{\mathrm{BA}}} v = \nabla_{\mathcal{D}} \widetilde{v},$$

where $\widetilde{v} \in X_{\mathcal{D},0}$ is defined by

$$\forall i \in I\,,\ \widetilde{v}_i = \begin{cases} v_i & \text{if } i \in I^{\mathrm{BA}}, \\ \displaystyle\sum_{j\in H_i} \beta_j^i v_j & \text{if } i \in I \setminus I^{\mathrm{BA}}. \end{cases} \tag{7.43}$$

(recall that $X_{\mathcal{D},0} = \{u = (u_i)_{i\in I} : u_i \in \mathbb{R}$ *for all* $i \in I$, $u_i = 0$ *for all* $i \in I_\partial\}$*, and notice that* $\widetilde{v}$ *indeed belongs to this space since* $I_\partial \subset I^{\mathrm{BA}}$ *and* $v_i = 0$ *if* $i \in I_\partial$*.)*

The regularity of the barycentric condensation $\mathcal{D}^{\mathrm{BA}}$ *is*

$$\mathrm{reg}_{\mathrm{BA}}(\mathcal{D}^{\mathrm{BA}}) = \max_{i\in I\setminus I^{\mathrm{BA}}} \left(\sum_{j\in H_i} |\beta_j^i| + \max_{K\in\mathcal{M}\,|\,i\in I_K} \max_{j\in H_i} \frac{\mathrm{dist}(\boldsymbol{x}_j, \boldsymbol{x}_i)}{\mathrm{diam}(K)} \right).$$

It is clear that $\mathcal{D}^{\mathrm{BA}}$ defined above is a GD. Indeed, if $\nabla_{\mathcal{D}^{\mathrm{BA}}} v = 0$ 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 $i \in I^{\mathrm{BA}}$, and thus that $\|\nabla_{\mathcal{D}^{\mathrm{BA}}}\cdot\|_{L^p(\Omega)^d}$ is a norm on $X_{\mathcal{D}^{\mathrm{BA}},0}$.

Note that $\mathrm{reg}_{\mathrm{BA}}(\mathcal{D}^{\mathrm{BA}})$ is always greater than or equal to 1 (take $i \in I \setminus I^{\mathrm{BA}}$ and write $1 = \sum_{j\in H_i} \beta_j^i \le \sum_{j\in H_i} |\beta_j^i|$). Let us, for a brief moment, confuse a DOF 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 $\mathrm{reg}_{\mathrm{BA}}(\mathcal{D}^{\mathrm{BA}})$ consists in requiring that, if $i \in I \setminus I^{\mathrm{BA}}$ 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 $\mathcal{O}(\mathrm{diam}(K))$ of any $j \in H_i$ used to eliminate i. This ensures that, after barycentric elimination, π_K and $\mathcal{G}_K$ are still computed using only degrees of freedom in a neighborhood of K.

The operation, performed to build a barycentric condensation of a given LLE GD and consisting in replacing some degrees of freedom with combinations of others, is called a *barycentric elimination*. These combinations are linearly exact thanks to (7.42). The LLE property is therefore preserved in the process, as formally stated in the lemma below.

Remark 7.39 (Barycentric elimination vs. static condensation)
A barycentric elimination 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 8.17 and 12.7.
A barycentric elimination occurs before a scheme is even written, and can also be performed for non-linear schemes; the replacement of DOFs in a barycentric elimination modifies the space and operators of the scheme independently of the model to which it is applied.

Lemma 7.40 (Barycentric elimination preserves the LLE property). *Let* $\mathcal{D}$ *be an LLE GD in the sense of Definition 7.33, and let* $\mathcal{D}^{\mathrm{BA}}$ *be a barycentric condensation of* $\mathcal{D}$ *in the sense of Definition 7.38. Then* $\mathcal{D}^{\mathrm{BA}}$ *is an LLE GD on the same mesh as* $\mathcal{D}$*, and*

$$\mathrm{reg}_{\mathrm{LLE}}(\mathcal{D}^{\mathrm{BA}}) \le \mathrm{reg}_{\mathrm{BA}}(\mathcal{D}^{\mathrm{BA}})\,\mathrm{reg}_{\mathrm{LLE}}(\mathcal{D}) + \mathrm{reg}_{\mathrm{BA}}(\mathcal{D}^{\mathrm{BA}}). \tag{7.44}$$

Proof. Let $\mathcal{M}$ be the mesh corresponding to $\mathcal{D}$, and let $K \in \mathcal{M}$. Take $v \in X_{\mathcal{D}^{\mathrm{BA}},0}$ and let $\widetilde{v} \in X_{\mathcal{D},0}$ be defined by (7.43). For any $K \in \mathcal{M}$, the values $(\widetilde{v}_i)_{i\in I_K}$ are computed as linear combinations of $(v_i)_{i\in I_K^{\mathrm{BA}}}$, with

$$I_K^{\mathrm{BA}} = (I_K \cap I^{\mathrm{BA}}) \cup \bigcup_{i\in I_K\setminus I^{\mathrm{BA}}} H_i. \tag{7.45}$$

By (7.33) in Definition 7.33 and the definition (7.43) of $\widetilde{v}$, for $K \in \mathcal{M}$ and a.e. $\boldsymbol{x} \in K$,

$$\begin{aligned} &\mathit{\Pi}_{\mathcal{D}^{\mathrm{BA}}} v(\boldsymbol{x}) = \mathit{\Pi}_{\mathcal{D}} \widetilde{v}(\boldsymbol{x}) = \sum_{i\in I_K} \widetilde{v}_i \pi_K^i(\boldsymbol{x}) \\ &= \sum_{i\in I_K\cap I^{\mathrm{BA}}} v_i \pi_K^i(\boldsymbol{x}) + \sum_{i\in I_K\setminus I^{\mathrm{BA}}} \Big(\sum_{j\in H_i} \beta_j^i v_j\Big) \pi_K^i(\boldsymbol{x}) = \sum_{j\in I_K^{\mathrm{BA}}} v_j \widetilde{\pi}_K^j(\boldsymbol{x}), \end{aligned} \tag{7.46}$$

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 (7.42) and (7.22) yields, for a.e. $\boldsymbol{x} \in K$,

$$\begin{aligned} \sum_{j\in I_K^{\mathrm{BA}}} \widetilde{\pi}_K^j(\boldsymbol{x}) &= \sum_{j\in I_K\cap I^{\mathrm{BA}}} \pi_K^j(\boldsymbol{x}) + \sum_{j\in I_K^{\mathrm{BA}}} \sum_{i\in I_K\setminus I^{\mathrm{BA}} \,|\, j\in H_i} \beta_j^i \pi_K^i(\boldsymbol{x}) \\ &= \sum_{j\in I_K\cap I^{\mathrm{BA}}} \pi_K^j(\boldsymbol{x}) + \sum_{i\in I_K\setminus I^{\mathrm{BA}}} \pi_K^i(\boldsymbol{x}) \sum_{j\in H_i} \beta_j^i \\ &= \sum_{i\in I_K\cap I^{\mathrm{BA}}} \pi_K^i(\boldsymbol{x}) + \sum_{i\in I_K\setminus I^{\mathrm{BA}}} \pi_K^i(\boldsymbol{x}) \\ &= \sum_{i\in I_K} \pi_K^i(\boldsymbol{x}) = 1. \end{aligned} \tag{7.47}$$

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 (7.46), $\mathit{\Pi}_{\mathcal{D}^{\mathrm{BA}}} v$ has the required form (7.33).
In a similar way as above, write

$$\begin{aligned} &\nabla_{\mathcal{D}^{\mathrm{BA}}} v(\boldsymbol{x}) = \nabla_{\mathcal{D}} \widetilde{v}(\boldsymbol{x}) = \sum_{i\in I_K} \widetilde{v}_i \mathcal{G}_K^i(\boldsymbol{x}) \\ &\quad = \sum_{i\in I_K\cap I^{\mathrm{BA}}} v_i \mathcal{G}_K^i(\boldsymbol{x}) + \sum_{i\in I_K\setminus I^{\mathrm{BA}}} \left(\sum_{j\in H_i} \beta_j^i v_j\right) \mathcal{G}_K^i(\boldsymbol{x}) = \sum_{j\in I_K^{\mathrm{BA}}} v_j \widetilde{\mathcal{G}}_K^j(\boldsymbol{x}), \end{aligned}$$

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

$$\begin{aligned} &\sum_{j \in I_K^{\mathrm{BA}}} A(\boldsymbol{x}_j) \widetilde{\mathcal{G}}_K^j(\boldsymbol{x}) \\ &\quad = \sum_{j \in I_K \cap I^{\mathrm{BA}}} A(\boldsymbol{x}_j) \mathcal{G}_K^j(\boldsymbol{x}) + \sum_{j \in I_K^{\mathrm{BA}}} A(\boldsymbol{x}_j) \sum_{i \in I_K \setminus I^{\mathrm{BA}} \,|\, j \in H_i} \beta_j^i \mathcal{G}_K^i(\boldsymbol{x}) \\ &\quad = \sum_{j \in I_K \cap I^{\mathrm{BA}}} A(\boldsymbol{x}_j) \mathcal{G}_K^j(\boldsymbol{x}) + \sum_{i \in I_K \setminus I^{\mathrm{BA}}} \left(\sum_{j \in H_i} A(\boldsymbol{x}_j) \beta_j^i \right) \mathcal{G}_K^i(\boldsymbol{x}). \end{aligned} \tag{7.48}$$

Since A is affine we have $A(\boldsymbol{x}) = A(\boldsymbol{x}_i) + \nabla A \cdot (\boldsymbol{x} - \boldsymbol{x}_i)$. Hence, (7.42) 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 (7.48) and using (7.25), 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 (7.34). 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 (7.47) but using absolute values and inequalities, we see that for any $K \in \mathcal{M}$ and a.e. $\boldsymbol{x} \in K$

$$\begin{aligned} \sum_{j \in I_K^{\mathrm{BA}}} |\widetilde{\pi}_K^j(\boldsymbol{x})| &\le \sum_{i \in I_K \cap I^{\mathrm{BA}}} |\pi_K^i(\boldsymbol{x})| + \sum_{i \in I_K \setminus I^{\mathrm{BA}}} |\pi_K^i(\boldsymbol{x})| \sum_{j \in H_i} |\beta_j^i| \\ &\le \mathrm{reg}_{\mathrm{BA}}(\mathcal{D}^{\mathrm{BA}}) \sum_{i \in I_K} |\pi_K^i(\boldsymbol{x})|. \end{aligned} \tag{7.49}$$

Take the $L^p(K)$ norm, multiply by $|K|^{-\frac{1}{p}}$ and recall the definition (7.23) of the norm of $\mathbb{P}_0$-exact function reconstructions to obtain

$$\|\widetilde{\pi}_K\|_p \le \mathrm{reg}_{\mathrm{BA}}(\mathcal{D}^{\mathrm{BA}})\,\|\pi_K\|_p\,. \tag{7.50}$$

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 (7.49) still holds with "$\mathcal{G}$" instead of "π" so that, taking the $L^p(K)$ norm and multiplying by $\mathrm{diam}(K)|K|^{-\frac{1}{p}}$,

$$\left\|\widetilde{\mathcal{G}}_K\right\|_p \le \mathrm{reg}_{\mathrm{BA}}(\mathcal{D}^{\mathrm{BA}})\,\|\mathcal{G}_K\|_p\,. \tag{7.51}$$

Finally, for $j \in I_K^{\mathrm{BA}}$ we estimate $\frac{\mathrm{dist}(\boldsymbol{x}_j,K)}{\mathrm{diam}(K)}$ by assuming first that $j \notin I_K$. Then, there exists $\ell \in I_K\backslash I^{\mathrm{BA}}$ such that $j \in H_\ell$, and thus $\frac{\mathrm{dist}(\boldsymbol{x}_j,\boldsymbol{x}_\ell)}{\mathrm{diam}(K)} \le \mathrm{reg}_{\mathrm{BA}}(\mathcal{D}^{\mathrm{BA}})$. This gives

$$\begin{aligned}\frac{\mathrm{dist}(\boldsymbol{x}_j,K)}{\mathrm{diam}(K)} &\le \frac{\mathrm{dist}(\boldsymbol{x}_j,\boldsymbol{x}_\ell)}{\mathrm{diam}(K)} + \frac{\mathrm{dist}(\boldsymbol{x}_\ell,K)}{\mathrm{diam}(K)}\\ &\le \mathrm{reg}_{\mathrm{BA}}(\mathcal{D}^{\mathrm{BA}}) + \max_{i\in I_K}\frac{\mathrm{dist}(\boldsymbol{x}_i,K)}{\mathrm{diam}(K)}.\end{aligned} \tag{7.52}$$

This last inequality obviously also holds if $j \in I_K$. The proof of (7.44) is completed by recalling the definition (7.35) of $\mathrm{reg}_{\mathrm{LLE}}$, by combining (7.50), (7.51) and (7.52), and by using $\mathrm{reg}_{\mathrm{BA}}(\mathcal{D}^{\mathrm{BA}}) \ge 1$. ∎

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 7.40 and Proposition 7.36, and the other three properties result from the fact that $X_{\mathcal{D}^{\mathrm{BA}},0}$ is (roughly) a subspace of $X_{\mathcal{D},0}$.

Theorem 7.41 (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, 2.6 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 $(\mathrm{reg}_{\mathrm{LLE}}(\mathcal{D}_m))_{m\in\mathbb{N}}$ is bounded. For any $m \in \mathbb{N}$ we take a barycentric condensation $\mathcal{D}_m^{\mathrm{BA}}$ of $\mathcal{D}_m$ in the sense of Definition 7.38, such that $(\mathrm{reg}_{\mathrm{BA}}(\mathcal{D}_m^{\mathrm{BA}}))_{m\in\mathbb{N}}$ is bounded.*
Then $(\mathcal{D}_m^{\mathrm{BA}})_{m\in\mathbb{N}}$ is also coercive, GD-consistent, limit-conforming, and compact. Moreover, we have

$$C_{\mathcal{D}_m^{\mathrm{BA}}} \le C_{\mathcal{D}_m} \quad \text{and} \quad W_{\mathcal{D}_m^{\mathrm{BA}}} \le W_{\mathcal{D}_m}. \tag{7.53}$$

Remark 7.42. 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 $(\mathcal{D}_m^{\mathrm{BA}})_{m\in\mathbb{N}}$ is also coercive.

Proof. For any $v \in X_{\mathcal{D}_m^{\mathrm{BA}},0}$, with $\widetilde{v}$ defined by (7.43) we have

$$\left\|\Pi_{\mathcal{D}_m^{\mathrm{BA}}} v\right\|_{L^p(\Omega)} = \|\Pi_{\mathcal{D}_m}\widetilde{v}\|_{L^p(\Omega)}$$

$$\leq C_{\mathcal{D}_m} \|\nabla_{\mathcal{D}_m} \widetilde{v}\|_{L^p(\Omega)^d} = C_{\mathcal{D}_m} \left\|\nabla_{\mathcal{D}_m^{\mathrm{BA}}} v\right\|_{L^p(\Omega)^d}.$$

This shows that $C_{\mathcal{D}_m^{\mathrm{BA}}} \leq C_{\mathcal{D}_m}$ and thus that $(\mathcal{D}_m^{\mathrm{BA}})_{m\in\mathbb{N}}$ is coercive if $(\mathcal{D}_m)_{m\in\mathbb{N}}$ is coercive.
To prove the compactness, we take $(\nabla_{\mathcal{D}_m^{\mathrm{BA}}} v_m)_{m\in\mathbb{N}} = (\nabla_{\mathcal{D}_m} \widetilde{v_m})_{m\in\mathbb{N}}$ bounded in $L^p(\Omega)^d$, and we use the compactness of $(\mathcal{D}_m)_{m\in\mathbb{N}}$ to see that $(\Pi_{\mathcal{D}_m} \widetilde{v_m})_{m\in\mathbb{N}} = (\Pi_{\mathcal{D}_m^{\mathrm{BA}}} v_m)_{m\in\mathbb{N}}$ is relatively compact in $L^p(\Omega)$.
The limit conformity follows by writing, for $v \in X_{\mathcal{D}_m^{\mathrm{BA}},0}$ and $\boldsymbol{\varphi} \in W^{\mathrm{div},p'}(\Omega)$,

$$\begin{aligned}
&\frac{1}{\left\|\nabla_{\mathcal{D}_m^{\mathrm{BA}}} v\right\|_{L^p(\Omega)^d}} \left| \int_\Omega \left(\nabla_{\mathcal{D}_m^{\mathrm{BA}}} v(\boldsymbol{x}) \cdot \boldsymbol{\varphi}(\boldsymbol{x}) + \Pi_{\mathcal{D}_m^{\mathrm{BA}}} v(\boldsymbol{x}) \mathrm{div}\boldsymbol{\varphi}(\boldsymbol{x}) \right) \mathrm{d}\boldsymbol{x} \right| \\
&\qquad = \frac{1}{\|\nabla_{\mathcal{D}_m} \widetilde{v}\|_{L^p(\Omega)^d}} \left| \int_\Omega \left(\nabla_{\mathcal{D}_m} \widetilde{v}(\boldsymbol{x}) \cdot \boldsymbol{\varphi}(\boldsymbol{x}) + \Pi_{\mathcal{D}_m} \widetilde{v}(\boldsymbol{x}) \mathrm{div}\boldsymbol{\varphi}(\boldsymbol{x}) \right) \mathrm{d}\boldsymbol{x} \right|,
\end{aligned}$$

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 7.40 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 7.36 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 7.43 (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 2.10.*

Example 7.44 (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 DOFs. Recalling the notations in Definition

2.10, 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 7.47 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 289, 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 7.45 (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 degrees of freedom. 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 2.10 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 7.46 (Mass lumping with respect to a non canonical basis)
The basis $(e_i)_{i\in B}$ of $X_{\mathcal{D},0}$ used in Definition 2.10 to perform a mass-lumping of $\mathcal{D}$ is usually a canonical basis, each vector in this basis corresponding to a natural degree of freedom 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 7.47 (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, 2.6 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*

$$\forall m \in \mathbb{N}\,,\ \forall v \in X_{\mathcal{D}_m,0}\,,\ \left\|\Pi_{\mathcal{D}_m}^{\mathrm{ML}} v - \Pi_{\mathcal{D}_m} v\right\|_{L^p(\Omega)} \leq \omega_m \left\|v\right\|_{\mathcal{D}_m}\,, \tag{7.54}$$

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 7.48 below, which gives a general setting for proving the properties of a GD by comparing it with another GD.

Theorem 7.48 (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*

$$\begin{aligned}&\lim_{m\to\infty}\omega_m = 0 \text{ and, for all } m\in\mathbb{N} \text{ and all } v\in X_{\mathcal{D}_m,0},\\ &\left\|\Pi^\star_{\mathcal{D}_m}v - \Pi_{\mathcal{D}_m}v\right\|_{L^p(\Omega)} \le \omega_m \|v\|_{\mathcal{D}_m}.\end{aligned} \tag{7.55}$$

   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, 2.6 and 2.8), then* $(\mathcal{D}^\star_m)_{m\in\mathbb{N}}$ *is also coercive (resp. GD-consistent, limit-conforming, or compact).*
2. *Reciprocally, if* $(\mathcal{D}_m)_{m\in\mathbb{N}}$ *and* $(\mathcal{D}^\star_m)_{m\in\mathbb{N}}$ *are both limit-conforming and compact in the sense of Definitions 2.6 and 2.8, then there exists* $(\omega_m)_{m\in\mathbb{N}}$ *such that* (7.55) *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 triangular inequality and invoking (7.55), we have, for any $v\in X_{\mathcal{D}_m,0}$,

$$\begin{aligned}\left\|\Pi^\star_{\mathcal{D}_m}v\right\|_{L^p(\Omega)} &\le \left\|\Pi^\star_{\mathcal{D}_m}v - \Pi_{\mathcal{D}_m}v\right\|_{L^p(\Omega)} + \|\Pi_{\mathcal{D}_m}v\|_{L^p(\Omega)}\\ &\le M\|\nabla_{\mathcal{D}_m}v\|_{L^p(\Omega)^d} + C_{\mathcal{D}_m}\|\nabla_{\mathcal{D}_m}v\|_{L^p(\Omega)^d}.\end{aligned}$$

The coercivity of $(\mathcal{D}^\star_m)_{m\in\mathbb{N}}$ follows, with $C_{\mathcal{D}^\star_m}\le M + C_{\mathcal{D}_m}\le M + C_P$.

GD-consistency: let us assume that $(\mathcal{D}_m)_{m\in\mathbb{N}}$ is consistent. Using the triangular inequality and (7.55), we write, for $v\in X_{\mathcal{D}_m,0}$ and $\varphi\in W^{1,p}_0(\Omega)$,

$$\begin{aligned}S_{\mathcal{D}^\star_m}(\varphi) \le{}& \left\|\Pi_{\mathcal{D}^\star_m}v - \varphi\right\|_{L^p(\Omega)} + \|\nabla_{\mathcal{D}_m}v - \nabla\varphi\|_{L^p(\Omega)^d}\\ \le{}& \omega_m\|\nabla_{\mathcal{D}_m}v\|_{L^p(\Omega)^d} + \|\Pi_{\mathcal{D}_m}v - \varphi\|_{L^p(\Omega)} + \|\nabla_{\mathcal{D}_m}v - \nabla\varphi\|_{L^p(\Omega)^d}\\ \le{}& \omega_m\|\nabla\varphi\|_{L^p(\Omega)^d} + \omega_m\|\nabla_{\mathcal{D}_m}v - \nabla\varphi\|_{L^p(\Omega)^d}\\ &+ \|\Pi_{\mathcal{D}_m}v - \varphi\|_{L^p(\Omega)} + \|\nabla_{\mathcal{D}_m}v - \nabla\varphi\|_{L^p(\Omega)^d}\\ \le{}& \omega_m\|\nabla\varphi\|_{L^p(\Omega)^d}\\ &+ (1+M)(\|\Pi_{\mathcal{D}_m}v - \varphi\|_{L^p(\Omega)} + \|\nabla_{\mathcal{D}_m}v - \nabla\varphi\|_{L^p(\Omega)^d}).\end{aligned}$$

Hence $S_{\mathcal{D}^\star_m}(\varphi)\le \omega_m\|\nabla\varphi\|_{L^p(\Omega)^d} + (1+M)S_{\mathcal{D}_m}(\varphi)$ and the consistency of $(\mathcal{D}^\star_m)_{m\in\mathbb{N}}$ 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 triangular inequality and (7.55), for any $\boldsymbol{\varphi}\in W^{\mathrm{div},p'}(\Omega)$,

$$\begin{aligned}&\left|\int_\Omega\Big(\nabla_{\mathcal{D}_m}v(\boldsymbol{x})\cdot\boldsymbol{\varphi}(\boldsymbol{x}) + \Pi^\star_{\mathcal{D}_m}v(\boldsymbol{x})\mathrm{div}\boldsymbol{\varphi}(\boldsymbol{x})\Big)\mathrm{d}\boldsymbol{x}\right|\\ &\qquad\le \|\mathrm{div}\boldsymbol{\varphi}\|_{L^{p'}(\Omega)}\,\omega_m\|\nabla_{\mathcal{D}_m}v\|_{L^p(\Omega)^d}\\ &\qquad\quad + \left|\int_\Omega(\nabla_{\mathcal{D}_m}v(\boldsymbol{x})\cdot\boldsymbol{\varphi}(\boldsymbol{x}) + \Pi_{\mathcal{D}_m}v(\boldsymbol{x})\mathrm{div}\boldsymbol{\varphi}(\boldsymbol{x}))\,\mathrm{d}\boldsymbol{x}\right|.\end{aligned}$$

We infer that $W_{\mathcal{D}_m^\star}(\boldsymbol{\varphi}) \le \omega_m \|\mathrm{div}\boldsymbol{\varphi}\|_{L^{p'}(\Omega)} + W_{\mathcal{D}_m}(\boldsymbol{\varphi}) \to 0$ 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 (7.55), 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 $(\mathcal{D}_m^\star)_{m\in\mathbb{N}}$ are both compact and limit-conforming, and that

$$\omega_m := \max_{v\in X_{\mathcal{D}_m,0}\setminus\{0\}} \frac{\|\Pi_{\mathcal{D}_m} v - \Pi^\star_{\mathcal{D}_m} v\|_{L^p(\Omega)}}{\|\nabla_{\mathcal{D}_m} v\|_{L^p(\Omega)^d}} \not\to 0 \text{ as } m \to \infty. \tag{7.56}$$

Then there exists $\varepsilon_0 > 0$, a subsequence of $(\mathcal{D}_m, \mathcal{D}_m^\star)_{m\in\mathbb{N}}$ (not denoted differently) and for each $m \in \mathbb{N}$ an element $v_m \in X_{\mathcal{D}_m,0}\setminus\{0\}$ such that $\left\|\Pi^\star_{\mathcal{D}_m} v_m - \Pi_{\mathcal{D}_m} v_m\right\|_{L^p(\Omega)} \ge \varepsilon_0 \|\nabla_{\mathcal{D}_m} v_m\|_{L^p(\Omega)^d}$. 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

$$\left\|\Pi^\star_{\mathcal{D}_m}\widetilde{v}_m - \Pi_{\mathcal{D}_m}\widetilde{v}_m\right\|_{L^p(\Omega)} \ge \varepsilon_0. \tag{7.57}$$

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 $(\mathcal{D}_m^\star)_{m\in\mathbb{N}}$, $\Pi_{\mathcal{D}_m}\widetilde{v}_m \to v$ 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 (7.57) we find $\|v - v^\star\|_{L^p(\Omega)} \ge \varepsilon_0$. 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 $\boldsymbol{\varphi} \in W^{\mathrm{div},p'}(\Omega)$,

$$\left|\int_{\mathbb{R}^d} \left(\nabla_{\mathcal{D}_m}\widetilde{v}_m(\boldsymbol{x})\cdot\boldsymbol{\varphi}(\boldsymbol{x}) + \Pi^\star_{\mathcal{D}_m}\widetilde{v}_m(\boldsymbol{x})\mathrm{div}\boldsymbol{\varphi}(\boldsymbol{x})\right)\mathrm{d}\boldsymbol{x}\right| \le W_{\mathcal{D}_m^\star}(\boldsymbol{\varphi}),$$

and

$$\left|\int_{\mathbb{R}^d} \left(\nabla_{\mathcal{D}_m}\widetilde{v}_m(\boldsymbol{x})\cdot\boldsymbol{\varphi}(\boldsymbol{x}) + \Pi_{\mathcal{D}_m}\widetilde{v}_m(\boldsymbol{x})\mathrm{div}\boldsymbol{\varphi}(\boldsymbol{x})\right)\mathrm{d}\boldsymbol{x}\right| \le W_{\mathcal{D}_m}(\boldsymbol{\varphi}).$$

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 $\|v - v^\star\|_{L^p(\Omega)} \ge \varepsilon_0$. Therefore the sequence $(\omega_m)_{m\in\mathbb{N}}$ defined by (7.56) satisfies (7.55). ∎

Remark 7.49. Three estimates obtained in the course of this proof deserve to be put forward. Under Assumption (7.55) and setting $M = \sup_{m\in\mathbb{N}} \omega_m$, we saw that

$$C_{\mathcal{D}_m^\star} \le M + C_{\mathcal{D}_m}, \tag{7.58}$$

that

$$\forall \varphi \in W_0^{1,p}(\Omega)\,,\ S_{\mathcal{D}_m^\star}(\varphi) \le \omega_m \left\|\nabla\varphi\right\|_{L^p(\Omega)^d} + (1+M) S_{\mathcal{D}_m}(\varphi), \tag{7.59}$$

and that

$$\forall \boldsymbol{\varphi} \in W^{\mathrm{div},p'}(\Omega)\,,\ W_{\mathcal{D}_m^\star}(\boldsymbol{\varphi}) \le \omega_m \left\|\mathrm{div}\boldsymbol{\varphi}\right\|_{L^{p'}(\Omega)} + W_{\mathcal{D}_m}(\boldsymbol{\varphi}). \tag{7.60}$$

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, (7.59) and (7.60) 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 3.2 and 3.28).
An example of this is given for mass-lumped $\mathbb{P}_1$ gradient discretisations in Remark 8.16.

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 degrees of freedom, the definition of a mass-lumped GD (Definition 7.43) 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 7.50 (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 2.18 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* (7.35).

Proposition 7.51 (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 7.50. 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* (2.16) *holds, and that*

$$\forall \varphi \in C^\infty(\overline{\Omega}),$$
$$\max_{K\in\mathcal{M}_m} \max_{i\in I_K\cap I_\partial} \frac{|(\mathcal{I}_{\mathcal{D}_m,\partial}\gamma(\varphi))_i - \varphi(\boldsymbol{x}_i)|}{\mathrm{diam}(K)} \to 0 \text{ as } m \to \infty. \quad (7.61)$$

Then $(\mathcal{D}_m)_{m\in\mathbb{N}}$ is GD-consistent in the sense of Definition 2.20.

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 (2.16) enables us to check the GD-consistency only on smooth functions (see Lemma 2.21). 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 7.36, 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 7.36 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 2.20.
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 $K \in \mathcal{M}_m$,

$$\omega_m(K) = \max_{i\in I_K} \frac{|w_i^m - v_i^m|}{\mathrm{diam}(K)}. \quad (7.62)$$

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),

$$\begin{aligned}
\|\mathcal{G}_K[(v_i^m)_{i\in I_K}] - \mathcal{G}_K[(w_i^m)_{i\in I_K}]\|_{L^p(K)^d} &\le \left\| \sum_{i\in I_K} |v_i^m - w_i^m|\, |\mathcal{G}_K^i| \right\|_{L^p(K)^d} \\
&\le \omega_m(K)\mathrm{diam}(K) \left\| \sum_{i\in I_K} |\mathcal{G}_K^i| \right\|_{L^p(K)^d} \\
&\le \|\mathcal{G}_K\|_p\, |K|^{\frac{1}{p}} \omega_m(K) \\
&\le \mathrm{reg}_{\mathrm{LLE}}(\mathcal{D}_m) |K|^{\frac{1}{p}} \omega_m(K). \quad (7.63)
\end{aligned}$$

Raising this estimate the power p and summing the result over $K \in \mathcal{M}_m$ gives

$$\|\nabla_{\mathcal{D}_m} v^m - \nabla_{\mathcal{D}_m} w^m\|_{L^p(\Omega)^d} \le \mathrm{reg}_{\mathrm{LLE}}(\mathcal{D}_m)|\Omega|^{\frac{1}{p}} \max_{K\in\mathcal{M}_m} \omega_m(K).$$

Since $(\mathrm{reg}_{\mathrm{LLE}}(\mathcal{D}_m))_{m\in\mathbb{N}}$ is bounded, Assumption (7.61) 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 7.38, 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 degrees of freedom are not eliminated ($I_\partial \subset I^{\mathrm{BA}}$).
Lemma 7.40, 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 7.41, is also valid provided that we assume (2.16) and (7.61) – to establish the GD-consistency by invoking Proposition 7.51.

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 7.48, and thus Theorem 7.47, still hold modulo a trivial adjustment of the space of degrees of freedom.

Neumann boundary conditions

LLE gradient discretisation

Definition 7.52 (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* (7.35).

The proof of the following proposition is identical to the proof of Proposition 7.36, in which we actually did not use the boundary value of the functions.

Proposition 7.53 (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 7.52. 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 2.27.*

Barycentric condensation

Starting from an LLE GD for Neumann conditions as defined above, a barycentric condensation is defined as in Definition 7.38, 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 (7.43). We note that, with the norm (2.25) 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 7.40) is still valid, as well as Theorem 7.41.

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 (C.6),

$$\left| \int_{\Omega} \Pi^{\star}_{\mathcal{D}} v(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \right| \le \left| \int_{\Omega} \Pi_{\mathcal{D}} v(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \right| + |\Omega|^{1/p'} \left\| \Pi^{\star}_{\mathcal{D}} v - \Pi_{\mathcal{D}} v \right\|_{L^p(\Omega)},$$

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 7.48 in which the norm $\|v\|_{\mathcal{D}_m}$ in (7.55) is defined by (2.25).
Theorem 7.47 then clearly holds, provided that we use the norm (2.25) in (7.54).

Remark 7.54 (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 7.48 and 7.47 to hold one must introduce this trace in (7.55) and (7.54). This latter formula, for example, would therefore become

$$\left\| \Pi^{\mathrm{ML}}_{\mathcal{D}_m} v - \Pi_{\mathcal{D}_m} v \right\|_{L^p(\Omega)} + \left\| \mathbb{T}^{\mathrm{ML}}_{\mathcal{D}_m} v - \mathbb{T}_{\mathcal{D}_m} v \right\|_{L^p(\partial\Omega)} \le \omega_m \|v\|_{\mathcal{D}_m}. \tag{7.64}$$

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 7.55 (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

$$\begin{aligned}\mathrm{reg}_{\mathrm{LLE}}(\mathcal{D}) = \max_{K\in\mathcal{M}} \left(\|\pi_K\|_p + \|\mathcal{G}_K\|_p + \max_{i\in I_K} \frac{\mathrm{dist}(\boldsymbol{x}_i, K)}{\mathrm{diam}(K)} \right) \\ + \max_{K_\partial\in\mathcal{M}_\partial} \left(\|\pi_{K_\partial}\|_p + \max_{i\in I_{K_\partial}} \frac{\mathrm{dist}(\boldsymbol{x}_i, K_\partial)}{\mathrm{diam}(K_\partial)} \right).\end{aligned} \tag{7.65}$$

The following proposition is then proved as Proposition 7.36, 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 7.56 (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 7.55. 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 2.49.*

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 7.38 to which we add the relation $\mathbb{T}_{\mathcal{D}^{\mathrm{BA}}} v = \mathbb{T}_{\mathcal{D}} \widetilde{v}$. With the norm (2.48) of a GD for Fourier boundary conditions, we still have $\|v\|_{\mathcal{D}^{\mathrm{BA}}} = \|\widetilde{v}\|_{\mathcal{D}}$.
The preservation of the LLE property (Lemma 7.40) is still valid, the trace reconstruction $\mathbb{T}_{\mathcal{D}^{\mathrm{BA}}}$ being dealt with as $\mathit{\Pi}_{\mathcal{D}^{\mathrm{BA}}}$. A barycentric condensation for Fourier boundary conditions preserves the properties of a GD (Theorem 7.41 holds).

Mass-lumping

The definition of a mass-lumped GD for Fourier boundary conditions is not different from Definition 7.43. In particular, if the trace reconstruction is not mass-lumped, Theorems 7.48 and 7.47 hold. If the trace reconstruction trace is mass-lumped, Assumption (7.54) must be replaced with (7.64).

8

Conforming methods and derived methods

8.1 Conforming Galerkin methods

8.1.1 Homogeneous Dirichlet boundary conditions

Conforming Galerkin methods are probably the simplest GDM there is. They simply consist in replacing the infinite-dimensional Sobolev space involved in the weak formulation (e.g. $H_0^1(\Omega)$ in (3.3)) with a finite dimensional subspace. We can therefore define a corresponding GD as follows.
Let $\mathcal{A} = (\varphi_i)_{i\in I}$ be a linearly independent 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, 2.6 and 2.8.

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's 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 the 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 $\varphi \in W_0^{1,p}(\Omega)$,

$$\begin{aligned}\|\Pi_{\mathcal{D}_m} v - \varphi\|_{L^p(\Omega)} &+ \|\nabla_{\mathcal{D}_m} v - \nabla\varphi\|_{L^p(\Omega)^d} \\ &= \|\Pi_{\mathcal{D}_m} v - \varphi\|_{L^p(\Omega)} + \|\nabla(\Pi_{\mathcal{D}_m} v) - \nabla\varphi\|_{L^p(\Omega)^d} \\ &\le (1 + \operatorname{diam}(\Omega)) \|\nabla(\Pi_{\mathcal{D}_m} v) - \nabla\varphi\|_{L^p(\Omega)^d}.\end{aligned}$$

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 $\boldsymbol{\varphi} \in W^{\operatorname{div},p'}(\Omega)$. 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 8.2. 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 Galerking 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 8.3 (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 2.32.*
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 2.33, 2.27, 2.34 and 2.36.

Remark 8.4 (Fourier boundary conditions). The relations (8.3) also define a conforming Galerkin GD for Fourier boundary conditions, and the equivalent of Theorem 8.3 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 7.5, 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 attached to the degrees of freedoms I, 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 DOFs 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 $I_\Omega = \mathcal{V}_{\rm int}^{(k)} := \mathcal{V}^{(k)} \cap \Omega$, $I_\partial = \mathcal{V}_{\rm ext}^{(k)} := \mathcal{V}^{(k)} \cap \partial\Omega$, and thus

$$X_{\mathcal{D},0} = \{v = (v_{\boldsymbol{s}})_{\boldsymbol{s}\in\mathcal{V}} \,:\, v_{\boldsymbol{s}} \in \mathbb{R} \text{ for all } \boldsymbol{s} \in \mathcal{V}_{\rm int}^{(k)}\,,\ v_{\boldsymbol{s}} = 0 \text{ for all } \boldsymbol{s} \in \mathcal{V}_{\rm ext}^{(k)}\}.$$

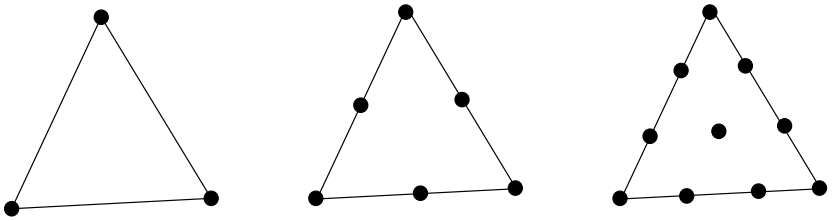


Fig. 8.1. Location of the degrees of freedom in each cell for the $\mathbb{P}_k$ finite element method.

2. For $K \in \mathcal{M}$, we let $I_K = \mathcal{V}_K^{(k)} := \mathcal{V}^{(k)} \cap \overline{K}$. The function reconstruction $\mathit{\Pi}_{\mathcal{D}}$ in (7.33) 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

$$\forall v \in X_{\mathcal{D},0}\,,\ \forall K \in \mathcal{M}\,,\ (\mathit{\Pi}_{\mathcal{D}} v)_{|K} = \sum_{\boldsymbol{s} \in \mathcal{V}_K^{(k)}} v_{\boldsymbol{s}} \pi_K^{\boldsymbol{s}}. \tag{8.6}$$

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 8.5 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 $v \in X_{\mathcal{D},0}$, $\mathit{\Pi}_{\mathcal{D}} v$ is polynomial of degree k or less in each cell, and satisfies $\mathit{\Pi}_{\mathcal{D}} v(\boldsymbol{s}) = v_{\boldsymbol{s}}$ for all $\boldsymbol{s} \in \mathcal{V}^{(k)}$. By Lemma 8.6, $\mathit{\Pi}_{\mathcal{D}} v$ is therefore continuous over Ω, and thus belong to $W^{1,p}(\Omega)$. Moreover, for any $\sigma \in \mathcal{F}_{\rm ext} \cap \mathcal{F}_K$, $\mathit{\Pi}_{\mathcal{D}} v$ 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 8.5 applied to σ instead of K we deduce that $\mathit{\Pi}_{\mathcal{D}} v = 0$ on the boundary faces, and thus that $\mathit{\Pi}_{\mathcal{D}} v \in W_0^{1,p}(\Omega)$.
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 8.5, 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 reasoning made in the construction above.

Lemma 8.5 ($\mathcal{V}_K^{(k)}$ 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 dimention $\mathrm{Card}(\mathcal{V}_K^{(k)})$. 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.

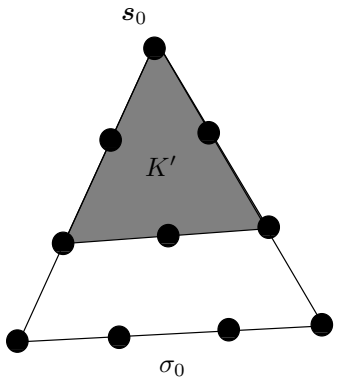


Fig. 8.2. Illustration of the construction in the proof of Lemma 8.5 for $k = 3$.

■

Lemma 8.6 (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 8.5 can be applied to σ, and shows that $(p_K)_{|\sigma}$ and $(p_L)_{|\sigma}$ are identical over the whole face σ.

■

8.2.2 Properties of $\mathbb{P}_k$ gradient discretisations

The properties of $\mathbb{P}_k$ GDs follow from their conformity and from Proposition 7.36, 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 8.7. *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*

$$|M^{-1}| \le \frac{d^{1/2}}{\omega_d \varrho^d} \ell^{-1}, \tag{8.9}$$

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 $\{\sum_{i=1}^d \lambda_i x_i \,:\, \lambda_i \ge 0\,,\ \sum_i \lambda_i \le 1\}$ of $\{0, \boldsymbol{x}_1, \dots, \boldsymbol{x}_d\}$. Therefore

$$|\det(M)| \ge \mathrm{vol}(B(0, \varrho\ell)) = \omega_d \varrho^d \ell^d. \tag{8.10}$$

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 $j = 1, \dots, d$,

$$\begin{aligned}\det(\boldsymbol{x}_1, \dots, \boldsymbol{x}_{j-1}, M\boldsymbol{\xi}, \boldsymbol{x}_{j+1}, \dots, \boldsymbol{x}_d) \\ = \det(\boldsymbol{x}_1, \dots, \boldsymbol{x}_{j-1}, \xi_j \boldsymbol{x}_j, \boldsymbol{x}_{j+1}, \dots, \boldsymbol{x}_d)\end{aligned}$$

$$= \xi_j \det(\boldsymbol{x}_1, \dots, \boldsymbol{x}_d) = \xi_j \det(M), \tag{8.11}$$

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

$$\begin{aligned}\ell^{d-1}|M\boldsymbol{\xi}| &\geq |\boldsymbol{x}_1| \dots |\boldsymbol{x}_{j-1}|\,|M\boldsymbol{\xi}|\,|\boldsymbol{x}_{j+1}| \dots |\boldsymbol{x}_d| \\ &\geq |\det(\boldsymbol{x}_1, \dots, \boldsymbol{x}_{j-1}, M\boldsymbol{\xi}, \boldsymbol{x}_{j+1}, \dots, \boldsymbol{x}_d)| \\ &\geq |\xi_j|\,|\det(M)| \geq \omega_d \varrho^d \ell^d |\xi_j|,\end{aligned}$$

that is, $|M\boldsymbol{\xi}| \geq \omega_d \varrho^d \ell |\xi_j|$. We square this relation, sum over $j = 1, \dots, d$, 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 8.8 (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 7.5, 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 only depends 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 $(0, \boldsymbol{s}_1, \dots, \boldsymbol{s}_d)$ 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 $(\boldsymbol{s}_1, \dots, \boldsymbol{s}_d)$. 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 8.7 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 only depends on d, and satisfies

$$\|\pi_K^{\boldsymbol{s}}(M\cdot)\|_{L^\infty(S_0)} \le C_4 \quad \text{and} \quad \|\nabla(\pi_K^{\boldsymbol{s}}(M\cdot))\|_{L^\infty(S_0)^d} \le C_4.$$

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 $|(M^T)^{-1}| = |M^{-1}| \le C_3 h_K^{-1}$. ∎

We can now prove the properties of $\mathbb{P}_k$ GDs.

Theorem 8.9 (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, 2.6 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, the 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 7.36 and Lemma 8.8. ∎

Remark 8.10 (Rates of convergence of the $\mathbb{P}_k$ GS)
The reasoning of Item 3 in the definition of the $\mathbb{P}_k$ GD shows that the reconstruction $\Pi_{\mathcal{D}}$ is based on local $\mathbb{P}_k$-exact function reconstructions (see Remark 7.37 for the definition). Since $\nabla_{\mathcal{D}} = \nabla(\Pi_{\mathcal{D}})$, this gradient reconstruction is also $\mathbb{P}_k$-exact. By Remark 7.37 we deduce that, under boundedness assumption on $\kappa_{\mathfrak{T}}$, a $\mathbb{P}_k$ GD satisfies

$$S_{\mathcal{D}}(\varphi) \le C h_{\mathcal{M}}^k \|\varphi\|_{W^{k+1,\infty}(\Omega)}.$$

Since $W_{\mathcal{D}} = 0$, Theorem 3.2 gives, as expected, $\mathcal{O}(h_{\mathcal{M}}^k)$ error estimates on the $\mathbb{P}_k$ method applied to the linear diffusion equation (3.1) (in the case $\overline{u} \in W^{k+1,\infty}(\Omega)$). We refer to [12, 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 7.50, 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

$$X_{\mathcal{D},\partial} = \{v \in X_{\mathcal{D}} \,:\, v_{\boldsymbol{s}} = 0 \text{ for all } \boldsymbol{s} \in \mathcal{V}^{(k)}_{\rm ext}\}.$$

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 [23], 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 2.19, 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 $\boldsymbol{s} \in \mathcal{V}^{(k)}_{\rm ext}$,

$$(\mathcal{I}_{\mathcal{D},\partial} g)_{\boldsymbol{s}} = g(\boldsymbol{s}). \tag{8.14}$$

We then have the following result.

Theorem 8.11 (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, 2.6 and 2.8. Moreover, with $S_{\mathcal{D}}$ defined by (2.14), we have $S_{\mathcal{D}_m}(\varphi) \to 0$ as $m \to \infty$, for all $\varphi \in W^{2,\infty}(\Omega)$.

Remark 8.12 (General GD-consistency property)
We state here a weaker version of the consistency (only for regular functions). Checking the consistency in the sense of Definition 2.20 on a dense subset in $W^{1,p}(\Omega)$ of smooth functions would require to ascertain that (2.16) holds. This is somewhat technical and requires the usage of Clément interpolator, with boundary interpolator $\mathcal{I}_{\mathcal{D},\partial}$ defined by (8.14). The literature does not seem to contain clear results in that direction.

Proof. The Poincaré's inequality, integration-by-parts and Rellich theorem in $W^{1,p}_0(\Omega)$ give the coercivity, limit-conformity and compactness as for homogeneous Dirichlet boundary conditions. Given the definition (8.14) 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 7.28 and 7.31 as in the proof of Proposition 7.36. ∎

8.3.2 Neumann boundary conditions

The modification for Neumann boundary conditions is natural. Following Definition 7.52, we simply enable boundary degrees of freedom to be non-zero, *i.e.* we take

$$X_{\mathcal{D}} = \{v = (v_{\boldsymbol{s}})_{\boldsymbol{s}\in\mathcal{V}^{(k)}} \,:\, v_{\boldsymbol{s}} \in \mathbb{R} \text{ for all } \boldsymbol{s} \in \mathcal{V}^{(k)}\}.$$

$\mathit{\Pi}_{\mathcal{D}}$ and $\nabla_{\mathcal{D}}$ are still defined by (8.6) and (8.8) (for all $v \in X_{\mathcal{D}}$).
The proof that (2.18) is a norm on $X_{\mathcal{D}}$ is straightforward. If $\|v\|_{\mathcal{D}} = 0$ then $\nabla_{\mathcal{D}} v = \nabla(\mathit{\Pi}_{\mathcal{D}} v) = 0$ and thus $\mathit{\Pi}_{\mathcal{D}} v$ is constant. As $\|v\|_{\mathcal{D}} = 0$ also implies $\int_\Omega \mathit{\Pi}_{\mathcal{D}} v(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = 0$, we infer that $\mathit{\Pi}_{\mathcal{D}} v = 0$. Then, for all $\boldsymbol{s} \in \mathcal{V}$, $v_{\boldsymbol{s}} = \mathit{\Pi}_{\mathcal{D}} v(\boldsymbol{s}) = 0$, 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

$$\mathbb{T}_{\mathcal{D}} v = \gamma(\mathit{\Pi}_{\mathcal{D}} v) = (\mathit{\Pi}_{\mathcal{D}} v)_{|\partial\Omega}. \tag{8.15}$$

Poincaré–Wirtinger's inequality in $W^{1,p}(\Omega)$ gives C depending only on Ω and p such that, for all $v \in X_{\mathcal{D}}$,

$$\|\mathit{\Pi}_{\mathcal{D}} v\|_{L^p(\Omega)} \le C \left(\|\nabla(\mathit{\Pi}_{\mathcal{D}} v)\|_{L^p(\Omega)^d} + \left| \int_\Omega \mathit{\Pi}_{\mathcal{D}} v(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \right| \right) = C \|v\|_{\mathcal{D}}$$

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 (2.26)) depending only on Ω and p. The choice (8.15) of the trace reconstruction shows that $W_{\mathcal{D}}$, defined by (2.28), is identically zero.
Proposition 7.53 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 8.13 (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 2.33, 2.27, 2.34 and 2.36.

8.3.3 Fourier conditions

For Fourier boundary conditions, the trace is still defined by (8.15) and clearly satisfies the conditions in Definition 7.55, with $\mathcal{M}_\partial = \mathcal{F}_{\rm ext}$ and $I_\sigma = \mathcal{V}_K^{(k)}$ for all $K \in \mathcal{M}$ and all $\sigma \in \mathcal{F}_K \cap \mathcal{F}_{\rm ext}$. A bound on the LLE regularity of the obtained GD can be established as in the proof of Lemma 8.8, 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 $\sigma \in \mathcal{F}_K \cap \mathcal{F}_{\rm ext}$ and all $\boldsymbol{s} \in \mathcal{V}_K^{(k)}$. To bound the quantities $\frac{\mathrm{dist}(\boldsymbol{x}_i, K_\partial)}{\mathrm{diam}(K_\partial)}$ in $\mathrm{reg}_{\rm LLE}(\mathcal{D})$ defined by (7.65), we also use the fact that $\mathrm{diam}(\sigma) \le C h_K$ whenever $\sigma \in \mathcal{F}_K$, with C depending only on an upper bound of $\kappa_{\mathfrak{T}}$.
As a conclusion, by Proposition 7.56, Theorem 8.13 remains valid in the context of Fourier boundary conditions (with Definition 2.27 replaced with Definition 2.49).

8.4 Mass-lumped $\mathbb{P}_1$ finite elements

It is obvious from (8.6) that the reconstruction $\mathit{\Pi}_{\mathcal{D}}$ 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 7.43.
Mass-lumping leads to a piecewise constant reconstruction $\mathit{\Pi}_{\mathcal{D}}^{\rm ML}$, 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 8.14 (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 7.5, and let $\mathcal{D} = (X_{\mathcal{D},0}, \mathit{\Pi}_{\mathcal{D}}, \nabla_{\mathcal{D}})$ 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 $\mathcal{D}^{\rm ML} = (X_{\mathcal{D},0}, \mathit{\Pi}_{\mathcal{D}}^{\rm ML}, \nabla_{\mathcal{D}})$ where $\mathit{\Pi}_{\mathcal{D}}^{\rm ML}$ is the piecewise constant reconstruction built from $(\Omega_{\boldsymbol{s}})_{\boldsymbol{s}\in\mathcal{V}}$, that is

$$\forall v \in X_{\mathcal{D},0}\,,\ \forall \boldsymbol{s} \in \mathcal{V}\,,\ \mathit{\Pi}_{\mathcal{D}}^{\rm ML} v = v_{\boldsymbol{s}} \text{ on } \Omega_{\boldsymbol{s}}.$$

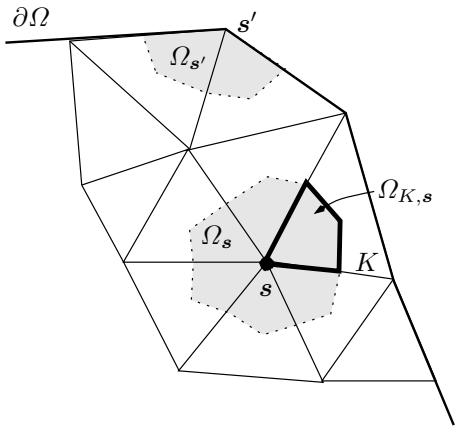


Fig. 8.3. Partitions for the mass-lumping of the $\mathbb{P}_1$ GD.

The mesh $(\Omega_{\boldsymbol{s}})_{\boldsymbol{s}\in\mathcal{V}}$ thus constructed is sometimes called the barycentric dual mesh 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 $\mathfrak{T}$.

The properties of this mass-lumped $\mathbb{P}_1$ GD are stated in the following theorem.

Theorem 8.15 (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 7.5, and let $(\mathcal{D}_m^{\mathrm{ML}})_{m\in\mathbb{N}}$ be the corresponding mass-lumped $\mathbb{P}_1$ GDs given by Definition 8.14. 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, 2.6, 2.8 and 2.10.

Proof. Let us assume that

$$\forall v\in X_{\mathcal{D}_m,0}\,,\ \left\|\Pi_{\mathcal{D}_m}v-\Pi_{\mathcal{D}_m}^{\mathrm{ML}}v\right\|_{L^p(\Omega)}\le h_{\mathcal{M}_m}\left\|\nabla_{\mathcal{D}_m}v\right\|_{L^p(\Omega)^d}. \tag{8.16}$$

Then the conclusion of the theorem follows from Theorem 8.9 (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 7.47.
The proof of (8.16) 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,

$$|\Pi^{\mathrm{ML}}_{\mathcal{D}_m} v(\boldsymbol{x}) - \Pi_{\mathcal{D}_m} v(\boldsymbol{x})| \le h_{\mathcal{M}_m} |\nabla_{\mathcal{D}_m} v(\boldsymbol{x})|. \tag{8.17}$$

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 $\boldsymbol{x} \in \Omega$, (8.17) gives (8.16). ∎

Remark 8.16. If $p > d/2$, by Proposition A.6 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 (8.16) shows that (7.55) holds (with $\mathcal{D}^\star_m = \mathcal{D}^{\mathrm{ML}}_m$) with $\omega_m = h_{\mathcal{M}_m}$. Combined with the previous estimates on $S_{\mathcal{D}_m}$ and $W_{\mathcal{D}_m}$, and with (7.59) and (7.60) in Remark 7.49, this proves that the mass-lumped $\mathbb{P}_1$ gradient discretisations satisfy

$$S_{\mathcal{D}^{\mathrm{ML}}_m}(\varphi) \le C' h_{\mathcal{M}_m} \|\varphi\|_{W^{2,p}(\Omega)}$$

(with C' not depending on m or φ), and

$$W_{\mathcal{D}^{\mathrm{ML}}_m}(\boldsymbol{\varphi}) \le h_{\mathcal{M}_m} \|\mathrm{div}\boldsymbol{\varphi}\|_{L^{p'}(\Omega)}.$$

Hence, as expected, mass-lumped $\mathbb{P}_1$ GSs are order 1 schemes. More precisely, if the exact solution to the linear elliptic problem (3.1) belongs to H^2 and $d = 1, 2, 3$, then the estimates (3.6) and (3.7) are $\mathcal{O}(h_{\mathcal{M}})$ when the mass-lumped $\mathbb{P}_1$ GD is used in the GS (3.4).

8.5 Vertex approximate gradient (VAG) methods

Successive versions of the VAG schemes have been described in several papers [50, 52]. 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 degrees of freedom 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 [52, 51, 53]; this is due to a specific mass-lumping that spreads the reconstructed function between the center of the control volumes and the vertices. Let us remark that the original version of the VAG scheme in [50] uses the same nodal formalism as in Chapter 13, but has been shown in the FVCA6 3D Benchmark [54] to be less precise than the version presented here.

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 condition and space dimension 3, this case being easy to adapt to other boundary conditions and to dimension 2.

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

$$\boldsymbol{x}_\sigma = \frac{1}{\mathrm{Card}(\mathcal{V}_\sigma)} \sum_{\boldsymbol{s} \in \mathcal{V}_\sigma} \boldsymbol{s}. \tag{8.18}$$

   We denote by $\mathfrak{T}^T$ the conforming simplicial mesh (as per Definition 7.5) defined by these thetrehedra these $T_{K,\sigma,\boldsymbol{s},\boldsymbol{s}'}$. 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 centers 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

$$\mathcal{V}^T = \mathcal{P} \cup \mathcal{V} \cup \{\boldsymbol{x}_\sigma \ : \ \sigma \in \mathcal{F}\}. \tag{8.19}$$

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 (8.19), for $\overline{\mathcal{D}}$ we can define the set I of geometrical entities attached to the DOFs 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 $\overline{\mathcal{D}}^{\mathrm{BA}}$ of $\overline{\mathcal{D}}$ (see Definition 7.38) which consists in eliminating the DOFs attached to the internal faces $\mathcal{F}_{\mathrm{int}}$ of $\mathfrak{T}$. Precisely, we let $I^{\mathrm{BA}} = \mathcal{M} \cup \mathcal{V} \cup \mathcal{F}_{\mathrm{ext}}$ and, for $\sigma \in \mathcal{F}_{\mathrm{int}}$, we set $H_\sigma = \mathcal{V}_\sigma$ and we define the coefficients $\beta^\sigma_{\boldsymbol{s}} = 1/\mathrm{Card}(\mathcal{V}_\sigma)$, for all $\boldsymbol{s} \in \mathcal{V}_\sigma$. These coefficients are precisely the ones appearing in (8.18). The mapping $v \in X_{\mathcal{D}^{\mathrm{BA}},0} \mapsto \widetilde{v} \in X_{\mathcal{D},0}$ described by (7.43) 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

$$\begin{aligned} &\forall K \in \mathcal{M}\,, && \widetilde{v}_K = v_K\,, \\ &\forall \boldsymbol{s} \in \mathcal{V}\,, && \widetilde{v}_{\boldsymbol{s}} = v_{\boldsymbol{s}}\,, \\ &\forall \sigma \in \mathcal{F}_{\mathrm{ext}}\,, && \widetilde{v}_\sigma = v_\sigma = 0\,, \\ &\forall \sigma \in \mathcal{F}_{\mathrm{int}}\,, && \widetilde{v}_\sigma = \frac{1}{\mathrm{Card}(\mathcal{V}_\sigma)} \sum_{\boldsymbol{s} \in \mathcal{V}_\sigma} v_{\boldsymbol{s}}. \end{aligned} \tag{8.20}$$

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 7.43. 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

$$\begin{aligned} X_{\mathcal{D},0} = \{ & v = ((v_K)_{K\in\mathcal{M}}, (v_{\boldsymbol{s}})_{\boldsymbol{s}\in\mathcal{V}}, (v_\sigma)_{\sigma\in\mathcal{F}_{\text{ext}}}) \,:\, v_K \in \mathbb{R} \text{ for all } K \in \mathcal{M}\,, \\ & v_{\boldsymbol{s}} \in \mathbb{R} \text{ for all } \boldsymbol{s} \in \mathcal{V} \cap \Omega\,,\ v_{\boldsymbol{s}} = 0 \text{ for all } \boldsymbol{s} \in \mathcal{V} \cap \partial\Omega\,, \\ & v_\sigma = 0 \text{ for all } \sigma \in \mathcal{F}_{\text{ext}}\}. \end{aligned}$$

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 2.10, 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

$$\forall v \in X_{\mathcal{D},0} \,:\, \Pi_{\mathcal{D}} v = \sum_{K\in\mathcal{M}} v_K \mathbf{1}_{\Omega_K} + \sum_{\boldsymbol{s}\in\mathcal{V}} v_{\boldsymbol{s}} \mathbf{1}_{\Omega_{\boldsymbol{s}}}. \tag{8.21}$$

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 $T_{K,\sigma,\boldsymbol{s},\boldsymbol{s}'}$.

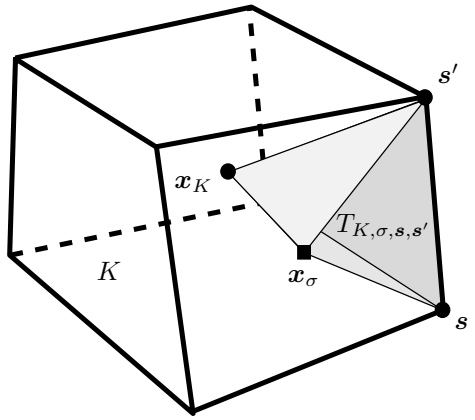


Fig. 8.4. Definition of a simplex $T_{K,\sigma,\boldsymbol{s},\boldsymbol{s}'}$ in a mesh cell K.

Remark 8.17 (Elimination of the cell DOFs in the VAG GS by static condensation) Apply the VAG GDM to obtain a GS (3.4) (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 (3.4) can be reduced to K. By letting α_K be the $\mathbb{P}_1$ Lagrange interpolator in the tetrahedra $(T_{K,\sigma,\boldsymbol{s},\boldsymbol{s}'})_{\sigma,\boldsymbol{s},\boldsymbol{s}'}$, that takes value 1 at $\boldsymbol{x}_K$ and 0 at all other vertices of these tetrahedra, we have, in K,

$$\nabla_{\mathcal{D}} u = u_K \nabla \alpha_K + \sum_{\boldsymbol{s}\in\mathcal{V}_K} u_{\boldsymbol{s}} \Theta_{\boldsymbol{s}}$$

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

$$\begin{aligned} u_K \int_K \Lambda(\boldsymbol{x}) \nabla\alpha_K(\boldsymbol{x}) \cdot \nabla\alpha_K(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \\ = \int_\Omega f(\boldsymbol{x})\alpha_K(\boldsymbol{x})\mathrm{d}\boldsymbol{x} - \sum_{\boldsymbol{s}\in\mathcal{V}_K} u_{\boldsymbol{s}} \int_K \Lambda(\boldsymbol{x})\Theta_{\boldsymbol{s}}(\boldsymbol{x}) \cdot \nabla\alpha_K(\boldsymbol{x})\mathrm{d}\boldsymbol{x}. \end{aligned}$$

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 in a GS for a linear elliptic problem, the cell DOFs can be locally eliminated and expressed in terms of the neighbouring vertex unknowns.

Lemma 8.18 (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 elimination, defined in Item 3 above, of the $\mathbb{P}_1$ GD on $\mathfrak{T}^T$.*
Then there exists C_5 depending only on ϱ such that $\mathrm{reg}_{\mathrm{BA}}(\overline{\mathcal{D}}^{\mathrm{BA}}) \leq C_5$.

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 center of the largest ball included in T; by definition of $\kappa_{\mathfrak{T}^T}$ we have $\rho_T \approx h_T$. Let $(\boldsymbol{s}_i)_{i=1,\dots,d}$ 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 $i = 1, \dots, d$, 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

$$h_T \lesssim \lambda(\boldsymbol{s} - \boldsymbol{s}') \cdot \boldsymbol{n}_{T,\tau} \leq (\boldsymbol{s} - \boldsymbol{s}') \cdot \boldsymbol{n}_{T,\tau}. \tag{8.22}$$

Therefore, $h_T \lesssim |\boldsymbol{s}' - \boldsymbol{s}|$. Since we also have $|\boldsymbol{s}' - \boldsymbol{s}| \leq h_T$, we infer that

$$\begin{gathered}\text{The length of any edge of a tetrahedron } T \in \mathcal{M}^T \\ \text{is comparable to } h_T.\end{gathered} \tag{8.23}$$

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 (8.23),

$$h_{T_1} \approx |\boldsymbol{x}_K - \overline{\boldsymbol{x}}_\sigma| \approx h_{T_2}. \tag{8.24}$$

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 (8.24) we have

$$h_\sigma = |\boldsymbol{s}_1 - \boldsymbol{s}_2| \leq |\boldsymbol{s}_1 - \overline{\boldsymbol{x}}_\sigma| + |\overline{\boldsymbol{x}}_\sigma - \boldsymbol{s}_2| \leq h_{T_1} + h_{T_2} \approx h_T. \tag{8.25}$$

Any edge of σ is also an edge of a tetrahedron with base on σ. Properties (8.23) and (8.25) therefore give

$$\text{The length of any edge of } \sigma \text{ is comparable to } h_\sigma. \tag{8.26}$$

Finally, T shares an edge with σ. Hence, (8.23) and (8.26) show that

$$\text{For any tetrahedron } T \in \mathcal{M}^T \text{ having its base on } \sigma,\ h_T \approx h_\sigma. \tag{8.27}$$

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 degree of freedom that is eliminated in I_K (from the $\mathbb{P}_1$ GD) is the degree of freedom at $\overline{\boldsymbol{x}}_\sigma$. This elimination is done by using the vertices of σ and, by (8.27), 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 only depends on ϱ (through the relations $\approx$). ∎

Remark 8.19 (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 8.20 (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 $\mathfrak{T}_m^T$ of $\mathfrak{T}_m$ as in Item 1 above. Assume that $h_{\mathcal{M}_m} \to 0$ as $m \to \infty$, and that $(\kappa_{\mathfrak{T}_m^T})_{m\in\mathbb{N}}$ 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, 2.6, 2.8 and 2.10.

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 $\overline{\mathcal{D}}_m^{\mathrm{BA}}$ of $\overline{\mathcal{D}}_m$, the result follows from Theorems 7.41 and 7.47 if we can prove that $\mathrm{reg}_{\mathrm{LLE}}(\overline{\mathcal{D}}_m)$ and $\mathrm{reg}_{\mathrm{BA}}(\overline{\mathcal{D}}_m^{\mathrm{BA}})$ remain bounded, and that the following version of (7.54) holds:

$$\forall v \in X_{\mathcal{D}_m,0}\,,\ \left\|\Pi_{\mathcal{D}_m} v - \Pi_{\overline{\mathcal{D}}_m^{\mathrm{BA}}} v\right\|_{L^p(\Omega)} \le h_{\mathcal{M}_m} \left\|\nabla_{\overline{\mathcal{D}}_m^{\mathrm{BA}}} v\right\|_{L^p(\Omega)^d}. \tag{8.28}$$

Since $(\kappa_{\mathfrak{T}_m^T})_{m\in\mathbb{N}}$ is bounded, the boundedness of $\mathrm{reg}_{\mathrm{LLE}}(\overline{\mathcal{D}}_m)$ follows from Lemma 8.8. The bound on $\mathrm{reg}_{\mathrm{BA}}(\overline{\mathcal{D}}_m^{\mathrm{BA}})$ follows from Lemma 8.18. To prove (8.28), 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. $T^{\boldsymbol{s}}_{K,\sigma,\boldsymbol{s},\boldsymbol{s}'}$), $\Pi_{\overline{\mathcal{D}}_m^{\mathrm{BA}}} v$ is linear, $\nabla_{\overline{\mathcal{D}}_m^{\mathrm{BA}}} v = \nabla(\Pi_{\overline{\mathcal{D}}_m^{\mathrm{BA}}} v)$ and $\Pi_{\mathcal{D}_m} v$ is equal to $v_K = \Pi_{\overline{\mathcal{D}}_m^{\mathrm{BA}}} v(\boldsymbol{x}_K)$ (resp. $v_{\boldsymbol{s}} = \Pi_{\overline{\mathcal{D}}_m^{\mathrm{BA}}} v(\boldsymbol{s})$). 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,

$$\left|\Pi_{\mathcal{D}_m} v - \Pi_{\overline{\mathcal{D}}_m^{\mathrm{BA}}} v\right| \le h_{\mathcal{M}_m} \left|\nabla_{\overline{\mathcal{D}}_m^{\mathrm{BA}}} v\right|. \tag{8.29}$$

We then conclude the proof of (8.28) by taking the $L^p(\Omega)$ norms in (8.29). ∎

Theorem 8.21 (Estimate 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*

$$\varrho \ge \kappa_{\mathfrak{T}^T} + \max_{K\in\mathcal{M}} \mathrm{Card}(\mathcal{V}_K),$$

and we let $\mathcal{D}$ be the VAG GD built on $\mathfrak{T}$. Then, there exists C_6 depending only on d, p, Ω and ϱ such that

$$C_{\mathcal{D}} \le C_6, \tag{8.30}$$

$$\forall \boldsymbol{\varphi} \in W^{\mathrm{div},p'}(\Omega)\,,\ W_{\mathcal{D}}(\boldsymbol{\varphi}) \le h_{\mathcal{M}} \left\|\mathrm{div}\boldsymbol{\varphi}\right\|_{L^{p'}(\Omega)} \tag{8.31}$$

and

$$\forall \varphi \in W_0^{1,p}(\Omega) \cap W^{2,p}(\Omega)\,,\ S_{\mathcal{D}}(\varphi) \le C_6 h_{\mathcal{M}} \left\|\varphi\right\|_{W^{2,p}(\Omega)} \tag{8.32}$$

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 (7.58) and (8.28) (which shows that we can take $\omega_m = h_{\mathcal{M}_m} \le \mathrm{diam}(\Omega)$ in (7.55) with $\mathcal{D}_m^\star = \mathcal{D}$ and $\mathcal{D}_m = \overline{\mathcal{D}}^{\mathrm{BA}}$), we have $C_{\mathcal{D}} \le \mathrm{diam}(\Omega) + C_{\overline{\mathcal{D}}^{\mathrm{BA}}}$. We then use (7.53) to get $C_{\overline{\mathcal{D}}^{\mathrm{BA}}} \le C_{\overline{\mathcal{D}}} \le C_P$, where C_P only depends on d, p and Ω (we can actually take $C_P = \mathrm{diam}(\Omega)$, an upper bound of the Poincaré's constant in $W_0^{1,p}(\Omega)$ – remember that $\overline{\mathcal{D}}$ is the $\mathbb{P}_1$ GD). This gives (8.30).

Similarly, Estimate (8.31) follows from (7.60) (in which we can take $\omega_m = h_{\mathcal{M}_m}$ by (8.28)) and from (7.53), which shows that $W_{\overline{\mathcal{D}}^{\mathrm{BA}}} \le W_{\overline{\mathcal{D}}} = 0$.

Owing to (7.59) with $\mathcal{D}_m^\star = \mathcal{D}$ and $\mathcal{D}_m = \overline{\mathcal{D}}^{\mathrm{BA}}$, and to (8.28), to prove (8.32) it suffices to show that $S_{\overline{\mathcal{D}}^{\mathrm{BA}}}(\varphi) \le C_7 h_{\mathcal{M}} \left\|\varphi\right\|_{W^{2,p}(\Omega)}$ with C_7 only depends on p, d, Ω and ϱ. This estimate is obtained by using Proposition A.8 in the appendix, provided that we find sets $(V_K)_{K\in\mathcal{M}}$ that satisfy (A.23) and (A.24), with $\mathcal{D} = \overline{\mathcal{D}}^{\mathrm{BA}}$ and θ depending only on d, p, Ω and ϱ.

We first notice that the bound on $\mathrm{reg}_{\mathrm{LLE}}(\overline{\mathcal{D}}^{\mathrm{BA}})$ in (A.24) is a consequence of Lemma 7.40, Lemma 8.8 (with $\mathcal{D} = \overline{\mathcal{D}}$ the $\mathbb{P}_1$ GD on $\mathfrak{T}^T$), and Lemma 8.18.

Each cell of the mesh $\mathcal{M}^T$ associated to $\overline{\mathcal{D}}^{\mathrm{BA}}$ is a tetrahedron $T_{K,\sigma,\boldsymbol{s},\boldsymbol{s}'}$ in a certain cell $K \in \mathcal{M}$. In Proposition A.8, 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 $\mathrm{Card}(\mathcal{V}_K) \le \varrho$ ensures that the number of such tetrahedra, and thus $\mathrm{Card}(\{T_{K,\sigma,\boldsymbol{s},\boldsymbol{s}'} \in \mathcal{M}^T : \boldsymbol{x} \in V_{T_{K,\sigma,\boldsymbol{s},\boldsymbol{s}'}}\})$, is bounded above by some constant depending only on ϱ. This takes care of the last term in (A.24). 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 $\mathrm{diam}(B_{T_{K,\sigma,\boldsymbol{s},\boldsymbol{s}'}}) \ge C_8 \mathrm{diam}(K)$ with C_8 depending only on ϱ. As in the proof of Lemma 8.18, 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 (8.22) 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 $\mathrm{Card}(\mathcal{V}_K) \lesssim 1$ we have $\mathrm{Card}(\mathcal{F}_K) \lesssim 1$ and Remark 8.19 can be invoked. This gives

$$h_K \approx h_T \lesssim d_{K,\tau}. \tag{8.33}$$

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).

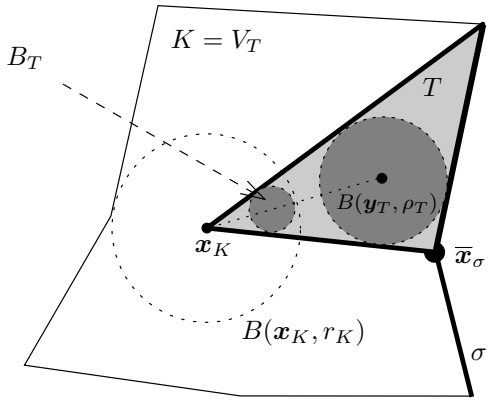


Fig. 8.5. Illustration of the proof of Theorem 8.21. This figure is a planar section of a cell K.

Since $\kappa_{\mathfrak{T}^T} \le \varrho$, T contains a ball $B(\boldsymbol{y}_T, \rho_T)$ with, owing to (8.33),

$$\rho_T \approx h_T \approx h_K \approx r_K. \tag{8.34}$$

We now find B_T – mentioned in (A.23) – 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 $\operatorname{diam}(B_T) = 2\mu\rho_T \approx \mu h_K = \mu \operatorname{diam}(V_T)$, this shows that the second term in the right-hand side of (A.24) is bounded by $\mu^{-1}C_9$ with C_9 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

$$|\boldsymbol{z} - \boldsymbol{x}_K| \le \mu|\boldsymbol{y}_T - \boldsymbol{x}_K| + \mu\rho_T < \mu(h_T + \rho_T) \le C_{10}\mu r_K$$

with C_{10} depending only on ϱ (we used (8.34)). Taking $\mu = 1/C_{10}$ ensures that $B_T \subset B(\boldsymbol{x}_K, r_K)$ and concludes the proof. ∎

9

Non-conforming finite element methods and derived methods

As briefly seen in Chapter 1, the non-conforming $\mathbb{P}_1$ finite element method can be recast in the GDM framework. Let us develop this in more details, using the notions developed in Chapter 7.

9.1 Non-conforming $\mathbb{P}_1$ finite element method for homogeneous Dirichlet boundary conditions

9.1.1 Definition of the non-conforming $\mathbb{P}_1$ gradient discretisation

The non-conforming $\mathbb{P}_1$ finite element method approximates solutions to PDEs with functions that are piecewise linear on a conforming simplicial mesh of Ω, and continuous at the centers of mass of the faces – but not necessarily on the whole edge. This approximation is "non-conforming" in the sense that it approximates the solution $u \in H_0^1(\Omega)$ by functions that are not in $H_0^1(\Omega)$, and therefore do not satisfy in particular the Stokes' formula.
This scheme is often called the Crouzeix-Raviart element, although this name usually pertains to the usage of this method for the Stokes problem (see [26] for the seminal paper and, for instance, [43, pp.25–26 and 199–201] for a synthetic presentation).

Let $\mathfrak{T} = (\mathcal{M}, \mathcal{F}, \mathcal{P}, \mathcal{V})$ be a conforming simplicial mesh of Ω in the sense of Definition 7.5. 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 DOFs 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

   $$X_{\mathcal{D},0} = \{v = (v_\sigma)_{\sigma\in\mathcal{F}} \,:\, v_\sigma \in \mathbb{R} \text{ for all } \sigma \in \mathcal{F}_{\rm int}\,,\ v_\sigma = 0 \text{ for all } \sigma \in \mathcal{F}_{\rm ext}\}.$$

   For all $K \in \mathcal{M}$, we let $I_K = \mathcal{F}_K$.

2. The reconstruction $\Pi_{\mathcal{D}}$ in (7.33) 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

$$\begin{aligned}&\forall\sigma\in\mathcal{F}_K\,,\ \pi_K^\sigma \text{ is affine on } K,\ \pi_K^\sigma(\overline{\boldsymbol{x}}_\sigma)=1,\\ &\text{and } \pi_K^\sigma(\overline{\boldsymbol{x}}_{\sigma'})=0 \text{ for all } \sigma'\in\mathcal{F}_K\backslash\{\sigma\}.\end{aligned}\tag{9.1}$$

   This leads to

$$\forall v\in X_{\mathcal{D},0}\,,\ \forall K\in\mathcal{M}\,,\ \text{ for a.e. } \boldsymbol{x}\in K\,,\ \Pi_{\mathcal{D}}v(\boldsymbol{x})=\sum_{\sigma\in\mathcal{F}_K}v_\sigma\pi_K^\sigma(\boldsymbol{x}).$$

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 (7.34) are the constant functions on the cells given by

$$\mathcal{G}_K^\sigma=\nabla\pi_K^\sigma.\tag{9.2}$$

   Hence, the reconstructed gradients are piecewise constant on the cells:

$$\forall v\in X_{\mathcal{D},0}\,,\ \forall K\in\mathcal{M}\,,\ (\nabla_{\mathcal{D}}v)_{|K}=\sum_{\sigma\in\mathcal{F}_K}v_\sigma\nabla\pi_K^\sigma=\nabla\left[(\Pi_{\mathcal{D}}v)_{|K}\right].\tag{9.3}$$

4. The existence and properties of $(\pi_K^\sigma)_{\sigma\in\mathcal{F}_K}$ and $(\mathcal{G}_K^\sigma)_{\sigma\in\mathcal{F}_K}$, and the fact that $\|\nabla_{\mathcal{D}}\cdot\|_{L^p(\Omega)^d}$ is a norm on $X_{\mathcal{D},0}$, are given by Lemma 9.1 below.

Contrary to the basis functions for the $\mathbb{P}_k$ finite elements, the global basis functions $(\pi^\sigma)_{\sigma\in\mathcal{F}}$ defined as in Section 7.3.3 (that is, $(\pi^\sigma)_{|K}=\pi_K^\sigma$ on K for all $K\in\mathcal{M}$) for non-conforming finite elements have jumps across the faces of the mesh. They are only continuous at the face centers. The function $\Pi_{\mathcal{D}}v=\sum_{\sigma\in\mathcal{F}_{\rm int}}v_\sigma\pi^\sigma$ is therefore also, in general, only continuous at the face centers $(\overline{\boldsymbol{x}}_\sigma)_{\sigma\in\mathcal{F}_{\rm int}}$ – at which it takes the values $(v_\sigma)_{\sigma\in\mathcal{F}_{\rm int}}$.
The gradient (9.3) is defined cell-by-cell and does not account for the jumps of $\Pi_{\mathcal{D}}v$. Hence, the reconstructed gradient $\nabla_{\mathcal{D}}v$ is not $\nabla(\Pi_{\mathcal{D}}v)$ in the sense of distributions on Ω. We nonetheless have $\nabla(\Pi_{\mathcal{D}}v)=\nabla_{\mathcal{D}}v$ in each $K\in\mathcal{M}$, and $\nabla_{\mathcal{D}}v$ is therefore usually refer to as the "broken gradient" of $\Pi_{\mathcal{D}}v$.

9.1.2 Preliminary lemmas

Lemma 9.1. *Let $\mathfrak{T}=(\mathcal{M},\mathcal{F},\mathcal{P},\mathcal{V})$ be a conforming simplicial mesh in the sense of Definition 7.5. Let $K\in\mathcal{M}$, $\pi_K=(\pi_K^\sigma)_{\sigma\in\mathcal{F}_K}$ be given by (9.1), and $\mathcal{G}_K=(\mathcal{G}_K^\sigma)_{\sigma\in\mathcal{F}_K}$ be given by (9.2). 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 $(\overline{\boldsymbol{x}}_\sigma)_{\sigma\in\mathcal{F}_K}$.*
Moreover, $\mathcal{D}$ defined above is an LLE GD, and there exists C_{11}, depending only on d and $\varrho\geq\kappa_{\mathfrak{T}}$ (see (7.10)), such that

$$\mathrm{reg}_{\rm LLE}(\mathcal{D})\leq C_{11}.\tag{9.4}$$

Proof. Let $K \in \mathcal{M}$. The convex hull $\widetilde{K}$ of the centers 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 8.5 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 centers $(\overline{\boldsymbol{x}}_\sigma)_{\sigma\in\mathcal{F}_K}$. This proves, in particular, that the basis functions $(\pi_K^\sigma)_{\sigma\in\mathcal{F}_K}$ are well-defined by (9.1).

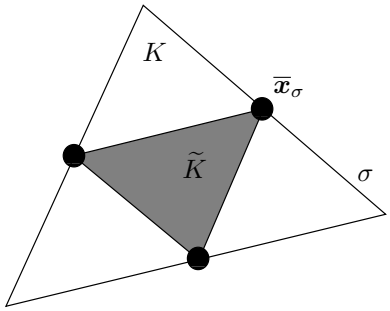


Fig. 9.1. A simplex K and the convex hull $\widetilde{K}$ of its face centers

The map $\sum_{\sigma\in\mathcal{F}_K}\pi_K^\sigma$ is affine and takes the value 1 at each of the face centers 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 values of A at $(\overline{\boldsymbol{x}}_\sigma)_{\sigma\in\mathcal{F}_K}$, and is therefore equal to A. As a consequence, on K,

$$\sum_{\sigma\in\mathcal{F}_K} A(\overline{\boldsymbol{x}}_\sigma)\mathcal{G}_K^\sigma = \sum_{\sigma\in\mathcal{F}_K} A(\overline{\boldsymbol{x}}_\sigma)\nabla\pi_K^\sigma = \nabla\sum_{\sigma\in\mathcal{F}_K} A(\overline{\boldsymbol{x}}_\sigma)\pi_K^\sigma = \nabla A.$$

Hence, $\mathcal{G}_K$ is a $\mathbb{P}_1$-exact gradient reconstruction on K upon $(\overline{\boldsymbol{x}}_\sigma)_{\sigma\in\mathcal{F}_K}$.
To complete the proof that $\mathcal{D}$ is an LLE GD, we need to show that $\|\nabla_{\mathcal{D}}\cdot\|_{L^p(\Omega)^d}$ is a norm on $X_{\mathcal{D},0}$. Let v be in this space and such that $\nabla_{\mathcal{D}}v = 0$. Then, on each cell K, the affine map $\Pi_{\mathcal{D}}v$ has zero gradient (see (9.3)), and is therefore constant on K. Since this map is continuous at the face centers, its constant values in two neighbouring cells must be the same. This shows that $\Pi_{\mathcal{D}}v$ is actually constant on each connected component of Ω. Any such component touches an face $\sigma \in \mathcal{F}_{\rm ext}$, at the center of which $\Pi_{\mathcal{D}}v$ equals $v_\sigma = 0$ (since $v \in X_{\mathcal{D},0}$). Hence, $\Pi_{\mathcal{D}}v = 0$ and all its values $(v_\sigma)_{\sigma\in\mathcal{F}}$ at the face centers are equal to 0. This shows that $v = 0$.

We now establish the upper bound on $\mathrm{reg}_{\rm LLE}(\mathcal{D})$. If $K \in \mathcal{M}$, the simplex $\widetilde{K} \subset K$ created by $(\overline{\boldsymbol{x}}_\sigma)_{\sigma\in\mathcal{F}_K}$ is a rotation 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 8.8 we can use Lemma 8.7 with the vertices of $\widetilde{K}$ as points $\boldsymbol{s}_i$ to see that

$$\|\pi_K^\sigma\|_{L^\infty(\widetilde{K})} \le C_{12} \quad \text{and} \quad \|\nabla\pi_K^\sigma\|_{L^\infty(\widetilde{K})} \le C_{12} h_{\widetilde{K}}^{-1}. \tag{9.5}$$

where C_{12} 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

$$\|\nabla\pi_K^\sigma\|_{L^\infty(K)} \le C_{12} d h_K^{-1}. \tag{9.6}$$

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 (9.5) and (9.6) that

$$\|\pi_K^\sigma\|_{L^\infty(K)} \le C_{12} + C_{12} d. \tag{9.7}$$

Remark 7.32 and Estimates (9.6) and (9.7) give an upper bound on the first two terms in the definition (7.35) 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. ∎

The following lemma provides a control of the non-conforming $\mathbb{P}_1$ GD by a polytopal toolbox, with proper bounds under the usual non-degeneracy assumption on the conforming simplicial meshes.

Lemma 9.2 (Control of the non-conforming $\mathbb{P}_1$ GD by a polytopal toolbox). *Let $\mathfrak{T} = (\mathcal{M}, \mathcal{F}, \mathcal{P}, \mathcal{V})$ be a conforming simplicial mesh of Ω in the sense of Definition 7.5, such that $\mathcal{P}$ are the centers of mass of the cells. Let $\boldsymbol{\Phi} : X_{\mathcal{D},0} \to X_{\mathfrak{T},0}$ be the control of $\mathcal{D}$ by $\mathfrak{T}$ (see Definition 7.10) defined by: for any $v \in X_{\mathcal{D},0}$,*

$$\boldsymbol{\Phi}(v)_K = \frac{1}{d+1}\sum_{\sigma\in\mathcal{F}_K} v_\sigma \quad \text{and} \quad \boldsymbol{\Phi}(v)_\sigma = v_\sigma. \tag{9.8}$$

Then

$$\|\boldsymbol{\Phi}\|_{\mathcal{D},\mathfrak{T}} \le \kappa_{\mathfrak{T}} d^{1/p}, \tag{9.9}$$

$$\omega^{\Pi}(\mathcal{D},\mathfrak{T},\boldsymbol{\Phi}) \le h_{\mathcal{M}}, \tag{9.10}$$

$$\omega^{\nabla}(\mathcal{D},\mathfrak{T},\boldsymbol{\Phi}) = 0. \tag{9.11}$$

Proof. For a given $v \in X_{\mathcal{D},0}$, set $\widehat{v} = \boldsymbol{\Phi}(v)$. Using $\Pi_{\mathcal{D}} v(\overline{\boldsymbol{x}}_\sigma) = v_\sigma$ shows that

$$\widehat{v}_\sigma = \Pi_{\mathcal{D}} v(\overline{\boldsymbol{x}}_\sigma).$$

The center of mass $\overline{\boldsymbol{x}}_K$ of K is given as the convex combination $\frac{1}{d+1}\sum_{\sigma\in\mathcal{F}_K}\overline{\boldsymbol{x}}_\sigma$. Hence, since $\Pi_{\mathcal{D}} v$ is affine in K,

$$\widehat{v}_K = \frac{1}{d+1}\sum_{\sigma\in\mathcal{F}_K} \Pi_{\mathcal{D}} v(\overline{\boldsymbol{x}}_\sigma) = \Pi_{\mathcal{D}} v\left(\frac{1}{d+1}\sum_{\sigma\in\mathcal{F}_K} \overline{\boldsymbol{x}}_\sigma\right) = \Pi_{\mathcal{D}} v(\overline{\boldsymbol{x}}_K).$$

Lemma B.4 page 374 yields $h_K \le \kappa_{\mathfrak{T}}\rho_K \le \kappa_{\mathfrak{T}} d_{K,\sigma}$. Since $\Pi_{\mathcal{D}} v$ is affine in K with constant gradient $(\nabla_{\mathcal{D}} v)_K$, we infer

$$\begin{aligned}|\widehat{v}_\sigma - \widehat{v}_K| = |\Pi_{\mathcal{D}} v(\overline{\boldsymbol{x}}_\sigma) - \Pi_{\mathcal{D}} v(\overline{\boldsymbol{x}}_K)| &\le |\overline{\boldsymbol{x}}_\sigma - \overline{\boldsymbol{x}}_K|\,|(\nabla_{\mathcal{D}} v)_K| \\ &\le h_K\,|(\nabla_{\mathcal{D}} v)_K| \le \kappa_{\mathfrak{T}} d_{K,\sigma}|(\nabla_{\mathcal{D}} v)_K|.\end{aligned}$$

Recalling the definition (7.7f) of $|\cdot|_{\mathfrak{T},p}$, divide the previous inequality by $d_{K,\sigma}$, raise to the power p, multiply by $|\sigma| d_{K,\sigma}$ and sum over $\sigma\in\mathcal{F}_K$ and $K\in\mathcal{M}$ to find

$$|\widehat{v}|_{\mathfrak{T},p}^p \le \kappa_{\mathfrak{T}}^p \sum_{K\in\mathcal{M}}\left(\sum_{\sigma\in\mathcal{F}_K} |\sigma| d_{K,\sigma}\right)|(\nabla_{\mathcal{D}} v)_K|^p.$$

Relation (B.1) p372 gives $\sum_{\sigma\in\mathcal{F}_K}|\sigma| d_{K,\sigma} = d|K|$, and thus

$$|\widehat{v}|_{\mathfrak{T},p}^p \le \kappa_{\mathfrak{T}}^p d\,\|\nabla_{\mathcal{D}} v\|_{L^p(\Omega)^d}^p.$$

This completes the proof of (9.9).
Observe next that, since $\Pi_{\mathcal{D}} v$ is affine in K, for any $\boldsymbol{x}\in K$,

$$\begin{aligned}|\Pi_{\mathcal{D}} v(\boldsymbol{x}) - \Pi_{\mathfrak{T}}\widehat{v}(\boldsymbol{x})| = |\Pi_{\mathcal{D}} v(\boldsymbol{x}) - \widehat{v}_K| &= |\Pi_{\mathcal{D}} v(\boldsymbol{x}) - \Pi_{\mathcal{D}} v(\overline{\boldsymbol{x}}_K)| \\ &\le |\boldsymbol{x} - \overline{\boldsymbol{x}}_K|\,|(\nabla_{\mathcal{D}} v)_K| \le h_{\mathcal{M}}|\nabla_{\mathcal{D}} v(\boldsymbol{x})|.\end{aligned}$$

Raising this last inequality to the power p and integrating over $\boldsymbol{x}\in\Omega$ gives $\|\Pi_{\mathcal{D}} v - \Pi_{\mathfrak{T}}\widehat{v}\|_{L^p(\Omega)} \le h_{\mathcal{M}}\|\nabla_{\mathcal{D}} v\|_{L^p(\Omega)^d}$, which is exactly (9.10).
The relation (9.11) follows from Item 1 in Lemma B.6 (see p376) applied, for any $K\in\mathcal{M}$, to the affine mapping $A = (\Pi_{\mathcal{D}} v)_{|K}$. Indeed, since $(\widehat{v}_K, (\widehat{v}_\sigma)_{\sigma\in\mathcal{F}_K})$ are the values of A at $\overline{\boldsymbol{x}}_K$ and $(\overline{\boldsymbol{x}}_\sigma)_{\sigma\in\mathcal{F}_K}$, this corollary yields $(\overline{\nabla}_{\mathfrak{T}}\widehat{v})_{|K} = \overline{\nabla}_K\widehat{v} = \nabla A = \nabla(\Pi_{\mathcal{D}} v)_{|K} = (\nabla_{\mathcal{D}} v)_{|K}$. ∎

9.1.3 Properties of the non-conforming $\mathbb{P}_1$ finite element method

Thanks to the previous lemmas, the proof of the properties of non-conforming $\mathbb{P}_1$ GDs is straightforward.

Theorem 9.3 (Properties of the non-conforming $\mathbb{P}_1$ GDs for homogeneous Dirichlet BCs). *Let $(\mathcal{D}_m)_{m\in\mathbb{N}}$ be a sequence of non-conforming $\mathbb{P}_1$ GDs, as in Section 9.1.1, defined from underlying conforming simplicial meshes $(\mathfrak{T}_m)_{m\in\mathbb{N}}$ in the sense of Definition 7.5. Assume that the sequence $(\kappa_{\mathfrak{T}_m})_{m\in\mathbb{N}}$ 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.

Proof. Owing to Lemma 9.2, the limit-conformity, coercivity and compactness follow from Corollary 7.13 (use Lemma B.4 to control $\theta_{\mathfrak{T}_m}+\eta_{\mathfrak{T}_m}$ by $\kappa_{\mathfrak{T}_m}$). The consistency is obtained by applying Proposition 7.36, thanks to Lemma 9.1. ∎

The following two propositions are easy consequences of the preliminary results and of some estimates in Appendix A. These propositions are useful to establish precise error estimates for non-conforming $\mathbb{P}_1$ gradient schemes.

Proposition 9.4 (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 7.5, and $\mathcal{D}$ be the non-conforming $\mathbb{P}_1$ GD on $\mathfrak{T}$ as in Section 9.1.1. Assume $p > d/2$ and take $\varrho \geq \kappa_\mathfrak{T}$ (see* (7.10)*). Then there exists $C_{13} > 0$, depending only on p, d, Ω and ϱ, such that, for all $\varphi \in W^{2,p}(\Omega) \cap W_0^{1,p}(\Omega)$,*

$$S_\mathcal{D}(\varphi) \leq C_{13} h_\mathcal{M} \|\varphi\|_{W^{2,p}(\Omega)}.$$

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 $\boldsymbol{x}_i \in \overline{K}$. Using Lemma 9.1, Lemma B.1 and Lemma B.4, we can invoke Proposition A.6 and the conclusion follows. ∎

Proposition 9.5 (Estimate on $W_\mathcal{D}$ for non-conforming $\mathbb{P}_1$ GD). *Let $\mathfrak{T}$ be a conforming simplicial mesh of Ω in the sense of Definition 7.5, and let $\mathcal{D}$ be the non-conforming $\mathbb{P}_1$ GD on $\mathfrak{T}$ as in Section 9.1.1. Take $\varrho \geq \kappa_\mathfrak{T}$ (see* (7.10)*). Then there exists C_{14} depending only on Ω, p, any ϱ, such that*

$$C_\mathcal{D} \leq C_{14} \tag{9.12}$$

and, for all $\boldsymbol{\varphi} \in W^{1,p'}(\Omega)^d$,

$$W_\mathcal{D}(\boldsymbol{\varphi}) \leq \|\boldsymbol{\varphi}\|_{W^{1,p'}(\Omega)^d} C_{14} h_\mathcal{M}. \tag{9.13}$$

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 7.12, Lemma 9.2 and Lemma B.4 (to bound $\theta_\mathfrak{T} + \eta_\mathfrak{T}$). ∎

The application of Propositions 9.4 and 9.5 to the error estimates (3.6) and (3.7) in Theorem 3.2 provides an error in $h_\mathcal{M}$ in the case of a linear elliptic problem in two or three space dimensions, if the exact solution $\overline{u}$ belongs to $H^2(\Omega)$.

9.2 Non-conforming $\mathbb{P}_1$ methods for Neumann and Fourier BCs

9.2.1 Neumann boundary conditions

Definition 7.52 of LLE GDs for Neumann boundary conditions provide 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 9.1.1.

Inequality (9.9) is valid even if v_σ are not zero for boundary edges. If $v \in X_\mathcal{D}$ is such that $\|\nabla_\mathcal{D} v\|_{L^p(\Omega)^d} = 0$, this inequality and the definition (7.7f) of $|\cdot|_{\mathfrak{T},p}$ show that all $(v_\sigma)_{\sigma\in\mathcal{F}}$ are identical, equal to some $c \in \mathbb{R}$. Then, $\mathit{\Pi}_\mathcal{D} v = c$ over Ω and thus, if $\int_\Omega \mathit{\Pi}_\mathcal{D} v(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = 0$, c must be equal to 0. This shows that the the quantity (2.18) 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,

$$\forall v \in X_\mathcal{D}\ ,\ \forall \sigma \in \mathcal{F}_{\rm ext}\ :\ \mathbb{T}_\mathcal{D} v = v_\sigma \text{ on } \sigma. \tag{9.14}$$

Since the regularity factor $\mathrm{reg}_{\rm LLE}(\mathcal{D})$ for Neumann BCs is defined as for Dirichlet BCs, Lemma 9.1 still applies and shows that this factor remains bounded if $\kappa_\mathfrak{T}$ is bounded. Defining the control $\boldsymbol{\Phi} : X_\mathcal{D} \longrightarrow X_\mathfrak{T}$ as in Lemma 9.2, we see that this emma still holds, with moreover $\omega^{\mathbb{T}}(\mathcal{D}, \mathfrak{T}, \boldsymbol{\Phi}) = 0$ in the case of non-homogeneous Neumann BCs. Hence, Corollary 7.19 and Proposition 7.53 give the following theorem.

Theorem 9.6 (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 2.33, 2.27, 2.34 and 2.36.

Proposition A.12 and Theorem 7.18 also give estimates on $S_\mathcal{D}$, $C_\mathcal{D}$ and $W_\mathcal{D}$ that are similar to those in Propositions 9.4 and 9.5. The constants depend only on Ω, p and an upper bound of $\kappa_\mathfrak{T}$.

Remark 9.7 (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 (9.14) by

$$\mathbb{T}^*_\mathcal{D} v = \sum_{\sigma\in\mathcal{F}_{\rm ext}} v_\sigma (\pi^\sigma)_{|\partial\Omega} = (\mathit{\Pi}_\mathcal{D} v)_{|\partial\Omega}.$$

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 $\nabla_\mathcal{D} v = \nabla(\mathit{\Pi}_\mathcal{D} v)$ is constant in K,

$$|\mathbb{T}^*_\mathcal{D} v(\boldsymbol{x}) - \mathbb{T}_\mathcal{D} v(\boldsymbol{x})| = |\mathit{\Pi}_\mathcal{D} v(\boldsymbol{x}) - \mathit{\Pi}_\mathcal{D} v(\overline{\boldsymbol{x}}_\sigma)| \le h_K |(\nabla_\mathcal{D} v)_K|.$$

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 $h_K|\sigma| \le C_{15}|K|$, where C_{15} only depends on an upper bound on $\theta_\mathfrak{T}$,

$$\int_\sigma |\mathbb{T}^*_{\mathcal{D}} v(\boldsymbol{x}) - \mathbb{T}_{\mathcal{D}} v(\boldsymbol{x})|^p \mathrm{d}\boldsymbol{x} \leq C_{15} h_{\mathcal{M}}^{p-1} |K| \, |(\nabla_{\mathcal{D}} v)_K|^p.$$

Sum this estimate over $\sigma \in \mathcal{F}_{\mathrm{ext}}$. A given K can have at most $d+1$ boundary faces (and only in the trivial case where $\Omega = K$, otherwise $\mathrm{Card}(\mathcal{F}_K \cap \mathcal{F}_{\mathrm{ext}}) \leq d$), and thus

$$\begin{aligned}
\|\mathbb{T}^*_{\mathcal{D}} v - \mathbb{T}_{\mathcal{D}}\|^p_{L^p(\partial\Omega)} &\leq (d+1) C_{15} h_{\mathcal{M}}^{p-1} \sum_{K \in \mathcal{M},\, \partial K \cap \partial\Omega \neq \emptyset} |K| \, |(\nabla_{\mathcal{D}} v)_{|K}|^p \\
&\leq (d+1) C_{15} h_{\mathcal{M}}^{p-1} \sum_{K \in \mathcal{M}} |K| \, |(\nabla_{\mathcal{D}} v)_{|K}|^p \\
&= (d+1) C_{15} h_{\mathcal{M}}^{p-1} \, \|\nabla_{\mathcal{D}} v\|^p_{L^p(\Omega)^d} \, .
\end{aligned}$$

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.

9.2.2 Fourier boundary conditions

Starting from the non-conforming $\mathbb{P}_1$ GD for Dirichlet boundary conditions, we follow Definition 7.55 in Section 7.3.6 to define an non-conforming $\mathbb{P}_1$ GD for Fourier boundary conditions.

The boundary mesh $\mathcal{M}_\partial$ is simply $\mathcal{F}_{\mathrm{ext}}$, and the trace reconstruction (9.14) corresponds, for $K_\partial = \sigma \in \mathcal{F}_{\mathrm{ext}}$, to $I_\sigma = \{\sigma\}$ and $\pi^\sigma_\sigma = 1$ on σ, $\pi^\sigma_\sigma = 0$ outside σ. The bound on $\mathrm{reg}_{\mathrm{LLE}}(\mathcal{D})$ 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 7.56.

As noticed in Remark 7.21, the work done for Neumann boundary conditions then immediately show that Theorem 9.6 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 9.4 and 9.5.

We finally remark that, instead of $\mathbb{T}_{\mathcal{D}}$ defined by (9.14), we can also use $\mathbb{T}^*_{\mathcal{D}}$ defined in Remark 9.7

9.3 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

$$\forall g \in W^{1-\frac{1}{p},p}(\partial\Omega)\,,\ \forall \sigma \in \mathcal{F}_{\mathrm{ext}} \,:\, (\mathcal{I}_{\mathcal{D},\partial} g)_\sigma = \frac{1}{|\sigma|} \int_\sigma g(\boldsymbol{x}) \mathrm{d}s(\boldsymbol{x}). \tag{9.15}$$

This interpolant clearly satisfies (7.61) since, for any $i = \sigma \in I_\partial = \mathcal{F}_{\rm ext}$, $\boldsymbol{x}_i = \overline{\boldsymbol{x}}_\sigma$ is the center of mass of σ and therefore, if $\varphi \in C^\infty(\overline{\Omega})$, $\varphi(\boldsymbol{x}_i) = \frac{1}{|\sigma|}\int_\sigma \varphi(\boldsymbol{x})\mathrm{d}s(\boldsymbol{x}) + \mathcal{O}(\mathrm{diam}(\sigma)^2)$.
We now check that (2.16) holds with C_1 depending only on Ω, p and an upper bound of $\kappa_{\mathfrak{T}}$. To this end, take $\varphi \in W^{1,p}(\Omega)$, define $v \in X_{\mathfrak{T}}$ by (7.17) in Proposition 7.15, and $w \in X_{\mathcal{D}}$ by $w_\sigma = v_\sigma$ for all $\sigma \in \mathcal{F}$. Since $w - \mathcal{I}_{\mathcal{D},\partial}\gamma\varphi \in X_{\mathcal{D},0}$, (2.16) is proved if we can establish that

$$\|\Pi_{\mathcal{D}} w\|_{L^p(\Omega)} + \|\nabla_{\mathcal{D}} w\|_{L^p(\Omega)^d} \le C_1 \|\varphi\|_{W^{1,p}(\Omega)} \tag{9.16}$$

Let $\widehat{w} \in X_{\mathfrak{T}}$ be given by $\widehat{w} = \boldsymbol{\Phi}(w)$ as defined by (9.8). Then $\widehat{w}$ and v have the same face values, and since the gradient $\overline{\nabla}_{\mathfrak{T}}$ depends only on the face values (see (7.7e)), we infer that $\overline{\nabla}_{\mathfrak{T}}\widehat{w} = \overline{\nabla}_{\mathfrak{T}} v$. As seen at the end of the proof of Lemma 9.2, $\nabla_{\mathcal{D}} w = \overline{\nabla}_{\mathfrak{T}}\widehat{w}$ (this also holds for non-zero boundary values). Item 2 of Lemma B.6 page 376 gives a bound on $\left\|\overline{\nabla}_{\mathfrak{T}} v\right\|_{L^p(\Omega)^d}$ by $|v|_{\mathfrak{T},p}$, and Estimate (7.18) in Proposition 7.15 yields a bound on $|v|_{\mathfrak{T},p}$ by $\|\varphi\|_{W^{1,p}(\Omega)}$. Gathering all these results show that

$$\|\nabla_{\mathcal{D}} w\|_{L^p(\Omega)^d} = \left\|\overline{\nabla}_{\mathfrak{T}} v\right\|_{L^p(\Omega)^d} \le d^{\frac{p-1}{p}} |v|_{\mathfrak{T},p} \le C_{16} \|\nabla\varphi\|_{L^p(\Omega)^d} \tag{9.17}$$

where C_{16} 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 7.15). Since $w_\sigma = v_\sigma$ for all $\sigma \in \mathcal{F}$ we can use the definitions (9.8) of $\widehat{w}_K$ and (7.7c) of $\Pi_{\mathfrak{T}} v$, and Estimate (B.11) p377, to see that for any $K \in \mathcal{M}$ and any $\boldsymbol{x} \in K$,

$$|\Pi_{\mathfrak{T}}\widehat{w}(\boldsymbol{x}) - \Pi_{\mathfrak{T}} v(\boldsymbol{x})|^p \le \frac{C_{17} h_K^p}{|K|} \int_K |\nabla\varphi(\boldsymbol{y})|^p \mathrm{d}\boldsymbol{y}$$

with C_{17} depending only on Ω, p and an upper bound on $\kappa_{\mathfrak{T}}$. Integrating this over K and summing over $K \in \mathcal{M}$ gives

$$\|\Pi_{\mathfrak{T}}\widehat{w} - \Pi_{\mathfrak{T}} v\|_{L^p(\Omega)} \le \mathrm{diam}(\Omega) C_{17}^{1/p} \|\nabla\varphi\|_{L^p(\Omega)^d}.$$

Moreover, (9.10) (also valid for vectors with non-zero boundary values) for w yields

$$\|\Pi_{\mathcal{D}} w - \Pi_{\mathfrak{T}}\widehat{w}\|_{L^p(\Omega)} \le \mathrm{diam}(\Omega) \|\nabla_{\mathcal{D}} w\|_{L^p(\Omega)^d}.$$

Invoking all these estimates, (7.18) and (9.17) enable us to infer that

$$\|\Pi_{\mathcal{D}} w\|_{L^p(\Omega)} \le \mathrm{diam}(\Omega)(C_{16} + C_{17}^{1/p}) \|\nabla\varphi\|_{L^p(\Omega)^d} + \|\varphi\|_{L^p(\Omega)}. \tag{9.18}$$

Gathering (9.17) and (9.18) proves (9.16).

As a consequence, since (7.61) and (2.16) hold, Proposition 7.51 can be invoked (using Lemma 9.1 to bound $\mathrm{reg}_{\rm LLE}(\mathcal{D})$) and shows that sequences of non-conforming $\mathbb{P}_1$ GDs for non-homogeneous Dirichlet BCs are 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. For non-conforming $\mathbb{P}_1$ GDs with non-homogeneous Dirichlet conditions, these properties therefore follow from Theorem 9.3.

9.4 Mass-lumped non-conforming $\mathbb{P}_1$ reconstruction

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 2.11 are not satisfied, which might prevent the usage of the non-conforming $\mathbb{P}_1$ scheme for some nonlinear problems. To recover a piecewise constant reconstruction, we apply to the non-conforming $\mathbb{P}_1$ GD the mass lumping process as in Definition 7.43.

Definition 9.8 (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 7.5, and let $\mathcal{D} = (X_{\mathcal{D},0}, \Pi_{\mathcal{D}}, \nabla_{\mathcal{D}})$ be the non-conforming $\mathbb{P}_1$ GD built on $\mathfrak{T}$ as in Section 9.1.1.*
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

$$\forall v \in X_{\mathcal{D},0}\,,\ \forall \sigma \in \mathcal{F}\,,\ \Pi_{\mathcal{D}}^{\rm ML} v = v_\sigma \text{ on } \Omega_\sigma.$$

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 7.47.

Theorem 9.9 (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 7.5, 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 9.8. 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, 2.6, 2.8 and 2.10.

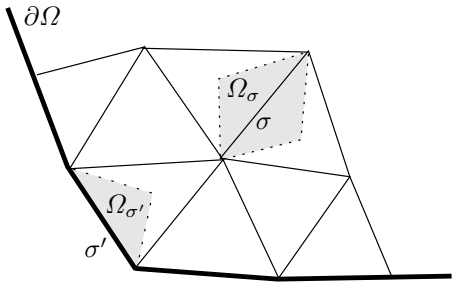


Fig. 9.2. Partition for the mass-lumping of the non-conforming $\mathbb{P}_1$ finite element method.

Proof. In each $D_{K,\sigma}$, $\Pi_{\mathcal{D}_m} v$ is linear and $\Pi^{\mathrm{ML}}_{\mathcal{D}_m} v = \Pi_{\mathcal{D}_m} v(\overline{\boldsymbol{x}}_\sigma)$. Hence, for $\boldsymbol{x} \in D_{K,\sigma}$,

$$\begin{aligned}|\Pi_{\mathcal{D}_m} v(\boldsymbol{x}) - \Pi^{\mathrm{ML}}_{\mathcal{D}_m} v(\boldsymbol{x})| &= |\Pi_{\mathcal{D}_m} v(\boldsymbol{x}) - \Pi_{\mathcal{D}_m} v(\overline{\boldsymbol{x}}_\sigma)| \\ &\le h_{\mathcal{M}} |(\nabla_{\mathcal{D}} v)_{|D_{K,\sigma}}| = h_{\mathcal{M}} |\nabla_{\mathcal{D}} v(\boldsymbol{x})|.\end{aligned}$$

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

$$\left\| \Pi_{\mathcal{D}_m} v - \Pi^{\mathrm{ML}}_{\mathcal{D}_m} v \right\|_{L^p(\Omega)} \le h_{\mathcal{M}_m} \left\| \nabla_{\mathcal{D}_m} v \right\|_{L^p(\Omega)^d}. \tag{9.19}$$

The conclusion then follows from the properties of the non-conforming $\mathbb{P}_1$ GDs (Theorem 9.3) and from Theorem 7.47. ∎

Remark 9.10. As in Remark 8.16, Propositions 9.4 and 9.5, Estimate (9.19) and Remark 7.49 show that, for $p > d/2$,

$$S_{\mathcal{D}^{\mathrm{ML}}_m}(\varphi) \le C h_{\mathcal{M}_m} \|\varphi\|_{W^{2,p}(\Omega)}$$

(with C not depending on m or φ) and

$$W_{\mathcal{D}^{\mathrm{ML}}_m}(\boldsymbol{\varphi}) \le h_{\mathcal{M}_m} \|\mathrm{div}\boldsymbol{\varphi}\|_{L^{p'}(\Omega)}.$$

Mass-lumped non-conforming $\mathbb{P}_1$ are thus order 1 schemes: if the exact solution to the linear elliptic problem (3.1) belongs to H^2 and $d = 1, 2, 3$, then the estimates (3.6) and (3.7) are $\mathcal{O}(h_{\mathcal{M}})$ when the mass-lumped non-conforming $\mathbb{P}_1$ GD is used in the GS (3.4).

10

Mixed finite element $\mathbb{RT}_k$ schemes

In this chapter, we only consider the case $p = 2$ and homogeneous Dirichlet boundary conditions. We establish that both primal and dual forms of the $\mathbb{RT}_k$ mixed finite elements are GDMs. In the primal form, the computation of the reconstructed gradient, which has to be $H_{\mathrm{div}}(\Omega)$ conforming, implies the resolution of a global linear system (which is actually part of the $\mathbb{RT}_k$ linear system for approximating the diffusion equation (3.1)). On the contrary, in the dual form, the reconstructed gradient is computed locally, as for the GDMs studied in the other chapters.

10.1 The $\mathbb{RT}_k$ mixed finite element scheme for linear elliptic problems

Let us first recall the primal and dual formulations of the mixed finite element method [41] for the linear anisotropic diffusion problem. We consider a slighlty simplified version of this problem, corresponding to $\boldsymbol{F} = 0$: find $\bar{u} \in H_0^1(\Omega)$ the solution to

$$\begin{aligned} &\overline{u} \in H_0^1(\Omega) \text{ such that, for all } v \in H_0^1(\Omega), \\ &\int_\Omega \Lambda(\boldsymbol{x}) \nabla \overline{u}(\boldsymbol{x}) \cdot \nabla v(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} = \int_\Omega f(\boldsymbol{x}) v(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}. \end{aligned} \tag{10.1}$$

The assumptions are as usual:

- Ω is an open polytopal bounded connected subset of $\mathbb{R}^d$,
- Λ is a measurable function from Ω to $\mathcal{M}_d(\mathbb{R})$ and there exists $\underline{\lambda}, \overline{\lambda} > 0$ such that, for a.e. $\boldsymbol{x} \in \Omega$, $\Lambda(\boldsymbol{x})$ is symmetric with eigenvalues in $[\underline{\lambda}, \overline{\lambda}]$,
- $f \in L^2(\Omega)$.

Take a conforming simplicial mesh $\mathfrak{T} = (\mathcal{M}, \mathcal{F}, \mathcal{P}, \mathcal{V})$ in the sense of Definition 7.5, and let

$$H_{\mathrm{div}}(\Omega) = \{\boldsymbol{\varphi} \in L^2(\Omega)^d \,:\, \mathrm{div}\boldsymbol{\varphi} \in L^2(\Omega)\},$$

$$\boldsymbol{V}_h = \{\boldsymbol{v} \in (L^2(\Omega))^d \,:\, \boldsymbol{v}|_K \in \mathbb{RT}_k(K)\,,\ \forall K \in \mathcal{M}\}, \tag{10.2}$$

$$\boldsymbol{V}_h^{\mathrm{div}} = \boldsymbol{V}_h \cap H_{\mathrm{div}}(\Omega), \tag{10.3}$$

$$W_h = \{p \in L^2(\Omega) \,:\, p|_K \in \mathbb{P}_k(K)\,,\ \forall K \in \mathcal{M}\}, \tag{10.4}$$

$$M_h^0 = \left\{\mu : \bigcup_{\sigma\in\mathcal{F}} \overline{\sigma} \to \mathbb{R} \,:\, \mu|_\sigma \in \mathbb{P}_k(\sigma)\,,\ \forall \sigma \in \mathcal{F}\,;\ \mu|_{\partial\Omega} = 0\right\}, \tag{10.5}$$

where

- $\mathbb{P}_k(K)$ is the space of polynomials, on K, of d variables and having degree less than or equal to k,
- $\mathbb{P}_k(\sigma)$ is the space of polynomials, on K, of $d-1$ variables and having degree less than or equal to k,
- $\mathbb{RT}_k(K) = \mathbb{P}_k(K)^d + \boldsymbol{x}\overline{\mathbb{P}}_k(K)$ is the Raviart-Thomas space, on K, of order k (here, $\overline{\mathbb{P}}_k(K)$ is the set of homogeneous polynomials of degree k, including the zero polynomial).

The primal formulation of the $\mathbb{RT}_k$ scheme for (10.1) reads

$$(\boldsymbol{v}, q) \in \boldsymbol{V}_h^{\mathrm{div}} \times W_h, \tag{10.6a}$$

$$\int_\Omega \boldsymbol{w}(\boldsymbol{x}) \cdot \Lambda^{-1}(\boldsymbol{x})\boldsymbol{v}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} - \int_\Omega q(\boldsymbol{x})\mathrm{div}\boldsymbol{w}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = 0, \quad \forall \boldsymbol{w} \in \boldsymbol{V}_h^{\mathrm{div}}, \tag{10.6b}$$

$$\int_\Omega \psi(\boldsymbol{x})\mathrm{div}\boldsymbol{v}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = \int_\Omega \psi(\boldsymbol{x}) f(\boldsymbol{x})\mathrm{d}\boldsymbol{x}, \qquad \forall \psi \in W_h. \tag{10.6c}$$

The dual, or Arnold–Brezzi, formulation [5, 66] corresponds to an hybridation of the primal formulation:

$$(\boldsymbol{v}, q, \lambda) \in \boldsymbol{V}_h \times W_h \times M_h^0, \tag{10.7a}$$

$$\int_K \boldsymbol{w}(\boldsymbol{x}) \cdot \Lambda^{-1}(\boldsymbol{x})\boldsymbol{v}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} - \int_K q(\boldsymbol{x})\mathrm{div}\boldsymbol{w}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} + \sum_{\sigma\in\mathcal{F}_K} \int_\sigma \lambda(\boldsymbol{x})\ \boldsymbol{w}|_K(\boldsymbol{x}) \cdot \boldsymbol{n}_{K,\sigma}\mathrm{d}\gamma(\boldsymbol{x}) = 0,\ \forall \boldsymbol{w} \in \boldsymbol{V}_h, \tag{10.7b}$$

$$\int_K \psi(\boldsymbol{x})\mathrm{div}\boldsymbol{v}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = \int_K \psi(\boldsymbol{x}) f(\boldsymbol{x})\mathrm{d}\boldsymbol{x},\ \forall \psi \in W_h, \forall K \in \mathcal{M}, \tag{10.7c}$$

$$\int_\sigma \mu(\boldsymbol{x})\ \boldsymbol{v}|_K(\boldsymbol{x}) \cdot \boldsymbol{n}_{K,\sigma}\mathrm{d}\gamma(\boldsymbol{x}) + \int_\sigma \mu(\boldsymbol{x})\ \boldsymbol{v}|_L(\boldsymbol{x}) \cdot \boldsymbol{n}_{L,\sigma}\mathrm{d}\gamma(\boldsymbol{x}) = 0, \quad \forall \sigma \in \mathcal{F}_{\mathrm{int}} \text{ with } \mathcal{M}_\sigma = \{K, L\},\ \forall \mu \in M_h^0. \tag{10.7d}$$

It is shown in e.g. [5] that the problems (10.6) and (10.7) admit a unique solution, and that the solutions $(\boldsymbol{v}, q)$ to (10.6) and (10.7) are identical.
Moreover, the following error estimate holds [41, Theorem 5.3 p.39]: there exists δ, depending only on Ω, $\underline{\lambda}$, $\overline{\lambda}$, and an upper bound of $\kappa_{\mathfrak{T}}$ such that

$$\|q-\overline{u}\|_{L^2(\Omega)}+\|\boldsymbol{v}+\Lambda\nabla\overline{u}\|_{H_{\rm div}(\Omega)} \leq \delta\left(\inf_{\psi\in W_h}\|\psi-\overline{u}\|_{L^2(\Omega)}+\inf_{\boldsymbol{w}\in\boldsymbol{V}_h^{\rm div}}\|\boldsymbol{w}-\Lambda\nabla\overline{u}\|_{H_{\rm div}(\Omega)}\right). \quad (10.8)$$

10.2 Gradient discretisation from primal mixed finite element

We construct a GD (in the sense of Section 2.1) inspired from the primal mixed finite element formulation (10.6) of Problem (10.1). Let W_h be defined by (10.4) and let $(\chi_i)_{i\in I}$ be a family of piecewise polynomial basis functions of degree k on each cell, spanning W_h. Define the gradient discretisation $\mathcal{D}=(X_{\mathcal{D},0},\Pi_{\mathcal{D}},\nabla_{\mathcal{D}})$ by:

$$X_{\mathcal{D},0}=\{v=(v_i)_{i\in I}\,:\,v_i\in\mathbb{R}\text{ for all }i\in I\}, \quad (10.9a)$$

$$\forall u\in X_{\mathcal{D},0}\,,\ \Pi_{\mathcal{D}}u=\sum_{i\in I}u_i\chi_i, \quad (10.9b)$$

$$\forall u\in X_{\mathcal{D},0}\,,\ \Lambda\nabla_{\mathcal{D}}u\in\boldsymbol{V}_h^{\rm div}\text{ and} \quad (10.9c)$$

$$\int_\Omega\boldsymbol{v}(\boldsymbol{x})\cdot\nabla_{\mathcal{D}}u(\boldsymbol{x})\mathrm{d}\boldsymbol{x}+\int_\Omega\Pi_{\mathcal{D}}u(\boldsymbol{x})\mathrm{div}\boldsymbol{v}(\boldsymbol{x})\mathrm{d}\boldsymbol{x}=0\,,\ \forall\boldsymbol{v}\in\boldsymbol{V}_h^{\rm div}. \quad (10.9d)$$

Remark 10.1 (Λ-dependent gradient discretisation)
A GD only relies on the definition of discrete operators and should be problem independent. However, the tensor Λ appears here in the definition of the gradient reconstruction. This is done to ensure that (10.9c) holds, but it also means that the GD defined above is problem-dependent.
An alternate option, that would lead to a problem-independent GD, is to perform the same construction without Λ. In this case, the convergence can still be proved, and the gradient is in $H_{\rm div}$, but $\Lambda\nabla_{\mathcal{D}}u$ is not. In the case of highly anisotropic and heterogeneous problems, one can therefore expect lower convergence rates for this alternate construction.

In order for (10.9) to define a GD, the system (10.9c)-(10.9d) should define one and only one $\nabla_{\mathcal{D}}u$, and $\|\cdot\|_{\mathcal{D}}:=\|\nabla_{\mathcal{D}}\cdot\|_{L^2(\Omega)^d}$ has to be a norm on $X_{\mathcal{D},0}$. The existence and uniqueness of $\nabla_{\mathcal{D}}u$ results from the fact that (10.9d) can be written as the square linear system

$$\int_\Omega\boldsymbol{v}(\boldsymbol{x})\cdot\Lambda^{-1}(\boldsymbol{x})(\Lambda\nabla_{\mathcal{D}}u)(\boldsymbol{x})\mathrm{d}\boldsymbol{x}=-\int_\Omega\Pi_{\mathcal{D}}u(\boldsymbol{x})\mathrm{div}\boldsymbol{v}(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\,,\ \forall\boldsymbol{v}\in\boldsymbol{V}_h^{\rm div}$$

on $\Lambda\nabla_{\mathcal{D}}u$, whose solution vanishes if the right-hand-side vanishes (consider $\Pi_{\mathcal{D}}u=0$, take $\boldsymbol{v}=\Lambda\nabla_{\mathcal{D}}u$ and use the coercivity of Λ^{-1}). The fact that $\|\nabla_{\mathcal{D}}\cdot\|_{L^2(\Omega)^d}$ defines a norm results from the coercivity property shown in the Theorem 10.2 below.

Before analysing this GD, we need to 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 $H^1(K)$. For $(\boldsymbol{V}_h^{\rm div}, W_h)$ defined by (10.3)–(10.4), by [41, Theorem 3.1 and Lemma 3.5] the interpolation operator P_k, defined by

$$\begin{aligned} &P_k \ : \ \boldsymbol{H}_{\mathcal{M}} := H_{\rm div}(\Omega) \cap (H^1(\mathcal{M}))^d \to \boldsymbol{V}_h^{\rm div}, \\ &\forall p \in W_h, \ \forall \boldsymbol{v} \in \boldsymbol{H}_{\mathcal{M}}, \ \int_\Omega p(\boldsymbol{x}) {\rm div}(\boldsymbol{v} - P_k \boldsymbol{v})(\boldsymbol{x}) {\rm d}\boldsymbol{x} = 0, \end{aligned} \tag{10.10}$$

satisfies

$$\forall \boldsymbol{v} \in \boldsymbol{H}_{\mathcal{M}}, \ \|\boldsymbol{v} - P_k \boldsymbol{v}\|_{L^2(\Omega)^d} \le \alpha h_{\mathcal{M}} \left(\sum_{K \in \mathcal{M}} \|\boldsymbol{v}\|_{H^1(K)}^2 \right)^{1/2}, \tag{10.11}$$

where $\alpha > 0$ depends only on an upper bound of $\kappa_{\mathfrak{T}}$.
The standard "inf-sup" condition [65] 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 $w \in H_0^1(B)$ such that

$$\forall q \in H_0^1(B), \int_B \nabla w(\boldsymbol{x}) \cdot \nabla q(\boldsymbol{x}) {\rm d}\boldsymbol{x} = \int_B p(\boldsymbol{x}) q(\boldsymbol{x}) {\rm d}\boldsymbol{x}. \tag{10.12}$$

Moreover, $w \in H^2(B)$ and, for some β depending only on d and R,

$$\|w\|_{H^2(B)} \le \beta \|p\|_{L^2(\Omega)}. \tag{10.13}$$

Therefore, since $\nabla w \in \boldsymbol{H}_{\mathcal{M}}$, Estimate (10.11) yields

$$\|\nabla w - P_k \nabla w\|_{L^2(\Omega)^d} \le \alpha \beta h_{\mathcal{M}} \|p\|_{L^2(\Omega)}.$$

Since $h_{\mathcal{M}} \le {\rm diam}(\Omega) \le 2R$, this shows that

$$\|P_k \nabla w\|_{L^2(\Omega)^d} \le (2R\alpha + 1)\beta \|p\|_{L^2(\Omega)}. \tag{10.14}$$

The inf-sup condition follows by writing, for any $p \in W_h$, thanks to (10.10) and since $P_k \nabla w \in \boldsymbol{V}_h^{\rm div}$,

$$\begin{aligned} \sup_{\boldsymbol{v} \in \boldsymbol{V}_h^{\rm div}} \frac{\int_\Omega p(\boldsymbol{x}) {\rm div}\boldsymbol{v}(\boldsymbol{x}) {\rm d}\boldsymbol{x}}{\|\boldsymbol{v}\|_{H_{\rm div}(\Omega)}} &\ge \frac{-\int_\Omega p(\boldsymbol{x}) {\rm div}(P_k \nabla w)(\boldsymbol{x}) {\rm d}\boldsymbol{x}}{\|P_k \nabla w\|_{H_{\rm div}(\Omega)}} \\ &\ge \frac{-\int_\Omega p(\boldsymbol{x}) {\rm div}(\nabla w)(\boldsymbol{x}) {\rm d}\boldsymbol{x}}{\|P_k \nabla w\|_{H_{\rm div}(\Omega)}} \\ &\ge \frac{1}{(2R\alpha+1)\beta} \frac{\int_\Omega p(\boldsymbol{x})^2 {\rm d}\boldsymbol{x}}{\|p\|_{L^2(\Omega)}} = \frac{1}{(2R\alpha+1)\beta} \|p\|_{L^2(\Omega)}. \end{aligned}$$

We can now state and prove the properties of the primal $\mathbb{RT}_k$ gradient discretisation.

Theorem 10.2 (Properties of the primal $\mathbb{RT}_k$ GDs). *Let $(\mathfrak{T}_m)_{m\in\mathbb{N}}$ be a sequence of conforming simplicial meshes in the sense of Definition 7.5, 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 $\mathcal{D}_m = (X_{\mathcal{D}_m,0}, \Pi_{\mathcal{D}_m}, \nabla_{\mathcal{D}_m})$ be the gradient discretisation defined by (10.9) for each $m \in \mathbb{N}$.*
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.

Proof.
Coercivity. The proof essentially corresponds to establishing the inf-sup condition as above. Let $u \in X_{\mathcal{D}_m,0}$ and set $p = \Pi_{\mathcal{D}_m} u \in (W_h)_m$. Extend p by 0 outside Ω to the ball B, take $w \in H^1_0(B)$ that satisfies (10.12), and let $\boldsymbol{v} = P_k \nabla w \in \boldsymbol{V}_h^{\mathrm{div}}$. The definition (10.10) of P_k yields

$$\|p\|^2_{L^2(\Omega)} = -\int_\Omega p(\boldsymbol{x})\mathrm{div}\boldsymbol{v}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = -\int_\Omega \Pi_{\mathcal{D}_m} u(\boldsymbol{x})\mathrm{div}\boldsymbol{v}(\boldsymbol{x})\mathrm{d}\boldsymbol{x}.$$

Use (10.9d) to infer

$$\|\Pi_{\mathcal{D}_m} u\|^2_{L^2(\Omega)} = \int_\Omega \boldsymbol{v}(\boldsymbol{x}) \cdot \nabla_{\mathcal{D}_m} u(\boldsymbol{x})\mathrm{d}\boldsymbol{x}.$$

Estimate (10.14) then gives

$$\|\Pi_{\mathcal{D}_m} u\|_{L^2(\Omega)} \le (2R\alpha + 1)\beta \|\nabla_{\mathcal{D}_m} u\|_{L^2(\Omega)^d}, \tag{10.15}$$

which proves the coercivity property.suf-cst

GD-consistency. By Lemma 10.3 below, the set

$$\mathcal{R} = \{\varphi \in H^1_0(\Omega) : \exists f \in C^\infty_c(\Omega) \text{ s.t. } \varphi \text{ is the solution to (10.1)}\} \tag{10.16}$$

is dense in $H^1_0(\Omega)$. Hence, the GD-consistency follows from Lemma 2.13 if we prove that, for all $\varphi \in \mathcal{R}$, $S_{\mathcal{D}_m}(\varphi) \to 0$ as $m \to \infty$. Consider the solution $(\boldsymbol{v}, q) \in (\boldsymbol{V}_h^{\mathrm{div}})_m \times (W_h)_m$ to (10.6) with $f = -\mathrm{div}(\Lambda\nabla\varphi)$. 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 $q = \sum_{i\in I} u_i\chi_i = \Pi_{\mathcal{D}_m} u$. Equation (10.6b) then shows that $\boldsymbol{v} = -\Lambda\nabla_{\mathcal{D}_m} u$. Using the error estimate (10.8) leads to

$$\begin{aligned}\|\Pi_{\mathcal{D}_m} u - \varphi\|_{L^2(\Omega)} + \| - \Lambda\nabla_{\mathcal{D}_m} u + \Lambda\nabla\varphi\|_{H_{\mathrm{div}}(\Omega)} \\ \le \delta \left(\inf_{\psi\in(W_h)_m} \|\psi - \varphi\|_{L^2(\Omega)} + \inf_{\boldsymbol{w}\in(\boldsymbol{V}_h^{\mathrm{div}})_m} \|\boldsymbol{w} - \Lambda\nabla\varphi\|_{H_{\mathrm{div}}(\Omega)} \right).\end{aligned}$$

Classical approximation properties of $(\boldsymbol{V}_h^{\mathrm{div}})_m$ and $(W_h)_m$ ensure that the right hand side of the above inequality tends to 0 as $m \to \infty$. The coercivity of Λ then shows that $S_{\mathcal{D}_m}(\varphi)$ tends to 0 as $m \to \infty$.

LIMIT-CONFORMITY. Let $(u_m)_{m\in\mathbb{N}}$ be such that $u_m \in X_{\mathcal{D}_m,0}$, and $\nabla_{\mathcal{D}_m} u_m$ remains bounded in $L^2(\Omega)^d$ as $m \to \infty$. Let $\boldsymbol{\varphi} \in H_{\rm div}(\Omega)$, and $\boldsymbol{\varphi}_m \in (\boldsymbol{V}_h^{\rm div})_m$ be an interpolation of $\boldsymbol{\varphi}$ such that $\|\boldsymbol{\varphi} - \boldsymbol{\varphi}_m\|_{H_{\rm div}(\Omega)} \to 0$ as $m \to \infty$. Then, recalling the definition (2.8) of $\widetilde{W}_{\mathcal{D}}$ and using (10.9),

$$\widetilde{W}_{\mathcal{D}_m}(\boldsymbol{\varphi}, u_m) = \int_\Omega (\nabla_{\mathcal{D}_m} u_m(\boldsymbol{x}) \cdot \boldsymbol{\varphi}(\boldsymbol{x}) + \Pi_{\mathcal{D}_m} u_m(\boldsymbol{x}) {\rm div}\boldsymbol{\varphi}(\boldsymbol{x}))\,{\rm d}\boldsymbol{x} =$$
$$\int_\Omega \Big(\nabla_{\mathcal{D}_m} u_m(\boldsymbol{x}) \cdot (\boldsymbol{\varphi}(\boldsymbol{x}) - \boldsymbol{\varphi}_m(\boldsymbol{x})) + \Pi_{\mathcal{D}_m} u_m(\boldsymbol{x})({\rm div}\boldsymbol{\varphi}(\boldsymbol{x}) - {\rm div}\boldsymbol{\varphi}_m(\boldsymbol{x}))\Big){\rm d}\boldsymbol{x}.$$

Apply the Cauchy–Schwarz inequality and the coercivity estimate (10.15) to deduce

$$|\widetilde{W}_{\mathcal{D}_m}(\boldsymbol{\varphi}, u_m)| \le \|\boldsymbol{\varphi} - \boldsymbol{\varphi}_m\|_{H_{\rm div}(\Omega)} (1 + (2R\alpha + 1)\beta) \|\nabla_{\mathcal{D}_m} u_m\|_{L^2(\Omega)^d}.$$

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 $\widetilde{W}(\boldsymbol{\varphi}, u_m) \to 0$ as $m \to \infty$.

COMPACTNESS. Let $(u_m)_{m\in\mathbb{N}}$ be such that $u_m \in X_{\mathcal{D}_m,0}$ for all $m \in \mathbb{N}$, and $(\|\nabla_{\mathcal{D}_m} u_m\|_{L^2(\Omega)^d})_{m\in\mathbb{N}}$ is bounded. The coercivity and limit-conformity just proved enable us to apply Lemma 2.12. Hence, up to a subsequence denoted the same way, there exists $\overline{u} \in H_0^1(\Omega)$ such that $\Pi_{\mathcal{D}_m} u_m \to \overline{u}$ weakly in $L^2(\Omega)$ and $\nabla_{\mathcal{D}_m} u_m \to \nabla\overline{u}$ weakly in $L^2(\Omega)^d$. Extend all these functions by 0 outside Ω to some ball B containing Ω.
Let $w_m \in H_0^1(B) \cap H^2(B)$ (resp. $w \in H_0^1(B) \cap H^2(B)$) be defined by (10.12) for $p = \Pi_{\mathcal{D}_m} u_m$ (resp. $p = u$). Since $\Pi_{\mathcal{D}_m} u_m \to \overline{u}$ weakly in $L^2(\Omega)$, it converges strongly in $H^{-1}(\Omega)$ (compact embedding of $L^2(\Omega)$ into $H^{-1}(\Omega)$) and thus $w_m \to w$ strongly in $H_0^1(B)$.
Applying (10.13) and (10.15) yields

$$\|\nabla w_m\|_{H^1(\Omega)} \le \beta \|\Pi_{\mathcal{D}_m} u_m\|_{L^2(\Omega)} \le (2R\alpha + 1)\beta^2 \|\nabla_{\mathcal{D}_m} u_m\|_{L^2(\Omega)^d}. \quad (10.17)$$

Let $\boldsymbol{v} = P_k \nabla w_m$ in (10.9d) to write

$$\int_\Omega P_k \nabla w_m(\boldsymbol{x}) \cdot \nabla_{\mathcal{D}_m} u_m(\boldsymbol{x}) {\rm d}\boldsymbol{x} + \int_\Omega \Pi_{\mathcal{D}_m} u_m(\boldsymbol{x}) {\rm div}(P_k \nabla w_m)(\boldsymbol{x}) {\rm d}\boldsymbol{x} = 0,$$

which provides, thanks to (10.10) and the fact that $-{\rm div}(\nabla w_m) = \Pi_{\mathcal{D}_m} u_m$,

$$\int_\Omega P_k \nabla w_m(\boldsymbol{x}) \cdot \nabla_{\mathcal{D}_m} u_m(\boldsymbol{x}) {\rm d}\boldsymbol{x} - \int_\Omega (\Pi_{\mathcal{D}_m} u_m(\boldsymbol{x}))^2 {\rm d}\boldsymbol{x} = 0. \quad (10.18)$$

Use now (10.11) and (10.17), and the convergence of ∇w_m to ∇w in $L^2(\Omega)^d$, to see that $P_k \nabla w_m$ converges in $L^2(\Omega)^d$ to ∇w. By weak-strong convergence (Lemma C.3 page 403) on the first term of (10.18),

$$\lim_{m\to\infty} \int_\Omega (\Pi_{\mathcal{D}_m} u_m(\boldsymbol{x}))^2 {\rm d}\boldsymbol{x} = \int_\Omega \nabla w(\boldsymbol{x}) \cdot \nabla u(\boldsymbol{x}) {\rm d}\boldsymbol{x} = \int_\Omega u(\boldsymbol{x})^2 {\rm d}\boldsymbol{x}.$$

For the last equality, we used the definition of w, solution to (10.12) with $p = u$ – recall also that ∇u and u vanish on B outside Ω. This shows that the convergence of $\Pi_{\mathcal{D}_m} u_m$ to u is actually strong in $L^2(\Omega)$, thus concluding the proof of the compactness of the sequence of GDs. ∎

Lemma 10.3 (A density result). *The set $\mathcal{R}$ defined by* (10.16) *is dense in $H_0^1(\Omega)$.*

Proof. Thanks to the Lax–Milgram lemma, the mapping $T : H^{-1}(\Omega) \to H_0^1(\Omega)$ defined by:

$$\forall \ell \in H^{-1}(\Omega)\,,\ v = T(\ell) \text{ is the solution in } H_0^1(\Omega) \text{ to } -\mathrm{div}(\Lambda \nabla v) = \ell$$

is well-defined, linear and continuous. Since $C_c^\infty(\Omega)$ is dense in $H^{-1}(\Omega)$ and $\mathcal{R} = T(C_c^\infty(\Omega))$, the conclusion follows. ∎

The following theorem establishes the link between the GD (10.9) and the primal form of the mixed finite element method.

Theorem 10.4 (Primal $\mathbb{RT}_k$ is a GDM). *Using the GD* (10.9)*, the GS* (3.4) *for Problem* (10.1) *is equivalent to the primal formulation* (10.6) *of the mixed finite element method.*

Proof. Let $u \in X_{\mathcal{D},0}$ be a solution to (3.4). Let us show that $(\boldsymbol{v}, q) = (-\Lambda \nabla_{\mathcal{D}} u, \Pi_{\mathcal{D}} u)$ is the solution of (10.6). We first observe that (10.9d) ensures (10.6b). Let us now consider $\psi \in W_h$, which can therefore be written $\psi = \Pi_{\mathcal{D}} v$, with $v \in X_{\mathcal{D},0}$. The GS (3.4) gives

$$\int_\Omega \Lambda(\boldsymbol{x}) \nabla_{\mathcal{D}} u(\boldsymbol{x}) \cdot \nabla_{\mathcal{D}} v(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} = \int_\Omega f(\boldsymbol{x}) \psi(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}. \tag{10.19}$$

Write (10.9d) with u replaced by v

$$\int_\Omega \boldsymbol{v}(\boldsymbol{x}) \cdot \nabla_{\mathcal{D}} v(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} + \int_\Omega \Pi_{\mathcal{D}} v(\boldsymbol{x}) \mathrm{div} \boldsymbol{v}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} = 0,$$

and substitute $\boldsymbol{v} = -\Lambda \nabla_{\mathcal{D}} u$ to obtain

$$\int_\Omega \Lambda(\boldsymbol{x}) \nabla_{\mathcal{D}} u(\boldsymbol{x}) \cdot \nabla_{\mathcal{D}} v(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} = \int_\Omega \psi(\boldsymbol{x}) \mathrm{div} \boldsymbol{v}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}.$$

Combined with (10.19), this completes the proof of (10.6c).
Reciprocally, consider the solution $(\boldsymbol{v}, q)$ to (10.6). Since $q \in W_h$, there exists a unique $u \in X_{\mathcal{D},0}$ such that $q = \Pi_{\mathcal{D}} u$. Comparing (10.6b) and (10.9d) yields $\boldsymbol{v} = -\Lambda \nabla_{\mathcal{D}} u$. Following the same computation as above and letting $\psi = \Pi_{\mathcal{D}} v$ for any $v \in X_{\mathcal{D},0}$, we see that (10.6c) implies (3.4). ∎

10.3 Gradient discretisation from dual mixed finite element formulation

We now construct a GD (in the sense of Section 2.1) inspired from the dual mixed finite element formulation (10.7) of Problem (10.1). Let W_h be defined by (10.4) and let again $(\chi_i)_{i\in I}$ be a a family of piecewise polynomial basis functions of degree k on each cell of the mesh, spanning W_h. Let M_h^0 be defined by (10.5) and let $(\xi_j)_{j\in J}$ be a family spanning M_h^0. To avoid confusions in the notations below, select index sets I and J that are disjoints. Recalling that $\boldsymbol{V}_h$ is given by (10.2), let the gradient discretisation $\widetilde{\mathcal{D}} = (X_{\widetilde{\mathcal{D}},0}, \Pi_{\widetilde{\mathcal{D}}}, \nabla_{\widetilde{\mathcal{D}}})$ be defined by:

$$X_{\widetilde{\mathcal{D}},0} = \{v = ((v_i)_{i\in I}, (v_j)_{j\in J}) \,:\, v_k \in \mathbb{R} \text{ for all } k \in I \cup J\}, \tag{10.20a}$$

$$\forall u \in X_{\widetilde{\mathcal{D}},0}\,,\ \Pi_{\widetilde{\mathcal{D}}} u = \sum_{i\in I} u_i \chi_i \text{ and } \Gamma_{\widetilde{\mathcal{D}}} u = \sum_{j\in J} u_j \xi_j, \tag{10.20b}$$

$$\forall u \in X_{\widetilde{\mathcal{D}},0}\,,\ \Lambda \nabla_{\widetilde{\mathcal{D}}} u \in \boldsymbol{V}_h \text{ and, for all } K \in \mathcal{M},$$
$$\int_K \boldsymbol{w}(\boldsymbol{x}) \cdot \nabla_{\widetilde{\mathcal{D}}} u(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} + \int_K \Pi_{\widetilde{\mathcal{D}}} u(\boldsymbol{x}) \mathrm{div} \boldsymbol{w}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}$$
$$- \sum_{\sigma\in\mathcal{F}_K} \int_\sigma \Gamma_{\widetilde{\mathcal{D}}} u(\boldsymbol{x})\ \boldsymbol{w}|_K(\boldsymbol{x}) \cdot \boldsymbol{n}_{K,\sigma} \mathrm{d}\gamma(\boldsymbol{x}) = 0,\ \forall \boldsymbol{w} \in \boldsymbol{V}_h. \tag{10.20c}$$

Remark 10.5. In the case $k = 0$, the GD (10.20) has the same DOFs and function reconstruction as the HMM scheme (see Chapter 12), which is also a GS. Nevertheless, the gradient reconstructions are different.

As in the previous section, in order for (10.20) to define a GD, the system (10.20c) should define one and only one reconstructed gradient $\nabla_{\widetilde{\mathcal{D}}} u$, and $\|\cdot\|_{\widetilde{\mathcal{D}}} := \|\nabla_{\widetilde{\mathcal{D}}} \cdot\|_{L^2(\Omega)^d}$ has to be a norm on $X_{\widetilde{\mathcal{D}},0}$. The existence and uniqueness of $\nabla_{\widetilde{\mathcal{D}}} u$ again results from the fact that (10.20c) provides a square linear system, whose solution vanishes if the right-hand-side vanishes. The fact that it defines a norm results, on one hand, from the coercivity property shown in Theorem 10.6 below, and on the other hand, on [66, Proposition 3.1 p.15], whose consequence is that for given $(u_i)_{i\in I}$ and $\nabla_{\widetilde{\mathcal{D}}} u$, there exists one and only one $(u_j)_{j\in J}$ such that (10.20c) hold.

Theorem 10.6 (Properties of the dual $\mathbb{RT}_k$ GDs). *Let $(\mathfrak{T}_m)_{m\in\mathbb{N}}$ be a sequence of conforming simplicial meshes in the sense of Definition 7.5, 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 (10.20) for each $m \in \mathbb{N}$.*
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.

Proof. We drop the index m for legibility reasons. Denote by $T : X_{\widetilde{\mathcal{D}},0} \to X_{\mathcal{D},0}$ the mapping $T(\widetilde{u}) = (\widetilde{u}_i)_{i\in I}$, where we use the gradient discretisation $\mathcal{D}$ defined by (10.9). We have $\Pi_{\widetilde{\mathcal{D}}}\widetilde{u} = \Pi_{\mathcal{D}}T(\widetilde{u})$ a.e. in Ω. By selecting $\boldsymbol{w} = \Lambda\nabla_{\mathcal{D}}T(\widetilde{u}) \in \boldsymbol{V}_h^{\rm div} \subset \boldsymbol{V}_h$ in (10.20c) and summing on $K \in \mathcal{M}$, all the integrals on $\sigma \in \mathcal{F}_{\rm int}$ vanish, and we obtain

$$\int_\Omega \Lambda\nabla_{\mathcal{D}}T(\widetilde{u})(\boldsymbol{x}) \cdot \nabla_{\widetilde{\mathcal{D}}}\widetilde{u}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} + \int_\Omega \Pi_{\widetilde{\mathcal{D}}}\widetilde{u}(\boldsymbol{x})\mathrm{div}(\Lambda\nabla_{\mathcal{D}}T(\widetilde{u}))(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = 0.$$

Using (10.9d) with $\boldsymbol{v} = \Lambda\nabla_{\mathcal{D}}T(\widetilde{u})$ and $T(\widetilde{u})$ instead of u then yields

$$\begin{aligned}\int_\Omega \Lambda\nabla_{\mathcal{D}}T(\widetilde{u})(\boldsymbol{x}) \cdot \nabla_{\widetilde{\mathcal{D}}}\widetilde{u}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} &= -\int_\Omega \Pi_{\mathcal{D}}T(\widetilde{u})(\boldsymbol{x})\mathrm{div}(\Lambda\nabla_{\mathcal{D}}T(\widetilde{u}))(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\\ &= \int_\Omega \Lambda\nabla_{\mathcal{D}}T(\widetilde{u})(\boldsymbol{x}) \cdot \nabla_{\mathcal{D}}T(\widetilde{u})(\boldsymbol{x})\mathrm{d}\boldsymbol{x}.\end{aligned}$$

The properties of Λ and the Cauchy–Schwarz inequality then lead to

$$\|\nabla_{\mathcal{D}}T(\widetilde{u})\|_{L^2(\Omega)^d}^2 \le \frac{\overline{\lambda}}{\underline{\lambda}}\|\nabla_{\widetilde{\mathcal{D}}}\widetilde{u}\|_{L^2(\Omega)^d}^2, \ \forall u \in X_{\widetilde{\mathcal{D}},0}. \tag{10.21}$$

Coercivity. The coercivity follows from $\Pi_{\widetilde{\mathcal{D}}}\widetilde{u} = \Pi_{\mathcal{D}}T(\widetilde{u})$, from the coercivity of the gradient discretisation $\mathcal{D}$, and from (10.21).

GD-consistency. Thanks to [66, Proposition 3.1 p.15], for any $u \in X_{\mathcal{D},0}$, there exists $\widetilde{u} \in X_{\widetilde{\mathcal{D}},0}$ such that $T(\widetilde{u}) = u$ and $\nabla_{\mathcal{D}}u = \nabla_{\widetilde{\mathcal{D}}}\widetilde{u}$. Take $\varphi \in H_0^1(\Omega)$ and

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

Then, since $\Pi_{\mathcal{D}}u = \Pi_{\mathcal{D}}T(\widetilde{u}) = \Pi_{\widetilde{\mathcal{D}}}\widetilde{u}$,

$$\begin{aligned}S_{\widetilde{\mathcal{D}}}(\varphi) &\le \left\|\Pi_{\widetilde{\mathcal{D}}}\widetilde{u} - \varphi\right\|_{L^2(\Omega)} + \left\|\nabla_{\widetilde{\mathcal{D}}}\widetilde{u} - \nabla\varphi\right\|_{L^2(\Omega)^d}\\ &= \|\Pi_{\mathcal{D}}u - \varphi\|_{L^2(\Omega)} + \|\nabla_{\mathcal{D}}u - \nabla\varphi\|_{L^2(\Omega)^d} \le S_{\mathcal{D}}(\varphi)\end{aligned}$$

and the consistency of $\widetilde{\mathcal{D}}$ follows from the consistency of $\mathcal{D}$.

Limit-conformity. If $\boldsymbol{w} \in \boldsymbol{V}_h^{\rm div}$ then writing (10.20c) 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 $\sigma \in \mathcal{F}_{\rm int}$ with $\mathcal{M}_\sigma = \{K, L\}$), and we see that (10.9d) holds with $\widetilde{\mathcal{D}}$ instead of $\mathcal{D}$. The limit-conformity can therefore be proved in a similar way as the limit-conformity of $\mathcal{D}_m$, by taking $\boldsymbol{\varphi}_m \in (\boldsymbol{V}_h^{\rm div})_m$ that converges to $\boldsymbol{\varphi}$ in $H_{\rm div}$ and by writing, for $\widetilde{u}_m \in X_{\mathcal{D}_m,0}$,

$$\widetilde{W}_{\widetilde{\mathcal{D}}_m}(\boldsymbol{\varphi}, \widetilde{u}_m) = \int_\Omega \left(\nabla_{\widetilde{\mathcal{D}}_m}\widetilde{u}_m(\boldsymbol{x}) \cdot \boldsymbol{\varphi}(\boldsymbol{x}) + \Pi_{\widetilde{\mathcal{D}}_m}\widetilde{u}_m(\boldsymbol{x})\mathrm{div}\boldsymbol{\varphi}(\boldsymbol{x})\right)\mathrm{d}\boldsymbol{x} =$$

$$\int_\Omega \Big(\nabla_{\widetilde{\mathcal{D}}_m} \widetilde{u}_m(\boldsymbol{x}) \cdot (\boldsymbol{\varphi}(\boldsymbol{x}) - \boldsymbol{\varphi}_m(\boldsymbol{x})) + \mathit{\Pi}_{\widetilde{\mathcal{D}}_m} \widetilde{u}_m(\boldsymbol{x}) (\mathrm{div}\boldsymbol{\varphi}(\boldsymbol{x}) - \mathrm{div}\boldsymbol{\varphi}_m(\boldsymbol{x})) \Big) \mathrm{d}\boldsymbol{x}.$$

Compactness. As the coercivity, this property is an immediate consequence of the compactness of the gradient discretisation $\mathcal{D}$, of (10.21) and of $\mathit{\Pi}_{\widetilde{\mathcal{D}}} \widetilde{u} = \mathit{\Pi}_{\mathcal{D}} T(\widetilde{u})$. ∎

We now check that the GD $\widetilde{\mathcal{D}}$ indeed corresponds to the dual $\mathbb{RT}_k$ scheme.

Theorem 10.7 (Dual $\mathbb{RT}_k$ is a GDM). *Using the GD* (10.20)*, the gradient scheme* (3.4) *for Problem* (10.1) *is equivalent to the Arnold–Brezzi formulation* (10.7) *of the mixed finite element method.*

Proof. Let $u \in X_{\widetilde{\mathcal{D}},0}$ be a solution to (3.4), and let us show that $(\boldsymbol{v}, q, \lambda) = (-\mathit{\Lambda} \nabla_{\widetilde{\mathcal{D}}} u, \mathit{\Pi}_{\widetilde{\mathcal{D}}} u, \mathit{\Gamma}_{\widetilde{\mathcal{D}}} u)$ is the solution of (10.7). We first observe that (10.20c) ensures (10.7b). Let $\psi \in W_h$ and $\mu \in M_h^0$, consider a particular $K \in \mathcal{M}$, and take in (3.4) a function test $v \in X_{\widetilde{\mathcal{D}},0}$ such that $\mathit{\Pi}_{\widetilde{\mathcal{D}}} v|_K = \psi|_K$, $\mathit{\Pi}_{\widetilde{\mathcal{D}}} v|_L = 0$ for all $L \in \mathcal{M} \setminus \{K\}$ and $\mathit{\Gamma}_{\widetilde{\mathcal{D}}} v = 0$. Thanks to (10.20c), the support of $\nabla_{\widetilde{\mathcal{D}}} v$ is also reduced to K and the GS (3.4) therefore gives

$$\int_K \mathit{\Lambda}(\boldsymbol{x}) \nabla_{\widetilde{\mathcal{D}}} u(\boldsymbol{x}) \cdot \nabla_{\widetilde{\mathcal{D}}} v(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} = \int_K f(\boldsymbol{x}) \psi(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}.$$

Setting $\boldsymbol{w} = \boldsymbol{v}$ in (10.20c) with u replaced by v, and using $\mathit{\Gamma}_{\widetilde{\mathcal{D}}} v = 0$, we get

$$\int_K \boldsymbol{v}(\boldsymbol{x}) \cdot \nabla_{\widetilde{\mathcal{D}}} v(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} + \int_K \mathit{\Pi}_{\widetilde{\mathcal{D}}} v(\boldsymbol{x}) \mathrm{div}\boldsymbol{v}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} = 0,$$

which implies

$$\begin{aligned} \int_K f(\boldsymbol{x}) \psi(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} &= \int_K \mathit{\Lambda}(\boldsymbol{x}) \nabla_{\widetilde{\mathcal{D}}} u(\boldsymbol{x}) \cdot \nabla_{\widetilde{\mathcal{D}}} v(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \\ &= -\int_K \boldsymbol{v}(\boldsymbol{x}) \cdot \nabla_{\widetilde{\mathcal{D}}} v(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \\ &= \int_K \mathit{\Pi}_{\widetilde{\mathcal{D}}} v(\boldsymbol{x}) \mathrm{div}\boldsymbol{v}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} = \int_K \psi(\boldsymbol{x}) \mathrm{div}\boldsymbol{v}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}. \end{aligned}$$

This completes the proof of (10.7c). Then, we take $\mu \in M_h^0$ and we let $v \in X_{\widetilde{\mathcal{D}},0}$ be such that $\mathit{\Pi}_{\widetilde{\mathcal{D}}} v = 0$ and $\mathit{\Gamma}_{\widetilde{\mathcal{D}}} v|_\sigma = \mu|_\sigma$ for a given $\sigma = K|L \in \mathcal{F}_{\mathrm{int}}$, and $\mathit{\Gamma}_{\widetilde{\mathcal{D}}} v|_{\sigma'} = 0$ for all $\sigma' \in \mathcal{F} \setminus \{\sigma\}$. Again setting $\boldsymbol{w} = \boldsymbol{v}$ in (10.20c) with u replaced by v, we get

$$\int_K \boldsymbol{v}(\boldsymbol{x}) \cdot \nabla_{\widetilde{\mathcal{D}}} v(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} - \int_\sigma \mathit{\Gamma}_{\widetilde{\mathcal{D}}} v(\boldsymbol{x})\ \boldsymbol{v}|_K(\boldsymbol{x}) \cdot \boldsymbol{n}_{K,\sigma} \mathrm{d}\gamma(\boldsymbol{x}) = 0,$$

and

$$\int_L \boldsymbol{v}(\boldsymbol{x}) \cdot \nabla_{\widetilde{\mathcal{D}}} v(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} - \int_\sigma \mathit{\Gamma}_{\widetilde{\mathcal{D}}} v(\boldsymbol{x})\ \boldsymbol{v}|_L(\boldsymbol{x}) \cdot \boldsymbol{n}_{L,\sigma} \mathrm{d}\gamma(\boldsymbol{x}) = 0.$$

Summing these relations and recalling that $\boldsymbol{v} = -\Lambda\nabla_{\widetilde{\mathcal{D}}}u$ gives

$$\int_{K\cup L} \Lambda\nabla_{\mathcal{D}}u(\boldsymbol{x})\cdot\nabla_{\widetilde{\mathcal{D}}}v(\boldsymbol{x})\mathrm{d}\boldsymbol{x} + \int_\sigma \mu(\boldsymbol{x})\ \boldsymbol{v}|_K(\boldsymbol{x})\cdot\boldsymbol{n}_{K,\sigma}\mathrm{d}\gamma(\boldsymbol{x}) + \int_\sigma \mu(\boldsymbol{x})\ \boldsymbol{v}|_L(\boldsymbol{x})\cdot\boldsymbol{n}_{L,\sigma}\mathrm{d}\gamma(\boldsymbol{x}) = 0. \quad (10.22)$$

Using the GS (3.4), 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 (10.22) vanishes. This proves (10.7d).

Conversely, considering the solution $(\boldsymbol{v}, q, \lambda)$ to (10.7), 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 (10.7b), we get that $\boldsymbol{v} = -\Lambda\nabla_{\widetilde{\mathcal{D}}}u$. 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 (10.7c) and (10.7d) imply (3.4), using (10.20c) where u is replaced by v.

■

Here again, the fact that we wish to obtain a mixed finite element scheme has led us to a problem dependent discretization. For the linear problem (10.1), this is not a difficulty. However if the diffusion tensor depends on the unknown, it becomes very intricate to ensure the H_{div} conformity, and in fact quite useless since one can get the approximate continuity of the flux from the GD itself. This line of thought may lead to consider the GD (10.20) with "$\Lambda\nabla_{\mathcal{D}}u \in \boldsymbol{V}_h$" replaced by "$\nabla_{\mathcal{D}}u \in \boldsymbol{V}_h$" in (10.20c).

11

The multi-point flux approximation MPFA-O scheme

The Two-Point Flux Approximation (TPFA) method was introduced in Section 1.1.3. This scheme is interesting because of its simplicity in the case of scalar diffusion operator, since it leads (in 2D) to a 5-point approximation for the Laplace operator after the elimination of the face unknowns. Its GDM version can be used 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 [1] 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.

11.1 MPFA methods for Dirichlet boundary conditions

11.1.1 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 7.5). In each of these cases, $\mathcal{P}$ are the centers 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 11.1):

- *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

$$|V_{K,\boldsymbol{s}}| = \frac{|K|}{\mathrm{Card}(\mathcal{V}_K)} \quad \text{and} \quad |\tau_{\sigma,\boldsymbol{s}}| = \frac{|\sigma|}{\mathrm{Card}(\mathcal{V}_\sigma)}. \tag{11.1}$$

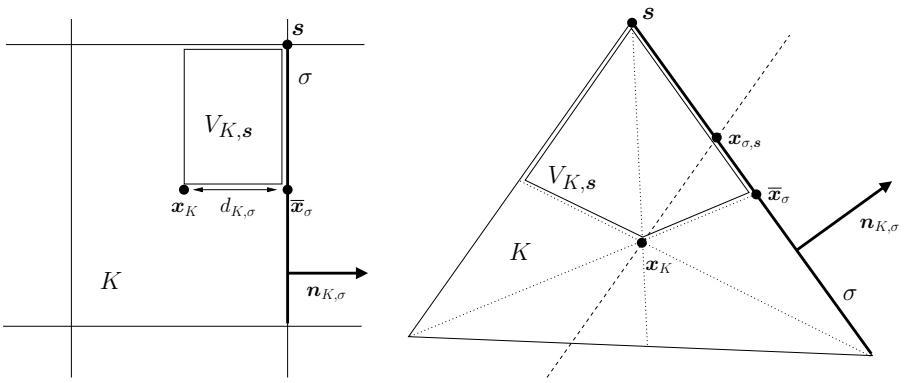


Fig. 11.1. Notations for MPFA-O schemes defined on Cartesian (left) and simplicial (right) meshes.

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 DOFs 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 $I_\Omega = \mathcal{M} \cup \{\tau_{\sigma,\boldsymbol{s}} : \sigma \in \mathcal{F}_{\rm int},\ \boldsymbol{s} \in \mathcal{V}_\sigma\}$ and $I_\partial = \{\tau_{\sigma,\boldsymbol{s}} : \sigma \in \mathcal{F}_{\rm ext},\ \boldsymbol{s} \in \mathcal{V}_\sigma\}$. This gives, with a slight abuse of notation (we should write $v_{\tau_{\sigma,\boldsymbol{s}}}$ instead of $v_{\sigma,\boldsymbol{s}}$),

$$\begin{aligned} X_{\mathcal{D},0} = \{&v = ((v_K)_{K\in\mathcal{M}}, (v_{\sigma,\boldsymbol{s}})_{\sigma\in\mathcal{F},\,\boldsymbol{s}\in\mathcal{V}_\sigma}) : \\ &v_K \in \mathbb{R} \text{ for all } K \in \mathcal{M},\ v_{\sigma,\boldsymbol{s}} \in \mathbb{R} \text{ for all } \sigma \in \mathcal{F}_{\rm int} \text{ and } \boldsymbol{s} \in \mathcal{V}_\sigma, \\ &v_{\sigma,\boldsymbol{s}} = 0 \text{ for all } \sigma \in \mathcal{F}_{\rm ext} \text{ and } \boldsymbol{s} \in \mathcal{V}_\sigma\}. \end{aligned}$$

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

$$\pi_K^i = 1 \text{ for } i = K, \text{ and } \pi_K^i = 0 \text{ for } i = \tau_{\sigma,\boldsymbol{s}}, \tag{11.2}$$

which means that

$$\forall v \in X_{\mathcal{D},0}\,,\ \forall K \in \mathcal{M}\,,\ \forall \boldsymbol{x} \in K\,,\ \Pi_{\mathcal{D}} v(\boldsymbol{x}) = v_K. \tag{11.3}$$

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. $\boldsymbol{x} \in V_{K,\boldsymbol{s}}$,

$$\begin{aligned}
&\mathcal{G}_K^K(\boldsymbol{x}) = -\frac{1}{|V_{K,\boldsymbol{s}}|}\sum_{\sigma\in\mathcal{V}_{K,s}} |\tau_{\sigma,\boldsymbol{s}}|\boldsymbol{n}_{K,\sigma},\\
&\forall \sigma \in \mathcal{F}_{K,\boldsymbol{s}}\,,\ \mathcal{G}_K^{\sigma,\boldsymbol{s}}(\boldsymbol{x}) = \frac{1}{|V_{K,\boldsymbol{s}}|}|\tau_{\sigma,\boldsymbol{s}}|\boldsymbol{n}_{K,\sigma}\,,\\
&\forall \sigma \in \mathcal{F}_{K,\boldsymbol{s}}\,,\ \mathcal{G}_K^{\sigma,\boldsymbol{s}} = 0 \text{ on } K \text{ outside } V_{K,\boldsymbol{s}}.
\end{aligned} \tag{11.4}$$

Hence, for all $v \in X_{\mathcal{D},0}$,

$$\begin{aligned}
&\forall K \in \mathcal{M}\,,\ \forall \boldsymbol{s} \in \mathcal{V}_K\,,\ \text{for a.e. } \boldsymbol{x} \in V_{K,\boldsymbol{s}}\,,\\
&\nabla_{\mathcal{D}} v(\boldsymbol{x}) = \frac{1}{|V_{K,\boldsymbol{s}}|}\sum_{\sigma\in\mathcal{F}_{K,\boldsymbol{s}}} |\tau_{\sigma,\boldsymbol{s}}|(v_{\sigma,\boldsymbol{s}} - v_K)\boldsymbol{n}_{K,\sigma}.
\end{aligned} \tag{11.5}$$

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 11.3 below.

Remark 11.1 (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 (3.4) is a finite volume scheme. Indeed, by selecting a test function with only non-zero value $v_K = 1$ in (3.4), we obtain the flux balance

$$\begin{aligned}
&\sum_{\sigma\in\mathcal{F}_K}\sum_{\boldsymbol{s}\in\mathcal{V}_\sigma} F_{K,\sigma,\boldsymbol{s}}(u) = \int_K f(\boldsymbol{x})\mathrm{d}\boldsymbol{x},\\
&\text{where } F_{K,\sigma,\boldsymbol{s}}(u) = \int_{\sigma_{\boldsymbol{s}}} \mathcal{G}_K u(\boldsymbol{x})\cdot\boldsymbol{n}_{K,\sigma}\mathrm{d}s(\boldsymbol{x}).
\end{aligned} \tag{11.6}$$

Selecting a test function with only non-zero value $v_{\sigma,\boldsymbol{s}} = 1$ in (3.4) leads to the conservativity of the fluxes:

$$\begin{aligned}
&F_{K,\sigma,\boldsymbol{s}}(u) + F_{L,\sigma,\boldsymbol{s}}(u) = 0\\
&\text{for all } \sigma \in \mathcal{F}_{\text{int}} \text{ with } \mathcal{M}_\sigma = \{K, L\}, \text{ and all } \boldsymbol{s} \in \mathcal{V}_\sigma.
\end{aligned} \tag{11.7}$$

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 (11.7) 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 (3.1) 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 [1]. The GD constructed above therefore gives indeed the MPFA-O scheme when used in the GS (3.4).

Remark 11.2 (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 (3.1), the gradient reconstruction defined by the MPFA-O scheme can be used in the finite volume scheme (11.6)-(11.7); however, the GS (3.4) built upon this gradient reconstruction cannot be expected to always converge, since the corresponding GD may fail to be limit-conforming and coercive.

11.1.2 Preliminary lemmas

Let us first prove that the GD constructed above is indeed an LLE GD, and let us estimate its regularity.

Lemma 11.3 (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 (11.2), and $\mathcal{G}_K = (\mathcal{G}_K^i)_{i\in I_K}$ be defined by (11.4). 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*

$$\forall \xi = (\xi_K, (\xi_{\sigma,\boldsymbol{s}})_{\sigma\in\mathcal{F}_K,\,\boldsymbol{s}\in\mathcal{V}_\sigma})\,,\ (\mathcal{G}_K\xi)_{|_{V_{K,\boldsymbol{s}}}} \cdot (\boldsymbol{x}_{\sigma,\boldsymbol{s}} - \boldsymbol{x}_K) = \xi_{\sigma,\boldsymbol{s}} - \xi_K. \quad (11.8)$$

Moreover, $\mathcal{D}$ is an LLE GE and there exists C_{18}, depending only on d and $\theta \geq \theta_{\mathfrak{T}}$ (see (7.8)), such that

$$\mathrm{reg}_{\mathrm{LLE}}(\mathcal{D}) \leq C_{18}. \quad (11.9)$$

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:

$$\forall \widetilde{\sigma} \in \mathcal{F}_{K,\boldsymbol{s}}\backslash\{\sigma\}\,,\ (\boldsymbol{x}_{\sigma,\boldsymbol{s}} - \boldsymbol{x}_K) \perp \boldsymbol{n}_{K,\widetilde{\sigma}},\ \text{and} \quad (11.10)$$

$$\frac{1}{|V_{K,\boldsymbol{s}}|}|\tau_{\sigma,\boldsymbol{s}}|\boldsymbol{n}_{K,\sigma} \cdot (\boldsymbol{x}_{\sigma,\boldsymbol{s}} - \boldsymbol{x}_K) = 1. \quad (11.11)$$

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

$$(\mathcal{G}_K\xi)_{|_{V_{K,\boldsymbol{s}}}} \cdot (\boldsymbol{x}_{\sigma,\boldsymbol{s}} - \boldsymbol{x}_K) = \frac{1}{|V_{K,\boldsymbol{s}}|}|\tau_{\sigma,\boldsymbol{s}}|(\xi_{\sigma,\boldsymbol{s}} - \xi_K)\boldsymbol{n}_{K,\sigma} \cdot (\boldsymbol{x}_{\sigma,\boldsymbol{s}} - \boldsymbol{x}_K)$$
$$= \xi_{\sigma,\boldsymbol{s}} - \xi_K,$$

which proves (11.8). 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 (11.10) and (11.11).

- *Cartesian mesh.* For a Cartesian mesh, (11.10) and (11.11) are rather straightforward by inspecting Figure 11.1, 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

$$\beta^K_{\boldsymbol{s}'}(\boldsymbol{x}_{\sigma,\boldsymbol{s}}) = 1/(d+1) \text{ for all } \boldsymbol{s}' \in \mathcal{V}_\sigma \setminus \{\boldsymbol{s}\},$$
$$\beta^K_{\boldsymbol{s}}(\boldsymbol{x}_{\sigma,\boldsymbol{s}}) = 2/(d+1)$$
$$\beta^K_{\overline{\boldsymbol{s}}}(\boldsymbol{x}_{\sigma,\boldsymbol{s}}) = 0.$$

  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 (11.10).
  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 (11.11) due to (11.1).

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 (11.8) 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 $|\tau_{\sigma,\boldsymbol{s}}| \le C_{19} h_K^{d-1}$ and $h_K^d \le C_{19}|V_{K,\boldsymbol{s}}|$ for some C_{19} depending only on θ and d, so that, by (11.4),

$$|\mathcal{G}_K^K| \le dC_{19}^2 h_K^{-1} \text{ and } |\mathcal{G}_K^{\sigma,\boldsymbol{s}}| \le C_{19}^2 h_K^{-1}.$$

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 11.4 (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 centers $(\mathcal{M}, \mathcal{P})$ 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 7.10) as the isomorphism $\boldsymbol{\Phi} : X_{\mathcal{D},0} \longrightarrow X_{\mathfrak{T}',0}$ given by $\boldsymbol{\Phi}(u)_K = u_K$ and $\boldsymbol{\Phi}(u)_{\tau_{\sigma,\boldsymbol{s}}} = u_{\sigma,\boldsymbol{s}}$.
Then, there exists C_{20}, constant if $\mathfrak{T}$ is Cartesian and depending only on $\theta \ge \kappa_{\mathfrak{T}}$ if $\mathfrak{T}$ is simplicial, such that

$$\|\boldsymbol{\Phi}\|_{\mathcal{D},\mathfrak{T}'} \le C_{20}, \tag{11.12}$$

$$\omega^{\Pi}(\mathcal{D}, \mathfrak{T}', \boldsymbol{\Phi}) = 0, \tag{11.13}$$

$$\omega^{\nabla}(\mathcal{D}, \mathfrak{T}', \boldsymbol{\Phi}) = 0. \tag{11.14}$$

Proof. Let $u \in X_{\mathcal{D},0}$ and apply (11.8) to $\xi = (u_K, (u_{\sigma,\boldsymbol{s}})_{\sigma\in\mathcal{F}_K, \boldsymbol{s}\in\mathcal{V}_\sigma})$ to deduce

$$\sum_{\sigma\in\mathcal{F}_K} \sum_{\boldsymbol{s}\in\mathcal{V}_\sigma} |\tau_{\sigma,\boldsymbol{s}}| d_{K,\sigma} \left| \frac{u_{\sigma,\boldsymbol{s}} - u_K}{d_{K,\sigma}} \right|^p \le C_{21} \int_K |\nabla_{\mathcal{D}} u(\boldsymbol{x})|^p \mathrm{d}\boldsymbol{x},$$

with $C_{21} = 1$ for parallelepipedic meshes, and $C_{21} > 0$ depends on $\theta \ge \kappa_{\mathfrak{T}}$ for simplicial meshes. Therefore $\|\boldsymbol{\Phi}(u)\|_{\mathfrak{T}',p}^p \le C_{21} \|\nabla_{\mathcal{D}} u\|_{L^p(\Omega)^d}^p$ and (11.12) is proved.
Relation (11.13) 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 (11.14). ∎

11.1.3 Properties of the MPFA-O gradient discretisation

Thanks to the previous lemmas, the proof of the properties of MPFA-O GDs is straightforward.

Theorem 11.5 (Properties of MPFA-O GDs). *Let $(\mathcal{D}_m)_{m\in\mathbb{N}}$ be a sequence of MPFA-O GDs, as in Section 11.1.1, 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 7.13, thanks to Lemma 11.4. The consistency is obtained by applying Proposition 7.36, thanks to Lemma 11.3. The piecewise constant reconstruction property is obvious from (11.3). ∎

The following two propositions, also direct consequences of results in the previous sections and in the appendix, are useful to establish precise error estimates for MPFA-O GSs.

Proposition 11.6 (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 11.1.1. Assume $p > d/2$ and take $\varrho \geq \theta_\mathfrak{T}$ (see (7.8)). Then there exists $C_{22} > 0$, depending only on Ω, p and ϱ, such that*

$$\forall \varphi \in W^{2,p}(\Omega) \cap W_0^{1,p}(\Omega)\,,\ S_\mathcal{D}(\varphi) \leq C_{22} h_\mathcal{M} \|\varphi\|_{W^{2,p}(\Omega)}\,,$$

where $S_\mathcal{D}$ is defined by (2.2).

Proof. For all $K \in \mathcal{M}$ and all $i \in I_K$, we have $\boldsymbol{x}_i \in \overline{K}$. By Lemmas 11.3 and B.1, the hypotheses of Proposition A.6 are satisfied with θ depending only on ϱ. This proposition yields the expected estimate on $S_\mathcal{D}$. ∎

Proposition 11.7 (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 11.1.1. 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_{23} depending only on Ω, p, and ϱ, such that*

$$C_\mathcal{D} \leq C_{23} \tag{11.15}$$

and, for all $\boldsymbol{\varphi} \in W^{1,p'}(\Omega)^d$,

$$W_\mathcal{D}(\boldsymbol{\varphi}) \leq C_{23} \|\boldsymbol{\varphi}\|_{W^{1,p'}(\Omega)^d} h_\mathcal{M}. \tag{11.16}$$

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 7.12 and Lemma 11.4. ∎

Note that the application of Lemmas 11.3 and 11.7 to the error estimate (3.6) and (3.7) in Theorem 3.2 provides an error in $h_{\mathcal{M}}$ in the case of a linear elliptic problem in one, two or three space dimensions, when the exact solution belongs to $H^2(\Omega)$.

11.2 MPFA-O methods for Neumann and Fourier boundary conditions

11.2.1 Neumann boundary conditions

We refer to Definition 7.52 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 11.1.1.
Defining $\mathfrak{T}'$ as in Lemma 11.4, for $v \in X_{\mathcal{D}} = X_{\mathfrak{T}'}$ such that $\|\nabla_{\mathcal{D}} v\|_{L^p(\Omega)^d} = 0$, Inequality (11.12) (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 (2.18) 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 $\mathfrak{T} = \mathfrak{T}'$):

$$\forall v \in X_{\mathcal{D}}\,,\ \forall \sigma \in \mathcal{F}_{\rm ext}\,,\ \forall \boldsymbol{s} \in \mathcal{V}_\sigma\ :\ \mathbb{T}_{\mathcal{D}} v = v_{\sigma,\boldsymbol{s}} \text{ on } \tau_{\sigma,\boldsymbol{s}}. \tag{11.17}$$

Since the regularity factor $\mathrm{reg}_{\rm LLE}(\mathcal{D})$ is defined as for Dirichlet boundary conditions, Lemma 11.3 still applies and show that this factor remains bounded if $\theta_{\mathfrak{T}}$ is bounded. Defining the control $\boldsymbol{\Phi} = \mathrm{Id} : X_{\mathcal{D}} \longrightarrow X_{\mathfrak{T}'}$ as in Lemma 11.4, we see that this lemma still holds and that $\omega^{\mathbb{T}}(\mathcal{D}, \mathfrak{T}', \boldsymbol{\Phi}) = 0$. Hence, Corollary 7.19 and Proposition 7.53 give the following theorem.

Theorem 11.8 (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 2.33, 2.27, 2.34 and 2.36. Moreover, each $\mathcal{D}_m$ has a piecewise constant reconstruction in the sense of Definition 2.10.

Proposition A.12 and Theorem 7.18 also give estimates on $S_{\mathcal{D}}$, $C_{\mathcal{D}}$ and $W_{\mathcal{D}}$ that are similar to those in Lemma 11.3 and Proposition 11.7. The constants only depend 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).

11.2.2 Fourier boundary conditions

Starting from an MPFA-O GD for Dirichlet boundary conditions, we follow Definition 7.55 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 (11.17) 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 7.56.
As noticed in Remark 7.21, the work done for Neumann boundary conditions then immediately show that Theorem 11.8 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 11.6 and 11.7.

12

Hybrid mimetic mixed schemes

Since the 50's, several schemes have been developed with the objective 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 13.2 and [37]) that design discrete operators and duality products to satisfy fully discrete calculus formula, MFD/CDO methods design discrete operators that satisfy a Stokes formula that involves both continuous and discrete functions. Depending on the choice of the location of the main degrees of freedom (faces or vertices), two different MFD/CDO families exist. We refer to [63] for a review on MFD methods, and to [11, 10] (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 [16, 15]. This requires to recast (3.1) 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 discretize this set of two equations. The resulting scheme takes a form that is apparently far from the GS (3.4). It was however proved in [35] that this form of hMFD can be actually embedded in a slightly larger family that also contains Hybrid Finite Volume (HFV, a particular case of "SUSHI") methods [49] and Mixed Finite Volume (MFV) methods [31, 32]. 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 (3.3) 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 [50] that this specific stabilisation term could be included in an augmented gradient, and thus that the HFV scheme is a GS. More surprisingly perhaps, [36] 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 we sections we detail the GD that leads to HMM methods when applied to linear diffusion equations, and we establish its proper-

ties. HMM methods correspond to LLE GDs. Following the nomenclature in Section 7.3.4, the general SUSHI methods 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 [30, 19, 48], but they are not GDMs and do not fully take advantage of the coercive gradient provided by HMM methods.

12.1 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.

12.1.1 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 7.50.

1. The geometrical entities attached to the DOFs 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

$$\pi_K^K = 1 \text{ and } \pi_K^\sigma = 0 \text{ for all } \sigma \in \mathcal{F}_K. \tag{12.1}$$

   Recalling the definition (7.7c) of $\mathit{\Pi}_\mathfrak{T}$, (7.33) therefore reads

$$\forall v \in X_\mathcal{D}\,,\ \forall K \in \mathcal{M}\,,\ \text{for a.e. } \boldsymbol{x} \in K\,,\ \mathit{\Pi}_\mathcal{D} v(\boldsymbol{x}) = \mathit{\Pi}_\mathfrak{T} v(\boldsymbol{x}) = v_K. \tag{12.2}$$

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 $(\mathcal{G}_K^i)_{i\in I_K}$. The polytopal gradient defined by (7.7e), that is,

$$\overline{\nabla}_K v = \frac{1}{|K|} \sum_{\sigma\in\mathcal{F}_K} |\sigma| v_\sigma \boldsymbol{n}_{K,\sigma}, \tag{12.3}$$

is $\mathbb{P}_1$-exact (Lemma B.6), but not "strong enough" to control all the DOFs in $X_{\mathfrak{T},0}$ (Remark 7.9). 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 DOFs in K and define, for $v \in X_K$, the function $\mathcal{G}_K v \in L^p(K)^d$ by

$$\begin{aligned} &\forall \sigma \in \mathcal{F}_K, \text{ for a.e. } \boldsymbol{x} \in D_{K,\sigma}, \\ &\mathcal{G}_K v(\boldsymbol{x}) = \overline{\nabla}_K v + \frac{\sqrt{d}}{d_{K,\sigma}} [\mathcal{L}_K R_K(v)]_\sigma \boldsymbol{n}_{K,\sigma}, \end{aligned} \tag{12.4}$$

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

$$\begin{aligned} &R_K(v) = (R_{K,\sigma}(v))_{\sigma\in\mathcal{F}_K} \text{ with} \\ &R_{K,\sigma}(v) = v_\sigma - v_K - \overline{\nabla}_K v \cdot (\overline{\boldsymbol{x}}_\sigma - \boldsymbol{x}_K), \end{aligned} \tag{12.5}$$

- $\mathcal{L}_K$ is an isomorphism of the vector space $\mathrm{Im}(R_K)$.

The gradient reconstruction $\nabla_\mathcal{D}$ is then defined by (7.34), which simply gives

$$\begin{aligned} &\forall v \in X_\mathcal{D}, \forall K \in \mathcal{M}, \forall \sigma \in \mathcal{F}_K, \text{ for a.e. } \boldsymbol{x} \in D_{K,\sigma}, \\ &\nabla_\mathcal{D} v(\boldsymbol{x}) = \overline{\nabla}_K v + \frac{\sqrt{d}}{d_{K,\sigma}} [\mathcal{L}_K R_K(v)]_\sigma \boldsymbol{n}_{K,\sigma}. \end{aligned} \tag{12.6}$$

The functions $(\mathcal{G}_K^i)_{i\in I_K}$ of $L^p(K)^d$ can be recovered through $\mathcal{G}_K$ defined by (12.4). 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,

$$\mathcal{G}_K^K = \mathcal{G}_K v^K \text{ and } \mathcal{G}_K^\sigma = \mathcal{G}_K v^\sigma \text{ for all } \sigma \in \mathcal{F}_K. \tag{12.7}$$

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

$$\forall g \in W^{1-\frac{1}{p},p}(\partial\Omega), \forall \sigma \in \mathcal{F}_{\mathrm{ext}}, (\mathcal{I}_{\mathcal{D},\partial} g)_\sigma = \frac{1}{|\sigma|} \int_\sigma g(\boldsymbol{x}) \mathrm{d}s(\boldsymbol{x}). \tag{12.8}$$

5. Lemma 12.8 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 12.1 (Hybrid method)
The face degree of freedom v_σ corresponds to the hybridisation of the hMFD methods.

Remark 12.2 (Simpler trace interpolation)
As explained in Remark 2.19, simpler trace interpolations can be considered the boundary condition g of the considered problem (e.g. in (3.22b)) is more regular that $W^{1-\frac{1}{p},p}(\partial\Omega)$. 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 [35]. 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 (3.25) of the linear problem (3.22) with $\mathbf{F} = 0$ can be written (see [35] in the case $g = 0$):

$$\begin{aligned} &\text{Find } u \in \mathcal{I}_{\mathcal{D},\partial} g + X_{\mathcal{D},0} \text{ such that, for all } v \in X_{\mathcal{D},0},\\ &\sum_{K\in\mathcal{M}} |K| \Lambda_K \overline{\nabla}_K u \cdot \overline{\nabla}_K v + \sum_{K\in\mathcal{M}} R_K(v)^T \mathbb{B}_K R_K(u)\\ &\qquad = \sum_{K\in\mathcal{M}} v_K \int_K f(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}, \end{aligned} \tag{12.9}$$

where Λ_K is the constant value of Λ on K (we assume that Λ is piecewise constant on $\mathcal{M}$ – see Remark 12.16 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 12.3 ($\mathbb{R}^{\mathrm{Card}(\mathcal{F}_K)}$ 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 (12.15) 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 12.4. *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 $K \in \mathcal{M}$,*

$$\overline{\nabla}_K v = \frac{1}{|K|} \int_K \nabla_{\mathcal{D}} v(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}. \tag{12.10}$$

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 (12.6), 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

$$\begin{aligned}\int_K \nabla_{\mathcal{D}}v(\boldsymbol{x})\mathrm{d}\boldsymbol{x} &= \sum_{\sigma\in\mathcal{F}_K} |D_{K,\sigma}|(\nabla_{\mathcal{D}}v)_{|D_{K,\sigma}}\\ &= |K|\overline{\nabla}_K v + \frac{1}{\sqrt{d}}\sum_{\sigma\in\mathcal{F}_K} |\sigma|[\mathcal{L}_K(R_K(v))]_\sigma\boldsymbol{n}_{K,\sigma}. \qquad (12.11)\end{aligned}$$

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 (12.11) vanishes. This proves that $\int_K \nabla_{\mathcal{D}}v(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = |K|\overline{\nabla}_K v$ as claimed. ∎

Proposition 12.5 (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* (3.4) *(with $\boldsymbol{F} = 0$) is the HMM scheme* (12.9) *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 (12.2) of $\Pi_{\mathcal{D}}$, the right-hand sides of (3.4) and (12.9) 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 $(u, v) \in X_{\mathcal{D}}^2$,

$$|K|\Lambda_K\overline{\nabla}_K u \cdot \overline{\nabla}_K v + R_K(v)^T\mathbb{B}_K R_K(u) = \int_K \Lambda_K \nabla_{\mathcal{D}} u(\boldsymbol{x}) \cdot \nabla_{\mathcal{D}} v(\boldsymbol{x})\mathrm{d}\boldsymbol{x}. \quad (12.12)$$

Summing (12.12) over $K \in \mathcal{M}$ then shows that the left-hand sides of (3.4) and (12.9) are identical.
Recall the definition (12.6) of $\nabla_{\mathcal{D}}$ and use $\sum_{\sigma\in\mathcal{F}_K} |D_{K,\sigma}| = |K|$ to write, by developing the scalar product,

$$\begin{aligned}\int_K \Lambda_K \nabla_{\mathcal{D}} u(\boldsymbol{x}) \cdot \nabla_{\mathcal{D}} v(\boldsymbol{x})\mathrm{d}\boldsymbol{x} &= \sum_{\sigma\in\mathcal{F}_K} |D_{K,\sigma}|\Lambda_K(\nabla_{\mathcal{D}} u)_{|K} \cdot (\nabla_{\mathcal{D}} v)_{|K} \\ &= |K|\Lambda_K\overline{\nabla}_K u \cdot \overline{\nabla}_K v \\ &\quad + \Lambda_K\overline{\nabla}_K u \cdot \sum_{\sigma\in\mathcal{F}_K} |D_{K,\sigma}|\frac{\sqrt{d}}{d_{K,\sigma}}[\mathcal{L}_K R_K(v)]_\sigma \boldsymbol{n}_{K,\sigma} \quad (12.13) \\ &\quad + \overline{\nabla}_K v \cdot \Lambda_K \sum_{\sigma\in\mathcal{F}_K} |D_{K,\sigma}|\frac{\sqrt{d}}{d_{K,\sigma}}[\mathcal{L}_K R_K(u)]_\sigma \boldsymbol{n}_{K,\sigma} \quad (12.14) \\ &\quad + \sum_{\sigma\in\mathcal{F}_K} |D_{K,\sigma}|\frac{d}{d_{K,\sigma}^2}\Lambda_K \boldsymbol{n}_{K,\sigma} \cdot \boldsymbol{n}_{K,\sigma}[\mathcal{L}_K R_K(u)]_\sigma[\mathcal{L}_K R_K(v)]_\sigma.\end{aligned}$$

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 12.4 shows that (12.13) and (12.14) vanish. Hence

$$\int_K \Lambda_K \nabla_{\mathcal{D}} u(\boldsymbol{x}) \cdot \nabla_{\mathcal{D}} v(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = |K|\Lambda_K\overline{\nabla}_K u \cdot \overline{\nabla}_K v + [\mathcal{L}_K R_K(v)]^T\mathbb{D}_K[\mathcal{L}_K R_K(u)] \quad (12.15)$$

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 (12.12) therefore holds provided that, for all $(\xi,\eta)\in(\mathrm{Im}(R_K))^2$,

$$\xi^T\mathbb{B}_K\eta = (\mathcal{L}_K(\xi))^T\mathbb{D}_K(\mathcal{L}_K(\eta)). \tag{12.16}$$

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 12.6 below then satisfies (12.17), which is precisely (12.16) with $x=\xi$ and $y=\eta$. ∎

Lemma 12.6. *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*

$$\text{for all } (x,y)\in E^2,\ \langle x,y\rangle_1 = \langle\mathcal{L}x,\mathcal{L}y\rangle_2. \tag{12.17}$$

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 $Y_e^TM_eX_e = (\mathcal{L}_eY_e)^T(\mathcal{L}_eX_e)$ translates into $\langle x,y\rangle_1 = \langle\mathcal{L}x,\mathcal{L}y\rangle_2$. ∎

Remark 12.7 (Elimination of the cell DOFs in the HMM GS by static condensation) By static condensation, the cell degrees of freedom can be eliminated when an HMM method is applied to a linear elliptic equation. This is done by taking, in (12.9), 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 (12.3) 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, (12.9) 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 (12.9) and provides a symmetric positive definite system only on $(u_\sigma)_{\sigma\in\mathcal{F}_{\mathrm{int}}}$.

12.1.2 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 12.1.1, we define the following measure of the invertibility properties of the isomorphisms $(\mathcal{L}_K)_{K\in\mathcal{M}}$:

$$\zeta_{\mathcal{D}} = \min\Bigg\{\zeta>0\,:\,\forall K\in\mathcal{M},\ \forall v\in X_K\,,\\ \zeta^{-1}\sum_{\sigma\in\mathcal{F}_K}|D_{K,\sigma}|\left|\frac{R_{K,\sigma}(v)}{d_{K,\sigma}}\right|^p \le \sum_{\sigma\in\mathcal{F}_K}|D_{K,\sigma}|\left|\frac{[\mathcal{L}_K R_K(v)]_\sigma}{d_{K,\sigma}}\right|^p \\ \le \zeta\sum_{\sigma\in\mathcal{F}_K}|D_{K,\sigma}|\left|\frac{R_{K,\sigma}(v)}{d_{K,\sigma}}\right|^p\Bigg\}. \tag{12.18}$$

The simplest way to choose $\mathcal{L}_K$ that satisfies the inequality in (12.18) is to take $\mathcal{L}_K = \beta_K\mathrm{Id}$, where $\beta_K\in[\zeta^{-1},\zeta]$. This corresponds to the original HFV method.
The following lemma states that HMM GDs are LLE GDs, and gives a control of their regularity $\mathrm{reg}_{\mathrm{LLE}}$ in terms of $\zeta_{\mathcal{D}}$ and geometric regularity factors.

Lemma 12.8 (Estimate on $\mathrm{reg}_{\mathbf{LLE}}(\mathcal{D})$ 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 12.1.1. 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 $\varrho\ge\theta_{\mathfrak{T}}+\zeta_{\mathcal{D}}$ (see (7.8) and (12.18)), there exists C_{24}, depending only on p, d and ϱ, such that

$$\mathrm{reg}_{\mathrm{LLE}}(\mathcal{D})\le C_{24}. \tag{12.19}$$

Proof. Let $K\in\mathcal{M}$. According to (12.1), $\sum_{i\in I_K}\pi_K^i=\pi_K^K=1$ so π_K is a $\mathbb{P}_0$-exact function reconstruction.
Lemma B.6 shows that $\overline{\nabla}_K$ is $\mathbb{P}_1$-exact gradient reconstruction. Hence, if v interpolates an affine mapping A at the approximation points $(\boldsymbol{x}_K,(\overline{\boldsymbol{x}}_\sigma)_{\sigma\in\mathcal{F}_K})$, (12.5) 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 (12.10) shows that $\overline{\nabla}_K v=0$ for all $K\in\mathcal{M}$ and thus, by (12.4) 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 (C.12) and the definitions (12.4) and (12.18) of $\mathcal{G}_K$ and $\zeta_\mathcal{D}$,

$$\begin{aligned}\|\mathcal{G}_K v\|_{L^p(K)^d}^p &= \sum_{\sigma\in\mathcal{F}_K} |D_{K,\sigma}|\,|(\mathcal{G}_K v)_{|D_{K,\sigma}}|^p\\ &\le 2^{p-1}\left(|K|\left|\overline{\nabla}_K v\right|^p + d^{\frac{p}{2}}\sum_{\sigma\in\mathcal{F}_K}|D_{K,\sigma}|\left|\frac{[\mathcal{L}_K R_K(v)]_\sigma}{d_{K,\sigma}}\right|^p\right)\\ &\le 2^{p-1}\left(|K|\left|\overline{\nabla}_K v\right|^p + \zeta_\mathcal{D} d^{\frac{p}{2}}\sum_{\sigma\in\mathcal{F}_K}|D_{K,\sigma}|\left|\frac{R_{K,\sigma}(v)}{d_{K,\sigma}}\right|^p\right).\end{aligned} \tag{12.20}$$

Set $V = \max\{|v_\sigma - v_K| : \sigma\in\mathcal{F}_K\}$. Since $\theta_\mathfrak{T}\frac{d_{K,\sigma}}{h_K}\ge 1$ and $\sum_{\sigma\in\mathcal{F}_K}|\sigma|d_{K,\sigma} = d|K|$ (see (B.1)), the definition (7.7e) of $\overline{\nabla}_K$ shows that

$$\left|\overline{\nabla}_K v\right| \le \frac{1}{|K|}\sum_{\sigma\in\mathcal{F}_K}|\sigma|V \le \frac{\theta_\mathfrak{T}}{h_K}V\frac{1}{|K|}\sum_{\sigma\in\mathcal{F}_K}|\sigma|d_{K,\sigma} \le d\theta_\mathfrak{T}\frac{V}{h_K}. \tag{12.21}$$

Using the definition (12.5) of $R_{K,\sigma}$ leads to

$$\left|\frac{R_{K,\sigma}(v)}{d_{K,\sigma}}\right| \le \frac{\theta_\mathfrak{T}}{h_K}|R_{K,\sigma}(v)| \le \frac{\theta_\mathfrak{T}}{h_K}(V + d\theta_\mathfrak{T}V). \tag{12.22}$$

Plugging (12.21) and (12.22) into (12.20) gives $C_{25} > 0$, depending only on ϱ, p and d, such that

$$\|\mathcal{G}_K v\|_{L^p(K)^d} \le C_{25}|K|^{1/p}h_K^{-1}\max\{|v_\sigma - v_K| : \sigma\in\mathcal{F}_K\}. \tag{12.23}$$

Applied to $v = v^K$ or $v = v^\sigma$ (defined in Item 3 of Section 12.1.1), this shows that $\left\|\mathcal{G}_K^i\right\|_{L^p(K)^d} \le C_{25}|K|^{1/p}h_K^{-1}$ for all $i\in I_K$. Recalling the definition (7.26) of $\|\mathcal{G}_K\|_p$, we infer that

$$\|\mathcal{G}_K\|_p \le C_{25}(1 + \mathrm{Card}(\mathcal{F}_K)) \le C_{25}(1+\varrho)$$

and the proof of (12.19) is complete. ∎

Lemma 12.9. *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 12.1.1. We take $\varrho \ge \theta_\mathfrak{T} + \zeta_\mathcal{D}$ (see (7.8) and (12.18)). Then, there exists $C_{26} > 0$ depending only on Ω, p and ϱ, such that*

$$\forall v\in X_\mathcal{D},\ |v|_{\mathfrak{T},p} \le C_{26}\left\|\nabla_\mathcal{D} v\right\|_{L^p(\Omega)^d}. \tag{12.24}$$

Proof. In this proof, $A \lesssim B$ means that $A \le MB$ for some M depending only on Ω, p and ϱ. Let $v\in X_\mathcal{D}$. Relation (12.10) and Jensen's inequality give

$$|\overline{\nabla}_K v|^p \le \frac{1}{|K|}\int_K |\nabla_{\mathcal{D}} v(\boldsymbol{x})|^p \mathrm{d}\boldsymbol{x}. \tag{12.25}$$

By definition (12.6) of $\nabla_{\mathcal{D}} v$ and by the power-of-sums inequality (C.12), we deduce that, for a.e. $\boldsymbol{y} \in D_{K,\sigma}$,

$$\begin{aligned}\left|\frac{\sqrt{d}}{d_{K,\sigma}}[\mathcal{L}_K R_K(v)]_\sigma \boldsymbol{n}_{K,\sigma}\right|^p &= \left|\nabla_{\mathcal{D}} v(\boldsymbol{y}) - \overline{\nabla}_K v\right|^p \\ &\lesssim |\nabla_{\mathcal{D}} v(\boldsymbol{y})|^p + \frac{1}{|K|}\int_K |\nabla_{\mathcal{D}} v(\boldsymbol{x})|^p \mathrm{d}\boldsymbol{x}.\end{aligned}$$

Integrating over $\boldsymbol{y} \in D_{K,\sigma}$ and summing over $\sigma \in \mathcal{F}_K$ leads to

$$\sum_{\sigma\in\mathcal{F}_K} |D_{K,\sigma}| \left|\frac{[\mathcal{L}_K R_K(v)]_\sigma}{d_{K,\sigma}}\right|^p \lesssim \int_K |\nabla_{\mathcal{D}} v(\boldsymbol{x})|^p \mathrm{d}\boldsymbol{x}. \tag{12.26}$$

Use then the definition (12.18) of $\zeta_{\mathcal{D}}$ to write

$$\sum_{\sigma\in\mathcal{F}_K} |D_{K,\sigma}| \left|\frac{R_{K,\sigma}(v)}{d_{K,\sigma}}\right|^p \lesssim \int_K |\nabla_{\mathcal{D}} v(\boldsymbol{x})|^p \mathrm{d}\boldsymbol{x}. \tag{12.27}$$

By definition (12.5) 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 (12.25) and using (12.27),

$$\sum_{\sigma\in\mathcal{F}_K} |D_{K,\sigma}| \left|\frac{v_\sigma - v_K}{d_{K,\sigma}}\right|^p \lesssim \int_K |\nabla_{\mathcal{D}} v(\boldsymbol{x})|^p \mathrm{d}\boldsymbol{x}.$$

Since $|D_{K,\sigma}| = \frac{|\sigma| d_{K,\sigma}}{d}$ (cf. (B.1)), summing this relation over $K \in \mathcal{M}$ and recalling the definition (7.7f) of $|\cdot|_{\mathfrak{T},p}$ proves (12.24). ∎

Remark 12.10 (Converse to (12.24)*)*
Under more restrictive hypotheses on the mesh, [49] also proves that $\|\nabla_{\mathcal{D}} v\|_{L^p(\Omega)^d} \le C_{26} |v|_{\mathfrak{T},p}$. This inequality is however not useful for the analysis of HMM GDs.

We can now define, and state estimates on, a control of an HMM GD by a polytopal toolbox.

Lemma 12.11 (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 12.1.1. Take $\varrho \ge \theta_{\mathfrak{T}} + \zeta_{\mathcal{D}}$ (see* (7.8) *and* (12.18)*) and define the control $\boldsymbol{\Phi} = \mathrm{Id} : X_{\mathcal{D},0} \longrightarrow X_{\mathfrak{T},0}$ of $\mathcal{D}$ by $\mathfrak{T}$ (see Definition 7.10). Then, there exists $C_{26} > 0$ depending only on Ω, p and ϱ, such that*

$$\|\boldsymbol{\Phi}\|_{\mathcal{D},\mathfrak{T}} \le C_{26}, \tag{12.28}$$

and

$$\omega^{\Pi}(\mathcal{D},\mathfrak{T},\boldsymbol{\Phi}) = 0\,, \quad \omega^{\nabla}(\mathcal{D},\mathfrak{T},\boldsymbol{\Phi}) = 0. \tag{12.29}$$

Proof. Estimate (12.28) is given by Lemma 12.9. The first relation in (12.29) follows from $\Pi_{\mathcal{D}} v = \Pi_{\mathfrak{T}} v = \Pi_{\mathfrak{T}} \boldsymbol{\Phi}(v)$ (see (12.2)). The second relation in (12.29) is a straightforward consequence of (12.10). ∎

12.1.3 Properties of HMM gradient discretisations

Thanks to the previous lemmas, the proof of the properties of HMM GDs is straightforward.

Theorem 12.12 (Properties of HMM GDs for Dirichlet BCs). *Let $(\mathcal{D}_m)_{m\in\mathbb{N}}$ be a sequence of HMM GDs, as in Section 12.1.1, 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 (12.18)), and that $h_{\mathcal{M}_m} \to 0$ as $m \to \infty$. Also assume that*

$$\sup_{m\in\mathbb{N}} \max \left\{ \frac{|K|}{h_K |\sigma|} \ : \ K \in \mathcal{M}_m,\, \sigma \in \mathcal{F}_K \right\} < +\infty. \tag{12.30}$$

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, 2.20, 2.6 and 2.8. Moreover, each $\mathcal{D}_m$ has a piecewise constant reconstruction in the sense of Definition 2.10.

Remark 12.13. The condition (12.30) is only useful to establish that (2.16) holds, which is used to prove the GD-consistency for non-homogeneous Dirichlet boundary conditions. In particular, for homogeneous Dirichlet BCs, the theorem holds without assuming (12.30).

Proof. The limit-conformity, coercivity and compactness are obtained by applying Corollary 7.13, thanks to Lemma 12.11. The property of piecewise constant reconstruction is also straightforward from (12.2) (using the notations in Definition 2.10, 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 7.51. The bound on $\mathrm{reg}_{\mathrm{LLE}}(\mathcal{D}_m)$ being provided by Lemma 12.8, we just have to check that (7.61) and (2.16) hold.
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 (12.8) of $\mathcal{I}_{\mathcal{D},\partial}$, since $\frac{1}{|\sigma|}\int_\sigma \boldsymbol{x}\mathrm{d}s(\boldsymbol{x}) = \overline{\boldsymbol{x}}_\sigma$,

$$|(\mathcal{I}_{\mathcal{D},\partial}\gamma(\varphi))_\sigma - \varphi(\overline{\boldsymbol{x}}_\sigma)| \le \frac{1}{2}\mathrm{diam}(\sigma)^2 \sup_{\overline{\Omega}} |D^2\varphi|.$$

Since $\mathrm{diam}(\sigma) \le \mathrm{diam}(K)$ for any $\sigma \in \mathcal{F}_K$, this proves that (7.61) holds.

It remains to prove (2.16). In the following, $A \lesssim B$ means that $A \le MB$ for some M not depending on m or the considered elements of $X_{\mathcal{D}_m,0}$. We also drop the index m. Let $\varphi \in W^{1,p}(\Omega)$ and define $v \in X_{\mathcal{D}}$ by

$$v_K = \frac{1}{|K|}\int_K \varphi(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \quad \text{and} \quad v_\sigma = \frac{1}{|\sigma|}\int_\sigma \varphi(\boldsymbol{x})\mathrm{d}s(\boldsymbol{x}).$$

We clearly have $v - \mathcal{I}_{\mathcal{D},\partial}\gamma(\varphi) \in X_{\mathcal{D},0}$. By Jensen's inequality, $|v_K|^p \le \frac{1}{|K|}\int_K |\varphi(\boldsymbol{x})|^p \mathrm{d}\boldsymbol{x}$ and therefore, multiplying by $|K|$ and summing over $K \in \mathcal{M}$,

$$\|\Pi_{\mathcal{D}} v\|_{L^p(\Omega)} \le \|\varphi\|_{L^p(\Omega)}. \tag{12.31}$$

Estimate (B.11), with $p = 1$, in Lemma B.7 (p.376) yields

$$|v_\sigma - v_K| \lesssim \frac{1}{|\sigma|}\int_K |\nabla\varphi(\boldsymbol{x})|\mathrm{d}\boldsymbol{x}.$$

Plugged into (12.23) this shows that

$$\|\mathcal{G}_K v\|_{L^p(K)^d} \lesssim \max_{\sigma\in\mathcal{F}_K} \frac{|K|^{1/p}}{h_K|\sigma|}\int_K |\nabla\varphi(\boldsymbol{x})|\mathrm{d}\boldsymbol{x}.$$

Raise to the power p, use Hölder's inequality (C.7) and the fact that $|K| \lesssim h_K|\sigma|$ for all $\sigma \in \mathcal{F}_K$ to obtain

$$\|\nabla_{\mathcal{D}} v\|^p_{L^p(K)^d} \lesssim \max_{\sigma\in\mathcal{F}_K} \frac{|K|^p}{h_K^p|\sigma|^p}\int_K |\nabla\varphi(\boldsymbol{x})|^p\mathrm{d}\boldsymbol{x} \lesssim \int_K |\nabla\varphi(\boldsymbol{x})|^p\mathrm{d}\boldsymbol{x}.$$

Summing this relation over $K \in \mathcal{M}$ gives $\|\nabla_{\mathcal{D}} v\|_{L^p(\Omega)^d} \lesssim \|\nabla\varphi\|_{L^p(\Omega)^d}$. Combined with (12.31), this establishes (2.16) and concludes the proof of the GD-consistency of $(\mathcal{D}_m)_{m\in\mathbb{N}}$. ∎

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 12.14 (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 12.1.1. Assume that $p > d/2$ and take $\varrho \ge \theta_{\mathfrak{T}} + \zeta_{\mathcal{D}}$ (see* (7.8) *and* (12.18)*). Let $\varphi \in W^{2,p}(\Omega)$ and, as in Remark 12.2, 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 $\sigma \in \mathcal{F}_{\mathrm{ext}}$ (this makes sense since $\varphi \in C(\overline{\Omega})$). Then, there exists $C_{27} > 0$, depending only on Ω, p and ϱ, such that*

$$S_{\mathcal{D}}(\varphi) \le C_{27} h_{\mathcal{M}} \|\varphi\|_{W^{2,p}(\Omega)},$$

where $S_{\mathcal{D}}$ is defined by (2.14).

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 (A.15) holds. Using Lemma 12.8, Proposition A.10 can be applied and the result follows immediately. ∎

Proposition 12.15 (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 12.1.1. Take $\varrho \geq \theta_{\mathfrak{T}} + \eta_{\mathfrak{T}} + \zeta_{\mathcal{D}}$ (see* (7.8)*,* (7.9) *and* (12.18)*). Then, there exists C_{28} depending only on Ω, p, any ϱ, such that*

$$C_{\mathcal{D}} \leq C_{28} \tag{12.32}$$

and

$$\forall \boldsymbol{\varphi} \in W^{1,p'}(\Omega)^d\,,\ W_{\mathcal{D}}(\boldsymbol{\varphi}) \leq C_{28} h_{\mathcal{M}} \left\|\boldsymbol{\varphi}\right\|_{W^{1,p'}(\Omega)^d}. \tag{12.33}$$

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 7.12 and Lemma 12.11. ∎

Note that the application of Propositions 12.14 and 12.15 to the error estimate (3.6) and (3.7) in Theorem 3.2 provides an $\mathcal{O}(h_{\mathcal{M}})$ error in the case of a linear elliptic problem in two or three space dimensions, when the solution belongs to $H^2(\Omega)$.

Remark 12.16 (Non piecewise constant diffusion tensor)
If Λ is not piecewise constant on $\mathcal{M}$, then (12.9) is the GS (3.4) for the problem (3.3) with Λ is replaced with its piecewise projection on the mesh, *i.e.* (12.9) is the GS for

$$\begin{aligned}&\overline{u}_{\mathcal{M}} \in H^1_0(\Omega),\ \forall v \in H^1_0(\Omega),\\ &\int_\Omega \Lambda_{\mathcal{M}}(\boldsymbol{x})\nabla\overline{u}_{\mathcal{M}}(\boldsymbol{x})\cdot\nabla v(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = \int_\Omega f(\boldsymbol{x})v(\boldsymbol{x})\mathrm{d}\boldsymbol{x} - \int_\Omega \boldsymbol{F}(\boldsymbol{x})\cdot\nabla v(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\end{aligned}$$

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 $\|\Lambda - \Lambda_{\mathcal{M}}\|_{L^\infty(\Omega)} \leq C h_{\mathcal{M}}$ and thus, denoting by $\overline{u}$ the solution to (3.3), 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

$$\begin{aligned}\underline{\lambda}\left\|\nabla\overline{u}_{\mathcal{M}} - \nabla\overline{u}\right\|^2_{L^2(\Omega)^d} &\leq \int_\Omega \Lambda_{\mathcal{M}}(\boldsymbol{x})(\nabla\overline{u}_{\mathcal{M}} - \nabla\overline{u})(\boldsymbol{x})\cdot(\nabla\overline{u}_{\mathcal{M}} - \nabla\overline{u})(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\\ &= \int_\Omega (\Lambda(\boldsymbol{x}) - \Lambda_{\mathcal{M}}(\boldsymbol{x}))\nabla\overline{u}(\boldsymbol{x})\cdot(\nabla\overline{u}_{\mathcal{M}} - \nabla\overline{u})(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\\ &\leq C h_{\mathcal{M}}\left\|\nabla\overline{u}\right\|_{L^2(\Omega)}\left\|\nabla\overline{u}_{\mathcal{M}} - \nabla\overline{u}\right\|_{L^2(\Omega)^d}.\end{aligned}$$

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 (12.9), the estimates in Chapter 3 and in Propositions 12.14 and 12.15

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 (3.3), 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.

12.2 HMM methods for Neumann and Fourier boundary conditions

12.2.1 Neumann boundary conditions

Following Definition 7.52, 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 12.1.1.
If $v \in X_{\mathcal{D}} = X_{\mathfrak{T}}$ and $\|\nabla_{\mathcal{D}} v\|_{L^p(\Omega)^d} = 0$, then Inequality (12.24) 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 (2.18) 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,

$$\forall v \in X_{\mathcal{D}}\,,\ \forall \sigma \in \mathcal{F}_{\rm ext}\ :\ \mathbb{T}_{\mathcal{D}} v = \mathbb{T}_{\mathfrak{T}} v = v_\sigma \text{ on } \sigma. \tag{12.34}$$

Since the regularity factor $\mathrm{reg}_{\rm LLE}(\mathcal{D})$ is defined as for Dirichlet boundary conditions, Lemma 12.8 still applies and shows that this factor remains bounded if $\theta_{\mathfrak{T}}$ and $\zeta_{\mathcal{D}}$ are bounded. Defining the control $\boldsymbol{\Phi} = \mathrm{Id} : X_{\mathcal{D}} \longrightarrow X_{\mathfrak{T}}$ of an HMM GD $\mathcal{D}$ for Neumann boundary conditions by $\mathfrak{T}$, we see that Lemma 12.11 still holds and that $\omega^{\mathbb{T}}(\mathcal{D}, \mathfrak{T}, \boldsymbol{\Phi}) = 0$. Hence, Corollary 7.19 and Proposition 7.53 give the following theorem.

Theorem 12.17 (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 (12.18)), 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 2.33, 2.27, 2.34 and 2.36. Moreover, each $\mathcal{D}_m$ has a piecewise constant reconstruction in the sense of Definition 2.10.

Proposition A.12 and Theorem 7.18 also give estimates on $S_{\mathcal{D}}$, $C_{\mathcal{D}}$ and $W_{\mathcal{D}}$ that are similar to those in Propositions 12.14 and 12.15. The constants only depend 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 $W_{\mathcal{D}}$).

12.2.2 Fourier boundary conditions

Starting from an HMM GD for Dirichlet boundary conditions, we follow Definition 7.55 in Section 7.3.6 to define an HMM GD for Fourier boundary conditions.
The boundary mesh $\mathcal{M}_\partial$ is simply $\mathcal{F}_{\rm ext}$, and the reconstructed trace (12.34) corresponds to $I_\sigma = \{\sigma\}$ and $\pi^\sigma_\sigma = 1$. The bound on $\mathrm{reg}_{\rm LLE}(\mathcal{D})$ 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 7.56.
As noticed in Remark 7.21, the work done for Neumann boundary conditions then immediately show that Theorem 12.17 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 12.14 and 12.15.

12.3 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

$$\begin{aligned}&\forall u, v \in X_\mathcal{D},\ \forall K \in \mathcal{M}\ :\\ &\sum_{\sigma\in\mathcal{F}_K} F_{K,\sigma}(u)(v_K - v_\sigma) = \int_K \Lambda(\boldsymbol{x})\nabla_\mathcal{D} u(\boldsymbol{x})\cdot\nabla_\mathcal{D} v(\boldsymbol{x})\mathrm{d}\boldsymbol{x}.\end{aligned} \tag{12.35}$$

The existence and uniqueness of these fluxes is ensured by the following proposition.

Proposition 12.18 (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* (12.35).

Proof. Let $u \in X_\mathcal{D}$ and assume that $(F_{K,\sigma}(u))_{K,\sigma}$ a solution of (12.35). 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 (12.35) gives

$$F_{K,\sigma}(u) = -\int_K \Lambda(\boldsymbol{x})\nabla_\mathcal{D} u(\boldsymbol{x})\cdot\nabla_\mathcal{D} w^\sigma(\boldsymbol{x})\mathrm{d}\boldsymbol{x}, \tag{12.36}$$

which determines uniquely $F_{K,\sigma}(u)$, since w^σ only depends 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 (12.36) satisfy (12.35). Fix a cell $K \in \mathcal{M}$ and let $v \in X_\mathfrak{T}$. Multiplying (12.36) by $v_K - v_\sigma$ and summing on $\sigma \in \mathcal{F}_K$ gives

$$\sum_{\sigma\in\mathcal{F}_K} F_{K,\sigma}(u)(v_K - v_\sigma)$$

$$= \int_K \Lambda(\boldsymbol{x})\nabla_{\mathcal{D}}u(\boldsymbol{x})\cdot\nabla_{\mathcal{D}}\left(\sum_{\sigma\in\mathcal{F}_K}(v_\sigma - v_K)w^\sigma\right)(\boldsymbol{x})\mathrm{d}\boldsymbol{x}$$

$$= \int_K \Lambda(\boldsymbol{x})\nabla_{\mathcal{D}}u(\boldsymbol{x})\cdot\nabla_{\mathcal{D}}V(\boldsymbol{x})\mathrm{d}\boldsymbol{x}, \tag{12.37}$$

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 (12.4) and (12.5) we see that $\nabla_{\mathcal{D}}V = \nabla_{\mathcal{D}}v$ on K. Equation (12.37) therefore shows that (12.35) is satisfied. ∎

The GS for (3.1) (with $\boldsymbol{F} = 0$) then corresponds to writing the flux conservativity and flux balances (see [35]):

$$\forall\sigma \in \mathcal{F}_{\mathrm{int}} \,:\, F_{K,\sigma}(u) + F_{L,\sigma}(u) = 0, \tag{12.38}$$

$$\forall K \in \mathcal{M} \,:\, \sum_{\sigma\in\mathcal{F}_K} F_{K,\sigma}(u) = \int_K f(\boldsymbol{x})\mathrm{d}\boldsymbol{x}. \tag{12.39}$$

The HMM method is therefore a Finite Volume scheme (more precisely, the Mixed Finite Volume scheme, see [35]).

For specific meshes and with $\Lambda = \mathrm{Id}$, the flux $F_{K,\sigma}(u)$ actually only depends on the values u_K and u_σ.

Lemma 12.19 (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:*

$$\forall K \in \mathcal{M}\,,\ \forall\sigma \in \mathcal{F}_K \,:\, \boldsymbol{n}_{K,\sigma} = \frac{\overline{\boldsymbol{x}}_\sigma - \boldsymbol{x}_K}{d_{K,\sigma}} \tag{12.40}$$

*(*i.e. *the orthogonal projection of $\boldsymbol{x}_K$ on each face $\sigma \in \mathcal{F}_K$ is the center of mass $\overline{\boldsymbol{x}}_\sigma$ of σ). Then the following property holds:*

$$\int_K \nabla_{\mathcal{D}}u(\boldsymbol{x})\cdot\nabla_{\mathcal{D}}u(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = \sum_{\sigma\in\mathcal{F}_K}\frac{|\sigma|}{d_{K,\sigma}}(u_K - u_\sigma)(v_K - v_\sigma). \tag{12.41}$$

Hence, if $\Lambda_K = \mathrm{Id}$ and $\mathcal{L}_K = \mathrm{Id}$, the fluxes defined by (12.35) *are given by*

$$F_{K,\sigma}(u) = \frac{|\sigma|}{d_{K,\sigma}}(u_K - u_\sigma).$$

A similar lemma can be proved [35] 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 center 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 center 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 (12.15) give

$$\int_K \nabla_{\mathcal{D}} u(\boldsymbol{x}) \cdot \nabla_{\mathcal{D}} v(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} = |K| \overline{\nabla}_K u \cdot \overline{\nabla}_K v + \sum_{\sigma \in \mathcal{F}_K} \frac{|\sigma|}{d_{K,\sigma}} R_{K,\sigma}(u) R_{K,\sigma}(v). \quad (12.42)$$

Thanks to Assumption (12.40), the reconstructed gradient may be written

$$\overline{\nabla}_K v = \frac{1}{|K|} \sum_{\sigma \in \mathcal{F}_K} \frac{|\sigma|}{d_{K,\sigma}} (v_\sigma - v_K)(\overline{\boldsymbol{x}}_\sigma - \boldsymbol{x}_K).$$

Usin again (12.40), 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 (12.5) of $R_{K,\sigma}$,

$$\sum_{\sigma \in \mathcal{F}_K} \frac{|\sigma|}{d_{K,\sigma}} R_{K,\sigma}(u) R_{K,\sigma}(v) = \sum_{\sigma \in \mathcal{F}_K} \frac{|\sigma|}{d_{K,\sigma}} (u_\sigma - u_K)(v_\sigma - v_K) - |K| \overline{\nabla}_K u \cdot \overline{\nabla}_K v.$$

Plugged into (12.42), this yields (12.41). The expressions of $F_{K,\sigma}$ are then obtained by comparing (12.41) and (12.35). ∎

12.4 A cell-centered variant of HMM schemes on Δ-admissible meshes

Let us consider a Δ-admissible mesh in the sense of [47]. We recall that in the case of a Δ-adapted polytopal mesh, the line $(\boldsymbol{x}_K, \boldsymbol{x}_L)$ is orthogonal to the interface σ. 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

$$\begin{array}{ll} d_\sigma = d_{K,\sigma} + d_{L,\sigma} & \text{and } \delta_{K,\sigma} u = u_L - u_K, \ \forall \sigma = K|L \in \mathcal{F}_{\mathrm{int}}, \\ d_\sigma = d_{K,\sigma} & \text{and } \delta_{K,\sigma} u = -u_K, \quad \forall \sigma \in \mathcal{F}_K \cap \mathcal{F}_{\mathrm{ext}}. \end{array} \quad (12.43)$$

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 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:

$$\nabla_K u = \frac{1}{|K|} \sum_{\sigma \in \mathcal{F}_K} |\sigma| (\overline{\boldsymbol{x}}_\sigma - \boldsymbol{x}_K) \frac{\delta_{K,\sigma} u}{d_\sigma}. \tag{12.44}$$

We then let

$$R_{K,\sigma}(u) = \frac{\delta_{K,\sigma} u}{d_\sigma} - \nabla_K u \cdot \boldsymbol{n}_{K,\sigma} \tag{12.45}$$

and

$$\nabla_{K,\sigma} u = \nabla_K u + \sqrt{d} R_{K,\sigma}(u)(\overline{\boldsymbol{x}}_\sigma - \boldsymbol{x}_K). \tag{12.46}$$

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 $\mathcal{L}_K = \mathrm{Id}$, this variant gives back the standard 2-point scheme for superadmissible meshes when $\Lambda = \mathrm{Id}$.

12.5 The SUSHI scheme for homogeneous Dirichlet conditions

The SUSHI scheme is nothing else than a barycentric condensation of HMM schemes, in which some of the face unknowns are eliminated. In its simplest form, a SUSHI GD is given by Definition 7.38 with $\mathcal{D}$ an HMM GD and $I^{\mathrm{Ba}} = \mathcal{M} \cup \mathcal{F}_{\mathrm{hyb}}$ for some $\mathcal{F}_{\mathrm{hyb}} \subset \mathcal{F}$. The face unknowns that are eliminated correspond to $\mathcal{F}_{\mathrm{bary}} = \mathcal{F} \backslash \mathcal{F}_{\mathrm{hyb}}$. If $\sigma \in \mathcal{F}_{\mathrm{bary}}$, 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 A.3 a linearly exact barycentric condensation as in Definition 7.38 leads to a poor approximation of the solution. The notion of $\mathcal{S}$-adapted barycentric condensation, that relaxes this requirement of a linearly exact condensation, is therefore particularly useful for the SUSHI GD.

12.5.1 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 (3.1) is smooth in the regions where Λ is smooth, and has continuous fluxes where Λ is discontinuous. This describes a subset $\mathcal{S}$ of $H^1_0(\Omega)$. We present here a SUSHI GD that is $\mathcal{S}$-adapted, and produces better approximation results in the case of heterogeneous material. The construction of the interpolation families is based on the following result.

Lemma 12.20. *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*

$$\boldsymbol{y}_\sigma = \frac{\lambda_L d_{K,\sigma} \boldsymbol{y}_L + \lambda_K d_{L,\sigma} \boldsymbol{y}_K}{\lambda_L d_{K,\sigma} + \lambda_K d_{L,\sigma}} + \frac{d_{K,\sigma} d_{L,\sigma}}{\lambda_L d_{K,\sigma} + \lambda_K d_{L,\sigma}} (\boldsymbol{\lambda}_L^t - \boldsymbol{\lambda}_K^t), \tag{12.47}$$

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

$$u(\boldsymbol{y}_\sigma) = \frac{\lambda_L d_{K,\sigma} u(\boldsymbol{x}_L) + \lambda_K d_{L,\sigma} u(\boldsymbol{x}_K)}{\lambda_L d_{K,\sigma} + \lambda_K d_{L,\sigma}}. \tag{12.48}$$

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

$$\begin{aligned} \forall \boldsymbol{y} \in \sigma \,:\, u(\boldsymbol{y}) &= u_K + d_{K,\sigma} g_K + (\boldsymbol{y} - \boldsymbol{y}_K) \cdot \boldsymbol{G}_K^t \\ &= u_L - d_{L,\sigma} g_L + (\boldsymbol{y} - \boldsymbol{y}_L) \cdot \boldsymbol{G}_L^t. \end{aligned} \tag{12.49}$$

This is equivalent to the two conditions $\boldsymbol{G}_K^t = \boldsymbol{G}_L^t =: \boldsymbol{g}^t$ and

$$d_{K,\sigma} g_K + d_{L,\sigma} g_L = u_L - u_K + (\boldsymbol{y}_K - \boldsymbol{y}_L) \cdot \boldsymbol{g}^t. \tag{12.50}$$

The condition $\Lambda_K \boldsymbol{G}_K \cdot \boldsymbol{n}_{KL} = \Lambda_L \boldsymbol{G}_L \cdot \boldsymbol{n}_{KL}$ can be written

$$g_K \lambda_K - g_L \lambda_L = \boldsymbol{g}^t \cdot (\boldsymbol{\lambda}_L^t - \boldsymbol{\lambda}_K^t). \tag{12.51}$$

From (12.50) and (12.51) 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 (12.49), this formula gives, for any $\boldsymbol{y} \in \sigma$,

$$u(\boldsymbol{y}) = u_K + d_{K,\sigma}\frac{\lambda_L(u_L-u_K)}{\lambda_L d_{K,\sigma}+\lambda_K d_{L,\sigma}} + d_{K,\sigma}\frac{\lambda_L(\boldsymbol{y}_K-\boldsymbol{y}_L)\cdot\boldsymbol{g}^t + d_{L,\sigma}\boldsymbol{g}^t\cdot(\boldsymbol{\lambda}_L^t-\boldsymbol{\lambda}_K^t)}{\lambda_L d_{K,\sigma}+\lambda_K d_{L,\sigma}} + (\boldsymbol{y}-\boldsymbol{y}_K)\cdot\boldsymbol{g}^t. \quad (12.52)$$

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 (12.47). Equation (12.52) 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 (12.48). ∎

This lemma justifies the following construction of interpolation families. We recall that $\mathcal{F}$ is split into $\mathcal{F}_{\rm hyb}$, corresponding to degrees of freedom that will remain in the SUSHI GD, and $\mathcal{F}_{\rm bary}$, corresponding to degrees of freedom 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 $\tau\in\mathcal{F}_{\rm hyb}$, then $\boldsymbol{y}_\tau=\overline{\boldsymbol{x}}_\tau$ and $w_\tau=u_\tau$;
2. if $\tau\in\mathcal{F}_{\rm bary}$ is a common face to grid cells M and N, then

$$\boldsymbol{y}_\tau = \frac{\lambda_N d_{M,\tau}\boldsymbol{y}_N+\lambda_M d_{N,\tau}\boldsymbol{y}_M + d_{M,\tau}d_{N,\tau}(\boldsymbol{\lambda}_N^\tau-\boldsymbol{\lambda}_M^\tau)}{\lambda_N d_{M,\tau}+\lambda_M d_{N,\tau}}, \quad (12.53)$$

$$w_\tau = \frac{\lambda_N d_{M,\tau}u_N+\lambda_M d_{N,\tau}u_M}{\lambda_N d_{M,\tau}+\lambda_M d_{N,\tau}}, \quad (12.54)$$

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

$$\lambda_M=\boldsymbol{n}_{MN}\cdot\Lambda_M\boldsymbol{n}_{MN},\ \ \boldsymbol{\lambda}_M^\tau=\Lambda_M\boldsymbol{n}_{MN}-\lambda_M\boldsymbol{n}_{MN},$$
$$\lambda_N=\boldsymbol{n}_{MN}\cdot\Lambda_N\boldsymbol{n}_{MN},\ \ \boldsymbol{\lambda}_N^\tau=\Lambda_N\boldsymbol{n}_{MN}-\lambda_N\boldsymbol{n}_{MN}$$

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 $\sigma\in\mathcal{F}_{\rm bary}$. 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:

$$\begin{gathered} w \text{ is affine in } K \text{ and in } L,\ w \text{ is continuous on } \sigma,\\ \Lambda_K(\nabla w)_K\cdot\boldsymbol{n}_{KL}=\Lambda_L(\nabla w)_L\cdot\boldsymbol{n}_{KL} \text{ and}\\ u_K=w(\boldsymbol{x}_K),\ u_L=w(\boldsymbol{x}_L),\ w_{\tau_i}=w(\boldsymbol{y}_{\tau_i}) \text{ for all } i=1,\dots,d. \end{gathered} \quad (12.55)$$

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

$$u_\sigma = \sum_{M\in H_\sigma} \beta_M^\sigma u_M. \tag{12.56}$$

This defines the family of barycentric coefficients $(\beta_i^\sigma)_{i\in H_\sigma}$. The degrees of freedom corresponding to $\mathcal{F}_{\rm hyb}$ are therefore not used to eliminate the degrees of freedom on $\mathcal{F}_{\rm bary}$.

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 $X = (u_K - u_\sigma, u_L - u_\sigma, (w_{\tau_i} - u_\sigma)_{i=1,\dots,d-1})$ 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 $(u_K, u_L, (w_{\tau_i})_{i=1,\dots,d-1})$ as expected.

In practice, the selection of the faces $(\tau_i)_{i=1,\dots,d-1}$ 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 each discretization is fully contained into only one P_ℓ), and that the sets of barycentric faces $\mathcal{F}_{\rm bary}$ are chosen such that

$$\forall \tau \in \mathcal{F}_{\rm bary} \,:\, \boldsymbol{y}_\tau \text{ defined by (12.53) belongs to } \tau \tag{12.57}$$

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 A.16 and therefore shows that SUSHI GDs constructed using the coefficients (12.56) are coercive, consistent, limit-conforming and compact. SUSHI GDs also obviously have piecewise constant reconstructions. We refer to Definition A.15 for the definition of the quantity $\mathrm{reg}_{\mathcal{S}}$ used in the next lemma.

Lemma 12.21. *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* (12.56).
Then $\mathcal{S}$ is dense in $H_0^1(\Omega)$ and, under Assumption (12.57)*, $\mathcal{S}$ 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 $\mathrm{reg}_s(\mathcal{D}^{\mathcal{S}})$) such that*

$$\forall \sigma \in \mathcal{F}_{\mathrm{bary}} \,:\, \left|\varphi(\overline{\boldsymbol{x}}_\sigma) - \sum_{K \in H_\sigma} \beta_K^\sigma \varphi(\boldsymbol{x}_K)\right| \leq C_\varphi R_{\mathcal{D}^{\mathcal{S}}} \mathrm{diam}(\sigma)^2. \tag{12.58}$$

Proof.
The density of $\mathcal{S}$ is established in [2, Lemma 3.2]. We consider therefore Property (12.58). For any $\tau = M|N \in \mathcal{F}_{\mathrm{bary}}$, 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 12.20,

$$\overline{\varphi}(\boldsymbol{y}_\tau) = \frac{\lambda_N d_{M,\tau}\overline{\varphi}(\boldsymbol{x}_N) + \lambda_M d_{N,\tau}\overline{\varphi}(\boldsymbol{x}_M)}{\lambda_N d_{M,\tau} + \lambda_M d_{N,\tau}}$$

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

$$\varphi(\boldsymbol{y}_\tau) = \frac{\lambda_N d_{M,\tau}\varphi(\boldsymbol{x}_N) + \lambda_M d_{N,\tau}\varphi(\boldsymbol{x}_M)}{\lambda_N d_{M,\tau} + \lambda_M d_{N,\tau}} + \mathcal{O}((h_M + h_N)^2). \tag{12.59}$$

Let us then consider the values w_τ constructed as above from the values $u_\tau = \varphi(\overline{\boldsymbol{x}}_\tau)$ if $\tau \in \mathcal{F}_{\mathrm{hyb}}$, and $u_N = \varphi(\boldsymbol{x}_N)$, $u_M = \varphi(\boldsymbol{x}_M)$ if $\tau = M|N \in \mathcal{F}_{\mathrm{bary}}$. 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 (12.55) 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

$$w - \varphi = \mathcal{O}(\mathrm{diam}(\sigma)^2) \tag{12.60}$$

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 (12.60) 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 (12.58). ∎

13

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 degrees of freedom 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 13.1.

13.1 Definition and properties of nMFD gradient discretisations

We first define the GD, and then prove that the corresponding GS (3.4) is indeed the nMFD scheme as defined in [14].
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

$$\int_K w(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \approx \sum_{\boldsymbol{s}\in\mathcal{V}_K} \omega_K^{\boldsymbol{s}} w(\boldsymbol{s}) \tag{13.1}$$

is exact for constant functions w, which means that

$$\sum_{\boldsymbol{s}\in\mathcal{V}_K} \omega_K^{\boldsymbol{s}} = |K|. \tag{13.2}$$

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

$$\int_\sigma w(\boldsymbol{x})\mathrm{d}s(\boldsymbol{x}) \approx \sum_{\boldsymbol{s}\in\mathcal{V}_\sigma} \omega_\sigma^{\boldsymbol{s}} w(\boldsymbol{s}) \tag{13.3}$$

is exact for affine functions w. This is equivalent to

$$\sum_{\boldsymbol{s}\in\mathcal{V}_\sigma} \omega_\sigma^{\boldsymbol{s}} = |\sigma| \quad \text{and} \quad \sum_{\boldsymbol{s}\in\mathcal{V}_\sigma} \omega_\sigma^{\boldsymbol{s}} \boldsymbol{s} = |\sigma| \overline{\boldsymbol{x}}_\sigma. \tag{13.4}$$

We also assume the following property on these weights.

$$\forall K \in \mathcal{M}\,,\ \forall \boldsymbol{s} \in \mathcal{V}_K\,,\ \exists \sigma \in \mathcal{F}_{K,\boldsymbol{s}} \text{ such that } \omega_\sigma^{\boldsymbol{s}} \neq 0, \tag{13.5}$$

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 (13.3) on the faces of K. (13.5) 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 center $\boldsymbol{x}_K \in \mathcal{P}$ by setting

$$\boldsymbol{x}_K = \frac{1}{|K|} \sum_{\boldsymbol{s}\in\mathcal{V}_K} \omega_K^{\boldsymbol{s}} \boldsymbol{s}, \tag{13.6}$$

and we select a partition $(V_{K,\boldsymbol{s}})_{\boldsymbol{s}\in\mathcal{V}_K}$ of K such that

$$\forall \boldsymbol{s} \in \mathcal{V}_K\,,\ |V_{K,\boldsymbol{s}}| = \sum_{\sigma\in\mathcal{F}_{K,\boldsymbol{s}}} \omega_\sigma^{\boldsymbol{s}} \frac{|D_{K,\sigma}|}{|\sigma|} = \frac{1}{d} \sum_{\sigma\in\mathcal{F}_{K,\boldsymbol{s}}} \omega_\sigma^{\boldsymbol{s}} d_{K,\sigma}. \tag{13.7}$$

The second equality follows from (B.1), and we note that (13.4) 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 DOFs 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

$$\forall \boldsymbol{s} \in \mathcal{V}_K\,,\ \text{for a.e. } \boldsymbol{x} \in K\,,\ \pi_K^{\boldsymbol{s}}(\boldsymbol{x}) := \frac{\omega_K^{\boldsymbol{s}}}{|K|}. \tag{13.8}$$

   Relation (7.33) gives

$$\forall v \in X_{\mathcal{D},0},\ \forall K \in \mathcal{M},\ \forall \boldsymbol{x} \in K,\ \Pi_{\mathcal{D}} v(\boldsymbol{x}) = v_K := \frac{1}{|K|} \sum_{\boldsymbol{s}\in\mathcal{V}_K} \omega_K^{\boldsymbol{s}} v_{\boldsymbol{s}}. \tag{13.9}$$

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 DOFs in K, and

$$\forall K \in \mathcal{M}\,,\ \forall v \in X_{\mathcal{V}_K}\,,\ \nabla_K v = \frac{1}{|K|}\sum_{\sigma\in\mathcal{F}_K}\left(\sum_{\boldsymbol{s}\in\mathcal{V}_\sigma}\omega_\sigma^{\boldsymbol{s}} v_{\boldsymbol{s}}\right)\boldsymbol{n}_{K,\sigma}. \tag{13.10}$$

Define then, for $v \in X_{\mathcal{V}_K}$, the function $\mathcal{G}_K v \in L^p(K)^d$ by

$$\begin{aligned}&\forall \boldsymbol{s} \in \mathcal{V}_K\,,\ \text{for a.e. } \boldsymbol{x} \in V_{K,\boldsymbol{s}}\,,\\ &\mathcal{G}_K v(\boldsymbol{x}) = \nabla_K v + \frac{1}{h_K}[\mathcal{L}_K R_K(v)]_{\boldsymbol{s}}\boldsymbol{N}_{K,\boldsymbol{s}}\end{aligned} \tag{13.11}$$

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

$$R_{K,\boldsymbol{s}}(v) = v_{\boldsymbol{s}} - v_K - \nabla_K v\cdot(\boldsymbol{s} - \boldsymbol{x}_K), \tag{13.12}$$

  where v_K is defined in (13.9) and $\boldsymbol{x}_K$ is given by (13.6),
- $\mathcal{L}_K$ is an isomorphism of the space $\mathrm{Im}(R_K)$.

By (7.34), we then have

$$\begin{aligned}&\forall v \in X_{\mathcal{D}}\,,\ \forall K \in \mathcal{M}\,,\ \forall \boldsymbol{s} \in \mathcal{V}_K\,,\ \text{for a.e. } \boldsymbol{x} \in V_{K,\sigma}\,,\\ &\nabla_{\mathcal{D}} v(\boldsymbol{x}) = \nabla_K v + \frac{1}{h_K}[\mathcal{L}_K R_K(v)]_{\boldsymbol{s}}\boldsymbol{N}_{K,\boldsymbol{s}}.\end{aligned} \tag{13.13}$$

The functions $(\mathcal{G}_K^{\boldsymbol{s}})_{\boldsymbol{s}\in\mathcal{V}_K}$ of $L^p(K)^d$ are recovered from the definition (13.11) 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

$$\mathcal{G}_K^{\boldsymbol{s}} = \mathcal{G}_K v^{\boldsymbol{s}} \text{ for all } \boldsymbol{s} \in \mathcal{V}_K. \tag{13.14}$$

4. The proof that π_K and $\mathcal{G}_K$ are exact reconstructions and that $\|\nabla_{\mathcal{D}}\cdot\|_{L^p(\Omega)^d}$ is a norm on $X_{\mathcal{D},0}$, is provided in Lemma 13.8 below.

Remark 13.1 (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 (3.4) 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 [14]. The space of degrees of freedom at the interior vertices of the mesh, denoted by $\mathcal{N}_0$ in [14], is simply $X_{\mathcal{D},0}$ defined above. The nMFD for (3.1) (with $\boldsymbol{F}=0$) is then written under the general form

$$\text{Find } u\in X_{\mathcal{D},0} \text{ such that, for all } v\in X_{\mathcal{D},0},\ [u,v]_{X_{\mathcal{D},0}} = \widetilde{f}(v), \tag{13.15}$$

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 (13.1), the linear form $\widetilde{f}$ is defined as

$$\widetilde{f}(v) = \sum_{K\in\mathcal{M}} \left(\frac{1}{|K|}\int_K f\right) \sum_{\boldsymbol{s}\in\mathcal{V}_K} \omega_K^{\boldsymbol{s}} v_{\boldsymbol{s}}. \tag{13.16}$$

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 [14] that this leads to the following generic form:

$$\begin{aligned}&\forall u\in X_{\mathcal{V}_K},\ \forall v\in X_{\mathcal{V}_K}\ :\\ &\qquad [u,v]_{\mathcal{V}_K} = u^T\mathbb{M}_K v \ \text{ with } \ \mathbb{M}_{\mathbb{K}} = \frac{1}{|K|}\mathbb{C}_K\Lambda_K^{-1}\mathbb{C}_K^T + \mathbb{D}_K\mathbb{K}_K\mathbb{D}_K^T,\end{aligned} \tag{13.17}$$

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 13.2 ($\mathbb{R}^{\mathcal{V}_K}$ vs. $\mathbb{R}^{\mathrm{Card}(\mathcal{V}_K)}$)
As in Remark 12.3, 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 method are GDMs, two technical results are required. The first one contains in particular results similar to those in Lemma 12.4, and the second one describes the kernel of $\mathbb{D}_K$.

Lemma 13.3. *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*

$$\sum_{\boldsymbol{s}\in\mathcal{V}_K} \frac{|V_{K,\boldsymbol{s}}|}{h_K} \beta_{\boldsymbol{s}} \boldsymbol{N}_{K,\boldsymbol{s}} = 0. \tag{13.18}$$

2. *For all $v \in X_{\mathcal{D},0}$ and all $K \in \mathcal{M}$,*

$$\nabla_K v = \frac{1}{|K|} \int_K \nabla_{\mathcal{D}} v(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}. \tag{13.19}$$

3. *∇_K is a $\mathbb{P}_1$-exact gradient reconstruction on K upon $\mathcal{V}_K$, in the sense of Definition 7.29.*
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 (13.4) 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

$$\begin{aligned}
\sum_{\boldsymbol{s}\in\mathcal{V}_K} \frac{|V_{K,\boldsymbol{s}}|}{h_K} \beta_{\boldsymbol{s}} \boldsymbol{N}_{K,\boldsymbol{s}} &= \frac{1}{d} \sum_{\boldsymbol{s}\in\mathcal{V}_K} \sum_{\sigma\in\mathcal{F}_{K,\boldsymbol{s}}} \beta_{\boldsymbol{s}} \omega_\sigma^{\boldsymbol{s}} \boldsymbol{n}_{K,\sigma} \\
&= \frac{1}{d} \sum_{\sigma\in\mathcal{F}_K} \left(\sum_{\boldsymbol{s}\in\mathcal{V}_\sigma} \omega_\sigma^{\boldsymbol{s}} \beta_{\boldsymbol{s}} \right) \boldsymbol{n}_{K,\sigma} \\
&= \frac{1}{d} \sum_{\sigma\in\mathcal{F}_K} \left(\sum_{\boldsymbol{s}\in\mathcal{V}_\sigma} \omega_\sigma^{\boldsymbol{s}} w_{\boldsymbol{s}} \right) \boldsymbol{n}_{K,\sigma} - \frac{1}{d} \sum_{\sigma\in\mathcal{F}_K} \left(\sum_{\boldsymbol{s}\in\mathcal{V}_\sigma} \omega_\sigma^{\boldsymbol{s}} \nabla_K w \cdot (\boldsymbol{s} - \boldsymbol{x}_K) \right) \boldsymbol{n}_{K,\sigma} \\
&= \frac{1}{d} \left(|K| \nabla_K w - T_1 \right).
\end{aligned} \tag{13.20}$$

We then use (13.4) and Lemma B.3 to write

$$\begin{aligned}
T_1 &= \sum_{\sigma\in\mathcal{F}_K} \left[\nabla_K w \cdot \left(\sum_{\boldsymbol{s}\in\mathcal{V}_\sigma} \omega_\sigma^{\boldsymbol{s}} (\boldsymbol{s} - \boldsymbol{x}_K) \right) \right] \boldsymbol{n}_{K,\sigma} \\
&= \sum_{\sigma\in\mathcal{F}_K} |\sigma| \left[\nabla_K w \cdot (\overline{\boldsymbol{x}}_\sigma - \boldsymbol{x}_K) \right] \boldsymbol{n}_{K,\sigma} = |K| \nabla_K w.
\end{aligned} \tag{13.21}$$

Substituted in (13.20) this shows that β satisfies (13.18). 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 (13.4) 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 (13.5), 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 (13.21), 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 (13.13) 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 (13.19).

Item 3. If $v = (A(\boldsymbol{s}))_{\boldsymbol{s}\in\mathcal{V}_K}$ for some affine map A, then (13.4) 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.6,

$$\nabla_K v = \frac{1}{|K|}\sum_{\sigma\in\mathcal{F}_K}|\sigma|A(\overline{\boldsymbol{x}}_\sigma)\boldsymbol{n}_{K,\sigma} = \overline{\nabla}_K u = \nabla A.$$

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 (13.7) 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 13.4. *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

$$\forall \boldsymbol{s} \in \mathcal{V}_K \ : \ v_{\boldsymbol{s}} = v_K + \nabla_K v \cdot (\boldsymbol{s} - \boldsymbol{x}_K). \tag{13.22}$$

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 13.3. The definitions (13.9) and (13.6) 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 (13.22) holds. Conversely, if (13.22) 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 13.5 (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* (13.2), (13.4) *and* (13.5), *and let* (13.15) *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* (3.4) *is identical to* (13.15).

Remark 13.6 (Non piecewise constant diffusion tensor)
As for the HMM method (see Remark 12.16), if Λ is not piecewise constant on $\mathcal{M}$, then (13.15) is the GS (3.4) 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 Chapter 3.

Proof. Given the definitions (13.9) of $\Pi_{\mathcal{D}}$ and (13.16) of $\widetilde{f}$, the right-hand sides of (3.4) and (13.15) 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 (13.11) 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

$$\int_K \Lambda(\boldsymbol{x}) \nabla_{\mathcal{D}} u(\boldsymbol{x}) \cdot \nabla_{\mathcal{D}} v(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} = u^T \mathbb{M}_K v. \tag{13.23}$$

Let $S_K(u) = \nabla_{\mathcal{D}} u - \nabla_K u$ be the stabilisation part of $\nabla_{\mathcal{D}} u$ on K. By (13.19), 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,

$$\begin{aligned}
&\int_K \Lambda(\boldsymbol{x})\nabla_{\mathcal{D}} u(\boldsymbol{x}) \cdot \nabla_{\mathcal{D}} v(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \\
&\quad = |K|\Lambda_K \nabla_K u \cdot \nabla_K v + \int_K \Lambda_K \nabla_K u \cdot S_K(v)(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \\
&\qquad + \int_K \Lambda_K S_K(u)(\boldsymbol{x}) \cdot \nabla_K v \mathrm{d}\boldsymbol{x} + \int_K \Lambda_K S_K(u)(\boldsymbol{x}) \cdot S_K(v)(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \\
&\quad = |K|\Lambda_K \nabla_K u \cdot \nabla_K v + \int_K \Lambda_K S_K(u)(\boldsymbol{x}) \cdot S_K(v)(\boldsymbol{x})\mathrm{d}\boldsymbol{x}. \qquad (13.24)
\end{aligned}$$

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 (13.24) therefore corresponds to the first term in the expression (13.17) 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 $u, v \in X_{\mathcal{V}_K}$,

$$u^T \mathbb{D}_K \mathbb{K}_K \mathbb{D}_K^T v = \int_K \Lambda_K S_K(u)(\boldsymbol{x}) \cdot S_K(v)(\boldsymbol{x})\mathrm{d}\boldsymbol{x}. \qquad (13.25)$$

By Lemma 13.4 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 13.7 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 (13.25) follows if we can establish the existence of an isomorphism $\mathcal{L}_K$ of $\mathrm{Im}(R_K)$ such that, for all $u, v \in X_{\mathcal{D},0}$,

$$\{R_K(v), R_K(v)\}_2 = \int_K \Lambda_K S_K(u)(\boldsymbol{x}) \cdot S_K(v)(\boldsymbol{x})\mathrm{d}\boldsymbol{x}. \qquad (13.26)$$

By definition of $S_K(u)$ (see (13.13)), we have

$$\int_K \Lambda_K S_K(u)(\boldsymbol{x}) \cdot S_K(v)(\boldsymbol{x})\mathrm{d}\boldsymbol{x}$$

$$= \sum_{\boldsymbol{s}\in\mathcal{V}_K} \frac{|V_{K,\boldsymbol{s}}|}{h_K^2} [\mathcal{L}_K R_K(u)]_{\boldsymbol{s}} [\mathcal{L}_K R_K(v)]_{\boldsymbol{s}} \Lambda_K \boldsymbol{N}_{K,\boldsymbol{s}} \cdot \boldsymbol{N}_{K,\boldsymbol{s}}$$
$$= \langle \mathcal{L}_K R_K(u), \mathcal{L}_K R_K(v) \rangle \tag{13.27}$$

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 13.3). Since $\{\cdot,\cdot\}_2$ and $\langle\cdot,\cdot\rangle$ are two scalar products on $\mathrm{Im}(R_K)$, Lemma 12.6 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 (13.27) shows that (13.26) holds for this choice of $\mathcal{L}_K$. ∎

The following lemma, used the in the above proof, is taken from [35].

Lemma 13.7. *Let X, Y and Z be finite dimension 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 $\{Bx, Bx'\}_Z = \{Ax, Ax'\}_Y$. ∎

13.1.1 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 $(\mathcal{L}_K)_{K\in\mathcal{M}}$:

$$\zeta_{\mathcal{D}} = \min\Bigg\{\zeta > 0 \,:\, \forall K \in \mathcal{M},\ \forall v \in X_{\mathcal{V}_K},\\ \zeta^{-1} \sum_{\boldsymbol{s}\in\mathcal{V}_K} |V_{K,\boldsymbol{s}}| \left|\frac{R_{K,\boldsymbol{s}}(v)}{h_K}\right|^p \le \sum_{\boldsymbol{s}\in\mathcal{V}_K} |V_{K,\boldsymbol{s}}| \left|\frac{[\mathcal{L}_K R_K(v)]_{\boldsymbol{s}}}{h_K}\right|^p \\ \le \zeta \sum_{\boldsymbol{s}\in\mathcal{V}_K} |V_{K,\boldsymbol{s}}| \left|\frac{R_{K,\boldsymbol{s}}(v)}{h_K}\right|^p \Bigg\}. \tag{13.28}$$

The boundedness of $\zeta_{\mathcal{D}}$ is a weaker assumption than the classical coercivity assumption of nMFD methods, see e.g. [14, Eq. (5.15)]. Choosing $\mathcal{L}_K = \beta_K \mathrm{Id}$ with $\beta_K \in [\zeta^{-1}, \zeta]$ ensures that the inequalities within (13.28) is satisfied.

Lemma 13.8 (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 13.1. 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 (13.28)) and

$$\varrho \ge \max_{K\in\mathcal{M}} \mathrm{Card}(\mathcal{V}_K), \tag{13.29}$$

then there exists C_{29} depending only on p, d and ϱ such that $\mathrm{reg}_{\mathrm{LLE}}(\mathcal{D}) \le C_{29}$.

Proof. By choice (13.2) of the weights and definition (13.8) 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 13.3 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 13.4, $v_K = A(\boldsymbol{x}_K)$ 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 (13.19) shows that $\nabla_K v = 0$ for all $K \in \mathcal{M}$ and thus, by (13.13), $R_K(v) = 0$. The definition (13.12) 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 nonnegative, $\sum_{i\in I_K} |\pi_K^i| = \sum_{i\in I_K} \pi_K^i = 1$ 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 (13.7) 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 (13.11) of $\mathcal{G}_K$ gives

$$\begin{aligned}\|\mathcal{G}_K v\|_{L^p(K)^d}^p &\le 2^{p-1}\left(|K|\,|\nabla_K v|^p + \sum_{\boldsymbol{s}\in\mathcal{V}_K}|V_{K,\boldsymbol{s}}|\left|\frac{[\mathcal{L}_K R_K(v)]_{\boldsymbol{s}}}{h_K}\right|^p|\boldsymbol{N}_{K,\sigma}|^p\right)\\ &\le 2^{p-1}\left(|K|\,|\nabla_K v|^p + \theta_{\mathfrak{T}}^p\sum_{\boldsymbol{s}\in\mathcal{V}_K}|V_{K,\boldsymbol{s}}|\left|\frac{[\mathcal{L}_K R_K(v)]_{\boldsymbol{s}}}{h_K}\right|^p\right)\\ &\le 2^{p-1}\left(|K|\,|\nabla_K v|^p + \theta_{\mathfrak{T}}^p\zeta_{\mathcal{D}}\sum_{\boldsymbol{s}\in\mathcal{V}_K}|V_{K,\boldsymbol{s}}|\left|\frac{R_{K,\boldsymbol{s}}(v)}{h_K}\right|^p\right). \quad (13.30)\end{aligned}$$

Let $V = \max_{\boldsymbol{s}\in\mathcal{V}_K}|v_{\boldsymbol{s}}|$. The definition (13.13) of $\nabla_K v$ yields, thanks to (B.1),

$$|\nabla_K v| \le \frac{V}{|K|}\sum_{\sigma\in\mathcal{F}_K}\sum_{\boldsymbol{s}\in\mathcal{V}_\sigma}\omega_\sigma^{\boldsymbol{s}} \le \frac{\theta_{\mathfrak{T}}V}{h_K|K|}\sum_{\sigma\in\mathcal{F}_K}|\sigma|d_{K,\sigma} = \frac{d\theta_{\mathfrak{T}}V}{h_K}. \quad (13.31)$$

Using the definition (13.9) of v_K, we infer

$$|R_{K,\boldsymbol{s}}(v)| \le \frac{1}{|K|}\sum_{\boldsymbol{s}\in\mathcal{V}_K}\omega_K^{\boldsymbol{s}}V + V + |\nabla_K v|h_K \le (2+d\theta_{\mathfrak{T}})V.$$

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 (13.31) into (13.30), this estimate gives

$$\|\mathcal{G}_K v\|_{L^p(K)^d}^p \le 2^{p-1}\left[(d\theta_{\mathfrak{T}})^p + \theta_{\mathfrak{T}}^p\zeta_{\mathcal{D}}(2+d\theta_{\mathfrak{T}})^p\right]h_K^{-p}|K|V^p.$$

Applied to $v = v^\sigma$ for all $\sigma \in \mathcal{F}_K$, and recalling the definition (12.7) of the functions $(\mathcal{G}_K^{\boldsymbol{s}})_{\boldsymbol{s}\in\mathcal{V}_K}$, we deduce $\|\mathcal{G}_K^{\boldsymbol{s}}\|_{L^p(K)^d} \le C_{30}h_K^{-1}|K|^{1/p}$ with C_{30} depending only on d, p and ϱ. The definition (7.26) of $\|\mathcal{G}_K\|_p$ and (13.29) then yield a bound on $\|\mathcal{G}_K\|_p$ that depends only on d, p and ϱ. ∎

Lemma 13.9 (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 13.1. We take $\varrho \ge \theta_{\mathfrak{T}} + \zeta_{\mathcal{D}}$ (see* (7.8) *and* (13.28)*). Then, there exists C_{31} depending only on Ω, p and ϱ such that*

$$\forall v \in X_{\mathcal{D},0},\ \sum_{K\in\mathcal{M}}\sum_{\boldsymbol{s}\in\mathcal{V}_K}|V_{K,\boldsymbol{s}}|\left|\frac{v_{\boldsymbol{s}}-v_K}{h_K}\right|^p \le C_{31}\|\nabla_{\mathcal{D}}v\|_{L^p(\Omega)}^p. \quad (13.32)$$

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 (13.19) and Jensen's inequality,

$$|\nabla_K v|^p \le \frac{1}{|K|}\int_K |\nabla_{\mathcal{D}} v(\boldsymbol{x})|^p \mathrm{d}\boldsymbol{x}. \tag{13.33}$$

Using Item 4 in Lemma 13.3 and the definition (13.13) of $\nabla_{\mathcal{D}}$, we infer that, for all $\boldsymbol{s} \in \mathcal{V}_K$ and a.e. $\boldsymbol{y} \in V_{K,\boldsymbol{s}}$,

$$\begin{aligned}\left|\frac{1}{h_K}[\mathcal{L}_K R_K(v)]_{\boldsymbol{s}}\right|^p &\le \left|\frac{1}{h_K}[\mathcal{L}_K R_K(v)]_{\boldsymbol{s}} \boldsymbol{N}_{K,\boldsymbol{s}}\right|^p \\ &\lesssim |\nabla_{\mathcal{D}} v(\boldsymbol{y})|^p + \frac{1}{|K|}\int_K |\nabla_{\mathcal{D}} v(\boldsymbol{x})|^p \mathrm{d}\boldsymbol{x}.\end{aligned}$$

Integrate over $\boldsymbol{y} \in V_{K,\boldsymbol{s}}$, sum over $\boldsymbol{s} \in \mathcal{V}_K$ and use the definition (13.28) 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 $|v_{\boldsymbol{s}} - v_K| \le |R_{K,\boldsymbol{s}}(v)| + h_K|\nabla_K v|$ and use (13.33) 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 (13.32). ∎

We now define a control of an nMFD GD by a polytopal toolbox, and we establish some estimates on this control.

Lemma 13.10 (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 13.1. 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 7.10) defined by: for $v \in X_{\mathcal{D},0}$,*

$$\begin{aligned}&\forall \sigma \in \mathcal{F}_K\,,\ \boldsymbol{\Phi}(v)_\sigma = \frac{1}{|\sigma|}\sum_{\boldsymbol{s}\in\mathcal{V}_\sigma} \omega_\sigma^{\boldsymbol{s}} v_{\boldsymbol{s}}\,, \quad \textit{and} \\ &\forall K \in \mathcal{M}\,,\ \boldsymbol{\Phi}(v)_K = v_K = \frac{1}{|K|}\sum_{\boldsymbol{s}\in\mathcal{V}_K} \omega_K^{\boldsymbol{s}} v_{\boldsymbol{s}}.\end{aligned} \tag{13.34}$$

Let $\varrho \ge \theta_{\mathfrak{T}} + \zeta_{\mathcal{D}}$ (see (7.8) and (13.28)). Then, there exists C_{32} depending only on Ω, p and ϱ such that

$$\|\boldsymbol{\Phi}\|_{\mathcal{D},\mathfrak{T}} \le C_{32}, \tag{13.35}$$

and

$$\omega^{\Pi}(\mathcal{D},\mathfrak{T},\boldsymbol{\Phi}) = 0\,, \quad \omega^{\nabla}(\mathcal{D},\mathfrak{T},\boldsymbol{\Phi}) = 0. \tag{13.36}$$

Proof. In this proof, $A \lesssim B$ means again that $A \le CB$ for some C depending only on Ω, p and ϱ. By definition of $\boldsymbol{\Phi}$, for $\sigma \in \mathcal{F}_K$,

$$|\boldsymbol{\Phi}(v)_\sigma - \boldsymbol{\Phi}(v)_K| = \left| \frac{1}{|\sigma|} \sum_{\boldsymbol{s}\in\mathcal{V}_\sigma} \omega_\sigma^{\boldsymbol{s}}(v_{\boldsymbol{s}} - v_K) \right| \le \frac{1}{|\sigma|} \sum_{\boldsymbol{s}\in\mathcal{V}_\sigma} \omega_\sigma^{\boldsymbol{s}} |v_{\boldsymbol{s}} - v_K|$$

Hence, the discrete Jensen inequality (C.11) (with the convex function $\Psi(s) = |s|^p$) and the definition of $\theta_{\mathfrak{T}}$ give

$$\left| \frac{\boldsymbol{\Phi}(v)_\sigma - \boldsymbol{\Phi}(v)_K}{d_{K,\sigma}} \right|^p \le \theta_{\mathfrak{T}}^p \left| \frac{\boldsymbol{\Phi}(v)_\sigma - \boldsymbol{\Phi}(v)_K}{h_K} \right|^p \lesssim \frac{1}{|\sigma|} \sum_{\boldsymbol{s}\in\mathcal{V}_\sigma} \omega_\sigma^{\boldsymbol{s}} \left| \frac{v_{\boldsymbol{s}} - v_K}{h_K} \right|^p.$$

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 (13.35) is completed by using (13.32) in Lemma 13.9 and by recalling the definition (7.12) of $\|\boldsymbol{\Phi}\|_{\mathcal{D},\mathfrak{T}}$.

We now turn to (13.36). By definitions (7.7c) of $\Pi_{\mathfrak{T}}$, (13.9) of $\Pi_{\mathcal{D}}$, and (13.34) 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, (13.19) 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 $\omega^{\nabla}(\mathcal{D}, \mathfrak{T}, \boldsymbol{\Phi}) = 0$. ∎

13.1.2 Properties of nMFD gradient discretisations

The theorems presented here follow immediately, as for HMM GDs, from the preliminary results above and from Propositions 7.36 and A.6, Theorem 7.12 and Corollary 7.13.

Theorem 13.11 (Properties of nMFD GDs). *Let $(\mathcal{D}_m)_{m\in\mathbb{N}}$ be a sequence of nMFD GDs, as in Section 13.1, 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 (13.28)), 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, 2.6 and 2.8.

Remark 13.12. 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 13.13 (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 13.1. Assume that $p > d/2$ and take $\varrho \geq \theta_{\mathfrak{T}} + \zeta_{\mathcal{D}}$ (see* (7.8) *and* (13.28)*) that also satisfies* (13.29)*. Then, there exists $C_{33} > 0$, depending only on Ω, p and ϱ, such that*

$$\forall \varphi \in W^{2,p}(\Omega) \cap W_0^{1,p}(\Omega)\,,\ S_{\mathcal{D}}(\varphi) \leq C_{33} h_{\mathcal{M}} \left\|\varphi\right\|_{W^{2,p}(\Omega)}\,,$$

where $S_{\mathcal{D}}$ is defined by (2.2)*.*

Proposition 13.14 (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 13.1. We take $\varrho \geq \theta_{\mathfrak{T}} + \eta_{\mathfrak{T}} + \zeta_{\mathcal{D}}$ (see* (7.8)*,* (7.9) *and* (13.28)*) that also satisfies* (13.29)*. Then, there exists C_{34} depending only on Ω, p and ϱ such that*

$$C_{\mathcal{D}} \leq C_{34} \tag{13.37}$$

and

$$\forall \boldsymbol{\varphi} \in W^{1,p'}(\Omega)^d\,,\ W_{\mathcal{D}}(\boldsymbol{\varphi}) \leq C_{34} h_{\mathcal{M}} \left\|\boldsymbol{\varphi}\right\|_{W^{1,p'}(\Omega)^d}\,. \tag{13.38}$$

Here, $C_{\mathcal{D}}$ and $W_{\mathcal{D}}$ are the coercivity constant and limit-conformity measure defined by (2.1) *and* (2.6)*.*

Remark 13.15 (Assumption on the weights)
As already mentioned, Assumption (13.5) 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.

13.2 Link with discrete duality finite volume methods

Let us consider the special case, in dimension $d = 3$, of an *octahedral* mesh. By that we mean 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 element 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,

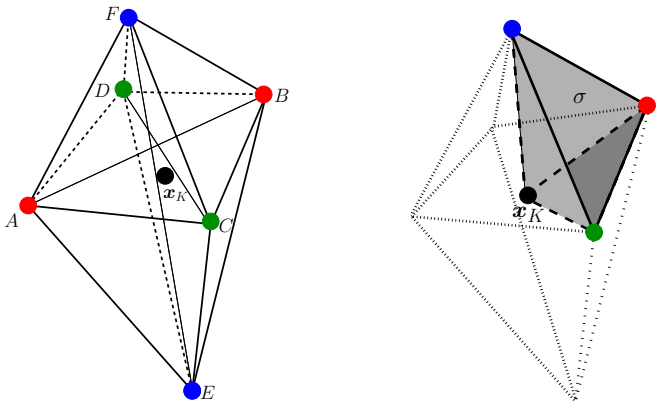


Fig. 13.1. Left: octahedral cell K. Right: illustration of $T_{K,\sigma}$ (greyed domain).

and each $\mathcal{V}_\sigma$ has 3 elements (see Figure 13.1, left). For any $K \in \mathcal{M}$, the center of K is defined by $\boldsymbol{x}_K = \frac{1}{6}\sum_{\boldsymbol{s}\in\mathcal{V}_K} \boldsymbol{s}$.
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 DOFs is unchanged, the gradient reconstruction is only built from the consistent part (13.10) 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 (13.3) 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 (13.10)).
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 13.1, right).

The following lemma characterises the reconstructed gradient.

Lemma 13.16. *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*

$$\text{For all opposite vertices } (\boldsymbol{s}_0, \boldsymbol{s}_1) \text{ of } K,\ \boldsymbol{\xi}\cdot(\boldsymbol{s}_0 - \boldsymbol{s}_1) = v_{\boldsymbol{s}_0} - v_{\boldsymbol{s}_1}. \quad (13.39)$$

Remark 13.17. The opposite vertices in the octahedra in Figure 13.1 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, (13.39) 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 (13.39). We have

$$(\nabla_{\mathcal{D}^\star} v)_{|K} = \frac{1}{|K|}\frac{1}{3}\sum_{\boldsymbol{s}\in\mathcal{V}_K} v_{\boldsymbol{s}} \sum_{\sigma\in\mathcal{F}_K|\boldsymbol{s}\in\mathcal{V}_\sigma} |\sigma|\boldsymbol{n}_{K,\sigma}. \tag{13.40}$$

Let us consider for example the case where $\boldsymbol{s} = A$ in Figure 13.1. 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 (13.40) 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 (13.39) 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, (13.39) is a well-known characterisation of the reconstructed gradient, piecewise constant on the so-called "diamond cells", of the CeVeFE Discrete Duality Finite Volume (DDFV) method [24, 25]. 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 in terms of GDMs is provided in [37]. We note that the same analysis also applies to the case $d = 2$, the octahedral mesh then becoming a quadrangular mesh (the cells still correspond to the "diamond cells" in the DDFV terminology).

Part IV

Appendix

In Parts I (or elliptic problems) and II (for parabolic problems), the properties (coercivity, GD-consistency, etc.) needed on gradient discretisations (GDs) to generate convergent gradient schemes (GSs) were introduced. This appendix introduces some technical tools which are used in Part III to prove that a given GD satisfies these core properties.

This appendix comprises two main chapters. Chapter A extends the analysis of LLE GDs done in Section 7.3 by establishing in particular, if $p > d/2$, explicit estimates on $S_{\mathcal{D}}$. These estimates are essential to obtain rates of convergence for GSs applied to linear elliptic and parabolic equations.
Chapter B is devoted to discrete functional analysis tools, that is, the translation to the discrete setting of classical results of functional analysis (Poincaré's inequality, compactness theorems, etc.). These tools are used, in Section 7.7 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 are build on results originally appeared in [49].
In these two chapters, 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 $\partial\Omega$.

A short final chapter, Chapter C, includes some classical technical results which are provided solely for the sake of completeness.

A

Complements on LLE GDs

This chapter deals with LLE GDs in the sense of Section 7.3. Section A.1 provides results enabling estimates on the consistency error $S_{\mathcal{D}}$, in the case where $p > d/2$ and the functions belong to $W^{2,p}(\Omega)$. We then generalise the notion of degree of freedom in Section A.2. Finally, we introduce the notion of non-linearly exact barycentric combinations in Section A.3, which may arise in the case of non homogeneous diffusion problems.

A.1 $W^{2,p}$ estimates for $S_{\mathcal{D}}$

Estimates on $S_{\mathcal{D}}(\varphi)$ are useful to obtain rates of convergences of GS for linear (and some non-linear) problems, see e.g. Theorem 3.2 and Theorem 3.28. The estimate (7.38) on $S_{\mathcal{D}}(\varphi)$ requires $\varphi \in W_0^{1,p}(\Omega) \cap W^{2,\infty}(\mathbb{R}^d)$. Hence, to use this estimate in Theorem 3.2 or Theorem 3.28, for example, the solution to the corresponding problem ((3.1) or (3.72)) 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 (7.38) 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 (3.1) or (3.72), than the $W^{2,\infty}$ regularity.

We start with a lemma that compares in $L^p(V)$ norm a function $\varphi \in W^{1,p}(V)$ with its average value on a ball in V.

Lemma A.1. *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_1 depending only on d and p such that, for any $\varphi \in W^{1,p}(V)$,*

$$\left\| \varphi - \frac{1}{|B|} \int_B \varphi(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \right\|_{L^p(V)} \leq C_1 \frac{\mathrm{diam}(V)^{\frac{d}{p}+1}}{\mathrm{diam}(B)^{\frac{d}{p}}} \| \, |\nabla \varphi| \, \|_{L^p(V)} \, . \tag{A.1}$$

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 $|\boldsymbol{x} - \boldsymbol{y}| \leq h_V$ gives

$$\begin{aligned}\left|\varphi(\boldsymbol{x}) - \frac{1}{|B|}\int_B \varphi(\boldsymbol{y})\mathrm{d}\boldsymbol{y}\right| &= \left|\frac{1}{|B|}\int_B \int_0^1 \nabla\varphi(t\boldsymbol{x} + (1-t)\boldsymbol{y}) \cdot (\boldsymbol{x} - \boldsymbol{y})\mathrm{d}t\mathrm{d}y\right| \\ &\leq \frac{h_V}{|B|}\int_B \int_0^1 |\nabla\varphi(t\boldsymbol{x} + (1-t)\boldsymbol{y})|\mathrm{d}t\mathrm{d}y.\end{aligned}$$

Taking the power p, using the Jensen inequality (C.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

$$\begin{aligned}\int_V \left|\varphi(\boldsymbol{x}) - \frac{1}{|B|}\int_B \varphi(\boldsymbol{y})\mathrm{d}\boldsymbol{y}\right|^p \mathrm{d}\boldsymbol{x} \\ \leq \frac{h_V^p}{|B|}\int_V \int_B \int_0^1 |\nabla\varphi(t\boldsymbol{x} + (1-t)\boldsymbol{y})|^p \mathrm{d}t\mathrm{d}y\mathrm{d}\boldsymbol{x}. \quad \text{(A.2)}\end{aligned}$$

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

$$\begin{aligned}\int_V \int_B \int_0^1 |\nabla\varphi(t\boldsymbol{x} + (1-t)\boldsymbol{y})|^p \mathrm{d}t\mathrm{d}\boldsymbol{y}\mathrm{d}\boldsymbol{x} \\ \leq \int_V |\nabla\varphi(\boldsymbol{z})|^p \int_B \int_{I(\boldsymbol{z},\boldsymbol{y})} t^{-d}\mathrm{d}t\mathrm{d}\boldsymbol{y}\mathrm{d}\boldsymbol{z}, \quad \text{(A.3)}\end{aligned}$$

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 $h_V t \geq |\boldsymbol{z} - \boldsymbol{y}|$. Hence $I(\boldsymbol{z}, \boldsymbol{y}) \subset [\frac{|\boldsymbol{z}-\boldsymbol{y}|}{h_V}, 1]$ and we deduce that (for $d > 1$)

$$\int_{I(\boldsymbol{z},\boldsymbol{y})} t^{-d}\mathrm{d}t \leq \int_{\frac{|\boldsymbol{z}-\boldsymbol{y}|}{h_V}}^1 t^{-d}\mathrm{d}t \leq \frac{1}{d-1}\frac{h_V^{d-1}}{|\boldsymbol{z} - \boldsymbol{y}|^{d-1}}. \quad \text{(A.4)}$$

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 $\boldsymbol{z} \in V$,

$$\int_B \int_{I(\boldsymbol{z},\boldsymbol{y})} t^{-d}\mathrm{d}t\mathrm{d}\boldsymbol{y} \leq \frac{h_V^{d-1}}{d-1}\int_B \frac{1}{|\boldsymbol{z} - \boldsymbol{y}|^{d-1}}\mathrm{d}\boldsymbol{y}$$

$$\leq \frac{h_V^{d-1}}{d-1} \int_{B(\boldsymbol{z},h_V)} |\boldsymbol{z}-\boldsymbol{y}|^{1-d} \mathrm{d}\boldsymbol{y}$$
$$\leq \frac{h_V^{d-1}}{d-1} \omega_d \int_0^{h_V} \rho^{1-d} \rho^{d-1} \mathrm{d}\rho \leq \frac{h_V^d}{d-1} \omega_d. \quad \text{(A.5)}$$

The proof is complete by plugging this estimate into (A.3), by using the resulting inequality in (A.2), and by recalling that

$$|B| = |B(0,1)| \left(\frac{\mathrm{diam}(B)}{2} \right)^d.$$

Note that in the case $d = 1$, (A.4) is modified and involves $\ln(\frac{h}{|\boldsymbol{z}-\boldsymbol{y}|})$ but the final estimate (A.5) is still in $\mathcal{O}(h^d)$. ∎

The following lemma is a simple technical result used in Lemma A.3 below.

Lemma A.2. *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*

$$\forall \boldsymbol{z} \in B(\boldsymbol{x},h),\ F_{\boldsymbol{x},h}(\boldsymbol{z}) = \int_{\frac{|\boldsymbol{x}-\boldsymbol{z}|}{h}}^1 s^{1-d} \mathrm{d}s. \quad \text{(A.6)}$$

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 $C_2 > 0$ depending only on d and q such that

$$\|F_{\boldsymbol{x},h}\|_{L^q(B(\boldsymbol{x},h))} \leq C_2 h^{d/q}. \quad \text{(A.7)}$$

Proof.
Case $d = 1$.
We have $|F_{\boldsymbol{x},h}(\boldsymbol{z})| \leq 1$ and therefore (A.7) is satisfied with $C_2 = 2^{1/q}$.

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

$$\|F_{\boldsymbol{x},h}\|^q_{L^q(B(\boldsymbol{x},h))} \leq 2\pi \int_0^h e^{-q} h q^q \mathrm{d}\rho = q^q e^{-q} h^2.$$

This proves (A.7) with $C_2 = (2\pi)^{1/q} q e^{-1}$.

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 $C_2 = \frac{\omega_d^{1/q}}{(d-2)[(2-d)q+d]^{1/q}}$. ∎

We can now state and prove a local $W^{2,p}$ interpolation estimate for $\mathbb{P}_1$-exact gradient reconstructions.

Lemma A.3 ($W^{2,p}$ estimates for $\mathbb{P}_1$-exact gradient reconstructions). *Assume that $p > \frac{d}{2}$, and let $B \subset K \subset V$ be bounded sets 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 7.29. Let $\theta \ge \mathrm{diam}(V)/\mathrm{diam}(B)$.*
Take $\varphi \in W^{2,p}(V) \cap C(\overline{V})$ and set $v = (\varphi(\boldsymbol{x}_i))_{i\in I}$. Then, there exists $C_3 > 0$, depending only on d, p and θ, and an affine function $A_\varphi : V \to \mathbb{R}$ such that

$$\sup_{\boldsymbol{x}\in\overline{V}} |\varphi(\boldsymbol{x}) - A_\varphi(\boldsymbol{x})| \le C_3 \mathrm{diam}(V)^{2-\frac{d}{p}} \left\| |D^2\varphi| \right\|_{L^p(V)}, \tag{A.8}$$

$$\|\nabla A_\varphi - \nabla\varphi\|_{L^p(V)^d} \le C_3 \mathrm{diam}(V) \left\| |D^2\varphi| \right\|_{L^p(V)}, \tag{A.9}$$

and

$$\|\mathcal{G}v - \nabla\varphi\|_{L^p(K)^d} \le C_3 \mathrm{diam}(V)(1 + \|\mathcal{G}\|_p) \left\| |D^2\varphi| \right\|_{L^p(V)}. \tag{A.10}$$

Remark A.4. If V is sufficiently regular, $W^{2,p}(V) \subset C(\overline{V})$ and we only need to assume that $\varphi \in W^{2,p}(V)$.

Remark A.5 (Averaged Taylor polynomial)
The affine mapping A_φ is similar to an averaged Taylor polynomial of φ as in [12], but with a simpler definition since we do not need here to approximate the higher order derivatives of φ.

Proof. To simplify the notations, we let $h_B = \mathrm{diam}(B)$ 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 $\boldsymbol{y} \in B$, we write the Taylor expansion

$$\varphi(\boldsymbol{x}) = \varphi(\boldsymbol{y}) + \nabla\varphi(\boldsymbol{y})\cdot(\boldsymbol{x}-\boldsymbol{y}) + \int_0^1 sD^2\varphi(\boldsymbol{x}+s(\boldsymbol{y}-\boldsymbol{x}))(\boldsymbol{x}-\boldsymbol{y})\cdot(\boldsymbol{x}-\boldsymbol{y})\mathrm{d}s. \quad (A.11)$$

Denote by $\overline{\boldsymbol{y}}$ the center 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 (A.11) 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,

$$|\varphi(\boldsymbol{x}) - A_\varphi(\boldsymbol{x})| \le |R_1(\boldsymbol{x})| + |R_2(\boldsymbol{x})|. \quad (A.12)$$

We now find bounds on 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 (A.6). Using Hölder's inequality, the inclusion $V\subset B(\boldsymbol{x},h_V)$ and Lemma A.2 we infer

$$|R_1(\boldsymbol{x})| \le \frac{h_V^2}{|B|}\left\|\,|D^2\varphi|\,\right\|_{L^p(V)}\|F_{\boldsymbol{x},h_V}\|_{L^{p'}(B(\boldsymbol{x},h_V))} \le C_4\frac{h_V^{2+\frac{d}{p'}}}{|B|}\left\|\,|D^2\varphi|\,\right\|_{L^p(V)}$$

where C_4 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_5 depending only on θ, p and d such that

$$|R_1(\boldsymbol{x})| \le C_5 h_V^{2-\frac{d}{p}}\left\|\,|D^2\varphi|\,\right\|_{L^p(V)}. \quad (A.13)$$

Bound on R_2.
By Hölder's inequality and $|B| = |B(0,1)|h_B^d$ we have

$$\begin{aligned}|R_2(\boldsymbol{x})| &\le h_B|B|^{\frac{1}{p'}-1}\left\|\nabla\varphi - \overline{\nabla\varphi}\right\|_{L^p(B)^d}\\ &\le |B(0,1)|^{-\frac{1}{p}}h_B^{1-\frac{d}{p}}\left\|\nabla\varphi - \overline{\nabla\varphi}\right\|_{L^p(B)^d}.\end{aligned}$$

Apply Lemma A.1 with $V = B$ and φ replaced by $\partial_i\varphi$ (for $i = 1,\dots,d$). This gives C_6 depending only on d and p such that

$$|R_2(\boldsymbol{x})| \le C_6 h_B^{2-\frac{d}{p}}\left\||D^2\varphi|\right\|_{L^p(B)}. \tag{A.14}$$

Conclusion.
Combining (A.12), (A.13) and (A.14) gives (A.8). To prove (A.9), notice that

$$\nabla A_\varphi = \overline{\nabla\varphi} = \frac{1}{|B|}\int_B \nabla\varphi(\boldsymbol{y})\mathrm{d}\boldsymbol{y}$$

and apply Lemma A.1 with φ replaced by $\partial_i\varphi$, for all $i = 1,\dots,d$. This gives C_7 depending only on d an p such that

$$\|\nabla A_\varphi - \nabla\varphi\|_{L^p(V)^d} \le C_7\frac{h_V^{d/p+1}}{h_B^{d/p}}\left\||D^2\varphi|\right\|_{L^p(V)}.$$

This completes the proof of (A.9) since $h_B \ge \theta^{-1}h_V$.
Let us now turn to (A.10). Define $\xi = (A_\varphi(\boldsymbol{x}_i))_{i\in I}$ and notice that $\mathcal{G}\xi = \nabla A_\varphi$ since $\mathcal{G}$ is a $\mathbb{P}_1$-exact gradient reconstruction upon $(\boldsymbol{x}_i)_{i\in I}$. The linearity of $\mathcal{G}$ and the definition of $\|\mathcal{G}\|_p$ show that

$$\begin{aligned}\|\mathcal{G}v - \mathcal{G}\xi\|_{L^p(K)^d} &= \left\|\sum_{i\in I}(v_i-\xi_i)\mathcal{G}^i\right\|_{L^p(K)^d}\\ &\le \left\|\sum_{i\in I}|\mathcal{G}^i|\right\|_{L^p(K)}\max_{i\in I}|v_i-\xi_i|\\ &= \|\mathcal{G}\|_p|K|^{\frac{1}{p}}\mathrm{diam}(K)^{-1}\max_{i\in I}|\varphi(\boldsymbol{x}_i) - A_\varphi(\boldsymbol{x}_i)|.\end{aligned}$$

Using (A.8) and the inequality $\mathrm{diam}(K)^{-1} \le h_B^{-1} \le \theta h_V^{-1}$, we deduce

$$\|\mathcal{G}v - \mathcal{G}\xi\|_{L^p(K)} \le C_3\|\mathcal{G}\|_p\,\theta h_V^{1-\frac{d}{p}}|K|^{\frac{1}{p}}\left\||D^2\varphi|\right\|_{L^p(V)}.$$

Since $|K| \le |B(0,1)|\mathrm{diam}(K)^d \le |B(0,1)|h_V^d$, this shows that there exists C_8 depending only on θ, d and p such that

$$\|\mathcal{G}v - \mathcal{G}\xi\|_{L^p(K)} \le C_8\|\mathcal{G}\|_p\,h_V\left\||D^2\varphi|\right\|_{L^p(V)}.$$

The proof of (A.10) is complete by recalling that $\mathcal{G}\xi = \nabla A_\varphi$, by using the triangle inequality, and by invoking (A.9).

We just proved the lemma for $\varphi \in C_c^2(\mathbb{R}^d)$. We notice that all quantities and norms involved in (A.8), (A.9) and (A.10) are continuous with respect to φ for the $W^{2,p}(V) \cap C(\overline{V})$ norm. Since V is star-shaped, by a classical dilatation and regularisation argument we see 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 (A.8), (A.9) and (A.10) are still valid for $\varphi \in W^{2,p}(V) \cap C(\overline{V})$, and the proof is complete.

■

The next proposition states our main bound on $S_{\mathcal{D}}(\varphi)$ for an LLE GD, in the case $p > d/2$ and $\varphi \in W^{2,p}(\Omega)$. This is established under a slightly restrictive assumption on the points $\boldsymbol{x}_i$, which holds for most of the schemes presented in Part III.

Proposition A.6 ($W^{2,p}$ 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*

$$\begin{aligned} &\textit{(i) For all } K \in \mathcal{M} \textit{ and all } i \in I_K,\ \boldsymbol{x}_i \in \overline{K}, \\ &\textit{(ii) For all } K \in \mathcal{M}\textit{, there exists a ball } B_K \subset K \textit{ such that} \\ &\qquad K \textit{ is star-shaped with respect to all points in } B_K. \end{aligned} \tag{A.15}$$

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 $C_9 > 0$, depending only on p, d, Ω and θ, such that

$$\forall \varphi \in W^{2,p}(\Omega) \cap W_0^{1,p}(\Omega)\,,\ S_{\mathcal{D}}(\varphi) \leq C_9 h_{\mathcal{M}} \|\varphi\|_{W^{2,p}(\Omega)}\,, \tag{A.16}$$

where $S_{\mathcal{D}}$ is defined by (2.2).

Remark A.7 (Broken $W^{2,p}$ estimates)
A close examination of the proof shows that Proposition A.6 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

$$W^{2,p}(\mathcal{M}) = \{\psi \in L^p(\Omega) \,:\, \forall K \in \mathcal{M}\,,\ \psi \in W^{2,p}(K)\}.$$

We just have to replace, in (A.16), the term $\|\varphi\|_{W^{2,p}(\Omega)}$ with the broken norm

$$\|\varphi\|_{W^{2,p}(\mathcal{M})} = \left(\sum_{K\in\mathcal{M}} \|\varphi\|_{W^{2,p}(K)}^p\right)^{1/p}.$$

Proof. The regularity assumption on Ω and 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 A.3 can be applied, for any $K \in \mathcal{M}$, with $V = K$ and $\mathcal{G} = \mathcal{G}_K$. Estimate (A.10) then yields, with $h_K = \mathrm{diam}(K)$,

$$\|\nabla_{\mathcal{D}} v - \nabla\varphi\|_{L^p(K)^d} \le C_3 h_K (1 + \|\mathcal{G}_K\|_p) \left\| |D^2\varphi| \right\|_{L^p(K)} \tag{A.17}$$

where C_3 depends only on p, d and θ. Raising to the power p and summing over $K \in \mathcal{M}$ leads to

$$\|\nabla_{\mathcal{D}} v - \nabla\varphi\|_{L^p(\Omega)^d} \le C_3 h_{\mathcal{M}} (1 + \mathrm{reg}_{\mathrm{LLE}}(\mathcal{D})) \left\| |D^2\varphi| \right\|_{L^p(\Omega)}. \tag{A.18}$$

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 A.3, write

$$\begin{aligned} |\varphi(\boldsymbol{x}) - \varphi(\boldsymbol{y})| &\le |\varphi(\boldsymbol{x}) - A_\varphi(\boldsymbol{x})| + |A_\varphi(\boldsymbol{x}) - A_\varphi(\boldsymbol{y})| + |A_\varphi(\boldsymbol{y}) - \varphi(\boldsymbol{y})| \\ &\le 2C_3 h_K^{2-\frac{d}{p}} \left\| |D^2\varphi| \right\|_{L^p(K)} + |\nabla A_\varphi| h_K. \end{aligned} \tag{A.19}$$

Since ∇A_φ is constant, (A.9) gives

$$\begin{aligned} |\nabla A_\varphi| &= |K|^{-\frac{1}{p}} \|\nabla A_\varphi\|_{L^p(K)^d} \\ &\le |K|^{-\frac{1}{p}} \|\nabla\varphi\|_{L^p(K)^d} + |K|^{-\frac{1}{p}} C_3 h_K \left\| |D^2\varphi| \right\|_{L^p(K)}. \end{aligned}$$

Plugged into (A.19), this yields

$$\begin{aligned} &|\varphi(\boldsymbol{x}) - \varphi(\boldsymbol{y})| \\ &\quad \le \left(2C_3 \mathrm{diam}(\Omega) h_K^{1-\frac{d}{p}} + (1 + C_3 \mathrm{diam}(\Omega)) h_K |K|^{-\frac{1}{p}} \right) \|\varphi\|_{W^{2,p}(K)}. \end{aligned}$$

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_{10} depending only on Ω, p and θ such that

$$|\varphi(\boldsymbol{x}) - \varphi(\boldsymbol{y})| \le C_{10} h_K |K|^{-\frac{1}{p}} \|\varphi\|_{W^{2,p}(K)}. \tag{A.20}$$

Recalling the relation (7.33) between $\Pi_{\mathcal{D}}$ and the elementary basis functions $(\pi_K^i)_{i \in I_K}$, (A.20) gives, for a.e. $\boldsymbol{x} \in K$,

$$\begin{aligned} |\Pi_{\mathcal{D}} v(\boldsymbol{x}) - \varphi(\boldsymbol{x})| &= \left| \sum_{i \in I_K} \pi_K^i(\boldsymbol{x}) (v_i - \varphi(\boldsymbol{x})) \right| \\ &\le \sup_{i \in I_K} |\varphi(\boldsymbol{x}_i) - \varphi(\boldsymbol{x})| \sum_{i \in I_K} |\pi_K^i(\boldsymbol{x})| \\ &\le C_{10} h_K \|\varphi\|_{W^{2,p}(K)} |K|^{-\frac{1}{p}} \sum_{i \in I_K} |\pi_K^i(\boldsymbol{x})|. \end{aligned}$$

Take the $L^p(K)$ norm over $\boldsymbol{x} \in K$ and recall the definition (7.23) of $\|\pi_K\|_p$ to deduce

$$\|\Pi_\mathcal{D} v - \varphi\|_{L^p(K)} \leq C_{10} h_K \|\varphi\|_{W^{2,p}(K)} \|\pi_K\|_p . \tag{A.21}$$

As (A.17), this estimate in K translates into the global estimate

$$\|\Pi_\mathcal{D} v - \varphi\|_{L^p(\Omega)} \leq C_{10} h_\mathcal{M} \|\varphi\|_{W^{2,p}(\Omega)} \operatorname{reg}_{\mathrm{LLE}}(\mathcal{D}). \tag{A.22}$$

The proof is complete by combining (A.18) and (A.22). ∎

Assumption (A.15) ensures that local errors estimates can be computed on a *mesh* of the domain (with non-overlapping sets). This ensure that, when added up, 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 in Part III, noticeably the SUSHI and VAG schemes (and any other barycentric condensation of an LLE GD, when some DOFs in I_K are eliminated by using other DOFs that lie outside K – see Definition 7.38).

Proposition A.8 ($W^{2,p}$ 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

$$\begin{aligned} &\textit{(i) For all } i \in I_K,\ \boldsymbol{x}_i \in \overline{V_K}, \\ &\textit{(ii) There exists a ball } B_K \subset K \textit{ such that } V_K \textit{ is star-shaped} \\ &\qquad \textit{with respect to all points of } B_K. \end{aligned} \tag{A.23}$$

Let

$$\theta \geq \operatorname{reg}_{\mathrm{LLE}}(\mathcal{D}) + \max_{K\in\mathcal{M}} \frac{\operatorname{diam}(V_K)}{\operatorname{diam}(B_K)} + \operatorname*{esssup}_{\boldsymbol{x}\in\Omega} \operatorname{Card}(\{K \in \mathcal{M} : \boldsymbol{x} \in V_K\}). \tag{A.24}$$

Then, there exists $C_{11} > 0$, depending only on p, d, Ω and θ, such that

$$\forall \varphi \in W^{2,p}(\Omega) \cap W_0^{1,p}(\Omega),\ S_\mathcal{D}(\varphi) \leq C_{11} h_\mathcal{M} \|\varphi\|_{W^{2,p}(\Omega)},$$

where $S_\mathcal{D}$ is defined by (2.2).

Remark A.9. 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 A.6 and use Lemma A.3 with $V = V_K$ to arrive, in a similar way as for (A.17) and (A.21), to

$$\|\Pi_{\mathcal{D}} v - \varphi\|_{L^p(K)} + \|\nabla_{\mathcal{D}} v - \nabla\varphi\|_{L^p(K)^d} \le C_{12}\mathrm{diam}(V_K)\,\|\varphi\|_{W^{2,p}(V_K)}\,,$$

where C_{12} depends only on p, d, Ω and θ. Since $\mathrm{diam}(V_K) \le \theta\mathrm{diam}(B_K) \le \theta h_{\mathcal{M}}$, raising to the power p gives C_{13} depending only on p, d, Ω and θ such that

$$\|\Pi_{\mathcal{D}} v - \varphi\|^p_{L^p(K)} + \|\nabla_{\mathcal{D}} v - \nabla\varphi\|^p_{L^p(K)^d} \le C_{13} h^p_{\mathcal{M}}\,\|\varphi\|^p_{W^{2,p}(V_K)}\,.$$

Summing over $K \in \mathcal{M}$ yields

$$\|\Pi_{\mathcal{D}} v - \varphi\|^p_{L^p(\Omega)} + \|\nabla_{\mathcal{D}} v - \nabla\varphi\|^p_{L^p(\Omega)^d} \le C^p_{13} h^p_{\mathcal{M}} \sum_{K\in\mathcal{M}} \|\varphi\|^p_{W^{2,p}(V_K)}\,. \quad \text{(A.25)}$$

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 $\sum_{K\in\mathcal{M}} \mathbf{1}_{V_K}(\boldsymbol{x}) = \mathrm{Card}(\{K \in \mathcal{M} : \boldsymbol{x} \in V_K\}) \le \theta$ for a.e. $\boldsymbol{x} \in \Omega$. Hence,

$$\sum_{K\in\mathcal{M}} \|g\|^p_{L^p(V_K)} \le \theta \int_\Omega |g|^p \mathrm{d}\boldsymbol{x} = \theta\,\|g\|^p_{L^p(\Omega)}\,.$$

The proof is complete by using this estimate in (A.25) 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 A.10 ($W^{2,p}$ 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 7.50. 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* (A.15) *holds and take $\theta \ge \mathrm{reg}_{\mathrm{LLE}}(\mathcal{D}) + \max_{K\in\mathcal{M}} \frac{\mathrm{diam}(K)}{\mathrm{diam}(B_K)}$. Take $\varphi \in W^{2,p}(\Omega)$ and assume that*

$$\forall i \in I_\partial\,,\ (\mathcal{I}_{\mathcal{D},\partial}\gamma(\varphi))_i = \varphi(\boldsymbol{x}_i). \quad \text{(A.26)}$$

Then, there exists $C_{14} > 0$, depending only on p, d, Ω and θ, such that

$$S_{\mathcal{D}}(\varphi) \le C_{14} h_{\mathcal{M}}\,\|\varphi\|_{W^{2,p}(\Omega)}\,, \quad \text{(A.27)}$$

where $S_{\mathcal{D}}$ is defined by (2.14).

Proof. Assumption (A.26) 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 (2.14) of $S_{\mathcal{D}}$. Since v satisfies the estimates (A.18) and (A.22) (which have been established without using the boundary value of φ), this completes the proof. ∎

Proposition A.11 ($W^{2,p}$ 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 A.10, except* (A.26) *which is replaced by*

$$\begin{aligned}&\forall K \in \mathcal{M}\,,\ \textit{there exists } C_K(\varphi) \geq 0 \ \textit{s.t.}\\ &\max_{i\in I_K\cap I_\partial} |(\mathcal{I}_{\mathcal{D},\partial}\gamma\varphi)_i - \varphi(\boldsymbol{x}_i)| \leq h_{\mathcal{M}}\mathrm{diam}(K)|K|^{-\frac{1}{p}}C_K(\varphi).\end{aligned} \tag{A.28}$$

Then, there exists C_{15} depending only on p, d, Ω, and θ, such that

$$S_{\mathcal{D}}(\varphi) \leq C_{15}h_{\mathcal{M}}\left(\|\varphi\|_{W^{2,p}(\Omega)} + \left(\sum_{K\in\mathcal{M}} C_K(\varphi)^p\right)^{1/p}\right). \tag{A.29}$$

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 A.11 is such that $\mathcal{I}_{\mathcal{D},\partial}\gamma(\varphi)$ is usually re-defined so that (A.26) holds. See Remarks 2.19 and 12.2.

Proof. The estimates established in the proof of Proposition A.6 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,

$$\|\Pi_{\mathcal{D}}v - \varphi\|_{L^p(\Omega)} + \|\nabla_{\mathcal{D}}v - \nabla\varphi\|_{L^p(\Omega)^d} \leq C_{16}h_{\mathcal{M}}\|\varphi\|_{W^{2,p}(\Omega)}\,, \tag{A.30}$$

where C_{16} depends only on d, p, Ω and θ.
Let us now consider $w \in X_{\mathcal{D}}$ as in the proof of Proposition 7.51, 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 (A.28) the quantity $\omega(K)$ defined by (7.62) satisfies $\omega(K) \leq h_{\mathcal{M}}|K|^{-\frac{1}{p}}C_K(\varphi)$. Plug this estimate into (7.63), raise the result to the power p and sum over $K \in \mathcal{M}$. This gives

$$\|\nabla_{\mathcal{D}}v - \nabla_{\mathcal{D}}w\|_{L^p(\Omega)^d} \leq \theta h_{\mathcal{M}}C_\Omega(\varphi), \tag{A.31}$$

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. $\boldsymbol{x} \in K$,

$$|\Pi_{\mathcal{D}} v(\boldsymbol{x}) - \Pi_{\mathcal{D}} w(\boldsymbol{x})| \le \sum_{i \in I_K} |\pi_K^i(\boldsymbol{x})| \, |v_i - w_i|$$
$$\le h_{\mathcal{M}} \mathrm{diam}(K) C_K(\varphi) |K|^{-\frac{1}{p}} \sum_{i \in I_K} |\pi_K^i(\boldsymbol{x})|.$$

Taking the $L^p(K)$ norm, recalling the definition (7.23) of $\|\pi_K\|_p$, raising to the power p and summing on $K \in \mathcal{M}$ leads to $\|\Pi_{\mathcal{D}} v - \Pi_{\mathcal{D}} w\|_{L^p(\Omega)} \le \theta h_{\mathcal{M}}^2 C_\Omega(\varphi)$. The proof is complete by combining this estimate with (A.31) and (A.30). ∎

Since the estimates (A.18) and (A.22) were obtained without referring to the boundary values of φ, they immediately give the following result.

Proposition A.12 ($W^{2,p}$ 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 7.52. 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* (A.15) *holds and 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, depending only on p, d, Ω and θ, such that*

$$\forall \varphi \in W^{2,p}(\Omega), \ S_{\mathcal{D}}(\varphi) \le C h_{\mathcal{M}} \|\varphi\|_{W^{2,p}(\Omega)},$$

where $S_{\mathcal{D}}$ is defined by (2.20).

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 (A.15), then the estimates of $\mathbb{T}_{\mathcal{D}}$ can be obtained as the estimates on $\Pi_{\mathcal{D}}$ in the proof of Proposition A.6.

Proposition A.13 ($W^{2,p}$ 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 7.55. 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* (A.15) *holds and, with $H_1, \dots, H_r$ hyperplanes whose union covers $\partial\Omega$, that*

(i) For any $K_\partial \in \mathcal{M}_\partial$ there is $\ell_{K_\partial} \in \{1, \dots, r\}$ 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

$$\theta \ge \mathrm{reg}_{\mathrm{LLE}}(\mathcal{D}) + \max_{K \in \mathcal{M}} \frac{\mathrm{diam}(K)}{\mathrm{diam}(B_K)} + \max_{K_\partial \in \mathcal{M}_\partial} \frac{\mathrm{diam}(K_\partial)}{\mathrm{diam}(B_{K_\partial})}.$$

Then, there exists $C_{17} > 0$, 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 $\ell = 1, \dots, r$,

$$S_{\mathcal{D}}(\varphi) \leq C_{17} h_{\mathcal{M}} \left(\|\varphi\|_{W^{2,p}(\Omega)} + \sum_{\ell=1}^{r} \|\gamma(\varphi)\|_{W^{2,p}(\partial\Omega \cap H_\ell)} \right),$$

where $S_{\mathcal{D}}$ is defined by (2.49).

A.2 LLE GDs with generalised degrees of freedom

The definition 7.29 of $\mathbb{P}_1$-exact gradient reconstructions ($\mathbb{P}_1$-exact GR in short) implicitly assume that the DOFs 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 degrees of freedom; for example, degrees of freedom 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 degrees of freedom. It makes sense to also generalise the definition to higher order reconstructions.

Definition A.14 ($\mathbb{P}_{k+1}$-exact GR with generalised dof). *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 degrees of freedom 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 7.29 corresponds to $P_i(\varphi) = \varphi(\boldsymbol{x}_i)$.

In a similar way as in Definition A.14, these gradient reconstructions could be used to design a notion of "$\mathbb{P}_{k+1}$-exact GDs with generalised degrees of freedom" 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 do not pursue further this idea here, and we let the interested reader fill in the details.

A.3 Non-linearly exact barycentric combinations

Let us consider a heterogeneous material, with a discontinous diffusion tensor Λ which is smooth inside subdomains $P_1,\dots,P_k$ (partition of Ω). The solution to (3.1) 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 degrees of freedom $i \in I\backslash I^{\mathrm{BA}}$, 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 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 $j\in I^{\mathrm{BA}}$ then, for the degrees of freedom $i\in I\backslash I^{\mathrm{BA}}$ such that H_i spread over several subdomains, the values $\widetilde{v}_i$ defined by (7.43) are no longer good (order $\mathrm{diam}(K)^2$) approximations of $\overline{u}(\boldsymbol{x}_i)$, and therefore $(\nabla_{\mathcal{D}^{\mathrm{BA}}} v)_{|K} = (\nabla_{\mathcal{D}}\widetilde{v})_{|K}$ will not approximate $(\nabla\overline{u})_{|K}$ properly. This does not prevent the corresponding GS from converging, but leads to reduce accuracy on coarse meshes.

Barycentric condensation preserves the LLE property thanks to Assumption (7.42); it's this assumptions 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 assumptions and create barycentric condensations that do not satisfy (7.42), but rather relations that ensures that, if v interpolates $\overline{u}$ at $(\boldsymbol{x}_j)_{j\in I^{\mathrm{BA}}}$, the values $\widetilde{v}$ computed through (7.43) 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 A.15 ($\mathcal{S}$-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 7.38, without Assumption* (7.42) *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*

$$\forall K \in \mathcal{M}\,,\ \forall i \in I_K \backslash I^{\mathrm{BA}}\ : \quad \left|\varphi(\boldsymbol{x}_i) - \sum_{j \in H_i} \beta_j^i \varphi(\boldsymbol{x}_j)\right| \leq C_\varphi R_{\mathcal{D}^{\mathcal{S}}} \mathrm{diam}(K)^2. \quad \text{(A.32)}$$

The $\mathcal{S}$*-regularity of* $\mathcal{D}^{\mathcal{S}}$ *is then defined by*

$$\mathrm{reg}_{\mathcal{S}}(\mathcal{D}^{\mathcal{S}}) = \mathrm{reg}_{\mathrm{BA}}(\mathcal{D}^{\mathcal{S}}) + R_{\mathcal{D}^{\mathcal{S}}} + \max_{K \in \mathcal{M}} \frac{\mathrm{diam}(O_K)}{\mathrm{diam}(K)}.$$

Linearly exact barycentric condensations (*i.e.* in the sense of Definition 7.38) are $\mathcal{S}$-adapted barycentric condensations with $\mathcal{S} = C_c^\infty(\Omega)$ and O_K the interior of the convex hull of $(\boldsymbol{x}_i)_{i \in I_K^{\mathrm{BA}}}$.
The following theorem is an equivalent of Theorem 7.41 for $\mathcal{S}$-adapted barycentric condensations.

Theorem A.16 (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, 2.6 and 2.8. Fix a subset* $\mathcal{S}$ *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* $(\mathrm{reg}_{\mathrm{LLE}}(\mathcal{D}_m))_{m \in \mathbb{N}}$ *and* $(\mathrm{reg}_{\mathcal{S}}(\mathcal{D}_m^{\mathcal{S}}))_{m \in \mathbb{N}}$ *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 7.41 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 (7.42). Hence those 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 7.36, define the interpolant $v \in X_{\mathcal{D}^{\mathcal{S}},0}$ by $v_i = \varphi(\boldsymbol{x}_i)$ for all $i \in I^{\mathrm{BA}}$. Let $\widetilde{v} \in X_{\mathcal{D},0}$ be given by (7.43), that is $\widetilde{v}_i = v_i = \varphi(\boldsymbol{x}_i)$ if $i \in I^{\mathrm{BA}}$ and

$$\widetilde{v}_i = \sum_{j \in H_i} \beta_j^i v_j = \sum_{j \in H_i} \beta_j^i \varphi(\boldsymbol{x}_i) \quad \text{if } i \in I \backslash I^{\mathrm{BA}}.$$

By (A.32) we have $|\widetilde{v}_i - \varphi(\boldsymbol{x}_i)| \leq C_\varphi R_{\mathcal{D}^{\mathcal{S}}} \mathrm{diam}(K)^2$ if $i \in I_K$. Hence

$$\forall K \in \mathcal{M}\,,\ \forall i \in I_K\ :\ \widetilde{v}_i = \varphi(\boldsymbol{x}_i) + \mathcal{O}(\mathrm{diam}(K)^2). \quad \text{(A.33)}$$

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 (7.30) holds up to an additional term $\mathcal{O}(\mathrm{diam}(O_K)^2) = \mathcal{O}(\mathrm{diam}(K)^2)$. Still following the computations in the proof of Lemma 7.31, the $W^{2,\infty}(O_K)$-regularity of φ then shows that $\|\mathcal{G}_K\widetilde{v} - \nabla\varphi\|_{L^p(K)^d} = \mathcal{O}(|K|^{1/p}\mathrm{diam}(K))$ on K. This gives

$$\|\nabla_{\mathcal{D}^{\mathcal{S}}} v - \nabla\varphi\|_{L^p(\Omega)^d} = \|\nabla_{\mathcal{D}}\widetilde{v} - \nabla\varphi\|_{L^p(\Omega)^d} = \mathcal{O}(h_{\mathcal{M}}). \tag{A.34}$$

The property (A.33), the definition (7.33) of $\Pi_{\mathcal{D}}$ and the boundedness of $\mathrm{reg}_{\mathrm{LLE}}(\mathcal{D})$ also give

$$\|\Pi_{\mathcal{D}^{\mathcal{S}}} v - \varphi\|_{L^p(\Omega)} = \|\Pi_{\mathcal{D}}\widetilde{v} - \varphi\|_{L^p(\Omega)} = \mathcal{O}(h_{\mathcal{M}}). \tag{A.35}$$

Estimates (A.34) and (A.35) show that $S_{\mathcal{D}^{\mathcal{S}}}(\varphi) = \mathcal{O}(h_{\mathcal{M}})$ for any $\varphi \in \mathcal{S}$. The proof is complete by invoking Lemma 2.13 and the density of $\mathcal{S}$ in $W_0^{1,p}(\Omega)$. ∎

B

Discrete functional analysis

A number of numerical methods are based on primary unknowns located at the cells and/or at the faces of polytopal meshes. For many other methods, such secondary unknowns can be defined from primary unknowns located, e.g., at the vertices. The aim of this section is to introduce discrete functional analysis tools for the study of methods with (primary or secondary) unknowns at the cells and faces of a mesh. These tools are essential to the notion of polytopal toolbox, and to obtain the coercivity, limit-conformity and compactness of GDs by controlling them by polytopal toolboxes (see Chapter 7).

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 center $\boldsymbol{x}_K$ 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).
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 $\varrho_K \le d_{K,\sigma}$. Hence

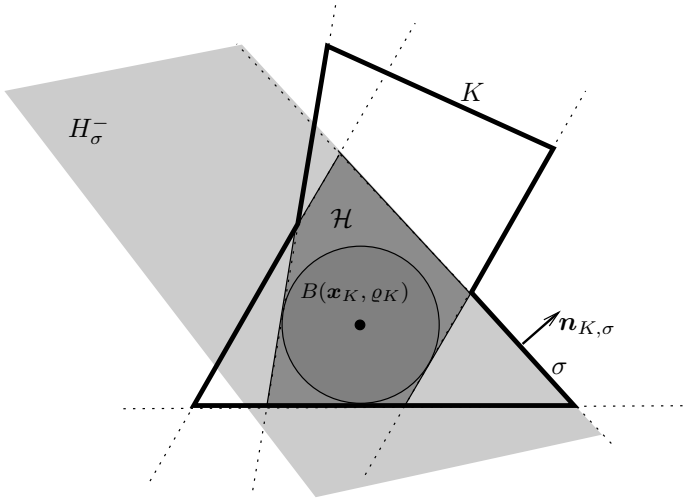


Fig. B.1. Illustration of the proof of Lemma B.1.

$B(\boldsymbol{x}_K, \varrho_K) \subset \cap_{\sigma\in\mathcal{F}_K} H_\sigma^- =: \mathcal{H}$. 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.
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

$$|D_{K,\sigma}| = \int_0^1 \int_\sigma t^{d-1} d_{K,\sigma} \mathrm{d}t\mathrm{d}s(\boldsymbol{y}) = \frac{1}{d} d_{K,\sigma}|\sigma|,$$

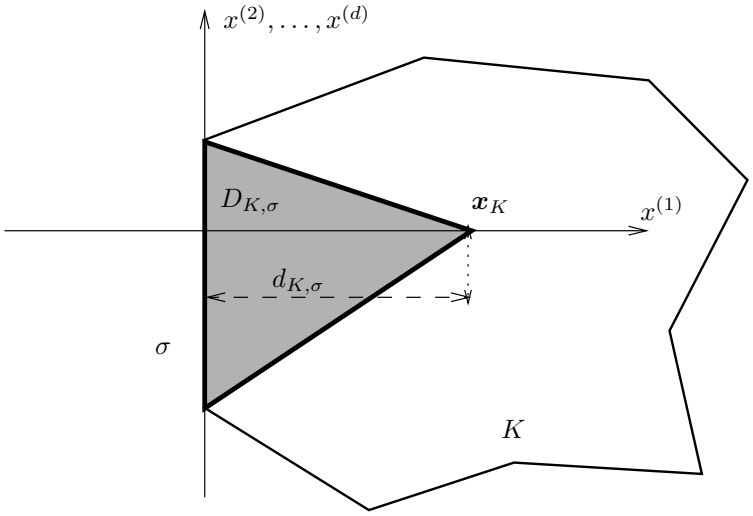


Fig. B.2. Illustration of the proof of Lemma B.2

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 barycenter 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 center 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 $i = 1,\dots,d$, 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 $\boldsymbol{\xi}\in\mathbb{R}^d$:

$$\left(\sum_{\sigma\in\mathcal{F}_K} |\sigma|\boldsymbol{n}_{K,\sigma}\right)\cdot\boldsymbol{\xi} = \sum_{\sigma\in\mathcal{F}_K}\int_\sigma \boldsymbol{\xi}\cdot\boldsymbol{n}_{K,\sigma}\mathrm{d}s(\boldsymbol{x}) = \int_K \mathrm{div}(\boldsymbol{\xi})\mathrm{d}\boldsymbol{x} = 0.$$

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 center of mass of K, and ρ_K be the maximum radius of the balls centered 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*

$$\rho_K = \min_{\sigma\in\mathcal{F}_K} d_{K,\sigma}\,, \tag{B.5}$$

$$\forall \boldsymbol{s}_0\neq\boldsymbol{s}_1 \text{ in } \mathcal{V}_K\,,\ \rho_K \le \frac{1}{d+1}\mathrm{dist}(\boldsymbol{s}_0,\boldsymbol{s}_1)\,, \tag{B.6}$$

$$\forall\sigma\in\mathcal{F}_K\,,\ \rho_K \le \frac{1}{d+1}\mathrm{diam}(\sigma). \tag{B.7}$$

As a consequence, if $\mathfrak{T}$ is a conforming simplicial mesh with $\mathcal{P}$ the centers of mass of the cells, then, recalling the definitions (7.8)–(7.10)*,*

$$\eta_{\mathfrak{T}} \le \frac{2\kappa_{\mathfrak{T}}^2}{d+1} \quad \text{and} \quad \theta_{\mathfrak{T}} \le \kappa_{\mathfrak{T}} + d + 1.$$

Proof. The inequality $\ge$ 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 neighbourhing 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 center 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 Approximation properties

The following result is the key to proving that several classical gradient discretisations are LLE GDs.

Lemma B.6 ($\mathbb{P}_1$-exactness of $\overline{\nabla}_{\mathfrak{T}}$, and estimate). *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 7.29. 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 $v \in X_{\mathfrak{T}}$,*
$$\left\|\overline{\nabla}_{\mathfrak{T}} v\right\|_{L^p(\Omega)^d} \le d^{\frac{p-1}{p}} |v|_{\mathfrak{T},p}\,. \tag{B.10}$$

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

$$|\overline{\nabla}_{\mathfrak{T}} v(\boldsymbol{x})|^p \le d^p \sum_{\sigma\in\mathcal{F}_K} \frac{|\sigma| d_{K,\sigma}}{d|K|}\left|\frac{v_\sigma - v_K}{d_{K,\sigma}}\right|^p = \frac{d^{p-1}}{|K|}\sum_{\sigma\in\mathcal{F}_K} |\sigma| d_{K,\sigma}\left|\frac{v_\sigma - v_K}{d_{K,\sigma}}\right|^p.$$

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.10). ∎

The following lemma is particularly useful to obtain estimates on interpolants of $W^{1,p}$ functions (see, e.g., the proofs of Theorem 7.12 and Proposition 7.15).

Lemma B.7. *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*

$$\max\left\{\frac{h_K}{d_{K,\sigma}} \,:\, K \in \mathcal{M}\,,\ \sigma \in \mathcal{F}_K\right\} \le \theta.$$

Then, there exists C_{18} depending only on d, p and θ such that, for any $\varphi \in W^{1,p}(\Omega)$ and any $K \in \mathcal{M}$,

$$\left|\frac{1}{|\sigma|}\int_\sigma \varphi(\boldsymbol{x})\mathrm{d}s(\boldsymbol{x}) - \frac{1}{|K|}\int_K \varphi(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\right|^p \le \frac{C_{18}h_K^{p-1}}{|\sigma|}\int_K |\nabla\varphi(\boldsymbol{x})|^p\mathrm{d}\boldsymbol{x} \quad \text{(B.11)}$$

and

$$\left\|\varphi - \frac{1}{|K|}\int_K \varphi(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\right\|_{L^p(K)} \le C_{18}h_K \,\|\,|\nabla\varphi|\,\|_{L^p(K)} \quad \text{(B.12)}$$

Proof. The proof is based on estimates first established in [31, 32]. Let us assume that there exists C_{19} depending only on d, p and θ such that, for all $K \in \mathcal{M}$ and all $\sigma \in \mathcal{F}_K$, setting $B_K = B(\boldsymbol{x}_K, \theta^{-1}h_K/2)$,

$$\left|\frac{1}{|\sigma|}\int_\sigma \varphi(\boldsymbol{x})\mathrm{d}s(\boldsymbol{x}) - \frac{1}{|B_K|}\int_{B_K} \varphi(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\right|^p \le C_{19}\frac{h_K^{p-1}}{|\sigma|}\int_K |\nabla\varphi(\boldsymbol{x})|^p\mathrm{d}\boldsymbol{x}, \quad \text{(B.13)}$$

$$\left|\frac{1}{|B_K|}\int_{B_K} \varphi(\boldsymbol{x})\mathrm{d}\boldsymbol{x} - \frac{1}{|K|}\int_K \varphi(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\right|^p \le C_{19}\frac{h_K^{p}}{|K|}\int_K |\nabla\varphi(\boldsymbol{x})|^p\mathrm{d}\boldsymbol{x}, \quad \text{(B.14)}$$

and

$$\left\|\varphi - \frac{1}{|B_K|}\int_{B_K} \varphi(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\right\|_{L^p(K)} \le C_{19}h_K \,\|\,|\nabla\varphi|\,\|_{L^p(K)}\,. \quad \text{(B.15)}$$

Then (B.11) follows from (B.13) and (B.14) by using the triangle inequality 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 (B.13), (B.14) and (B.15), notice first that, since $C^\infty(\overline{K})$ is dense in $W^{1,p}(K)$ (K is a polytopal set), these estimates only need to be establised for φ smooth.

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 $z \in B_K$ and $\boldsymbol{y} \in \sigma$ and using Jensen's inequality yields

$$L_{(\text{B.13})} \le \frac{h_K^p}{|\sigma|\,|B_K|}\int_0^1\int_\sigma\int_{B_K} |\nabla\varphi(\boldsymbol{z}+t(\boldsymbol{y}-\boldsymbol{z}))|^p\mathrm{d}\boldsymbol{z}\mathrm{d}s(\boldsymbol{y})\mathrm{d}t, \quad \text{(B.16)}$$

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,

$$|(\boldsymbol{y}-\boldsymbol{z})\cdot\boldsymbol{n}_{K,\sigma}| \geq |(\boldsymbol{y}-\boldsymbol{x}_K)\cdot\boldsymbol{n}_{K,\sigma}| - |\boldsymbol{z}-\boldsymbol{x}_K| \geq d_{K,\sigma} - \frac{\theta^{-1}h_K}{2} \geq \frac{\theta^{-1}h_K}{2}.$$

Hence,

$$J\psi \geq \left(\frac{|\boldsymbol{x}-\boldsymbol{z}|}{h_K}\right)^{d-1} \frac{\theta^{-1}}{2} h_K \geq (2\theta)^{-1} h_K^{2-d} |\boldsymbol{x}-\boldsymbol{z}|^{d-1}.$$

Using ψ in (B.16) therefore leads to

$$L_{(\mathrm{B.13})} \leq \frac{2\theta h_K^{p+d-2}}{|\sigma|\,|B_K|} \int_K |\nabla\varphi(\boldsymbol{x})|^p \int_{B_K} |\boldsymbol{x}-\boldsymbol{z}|^{1-d} \mathrm{d}\boldsymbol{z}\mathrm{d}\boldsymbol{x}. \tag{B.17}$$

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 $\mathbb{R}^d$,

$$\int_{B_K} |\boldsymbol{x}-\boldsymbol{z}|^{1-d}\mathrm{d}\boldsymbol{z} \leq \int_{B(\boldsymbol{x},h_K)} |\boldsymbol{x}-\boldsymbol{z}|^{1-d}\mathrm{d}\boldsymbol{z} = \omega_d \int_0^{h_K} \rho^{1-d}\rho^{d-1}\mathrm{d}\rho = \omega_d h_K.$$

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)

The proof follows similar ideas as in the proof of Lemma A.1. 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 $L_{(\mathrm{B.14})}$ the left-hand side of (B.14), Jensen's inequality gives

$$L_{(\mathrm{B.14})} \leq \frac{h_K^p}{|B_K|\,|K|} \int_{B_K}\int_K\int_0^1 |\nabla\varphi(t\boldsymbol{x}+(1-t)\boldsymbol{y})|^p \mathrm{d}t\mathrm{d}\boldsymbol{y}\mathrm{d}\boldsymbol{x}. \tag{B.19}$$

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

$$\int_{B_K}\int_K\int_0^1 |\nabla\varphi(t\boldsymbol{x}+(1-t)\boldsymbol{y})|^p \mathrm{d}t\mathrm{d}\boldsymbol{y}\mathrm{d}\boldsymbol{x} \leq \int_K |\nabla\varphi(\boldsymbol{z})|^p \int_K \int_{I(\boldsymbol{z},\boldsymbol{y})} t^{-d}\mathrm{d}t\mathrm{d}\boldsymbol{y}\mathrm{d}\boldsymbol{z} \tag{B.20}$$

where, as in the proof of Lemma A.1 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 (A.4) and (A.5), we arrive at

$$\int_K \int_{I(\boldsymbol{z},\boldsymbol{y})} t^{-d}\mathrm{d}t\mathrm{d}\boldsymbol{y} \leq \frac{h_K^d}{d-1}\omega_d. \tag{B.21}$$

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 A.1 since $V = K$ is star-shaped with respect to $B = B_K$, and $\mathrm{diam}(V) = h_K = \theta\mathrm{diam}(B)$. ∎

The following lemma is an enabler to prove the limit-conformity of a GDs controlled by a polytopal toolbox.

Lemma B.8 (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_{20} depending only on d, p and θ such that, for all $\boldsymbol{\varphi} \in W^{1,p'}(\Omega)^d$ and all $v \in X_{\mathfrak{T}}$,*

$$\Big| \int_\Omega \big(\overline{\nabla}_{\mathfrak{T}} v(\boldsymbol{x}) \cdot \boldsymbol{\varphi}(\boldsymbol{x}) + \Pi_{\mathfrak{T}} v(\boldsymbol{x}) \mathrm{div}\boldsymbol{\varphi}(\boldsymbol{x})\big)\,\mathrm{d}\boldsymbol{x} - \int_{\partial\Omega} \mathbb{T}_{\mathfrak{T}} v(\boldsymbol{x}) \gamma_{\mathbf{n}}(\boldsymbol{\varphi})(\boldsymbol{x}) \mathrm{d}s(\boldsymbol{x}) \Big| \leq C_{20} \,\| \,|\nabla\boldsymbol{\varphi}|\, \|_{L^{p'}(\Omega)} \,|v|_{\mathfrak{T},p}\, h_{\mathcal{M}}, \quad \text{(B.22)}$$

where $\gamma_{\mathbf{n}}(\boldsymbol{\varphi}) = \gamma(\boldsymbol{\varphi}) \cdot \boldsymbol{n}_{\partial\Omega}$ is the normal trace of $\boldsymbol{\varphi}$.

Remark B.9 (Broken $W^{1,p'}$ estimate)
The proof actually shows that the result still holds if we take $\boldsymbol{\varphi} \in W^{\mathrm{div},p'}(\Omega) \cap W^{1,p'}(\mathcal{M})^d$, 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.22), $\| \,|\nabla\boldsymbol{\varphi}|\, \|_{L^{p'}(\Omega)}$ must simply be replaced with

$$|\boldsymbol{\varphi}|_{W^{1,p'}(\mathcal{M})} = \left(\sum_{K\in\mathcal{M}} \| \,|\nabla\boldsymbol{\varphi}|\, \|_{L^{p'}(K)}^{p'} \right)^{1/p'}$$

(or $|\boldsymbol{\varphi}|_{W^{1,\infty}(\mathcal{M})} = \max_{K\in\mathcal{M}} \| \,|\nabla\boldsymbol{\varphi}|\, \|_{L^\infty(K)}$ if $p = 1$).

Proof. Set $\boldsymbol{\varphi}_\sigma = \frac{1}{|\sigma|}\int_\sigma \boldsymbol{\varphi}(\boldsymbol{x})\mathrm{d}s(\boldsymbol{x})$. Since $\boldsymbol{n}_{K,\sigma} = -\boldsymbol{n}_{L,\sigma}$ whenever σ is a face between K and L, gathering by faces shows that

$$\begin{aligned}
&\sum_{K\in\mathcal{M}} \sum_{\sigma\in\mathcal{F}_K} v_\sigma |\sigma| \boldsymbol{\varphi}_\sigma \cdot \boldsymbol{n}_{K,\sigma} \\
&\qquad = \sum_{\sigma\in\mathcal{F}_{\mathrm{int}},\, \mathcal{M}_\sigma = \{K,L\}} v_\sigma |\sigma| \left(\boldsymbol{\varphi}_\sigma \cdot \boldsymbol{n}_{K,\sigma} + \boldsymbol{\varphi}_\sigma \cdot \boldsymbol{n}_{L,\sigma}\right) \\
&\qquad\quad + \sum_{\sigma\in\mathcal{F}_{\mathrm{ext}},\, \mathcal{M}_\sigma = \{K\}} v_\sigma \int_\sigma \boldsymbol{\varphi}(\boldsymbol{x}) \cdot \boldsymbol{n}_{K,\sigma} \mathrm{d}s(\boldsymbol{x})
\end{aligned}$$

$$= \int_{\partial\Omega} \mathbb{T}_{\mathfrak{T}} v(\boldsymbol{x}) \boldsymbol{\varphi}(\boldsymbol{x}) \cdot \boldsymbol{n}_{\partial\Omega}(\boldsymbol{x}) \mathrm{d}s(\boldsymbol{x}).$$

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,

$$\begin{aligned}\int_\Omega \Pi_{\mathfrak{T}} v(\boldsymbol{x}) \mathrm{div}\boldsymbol{\varphi}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} &= \sum_{K\in\mathcal{M}} v_K \sum_{\sigma\in\mathcal{F}_K} |\sigma| \boldsymbol{\varphi}_\sigma \cdot \boldsymbol{n}_{K,\sigma} \\ &= \sum_{K\in\mathcal{M}} \sum_{\sigma\in\mathcal{F}_K} (v_K - v_\sigma) |\sigma| \boldsymbol{\varphi}_\sigma \cdot \boldsymbol{n}_{K,\sigma} + \int_{\partial\Omega} \mathbb{T}_{\mathfrak{T}} v(\boldsymbol{x}) \gamma_{\boldsymbol{n}}(\boldsymbol{\varphi})(\boldsymbol{x}) \mathrm{d}s(\boldsymbol{x}). \quad \text{(B.23)}\end{aligned}$$

Introduce $\boldsymbol{\varphi}_K = \frac{1}{|K|}\int_K \boldsymbol{\varphi}(\boldsymbol{x})\mathrm{d}\boldsymbol{x}$ and write, since $\sum_{\sigma\in\mathcal{F}_K} |\sigma|(v_\sigma - v_K)\boldsymbol{n}_{K,\sigma} = |K|\overline{\nabla}_K v$,

$$\begin{aligned}\int_\Omega \Pi_{\mathfrak{T}} v(\boldsymbol{x}) \mathrm{div}\boldsymbol{\varphi}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} - \int_{\partial\Omega} \mathbb{T}_{\mathfrak{T}} v(\boldsymbol{x}) \gamma_{\mathbf{n}}(\boldsymbol{\varphi})(\boldsymbol{x}) \mathrm{d}s(\boldsymbol{x}) \\ = \sum_{K\in\mathcal{M}} \sum_{\sigma\in\mathcal{F}_K} |\sigma|(v_K - v_\sigma)\boldsymbol{n}_{K,\sigma} \cdot \boldsymbol{\varphi}_K \\ + \sum_{K\in\mathcal{M}} \sum_{\sigma\in\mathcal{F}_K} |\sigma|(v_K - v_\sigma)(\boldsymbol{\varphi}_\sigma - \boldsymbol{\varphi}_K) \cdot \boldsymbol{n}_{K,\sigma} \\ = -\int_\Omega \overline{\nabla}_{\mathfrak{T}} v(\boldsymbol{x}) \cdot \boldsymbol{\varphi}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \\ + \sum_{K\in\mathcal{M}} \sum_{\sigma\in\mathcal{F}_K} |\sigma|(v_K - v_\sigma)(\boldsymbol{\varphi}_\sigma - \boldsymbol{\varphi}_K) \cdot \boldsymbol{n}_{K,\sigma}. \quad \text{(B.24)}\end{aligned}$$

Let T be the left-hand side of (B.22). Equation (B.24) and Hölder's inequality (C.3) show that, for $p > 1$,

$$\begin{aligned}T &\le \sum_{K\in\mathcal{M}} \sum_{\sigma\in\mathcal{F}_K} |\sigma| d_{K,\sigma} \left|\frac{v_\sigma - v_K}{d_{K,\sigma}}\right| |\boldsymbol{\varphi}_\sigma - \boldsymbol{\varphi}_K| \quad \text{(B.25)} \\ &\le \left(\sum_{K\in\mathcal{M}} \sum_{\sigma\in\mathcal{F}_K} |\sigma| d_{K,\sigma} \left|\frac{v_\sigma - v_K}{d_{K,\sigma}}\right|^p\right)^{\frac{1}{p}} \\ &\qquad \times \left(\sum_{K\in\mathcal{M}} \sum_{\sigma\in\mathcal{F}_K} |\sigma| d_{K,\sigma} |\boldsymbol{\varphi}_\sigma - \boldsymbol{\varphi}_K|^{p'}\right)^{\frac{1}{p'}}.\end{aligned}$$

Apply (B.11) in Lemma B.7 to each component of $\boldsymbol{\varphi}$, with p' instead of p. Since $d_{K,\sigma} \le h_K$, this gives C_{21} depending only on d, p and θ such that

$$\begin{aligned}T &\le C_{21} |v|_{\mathfrak{T},p} \left(\sum_{K\in\mathcal{M}} \sum_{\sigma\in\mathcal{F}_K} h_K^{p'} \int_K |\nabla\boldsymbol{\varphi}(\boldsymbol{x})|^{p'} \mathrm{d}\boldsymbol{x}\right)^{\frac{1}{p'}} \\ &\le C_{21} \theta^{\frac{1}{p'}} |v|_{\mathfrak{T},p} h_{\mathcal{M}} \| |\nabla\boldsymbol{\varphi}| \|_{L^{p'}(\Omega)}.\end{aligned}$$

This completes the proof in the case $p > 1$. If $p = 1$, simply write $|\boldsymbol{\varphi}_K - \boldsymbol{\varphi}_\sigma| \leq \| |\nabla \boldsymbol{\varphi}| \|_{L^\infty(\Omega)} h_{\mathcal{M}}$ in (B.25). ∎

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 generalizing 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 $1^\star = \frac{d}{d-1}$:

$$\forall w \in W^{1,1}(\mathbb{R}^d),\ \|w\|_{L^{1^\star}(\mathbb{R}^d)} \leq \frac{1}{2d} \sum_{i=1}^d \|\partial_i w\|_{L^1(\mathbb{R}^d)}. \tag{B.26}$$

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.26) 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.26) to $w = w_n$ to obtain

$$\|w_n\|_{L^{1^\star}(\mathbb{R}^d)} \leq \frac{1}{2d} \sum_{i=1}^d \|\partial_i w_n\|_{L^1(\mathbb{R}^d)} \leq \frac{1}{2d} \|w\|_{BV(\mathbb{R}^d)}.$$

Take now the inferior limit, using Fatou's lemma in the left-hand side, to see that

$$\forall w \in BV(\mathbb{R}^d),\ \|w\|_{L^{1^\star}(\mathbb{R}^d)} \leq \frac{1}{2d} \|w\|_{BV(\mathbb{R}^d)}. \tag{B.27}$$

Let us now state the discrete Sobolev embedding for $p = 1$.

Lemma B.10 (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*

$$\forall u \in X_{\mathfrak{T},0}\,,\ \|\Pi_{\mathfrak{T}} u\|_{L^{1^\star}(\Omega)} \le \frac{1}{2\sqrt{d}}\,|u|_{\mathfrak{T},1}\,. \tag{B.28}$$

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.23) 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

$$\begin{aligned}\int_{\mathbb{R}^d} \Pi_{\mathfrak{T}} u(\boldsymbol{x}) \mathrm{div}\boldsymbol{\varphi}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} &= \int_\Omega \Pi_{\mathfrak{T}} u(\boldsymbol{x}) \mathrm{div}\boldsymbol{\varphi}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \\ &= \sum_{K\in\mathcal{M}} \sum_{\sigma\in\mathcal{F}_K} |\sigma|(u_K - u_\sigma) \frac{1}{|\sigma|} \int_\sigma \boldsymbol{\varphi}(\boldsymbol{x}) \cdot \boldsymbol{n}_{K,\sigma} \mathrm{d}s(\boldsymbol{x}) \\ &\le \sqrt{d} \sum_{K\in\mathcal{M}} \sum_{\sigma\in\mathcal{F}_K} |\sigma| |u_K - u_\sigma| = \sqrt{d}\,|u|_{\mathfrak{T},1}\,. \end{aligned} \tag{B.29}$$

Hence, $\|\Pi_{\mathfrak{T}} u\|_{BV(\mathbb{R}^d)} \le \sqrt{d}\,|u|_{\mathfrak{T},1}$ and (B.27) leads to (B.28). ∎

Lemma B.11 (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_{22}, depending only on d, p and $\eta \ge \eta_{\mathfrak{T}}$ (see (7.9)), such that*

$$\forall u \in X_{\mathfrak{T},0}\,,\ \|\Pi_{\mathfrak{T}} u\|_{L^{p^\star}(\Omega)} \le C_{22}\,|u|_{\mathfrak{T},p}\,. \tag{B.30}$$

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.28) to $\widehat{u}$ and gathering the sums by edges gives

$$\Big(\int_\Omega |\Pi_{\mathfrak{T}} u(\boldsymbol{x})|^{p^\star} \mathrm{d}\boldsymbol{x}\Big)^{\frac{d-1}{d}} \le \frac{1}{2\sqrt{d}} \sum_{K\in\mathcal{M}} \sum_{\sigma\in\mathcal{F}_K} |\sigma|\,\big||u_K|^\alpha - \widehat{u}_\sigma\big| \tag{B.31}$$

$$\le \frac{1}{2\sqrt{d}} \sum_{\sigma\in\mathcal{F}_{\rm ext},\,\mathcal{M}_\sigma=\{K\}} |\sigma| |u_K|^\alpha + \frac{1}{2\sqrt{d}} \sum_{\sigma\in\mathcal{F}_{\rm int},\,\mathcal{M}_\sigma=\{K,L\}} |\sigma|\,\big||u_K|^\alpha - |u_L|^\alpha\big|.$$

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

$$\big|\,|u_K|^\alpha - |u_L|^\alpha\big| \le \alpha(|u_K|^{\alpha-1} + |u_L|^{\alpha-1})|u_K - u_L|. \tag{B.32}$$

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 (C.3) then yields

$$\begin{aligned}\left(\int_\Omega |\Pi_{\mathfrak{T}} u(\boldsymbol{x})|^{p^\star}\,\mathrm{d}\boldsymbol{x}\right)^{\frac{d-1}{d}} \le \frac{\alpha}{2\sqrt{d}} &\left(\sum_{K\in\mathcal{M}} \sum_{\sigma\in\mathcal{F}_K} |\sigma| d_{K,\sigma}|u_K|^{(\alpha-1)p'}\right)^{\frac{1}{p'}} \\ &\times \left(\sum_{K\in\mathcal{M}} \sum_{\sigma\in\mathcal{F}_K} |\sigma| d_{K,\sigma}\left|\frac{\delta_\sigma u}{d_{K,\sigma}}\right|^p\right)^{\frac{1}{p}}.\end{aligned} \tag{B.33}$$

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.33) and noticing that $\frac{d-1}{d} - \frac{1}{p'} = \frac{1}{p^\star}$, this shows that

$$\|\Pi_{\mathfrak{T}} u\|_{L^{p^\star}(\Omega)} \le \frac{\alpha d^{1/p'}}{2\sqrt{d}} \left(\sum_{K\in\mathcal{M}} \sum_{\sigma\in\mathcal{F}_K} |\sigma| d_{K,\sigma}\left|\frac{\delta_\sigma u}{d_{K,\sigma}}\right|^p\right)^{\frac{1}{p}}. \tag{B.34}$$

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,

$$\sum_{K\in\mathcal{M}} \sum_{\sigma\in\mathcal{F}_K} |\sigma| d_{K,\sigma}\left|\frac{\delta_\sigma u}{d_{K,\sigma}}\right|^p$$

$$\leq 2^{p-1}(1+\eta^{p-1}) \sum_{K\in\mathcal{M}} \sum_{\sigma\in\mathcal{F}_K} |\sigma| \left(d_{K,\sigma} \left| \frac{u_K - u_\sigma}{d_{K,\sigma}} \right|^p + d_{L,\sigma} \left| \frac{u_L - u_\sigma}{d_{L,\sigma}} \right|^p \right)$$

$$\leq 2^{p}(1+\eta^{p-1}) \sum_{K\in\mathcal{M}} \sum_{\sigma\in\mathcal{F}_K} |\sigma| d_{K,\sigma} \left| \frac{u_K - u_\sigma}{d_{K,\sigma}} \right|^p . \tag{B.35}$$

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.30) is deduced from (B.34), (B.35) 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 (C.3) with exponents $\frac{p}{q} > 1$ and $\frac{p}{p-q}$, we have

$$\begin{aligned} |u|_{\mathfrak{T},q} &= \left(\sum_{K\in\mathcal{M}} \sum_{\sigma\in\mathcal{F}_K} |\sigma| d_{K,\sigma} \left| \frac{u_\sigma - u_K}{d_{K,\sigma}} \right|^q \right)^{\frac{1}{q}} \\ &\leq \left(\sum_{K\in\mathcal{M}} \sum_{\sigma\in\mathcal{F}_K} |\sigma| d_{K,\sigma} \left| \frac{u_\sigma - u_K}{d_{K,\sigma}} \right|^p \right)^{\frac{1}{p}} \left(\sum_{K\in\mathcal{M}} \sum_{\sigma\in\mathcal{F}_K} |\sigma| d_{K,\sigma} \right)^{\frac{1}{q}-\frac{1}{p}} \\ &= |u|_{\mathfrak{T},p} \, (d|\Omega|)^{\frac{1}{q}-\frac{1}{p}}. \end{aligned} \tag{B.36}$$

In the last line, we invoked (B.1).

Lemma B.12 (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_{23}, depending only on Ω, p, q and η, such that*

$$\forall u \in X_{\mathfrak{T},0}\,, \ \|\Pi_{\mathfrak{T}} u\|_{L^q(\Omega)} \leq C_{23} \, |u|_{\mathfrak{T},p} \,. \tag{B.37}$$

If $p < d$ we can take $q = p^\star = \frac{pd}{d-p}$ and, if $p \geq d$, we can take any $q < +\infty$.

Proof. If $p = 1$, take $q = 1^\star$ and the result follows from Lemma B.10 (in this case, C_{23} does not depend on η). If $1 < p < d$, take $q = p^\star$ and the result is given by Lemma B.11.

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 only depends on q and d, and Lemma B.11 gives

$$\|\Pi_{\mathfrak{T}} u\|_{L^q(\Omega)} \leq C_{22} \, |u|_{\mathfrak{T},p_1}$$

for some C_{22} depending only on p_1, d and η. Inequality (B.37) follows from this estimate and (B.36) with $q = p_1$. ∎

B.2.2 Compactness in $L^p(\Omega)$

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.13 (Estimates of translations 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,*

$$\forall \boldsymbol{h} \in \mathbb{R}^d\,,\ \|\Pi_{\mathfrak{T}}u(\cdot+\boldsymbol{h}) - \Pi_{\mathfrak{T}}u\|_{L^1(\mathbb{R}^d)} \le |\boldsymbol{h}|\sqrt{d}\,|u|_{\mathfrak{T},1}\,. \tag{B.38}$$

Proof. Since $p = 1$, the proof could be done by following the technique in [47], and would lead to (B.38) without $\sqrt{d}$. We provide here another, more direct, proof based on the BV space, as in Lemma B.10.

Let $w \in C_c^\infty(\mathbb{R}^d)$. For $\boldsymbol{x}, \boldsymbol{h} \in \mathbb{R}^d$, write

$$|w(\boldsymbol{x}+\boldsymbol{h}) - w(\boldsymbol{x})| = \left|\int_0^1 \nabla w(\boldsymbol{x}+t\boldsymbol{h})\cdot\boldsymbol{h}\mathrm{d}t\right| \le |\boldsymbol{h}| \int_0^1 |\nabla w(\boldsymbol{x}+t\boldsymbol{h})|\mathrm{d}t.$$

Integrating with respect to $\boldsymbol{x} \in \mathbb{R}^d$ and using Fubini's Theorem gives the well kown result

$$\|w(\cdot+\boldsymbol{h}) - w\|_{L^1(\mathbb{R}^d)} \le |\boldsymbol{h}| \int_{\mathbb{R}^d} |\nabla w(\boldsymbol{x})|\mathrm{d}\boldsymbol{x} \le |\boldsymbol{h}| \sum_{i=1}^d \|\partial_i w\|_{L^1(\mathbb{R}^d)}\,. \tag{B.39}$$

By density of $C_c^\infty(\mathbb{R}^d)$ in $W^{1,1}(\mathbb{R}^d)$, Inequality (B.39) 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:

$$\forall w \in BV(\mathbb{R}^d)\,,\ \forall \boldsymbol{h} \in \mathbb{R}^d\,,\ \|w(\cdot+\boldsymbol{h}) - w\|_{L^1(\mathbb{R}^d)} \le |\boldsymbol{h}|\,\|w\|_{BV(\mathbb{R}^d)}\,. \tag{B.40}$$

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.10 that $\|\Pi_{\mathfrak{T}}u\|_{BV(\mathbb{R}^d)} \le \sqrt{d}\,|u|_{\mathfrak{T},1}$. The proof is therefore complete by applying (B.40) to $w = \Pi_{\mathfrak{T}}u$. ∎

The following compactness result in L^1 results from Lemmas B.10 and B.13, and the Kolmogorov compactness criterion.

Lemma B.14 (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.10 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 [13, Theorem 4.26] and Lemma B.13 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.15 (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.36) with $q = 1$ shows that $(|u_m|_{\mathfrak{T}_m,1})_{m\in\mathbb{N}}$ is bounded. By Lemma B.14, $(\Pi_{\mathcal{D}_m} u_m)_{m\in\mathbb{N}}$ is thus relatively compact in $L^1(\Omega)$ and, up to a subsequence denoted the same way, converges in this space. By Lemma B.12, $(\Pi_{\mathcal{D}_m} u_m)_{m\in\mathbb{N}}$ is also bounded by some C_{24} in $L^q(\Omega)$ for some $q > p$.
Recall now the interpolation inequality, consequence of Hölder's inequality (C.5) applied to $|f|^p = |f|^{\alpha p}|f|^{(1-\alpha)p}$ with $\alpha = \frac{q-p}{(q-1)p}$ and exponents $(r, r') = (\frac{1}{p\alpha}, \frac{q}{(1-\alpha)p})$:

$$\|f\|_{L^p(\Omega)} \le \|f\|_{L^1(\Omega)}^{\frac{q-p}{(q-1)p}} \|f\|_{L^q(\Omega)}^{\frac{q(p-1)}{(q-1)p}}.$$

Apply this estimate to $f = \Pi_{\mathfrak{T}_m} u_m - \Pi_{\mathfrak{T}_\ell} u_\ell$ and use $\|f\|_{L^q(\Omega)} \le 2C_{24}$. This gives

$$\|\Pi_{\mathfrak{T}_m} u_m - \Pi_{\mathfrak{T}_\ell} u_\ell\|_{L^p(\Omega)} \le \|\Pi_{\mathfrak{T}_m} u_m - \Pi_{\mathfrak{T}_\ell} u_\ell\|_{L^1(\Omega)}^{\frac{q-p}{(q-1)p}} (2C_{24})^{\frac{q(p-1)}{(q-1)p}}. \quad \text{(B.41)}$$

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.41) 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 the discrete version of a classical trace estimate.

Lemma B.16 (Discrete trace inequality). *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 $C_{25} > 0$, depending only on Ω, d, p and ϱ, such that*

$$\forall u \in X_{\mathfrak{T}}\,,\ \|\mathbb{T}_{\mathfrak{T}} u\|_{L^p(\partial\Omega)} \leq C_{25}(|u|_{\mathfrak{T},p} + \|\Pi_{\mathfrak{T}} u\|_{L^p(\Omega)}). \tag{B.42}$$

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 $\mathcal{C}(\tau_i, \boldsymbol{\xi}_i) = \{\boldsymbol{x} + t\boldsymbol{\xi}, t \in (0,1), \boldsymbol{x} \in \tau\}$ 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

$$\begin{aligned}\boldsymbol{x} &= \sum_{j=1}^d \alpha_j \boldsymbol{x}_j + \frac{t}{d+1}(\boldsymbol{x}_{d+1} - \boldsymbol{x}_i) \\ &= \sum_{j=1, j\neq i}^d \alpha_j \boldsymbol{x}_j + \left(\alpha_i - \frac{t}{d+1}\right)\boldsymbol{x}_i + \frac{t}{d+1}\boldsymbol{x}_{d+1}.\end{aligned}$$

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 (only depending 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 for $p = 1$.
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:

- 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. $D(\boldsymbol{x}, \boldsymbol{\xi}_i)$ always exists a cell after having entered it and thus

$$\forall K \in \mathcal{M}\,, \sum_{\sigma \in \mathcal{F}_K} \chi_{K,\sigma}(\boldsymbol{x}) = 0. \tag{B.43}$$

Define

$$\forall \sigma \in \mathcal{F}\,,\ \beta_\sigma(\boldsymbol{x}) = \max\left(1 - \frac{(\boldsymbol{y}_\sigma(\boldsymbol{x}) - \boldsymbol{x}) \cdot \boldsymbol{\xi}_i}{|\boldsymbol{\xi}_i|^2}, 0\right),$$
$$\forall K \in \mathcal{M}\,,\ \beta_K(\boldsymbol{x}) = \max\left(1 - \frac{(\boldsymbol{x}_K - \boldsymbol{x}) \cdot \boldsymbol{\xi}_i}{|\boldsymbol{\xi}_i|^2}, 0\right).$$

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 $\sigma \in \mathcal{F}$,

$$\begin{aligned} &\text{If } \boldsymbol{x} \notin \sigma \text{ then } \sum_{K \in \mathcal{M}_\sigma} \chi_{K,\sigma}(\boldsymbol{x}) \beta_\sigma(\boldsymbol{x}) = 0\,, \\ &\text{If } \boldsymbol{x} \in \sigma \text{ then } \sum_{K \in \mathcal{M}_\sigma} \chi_{K,\sigma}(\boldsymbol{x}) \beta_\sigma(\boldsymbol{x}) = 1 \end{aligned} \tag{B.44}$$

(note that the second situation only happens for a single $\sigma \in \mathcal{F}_{\mathrm{ext}}$ since $\boldsymbol{x} \in \partial\Omega$). Relations (B.43) and (B.44) show that

$$\sum_{K \in \mathcal{M}} \sum_{\sigma \in \mathcal{F}_K} \chi_{K,\sigma}(\boldsymbol{x})(\beta_\sigma(\boldsymbol{x}) u_\sigma - \beta_K(\boldsymbol{x}) u_K)$$

$$= \sum_{\sigma\in\mathcal{F}} u_\sigma \sum_{K\in\mathcal{M}_\sigma} \chi_{K,\sigma}(\boldsymbol{x})\beta_\sigma(\boldsymbol{x}) - \sum_{K\in\mathcal{M}} \beta_K(\boldsymbol{x})u_K \sum_{\sigma\in\mathcal{F}_K} \chi_{K,\sigma}(\boldsymbol{x})$$
$$= u_{\sigma_{\boldsymbol{x}}}$$

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

$$\begin{aligned}
|\mathbb{T}_{\mathfrak{T}}u(\boldsymbol{x})| &= \left|\sum_{K\in\mathcal{M}}\sum_{\sigma\in\mathcal{F}_K} \chi_{K,\sigma}(\boldsymbol{x})(\beta_\sigma(\boldsymbol{x})u_\sigma - \beta_K(\boldsymbol{x})u_K)\right| \\
&= \left|\sum_{K\in\mathcal{M}}\sum_{\sigma\in\mathcal{F}_K} \chi_{K,\sigma}(\boldsymbol{x})\Big[\beta_\sigma(\boldsymbol{x})(u_\sigma - u_K) + (\beta_\sigma(\boldsymbol{x}) - \beta_K(\boldsymbol{x}))u_K\Big]\right| \\
&\le \sum_{K\in\mathcal{M}}\sum_{\sigma\in\mathcal{F}_K} |\chi_{K,\sigma}(\boldsymbol{x})|\Big[\beta_\sigma(\boldsymbol{x})|u_\sigma - u_K| + |\beta_\sigma(\boldsymbol{x}) - \beta_K(\boldsymbol{x})||u_K|\Big].
\end{aligned}$$

Integrating over τ_i gives

$$\begin{aligned}
\|\mathbb{T}_{\mathfrak{T}}u\|_{L^1(\tau_i)} \le{}& \sum_{K\in\mathcal{M}}\sum_{\sigma\in\mathcal{F}_K} |u_\sigma - u_K| \int_{\tau_i} |\chi_{K,\sigma}(\boldsymbol{x})|\beta_\sigma(\boldsymbol{x})\mathrm{d}s(\boldsymbol{x}) \\
&+ \sum_{K\in\mathcal{M}} |u_K| \sum_{\sigma\in\mathcal{F}_K} \int_{\tau_i} |\chi_{K,\sigma}(\boldsymbol{x})|\,|\beta_\sigma(\boldsymbol{x}) - \beta_K(\boldsymbol{x})|\mathrm{d}s(\boldsymbol{x}).
\end{aligned} \tag{B.45}$$

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 $|\boldsymbol{\xi}_i \cdot \boldsymbol{n}_{\tau_i}| \ge \alpha|\boldsymbol{\xi}_i|$, we have $|\sigma|/|\widehat{\boldsymbol{\xi}_i} \cdot \boldsymbol{n}_{\tau_i}| \le |\sigma|/\alpha$. Hence, using $\beta_\sigma(\boldsymbol{x}) \le 1$,

$$\int_{\tau_i} |\chi_{K,\sigma}(\boldsymbol{x})|\beta_\sigma(\boldsymbol{x})\mathrm{d}s(\boldsymbol{x}) \le \frac{|\sigma|}{\alpha}. \tag{B.46}$$

Noticing that $|\beta_\sigma(\boldsymbol{x}) - \beta_K(\boldsymbol{x})| \le \frac{|(\boldsymbol{y}_\sigma(\boldsymbol{x}) - \boldsymbol{x}_K)\cdot\boldsymbol{\xi}_i|}{|\boldsymbol{\xi}_i|^2} \le \frac{h_K}{|\boldsymbol{\xi}_i|} \le \frac{\varrho d_{K,\sigma}}{|\boldsymbol{\xi}_i|}$, we also have

$$\int_{\tau_i} |\chi_{K,\sigma}(\boldsymbol{x})|\,|\beta_\sigma(\boldsymbol{x}) - \beta_K(\boldsymbol{x})|\mathrm{d}s(\boldsymbol{x}) \le \frac{|\sigma|}{\alpha}\frac{\varrho d_{K,\sigma}}{|\boldsymbol{\xi}_i|}. \tag{B.47}$$

Plugging (B.46) and (B.47) into (B.45), and recalling (B.1), provides C_{26} depending only on α, ϱ, $\boldsymbol{\xi}_i$ and d such that

$$\|\mathbb{T}_{\mathfrak{T}}u\|_{L^1(\tau_i)} \le C_{26}(|u|_{\mathfrak{T},1} + \|\Pi_{\mathfrak{T}}u\|_{L^1(\Omega)}).$$

The trace inequality (B.42) for $p = 1$ follows by summing these estimates over $i = 1, \ldots, M$.

Step 3: proof of the trace inequality, $p > 1$.

Let $u \in X_{\mathfrak{T}}$ and, in a similar way as in the proof of Lemma B.11, apply (B.42) with $p = 1$ to $\widehat{u} = ((|u_K|^p)_{K\in\mathcal{M}}, (\widehat{u}_\sigma)_{\sigma\in\mathcal{F}})$ with

$$\widehat{u}_\sigma = \frac{1}{2}(|u_K|^p + |u_L|^p) \quad \text{if } \mathcal{M}_\sigma = \{K, L\},$$
$$\widehat{u}_\sigma = |u_\sigma|^p \quad \text{if } \sigma \in \mathcal{F}_{\rm ext}.$$

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 C_{27} depending only on Ω, d and ϱ such that

$$\|\mathbb{T}_{\mathfrak{T}}u\|^p_{L^p(\partial\Omega)} \le C_{27}(|\widehat{u}|_{\mathfrak{T},1} + \|\Pi_{\mathfrak{T}}u\|^p_{L^p(\Omega)}). \tag{B.48}$$

Suppose that we establish the existence of C_{28}, depending only on Ω, d, p and ϱ, such that

$$|\widehat{u}|_{\mathfrak{T},1} \le C_{28}\,|u|_{\mathfrak{T},p}\,(\|\Pi_{\mathfrak{T}}u\|^{p-1}_{L^p(\Omega)} + \|\mathbb{T}_{\mathfrak{T}}u\|^{p-1}_{L^p(\partial\Omega)}). \tag{B.49}$$

Then, by Young's inequality (C.9),

$$|\widehat{u}|_{\mathfrak{T},1} \le \frac{2C_{28}^p}{p\varepsilon^{p'/p}}\,|u|^p_{\mathfrak{T},p} + \frac{\varepsilon}{p'}\,\|\Pi_{\mathfrak{T}}u\|^p_{L^p(\Omega)} + \frac{\varepsilon}{p'}\,\|\mathbb{T}_{\mathfrak{T}}u\|^p_{L^p(\partial\Omega)}\,. \tag{B.50}$$

Taking $\varepsilon > 0$ such that $\frac{C_{27}\varepsilon}{p'} = \frac{1}{2}$ and plugging the result in (B.48) gives (B.42).
Let us now prove (B.49). If $\mathcal{M}_\sigma = \{K, L\}$, owing to (B.32),

$$|\widehat{u}_K - \widehat{u}_\sigma| = \frac{1}{2}\big||u_K|^p - |u_L|^p\big| \le \frac{p}{2}(|u_K|^{p-1} + |u_L|^{p-1})|u_K - u_L|.$$

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 $\mathcal{M}_\sigma = \{K, L\}$,

$$|\widehat{u}|_{\mathfrak{T},1} \le p \sum_{\sigma\in\mathcal{F}_{\rm ext}} |\sigma|\,|u_\sigma|^{p-1}\delta_\sigma u + p \sum_{K\in\mathcal{M}}\sum_{\sigma\in\mathcal{F}_K} |\sigma|\,|u_K|^{p-1}\delta_\sigma u. \tag{B.51}$$

Let $w_i, F_i, G_i \ge 0$ and $H_i > 0$. Applying the Hölder inequality (C.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

$$\sum_{i\in I} w_i F_i^{p-1} G_i \le \left(\sum_{i\in I} w_i \frac{G_i^p}{H_i^{p-1}}\right)^{\frac{1}{p}} \left(\sum_{i\in I} w_i H_i F_i^p\right)^{\frac{p-1}{p}}.$$

Applied with $w_i = |\sigma|$, $H_i = 1$, $F_i = |u_\sigma|$ and $G_i = \delta_\sigma u$ in the first term of (B.51), 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.51), this gives

$$
\begin{aligned}
|\widehat{u}|_{\mathfrak{T},1} &\le p\left(\sum_{\sigma\in\mathcal{F}_{\rm ext}}|\sigma|(\delta_\sigma u)^p\right)^{\frac1p}\left(\sum_{\sigma\in\mathcal{F}_{\rm ext}}|\sigma|\,|u_\sigma|^p\right)^{\frac{p-1}{p}}\\
&\quad + p\left(\sum_{K\in\mathcal{M}}\sum_{\sigma\in\mathcal{F}_K}|\sigma|\frac{(\delta_\sigma u)^p}{d_{K,\sigma}^{p-1}}\right)^{\frac1p}\left(\sum_{K\in\mathcal{M}}\sum_{\sigma\in\mathcal{F}_K}|\sigma|d_{K,\sigma}|u_K|^p\right)^{\frac{p-1}{p}}\\
&= T_1+T_2. \qquad\text{(B.52)}
\end{aligned}
$$

Write $(\delta_\sigma u)^p = d_{K,\sigma}^p\frac{(\delta_\sigma u)^p}{d_{K,\sigma}^p} \le \mathrm{diam}(\Omega)^{p-1}d_{K,\sigma}\frac{(\delta_\sigma u)^p}{d_{K,\sigma}^p}$ to notice that

$$
T_1 \le p\mathrm{diam}(\Omega)^{\frac{1}{p'}}\,|u|_{\mathfrak{T},p}\,\|\mathbb{T}_{\mathfrak{T}}u\|_{L^p(\partial\Omega)}^{p-1}. \qquad\text{(B.53)}
$$

To estimate T_2, first use the triangle inequality to write, if $\mathcal{M}_\sigma=\{K,L\}$, $\delta_\sigma u\le|u_K-u_\sigma|+|u_L-u_\sigma|$. Then, by definition of $\varrho\ge\eta_{\mathfrak{T}}$ and invoking the power-of-sums inequality (C.12),

$$
\frac{(\delta_\sigma u)^p}{d_{K,\sigma}^{p-1}} \le 2^{p-1}d_{K,\sigma}\left|\frac{u_K-u_\sigma}{d_{K,\sigma}}\right|^p + 2^{p-1}\varrho^{p-1}d_{L,\sigma}\left|\frac{u_L-u_\sigma}{d_{L,\sigma}}\right|^p.
$$

This also holds, dropping the second addend, if $\mathcal{M}_\sigma=\{K\}$. Using this estimate in the first factor in T_2, the term $d_{K,\sigma}\left|\frac{u_K-u_\sigma}{d_{K,\sigma}}\right|^p$ 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,

$$
\begin{aligned}
\sum_{K\in\mathcal{M}}\sum_{\sigma\in\mathcal{F}_K}|\sigma|\frac{(\delta_\sigma u)^p}{d_{K,\sigma}^{p-1}} &\le 2^{p-1}(1+\varrho^{p-1})\sum_{K\in\mathcal{M}}\sum_{\sigma\in\mathcal{F}_K}|\sigma|d_{K,\sigma}\left|\frac{u_K-u_\sigma}{d_{K,\sigma}}\right|^p\\
&= 2^{p-1}(1+\varrho^{p-1})\,|u|_{\mathfrak{T},p}^p.
\end{aligned}
$$

Invoke (B.1) to bound the second factor in T_2 and write

$$
T_2 \le p(2d)^{\frac{1}{p'}}(1+\varrho^{p-1})^{\frac1p}\,|u|_{\mathfrak{T},p}\,\|\Pi_{\mathfrak{T}}u\|_{L^p(\Omega)}^{p-1}. \qquad\text{(B.54)}
$$

Estimates (B.52), (B.53) and (B.54) complete the proof of (B.49). ∎

The following lemma is particularly useful when dealing with Fourier boundary conditions.

Lemma B.17. *Let $p\in[1,+\infty)$, $\mathfrak{T}$ be a polytopal mesh of Ω in the sense of Definition 7.2 and $\varrho\ge\theta_{\mathfrak{T}}+\eta_{\mathfrak{T}}$. Then, there exists $C_{29}>0$, depending only on Ω, d, p and ϱ, such that*

$$
\forall u\in X_{\mathfrak{T}},\ \|\Pi_{\mathfrak{T}}u\|_{L^p(\Omega)} \le C_{29}(|u|_{\mathfrak{T},p}+\|\mathbb{T}_{\mathfrak{T}}u\|_{L^p(\partial\Omega)}).
$$

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.16, 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.16, $\chi_{K,\sigma}(\boldsymbol{x})$ is here defined for all $\boldsymbol{x} \in \Omega$. Since $\chi_{\sigma,K}$ is non-zero (and equal to ± 1) only in the cylinder with base σ and axis $\mathbf{e}$,

$$\int_\Omega |\chi_{K,\sigma}(\boldsymbol{x})| \mathrm{d}\boldsymbol{x} \le |\sigma| \mathrm{diam}(\Omega). \tag{B.55}$$

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 face $\sigma \in \mathcal{F}$ 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.55) to deduce

$$\begin{aligned} \|\Pi_{\mathfrak{T}} u\|_{L^1(\Omega)} &\le \mathrm{diam}(\Omega) \sum_{K \in \mathcal{M}} \sum_{\sigma \in \mathcal{F}_K} |\sigma| \, |u_K - u_\sigma| + \mathrm{diam}(\Omega) \sum_{\sigma \in \mathcal{F}_{\mathrm{ext}}} |\sigma| \, |u_\sigma| \\ &= \mathrm{diam}(\Omega)(|u|_{\mathfrak{T},1} + \|\mathbb{T}_{\mathfrak{T}} u\|_{L^1(\partial\Omega)}). \end{aligned} \tag{B.56}$$

This concludes the proof in the case $p = 1$. The general case $p > 1$ follows by applying (B.56) to $\widehat{u}$ defined as in Step 3 of the proof of Lemma B.16, and by using (B.50) with $\varepsilon = p'/2$. ∎

B.3.2 Discrete Sobolev embeddings

Lemma B.18 (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_{30} depending only on Ω and d such that*

$$\forall u \in X_{\mathfrak{T}}\,, \ \left\| \Pi_{\mathfrak{T}} u - \overline{\Pi_{\mathfrak{T}} u} \right\|_{L^{1^\star}(\Omega)} \le C_{30} \, |u|_{\mathfrak{T},1}\,, \tag{B.57}$$

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 $\|w - \overline{w}\|_{L^{1^\star}(\Omega)} \le C_{31} \|\nabla w\|_{L^1(\Omega)^d}$ for all $w \in W^{1,1}(\Omega)$, where C_{31} 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:

$$\left\| \Pi_{\mathfrak{T}} u - \overline{\Pi_{\mathfrak{T}} u} \right\|_{L^{1^\star}(\Omega)} \le C_{31} \, |\Pi_{\mathfrak{T}} u|_{BV(\Omega)}\,, \tag{B.58}$$

where

$$|w|_{BV(\Omega)} = \sup \left\{ \int_\Omega w(\boldsymbol{x}) \mathrm{div}\boldsymbol{\varphi}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \,:\, \boldsymbol{\varphi} \in C_c^\infty(\Omega, \mathbb{R}^d), \|\boldsymbol{\varphi}\|_{L^\infty(\Omega)^d} \le 1 \right\}.$$

Write (B.23) with $v = u$. The integral term on $\partial\Omega$ can be dropped since $\boldsymbol{\varphi}$ vanishes on the boundary. Reason then as in (B.29) in Lemma B.10 to obtain $|\Pi_{\mathfrak{T}} u|_{BV(\Omega)} \le \sqrt{d}\,|u|_{\mathfrak{T},1}$, and the conclusion follows from (B.58). ∎

Lemma B.19 (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 $\varrho \ge \theta_{\mathfrak{T}} + \eta_{\mathfrak{T}}$. Then, there exists C_{32}, depending only on Ω, d, p and ϱ, such that*

$$\forall u \in X_{\mathfrak{T}}\,,\ \left\|\Pi_{\mathfrak{T}} u - \overline{\Pi_{\mathfrak{T}} u}\right\|_{L^{p^\star}(\Omega)} \le C_{32}\,|u|_{\mathfrak{T},p}\,,$$

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 $A \le MB$ 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

$$\begin{aligned}
&\widehat{u}_\sigma = \frac{1}{2}(|u_K|^\alpha + |u_L|^\alpha) \quad \text{if } \mathcal{M}_\sigma = \{K,L\},\\
&\widehat{u}_\sigma = |u_K|^\alpha \quad \text{if } \mathcal{M}_\sigma = \{K\}.
\end{aligned}$$

Since $|\overline{\Pi_{\mathfrak{T}}\widehat{u}}| \le \frac{1}{|\Omega|}\,\|\Pi_{\mathfrak{T}} u\|^\alpha_{L^\alpha(\Omega)}$, Inequality (B.57) applied to $\widehat{u}$ yields

$$\begin{aligned}
\|\Pi_{\mathfrak{T}} u\|^\alpha_{L^{\alpha 1^\star}(\Omega)} &= \|\Pi_{\mathfrak{T}}\widehat{u}\|_{L^{1^\star}(\Omega)}\\
&\le \left\|\Pi_{\mathfrak{T}}\widehat{u} - \overline{\Pi_{\mathfrak{T}}\widehat{u}}\right\|_{L^{1^\star}(\Omega)} + |\Omega|^{\frac{1}{1^\star}}\,|\overline{\Pi_{\mathfrak{T}}\widehat{u}}|\\
&\lesssim |\widehat{u}|_{\mathfrak{T},1} + \|\Pi_{\mathfrak{T}} u\|^\alpha_{L^\alpha(\Omega)}\,. \qquad \text{(B.59)}
\end{aligned}$$

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.11 (passage from (B.31) to (B.33)) 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.59) and taking the power $1/\alpha$ (thanks to the power-of-sums inequality (C.13)) yields

$$\|\Pi_{\mathfrak{T}} u\|_{L^{\alpha 1^\star}(\Omega)} \lesssim |u|^{\frac{1}{\alpha}}_{\mathfrak{T},r}\left\||\Pi_{\mathfrak{T}} u|^{\alpha-1}\right\|^{\frac{1}{\alpha}}_{L^{r'}(\Omega)} + \|\Pi_{\mathfrak{T}} u\|_{L^\alpha(\Omega)}\,. \qquad \text{(B.60)}$$

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 (C.9) with exponent α, and ε small enough (depending only on the constants hidden in $\lesssim$), to deduce from (B.60) that

$$\|\Pi_{\mathfrak{T}} u\|_{L^{\alpha 1^\star}(\Omega)} \lesssim |u|_{\mathfrak{T},r} + \|\Pi_{\mathfrak{T}} u\|_{L^\alpha(\Omega)}.$$

If $r \le p$, that is if $r' = \frac{\alpha 1^\star}{\alpha - 1} \ge p'$, then (B.36) shows that

$$\|\Pi_{\mathfrak{T}} u\|_{L^{\alpha 1^\star}(\Omega)} \lesssim |u|_{\mathfrak{T},p} + \|\Pi_{\mathfrak{T}} u\|_{L^\alpha(\Omega)}. \tag{B.61}$$

The estimate (B.57) 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.61) applied with $\alpha = 1^\star$, $(1^\star)^2$, ... then establishes that, for any $k \in \mathbb{N}$ such that $\frac{(1^\star)^{k+1}}{(1^\star)^k - 1} \ge p'$,

$$\|\Pi_{\mathfrak{T}} u\|_{L^{(1^\star)^{k+1}}(\Omega)} \lesssim |u|_{\mathfrak{T},p}. \tag{B.62}$$

Select k as the largest integer such that $\frac{(1^\star)^{k+1}}{(1^\star)^k - 1} \ge p'$. 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

$$\frac{\alpha 1^\star}{\alpha - 1} = \frac{1^\star p^\star}{p^\star - 1^\star} \ge p' \tag{B.63}$$

and

$$\alpha = \frac{p^\star}{1^\star} \le (1^\star)^{k+1}. \tag{B.64}$$

Inequality (B.63) allows us to apply (B.61), which gives

$$\|\Pi_{\mathfrak{T}} u\|_{L^{p^\star}(\Omega)} \lesssim |u|_{\mathfrak{T},p} + \|\Pi_{\mathfrak{T}} u\|_{L^\alpha(\Omega)}.$$

By (B.64), $\|\Pi_{\mathfrak{T}} u\|_{L^\alpha(\Omega)} \lesssim \|\Pi_{\mathfrak{T}} u\|_{L^{(1^\star)^{k+1}}(\Omega)}$ and (B.62) then concludes the proof.

It remains to check (B.63) and (B.64). We have $p = \frac{dp^\star}{d + p^\star}$ so (B.63) boils down to $\frac{1^\star p^\star}{p^\star - 1^\star} \ge \frac{dp^\star}{dp^\star - d - p^\star}$, that is to say $1^\star(dp^\star - d - p^\star) \ge d(p^\star - 1^\star)$, or $1^\star(d-1)p^\star \ge dp^\star$. This last relation is obvious since $1^\star(d-1) = d$ (we thus even have equality in (B.63)). To check (B.64), we start by writing that, by definition of k, $\frac{(1^\star)^{k+2}}{(1^\star)^{k+1} - 1} \le p'$, which can be recast as $1 - \frac{1}{p} = \frac{1}{p'} \le \frac{1}{1^\star} - \frac{1}{(1^\star)^{k+2}}$. But $\frac{1}{p^\star} = \frac{1}{p} - \frac{1}{d}$ and $\frac{1}{1^\star} = 1 - \frac{1}{d}$, so

$$\frac{1}{p^\star} \ge 1 - \frac{1}{1^\star} + \frac{1}{(1^\star)^{k+2}} - \frac{1}{d} = \frac{1}{(1^\star)^{k+2}},$$

which is equivalent to (B.64). ∎

The proof of the following lemma is similar to the proof of Lemma B.12, using Lemmas B.18 and B.19.

Lemma B.20 (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_{33}, depending only on Ω, d, p and ϱ, such that*

$$\forall u \in X_{\mathfrak{T}}\,,\ \left\|\Pi_{\mathfrak{T}} u - \overline{\Pi_{\mathfrak{T}} u}\right\|_{L^q(\Omega)} \leq C_{33}\,|u|_{\mathfrak{T},p}\,,$$

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 in $L^p(\Omega)$

Lemma B.21. *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_{34}, depending only on Ω and ϱ, such that*

$$\forall u \in X_{\mathfrak{T}}\,,\ \forall \boldsymbol{h} \in \mathbb{R}^d\,,\ \|\Pi_{\mathfrak{T}} u(\cdot + \boldsymbol{h}) - \Pi_{\mathfrak{T}} u\|_{L^1(\mathbb{R}^d)} \leq |\boldsymbol{h}| C_{34}(|u|_{\mathfrak{T},1} + |\overline{\Pi_{\mathfrak{T}} u}|),$$

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.23) 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 $|\boldsymbol{\varphi}| \leq \sqrt{d}$),

$$\begin{aligned}\int_{\mathbb{R}^d} \Pi_{\mathfrak{T}} u(\boldsymbol{x})\mathrm{div}\boldsymbol{\varphi}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} &\leq \sqrt{d}\sum_{K\in\mathcal{M}}\sum_{\sigma\in\mathcal{F}_K} |\sigma||u_K - u_\sigma| + \sqrt{d}\int_{\partial\Omega} |\mathbb{T}_{\mathfrak{T}} u(\boldsymbol{x})|\mathrm{d}s(\boldsymbol{x}) \\ &\leq \sqrt{d}\,|u|_{\mathfrak{T},1} + \sqrt{d}\,\|\mathbb{T}_{\mathfrak{T}} u\|_{L^1(\partial\Omega)}\,.\end{aligned}$$

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.16 and B.18 then provide C_{35} depending only on Ω, d and ϱ such that

$$\|\Pi_{\mathfrak{T}} u\|_{BV(\mathbb{R}^d)} \leq C_{35}(|u|_{\mathfrak{T},1} + |\overline{\Pi_{\mathfrak{T}} u}|).$$

The inequality (B.40) concludes the proof. ∎

As for Dirichlet boundary conditions, the following compactness result is an immediate consequence of Lemmas B.20 and B.21.

Lemma B.22 (Discrete Rellich theorem 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 $L^p(\Omega)$.*

B.4 Discrete functional analysis for mixed boundary condition

We consider here that Assumption (7.2) on Ω and Assumption (2.52) 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

$$\begin{aligned} X_{\mathfrak{T},\Gamma_d} &= \{v \in X_{\mathfrak{T},\partial} \,:\, v_\sigma = 0 \text{ for all } \sigma \in \mathcal{F}_{\rm ext} \text{ such that } \\ &\qquad \sigma \cap \Gamma_d = \emptyset.\}, \\ X_{\mathfrak{T},\Omega,\Gamma_n} &= \{v \in X_{\mathfrak{T}} \,:\, v_\sigma = 0 \text{ for all } \sigma \in \mathcal{F}_{\rm ext} \text{ such that } \\ &\qquad \sigma \cap \Gamma_d \neq \emptyset\}. \end{aligned} \tag{B.65}$$

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.23. *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_{36} depending only on $\widetilde{\Omega}$ and A such that, for all $w \in BV(\widetilde{\Omega})$ satisfying $\int_A w(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = 0$,*

$$\|w\|_{L^{1^\star}(\widetilde{\Omega})} \le C_{36}\, |w|_{BV(\widetilde{\Omega})}\,, \tag{B.66}$$

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_{37} depending only on $\widetilde{\Omega}$ such that

$$\forall w \in BV(\widetilde{\Omega})\,,\ \|w\|_{L^{1^\star}(\widetilde{\Omega})} \le C_{37}(|w|_{BV(\widetilde{\Omega})} + \|w\|_{L^1(\widetilde{\Omega})}).$$

Estimate (B.66) is proved if we establish the following Poincaré's inequality: there exists C_{38} depending only on $\widetilde{\Omega}$ and A such that, for any $w \in BV(\widetilde{\Omega})$ satisfying $\int_A w(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = 0$,

$$\|w\|_{L^1(\widetilde{\Omega})} \le C_{38}\, |w|_{BV(\widetilde{\Omega})}\,. \tag{B.67}$$

The proof of (B.67) 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.67) holds and so does (B.66). ∎

Under Assumptions (7.2) and (2.52), 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.3. $\widetilde{\Omega}$ and A depend only on Ω and Γ_d.

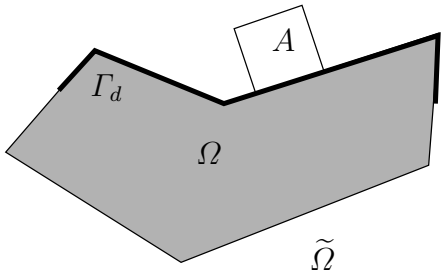


Fig. B.3. Extension of Ω.

Lemma B.24. *Under Assumptions* (7.2) *and* (2.52)*, 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*

$$\forall u \in X_{\mathfrak{T},\Omega,\Gamma_n}\,,\ \left|\widetilde{\Pi}_{\mathcal{D}} u\right|_{BV(\widetilde{\Omega})} \leq \sqrt{d}\,|u|_{\mathfrak{T},1}\,. \tag{B.68}$$

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.23) written for $v = u$ vanishes, and the same computations as in (B.29) lead to (B.68). ∎

The following Sobolev embeddings are a straitghforward consequence of Lemma B.23 and B.24.

Lemma B.25 (Discrete embedding of $W^{1,1}(\Omega)$ in $L^{1^\star}(\Omega)$, mixed BCs). *Under Assumptions* (7.2) *and* (2.52)*, let $\mathfrak{T}$ be a polytopal mesh of Ω in the sense of Definition 7.2. Then, there exists C_{39} depending only on Ω and Γ_d such that*

$$\forall u \in X_{\mathfrak{T},\Omega,\Gamma_n}\,,\ \|\Pi_{\mathfrak{T}} u\|_{L^{1^\star}(\Omega)} \le C_{39}\,|u|_{\mathfrak{T},1}\,.$$

The following results can be then proved from Lemma B.25 by using the same trick as in the proof of Lemma B.11 and Lemma B.12.

Lemma B.26 (Discrete embedding of $W^{1,p}(\Omega)$ in $L^{p^\star}(\Omega)$, mixed BCs, $p \in (1,d)$). *Under Assumptions* (7.2) *and* (2.52)*, let $\mathfrak{T}$ be a polytopal mesh of Ω in the sense of Definition 7.2, $p \in (1,d)$ and $\eta \ge \eta_{\mathfrak{T}}$. Then, there exists C_{40} depending only on Ω, Γ_d and η such that*

$$\forall u \in X_{\mathfrak{T},\Omega,\Gamma_n}\,,\ \|\Pi_{\mathfrak{T}} u\|_{L^{p^\star}(\Omega)} \le C_{40}\,|u|_{\mathfrak{T},p}\,,$$

where $p^\star = \frac{dp}{d-p}$.

Lemma B.27 (Discrete embedding of $W^{1,p}(\Omega)$ in $L^q(\Omega)$ for some $q > p$, mixed BCs). *Under Assumptions* (7.2) *and* (2.52)*, let $\mathfrak{T}$ be a polytopal mesh of Ω in the sense of Definition 7.2, $p \in [1,+\infty)$ and $\eta \ge \eta_{\mathfrak{T}}$. Then, there exists $q > p$, depending only on p and d, and C_{41}, depending only on Ω, d, p, Γ_d and η, such that*

$$\forall u \in X_{\mathfrak{T},\Omega,\Gamma_n}\,,\ \|\Pi_{\mathfrak{T}} u\|_{L^{q}(\Omega)} \le C_{41}\,|u|_{\mathfrak{T},p}\,.$$

If $p < d$ we can take $q = p^\star = \frac{pd}{d-p}$. If $p \ge d$, we can take any $q < +\infty$.

B.4.2 Compactness in $L^p(\Omega)$

Lemma B.28 (Discrete Rellich theorem, mixed BCs). *Under Assumptions* (7.2) *and* (2.52)*, let $p \in [1,+\infty)$ and $(\mathfrak{T}_m)_{m\in\mathbb{N}}$ be a sequence of polytopal meshes of Ω. Assume that $\sup_{m\in\mathbb{N}}(\theta_{\mathfrak{T}_m} + \eta_{\mathfrak{T}_m}) < +\infty$. Then, for any $u_m \in X_{\mathfrak{T}_m,\Omega,\Gamma_n}$ 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 $L^p(\Omega)$.*

Proof. By Lemma B.27, 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.22 gives the relative compactness of $(\Pi_{\mathfrak{T}_m} u_m)_{m\in\mathbb{N}}$ in $L^p(\Omega)$. ∎

C

Technical results

C.1 Standard notations, inequalities and relations

We gather here a few notations and standard inequalities that are used throughout the book, sometimes implicitly.

C.1.1 Notations

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 $\boldsymbol{\xi}$.
The Lebesgue measure of a subset A of $\mathbb{R}^d$ is written $|A|$.

C.1.2 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

$$\sum_{i\in I} |a_i b_i| \leq \left(\sum_{i\in I} |a_i|^p \right)^{\frac{1}{p}} \left(\sum_{i\in I} |b_i|^{p'} \right)^{\frac{1}{p'}}. \tag{C.1}$$

It is frequently used after the introduction of some non-zero real numbers $(d_i)_{i\in I}$ in the product $a_i b_i$. More precisely, writing $a_i b_i = (a_i d_i)(\frac{b_i}{d_i})$ and applying (C.1) to this new product, we have

$$\sum_{i\in I} |a_i b_i| \leq \left(\sum_{i\in I} |a_i|^p |d_i|^p \right)^{\frac{1}{p}} \left(\sum_{i\in I} \frac{|b_i|^{p'}}{|d_i|^{p'}} \right)^{\frac{1}{p'}}. \tag{C.2}$$

Another frequent use is to evenly split an existing weight. If $(w_i)_{i\in I}$ are non-negative numbers, writing $w_i |a_i b_i| = (w_i^{1/p} |a_i|)(w_i^{1/p'} |b_i|)$ and using (C.1) leads to

$$\sum_{i\in I} w_i|a_i b_i| \leq \left(\sum_{i\in I} w_i|a_i|^p\right)^{\frac{1}{p}} \left(\sum_{i\in I} w_i|b_i|^{p'}\right)^{\frac{1}{p'}}. \tag{C.3}$$

Using both weights and the introduction of non-zero numbers, we also have

$$\sum_{i\in I} w_i|a_i b_i| \leq \left(\sum_{i\in I} w_i|a_i|^p|d_i|^p\right)^{\frac{1}{p}} \left(\sum_{i\in I} w_i \frac{|b_i|^{p'}}{|d_i|^{p'}}\right)^{\frac{1}{p'}}. \tag{C.4}$$

The Hölder inequalities are also valid in Lebesgue spaces over a measurable set (X, μ). For example, the equivalent of (C.1) for integrals is: if $f, g : X \to \mathbb{R}$ are measurable functions, then

$$\int_X |fg|\mathrm{d}\mu \leq \left(\int_X |f|^p \mathrm{d}\mu\right)^{\frac{1}{p}} \left(\int_X |g|^{p'} \mathrm{d}\mu\right)^{\frac{1}{p'}}. \tag{C.5}$$

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

$$\int_X |f|\mathrm{d}\mu \leq \left(\int_X |f|^p \mathrm{d}\mu\right)^{\frac{1}{p}} \mu(X)^{\frac{1}{p'}} = \left(\int_X |f|^p \mathrm{d}\mu\right)^{\frac{1}{p}} \mu(X)^{1-\frac{1}{p}}. \tag{C.6}$$

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

$$\|f\|_{L^r(X)} \leq \mu(X)^{\frac{1}{r}-\frac{1}{q}} \|f\|_{L^q(X)}. \tag{C.7}$$

C.1.3 Young inequality

For $a, b \geq 0$ and (p, p') conjugate exponents, the Young inequality reads

$$ab \leq \frac{1}{p}a^p + \frac{1}{p'}b^{p'}. \tag{C.8}$$

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

$$ab \leq \frac{\varepsilon}{p}a^p + \frac{1}{p'\varepsilon^{p'/p}}b^{p'}. \tag{C.9}$$

C.1.4 Jensen inequality

Let A be a measurable set 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

$$\Psi\left(\frac{1}{|A|}\int_A f(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\right) \leq \frac{1}{|A|}\int_A \Psi(f(\boldsymbol{x}))\mathrm{d}\boldsymbol{x}. \tag{C.10}$$

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

$$\Psi\left(\frac{1}{W}\sum_{i\in I} w_i a_i\right) \leq \frac{1}{W}\sum_{i\in I} w_i \Psi(a_i). \tag{C.11}$$

C.1.5 Power of sums

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

$$\forall \alpha \geq 1\,,\ (a+b)^\alpha \leq 2^{\alpha-1}a^\alpha + 2^{\alpha-1}b^\alpha. \tag{C.12}$$

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

$$\forall \alpha \leq 1\,,\ (a+b)^\alpha \leq a^\alpha + b^\alpha. \tag{C.13}$$

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,

$$\left(\sum_{i=1}^{\ell} a_i\right)^\alpha \leq \ell^{\alpha-1}\sum_{i=1}^{\ell} a_i^\alpha. \tag{C.14}$$

C.1.6 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

$$\begin{aligned}\sum_{n=0}^{N-1}(a_{n+1}-a_n)b_n &= \sum_{n=0}^{N-1} a_{n+1}b_n - \sum_{n=0}^{N-1} a_n b_n \\ &= \sum_{n=0}^{N-1} a_{n+1}b_n - \left(a_0 b_0 + \sum_{n=0}^{N-1} a_{n+1}b_{n+1} - a_N b_N\right)\end{aligned}$$

$$= \sum_{n=0}^{N-1} a_{n+1}(b_n - b_{n+1}) + a_N b_N - a_0 b_0.$$

To summarise,

$$\sum_{n=0}^{N-1} (a_{n+1} - a_n) b_n = -\sum_{n=0}^{N-1} a_{n+1}(b_{n+1} - b_n) + a_N b_N - a_0 b_0. \tag{C.15}$$

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 (C.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 (C.15) to $(\widetilde{b}_n)_{n=0,\dots,N}$ instead of $(b_n)_{n=0,\dots,N}$ gives

$$\begin{aligned}
\sum_{n=0}^{N-1} (a_{n+1} - a_n) b_{n+1} &= \\
&= -\sum_{n=0}^{N-1} a_{n+1}(b_{n+2} - b_{n+1}) - a_0 b_1 \\
&= -\sum_{n=1}^{N} a_n (b_{n+1} - b_n) - a_0 b_1 \\
&= -\sum_{n=0}^{N-1} a_n (b_{n+1} - b_n) + a_0(b_1 - b_0) - a_N(b_{N+1} - b_N) - a_0 b_1.
\end{aligned}$$

In other words,

$$\sum_{n=0}^{N-1} (a_{n+1} - a_n) b_{n+1} = -\sum_{n=0}^{N-1} a_n (b_{n+1} - b_n) + a_N b_N - a_0 b_0. \tag{C.16}$$

This is the equivalent of (C.15) with an offset of the second family $(b_n)_{n=0,\dots,N}$.

By creating a convex combination of (C.15) and (C.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 $\nu \times$ (C.16) and $(1-\nu) \times$ (C.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,

$$\sum_{n=0}^{N-1} (a_{n+1} - a_n) b_{n+\nu} = -\sum_{n=0}^{N-1} a_{n+(1-\nu)} (b_{n+1} - b_n) + a_N b_N - a_0 b_0. \tag{C.17}$$

C.2 Topological degree

The following theorem is a consequence of the theory of the topological degree [27].

Theorem C.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 C.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 centered 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 C.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 $v = F(v)$. ∎

C.3 Weak and strong convergences in integrals

The following lemma is used throughout the book.

Lemma C.3 (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,

$$\begin{aligned}\left|\int_X f_n\cdot g_n\mathrm{d}\mu- \int_X f\cdot g\mathrm{d}\mu\right| &\\ &= \left|\int_X (f_n-f)\cdot g_n\mathrm{d}\mu + \int_X f\cdot(g_n-g)\mathrm{d}\mu\right| \\ &\le \|f_n-f\|_{L^p(X)^d}\|g_n\|_{L^{p'}(X)^d} + \left|\int_X f\cdot(g_n-g)\mathrm{d}\mu\right| \\ &\le C\|f_n-f\|_{L^p(X)^d} + \left|\int_X f\cdot(g_n-g)\mathrm{d}\mu\right|.\end{aligned}$$

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 C.4 (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. ∎

C.4 Minty trick and convexity inequality

The next lemma, whose proof is based on a technique called in the literature as the Minty trick, is used to identify limits of non-linear functions of weakly convergent sequences.

Lemma C.5 (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:

$$\liminf_{n\to\infty} \int_X \beta(w_n)\zeta(w_n)\mathrm{d}\mu \leq \int_X \overline{\beta}\,\overline{\zeta}\mathrm{d}\mu. \tag{C.18}$$

Then,

$$\overline{\beta} = \beta(w) \text{ and } \overline{\zeta} = \zeta(w) \text{ a.e. in } X, \tag{C.19}$$

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,

$$\int_X \left[\beta(w_m) - \beta(w)\right]\left[\zeta(w_m) - \zeta(w)\right]\mathrm{d}\mu \geq 0.$$

Develop this relation and use (C.18) and the weak convergences of $\beta(w_m)$ and $\zeta(w_m)$ to take the inferior limit as $m \to \infty$. This gives

$$\int_X \left[\overline{\beta} - \beta(w)\right]\left[\overline{\zeta} - \zeta(w)\right]\mathrm{d}\mu \geq 0. \tag{C.20}$$

With w defined as in the lemma,

$$\frac{\overline{\beta}+\overline{\zeta}}{2} = \frac{\beta(w)+\zeta(w)}{2}. \tag{C.21}$$

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 (C.20), 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 (C.19) follows by adding and subtracting this relation to/from (C.21). ∎

The proof of this lemma is classical, and only given for the convenience of the reader.

Lemma C.6 (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 [42]. The result of the lemma is just the translation of this weak lower semi-continuity of Φ. ∎

References

1. 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.
2. 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.
3. 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.
4. 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*, page 30p, 2015. To appear.
5. 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.
6. 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.
7. 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.
8. 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.
9. M. Bertsch, R. Kersner, and L. A. Peletier. Positivity versus localization in degenerate diffusion equations. *Nonlinear Anal.*, 9(9):987–1008, 1985.
10. J. Bonelle. *Compatible Discrete Operator schemes on polyhedral meshes for elliptic and Stokes equations*. PhD thesis, University of Paris-Est, 2014.
11. 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.

12. 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.
13. H. Brezis. *Functional analysis, Sobolev spaces and partial differential equations*. Universitext. Springer, New York, 2011.
14. F. Brezzi, A. Buffa, and K. Lipnikov. Mimetic finite differences for elliptic problems. *M2AN Math. Model. Numer. Anal.*, 43(2):277–295, 2009.
15. 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.
16. 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.
17. J. Carrillo. Entropy solutions for nonlinear degenerate problems. *Arch. Ration. Mech. Anal.*, 147(4):269–361, 1999.
18. 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.
19. 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.
20. Y. Chen, B. Vemuri, and L. Wang. Image denoising and segmentation via nonlinear diffusion. *Computers and Mathematics with Applications*, 39:131–149, 2000.
21. P. Ciarlet. The finite element method for elliptic problems. North-Holland, 1978.
22. 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.
23. P. Clément. Approximation by finite element functions using local regularization. *RAIRO Anal. Numér.*, 9:77–84, 1975.
24. 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.
25. 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.
26. 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.
27. K. Deimling. *Nonlinear functional analysis*. Springer-Verlag, Berlin, 1985.
28. 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.
29. J. Droniou. Intégration et espaces de sobolev à valeurs vectorielles. Polycopiés de l'Ecole Doctorale de Maths-Info de Marseille. `http://www-gm3.univ-mrs.fr/polys/gm3-02/`, 2001.
30. J. Droniou. Finite volume schemes for fully non-linear elliptic equations in divergence form. *ESAIM: Mathematical Modelling and Numerical Analysis*, 40(6):1069, 2006.

31. J. Droniou and R. Eymard. A mixed finite volume scheme for anisotropic diffusion problems on any grid. *Numer. Math.*, 105(1):35–71, 2006.
32. 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.
33. 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.
34. J. Droniou, R. Eymard, and P. Feron. Gradient Schemes for Stokes problem. *IMA J. Numer. Anal.*, 36(4):1636–1669, 2016.
35. 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.
36. 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.
37. 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.
38. J. Droniou, R. Eymard, and K. S. Talbot. Uniform temporal stability of solutions to doubly nonlinear degenerate parabolic equations. *J. Differential Equations*, 260:7821–7860, 2016.
39. J. Droniou and B. P. Lamichhane. Gradient schemes for linear and non-linear elasticity equations. *Numer. Math.*, 129(2):251–277, 2015.
40. J. Droniou and N. Nataraj. Improved l^2 estimate for gradient schemes, and super-convergence of the TPFA finite volume scheme. 2016. Submitted.
41. 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.
42. 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.
43. A. Ern and J.-L. Guermond. *Theory and practice of finite elements*, volume 159. Springer, 2004.
44. A. Ern and J.-L. Guermond. *Finite elements methods (bis, titre à vérifier)*. Springer, 2017.
45. L. C. Evans. *Partial differential equations*, volume 19 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, second edition, 2010.
46. L. C. Evans and J. Spruck. Motion of level sets by mean curvature I. *J. Differential Geometry*, 33:635–681, 1991.
47. 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.
48. 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.

49. 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.
50. R. Eymard, C. Guichard, and R. Herbin. Small-stencil 3d schemes for diffusive flows in porous media. *M2AN*, 46:265–290, 2012.
51. 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.
52. 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.
53. 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.
54. 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.
55. P. Féron, R. Eymard, and C. Guichard. Gradient schemes for navier–stokes equations. submitted.
56. T. Gallouët and R. Herbin. Partial differential equations. Master course, 2015.
57. 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*, to appear.
58. 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.
59. M. Guedda, D. Hilhorst, and M. A. Peletier. Disappearing interfaces in nonlinear diffusion. *Adv. Math. Sci. Appl.*, 7(2):695–710, 1997.
60. J. Hadamard. *Lectures on Cauchy's Problem in Linear Partial Differential Equations*. Dover Phoenix, New York, NY, USA, 1923.
61. 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.
62. 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.
63. K. Lipnikov, G. Manzini, and M. Shaskov. Mimetic finite difference method. *J. Comput. Phys.* to appear.
64. 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.
65. J. E. Roberts and J.-M. Thomas. Mixed and hybrid methods. In *Handbook of numerical analysis, Vol. II*, Handb. Numer. Anal., II, pages 523–639. North-Holland, Amsterdam, 1991.
66. F.-J. Sayas. From Raviart-Thomas to HDG, `http://arxiv.org/pdf/1307.2491v1.pdf`. 2013.

67. G. Strang. Variational crimes in the finite element method. *The mathematical foundations of the finite element method with applications to partial differential equations*, pages 689–710, 1972.
68. 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.
69. J. Weickert. Coherence-enhancing diffusion filtering. *International Journal of Computer Vision*, 31(2/3):111–127, 1999.