



**HAL**  
open science

# Error estimates for finite differences approximations of the total variation

Corentin Caillaud, Antonin Chambolle

► **To cite this version:**

Corentin Caillaud, Antonin Chambolle. Error estimates for finite differences approximations of the total variation. *IMA Journal of Numerical Analysis*, 2023, 43 (2), pp.692–736. 10.1093/imanum/drac001 . hal-02539136v2

**HAL Id: hal-02539136**

**<https://hal.science/hal-02539136v2>**

Submitted on 7 Oct 2021

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

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

# Error estimates for finite differences approximations of the total variation

Corentin Caillaud\*      Antonin Chambolle†

September 15, 2021

We present a convergence rate analysis of the Rudin-Osher-Fatemi (ROF) denoising problem for two different discretizations of the total variation. The first is the **standard discretization, which induces blurring in some particular diagonal directions**. We prove that in a **simplified** setting corresponding to such a direction, the discrete ROF energy converges to the continuous one **with the rate  $h^{2/3}$** . The second **discretization** is based on dual Raviart-Thomas fields and achieves **an optimal  $O(h)$  convergence rate** for the same quantity, **for discontinuous solutions** with some standard hypotheses.

**Keywords:** total variation, finite differences, finite elements, error bounds, convergence rates, image denoising

**MSC (2020):** 35A21 35A35 65D18 65N06 65N15 65N30

## 1 Introduction

Since its introduction by Rudin, Osher and Fatemi in [31] the use of total variation as a regularizer for denoising and inverse problems has proven to be effective in removing noise while preserving sharp edges. In the continuous setting, the denoising model consists in solving the so-called “ROF” problem:

$$\bar{u} = \arg \min_{u \in BV(\Omega) \cap L^2(\Omega)} \frac{1}{2\lambda} \|u - g\|_{L^2}^2 + \text{TV}(u) =: E(u) \quad (1)$$

where  $\Omega = [0, 1] \times [0, 1]$  is the domain of a noisy image  $g \in L^\infty(\Omega)$  and  $\lambda > 0$  is a regularizing parameter. Here TV stands for the continuous total variation given by  $\text{TV}(u) = \int_\Omega |\nabla u|$  when  $u$  is regular, and with  $|\cdot|$  denoting the euclidean norm in  $\mathbb{R}^2$ .

---

\*CMAP, Ecole Polytechnique, CNRS, Institut Polytechnique de Paris, Palaiseau, France.

†CEREMADE, CNRS and Université Paris-Dauphine, PSL Research University, Paris, France.

To be more precise, we consider both Neumann and Dirichlet boundary conditions to this setting. This will result in two different variants of (1): in the Neumann setting, <sup>1</sup>

$$\begin{aligned} \bar{u}_N &= \arg \min_{u \in BV \cap L^2(\Omega)} \frac{1}{2\lambda} \|u - g\|_{L^2}^2 + \text{TV}_N(u) =: E_N(u) \text{ where} & (2) \\ \text{TV}_N(u) &= \sup \left\{ - \int_{\Omega} u \operatorname{div} \phi : \phi \in \mathcal{C}_c^1(\Omega, \mathbb{R}^2), \|\phi\|_{\infty} \leq 1 \right\} \end{aligned}$$

where  $\mathcal{C}_c^1(\Omega, \mathbb{R}^2)$  is the space of continuously differentiable and compactly supported fields from  $\Omega$  to  $\mathbb{R}^2$ , and  $\|\phi\|_{\infty} = \sup_{x \in \Omega} |\phi(x)|$ . It is standard that the expression  $\text{TV}_N(u)$  is finite if and only if the distributional derivative  $Du$  is a Radon measure bounded in  $\Omega$ , in which case it coincides with the total mass  $|Du|(\Omega)$ , see for instance [1] for details.

In the Dirichlet setting, we add the constraint that  $u = b$  on  $\partial\Omega$  for some  $b \in L^{\infty}(\partial\Omega)$  (one usually takes  $b = g|_{\partial\Omega}$ ) and replace  $\text{TV}_N$  by

$$\text{TV}_D(u) = \sup \left\{ - \int_{\Omega} u \operatorname{div} \phi + \int_{\partial\Omega} b \langle \phi, \bar{n} \rangle : \phi \in \mathcal{C}^1(\Omega, \mathbb{R}^2), \|\phi\|_{\infty} \leq 1 \right\}$$

where  $\bar{n}$  denotes the outer normal unit vector. This can be proved to coincide  $|Du|(\Omega) + \int_{\partial\Omega} |u - b|$  for  $u$  with bounded variation, expressing the fact that if the trace of  $u$  fails to satisfy the boundary condition  $u = b$ , the energy has to penalize the corresponding jump. Then, we formulate the Dirichlet problem as:

$$\bar{u}_D = \arg \min_{u \in BV \cap L^2(\Omega)} \frac{1}{2\lambda} \|u - g\|_{L^2}^2 + \text{TV}_D(u) =: E_D(u). \quad (3)$$

In the following, we will denote for  $B \in \{N, D\}$  the optimal value of the continuous problems  $\bar{E}_B = E_B(\bar{u}_B)$ . When no subscript ( $N$  or  $D$ ) is used, it means our statement is valid under both boundary conditions.

In practice,  $\Omega$  is discretized into  $N \times N$  square pixels of size  $h = 1/N$ , namely  $\Omega = \cup_{1 \leq i, j \leq N} C_{i,j}$  with  $C_{i,j} = [(i-1)h, ih] \times [(j-1)h, jh]$ . Images are now elements of  $P0 = \{u : \Omega \rightarrow \mathbb{R} : \forall 1 \leq i, j \leq N, \exists u_{i,j} \in \mathbb{R}, u = u_{i,j} \text{ a.e. in } C_{i,j}\}$ . One introduces the projection of the continuous image  $g^h = \Pi_{P0}(g)$  given by  $(g^h)_C = \frac{1}{h^2} \int_C g$  for every square pixel  $C$ , and the discrete counterpart of (1) is the following:

$$\bar{u}^h = \arg \min_{u^h \in P0} \frac{1}{2\lambda} \|u^h - g^h\|_{L^2}^2 + \text{TV}^h(u^h) =: E^h(u^h) \quad (4)$$

where  $\text{TV}^h$  is some discretization of the total variation defined on  $P0$ . In the Dirichlet setting,  $\text{TV}^h$  can involve the discretization  $b^h$  of  $b$  given by  $(b^h)_e = \frac{1}{h} \int_e b$  for every

---

<sup>1</sup>Throughout the paper, the integrals are assumed to be with respect to the 2-dimensional Lebesgue measure, and the boundary integrals with respect to 1-dimensional Hausdorff measure on the boundary.

boundary edge  $e$ . This article deals with the study of the convergence rate of  $\bar{E}^h := E^h(\bar{u}^h)$  towards  $\bar{E}$  for two different discretizations  $\text{TV}^h$ .

A widely used choice for  $\text{TV}^h$  is the so called ‘‘isotropic’’ total variation which discretizes the expression  $\text{TV}(u)$  (which, for  $u$  with integrable gradient and satisfying, when required, the boundary condition, coincides with  $\int_{\Omega} |\nabla u|$ ) using a finite difference operator  $D$ . It is given by:

$$\text{TV}_i^h(u^h) = h \sum_{1 \leq i, j \leq N} |(Du^h)_{i,j}| \quad \text{where } (Du^h)_{i,j} = \begin{pmatrix} u_{i+1,j}^h - u_{i,j}^h \\ u_{i,j+1}^h - u_{i,j}^h \end{pmatrix} \quad (5)$$

(with either  $u_{N+1,j}^h = b_{N+\frac{1}{2},j}^h$ ,  $u_{i,N+1}^h = b_{i,N+\frac{1}{2}}^h$  in the case of a Dirichlet boundary condition or  $u_{N+1,j}^h - u_{N,j}^h = u_{i,N+1}^h - u_{i,N}^h = 0$  in the Neumann b.c. case). The scaling  $h = h^2/h$  in (5) corresponds to the size of an elementary pixel,  $h^2$ , divided by the length  $h$  which appears in the denominator when discretising the derivative of  $u$  with finite differences. The term ‘‘isotropic’’ refers to the behavior of this functional as the mesh size  $h$  tends to zero. One can indeed show (see [18] where this is proven for a more complicated TV) that the functional  $u \mapsto \text{TV}_i^h(u^h)$  if  $u = u^h \in P_0$ ,  $+\infty$  otherwise  $\Gamma$ -converges to TV, so that the minimizers  $\bar{u}^h$  converge (for instance in  $L^2$ ) to  $\bar{u}$ , the minimizer of (1). This convergence leads to thinking that  $\text{TV}_i^h$  inherits of the isotropy of TV for denoising problems such as ROF. We recall below the standard example of this isotropy of the continuous total variation: the denoising of the characteristic of a half plane in the Dirichlet setting.

Given a direction  $\nu \in \mathbb{R}^2$  with  $|\nu| = 1$ , take  $g = g_\nu$  defined by  $g_\nu(x) = 1$  if  $\langle x|\nu \rangle \geq a$  and  $g_\nu(x) = 0$  otherwise where  $a$  is some fixed real number (for instance  $a = \langle (1/2, 1/2)|\nu \rangle$ ). Then, problem (3) with boundary condition  $b = g_\nu|_{\partial\Omega}$  has solution  $\bar{u}_D = g_\nu$ , no matter the orientation of  $\nu$ . This comes from the following important fact:

**Claim.** Fix  $\nu \in \mathbb{R}^2$  with  $|\nu| = 1$ . Given the boundary condition  $b = g_\nu|_{\partial\Omega}$ , the minimal value of  $\text{TV}_D$  is  $\text{TV}_D(g_\nu) = \int_{\partial\Omega} g_\nu \langle \nu|\vec{n} \rangle$ .

This follows from two remarks: first, using Green’s theorem, one easily shows that  $\phi \equiv \nu$  reaches the maximum in the definition of  $\text{TV}_D(g_\nu)$ . We deduce that  $\text{TV}_D(g_\nu) = \int_{\partial\Omega} g_\nu \langle \nu|\vec{n} \rangle$ . Then, given  $u \in BV \cap L^2(\Omega)$ , taking again the admissible field  $\phi \equiv \nu$  in the definition of  $\text{TV}_D(u)$  yields

$$\text{TV}_D(u) \geq \int_{\partial\Omega} g_\nu \langle \nu|\vec{n} \rangle = \text{TV}_D(g_\nu)$$

and the claim follows.

However, this convergence result does not guarantee the isotropy of the discrete isotropic TV itself. In fact  $\text{TV}_i^h(g_\nu^h)$  can be quite far from the length of the continuous line  $\text{TV}(g_\nu)$ . What is worse is that the value of  $\text{TV}_i^h(g_\nu^h)$  actually depends on the

orientation of  $\nu$ . The case of the  $45^\circ$  diagonal is eloquent: as noted for instance in [15], the choice of the finite difference operator  $D$  induces a difference of roughly 40% between the main diagonal, that is  $\nu = \frac{1}{\sqrt{2}}(1, 1)$  and its flipped version that is  $\nu = \frac{1}{\sqrt{2}}(-1, 1)$ :

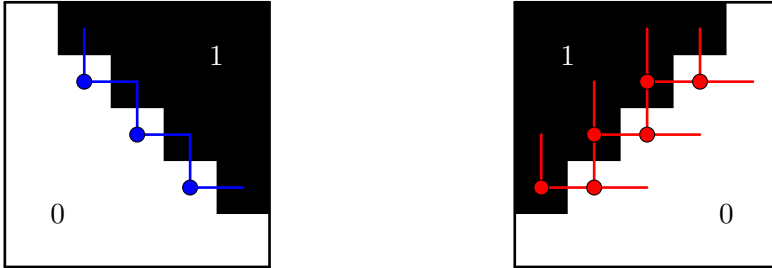


Figure 1: On the left image  $\text{TV}_i^h \simeq N\sqrt{2}$  while on the right  $\text{TV}_i^h \simeq 2N$

This differentiation breaks the isotropy of  $\text{TV}_i^h$  for a fixed  $h > 0$  leading to artefacts depending on the direction in denoising problems such as the denoising of a circle: the edges oriented along the more penalized diagonal are blurred (see Figure 2). Going back to the case  $g = g_\nu$ , even if one always has  $\bar{u}_h \rightarrow g_\nu$ , the speed of this convergence may vary with  $\nu$ . We take again the example of the two mirror diagonals for which denoising with  $\text{TV}_i^h$  for different step sizes  $h$  are shown Figure 2. One notices that the denoising is achieved correctly for the  $135^\circ$  diagonal  $\triangleright$  (which we will now call consequently the “good” diagonal) whereas one needs to take  $h$  very small before obtaining a sharp looking discontinuity with the other diagonal  $\triangleleft$  (the “bad” one).

To mitigate these issues, many different paths have been pursued in the imaging community. Graph-based approaches have been quite successful and may be enriched [10, 27, 13] to yield a more isotropic behaviour. Yet, most of the time they aim at obtaining in the continuous limit a “crystalline” total variation or perimeter which approximates the Euclidean length, however with a behaviour quite different from a “true” isotropic approximation (see also [28] where this phenomenon is studied on regular finite element meshes). We may refer to [20, 33] for attempts to design isotropic perimeters or total variations on graphs or using finite differences. An exhaustive study and comparison of discrete and continuous graph-based approaches (with an interesting experimental study of the convergence rates) is developed in [27].

While [20] (see also [2] which relies on a similar discretization of the constraints) share some similarities with the Raviart-Thomas approximation we introduce below, it is different and does not come with a numerical analysis, on the other hand, [33] provide an analysis of their approximations but the precision is low and we may expect it to be as “bad” as for  $\text{TV}_i^h$  above, in most directions. Also [32], based on a discontinuous Galerkin approximation and some mesh adaptation, seems to perform well but its numerical analysis seems difficult. A few approaches based on P1 [3] or discontinuous P1 (yet still

conforming) [23] yield a better isotropic behaviour, and can include mesh optimization, yet we also expect with these approaches that the local directions of the mesh have an influence on the precision of the result, see in particular [17] for a discussion. Most of the literature mentioned above is qualitative (with the important exception of [33, 3]), and despite its success, does not come with quantitative estimates. We address here this problem from the point of view of numerical analysis, trying to understand how (un)precise the standard approximation is for simple problems, and what precision can be expected with a slightly better designed variant.

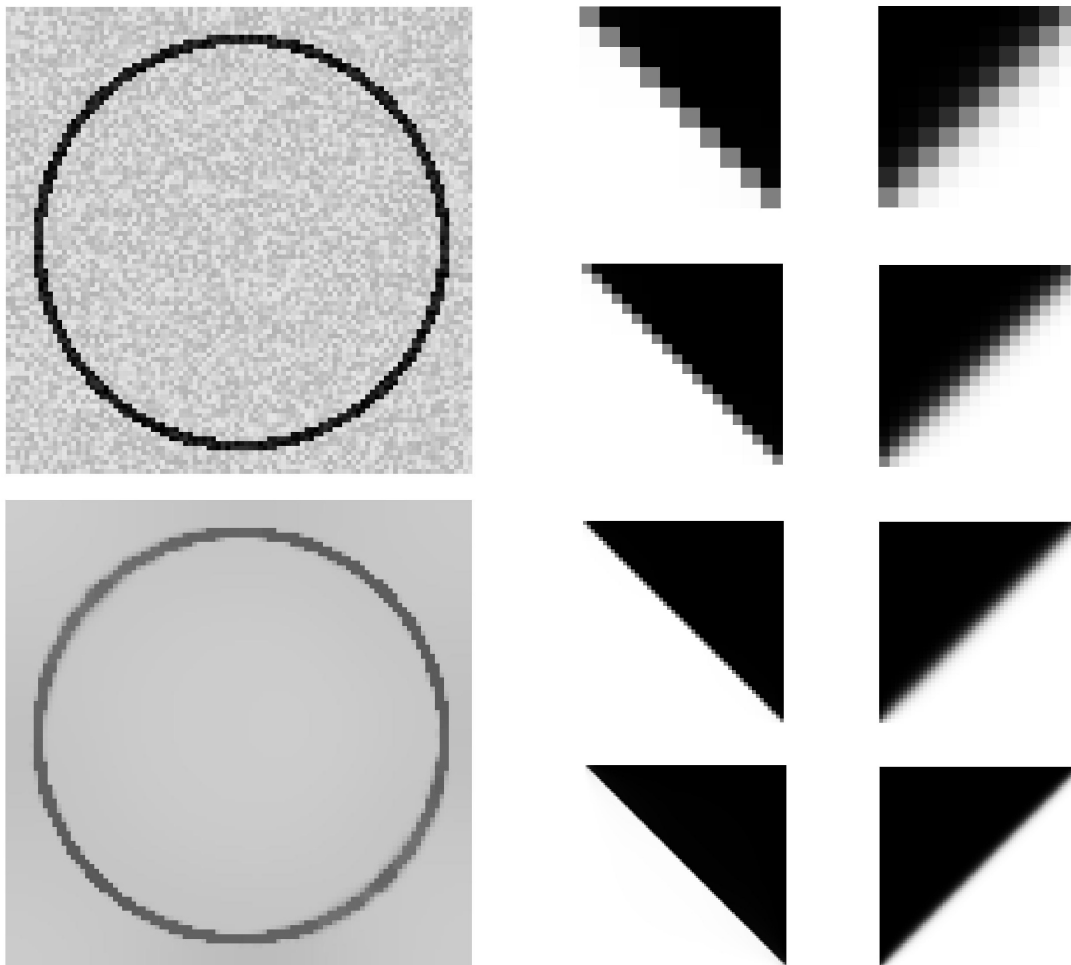


Figure 2: Denoising with  $TV_i^h$ : noisy and denoised circle with Neumann b.c., good (2nd col.) and bad (3rd col.) diagonals with Dirichlet b.c. and  $N = 10, 20, 50, 100$ .

The first goal, in this paper, is to quantitatively study the error resulting from the use of the discretization  $TV_i^h$  of the isotropic total variation in the “bad” diagonal

directions (Fig. 2, last column). To this end, we estimate the **rate of convergence** of the optimal discrete energy of problem (4) towards the optimal continuous energy in (1) **in simplified settings (in particular, translation invariant domains) which make the analysis tractable**. Up to a change of the domain **which reduces the problem to the study of a one-dimensional problem**, we show that it is of order  $O(h^{2/3})$  in the “bad” direction. More precisely we will prove the following theorem:

**Theorem 1.** *On an appropriate domain (depending on the orientation  $\nu$ ) we have:*

1. For  $\nu = \frac{1}{\sqrt{2}}(1, 1)$  the denoising is exact, meaning that  $\bar{u}^h = \Pi_{P0}(\bar{u})$ ,  $\bar{E}^h = \bar{E}$ .
2. For  $\nu = \frac{1}{\sqrt{2}}(-1, 1)$ ,  $\exists \underline{h}, c, c' > 0$  depending only on  $\lambda$  such that

$$\forall h \leq \underline{h}, \quad ch^{2/3} \leq \bar{E}^h - \bar{E} \leq c'h^{2/3}.$$

The domains on which these estimates hold are given, for  $\nu = (\pm 1, 1)/\sqrt{2}$ , by the periodic strips:  $\Omega_{per} = \{(i, j) : -N \leq i \pm j \leq N\}$  with, for all  $(i, j) \in \Omega_{per}$ , the point  $(i \mp D, j \mp D)$  identified with  $(i, j)$ . Here  $N \geq 1$  and  $D \geq 1$  are given integers.

These rates ought to be compared with results obtained by Lucier and co-authors in [25] and [33]. In [25] the authors give a bound of type  $|\bar{E}^h - \bar{E}| \leq c\sqrt{h}$  (as well as  $\|\bar{u}^h - \bar{u}\|^2 \leq c\sqrt{h}$ ) for a so called central-difference discretization of the ROF model meaning that they use the following discrete total variation  $\text{TV}^h = \text{TV}_c^h$  with

$$\text{TV}_c^h(u^h) = h \sum_{i,j} \sqrt{\left(\frac{u_{i+1,j}^h - u_{i-1,j}^h}{2}\right)^2 + \left(\frac{u_{i,j+1}^h - u_{i,j-1}^h}{2}\right)^2}.$$

In [33], errors in  $h^{\frac{\alpha}{\alpha+1}}$  are given, where  $\alpha$  is the Lipschitz order of  $g$ , and the discrete total variation at stake is an average of the four possible isotropic total variations obtained by the finite difference approximations of the gradient: forward/forward (which is  $\text{TV}_i^h$ ), forward/backward, backward/forward and backward/backward. **We expect that the analysis provided in this paper could be adapted to these variants and yield also sub-optimal approximation errors in most directions.**

In a second part of this paper, starting in Section 3, we establish a convergence rate (valid under some hypothesis)  $|\bar{E}^h - \bar{E}| \leq ch$  for another discrete total variation denoted  $\text{TV}_{RT}^h$ : this is of course much better. It also comes with an improved rate of convergence for the solutions. A similar error is obtained in [15] for a non-conforming P1 finite-elements based approximation of the total variation (see also the extensions [4, 5, 7]). The idea behind the  $\text{TV}_{RT}^h$  total variation is to gain isotropy in the discretization of the continuous TV by allowing any direction  $\nu$  to be an admissible discrete field  $\phi$ . We propose to mimic the dual definition of the continuous total variation, in the Dirichlet setting  $\text{TV}_D$ , **replacing the smooth dual vector fields by** Raviart-Thomas fields (RT0) [29]

which are piecewise affine fields (whose precise definitions will be given later on), and define for either continuous or discrete functions  $u$  and boundary term  $b$ :

$$\mathrm{TV}_{RT,D}^h(u) = \sup \left\{ - \int_{\Omega} u \operatorname{div} \phi + \int_{\partial\Omega} b \langle \phi | \vec{n} \rangle : \phi \in RT0, \|\phi\|_{\infty} \leq 1 \right\}. \quad (6)$$

The fact that a constant  $\nu \in RT0$  allows to show that this total variation is “isotropic” in the sense that when taking  $b = g_{\nu}$  one has **that for any  $\nu \in \mathbb{S}^1$ ,  $\bar{u} = g_{\nu}$  is a minimizer of the ROF model with data term  $g_{\nu}$ :**

$$\min_{\substack{u \in BV \cap L^2(\Omega) \\ u|_{\partial\Omega} = g_{\nu}|_{\partial\Omega}}} \frac{1}{2\lambda} \|u - g_{\nu}\|_{L^2}^2 + \mathrm{TV}_{RT,D}^h(u).$$

Surprisingly, a similar result even holds for a purely Dirichlet problem (with no quadratic penalty, corresponding to  $\lambda = +\infty$ ), see [17, Prop. 4.1].

We define the Raviart-Thomas ROF discrete problem as:

$$\bar{u}^h = \arg \min_{u^h \in P0} \frac{1}{2\lambda} \|u^h - g\|_2^2 + \mathrm{TV}_{RT}^h(u^h) \quad (7)$$

where  $\mathrm{TV}_{RT}^h$  stands for  $\mathrm{TV}_{RT,D}^h$  given by (6) in the Dirichlet setting, and is replaced by

$$\mathrm{TV}_{RT,N}^h(u^h) = \sup \left\{ - \int_{\Omega} u^h \operatorname{div} \phi : \phi \in RT0_0, \|\phi\|_{\infty} \leq 1 \right\}$$

in the Neumann setting: in this latter definition, the fields  $RT0_0$  are fields  $RT0$  with vanishing flux through the boundary  $\partial\Omega$ . In this new framework, we can show the following result, in contrast with Theorem 1:

**Theorem 2.** *Let  $g \in L^{\infty}(\Omega)$ ,  $\bar{u}$  be the solution of (2) or (3), and, for  $h > 0$  small,  $\bar{u}^h$  be the solution of the corresponding discrete problem (7). Denote  $\bar{E}$ ,  $\bar{E}^h$  the respective minimal energies. In the Dirichlet case, assume in addition  $b \in BV(\partial\Omega)$ . Then if the dual problem of (2)-(3) admits a Lipschitz-continuous solution, there exists  $c$  (depending on  $g$ ,  $\lambda$  and the Lipschitz constant of the dual solution) such that:*

$$\begin{aligned} |\bar{E} - \bar{E}^h| &\leq ch \\ \|\bar{u}^h - \bar{u}\|_{L^2} &\leq c\sqrt{h}. \end{aligned}$$

The assumption on the dual solution will be made clear in Section 3 where this result is proved. It is satisfied for simple test cases, such as the Dirichlet case  $g = \chi_B$ ,  $b = 0$  where  $B \subset \Omega$  is a ball. In this case, the estimate above is optimal, since the projection of  $\bar{u}$  on piecewise constant functions with scale  $h$  is at  $L^2$  distance  $O(\sqrt{h})$  from  $\bar{u}$ . The result is inspired from [15] where similar estimates (and the same hypotheses) are shown in a slightly different context (see also [4, 5, 8]).



The paper is organised as follows: The estimate in Theorem 1 is proved in Section 2, which is divided in many parts. Then, Section 3 introduces the Raviart-Thomas total variation and proves Theorem 2. In Section 4, we show numerical results comparing on test images the total variations  $\text{TV}_i^h$ ,  $\text{TV}_{RT}^h$  and a state-of-the-art variant initially proposed in [24], and analysed and implemented by Condat [19] which seems to perform better, but for which we do not have error bounds up to now, while consistency has only been established recently [16].

## 2 Proof of Theorem 1

In this section we prove, in many steps, the main estimate in Theorem 1. First, we introduce the setting, which consists in considering a domain made of an infinite periodic strip in order to reduce the 2D case to a one-dimensional problem: this is developed in Section 2.1. We then analyse the continuous limit of this one-dimensional problem and exhibit the main symmetry properties of the discrete and continuous solutions.

Section 2.2 is devoted to the proof of the upper bound in (the second point of) Theorem 1. We rapidly show that a sharp upper bound cannot be obtained by a too simple linear ramp and develop then a more refined strategy to build a discrete approximation which reaches the energy error  $O(h^{2/3})$ .

We prove then the lower bound in Section 2.3. This is quite difficult. As we need to bound the energy from below, we first introduce a dual maximization problem. Then, a change of variable allows to identify a candidate for a continuum limit of this problem. We study this limit (Section 2.3.3) and show (by passing to a simpler “dual of the dual”!) that it admits solutions. Eventually in Section 2.3.4, we discretize back these solutions, and show that it gives the desired bound for the dual problem.

### 2.1 Reduction to a 1D TV denoising problem

To study the orientation dependent error of the isotropic TV, we introduce the following experiment. Placing ourselves in a well-chosen periodic domain  $\Omega = \Omega_{per}$ , we reduce the two-dimensional  $\text{TV}_i^h$  denoising problem in the case of a diagonal image  $g = g_\nu$  with  $\nu = \frac{1}{\sqrt{2}}(-1, 1)$  to a one-dimensional problem. In the following, we will denote respectively TV and tv the 2D and 1D total variations. The first point of Theorem 1, which is the case  $\nu = \frac{1}{\sqrt{2}}(1, 1)$ , will be quickly obtained. We next present some general results about the case  $\nu = \frac{1}{\sqrt{2}}(-1, 1)$  that will be useful to prove the second point of Theorem 1 in the following sections.

#### 2.1.1 The domain $\Omega_{per}$

We actually do not consider the ROF model (4) on a square domain, but on a periodic strip oriented along the diagonal at stake, see the drawing below in which each square

pixel is of size  $h = 1/N$  and where the (green) dotted lines are to be glued together. For  $\nu = \frac{1}{\sqrt{2}}(-1, 1)$ , we now work with a variable  $u_{i,j}^h$  defined in the domain  $\Omega_{per}$  introduced in Theorem 1, that is, for  $(i, j) \in \mathbb{Z}^2$  such that  $-N \leq i - j \leq N$ ;  $0 \leq i + j \leq 2D$  and satisfying  $u_{i+D,j+D}^h = u_{i,j}^h$  for any  $(i, j)$ . Introducing the change of variables  $n = i - j$ ,  $d = \lfloor \frac{i+j}{2} \rfloor$ , the domain is represented by:

$$\Omega_{per} = \left\{ (n, d) : -N \leq n \leq N, d \in \mathbb{Z}/D\mathbb{Z} \right\}.$$

The source term  $g^h : \Omega_{per} \rightarrow \mathbb{R}$  is the projection of  $g_\nu$  ( $\nu = \frac{1}{\sqrt{2}}(1, -1)^2$ ) on piecewise constant functions, which in the new variables is given by:

$$g^h(n, d) = g_n^h = \begin{cases} 0 & \text{if } n < 0, \\ 1/2 & \text{if } n = 0, \\ 1 & \text{if } n > 0 \end{cases} \quad \text{for all } d \in \mathbb{Z}/D\mathbb{Z}. \quad (8)$$

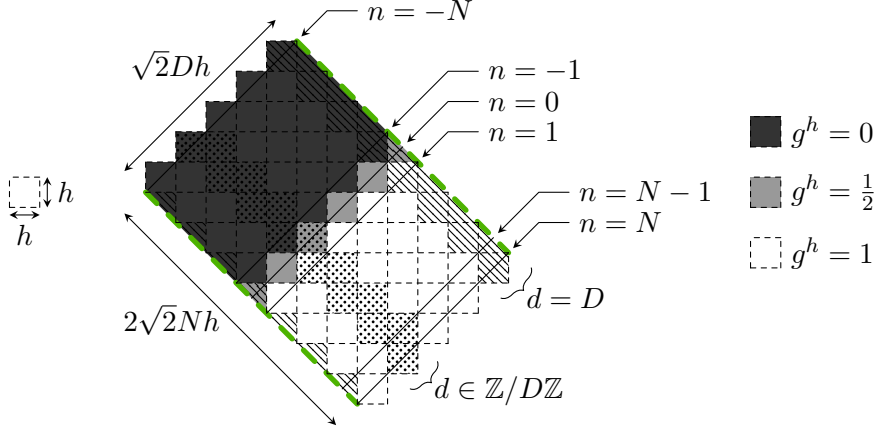


Figure 3: Setting for the lower error bound estimate.

Then the problem (4) is to solve

$$\bar{u}^h = \arg \min_{u^h: \Omega_{per} \rightarrow \mathbb{R}} \frac{h^2}{2\lambda} \sum_{(n,d) \in \Omega_{per}} |u^h(n, d) - g^h(n, d)|^2 + \text{TV}_i^h(u) := E^h(u^h)$$

where  $\text{TV}_i^h$  stands for the isotropic TV on this particular domain. If  $n = i - j$  and  $d = \lfloor \frac{i+j}{2} \rfloor$  so that  $u^h(n, d)$  codes for the value of  $u_{i,j}^h$ , then one finds that  $u_{i+1,j}^h$  (respectively  $u_{i,j+1}^h$ ) is represented by  $u^h(n + 1, d')$  (respectively  $u^h(n - 1, d')$ ) with  $d' = d$  for  $n$  even

<sup>2</sup>For some reason, we considered here, in fact, the case  $\nu = \frac{1}{\sqrt{2}}(1, -1)$ , which is of course equivalent to the opposite direction—and whose solutions are nondecreasing in the new variable  $n$ .

and  $d' = d + 1$  for  $n$  odd. **The change of variable  $(i, j) \rightarrow (n, d)$  leads in eq. (5) to the following expression of the isotropic total variation:**

$$\text{TV}_i^h(u^h) = h \sum_{\substack{d \in \mathbb{Z}/D\mathbb{Z} \\ n \text{ even}}} \left| \begin{pmatrix} u^h(n+1, d) - u^h(n, d) \\ u^h(n-1, d) - u^h(n, d) \end{pmatrix} \right| + h \sum_{\substack{d \in \mathbb{Z}/D\mathbb{Z} \\ n \text{ odd}}} \left| \begin{pmatrix} u^h(n+1, d+1) - u^h(n, d) \\ u^h(n-1, d+1) - u^h(n, d) \end{pmatrix} \right|.$$

We will first study the case of Dirichlet boundary conditions, meaning that we impose (both on the definition of  $\text{TV}_i^h$  and on the optimization problem) that for all  $d \in \mathbb{Z}/D\mathbb{Z}$ :

$$u^h(N+1, d) = u^h(N, d) = g^h(N, d) = 1, \quad u^h(-N-1, d) = u^h(-N, d) = g^h(-N, d) = 0.$$

Later on we will deduce from the Dirichlet setting the same rate for the Neumann boundary conditions:

$$u^h(N+1, d) = u^h(N, d), \quad u^h(-N-1, d) = u^h(-N, d).$$

The benefit of this periodic setting is to reduce the problem **to a one-dimensional study, since** at the optimum one has  $\bar{u}^h(n, d) = \bar{u}^h(n, d')$  for all  $n$  and  $d, d' \in \mathbb{Z}/D\mathbb{Z}$ . Indeed, as all the terms in the objective are invariant when changing  $d$  to  $d + 1$ , the shifted image  $\tilde{u}^h : (n, d) \mapsto \bar{u}^h(n, d + 1)$  has the same energy  $E^h$ , hence  $\tilde{u}^h = \bar{u}^h$  by uniqueness of the optimizer.

We keep the letter  $u$  for **the variable of the 1D problem**, and **renormalize the energy** by a factor  $\sqrt{2}Dh$  **corresponding to the width of the periodic strip. We obtain:**

$$\bar{u}^h = \arg \min_{\substack{u^h \in \mathbb{R}^{2N+1} \\ \text{with B.C.}}} E^h(u^h) := \frac{h}{2\sqrt{2}\lambda} \|u^h - g^h\|_2^2 + \text{tv}_i^h(u^h) \quad (9)$$

**with  $g^h$  given by (8), where we defined:**

$$\begin{cases} \|u^h - g^h\|_2^2 = \sum_{n=-N}^N (u_n^h - g_n^h)^2 \\ \text{tv}_i^h(u^h) = \frac{1}{\sqrt{2}} \sum_{n=-N}^N \sqrt{(u_{n+1}^h - u_n^h)^2 + (u_n^h - u_{n-1}^h)^2}, \end{cases}$$

and where BC stands for the following boundary conditions:

$$\begin{cases} u_{N+1}^h = u_N^h = 1 \text{ and } u_{-N-1}^h = u_{-N}^h = 0 \text{ for Dirichlet} \\ u_{N+1}^h = u_N^h \text{ and } u_{-N-1}^h = u_{-N}^h \text{ for Neumann.} \end{cases}$$

**The resulting problem is a one-dimensional TV-denoising problem which relies on a modified 1D total variation:**  $\frac{1}{\sqrt{2}} \sum_n \sqrt{(u_{n+1}^h - u_n^h)^2 + (u_n^h - u_{n-1}^h)^2}$ . **The interaction**

between two consecutive differences is responsible for the bad behavior of  $\text{TV}_i^h$  on this diagonal.

On the other hand, for the direction  $\nu = \frac{1}{\sqrt{2}}(1, 1)$ , the domain  $\Omega_{per}$  defined in Theorem 1 is now oriented along the other diagonal. One can check that a similar change of variables leads now to a one-dimensional TV-denoising problem with the classical 1D discrete total variation  $\text{tv}^h(u^h) = \sum_n |u_{n+1}^h - u_n^h|$ . As a consequence, the denoising is exact:  $\bar{u}^h = g^h$ . Indeed, the problem (in the Dirichlet setting) is to minimize  $\|u^h - g^h\|^2 + c \text{tv}^h(u^h)$  for some constant  $c > 0$  and under the constraint that  $u_{N+1}^h = u_N^h = 1$  and  $u_{-N}^h = 0$ . This constraint yields  $\text{tv}^h(u^h) \geq \left| \sum_{n=-N}^N u_{n+1}^h - u_n^h \right| = 1 = \text{tv}^h(g^h)$ , showing the first point in Theorem 1.

### 2.1.2 Solution of the continuous limit problem

In this section we investigate the continuous 1D denoising problem obtained when passing to the limit  $h \rightarrow 0$  in (9). Assuming  $u^h$  is the discretization of a smooth function  $u$  defined on  $[-1, 1]$ , we write:

$$E^h(u^h) = \frac{1}{N\sqrt{2}} \sum_{n=-N}^N \frac{1}{2\lambda} (u(nh) - g_n^h)^2 + \sqrt{\left( \frac{u(nh+h) - u(nh)}{h} \right)^2 + \left( \frac{u(nh) - u(nh-h)}{h} \right)^2}$$

and we see that this converges as  $h \rightarrow 0$  to

$$E(u) = \int_{-1}^1 \frac{1}{2\sqrt{2}\lambda} (u - g)^2 + |u'| \quad (10)$$

with  $\int_{-1}^1 |u'| =: \text{tv}(u)$  being the continuous 1D total variation.

It is easily shown that (10) is also the  $\Gamma$ -limit of the discrete problem, so that the minimizers  $\bar{u}^h$  of (9) will converge to the minimizer of (10). Indeed, the above discussion for a smooth function  $u$  establishes a “ $\Gamma$ -lim sup” inequality. For the “ $\Gamma$ -lim inf”, consider  $(u^h)$  such that  $\sum_{n=-N}^N u(nh) \chi_{(nh-\frac{1}{2}, nh+\frac{1}{2})} \rightarrow u$  as  $h = 1/N \rightarrow 0$  in  $L^2(-1, 1)$ , and let  $\phi \in C_c^\infty((-1, 1); [-1, 1])$ . Then,

$$E^h(u^h) \geq \frac{1}{N\sqrt{2}} \sum_{n=-N}^N \frac{1}{2\lambda} (u(nh) - g_n^h)^2 + \frac{1}{\sqrt{2}} \phi(nh + \frac{h}{2}) \frac{u(nh+h) - u(nh)}{h} + \frac{1}{\sqrt{2}} \phi(nh - \frac{h}{2}) \frac{u(nh) - u(nh-h)}{h}.$$

The first term clearly goes to  $\frac{1}{2\sqrt{2}\lambda} \int_{-1}^1 (u - g)^2$ , while the second is:

$$\frac{1}{2N} \sum_n 2 \frac{\phi(nh - \frac{h}{2}) - \phi(nh + \frac{h}{2})}{h} u(nh) \rightarrow - \int_{-1}^1 \phi' u$$

as  $h \rightarrow 0$ . Taking the supremum with respect to  $\phi$ , we recover  $E(u) \leq \liminf_{h \rightarrow 0} E^h(u^h)$ . The proof in the Dirichlet setting needs a bit of adaption, using  $\phi \in C^\infty([-1, 1]; [-1, 1])$  and some extra care at the boundary points  $n = \pm N$ .

For the Dirichlet setting, we enforce the constraint  $u = g$  at the boundary of the domain i.e.  $u(-1) = 0$  and  $u(1) = 1$ . In that situation, for any admissible  $u$  we have:

$$\int_{-1}^1 |u'| \geq \left| \int_{-1}^1 u' \right| = |u(1) - u(-1)| = 1 = \int_{-1}^1 |g'|$$

which directly shows that the energy (10) is minimal for  $u = g$  with value  $\bar{E}_D = 1$ .

In the Neumann setting however, no boundary condition is **enforced**. To find the solution, one writes the optimality conditions given by duality theory (see [11]):

$$\text{tv}(u) = - \int_{-1}^1 uz' \quad \text{and} \quad \frac{1}{\sqrt{2\lambda}}(u - g) - z' = 0$$

for some function  $z$  such that  $|z| \leq 1$  and  $z(-1) = z(1) = 0$ . If these equations are met for some couple  $(u, z)$  then  $u$  is optimal in problem (10). We search for  $u$  of the form  $u = u_a$  for some  $a \in \mathbb{R}$  with  $u_a(x) = a$  if  $x \in (-1, 0)$  and  $u_a(x) = 1 - a$  if  $x \in (0, 1)$ . This leads to taking  $z(x) = \frac{a}{\sqrt{2\lambda}}(x+1)$  if  $x \in (-1, 0)$  and  $z(x) = \frac{a}{\sqrt{2\lambda}}(1-x)$  if  $x \in (0, 1)$ . Then one must try to fulfill the equations  $\text{tv}(u_a) = - \int_{-1}^1 u_a z'$  that is  $|1 - 2a| = \frac{1}{\sqrt{2\lambda}}a(1 - 2a)$  and  $|z| \leq 1$  that is  $|a| \leq \sqrt{2\lambda}$ . These two equations on  $a$  always give rise to a unique solution: if  $\lambda \leq \lambda^* := \frac{\sqrt{2}}{4}$  then  $u_a$  is optimal with  $a = a_{opt} := \sqrt{2\lambda}$  and the minimal energy is  $\bar{E}_N^{\leq} := 1 - \sqrt{2\lambda}$ . If  $\lambda > \lambda^*$  then  $u_a$  is optimal with  $a = \frac{1}{2}$  and the minimal energy is  $\bar{E}_N^{>} := \frac{1}{4\sqrt{2\lambda}}$ . In the following, we will see that in the case  $\lambda > \lambda^*$  the discrete problem is exact ( $\bar{u}^h \equiv \frac{1}{2}$ ), therefore we will always place ourselves in the case  $\lambda \leq \lambda^*$ , and we denote  $\bar{E}_N := \bar{E}_N^{\leq} = 1 - \sqrt{2\lambda}$ .

### 2.1.3 Qualitative properties of the solution

Before turning to the proof of the  $O(h^{2/3})$  bounds, we **show some properties** of the **minimizer** of (9), for  $g$  given by (8):

**Proposition 1.** *The solution  $\bar{u}^h$  of problem (9) (either with Dirichlet or Neumann boundary conditions) satisfies:*

1.  $\forall n, \bar{u}_{-n}^h = 1 - \bar{u}_n^h$ , in particular  $\bar{u}_0^h = \frac{1}{2}$ .
2.  $\forall n > 0, 1 \geq \bar{u}_n^h \geq \frac{1}{2}$ , hence  $\forall n < 0, 0 \leq \bar{u}_n^h \leq \frac{1}{2}$ .
3.  $\bar{u}^h$  is non-decreasing:  $\forall n, \bar{u}_{n+1}^h \geq \bar{u}_n^h$ .

*Proof.* For the first point, the symmetry of  $g^h$  ( $g_{-n}^h = 1 - g_n^h$ , see (8)) and  $\text{tv}_i^h$  yields that  $\tilde{u}_n^h = 1 - \bar{u}_{-n}^h$  satisfies  $E^h(\tilde{u}^h) = E^h(\bar{u}^h)$ . By uniqueness of the minimizer,  $\tilde{u}^h = \bar{u}^h$ .

For the second point, the truncated variable  $\hat{u}_n^h = \max(g_n^h, \min(\bar{u}_n^h, \frac{1}{2}))$  for  $n \leq 0$  satisfies  $|\hat{u}_n^h - g_n^h| \leq |\bar{u}_n^h - g_n^h|$  and  $|\hat{u}_{n+1}^h - \hat{u}_n^h| \leq |\bar{u}_{n+1}^h - \bar{u}_n^h|$  for any  $n$ , hence  $E^h(\hat{u}^h) \leq E^h(\bar{u}^h)$  and  $\bar{u}^h = \hat{u}^h$ .

For the third point, consider the staircase version of  $\bar{u}^h$  given by:  $\check{u}_n^h = \max\{\bar{u}_k^h, 0 \leq k \leq n\}$  if  $n > 0$ ,  $\check{u}_0^h = \frac{1}{2}$  and  $\check{u}_n^h = \min\{\bar{u}_k^h, n \leq k \leq 0\}$  if  $n < 0$ . As  $\bar{u}_n^h \in [0, 1]$  we have  $|\check{u}_n^h - g_n^h| \leq |\bar{u}_n^h - g_n^h|$ , and again  $|\check{u}_{n+1}^h - \check{u}_n^h| \leq |\bar{u}_{n+1}^h - \bar{u}_n^h|$  for any  $n$ , hence  $E^h(\check{u}^h) \leq E^h(\bar{u}^h)$  and  $\bar{u}^h = \check{u}^h$ .  $\square$

**Proposition 2.** We denote  $\lambda^* = \frac{\sqrt{2}}{4}$ . The solution  $\bar{u}^h$  of problem (9) is such that:

1. With Dirichlet boundary conditions,  $\bar{u}_1^h > \frac{1}{2}$  for any  $\lambda$ .
2. With Neumann boundary conditions,  $\bar{u}^h \equiv \frac{1}{2}$  for any  $\lambda \geq \lambda^*$  and  $\bar{u}_1^h > \frac{1}{2}$  for any  $\lambda < \lambda_h^*$  for some  $\lambda_h^*$  such that  $|\lambda_h^* - \lambda^*| \leq ch^{1/3}$  for some constant  $c > 0$ . In particular, for any  $\lambda < \lambda^*$  one has  $\bar{u}_1^h > \frac{1}{2}$  for  $h$  small enough.

*Proof.* For  $u \in \mathbb{R}^{2N+1}$  satisfying the three properties of Proposition 1 and such that  $u_1 = \frac{1}{2}$ , we define  $k \in \{1, \dots, N\}$  such that  $u_{-1} = u_0 = \dots = u_k = \frac{1}{2}$  and  $u_{k+1} > \frac{1}{2}$ . Suppose first that  $k \leq N - 2$  then the energy of  $u$  can be written

$$\begin{aligned} E^h(u) &= \frac{h}{2\sqrt{2}\lambda}(u_k - 1)^2 + \frac{1}{\sqrt{2}}|u_k - \frac{1}{2}| + \frac{1}{\sqrt{2}}\sqrt{(u_{k+1} - u_k)^2 + (u_k - \frac{1}{2})^2} \\ &\quad + \frac{1}{\sqrt{2}}\sqrt{(u_{k+2} - u_{k+1})^2 + (u_{k+1} - u_k)^2} + R(u) \end{aligned}$$

where  $R(u)$  does not depend on  $u_k$ . As  $u_{k+1} > \frac{1}{2}$ , we have the following derivatives or subgradients:

$$\begin{aligned} \frac{\partial}{\partial u_k} \left( \sqrt{(u_{k+1} - u_k)^2 + (u_k - \frac{1}{2})^2} \right) \Big|_{u_k = \frac{1}{2}} &= \left( \frac{(u_k - u_{k+1}) + (u_k - \frac{1}{2})}{\sqrt{(u_{k+1} - u_k)^2 + (u_k - \frac{1}{2})^2}} \right) \Big|_{u_k = \frac{1}{2}} = -1 \\ \frac{\partial}{\partial u_k} \left( \sqrt{(u_{k+2} - u_{k+1})^2 + (u_{k+1} - u_k)^2} \right) \Big|_{u_k = \frac{1}{2}} &= \frac{\frac{1}{2} - u_{k+1}}{\sqrt{(\frac{1}{2} - u_{k+1})^2 + (u_{k+2} - u_{k+1})^2}} = d < 0 \\ \frac{\partial}{\partial u_k} ((u_k - 1)^2) \Big|_{u_k = \frac{1}{2}} &= -1 \text{ and } \frac{\partial}{\partial u_k} (|u_k - \frac{1}{2}|) \Big|_{u_k = \frac{1}{2}} = [-1, 1]. \end{aligned}$$

Finally  $\frac{\partial E^h}{\partial u_k} \Big|_{u_k = \frac{1}{2}} = -\frac{h}{2\sqrt{2}\lambda} + \frac{1}{\sqrt{2}}[-1, 1] - \frac{1}{\sqrt{2}}(1 - d) \subset \mathbb{R}_*^-$  so that  $0 \notin \frac{\partial E^h}{\partial u_k} \Big|_{u_k = \frac{1}{2}}$  hence  $u$  is not optimal. For  $k = N - 1$  the same reasoning is correct in the Dirichlet setting noting that  $u_{k+2} = 1$  whereas in the Neumann setting it is changed to

$$\begin{aligned} E^h(u) &= \frac{h}{2\sqrt{2}\lambda}(u_k - 1)^2 + \frac{1}{\sqrt{2}}|u_k - \frac{1}{2}| \\ &\quad + \frac{1}{\sqrt{2}}\sqrt{(u_{k+1} - u_k)^2 + (u_k - \frac{1}{2})^2} + \frac{1}{\sqrt{2}}|u_{k+1} - u_k| + R(u) \end{aligned}$$

for which one computes  $\frac{\partial E^h}{\partial u_k} |_{u_k=\frac{1}{2}} = -\frac{h}{2\sqrt{2}\lambda} + \frac{1}{\sqrt{2}}[-1, 1] - \frac{2}{\sqrt{2}} \in \mathbb{R}_*^-$  and gets the same conclusion. This concludes the proof in the Dirichlet setting as in this case  $k < N$ .

In the Neumann setting, the case  $k = N$  corresponds to our alternative  $\bar{u}^h \equiv \frac{1}{2}$  so that we only have to exhibit an admissible  $u^h$  such that  $E^h(u^h) < E(\frac{1}{2})$  to prove that  $\bar{u}_1^h > \frac{1}{2}$ . We postpone this construction to Section 2.2.3 where, provided that  $\lambda < \lambda^*$ , we will explicitly build a  $u^h$  such that  $E^h(u^h) \leq 1 - \lambda\sqrt{2} + ch^{2/3}$  for some constant  $c > 0$ . In comparison the energy of the constant  $u^h \equiv \frac{1}{2}$  is  $E^h(\frac{1}{2}) = \frac{h}{2\sqrt{2}\lambda} \times 2N \times (\frac{1}{2})^2 = \frac{\sqrt{2}}{8\lambda}$ . The conclusion comes from studying when  $1 - \lambda\sqrt{2} + ch^{2/3} < \frac{\sqrt{2}}{8\lambda}$ .

Finally, suppose now that  $\lambda \geq \lambda^*$ , we want to prove that  $\bar{u}^h \equiv \frac{1}{2}$ . For any  $u \in \mathbb{R}^{2N+1}$  satisfying the three properties of Proposition 1, denoting  $a = u_{-N} \in [0, \frac{1}{2}]$  we form the following estimate. On one hand, as  $u$  is non-decreasing, the  $L^2$  term  $\|u - g^h\|^2$  is bounded below by  $\|u^a - g^h\|^2$  where  $u_n^a = a$  for  $n < 0$ ,  $u_0^a = \frac{1}{2}$  and  $u_n^a = 1 - a$  for  $n > 0$ . On the other hand, **one has**  $\sqrt{(u_n - u_{n+1})^2 + (u_n - u_{n-1})^2} \geq |u_{n+1} - u_{n-1}|/\sqrt{2} = \sqrt{2}(u_{n+1} - u_{n-1})/2$ . We obtain:

$$E^h(u) \geq \frac{h}{2\sqrt{2}\lambda} \times 2Na^2 + \frac{1}{2}(u_{N+1} + u_N - u_{-N-1} - u_{-N}) = \frac{\sqrt{2}}{2\lambda}a^2 + 1 - 2a.$$

As  $\lambda \geq \lambda^* = \frac{\sqrt{2}}{4}$ , minimizing this quantity over  $a \in [0, \frac{1}{2}]$  leads to taking  $a = \frac{1}{2}$ , and we get  $E^h(u) \geq \frac{\sqrt{2}}{8\lambda} = E^h(\frac{1}{2})$ , hence  $\bar{u}^h \equiv \frac{1}{2}$ .  $\square$

## 2.2 Upper bound for the primal energy

In this section we prove the upper bound of the point 2 of Theorem 1, that is:  $\exists \underline{h}, c > 0$  such that

$$\forall h \leq \underline{h}, \bar{E}^h - \bar{E} \leq ch^{2/3}.$$

We first focus on the Dirichlet case and later on present the modifications needed for Neumann boundary conditions. As no reference to the continuous problem will appear in this section (except from its value  $\bar{E}$ ) we drop the exponent  $h$  and denote the variables  $u^h, g^h$  simply by  $u, g \in \mathbb{R}^{2N+1}$ . Recall that the primal problem in the Dirichlet setting is:

$$\begin{aligned} \bar{u} &= \arg \min_{\substack{(u_n)_{-2N \leq n \leq 2N} \\ u_{2N+1} = u_{2N} = 1 \\ u_{-2N-1} = u_{-2N} = 0}} \frac{h}{2\sqrt{2}\lambda} \|u - g\|_2^2 + \text{tv}_i^h(u) := E^h(u) \\ \text{with } \text{tv}_i^h(u) &= \frac{1}{\sqrt{2}} \sum_{n=-2N}^{2N} \sqrt{(u_{n+1} - u_n)^2 + (u_n - u_{n-1})^2} \end{aligned}$$

and where  $g_n = 0$  for  $n < 0$ ,  $g_n = 1$  for  $n > 0$  and  $g_0 = 1/2$ . The limit continuous energy is  $\bar{E} = \bar{E}_D = 1$ . In the following we build an admissible  $u$  of a particular form to

establish an upper bound estimate of the type

$$\bar{E}^h \leq E^h(u) \leq \bar{E} + ch^\theta,$$

for some  $0 < \theta < 1$ .

### 2.2.1 General construction

The idea is to take a function  $u$  such that  $u - g$  has a compact support of vanishing size but containing a number of points going to infinity. This is achieved by taking  $u_n$ , for  $-N \leq n \leq N$ , of the form (remember that  $N = \frac{1}{h}$ ):

$$u_n = f\left(\frac{n}{N_\alpha}\right) \text{ with } N_\alpha = \lceil h^{-\alpha} \rceil \text{ and } 0 < \alpha < 1$$

where  $f$  is some continuous function increasing from  $f(x) = 0$  for  $x \leq -1$  to  $f(x) = 1$  for  $x \geq 1$ . We also suppose in all what follows that  $f$  satisfies  $f(-x) = 1 - f(x)$  for any  $x \in \mathbb{R}$  to fulfill the conclusions of Proposition 1.

As  $u = g$  is constant for  $|n| \geq N_\alpha$ , one only has to consider what is happening in the transition phase, that is for  $|n| < N_\alpha$  for the  $L^2$  terms, and for  $|n| \leq N_\alpha$  for the tv terms. To understand what is at stake, let us first try with the piecewise affine function

$$f(x) = \begin{cases} 0 & \text{if } x < -1 \\ \frac{x+1}{2} & \text{if } -1 \leq x \leq 1 \\ 1 & \text{if } x > 1. \end{cases}$$

First compute the fidelity term:

$$\begin{aligned} \frac{h}{2} \|u - g\|_2^2 &= h \sum_1^{N_\alpha-1} \left(f\left(\frac{n}{N_\alpha}\right) - 1\right)^2 \\ &= \frac{h}{4N_\alpha^2} \sum_1^{N_\alpha-1} n^2 \\ &= \frac{hN_\alpha}{12} - \frac{h}{8} + \frac{h}{24N_\alpha} \end{aligned}$$

and then the tv term:

$$\begin{aligned} \text{tv}_i^h(u) &= \frac{1}{\sqrt{2}} \sum_{-N_\alpha+1}^{N_\alpha-1} \sqrt{\left(\frac{n+1}{2N_\alpha} - \frac{n}{2N_\alpha}\right)^2 + \left(\frac{n}{2N_\alpha} - \frac{n-1}{2N_\alpha}\right)^2} \\ &\quad + \frac{1}{\sqrt{2}} \left| 1 - \frac{1}{2} \left( \frac{N_\alpha-1}{N_\alpha} + 1 \right) \right| + \frac{1}{\sqrt{2}} \left| \frac{1}{2} \left( \frac{-N_\alpha+1}{N_\alpha} + 1 \right) \right| \\ &= \frac{1}{\sqrt{2}} \left( (2N_\alpha - 1) \times \sqrt{2 \times \frac{1}{4N_\alpha^2} + \frac{1}{N_\alpha}} \right) \\ &= 1 + \frac{\sqrt{2} - 1}{2N_\alpha}. \end{aligned}$$



Note that, the limit energy appears in the above expression as  $1 = \bar{E}$ . This finally leads to

$$\begin{aligned} E^h(u) - \bar{E} &= \frac{\sqrt{2}-1}{2N_\alpha} + \frac{hN_\alpha}{12\sqrt{2}\lambda} - \frac{h}{8\sqrt{2}\lambda} + \frac{h}{24\sqrt{2}\lambda N_\alpha} \\ &\leq \frac{\sqrt{2}-1}{2}h^\alpha + \frac{h(h^{-\alpha}+1)}{12\sqrt{2}\lambda} - \frac{h}{8\sqrt{2}\lambda} + \frac{h^{\alpha+1}}{24\sqrt{2}\lambda} \\ &\leq \frac{\sqrt{2}-1}{2}h^\alpha + \frac{h^{1-\alpha}}{12\sqrt{2}\lambda} + \frac{h^{\alpha+1}}{24\sqrt{2}\lambda}. \end{aligned}$$

The optimal choice of  $\alpha$  is then to make the two dominant terms in  $h^\alpha$  and  $h^{1-\alpha}$  of the same order, hence  $\alpha = 1/2$ . We conclude that, for any  $c > \frac{\sqrt{2}-1}{2} + \frac{1}{12\sqrt{2}\lambda}$ , one has for  $h$  small enough

$$E^h(u) - \bar{E} \leq c\sqrt{h}.$$

In the [next section](#), we show that a cubic function  $f$ , realising a smoother transition, leads to a better result: there exist constants  $c > 0$  and  $\underline{h} > 0$  depending only on  $\lambda$  such that:

$$\forall h \leq \underline{h}, E^h(u) - \bar{E} \leq ch^{2/3}. \quad (11)$$

This is optimal, as we then prove in [Section 2.3](#) a lower bound of the same order.

### 2.2.2 Analysis for an appropriate function $f$

In fact for any function regular enough ( $C^1$ )  $f$ , when  $h \rightarrow 0$  we have:  $u^h$  converges to  $g$  in  $L^2$  so  $h\|u^h - g^h\|_2^2 \rightarrow 0$ , and  $\text{tv}_i^h(u) \rightarrow \text{tv}(g) = 1$ . So  $E(u) \rightarrow \bar{E}$ . [So the difficulty is to find a particular  \$f\$  which reaches the optimal rate. We show in this section that it is obtained by a \(specific\) function, regular enough as it reaches the values 0 and 1.](#)

Our first remark in this section is that the approximation of the  $L^2$  term of the energy cannot really be improved, so we quickly switch to a detailed analysis of the total variation term. Then we split this total variation term into a term which accounts for the variation itself (up to a high order error term) and a perturbation, denoted  $d_n$  below. Eventually, we estimate the excess of energy coming from this perturbation, and we show that for a specific choice of  $f$ , it is of order  $O(1/N_\alpha^2)$  (Lemma 2) where as before  $2N_\alpha$  is the number of coefficients forming the ‘‘ramp’’ from 0 to 1 in the discrete approximation. Choosing then as before the best  $\alpha$  in the global rate yields (11). The adaption to the case of Neumann boundary conditions is explained in the next section. We now give the computational details.

As mentioned, the  $L^2$  term is easy to estimate:

$$\begin{aligned} \frac{h}{2} \|u - g\|_2^2 &= h \sum_{n=1}^{N_\alpha-1} \left(f\left(\frac{n}{N_\alpha}\right) - 1\right)^2 \\ &= h N_\alpha \frac{1}{N_\alpha} \sum_1^{N_\alpha-1} \left(f\left(\frac{n}{N_\alpha}\right) - 1\right)^2 \\ &\sim h^{1-\alpha} \int_0^1 (f - 1)^2 \text{ when } N_\alpha \rightarrow \infty \end{aligned}$$

hence for any  $c_1 > \frac{1}{\sqrt{2\lambda}} \int_0^1 (f - 1)^2$ , we have for  $h$  small enough:

$$\frac{h}{2\sqrt{2\lambda}} \|u - g\|_2^2 \leq c_1 h^{1-\alpha}. \quad (12)$$

The total variation term is trickier. We have:

$$\begin{aligned} \text{tv}_i^h(u) &= \frac{1}{\sqrt{2}} \sum_{n=-N_\alpha+1}^{N_\alpha-1} \sqrt{\left(f\left(\frac{n+1}{N_\alpha}\right) - f\left(\frac{n}{N_\alpha}\right)\right)^2 + \left(f\left(\frac{n}{N_\alpha}\right) - f\left(\frac{n-1}{N_\alpha}\right)\right)^2} \\ &\quad + \frac{1}{\sqrt{2}} \left|1 - f\left(\frac{N_\alpha-1}{N_\alpha}\right)\right| + \frac{1}{\sqrt{2}} \left|f\left(\frac{-N_\alpha+1}{N_\alpha}\right)\right|. \end{aligned}$$

The boundary terms simplify to:

$$\frac{1}{\sqrt{2}} \left|1 - f\left(\frac{N_\alpha-1}{N_\alpha}\right)\right| + \frac{1}{\sqrt{2}} \left|f\left(\frac{-N_\alpha+1}{N_\alpha}\right)\right| = \sqrt{2} \left(1 - f\left(1 - \frac{1}{N_\alpha}\right)\right).$$

For the middle terms, we use the following lemma with  $u_n = f\left(\frac{n}{N_\alpha}\right)$ :

**Lemma 1.** *If  $(u_n)$  is an increasing sequence, then for any  $n$ :*

$$\frac{1}{\sqrt{2}} \sqrt{(u_{n+1} - u_n)^2 + (u_n - u_{n-1})^2} \leq \frac{1}{2} (u_{n+1} - u_{n-1}) + d_n$$

with

$$d_n = \frac{1}{4} (u_{n+1} - u_{n-1}) (2u_n - u_{n+1} - u_{n-1}) \left( \frac{1}{u_{n+1} - u_n} - \frac{1}{u_n - u_{n-1}} \right) \quad (13)$$

$$= \frac{(u_{n+1} - u_{n-1}) (2u_n - u_{n+1} - u_{n-1})^2}{4(u_{n+1} - u_n)(u_n - u_{n-1})}. \quad (14)$$

*Proof.* Denote  $A = \sqrt{(u_{n+1} - u_n)^2 + (u_n - u_{n-1})^2}$  the quantity we want to estimate. Using  $\sqrt{x+h} \leq \sqrt{x} + \frac{1}{2\sqrt{x}}h$  we get:

$$\begin{aligned} A &= \sqrt{2(u_{n+1} - u_n)^2 + (u_n - u_{n-1})^2 - (u_{n+1} - u_n)^2} \\ &= \sqrt{2(u_{n+1} - u_n)^2 + (u_{n+1} - u_{n-1})(2u_n - u_{n+1} - u_{n-1})} \\ &\leq \sqrt{2}(u_{n+1} - u_n) + \frac{1}{2\sqrt{2}(u_{n+1} - u_n)} (u_{n+1} - u_{n-1})(2u_n - u_{n+1} - u_{n-1}). \end{aligned}$$

And similarly

$$\begin{aligned} A &= \sqrt{2(u_n - u_{n-1})^2 + (u_{n+1} - u_n)^2 - (u_n - u_{n-1})^2} \\ &\leq \sqrt{2}(u_n - u_{n-1}) - \frac{1}{2\sqrt{2}(u_n - u_{n-1})}(u_{n+1} - u_n)(2u_n - u_{n+1} - u_{n-1}). \end{aligned}$$

The result is obtained as the average of these two estimates.  $\square$

The term in  $\frac{1}{2}(u_{n+1} - u_{n-1}) = \frac{1}{2}(f(\frac{n+1}{N_\alpha}) - f(\frac{n-1}{N_\alpha}))$  is responsible for the convergence towards 1 as

$$\begin{aligned} \sum_{-N_\alpha+1}^{N_\alpha-1} \frac{1}{2} \left( f\left(\frac{n+1}{N_\alpha}\right) - f\left(\frac{n-1}{N_\alpha}\right) \right) &= \frac{1}{2} \left( f(1) + f\left(1 - \frac{1}{N_\alpha}\right) - f(-1) - f\left(-1 + \frac{1}{N_\alpha}\right) \right) \\ &= f\left(1 - \frac{1}{N_\alpha}\right). \end{aligned}$$

For the term in  $d_n$  note that the symmetry of  $f$  gives  $d_0 = 0$  and  $d_{-n} = d_n$  so that the sum is reduced to  $n \in [1, N_\alpha - 1]$  and we get the expression:

$$\text{tv}_i^h(u) \leq 1 + (\sqrt{2} - 1)(1 - f(1 - \frac{1}{N_\alpha})) + \sum_1^{N_\alpha-1} d_n. \quad (15)$$

Next we pursue our analysis for a particular function  $f$  given by

$$f(t) = \begin{cases} 0 & \text{if } t \leq -1 \\ \frac{1}{2}(1+t)^3 & \text{if } -1 \leq t \leq 0 \\ 1 - \frac{1}{2}(1-t)^3 & \text{if } 0 \leq t \leq 1 \\ 1 & \text{if } t \geq 1. \end{cases} \quad (16)$$

Then the term  $1 - f(1 - \frac{1}{N_\alpha})$  equals  $\frac{1}{N_\alpha^3}$  while **we can show:**

**Lemma 2.** *For the choice of  $f$  given by (16), one has*

$$\sum_1^{N_\alpha-1} d_n \leq \frac{6}{N_\alpha^2}.$$

*Proof.* Let us denote

$$\begin{aligned} \Delta_+ &:= f\left(\frac{n+1}{N_\alpha}\right) - f\left(\frac{n}{N_\alpha}\right) = \frac{1}{2} \left( \left(1 - \frac{n}{N_\alpha}\right)^3 - \left(1 - \frac{n+1}{N_\alpha}\right)^3 \right) \\ &= \frac{1}{2} \left( 3\left(1 - \frac{n}{N_\alpha}\right)^2 \frac{1}{N_\alpha} - 3\left(1 - \frac{n}{N_\alpha}\right) \frac{1}{N_\alpha^2} + \frac{1}{N_\alpha^3} \right) \\ &= \frac{3}{2N_\alpha} \left( \left(1 - \frac{n}{N_\alpha}\right)^2 - \left(1 - \frac{n}{N_\alpha}\right) \frac{1}{N_\alpha} + \frac{1}{3N_\alpha^2} \right). \end{aligned}$$

Similarly (that is, taking  $n \leftarrow n - 1$ ),

$$\Delta_- := f\left(\frac{n}{N_\alpha}\right) - f\left(\frac{n-1}{N_\alpha}\right) = \frac{3}{2N_\alpha} \left( \left(1 - \frac{n}{N_\alpha}\right)^2 + \left(1 - \frac{n}{N_\alpha}\right) \frac{1}{N_\alpha} + \frac{1}{3N_\alpha^2} \right)$$

so that

$$\begin{aligned}\Delta_+ + \Delta_- &= f\left(\frac{n+1}{N_\alpha}\right) - f\left(\frac{n-1}{N_\alpha}\right) = \frac{3}{N_\alpha} \left( \frac{1}{3N_\alpha^2} + \left(1 - \frac{n}{N_\alpha}\right)^2 \right) \\ \Delta_- - \Delta_+ &= 2f\left(\frac{n}{N_\alpha}\right) - f\left(\frac{n+1}{N_\alpha}\right) - f\left(\frac{n-1}{N_\alpha}\right) = \frac{3}{N_\alpha^2} \left(1 - \frac{n}{N_\alpha}\right)\end{aligned}$$

and

$$\begin{aligned}\Delta_+ \times \Delta_- &= \left(f\left(\frac{n+1}{N_\alpha}\right) - f\left(\frac{n}{N_\alpha}\right)\right) \left(f\left(\frac{n}{N_\alpha}\right) - f\left(\frac{n-1}{N_\alpha}\right)\right) \\ &= \frac{9}{4N_\alpha^2} \left( \left(1 - \frac{n}{N_\alpha}\right)^2 + \frac{1}{N_\alpha^3} - \left(1 - \frac{n}{N_\alpha}\right) \frac{1}{N} \right) \\ &\quad \times \left( \left(1 - \frac{n}{N_\alpha}\right)^2 + \frac{1}{N_\alpha^3} + \left(1 - \frac{n}{N_\alpha}\right) \frac{1}{N} \right) \\ &= \frac{9}{4N_\alpha^2} \left( \left( \left(1 - \frac{n}{N_\alpha}\right)^2 + \frac{1}{3N_\alpha^2} \right)^2 - \left( \left(1 - \frac{n}{N_\alpha}\right) \frac{1}{N_\alpha} \right)^2 \right).\end{aligned}$$

We can now estimate  $d_n$  thanks to expression (14):

$$\begin{aligned}d_n &= \frac{1}{4} \times \frac{\left( \frac{3}{N_\alpha} \left( \frac{1}{3N_\alpha^2} + \left(1 - \frac{n}{N_\alpha}\right)^2 \right) \right) \times \left( \frac{3}{N_\alpha^2} \left(1 - \frac{n}{N_\alpha}\right) \right)^2}{\frac{9}{4N_\alpha^2} \left( \left( \left(1 - \frac{n}{N_\alpha}\right)^2 + \frac{1}{3N_\alpha^2} \right)^2 - \left( \left(1 - \frac{n}{N_\alpha}\right) \frac{1}{N_\alpha} \right)^2 \right)} \\ &= \frac{3}{N_\alpha^3} \left(1 - \frac{n}{N_\alpha}\right)^2 \times \frac{\frac{1}{3N_\alpha^2} + \left(1 - \frac{n}{N_\alpha}\right)^2}{\left( \left(1 - \frac{n}{N_\alpha}\right)^2 + \frac{1}{3N_\alpha^2} \right)^2 - \left(1 - \frac{n}{N_\alpha}\right)^2 \frac{1}{N_\alpha^2}}.\end{aligned}$$

Then as  $n \leq N_\alpha - 1$  we can use

$$\left( \left(1 - \frac{n}{N_\alpha}\right)^2 + \frac{1}{3N_\alpha^2} \right)^2 - \left(1 - \frac{n}{N_\alpha}\right)^2 \frac{1}{N_\alpha^2} \geq \left(1 - \frac{n}{N_\alpha}\right)^4 - \frac{1}{3N_\alpha^2} \left(1 - \frac{n}{N_\alpha}\right)^2 > 0$$

and make the **change of variable**  $n \leftarrow N - n$  to get:

$$\sum_{n=1}^{N_\alpha-1} d_n \leq \frac{3}{N_\alpha^3} \sum_{n=1}^{N_\alpha-1} n^2 \frac{\frac{1}{3N_\alpha^2} + n^2}{n^4 - \frac{1}{3N_\alpha^2} n^2} \leq \frac{3}{N_\alpha^3} \sum_{n=1}^{N_\alpha-1} 2 \leq \frac{6}{N_\alpha^2}$$

because  $\frac{1}{3N_\alpha^2} + n^2 \leq 2(n^2 - \frac{1}{3N_\alpha^2})$ . □

This concludes our proof of the upper bound inequality in Theorem 1. Indeed, when combining the tv estimate (15) with the  $L^2$  estimate (12), we finally are able to state the following: for any  $c_1 > \frac{1}{\sqrt{2\lambda}} \int_0^1 (f-1)^2 = \frac{1}{28\sqrt{2\lambda}}$  and  $c_2 > 6$ , there exists  $\underline{h} > 0$  such that

$$\forall h \leq \underline{h}, E^h(u) \leq \bar{E} + c_1 h^{1-\alpha} + c_2 h^{2\alpha}. \quad (17)$$

Taking  $\alpha = 1/3$  then proves (11). More precisely, given  $c_1, c_2$  and  $h > 0$ , the best  $\alpha$  in (17) must satisfy  $-c_1 h^{1-\alpha} \log h + 2c_2 h^{2\alpha} \log h = 0$  which leads to  $\alpha = \frac{1}{3} - \frac{\log(2c_2/c_1)}{3 \log h}$  and gives the upper bound  $E^h(u) \leq \bar{E} + ch^{2/3}$  with  $c = (2^{1/3} + 2^{-2/3}) c_1^{2/3} c_2^{1/3}$  (note that  $c$  varies in  $\lambda^{-2/3}$ ).

### 2.2.3 Upper bound for Neumann boundary conditions

In this section we adjust the admissible variable  $u$  from the previous section to explain why the upper bound result (11) remains valid for Neumann boundary conditions. In the following,  $c$  denotes a constant depending only on  $\lambda$  that can change from line to line.

Remember from Section 2.1.2 that with the Neumann boundary conditions, the limit continuous value of the energy is changed to  $\bar{E} = \bar{E}_N = 1 - \sqrt{2}\lambda$  when  $\lambda \leq \lambda^* = \frac{\sqrt{2}}{4}$ . Because of the form of this continuous solution, it is natural to consider, for  $u$  the cubic transition in the Dirichlet setting of the previous section, the variable  $v$  given by

$$\forall -N \leq n \leq N, v_n = \frac{1}{2} + \mu(u_n - \frac{1}{2}).$$

Here  $\mu \in (0, 1)$  is a shrinking parameter that we adjust so that  $v_N = 1 - a_{opt} = 1 - \sqrt{2}\lambda$ : as  $u_N = 1$  this corresponds to taking  $\mu = 1 - 2\sqrt{2}\lambda$ .

We write  $v_n = f_\mu(\frac{n}{N_\alpha})$  for the function  $f_\mu = \frac{1}{2} + \mu(f - \frac{1}{2})$  which is such that  $f_\mu(x) = \frac{1+\mu}{2} = 1 - \sqrt{2}\lambda$  for  $x \geq 1$ . This leads to splitting the fidelity term into:

$$\frac{h}{2} \|v - g\|_2^2 = h \sum_{n=1}^{N_\alpha} (v_n - 1)^2 + h \sum_{n=N_\alpha+1}^N (v_n - 1)^2.$$

Then on one hand when  $N_\alpha \rightarrow \infty$ ,

$$h \sum_{n=1}^{N_\alpha} (v_n - 1)^2 \sim h^{1-\alpha} \int_0^1 (f_\mu - 1)^2 \quad \text{so} \quad h \sum_{n=1}^{N_\alpha} (v_n - 1)^2 \leq ch^{1-\alpha}$$

and on the other hand

$$h \sum_{n=N_\alpha+1}^N (v_n - 1)^2 = h(N - N_\alpha) \times 2\lambda^2 \leq 2\lambda^2.$$

For the tv term, we have

$$\begin{aligned} \text{tv}_i^h(v) &= \mu \text{tv}_i^h(u) = (1 - 2\sqrt{2}\lambda) \text{tv}_i^h(u) \\ &\leq (1 - 2\sqrt{2}\lambda)(1 + ch^{2\alpha}) \\ &\leq 1 - 2\sqrt{2}\lambda + ch^{2\alpha} \end{aligned}$$

so finally

$$\begin{aligned} E^h(v) &= \frac{h}{2\sqrt{2}\lambda} \|v - g\|_2^2 + \text{tv}_i^h(v) \\ &\leq \sqrt{2}\lambda + ch^{1-\alpha} + 1 - 2\sqrt{2}\lambda + ch^{2\alpha} \\ &\leq \bar{E} + ch^{2/3} \end{aligned}$$

when taking  $\alpha = 1/3$ .

### 2.3 Lower bound estimate

In this section we now prove the lower bound of the point 2 of Theorem 1, that is:  $\exists \underline{h}, c > 0$  such that

$$\forall h \leq \underline{h}, ch^{2/3} \leq \bar{E}^h - \bar{E}.$$

Symmetrically to what we did in the previous section, we will obtain this bound by proposing an admissible solution, but for the dual *maximization* problem. **The proof is quite long and split into many intermediate steps. We first introduce the discrete dual problem.**

#### 2.3.1 Dual problem

**Lemma 3.** *Problem (9) admits the dual representation:*

$$\bar{E}^h = \max \left\{ \frac{1}{2} + \frac{1}{\sqrt{2}} p_1 - \frac{\lambda}{\sqrt{2}h} \sum_{n=1}^{N-1} (p_{-n+1} - p_{-n} + p_n - p_{n+1})^2 : \right. \\ \left. p_n^2 + p_{-n}^2 \leq 1 \ \forall n = 1, \dots, N, \quad p_0 = \frac{\sqrt{2}}{2} \right\}. \quad (18)$$

*Proof.* Using:

$$\sqrt{(u_{n+1} - u_n)^2 + (u_n - u_{n-1})^2} = \max_{p_n^2 + q_n^2 \leq 1} q_n(u_{n+1} - u_n) + p_n(u_n - u_{n-1}) \quad (19)$$

and standard convex duality results (see for instance [30, Cor. 31.2.1]), we first obtain the following dual problem for (9):

$$\begin{aligned} & \max_{\substack{p_n^2 + q_n^2 \leq 1 \\ -N \leq n \leq N}} \min_{u \in \mathbb{R}^{2N+1}} \sum_{n=-N}^N \frac{h}{2\lambda} (u_n - g_n)^2 + q_n(u_{n+1} - u_n) + p_n(u_n - u_{n-1}) \\ &= \max_{\substack{p_n^2 + q_n^2 \leq 1 \\ -N \leq n \leq N}} \min_{u \in \mathbb{R}^{2N+1}} \frac{1}{\sqrt{2}} \left\{ \sum_{n=-N}^N \frac{h}{2\lambda} (u_n - g_n)^2 + \sum_{n=-N+1}^{N-1} u_n (q_{n-1} - q_n + p_n - p_{n+1}) \right. \\ & \quad \left. + u_N (q_{N-1} - q_N + p_N) + u_{-N} (-q_{-N} + p_{-N} - p_{-N+1}) \right. \\ & \quad \left. + u_{N+1} q_N - u_{-N-1} p_{-N} \right\}. \end{aligned}$$

From this point on, we focus exclusively on Dirichlet boundary conditions, that is  $u_N = u_{N+1} = 1$  ;  $u_{-N} = u_{-N-1} = 0$ . See Section 2.3.5 for Neumann boundary conditions.

For  $|n| < N$ , we find that  $u_n = g_n - \frac{\lambda}{h}(q_{n-1} - q_n + p_n - p_{n+1})$ , and the value of the dual problem is consequently (after simplification using the value of  $g_n$ ):

$$\max_{\substack{p_n^2 + q_n^2 \leq 1 \\ -N \leq n \leq N}} \frac{1}{\sqrt{2}} \left\{ \frac{1}{2}(q_{-1} + q_0 + p_0 + p_1) - \frac{\lambda}{2h} \sum_{n=-N+1}^{N-1} (q_{n-1} - q_n + p_n - p_{n+1})^2 \right\}.$$

Now we make two more simplifications before turning to an evaluation of the convergence rate of this quantity. First, one easily checks that the objective is concave and invariant by the change  $(q_n, p_n) \rightarrow (p_{-n}, q_{-n})$ : as a consequence, one can find a solution satisfying  $q_n = p_{-n}$  for all  $n$ .

Second, **the optimality conditions at the saddle point for the above min-max problem guarantee that for any optimal  $u$  and  $(q, p)$ ,  $(q_n, p_n)$  should reach the maximum in (19).** For  $n = 0$  we find, thanks to Proposition 2, that  $\sqrt{2}|u_1 - u_0| = (q_0 + p_0)(u_1 - u_0)$  so that  $q_0 = p_0 = \frac{\sqrt{2}}{2}$ . Simplifying the term  $(q_{n-1} - q_n + p_n - p_{n+1})^2$  which is invariant by  $n \rightarrow -n$  and vanishes at  $n = 0$ , we finally get (18), which shows the lemma.  $\square$

**In the next step, we introduce a change of variables which allow to identify a continuum limit of the dual energy, up to a renormalization.**

### 2.3.2 A change of variables

We are interested in the evaluation of the convergence rate of the value of the problem (18) towards its continuous limit  $\bar{E} = \bar{E}_D = 1$ . First let us notice that taking  $p_n \equiv \sqrt{2}/2$  gives  $\bar{E}^h \geq \bar{E}$ . Consequently, we expect the optimal value of  $p$  to be close to  $\sqrt{2}/2$  for  $N$  large. Together with the symmetry regarding  $n \rightarrow -n$  of the objective, this leads us to proposing the following change of variable: for  $0 \leq n \leq N$

$$s_n = \frac{1}{\sqrt{2}}(p_n + p_{-n}) - 1 ; r_n = \frac{1}{\sqrt{2}}(p_n - p_{-n})$$

for which we calculate

$$p_{-n+1} - p_{-n} + p_n - p_{n+1} = \frac{1}{\sqrt{2}}(s_{n-1} - s_{n+1} + 2r_n - r_{n-1} - r_{n+1})$$

$$p_n^2 + p_{-n}^2 \leq 1 \iff s_n^2 + 2s_n + r_n^2 \leq 0$$

and (18) becomes:

$$\bar{E}^h - \bar{E} = \max_{\substack{(s_n, r_n)_{0 \leq n \leq N} \\ s_0 = r_0 = 0 \\ s_n^2 + 2s_n + r_n^2 \leq 0}} \frac{1}{\sqrt{2}} \left\{ s_1 + r_1 - \frac{\lambda}{2h} \sum_{n=1}^{N-1} (s_{n-1} - s_{n+1} + 2r_n - r_{n-1} - r_{n+1})^2 \right\}.$$

We would like to show that  $\bar{E}^h - \bar{E} \geq cN^{-\alpha}$  for some exponent  $0 < \alpha < 1$ . If we introduce  $\tau = 1/N^\beta$  for some  $\beta \in (0, \alpha)$  and  $\sigma_n = N^\alpha s_n$ ,  $\rho_n = N^{\alpha-\beta} r_n$ , then we can force the appearance of first and second discrete derivatives for  $\sigma$  and  $\rho$  as

$$(\bar{E}^h - \bar{E})N^\alpha = \max_{\substack{(\sigma_n, \rho_n) \\ 0 \leq n \leq N}} \frac{1}{\sqrt{2}} \left\{ \sigma_1 + \frac{\rho_1}{\tau} - \frac{\lambda}{2} N^{1-\alpha-\beta} \tau \sum_{n=1}^{N-1} \left( \frac{\sigma_{n-1} - \sigma_{n+1}}{\tau} + \frac{2\rho_n - \rho_{n-1} - \rho_{n+1}}{\tau^2} \right)^2 \right\} \quad (20)$$

along with the constraints  $\sigma_0 = \rho_0 = 0$  and  $N^{-\alpha} \sigma_n^2 + 2\sigma_n + N^{2\beta-\alpha} \rho_n^2 \leq 0$ .

If  $1-\alpha-\beta = 0$ , we find that as  $N \rightarrow \infty$ , the limiting energy in the variational problem should be of the form,

$$\max \frac{1}{\sqrt{2}} \left\{ \rho'(0) - \frac{\lambda}{2} \int_0^\infty |2\sigma' + \rho''|^2 \right\}$$

for functions  $\sigma, \rho : [0, \infty) \rightarrow \mathbb{R}$  with  $\sigma(0) = \rho(0) = 0$ . The constraint, on the other hand, becomes

$$\begin{cases} \rho^2 = 0 & \text{if } 2\beta - \alpha > 0 \\ 2\sigma + \rho^2 \leq 0 & \text{if } 2\beta - \alpha = 0 \Leftrightarrow \beta = 1/3, \alpha = 2/3 \\ 2\sigma \leq 0 & \text{if } 2\beta - \alpha < 0. \end{cases}$$

In the first case, which is when  $\alpha < 2/3$ , we may expect that the discrete energy goes to zero, and we expect that  $\bar{E}^h - \bar{E} = o(N^{-\alpha})$  as  $N \rightarrow \infty$ . In the third case, the continuous problem has value  $+\infty$  and we expect that  $N^\alpha(\bar{E}^h - \bar{E}) \rightarrow \infty$  for  $\alpha > 2/3$ . We would like to show that in the second case, that is  $\alpha = 2/3$ , the limiting problem has a positive value  $c$  so that  $\bar{E}^h - \bar{E} \geq cN^{-2/3}$  for sufficiently large  $N$ . Consequently we study the problem

$$\max_{(\sigma, \rho) \in S} \frac{1}{\sqrt{2}} \left\{ \rho'(0) - \frac{\lambda}{2} \int_0^\infty (2\sigma' + \rho'')^2 \right\} =: D(\sigma, \rho) \quad (21)$$

where  $S$  is the set of couples of functions  $\sigma, \rho : [0, \infty) \rightarrow \mathbb{R}$  such that:  $\sigma(0) = \rho(0) = 0$ ,  $2\sigma + \rho^2 \leq 0$ ,  $\rho$  admits a right derivative at 0 and the distributional derivative  $2\sigma' + \rho''$  is in  $L^2(0, \infty)$ .

Our strategy is now the following: in Section 2.3.3 we prove that Problem (21) has a positive value and investigate the form of the solution  $(\sigma, \rho)$ . Then in Section 2.3.4 we explain how to discretize it in order to get the positivity, for  $h$  small enough, of the discrete problem:

$$(\bar{E}^h - \bar{E})h^{-2/3} = \max_{\substack{(\sigma_n, \rho_n)_{0 \leq n \leq N} \\ \sigma_0 = \rho_0 = 0 \\ N^{-2/3} \sigma_n^2 + 2\sigma_n + \rho_n^2 \leq 0}} D^h(\sigma, \rho) \quad (22)$$



where

$$D^h(\sigma, \rho) := \frac{1}{\sqrt{2}} \left\{ \sigma_1 + \frac{\rho_1}{\tau} - \frac{\lambda}{2} \tau \sum_{n=1}^{N-1} \left( \frac{\sigma_{n-1} - \sigma_{n+1}}{\tau} + \frac{2\rho_n - \rho_{n-1} - \rho_{n+1}}{\tau^2} \right)^2 \right\} \quad (23)$$

which is the expression in (20) for  $\alpha = 2/3$  and  $\beta = 1/3$ . Showing that (22) is positive for  $h$  small enough establishes the desired lower bound.

### 2.3.3 Study of the limit problem

First, the change of variable  $\hat{\sigma}(t) = \lambda^{-2/3} \sigma(t\lambda^{-1/3})$ ,  $\hat{\rho}(t) = \lambda^{-1/3} \rho(t\lambda^{-1/3})$  shows that (adding the parameter  $\lambda$  to the arguments of  $D$ )

$$\max D(\sigma, \rho, \lambda) = \lambda^{-2/3} \max D(\sigma, \rho, 1).$$

Consequently we suppose  $\lambda = 1$  in all of the following.

**A dual of the dual.** We build a solution of problem (21) by studying a dual problem, which we derive as follows. We write:

$$-\frac{1}{2} \int_0^\infty (2\sigma' + \rho'')^2 = \inf_{\psi} \int_0^\infty (2\sigma' + \rho'')\psi + \frac{1}{2} \int_0^\infty \psi^2$$

where the infimum is taken over the functions  $\psi \in \mathcal{C}_c^\infty([0, \infty))$ . (Note that if  $\sigma, \rho$  were regular enough one would have at the optimum  $\psi = -(2\sigma' + \rho'')$ .) Integrating by parts and using that  $\sigma(0) = \rho(0) = 0$  for any  $(\sigma, \rho) \in S$ , we obtain the dual problem

$$\frac{1}{\sqrt{2}} \inf_{\psi} \frac{1}{2} \int_0^\infty \psi^2 + \sup_{(\sigma, \rho) \in S} (1 - \psi(0))\rho'(0) + \int_0^\infty (\rho\psi'' - 2\sigma\psi').$$

First, taking for  $\rho$  a bounded smooth function with  $|\rho'(0)|$  as large as we want, we see that one must have  $\psi(0) = 1$ . Second, we relax the constraint  $(\sigma, \rho) \in S$  in the remaining integral into just  $2\sigma + \rho^2 \leq 0$  (we will show below that strong duality with problem (21) actually occurs) to get:

$$\frac{1}{\sqrt{2}} \inf_{\psi(0)=1} \frac{1}{2} \int_0^\infty |\psi|^2 + \int_0^\infty H(\psi', \psi'')$$

where the function  $H$  is defined for  $x, y \in \mathbb{R}$  by

$$H(x, y) = \sup_{2\sigma + \rho^2 \leq 0} -2\sigma x + \rho y = \begin{cases} +\infty & \text{if } x > 0 \text{ or } x = 0, y \neq 0, \text{ (via } \rho = 0, \sigma \rightarrow -\infty) \\ 0 & \text{if } (x, y) = (0, 0), \\ \frac{y^2}{4|x|} & \text{if } x < 0 \text{ (via } \rho = -y/2x, \sigma = -\rho^2/2). \end{cases}$$

Observe that necessarily  $\psi' \leq 0$ . Denoting  $\phi = \sqrt{-\psi'}$  gives  $\phi' = -\psi''/(2\sqrt{-\psi'})$  so that  $H(\psi', \psi'') = |\phi'|^2$ . Then, one has  $\psi(x) = 1 - \int_0^x \phi(t)^2 dt$ . In particular as  $\psi^2$  is integrable,

one must have  $\int_0^\infty \phi(t)^2 dt = 1$  and  $\psi(x) = \int_x^\infty \phi(t)^2 dt$ . Hence the dual problem can be rewritten (extending the search of  $\phi$  to  $H^1(0, \infty)$  by density)

$$\frac{1}{\sqrt{2}} \inf_{(\phi, \psi) \in S'} \left\{ \frac{1}{2} \int_0^\infty |\psi|^2 + \int_0^\infty |\phi'|^2 \right\} \quad (24)$$

where  $S' = \{(\phi, \psi) : \phi \in H^1(0, \infty), \|\phi\|_2^2 = 1 \text{ and } \psi(x) = \int_x^\infty \phi(t)^2 dt\}$ .

It turns out this problem has a positive value:

**Proposition 3.** *Problem (24) has a minimizer  $(\psi, \phi) \in W^{2,1}(0, \infty) \times H^1(0, \infty)$ .*

*Proof.* Consider a minimizing sequence  $(\phi_n, \psi_n)$ : as  $\phi_n$  is bounded in  $H^1(0, +\infty)$ , up to a subsequence it converges to some  $\phi$ , moreover the convergence is strong in  $L^2(0, T)$  for any  $T > 0$ , and  $\int_0^\infty \phi^2 \leq 1$ . We also assume that  $\psi_n$  converges, weakly in  $L^2(0, +\infty)$ , to some  $\psi$ . In addition,  $\psi_n(x) = 1 - \int_0^x \phi_n^2 \rightarrow 1 - \int_0^x \phi^2 =: \tilde{\psi}(x)$  for any  $x \geq 0$ , and one even has  $|\psi_n(x) - \tilde{\psi}(x)| = |\int_0^x (\phi_n - \phi)(\phi_n + \phi)| \leq 2\|\phi_n - \phi\|_{L^2(0, x)}$  hence the convergence is locally uniform. Consequently, it must be that  $\tilde{\psi} = \psi$ . As  $\int_0^\infty |\psi|^2 < +\infty$ , we deduce that  $\psi$  (which is nonincreasing) goes to 0 at  $\infty$ , hence  $\int_0^\infty \phi^2 = 1$ . It follows that  $(\psi, \phi)$  is a minimizer of (24).  $\square$

To recover the positive value of problem (21), we now need to show that strong duality holds. To do that we first prove some properties of the minimizer  $(\psi, \phi)$ .

**Proposition 4.** *The minimizer  $(\psi, \phi)$  of problem (24) satisfies:*

1.  $\psi, \phi \in C^\infty([0, \infty)) \cap L^2(0, \infty)$ .
2.  $\phi'(0) = 0$  and  $\phi'' = k\phi$  where  $k(t) = \int_0^t \psi - A$  with  $A = \|\phi'\|_2^2 + \|\psi\|_2^2$  satisfies  $k' = \psi$ .
3.  $\phi \geq 0$ ,  $\phi(0) > 0$ ,  $\phi$  is nonincreasing and tends to zero at infinity.

*Proof.* One has  $\psi' = -\phi^2 \in L^1(0, \infty)$  and  $\psi'' = -2\phi\phi' \in L^1(0, \infty)$  (hence  $\psi \in W^{2,1}(0, +\infty)$  and is at least  $C^1$ ). Moreover, if  $(\psi, \phi)$  is a minimizer, so is  $(\psi, |\phi|)$ . The **minimizer** of (24) being unique, one has  $\phi \geq 0$ .

From this solution  $(\psi, \phi)$ , let us form for  $\varepsilon \in \mathbb{R}$  and for a test function  $\eta$

$$\phi_\varepsilon = \frac{\phi + \varepsilon\eta}{\|\phi + \varepsilon\eta\|_2} ; \psi_\varepsilon(x) = \int_x^\infty \phi_\varepsilon^2.$$

Then  $(\phi_\varepsilon, \psi_\varepsilon)$  are admissible in the dual of the dual problem and one computes:

$$\begin{aligned} \phi_\varepsilon^2 &= \phi^2 + 2\varepsilon\eta\phi - 2\varepsilon\phi^2 \int_0^\infty \phi\eta + O(\varepsilon^2) \\ \psi_\varepsilon^2(x) &= \psi^2(x) + 4\varepsilon\psi(x) \int_x^\infty \phi\eta - 4\varepsilon\psi^2(x) \int_0^\infty \phi\eta + O(\varepsilon^2) \\ \phi'_\varepsilon &= \phi' + \varepsilon\eta' - \varepsilon\phi' \int_0^\infty \phi\eta + O(\varepsilon^2) \end{aligned}$$

so that, after noting that  $\int_0^\infty \psi(x) \int_x^\infty \phi \eta \, dx = \int_0^\infty \phi \eta \nu$  with  $\nu(t) = \int_0^t \psi$  one has

$$\begin{aligned} \int_0^\infty |\phi'_\varepsilon|^2 &= \int_0^\infty |\phi'|^2 - 2\varepsilon \int_0^\infty |\phi'|^2 \int_0^\infty \phi \eta + 2\varepsilon \int_0^\infty \phi' \eta' + O(\varepsilon^2) \\ \int_0^\infty |\psi_\varepsilon|^2 &= \int_0^\infty |\psi|^2 - 4\varepsilon \int_0^\infty |\psi|^2 \int_0^\infty \phi \eta + 4\varepsilon \int_0^\infty \phi \eta \nu + O(\varepsilon^2). \end{aligned}$$

Now the optimality of  $(\psi, \phi)$  in problem (24) leads to

$$\int_0^\infty \phi \eta \nu - \int_0^\infty |\psi|^2 \int_0^\infty \phi \eta + \int_0^\infty \phi' \eta' - \int_0^\infty |\phi'|^2 \int_0^\infty \phi \eta = 0. \quad (25)$$

First, as this relation holds for any  $\eta \in \mathcal{C}_c^\infty(0, \infty)$ , we have  $\phi'' = k\phi$  (with  $k = \nu - A$  where  $A = \|\psi\|^2 + \|\phi'\|^2$ ) in the weak sense. However this relation induces the regularity of  $\phi$  and  $\psi$  which are finally  $\mathcal{C}^\infty$ . What is more is that, re-evaluating the relation (25) with now  $\eta \in \mathcal{C}_c^\infty([0, \infty))$ , we also deduce that  $\phi'(0) = 0$ .

To finish with, one must have  $\phi(0) > 0$  as otherwise  $\phi$  would be zero everywhere as solution of  $\phi'' = k\phi$ ,  $\phi'(0) = \phi(0) = 0$ . **Additionally**, note that  $\phi'' = k\phi$  has the sign of  $k$  which is nonincreasing since  $k' = \psi \geq 0$ . Hence  $\phi''$  is first nonpositive (starting at  $\phi''(0) = -A\phi(0) \leq 0$ ) then possibly nonnegative. As a consequence,  $\phi'$  is first nonincreasing, and hence nonpositive since  $\phi'(0) = 0$ , then can become nondecreasing. But even in that case,  $\phi'$  has to remain nonpositive otherwise one has  $\phi'(t) \geq c > 0$  for  $t$  large enough so  $\phi(t) \geq ct + c'$  which contradicts the fact that  $\phi^2$  is integrable. This concludes the proof.  $\square$

In the following we show that strong duality holds between problems (21) and (24). To do so we divide our study in two cases: either  $\phi > 0$  on  $\mathbb{R}^+$  (the “positive” case), or  $\phi > 0$  on  $[0, a)$  and  $\phi = 0$  on  $[a, +\infty[$  for some  $a > 0$  (the “compact support” case). Note that numerical experiments seem to show we actually are in the “compact support” case, see Figure 4.

**Strong duality holds.** We now show that it is possible to build, from the minimizer  $(\psi, \phi)$  of (24), a pair  $(\sigma, \rho)$  which solves (21) with equality of the optimal value  $D(\sigma, \rho)$ .

In the “positive” case, recalling how the dual problem was obtained, one defines  $\sigma = -\rho^2/2$  and  $\rho = -\phi'/\phi$  and then checks that  $2\sigma + \rho^2 \leq 0$ ,  $\sigma(0) = \rho(0) = 0$ ,  $\rho'(0) = A$  and  $2\sigma' + \rho'' = -\psi$  so that

$$\frac{1}{\sqrt{2}} \left\{ \rho'(0) - \frac{1}{2} \int_0^\infty |2\sigma' + \rho''|^2 \right\} = \frac{1}{\sqrt{2}} \left\{ \int_0^\infty \phi'^2 + \frac{1}{2} \int_0^\infty \psi^2 \right\}$$

and strong duality holds.

In the “compact support” case, one still defines  $\rho = -\phi'/\phi$  and  $\sigma = -\rho^2/2$  on  $[0, a)$ . Then one has to decide what to do on  $[a, +\infty)$ . First, for  $t < a$ :

$$\rho(t) = -\frac{\phi'(t)}{\phi(t)} = \frac{1}{\phi(t)} \int_t^a \phi''(s) ds = \int_t^a \frac{\phi(s)}{\phi(t)} k(s) ds.$$

Since  $\phi$  is nonincreasing,  $\frac{\phi(s)}{\phi(t)} \leq 1$  in the above integral and we deduce

$$|\rho(t)| \leq \int_t^a k(s) ds \rightarrow 0 \text{ when } t \rightarrow a$$

and also  $\sigma(t) = -\rho(t)^2/2 \rightarrow 0$  when  $t \rightarrow a$ . The first guess would then consist in extending  $\sigma$  and  $\rho$  by continuity one could set  $\sigma = \rho = 0$  on  $[a, +\infty)$ .

This would actually lead to a discontinuous  $\rho'$ . Indeed  $\rho$  is differentiable in  $a^+$  with  $\rho'(a^+) = 0$ ; furthermore  $\rho'(t) = \rho^2(t) - k(t)$  for  $t \in (0, a)$ , and  $\rho(t) \rightarrow 0$ ,  $k(t) \rightarrow -k(a)$  when  $t \rightarrow a$ . Hence  $\rho$  is differentiable in  $a^-$  with  $\rho'(a^-) = -k(a)$ . Anyway  $\rho'$  is discontinuous at  $a$  (and  $\mathcal{C}^\infty$  elsewhere), so  $\rho''$  has a Dirac mass at  $a$ . Whereas  $\sigma = -\rho^2/2$  on  $(0, a)$  as well as on  $[a, +\infty)$  is continuous and has derivative  $\sigma' = -\rho'\rho$  also continuous at  $a$  as  $\rho(a) = 0$ . Finally  $2\sigma' + \rho'' \notin L^2$ .

This is why one should not take  $\sigma = 0$  but rather  $\sigma = -k(a)/2$  on  $(a, +\infty)$  and still  $\rho = 0$ . This is an admissible choice since  $k(a) > 0$ : indeed this comes again from the fact that  $\phi'' = k\phi$ : if  $k(a) < 0$  then, as  $\phi > 0$  on  $[0, a)$  and  $k$  is nondecreasing, one would obtain that  $\phi'$  is (strictly) decreasing on  $[0, a)$ . Starting with  $\phi'(0) = 0$  we obtain that  $\phi'(a) < 0$ , but  $\phi = 0$  on  $[a, +\infty)$  so one should have  $\phi'(a) = 0$ . With this setting,  $2\sigma + \rho'$  is continuous at  $a$  so that  $2\sigma' + \rho'' \in L^2$  (the two Dirac masses compensate each other in the derivatives). And, just as before,  $2\sigma' + \rho'' = -\psi$  so strong duality holds.

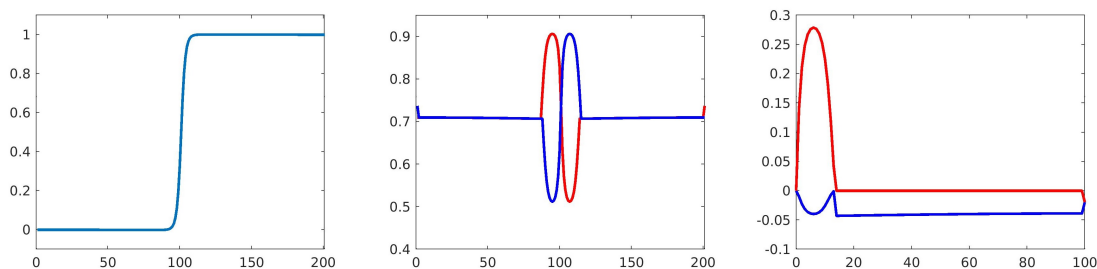


Figure 4: Primal solution  $u$  (left), dual solutions  $p$  and  $q$  (center), corresponding  $\sigma$  (blue) and  $\rho$  (red) (right) in the Dirichlet setting with  $N = 100$ .

### 2.3.4 Back to the discrete problem

Now, the previous analysis performed in the continuous setting will allow to build an approximate discrete solution  $(\sigma, \rho)$  with a value of  $D^h(\sigma, \rho)$  close to  $D(\sigma, \rho) > 0$ , showing that (22) is positive.

We recall the definitions (21) and (23) of  $D(\sigma, \rho)$  and  $D^h(\sigma, \rho)$ . In the following,

$\lambda = 1$ , and  $\tau = N^{-1/3}$ . The constraints on  $\sigma, \rho : \mathbb{R}^+ \rightarrow \mathbb{R}$  and  $\boldsymbol{\sigma}, \boldsymbol{\rho} \in \mathbb{R}^{N+1}$  are

$$\begin{aligned} \sigma(0) = \rho(0) = 0 \text{ and } 2\sigma + \rho^2 \leq 0 \text{ on } \mathbb{R}^+ \\ \boldsymbol{\sigma}_0 = \boldsymbol{\rho}_0 = 0 \text{ and } \forall n \geq 1, N^{-2/3}\boldsymbol{\sigma}_n^2 + 2\boldsymbol{\sigma}_n + \boldsymbol{\rho}_n^2 \leq 0. \end{aligned}$$

Given an admissible  $(\sigma, \rho)$  of the continuous problem with  $D(\sigma, \rho) > 0$  we chose the following discretization: set  $\boldsymbol{\sigma}_0 = 0$  and  $\forall n \geq 1$ ,  $\boldsymbol{\sigma}_n = \sigma(\tau n) - \tau$  and  $\boldsymbol{\rho}_n = \rho(\tau n)$  for all  $n$ . Then, provided  $\sigma$  is bounded,  $(\boldsymbol{\sigma}, \boldsymbol{\rho})$  is indeed admissible in the discrete problem as  $\boldsymbol{\sigma}_0 = \boldsymbol{\rho}_0 = 0$  and

$$\begin{aligned} N^{-2/3}\boldsymbol{\sigma}_n^2 + 2\boldsymbol{\sigma}_n + \boldsymbol{\rho}_n^2 &= N^{-2/3}(\sigma(\tau n) - \tau)^2 - 2\tau + 2\sigma(\tau n) + \rho(\tau n)^2 \\ &\leq N^{-2/3}(\sigma(\tau n) - \tau)^2 - 2N^{-1/3} \end{aligned}$$

with this quantity being nonpositive as soon as  $|\sigma(\tau n) - N^{-1/3}| \leq \sqrt{2}N^{1/6}$  which is true for  $N$  sufficiently large when  $\sigma$  is bounded.

Therefore we just need to check that with this discretization  $D^h(\boldsymbol{\sigma}, \boldsymbol{\rho})$  converges to  $D(\sigma, \rho)$  when  $N \rightarrow \infty$  as expected in the first place. First note that  $\boldsymbol{\sigma}_1 = \sigma(\tau) - \tau \rightarrow \sigma(0) = 0$  (as long as  $\sigma$  is continuous) and that  $\frac{1}{\tau}\boldsymbol{\rho}_1 = \frac{\rho(\tau) - \rho(0)}{\tau} \rightarrow \rho'(0)$ . As a result we focus next on the convergence of the Riemann sum towards the desired integral.

To simplify, inside the Riemann sum in  $D^h(\boldsymbol{\sigma}, \boldsymbol{\rho})$ , we replace  $\boldsymbol{\sigma}$  as defined above with  $\boldsymbol{\sigma}_n = \sigma(\tau n)$ . This introduces a small error which affects only the first term of the sum:

$$-\frac{\tau}{2} \left| \frac{\sigma(2\tau) - \tau}{\tau} + \frac{\rho(2\tau) - 2\rho(\tau) + \rho(0)}{\tau^2} \right|^2 + \frac{\tau}{2} \left| \frac{\sigma(2\tau) - \sigma(0)}{\tau} + \frac{\rho(2\tau) - 2\rho(\tau) + \rho(0)}{\tau^2} \right|^2$$

and which clearly vanishes as  $\tau \rightarrow 0$ , since  $\frac{\rho(2\tau) - 2\rho(\tau) + \rho(0)}{\tau^2} \rightarrow \rho''(0)$ ,  $\frac{\sigma(2\tau) - \sigma(0)}{\tau} \rightarrow 2\sigma'(0)$  and  $\frac{\sigma(2\tau) - \tau}{\tau} \rightarrow 2\sigma'(0) - 1$ .

To ensure the convergence of the sum, we need some regularity on  $\sigma$  and  $\rho$ . In the positive case the regularity of the optimal  $(\sigma, \rho)$  defined above is sufficient but we must show an exponential decay of the terms in the sum to ensure convergence. This will be done after the next paragraph, which deals with the compact support case. In the compact support case, we regularize the optimal pair to get rid of the singularity at  $a$ , as we now explain.

**Compact support case:** In this case we have  $\sigma, \rho : \mathbb{R}^+ \rightarrow \mathbb{R}$  satisfying  $D(\sigma, \rho) > 0$  with  $\rho = \sigma' = 0$  on  $(a, +\infty)$  and  $\sigma, \rho$  of class  $\mathcal{C}^\infty$  on  $[0, \infty) \setminus \{a\}$ . We extend  $\sigma$  and  $\rho$  to  $\mathbb{R}^-$  by 0 and regularize them into  $\mathcal{C}^\infty$  functions on  $[0, \infty)$  while keeping their admissibility in problem (21) as well as the compactness of their support and the value of  $\rho'(0)$ .

To this end, we first regularize by convolution with a function  $\eta \in \mathcal{C}_c^\infty(\mathbb{R})$  with  $\eta \geq 0$ ,  $\int_0^\infty \eta = 1$ , and  $\eta(x) = 0$  for any  $x \notin (0, 1)$ : we obtain functions  $\rho_\varepsilon = \int_{\mathbb{R}} \rho(\cdot + \varepsilon t)\eta(t)dt$  and  $\sigma_\varepsilon = \int_{\mathbb{R}} \sigma(\cdot + \varepsilon t)\eta(t)dt$  which are  $\mathcal{C}^\infty$  on  $[0, \infty)$  and satisfy  $\rho_\varepsilon = \sigma'_\varepsilon = 0$  on  $(a, \infty)$  as

well as  $2\sigma_\varepsilon + \rho_\varepsilon^2 \leq 0$  since this constraint is convex, that is  $C = \{(s, r) \in \mathbb{R}^2 : 2s + r^2 \leq 0\}$  is a convex set.

However, we lost the values of  $\rho(0), \sigma(0)$  and more importantly of  $\rho'(0)$  which appears in problem (21). To this end, **consider**  $\nu \in \mathcal{C}^\infty$  a **smooth, nonincreasing** plateau function such that  $\nu = 1$  on  $(-\infty, \frac{a}{3})$  and  $\nu = 0$  on  $(\frac{2a}{3}, +\infty)$ , and set  $\hat{\sigma}_\varepsilon = \nu\sigma + (1 - \nu)\sigma_\varepsilon$ ,  $\hat{\rho}_\varepsilon = \nu\rho + (1 - \nu)\rho_\varepsilon$ . As  $\sigma$  and  $\rho$  are  $\mathcal{C}^\infty$  on  $[0, +\infty)$  except in  $a$  which is avoided,  $\hat{\sigma}_\varepsilon$  and  $\hat{\rho}_\varepsilon$  are  $\mathcal{C}^\infty$  on  $[0, +\infty)$ , and as  $\hat{\rho}_\varepsilon = \rho$ ,  $\hat{\sigma}_\varepsilon = \sigma$  near 0 we keep  $\hat{\sigma}_\varepsilon(0) = \hat{\rho}_\varepsilon(0) = 0$  and  $\hat{\rho}'_\varepsilon(0) = \rho'(0)$ . Furthermore, the constraint  $2\hat{\sigma}_\varepsilon + \hat{\rho}_\varepsilon^2 \leq 0$  is still fulfilled by convexity. Finally one checks that:

$$2\hat{\sigma}'_\varepsilon + \hat{\rho}'_\varepsilon{}^2 = 2\sigma'_\varepsilon + \rho_\varepsilon'' + \{2(\sigma' - \sigma'_\varepsilon) + (\rho'' - \rho_\varepsilon'')\}\nu + \{(\sigma - \sigma_\varepsilon) + 2(\rho' - \rho'_\varepsilon)\}\nu' + \{\rho - \rho_\varepsilon\}\nu''$$

so that when  $\varepsilon$  goes to 0:

- $2\sigma'_\varepsilon + \rho_\varepsilon''$  converges to  $2\sigma' + \rho''$  in  $L^2(0, \infty)$ .
- $\sigma', \rho''$  are continuous on  $[0, \frac{2a}{3}]$  hence  $2(\sigma' - \sigma'_\varepsilon) + (\rho'' - \rho_\varepsilon'')$  converges to 0 uniformly on  $[0, \frac{2a}{3}]$ . As  $\nu = 0$  on  $(\frac{2a}{3}, +\infty)$  this implies that  $\{2(\sigma' - \sigma'_\varepsilon) + (\rho'' - \rho_\varepsilon'')\}\nu$  converges to 0 in  $L^2(0, \infty)$ .
- $\nu' = \nu'' = 0$  outside of  $[\frac{a}{3}, \frac{2a}{3}]$  where  $\sigma, \sigma'$  and  $\rho'$  are continuous hence  $\{(\sigma - \sigma_\varepsilon) + 2(\rho' - \rho'_\varepsilon)\}\nu' + \{\rho - \rho_\varepsilon\}\nu''$  converges to 0 uniformly hence in  $L^2(0, \infty)$ .

To conclude,  $D(\hat{\sigma}_\varepsilon, \hat{\rho}_\varepsilon) \rightarrow D(\sigma, \rho)$ . This shows that one can find  $(\sigma, \rho)$  admissible in the continuous problem such that  $D(\sigma, \rho) > 0$  and  $\sigma, \rho$  are  $\mathcal{C}^\infty$  on  $[0, +\infty)$ , with  $\rho$  and  $\sigma'$  having compact supports. In particular all the functions  $\sigma, \sigma', \sigma'', \rho, \rho', \rho''$  and  $\rho'''$  can be uniformly bounded by some constant  $M > 0$ .

Then to estimate convergence of  $D^h(\sigma, \rho)$  towards  $D(\sigma, \rho)$  we can truncate the Riemann sum at  $n = \lfloor \frac{a}{\tau} \rfloor$  where the supports of  $\sigma'$  and  $\rho$  are included in  $[0, a]$ . Doing so it is easy to show that

$$\tau \sum_{n=1}^{N-1} \left| \frac{\sigma_{n+1} - \sigma_{n-1}}{\tau} + \frac{\rho_{n+1} - 2\rho_n + \rho_{n-1}}{\tau^2} \right|^2 = \tau \sum_{n=1}^{\lfloor \frac{a}{\tau} \rfloor} |2\sigma'(\tau n) + \rho''(\tau n)|^2 + O(\tau).$$

We conclude **by observing that since**  $(2\sigma' + \rho'')$  is smooth, one has:

$$\tau \sum_{n=1}^{\lfloor \frac{a}{\tau} \rfloor} |2\sigma'(\tau n) + \rho''(\tau n)|^2 \rightarrow \int_0^a (2\sigma' + \rho'')^2 = \int_0^\infty (2\sigma' + \rho'')^2$$

hence the desired convergence. **It follows that for  $h > 0$  small enough, (22) is positive.**

**Positive case:** Recall that in this case we have  $\sigma, \rho : \mathbb{R}^+ \rightarrow \mathbb{R}$  satisfying  $D(\sigma, \rho) > 0$  with  $\sigma = -\rho^2/2$  and  $\rho = -\phi'/\phi$  for some  $\phi > 0$   $C^\infty$  on  $\mathbb{R}^+$ . We also had that  $\phi' \leq 0$  and  $\phi'' = k\phi$  with  $k(t) = \int_0^t s\phi^2(s)ds + t\psi(t) - A$  nondecreasing. Therefore  $\rho$  satisfies on  $\mathbb{R}^+$

$$\rho' = -\frac{\phi''}{\phi} + \frac{\phi'^2}{\phi^2} = \rho^2 - k.$$

This relation allows us to show that the derivatives of  $\rho$  tends to 0 exponentially fast, which will compensate the non compactness of their support. It is important to note that the key argument in the following proofs is that this relation holds on the whole  $\mathbb{R}^+$ : in the case of compact support it only holds on  $[0, a)$  and one cannot obtain the same conclusions (especially, in the compact support case, we cannot have  $\rho'(t) \geq 0$  for all  $t \geq 0$  as shown below). Our analysis begins with the two following lemmas that derive from easy manipulations and antidifferentiation and for which we only **sketch the proofs**.

**Lemma 4.** *Let  $\rho, k : \mathbb{R}^+ \rightarrow \mathbb{R}$  be  $C^1$  functions such that for all  $t \geq 0$ ,  $\rho'(t) = \rho^2(t) - k(t)$ ,  $\rho(t) \geq 0$  and  $k'(t) \geq 0$ . Then for all  $t \geq 0$ ,  $\rho'(t) \geq 0$ .*

*Proof.* Suppose  $\rho'(t) = -r < 0$  for some  $t \geq 0$ , then one can prove that  $\rho$  is nonincreasing on  $(t, \infty)$ . But then so is  $\rho' = \rho^2 - k$  as  $\rho, k' \geq 0$ . Consequently,  $\rho'(s) \leq -r$  for any  $s \geq t$  which cannot stand with the hypothesis that  $\rho \geq 0$ .  $\square$

**Lemma 5.** *Let  $t_1 \in \mathbb{R}$  and let  $\rho : [t_1, +\infty[ \rightarrow \mathbb{R}^+$  be a  $C^1$  function. There is no  $L \in \mathbb{R}$  such that  $\forall t \geq t_1$*

$$\rho^2(t) - L \neq 0 \quad \text{and} \quad \frac{\rho'(t)}{\rho^2(t) - L} \geq 1.$$

*Proof.* The case  $L = 0$  is clear. Otherwise, one integrates  $\frac{\rho'}{\rho^2 - L}$  as  $\log \left| \frac{\rho - \sqrt{L}}{\rho + \sqrt{L}} \right|$  if  $L > 0$  or as  $\frac{1}{\sqrt{-L}} \arctan \left( \frac{\rho}{\sqrt{-L}} \right)$  if  $L < 0$ . In either cases, taking the limit at infinity leads to a contradiction.  $\square$

Thanks to the first lemma,  $\rho$  is nonnegative and nondecreasing (and not zero everywhere), so  $\rho(t) \rightarrow R \in (0, +\infty]$  when  $t \rightarrow \infty$ . In particular there exists  $c > 0$  and  $t_0 > 0$  such that  $\forall t \geq t_0$ ,  $-\frac{\phi'(t)}{\phi(t)} = \rho(t) \geq c > 0$  which leads to  $\phi(t) \leq \phi(t_0) \exp(-c(t - t_0))$ . As a consequence,  $k(t) = \int_0^t s\phi^2(s)ds + t \int_t^\infty \phi^2(s)ds - A$  is bounded and increasing so converges to some  $L \in \mathbb{R}$  and the convergence is exponential since :

$$L - k(t) = \int_t^\infty (s - t)\phi^2(s)ds \leq M \exp(-2ct) \text{ for some } M > 0.$$

Next we must have  $R < +\infty$ . Indeed, otherwise we would have a  $t_1 > 0$  such that  $\forall t \geq t_1$ ,  $\rho'(t) = \rho^2(t) - k(t) \geq \rho^2(t) - L > 0$  hence  $\frac{\rho'(t)}{\rho^2(t) - L} \geq 1$  which is not possible according to the second lemma.

Hence  $R^2 \leq L$ , while since  $\rho' = \rho^2 - k$  remains nonnegative and converges to  $R^2 - L$ ,  $R^2 = L$  and finally,

$$\forall t \geq 0, \rho'(t) = \rho^2(t) - L + L - k(t) \leq L - k(t) \leq M \exp(-2ct).$$

As a consequence,  $\sigma', \rho'', \sigma''$  and  $\rho'''$  decrease exponentially to zero. Indeed:

- $\sigma' = -\rho' \rho$  with  $\rho$  bounded.
- $\rho'' = 2\sigma' \sigma - \psi$  with  $\sigma = -\rho^2/2$  bounded and  $\psi$  decreasing exponentially to zero (as  $\psi(t) = \int_t^\infty \phi^2$  with  $\phi$  decreasing exponentially).
- $\sigma'' = -\rho'^2 - \rho'' \rho$ .
- $\rho''' = 2\rho'' \rho + 2\rho'^2 + \phi^2$ .

Then we get the following estimate for our discretization: write for  $1 \leq n \leq N-1$

$$\frac{\sigma_{n+1} - \sigma_{n-1}}{\tau} = 2\sigma'(\tau n + \eta_n) \quad \text{and} \quad \frac{\rho_{n+1} - 2\rho_n + \rho_{n-1}}{\tau^2} = \rho''(\tau n + \tilde{\eta}_n)$$

for some  $\eta_n, \tilde{\eta}_n \in (-\tau, \tau)$ , so that we have:

$$\left| \left| \frac{\sigma_{n+1} - \sigma_{n-1}}{\tau} + \frac{\rho_{n+1} - 2\rho_n + \rho_{n-1}}{\tau^2} \right|^2 - \tau \sum_{n=1}^{N-1} |2\sigma'(\tau n) - \rho''(\tau n)|^2 \right| = \Delta_n^- \times \Delta_n^+$$

with

$$\begin{aligned} \Delta_n^- &:= |2\sigma'(\tau n + \eta_n) - 2\sigma'(\tau n) + \rho''(\tau n + \tilde{\eta}_n) - \rho''(\tau n)| \\ &\leq 2\tau \times (2\|\sigma''\|_{\infty, (\tau n - \tau, \tau n + \tau)} + \|\rho'''\|_{\infty, (\tau n - \tau, \tau n + \tau)}) \\ &\leq \tau M \exp(-c(\tau n - \tau)) \\ \Delta_n^+ &:= |2\sigma'(\tau n + \eta_n) + 2\sigma'(\tau n) + \rho''(\tau n + \tilde{\eta}_n) + \rho''(\tau n)| \\ &\leq 4\|\sigma'\|_{\infty, (\tau n - \tau, \tau n + \tau)} + 2\|\rho'''\|_{\infty, (\tau n - \tau, \tau n + \tau)} \\ &\leq \tau M \exp(-c(\tau n - \tau)) \end{aligned}$$

for some constants  $M, c > 0$  and finally one can write (for other constants  $M, c > 0$ ):

$$\begin{aligned} &\left| \tau \sum_{n=1}^{N-1} \left| \frac{\sigma_{n+1} - \sigma_{n-1}}{\tau} + \frac{\rho_{n+1} - 2\rho_n + \rho_{n-1}}{\tau^2} \right|^2 - \tau \sum_{n=1}^{N-1} |2\sigma'(\tau n) - \rho''(\tau n)|^2 \right| \\ &\leq \tau^2 \sum_{n=1}^{N-1} M \exp(-c(\tau n - \tau)) \\ &\leq M\tau^2 \sum_{n=0}^{\infty} \exp(-c\tau)^n = M \frac{\tau^2}{1 - \exp(-c\tau)} \sim M \frac{\tau^2}{c\tau} \rightarrow 0 \text{ as } N \rightarrow \infty. \end{aligned}$$



To conclude (i.e. to obtain  $D^h(\boldsymbol{\sigma}, \boldsymbol{\rho}) \rightarrow D(\boldsymbol{\sigma}, \boldsymbol{\rho})$ ), we state that

$$\tau \sum_{n=1}^{N-1} (2\sigma'(\tau n) + \rho''(\tau n))^2 \rightarrow \int_0^\infty (2\sigma' + \rho'')^2 \text{ as } N \rightarrow \infty.$$

This comes from taking  $f = (2\sigma' + \rho'')^2 = \psi^2$  – which is indeed nonincreasing as  $\psi' = -\phi^2 \leq 0$  and  $\psi \geq 0$  – in the following easy result:

**Lemma 6.** *Let  $f : \mathbb{R}^+ \rightarrow \mathbb{R}$  be a continuous and nonincreasing function such that  $\int_0^\infty f$  converges. Let  $a > b > 0$  and  $c_1, c_2, c_3 \in \mathbb{R}$  constants. Then*

$$\frac{1}{N^b} \sum_{l=[c_1]}^{[c_2 N^a + c_3]} f\left(\frac{l}{N^b}\right) \rightarrow \int_0^\infty f \text{ when } N \rightarrow \infty.$$

As for the compact support case, we deduce that for  $h > 0$  small enough, (22) is positive.

### 2.3.5 Neumann boundary conditions

Dealing with Neumann boundary conditions takes us back to the 1D problem (9), where we now take  $u_{N+1} = u_N$  and  $u_{-N-1} = u_{-N}$ . We also suppose  $\lambda < \lambda^*$  so that  $\bar{u} \neq \frac{1}{2}$ . Thanks to Proposition 2, we can suppose  $p_0 = q_0 = \sqrt{2}/2$  in the dual problem (18), and one checks that it is changed into

$$\bar{E}^h = \max_{\substack{p_n^2 + p_{-n}^2 \leq 1 \\ -N \leq n \leq N \\ p_0 = \sqrt{2}/2}} \frac{1}{2} + \frac{1}{\sqrt{2}} p_1 - \frac{\lambda}{h\sqrt{2}} \sum_{n=1}^{N-1} (p_{-n+1} - p_{-n} + p_n - p_{n+1})^2 - \frac{\lambda}{h\sqrt{2}} (p_{-N+1} + p_N)^2.$$

Remember from Section 2.1.2 that the limit value when  $h = \frac{1}{N} \rightarrow 0$  is  $\bar{E} = \bar{E}_N = 1 - \sqrt{2}\lambda$ . This value is (almost) achieved when taking  $p_n = \sqrt{2}/2 - |n|/\sqrt{2}N$  as it gives  $\bar{E}^h \geq 1 - \sqrt{2}\lambda + \frac{3\lambda - \sqrt{2}}{2\sqrt{2}}h$  (but  $3\lambda - \sqrt{2} < 0$ ). Let us denote

$$F(p, \lambda) = \frac{1}{2} + \frac{1}{\sqrt{2}} p_1 - \frac{\lambda}{\sqrt{2}h} \sum_{n=1}^{N-1} (p_{-n+1} - p_{-n} + p_n - p_{n+1})^2$$

$$\tilde{F}(\tilde{p}, \lambda) = \frac{1}{2} + \frac{1}{\sqrt{2}} \tilde{p}_1 - \frac{\lambda}{\sqrt{2}h} \sum_{n=1}^{N-1} (\tilde{p}_{-n+1} - \tilde{p}_{-n} + \tilde{p}_n - \tilde{p}_{n+1})^2 - \frac{\lambda}{\sqrt{2}h} (\tilde{p}_{-N+1} + \tilde{p}_N)^2.$$

Note that the constraint on  $p$  in Dirichlet and Neumann problems is the same:  $p_0 = \sqrt{2}/2$  and  $p_n^2 + p_{-n}^2 \leq 1$ . Now suppose  $p$  is the Dirichlet variable constructed in the previous sections, and form  $\tilde{p}_n = p_n - \frac{|n|}{\sqrt{2}N}$ . We want to compare  $\tilde{F}(\tilde{p}, \lambda) - \bar{E}_N$  to

$F(p, \lambda) - \bar{E}_D$ . As  $\bar{E}_N = 1 - \lambda\sqrt{2} = \bar{E}_D - \lambda\sqrt{2}$ , we split  $\lambda\sqrt{2}$  into  $N \times \frac{\lambda}{\sqrt{2}h} \times \frac{2}{N^2}$  and allocate each  $\frac{2}{N^2}$  to a term involving  $p^2$  in the expression of  $\tilde{E}$ . We obtain:

$$(\tilde{p}_{-n+1} - \tilde{p}_{-n} + \tilde{p}_n - \tilde{p}_{n+1})^2 - \frac{2}{N^2} = (p_{-n+1} - p_{-n} + p_n - p_{n+1} + \frac{\sqrt{2}}{N})^2 - \frac{2}{N^2} = x_n^2 + \frac{2\sqrt{2}}{N}x_n$$

where we denoted  $x_n = p_{-n+1} - p_{-n} + p_n - p_{n+1}$ . When summing, we will recover the term in  $x_n^2$  appearing in  $E(p, \lambda)$ , along with

$$\sum_{n=1}^{N-1} x_n = p_1 - p_N + p_0 - p_{-N+1} = (p_1 - \frac{\sqrt{2}}{2}) - (p_N + p_{N-1} - \sqrt{2}).$$

Besides, one has

$$(\tilde{p}_{-N+1} + \tilde{p}_N)^2 - \frac{2}{N^2} = (p_{-N+1} + p_N - \sqrt{2})^2 + \frac{\sqrt{2}}{N}(p_{-N+1} + p_N - \sqrt{2}) - \frac{3}{N^2}.$$

Then we obtain:

$$\begin{aligned} \tilde{F}(\tilde{p}, \lambda) - \bar{E}_N &= \frac{1}{2} + \frac{1}{\sqrt{2}}p_1 - \frac{1}{2N} - 1 - \frac{\lambda}{\sqrt{2}h} \sum_{n=1}^{N-1} x_n^2 \\ &\quad - \frac{\lambda}{\sqrt{2}h} \times \frac{2\sqrt{2}}{N} (p_1 - \frac{\sqrt{2}}{2}) - \frac{\lambda}{\sqrt{2}h} (p_{-N+1} + p_N - \sqrt{2})^2 \\ &\quad - \frac{\lambda}{\sqrt{2}h} \times \frac{\sqrt{2}}{N} (p_{-N+1} + p_N - \sqrt{2}) + \frac{\lambda}{\sqrt{2}h} \frac{3}{N^2} \\ &= F(p, \lambda) - \bar{E}_D - 2\lambda(p_1 - \frac{\sqrt{2}}{2}) + R \end{aligned} \tag{26}$$

where  $R = \lambda(p_N + p_{-N+1} - \sqrt{2}) - \frac{\lambda}{\sqrt{2}h} (p_N + p_{-N+1} - \sqrt{2})^2 + \frac{3\sqrt{2}\lambda-1}{2N}$ .

At this point, remember  $p$  was obtained from continuous functions  $\sigma$  and  $\rho$  through

$$\begin{cases} p_n = \frac{1}{\sqrt{2}}(\sigma_n + 1 + \rho_n) ; p_{-n} = \frac{1}{\sqrt{2}}(\sigma_n + 1 - \rho_n) \\ \text{with } \sigma_n = N^{-2/3}(\sigma(\tau n) - \tau) ; \rho_n = N^{-1/3}\rho(\tau n) \end{cases}$$

As  $\rho$  and  $\sigma$  are bounded, one sees that  $p_n$  converges to  $\frac{\sqrt{2}}{2}$  uniformly as  $N$  goes to infinity (that is  $\max_{-N \leq n \leq N} |p_n - \frac{\sqrt{2}}{2}| \rightarrow 0$  as  $N \rightarrow \infty$ ). This first shows that  $\tilde{p}$  is admissible in the dual problem (meaning that  $\tilde{p}_n^2 + \tilde{p}_{-n}^2 \leq 1$ ): indeed  $p$  is itself admissible and  $p_n \geq \tilde{p}_n \geq -1 \geq -p_n$  for  $N$  sufficiently large. Second, remember that, at infinity,  $\sigma$  converges to  $-k(a) < 0$  or 0, and  $\rho$  converges to 0. Writing

$$p_{-N+1} + p_N - \sqrt{2} = \frac{1}{\sqrt{2}}(\sigma_N + \sigma_{N-1} + \rho_N - \rho_{N-1})$$

one sees that  $N^{2/3}R \rightarrow 0$  when  $N \rightarrow \infty$ . Then we apply a last trick to include  $2\lambda(p_1 - \frac{\sqrt{2}}{2})$  from (26) into our energies: we remark that

$$F(p, \lambda) - \bar{E}_D - 2\lambda(p_1 - \frac{\sqrt{2}}{2}) = (1 - 2\sqrt{2}\lambda)(F(p, \frac{\lambda}{1 - 2\sqrt{2}\lambda}) - \bar{E}_D).$$

This finally shows that

$$N^{2/3}(\tilde{F}(\tilde{p}, \lambda) - \bar{E}_N) = N^{2/3}\left((1 - 2\sqrt{2}\lambda)(F(p, \frac{\lambda}{1 - 2\sqrt{2}\lambda}) - \bar{E}_D)\right) + N^{2/3}R.$$

converges to a positive value when  $N$  tends to infinity; hence the  $O(h^{2/3})$  rate is also true in the Neumann setting.

### 3 The Raviart-Thomas total variation

In this section, we show that relying on a simple construction we can define a discretization of the total variation with a convergence behaviour much better than (5). It is built upon a finite-element approximation of the dual fields. We first introduce the main definition, and then study the convergence rate for the discrete ROF problems and prove Theorem 2.

#### 3.1 Definitions

The idea behind the definition of the isotropic total variation (5) is of course to catch the  $L^1$  norm of the gradient of  $u$  based on a discretization of the expression  $\text{TV}(u) = \int_{\Omega} |\nabla u|$ . To do so, one chooses a finite differences operator  $D$ , defined on the mesh  $\Omega = \cup C_{i,j}$ , to approximate  $\nabla$ . However, the non isotropy of the grid itself prevents  $D$  from being isotropic, as it has to involve a notion of neighbour on this two-directional grid. On the contrary, the dual definition of TV offers the possibility to discretize a field rather than an operator. In the formulas

$$\text{TV}_N(u) = \sup \left\{ - \int_{\Omega} u \operatorname{div} \phi : \phi \in \mathcal{C}_c^1(\Omega, \mathbb{R}^2), \|\phi\|_{\infty} \leq 1 \right\} \quad (27)$$

$$\text{TV}_D(u) = \sup \left\{ - \int_{\Omega} u \operatorname{div} \phi + \int_{\partial\Omega} b \langle \phi | \vec{n} \rangle : \phi \in \mathcal{C}^1(\Omega, \mathbb{R}^2), \|\phi\|_{\infty} \leq 1 \right\} \quad (28)$$

we can keep the exact operator  $\operatorname{div}$  but replace the spaces  $\mathcal{C}_c^1(\Omega, \mathbb{R}^2)$  and  $\mathcal{C}^1(\Omega, \mathbb{R}^2)$  of (compactly supported)  $\mathcal{C}^1$  fields from  $\Omega$  to  $\mathbb{R}^2$ , by a space of discrete fields favouring no direction.

The most simple space available is the so-called ‘‘Raviart-Thomas’’ finite elements space [29], which first seems to have been used in this context in [21, 22] (for a regularized variant of the total variation). Raviart-Thomas fields are defined via their fluxes through

the edges of the squares, we will denote  $f_{i+\frac{1}{2},j}$  (resp.  $f_{i,j+\frac{1}{2}}$ ) the averaged flux through the edge between the squares  $C_{i,j}$  and  $C_{i+1,j}$  (resp.  $C_{i,j}$  and  $C_{i,j+1}$ ), and  $(x_{i,j}, y_{i,j})$  the center of the square  $C_{i,j}$ . Then the Raviart-Thomas fields are the elements of

$$RT0 = \left\{ \phi : \Omega \rightarrow \mathbb{R}^2 : \exists (f_{i+\frac{1}{2},j}, f_{i,j+\frac{1}{2}})_{i,j}, \forall 1 \leq i, j \leq N, \right. \\ \left. \phi(x, y) = \begin{pmatrix} \frac{f_{i+\frac{1}{2},j} + f_{i-\frac{1}{2},j}}{2} + (f_{i+\frac{1}{2},j} - f_{i-\frac{1}{2},j}) \frac{x - x_{i,j}}{h} \\ \frac{f_{i,j+\frac{1}{2}} + f_{i,j-\frac{1}{2}}}{2} + (f_{i,j+\frac{1}{2}} - f_{i,j-\frac{1}{2}}) \frac{y - y_{i,j}}{h} \end{pmatrix} \text{ in } C_{i,j} \right\}. \quad (29)$$

In the sequel, we will write  $\phi = \phi_f \in RT0$  to precise that  $f$  denotes the fluxes of the Raviart-Thomas fields  $\phi$  according to (29). In the Neumann setting, we use Raviart-Thomas fields vanishing on the boundary of  $\Omega$ , which we denote  $RT0_0$ :

$$RT0_0 = \{ \phi_f \in RT0 : \forall 1 \leq i, j \leq N, f_{\frac{1}{2},j} = f_{N+\frac{1}{2},j} = f_{i,\frac{1}{2}} = f_{i,N+\frac{1}{2}} = 0 \}.$$

Finally, in the Neumann setting, the Raviart-Thomas total variation we study is, for any  $u^h \in P0$ :

$$TV_{RT,N}^h(u^h) = \sup \left\{ - \int_{\Omega} u^h \operatorname{div} \phi : \phi = \phi_f \in RT0_0, \|\phi\|_{\infty} \leq 1 \right\} \quad (30)$$

while in the Dirichlet setting, we use the source term  $b^h$  of the ROF problem in the integral on  $\partial\Omega$ , **this term is obtained by averaging the boundary datum  $b$  on each facet of the boundary elements:**

$$TV_{RT,D}^h(u^h) = \sup \left\{ - \int_{\Omega} u^h \operatorname{div} \phi + \int_{\partial\Omega} b^h \langle \phi | \vec{n} \rangle : \phi = \phi_f \in RT0, \|\phi\|_{\infty} \leq 1 \right\}. \quad (31)$$

**(Remark here that since the flux  $\langle \phi | \vec{n} \rangle$  is constant on each facet, the expression is the same if we replace here  $b^h$  with  $b$ .)**

We stress the fact that no discontinuity jump appears in the calculus of  $\operatorname{div} \phi_f$  so that, for instance in the Neumann setting, for  $\phi_f \in RT0_0$ :

$$\begin{aligned} - \int_{\Omega} u^h \operatorname{div} \phi_f &= - \sum_{i,j} h^2 u_{i,j}^h \frac{1}{h} (f_{i+\frac{1}{2},j} - f_{i-\frac{1}{2},j} + f_{i,j+\frac{1}{2}} - f_{i,j-\frac{1}{2}}) \\ &= h \sum_{i,j} f_{i+\frac{1}{2},j} (u_{i+1,j}^h - u_{i,j}^h) + h \sum_{i,j} f_{i,j+\frac{1}{2}} (u_{i,j+1}^h - u_{i,j}^h) \\ &= h \sum_{i,j} \left\langle \begin{pmatrix} f_{i+\frac{1}{2},j} \\ f_{i,j+\frac{1}{2}} \end{pmatrix} \middle| \begin{pmatrix} (u^h)_{i+1,j} - (u^h)_{i,j} \\ (u^h)_{i,j+1} - (u^h)_{i,j} \end{pmatrix} \right\rangle = h \langle \langle f | Du^h \rangle \rangle. \end{aligned}$$

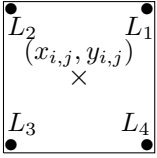
In particular, as noted by the authors of [26], the isotropic total variation (5) can be recovered in the context of Raviart-Thomas fields as:

$$TV_i^h(u^h) = \sup \left\{ - \int_{\Omega} u^h \operatorname{div} \phi : \phi = \phi_f \in RT0_0, \forall i, j, \left| \begin{pmatrix} f_{i+\frac{1}{2},j} \\ f_{i,j+\frac{1}{2}} \end{pmatrix} \right| \leq 1 \right\}.$$

In  $\text{TV}_{RT}^h$ , the constraint on  $\phi_f$  is the same as on  $\phi$  on the continuous TV, namely that  $|\phi_f(x)| \leq 1$  for all  $x \in \Omega$ . Note that since the two components of  $\phi_f$  are piecewise affine, the constraint of being less than 1 everywhere on  $\Omega$  reduces to being less than 1 in the corners of the mesh, that is

$$\begin{aligned} \text{TV}_{RT,N}^h(u_h) &= \sup \left\{ - \int_{\Omega} u_h \operatorname{div} \phi_f : \phi_f \in RT0_0, \right. \\ &\quad \left. \forall 1 \leq i, j \leq N, \max_{1 \leq k \leq 4} |(L_k \phi_f)_{i,j}| \leq 1 \right\} \\ \text{TV}_{RT,D}^h(u_h) &= \sup \left\{ - \int_{\Omega} u_h \operatorname{div} \phi_f + \int_{\partial\Omega} b^h \langle \phi_f | \vec{n} \rangle : \phi_f \in RT0, \right. \\ &\quad \left. \forall 1 \leq i, j \leq N, \max_{1 \leq k \leq 4} |(L_k \phi_f)_{i,j}| \leq 1 \right\} \end{aligned}$$

with:

$$\begin{aligned} (L_1 \phi_f)_{i,j} &= \phi_f(x_{i,j} + \frac{h}{2}^-, y_{i,j} + \frac{h}{2}^-) = (f_{i+\frac{1}{2},j}, f_{i,j+\frac{1}{2}})^T \\ (L_2 \phi_f)_{i,j} &= \phi_f(x_{i,j} - \frac{h}{2}^-, y_{i,j} + \frac{h}{2}^-) = (f_{i-\frac{1}{2},j}, f_{i,j+\frac{1}{2}})^T \\ (L_3 \phi_f)_{i,j} &= \phi_f(x_{i,j} - \frac{h}{2}^-, y_{i,j} - \frac{h}{2}^-) = (f_{i-\frac{1}{2},j}, f_{i,j-\frac{1}{2}})^T \\ (L_4 \phi_f)_{i,j} &= \phi_f(x_{i,j} + \frac{h}{2}^-, y_{i,j} - \frac{h}{2}^-) = (f_{i+\frac{1}{2},j}, f_{i,j-\frac{1}{2}})^T \end{aligned} \quad (32)$$


where we used the notation  $a^-$  ( $a^+$ ) to denote the left (right) limit. Other choices of constraints on  $\phi_f$  proposed in [19, 24] lead to (visually) better numerical results, however **no error estimates are proven yet for these**. Nevertheless, they also fit the framework of Raviart-Thomas total variations, see Section 4.

### 3.2 Convergence rate for $\text{TV}_{RT}^h$ and proof of Theorem 2

In [15] the authors have studied Crouzeix-Raviart finite elements based total variation on a triangular mesh, which can be computed by approximating the dual fields with Raviart-Thomas fields with a norm constraint only in the center of each triangle. Given a source term  $g \in L^\infty$ , and under a regularity assumption on the dual field, they show there exists a constant  $c$  (depending on  $g$  and the value of the continuous ROF problem) such that  $|\bar{E} - \bar{E}^h| \leq ch$ , where  $\bar{E}$  and  $\bar{E}^h$  are respectively the optimal values of the continuous and discrete problems. **In this section, we prove that a similar study is valid for the Raviart-Thomas total variation, and deduce Theorem 2. In this case, the continuous and discrete problems are:**

$$\bar{u} = \arg \min_{u \in BV(\Omega)} \frac{1}{2\lambda} \|u - g\|_{L^2}^2 + \text{TV}(u) =: E(u) \quad (33)$$

$$\bar{u}^h = \arg \min_{u^h \in P0} \frac{1}{2\lambda} \|u^h - g\|_{L^2}^2 + \text{TV}_{RT}^h(u^h) =: E^h(u^h) \quad (34)$$

with appropriate variants for Dirichlet and Neumann boundary conditions (recall that when no subscript  $N$  or  $D$  is specified, the proposed results are valid for both settings).

Observe that in (34) we may consider  $g^h$  or  $g$  in the  $L^2$  term without changing the optimal point  $\bar{u}^h$ , since by orthogonality  $\|u^h - g\|_{L^2}^2 = \|u^h - g^h\|_{L^2}^2 + \|g^h - g\|_{L^2}^2$ . The reason for our choice is that otherwise, the additional term  $\frac{1}{2\lambda}\|g^h - g\|_{L^2}^2$  (of order  $h$  whenever  $g \in BV(\Omega)$ , but which could be larger in general) appears in the difference of the energies  $\bar{E} - \bar{E}^h$ . Since we are of course interested in the approximation error for the solution  $\bar{u}$  and not for the data  $g$ , we found this choice more meaningful.

Thanks to the strong convexity of the energy these estimates also control the squared  $L^2$  error between  $\bar{u}$  and  $\bar{u}^h$ .

The proof of the rate is two-fold: a first estimate comes from the primal problems, a second from the dual. The first one relies on the conformal aspect of our discrete total variation  $TV_{RT}^h$  (30), (31) with respect to the continuous TV (27), (28). As in [6] (which addresses a different, anisotropic “ $\ell_1$ ” total variation), it follows from the TV-diminishing lemma (valid for both Dirichlet and Neumann cases):

**Lemma 7.** *For any  $u \in BV \cap L^2(\Omega)$  admissible in the continuous ROF problem (1), one has  $TV_{RT}^h(\Pi_{P_0}u) \leq TV(u)$ .*

*Proof.* The main argument is that if  $\phi \in RT_0$ , then  $\text{div } \phi$  is piecewise constant so that  $u^h = \Pi_{P_0}(u)$  satisfies

$$\int_{\Omega} u^h \text{div } \phi = \int_{\Omega} u \text{div } \phi.$$

On the other hand, any Raviart-Thomas field is easily approximated with smooth fields (with compact support if  $\phi \in RT_0$ ), with divergences uniformly bounded (in  $L^\infty$ ), hence we deduce that  $\int_{\Omega} u \text{div } \phi \leq TV(u)$ . The result follows by taking the supremum over admissible fields  $\phi$ .  $\square$

Another useful result is the following, valid for any function with bounded variation:

**Lemma 8.** *There exists a constant  $c > 0$  such that*

$$\forall f \in BV(\Omega), \|f - \Pi_{P_0}f\|_{L^1(\Omega)} \leq chTV(f).$$

*Proof.* The proof is classical and is obtained by integrating first over one element. For a smooth function with average  $\bar{f}$  in the unit square, one has:

$$\begin{aligned} \int_{[0,1]^2} |f(x) - \bar{f}| dx &= \int_{[0,1]^2} \left| \int_{[0,1]^2} f(x) - f(y) dy \right| dx \\ &\leq \int_{[0,1]^2} \int_{[0,1]^2} \int_0^1 |\nabla f(x + s(y-x)) \cdot (y-x)| ds dy dx \leq c \int_{[0,1]^2} |\nabla f(x)| dx. \end{aligned}$$

The estimate follows then, using an approximation of a generic  $f$  with smooth function, by a decomposition of the norms on each square and a scaling argument.  $\square$

Using strong convexity of the primal objectives leads to the first estimate:

**Proposition 5.** *The solutions  $\bar{u}, \bar{u}^h$  of (33), (34) satisfy*

$$\frac{1}{2\lambda} \|\bar{u}^h - \Pi_{P_0} \bar{u}\|_{L^2}^2 \leq \bar{E} - \bar{E}^h - \frac{1}{2\lambda} (\|\bar{u} - g\|_{L^2}^2 - \|\Pi_{P_0}(\bar{u}) - g\|_{L^2}^2).$$

*Proof.* This simply follows from Lemma 7 which implies that

$$\begin{aligned} TV(\bar{u}) + \frac{1}{2\lambda} \|\bar{u} - g\|^2 &\geq TV(\Pi_{P_0}(\bar{u})) + \frac{1}{2\lambda} \|\Pi_{P_0}(\bar{u}) - g\|_{L^2}^2 \\ &\quad + \frac{1}{2\lambda} (\|\bar{u} - g\|^2 - \|\Pi_{P_0}(\bar{u}) - g\|_{L^2}^2). \end{aligned}$$

One uses then the minimality of  $\bar{u}^h$  for the discrete energy.  $\square$

*Remark 1.* We have:

$$\|\bar{u} - g\|^2 - \|\Pi_{P_0}(\bar{u}) - g\|_{L^2}^2 = \int (\bar{u} - \Pi_{P_0}(\bar{u}))(\bar{u} + \Pi_{P_0}(\bar{u}) - 2g)$$

and since  $\bar{u} - \Pi_{P_0}(\bar{u})$  is orthogonal to  $P_0$  functions, this is also

$$\int (\bar{u} - \Pi_{P_0}(\bar{u}))(\bar{u} - \Pi_{P_0}(\bar{u}) - 2(g - \Pi_{P_0}(g))) = \|\bar{u} - \Pi_{P_0}(\bar{u})\|_{L^2}^2 - 2 \int (\bar{u} - \Pi_{P_0}(\bar{u}))(g - \Pi_{P_0}(g)),$$

hence using Lemma 8, Proposition 5 yields:

$$\begin{aligned} \frac{1}{4\lambda} \|\bar{u}^h - \bar{u}\|_{L^2}^2 &\leq \frac{1}{2\lambda} \|\bar{u}^h - \Pi_{P_0} \bar{u}\|_{L^2}^2 + \frac{1}{2\lambda} \|\bar{u} - \Pi_{P_0} \bar{u}\|_{L^2}^2 \\ &\leq \bar{E} - \bar{E}^h + h \frac{c}{\lambda} \|g - \Pi_{P_0}(g)\|_{L^\infty} TV(\bar{u}). \end{aligned} \quad (35)$$

In addition, if  $g$  is smooth (which is not really the case we are interested in in this study), one sees that the additional error term in the right-hand side is of higher order.

The second part of the estimate relies on the evaluation of the dual problems of (33) and (34). In the continuous setting, switching the min operator from (33) with the supremum defining the total variation leads to the following dual problems:

$$\bar{\phi}_N \in \arg \max_{\substack{\phi \in \mathcal{H}^0: \\ \|\phi\|_\infty \leq 1}} - \int_{\Omega} g \operatorname{div} \phi - \frac{\lambda}{2} \|\operatorname{div} \phi\|_{L^2}^2 =: D_N(\phi), \quad (36)$$

$$\bar{\phi}_D \in \arg \max_{\substack{\phi \in \mathcal{H}: \\ \|\phi\|_\infty \leq 1}} - \int_{\Omega} g \operatorname{div} \phi - \frac{\lambda}{2} \|\operatorname{div} \phi\|_{L^2}^2 + \int_{\partial\Omega} b \langle \phi | \vec{n} \rangle =: D_D(\phi) \quad (37)$$

where  $\mathcal{H} = \{\phi \in L^\infty(\Omega) : \operatorname{div} \phi \in L^2(\Omega)\}$  and  $\mathcal{H}^0$  is the subset of  $\mathcal{H}$  made of fields vanishing at the boundary in the weak sense:  $\mathcal{H}^0 = \{\phi \in \mathcal{H} : \forall u \in H^1(\Omega), \int_{\Omega} \langle \nabla u | \phi \rangle =$

$-\int_{\Omega} u \operatorname{div} \phi$ . Observe, for instance in the Neumann setting, that for any  $\phi \in \mathcal{H}^0$  such that  $\|\phi\|_{\infty} \leq 1$ , one has  $D_N(\phi) \leq E(\bar{u}) = \operatorname{TV}(\bar{u}) + \frac{1}{2\lambda}\|\bar{u} - g\|_{L^2}^2$ . The Euler-Lagrange equation for the ROF problem (see [12]) shows that  $\bar{u}$  is a minimizer of (33) if and only if there exists  $\bar{\phi} \in \mathcal{H}$  with  $\bar{u} - g = \lambda \operatorname{div} \bar{\phi}$ ,  $\|\bar{\phi}\|_{\infty} \leq 1$  and  $-\int_{\Omega} \bar{u} \operatorname{div} \bar{\phi} = \operatorname{TV}(\bar{u})$ . Choosing  $\phi = \bar{\phi}$  in the above inequality shows that strong duality between primal and dual problems holds. Finally,  $D(\bar{\phi}) = \bar{E}$  through the relation  $\bar{u} = g + \lambda \operatorname{div} \bar{\phi}$ . The same relations hold in the discrete case which is completely similar and where **the dual problems are given by**:

$$\begin{aligned} \bar{\phi}_N^h &\in \arg \max_{\substack{\phi^h \in RT0_0 \\ \|\phi^h\|_{\infty} \leq 1}} - \int_{\Omega} g^h \operatorname{div} \phi^h - \frac{\lambda}{2} \|\operatorname{div} \phi^h\|_{L^2}^2 + \frac{1}{2\lambda} \|g^h - g\|_{L^2}^2 =: D_N^h(\phi^h), \\ \bar{\phi}_D^h &\in \arg \max_{\substack{\phi^h \in RT0 \\ \|\phi^h\|_{\infty} \leq 1}} - \int_{\Omega} g^h \operatorname{div} \phi^h - \frac{\lambda}{2} \|\operatorname{div} \phi^h\|_{L^2}^2 + \int_{\partial\Omega} b^h \langle \phi^h | \vec{n} \rangle + \frac{1}{2\lambda} \|g^h - g\|_{L^2}^2 =: D_D^h(\phi^h). \end{aligned}$$

(Here, the terms  $\frac{1}{2\lambda}\|g^h - g\|_{L^2}^2$  appear because we used  $g$  in the definition (34) of the discrete energy.)

To estimate the discrete dual energy, one has to be able to get a discrete field from a continuous one through a projection operator. This is classically achieved by the operator  $\Pi_{RT0} : \mathcal{H} \rightarrow RT0$  which takes  $\phi = (\phi_1, \phi_2) : \Omega \rightarrow \mathbb{R}^2$  to  $\phi_f \in RT0$  where the fluxes through the edges of the mesh  $f$  are defined by

$$f_{i+1/2,j} = \frac{1}{h} \int_{E_{i+1/2,j}} \phi_1(x,y) dy, \quad f_{i,j+1/2} = \frac{1}{h} \int_{E_{i,j+1/2}} \phi_2(x,y) dx \quad (38)$$

where  $E_{i+1/2,j} = \partial C_{i,j} \cap \partial C_{i+1,j}$  and  $E_{i,j+1/2} = \partial C_{i,j} \cap \partial C_{i,j+1}$ . This projection operator enjoys two properties that derive from basic observations. First, using Green's formula, one has:

$$\frac{1}{h^2} \int_{C_{i,j}} \operatorname{div} z = \frac{f_{i+1/2,j} - f_{i-1/2,j}}{h} + \frac{f_{i,j+1/2} - f_{i,j-1/2}}{h}$$

which shows, according to the definition (29), the following result:

**Lemma 9.**  $\forall \phi \in \mathcal{H}, \quad \Pi_{P0}(\operatorname{div} \phi) = \operatorname{div} (\Pi_{RT0}(\phi))$ .

Next, we need to understand the behavior of  $\Pi_{RT0}$  with respect to the infinite norm. For a general  $\phi$ , one cannot expect better than  $\|\Pi_{RT0}(\phi)\|_{\infty} \leq \sqrt{2}\|\phi\|_{\infty}$ . However if in addition  $\phi$  is Lipschitz, one can show the following:

**Lemma 10.** *If  $\phi : \Omega \rightarrow \mathbb{R}^2$  is  $L$ -Lipschitz and if  $\|\phi\|_{\infty} \leq 1$  then its projection  $\phi^h = \Pi_{RT0}(\phi)$  satisfies  $\|\phi^h\|_{\infty} \leq 1 + Lh$ .*

*Proof.* It follows from the fact that if  $(x_c, y_c)$  is the center of the element  $C_{i,j}$ , then from (38) we get that  $|f_{i+1/2,j} - \phi_1(x_c, y_c)| \leq Lh/\sqrt{2}$  and  $|f_{i,j+1/2} - \phi_2(x_c, y_c)| \leq Lh/\sqrt{2}$ . We use that  $\phi_1^2(x_c, y_c) + \phi_2^2(x_c, y_c) \leq 1$  to conclude.  $\square$



In our analysis, we will consequently need a Lipschitz hypothesis to hold on the optimal dual field  $\bar{\phi}$ . As noticed by [15], this hypothesis is reasonable in the sense that it is known to hold when  $g$  is the characteristic of a disk and  $b = 0$  in the Dirichlet case (or  $\Omega = \mathbb{R}^2$ ), as well as in the case  $g = g_\nu$  (where one can even take  $L = 0$ , since  $\bar{\phi} = \nu$  is a dual solution). **On the other hand, [8] provides an example of a  $g \in L^\infty(\Omega)$  for which the optimal dual field is not Lipschitz continuous—only  $C^{0,1/2}$ .**

We now prove the following estimates:

**Proposition 6.** *Suppose the dual continuous problem (36), (37) admits a  $L$ -Lipschitz solution, then one has:*

$$\begin{aligned}\bar{E}_N &\leq (1 + Lh)\bar{E}_N^h \\ \bar{E}_D &\leq (1 + Lh)\bar{E}_D^h + \|b - b^h\|_{L^1(\partial\Omega)}.\end{aligned}$$

*Proof.* We consider the Dirichlet case (the other being similar). Let  $\bar{\phi}$  be a Lipschitz optimal dual field for the continuous problem,  $\phi^h$  its projection onto Raviart-Thomas fields, and  $\tilde{\phi}^h = \phi^h/(1 + Lh)$ : then thanks to Lemma 10 it is admissible in the discrete dual problem, so that  $\bar{E}_D^h \geq D_D^h(\tilde{\phi}^h)$ .

We rewrite

$$\begin{aligned}\bar{E}_D = D_D(\bar{\phi}) &= - \int_{\Omega} g \operatorname{div} \bar{\phi} - \frac{\lambda}{2} \|\operatorname{div} \bar{\phi}\|_{L^2}^2 + \int_{\partial\Omega} b \langle \bar{\phi} | \vec{n} \rangle \\ &= - \frac{1}{2\lambda} \|\lambda \operatorname{div} \bar{\phi} + g\|_{L^2}^2 + \frac{1}{2\lambda} \|g\|_{L^2}^2 + \int_{\partial\Omega} (b - b^h) \langle \bar{\phi} | \vec{n} \rangle + \int_{\partial\Omega} b^h \langle \phi^h | \vec{n} \rangle \\ &\leq - \frac{1}{2\lambda} \|\lambda \operatorname{div} \phi^h + g^h\|_{L^2}^2 + \frac{1}{2\lambda} \|g\|_{L^2}^2 + \int_{\partial\Omega} b^h \langle \phi^h | \vec{n} \rangle + \int_{\partial\Omega} |b - b^h|\end{aligned}$$

where we have used that  $\operatorname{div} \phi^h + g^h$  is the  $L^2$  projection of  $\operatorname{div} \bar{\phi} + g$  on piecewise constant functions (Lemma 9).

Then, we use that

$$\begin{aligned}- \frac{1}{2\lambda} \|\lambda \operatorname{div} \phi^h + g^h\|_{L^2}^2 + \frac{1}{2\lambda} \|g\|_{L^2}^2 &= - \int g^h \operatorname{div} \phi^h - \frac{\lambda}{2} \|\operatorname{div} \phi^h\|_{L^2}^2 + \frac{1}{2\lambda} (\|g\|_{L^2}^2 - \|g^h\|^2) \\ &\leq -(1 + Lh) \int g^h \operatorname{div} \tilde{\phi}^h - (1 + Lh)^2 \frac{\lambda}{2} \|\operatorname{div} \tilde{\phi}^h\|_{L^2}^2 + \frac{1}{2\lambda} \|g - g^h\|^2 \\ &\leq (1 + Lh) \left( - \int g^h \operatorname{div} \tilde{\phi}^h - \frac{\lambda}{2} \|\operatorname{div} \tilde{\phi}^h\|_{L^2}^2 + \frac{1}{2\lambda} \|g - g^h\|^2 \right),\end{aligned}$$

and we deduce the second inequality in the statement of the proposition.  $\square$

**Proof of Theorem 2.** Theorem 2 follows as a corollary of Propositions 5 (more precisely, see (35) in Remark 1) and 6. In the Dirichlet case, we need to assume  $b \in BV(\partial\Omega)$  to ensure that  $\|b - b^h\|_{L^1(\partial\Omega)} \leq ch$  (cf Lemma 8).

*Remark 2.* As mentioned, a similar result has been first established in [15], a non-conforming P1 approximation, this latter result has been generalized in the recent papers to other classes of discontinuous/mixed methods in [4, 5, 8].

*Remark 3.* We observe that similar rates could be obtained with a weaker TV diminishing lemma, if we had:  $\text{TV}^h(\Pi_{P0}(u)) \leq (1 + ch)\text{TV}(u)$ , which could be true for other discrete total variations.

## 4 Implementation and results

### 4.1 A united framework

As we have seen, the Raviart-Thomas fields offer a united framework to deal with different total variations. Indeed,  $\text{TV}_i^h$ ,  $\text{TV}_{RT}^h$  as well as the total variation proposed in [19, 24] (that we will refer to as ‘‘Condat TV’’, referring to the implementation in [19]) can all be expressed in the form:

$$\begin{aligned} \text{TV}_N^L(u^h) &= \sup \left\{ - \int_{\Omega} u^h \operatorname{div} \phi : \phi \in RT0_0, \| |L\phi| \|_{\infty} \leq 1 \right\} \\ \text{TV}_D^L(u^h) &= \sup \left\{ - \int_{\Omega} u^h \operatorname{div} \phi + \int_{\partial\Omega} b^h \langle \phi | \vec{n} \rangle : \phi \in RT0, \| |L\phi| \|_{\infty} \leq 1 \right\} \end{aligned}$$

where  $L : RT0 \rightarrow (\mathbb{R}^2)^{\mathcal{T}}$  is some linear operator giving the constraints that the dual field must satisfy, namely that  $\forall i, j, |(L\phi)_{i,j}| \leq 1$ .

In the case of the isotropic total variation, one has  $L = L_1$ , for Raviart-Thomas  $L = (L_1, L_2, L_3, L_4)$ , and for Condat  $L = (L_{\bullet}, L_{\leftrightarrow}, L_{\leftrightarrow})$  where these operators are given, for  $0 \leq i, j \leq N$ , by (32) and:

$$\begin{aligned} (L_{\bullet}\phi_f)_{i,j} &= \frac{1}{2} \begin{pmatrix} f_{i+\frac{1}{2},j} + f_{i-\frac{1}{2},j} \\ f_{i,j+\frac{1}{2}} + f_{i,j-\frac{1}{2}} \end{pmatrix} \\ (L_{\leftrightarrow}\phi_f)_{i,j} &= \begin{pmatrix} \frac{1}{4}(f_{i+\frac{1}{2},j} + f_{i-\frac{1}{2},j} + f_{i+\frac{1}{2},j+1} + f_{i-\frac{1}{2},j+1}) \\ f_{i,j+\frac{1}{2}} \end{pmatrix} \\ (L_{\updownarrow}\phi_f)_{i,j} &= \begin{pmatrix} f_{i+\frac{1}{2},j} \\ \frac{1}{4}(f_{i,j+\frac{1}{2}} + f_{i,j-\frac{1}{2}} + f_{i+1,j+\frac{1}{2}} + f_{i+1,j-\frac{1}{2}}) \end{pmatrix} \end{aligned}$$

with  $f_{k,l} = 0$  for couples  $(k, l)$  such that this quantity is not defined.

Note that the four variants of the isotropic total variation (obtained through the four combinations of directions selected to discretize the  $\nabla$  operator) correspond to enforcing the constraints  $\| |L_k(\phi_f)| \|_\infty \leq 1$  for  $1 \leq k \leq 4$  separately. On the contrary, the Raviart-Thomas total variation enforces the four of them simultaneously. **More recent results on such general forms of discrete total variations are found in [16, 17]**

## 4.2 Resolution by a primal-dual algorithm

While the one-dimensional problem 9 is easily solved by a dual method, possibly accelerated (but such a small dimensional problem is solved in less than some fractions of a second on a standard modern computer, up to high accuracy), the 2D problems, with their nested constraints, require a more involved implementation.

We write the (dual) ROF problem in the following way, for instance for Neumann boundary conditions:

$$\begin{aligned} & \min_{u^h \in P0} \frac{1}{2\lambda} \|u^h - g^h\|_{L^2}^2 + \sup \left\{ - \int_{\Omega} u^h \operatorname{div} \phi_f : \phi_f \in RT0_0, \| |L\phi_f| \|_\infty \leq 1 \right\} \\ &= \sup_{\phi_f \in RT0_0} \min_{u^h \in P0} \frac{1}{2\lambda} \|u^h - g^h\|_{L^2}^2 - \int_{\Omega} u^h \operatorname{div} \phi_f - F(L\phi_f) \\ &= - \min_{\phi_f \in RT0_0} G(\phi_f) + F(L\phi_f) \end{aligned}$$

where  $G(\phi_f) = \frac{\lambda}{2} \|\operatorname{div} \phi_f\|_{L^2}^2 + \int_{\Omega} g^h \operatorname{div} \phi_f$  and  $F : (\mathbb{R}^2)^{\mathcal{I}} \rightarrow \mathbb{R}$  is given by  $F(z) = 0$  if  $\| |z| \|_\infty \leq 1$ ,  $+\infty$  otherwise. Note that the optimal primal solution will be obtained from the optimal  $\bar{\phi}_f$  through  $\bar{u}^h = g^h + \lambda \operatorname{div} \bar{\phi}_f$ .

This allows one to use one of the primal-dual algorithm presented in [14] for which one needs to calculate the following proximal operators (we denote  $F^*$  the convex conjugate of  $F$  and use the Moreau identity to calculate its prox, see [9]):

$$\begin{aligned} (Id + \tau \partial G)^{-1}(\phi_f) &= \left( \frac{1}{\tau} Id + \lambda DD^* \right)^{-1} \left( \frac{1}{\tau} \phi_f + Dg^h \right) \\ (Id + \sigma \partial F^*)^{-1}(z) &= \left\{ \begin{array}{l} 0 \text{ if } |z_i| \leq \sigma \\ z_i \left(1 - \frac{\sigma}{|z_i|}\right) \text{ otherwise} \end{array} \right\}_{i \in \mathcal{I}} \end{aligned}$$

where  $D = -\operatorname{div}^*$  is the opposite of the dual operator of the divergence on the  $RT0$  fields, which corresponds to a finite difference approximation of the gradient. Finally, we use the simplest version of the proposed algorithm and obtain the following:

**Algorithm.** From  $\phi_f^0 \in RT0_0$ ,  $z^0 \in (\mathbb{R}^2)^{\mathcal{I}}$ , and  $\sigma, \tau > 0$  such that  $\sigma\tau \|L\|^2 \leq 1$ , set

$\bar{\phi}_f^0 = \phi_f^0$  and do  $\forall n \geq 0$ :

$$\begin{aligned} z^{n+1} &= (Id + \sigma \partial F^*)^{-1}(z^n + \sigma L^* \bar{\phi}_f^n) \\ \phi_f^{n+1} &= \left( \frac{1}{\tau} Id + \lambda DD^* \right)^{-1} \left( \frac{1}{\tau} \phi_f^n - L^* z^{n+1} + Dg^h \right) \\ \bar{\phi}_f^{n+1} &= 2\phi_f^{n+1} - \phi_f^n. \end{aligned}$$

One checks that in the Dirichlet setting, the function  $G$  is replaced by

$$G(\phi_f) = \frac{\lambda}{2} \|\operatorname{div} \phi_f\|_{L^2}^2 + \int_{\Omega} g^h \operatorname{div} \phi_f - \int_{\partial\Omega} b^h \langle \phi_f | \vec{n} \rangle$$

and that the same algorithm applies just replacing  $Dg^h$  with the appropriate correction to take into account the boundary term (namely in Neumann  $Dg^h = 0$  on the boundary edges while in Dirichlet  $Dg^h$  has value  $Dg_b^h$  such that  $\int_{\partial\Omega} \langle \phi_f | Dg_b^h \rangle = \int_{\partial\Omega} b^h \langle \phi_f | \vec{n} \rangle$ ).

### 4.3 Numerical results

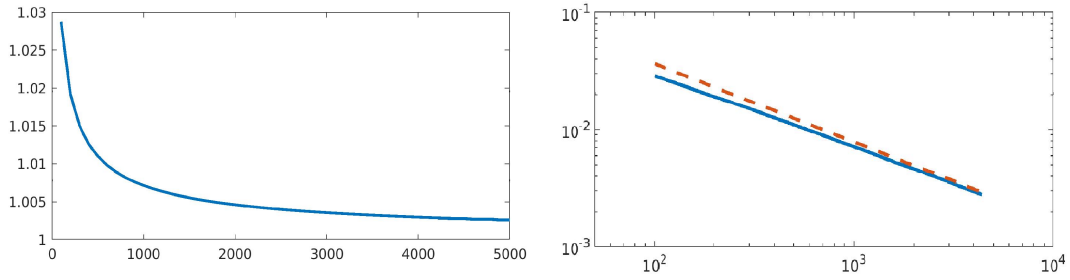


Figure 5:  $\bar{E}^h$  (plain) and  $h^{2/3}$  (dashed) as functions of  $N = \frac{1}{h}$  (Cartesian and log–log).

Numerical optimization of the one-dimensional problem 9 reveals that the  $O(h^{2/3})$  rate is almost observed in practice. In Figure 5 we plotted the value of the energy  $\bar{E}^h$  in the Dirichlet setting for  $N$  ranging in  $[100, 5000]$  with a stepsize of 100. The corresponding log – log graph exhibits an empirical convergence rate of  $h^\theta$  with  $\theta = 0.6240$ .

We present in Figure 6 the results for the denoising of a line, that is  $g = g_\nu$  in the Dirichlet setting for different orientations  $\nu$  and for the three total variations we considered: isotropic, Raviart-Thomas and “Condat”. We give also the results for the denoising of the circle we showed in the introduction, this time in the Neumann setting.

We see that the Raviart-Thomas TV performs as well as the TV of [19, 24]. However, it is important to notice that this good behavior relies heavily on the presence of the  $L^2$  term  $\|u - g\|_2^2$  in the problem we considered. Indeed, when tackling the inpainting

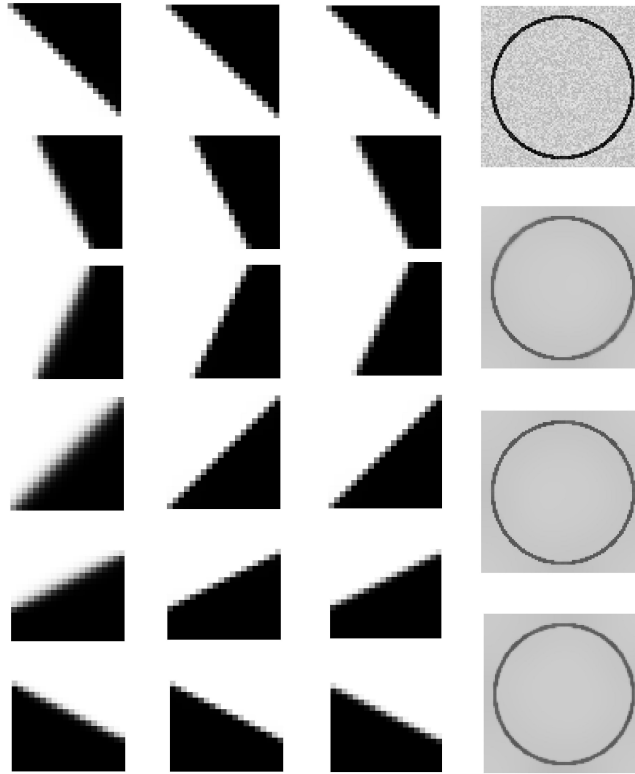


Figure 6: Denoising lines and a circle with isotropic (left col. and 2nd circle), RT (middle col. and 3rd circle) and Condat (right col. and 4th circle) total variations.

problem, that is the completion of a missing image (here, a plain discontinuity) from its boundary datum:

$$\arg \min_{\substack{u^h \in P0, \\ u^h|_B = g^h|_B}} \text{TV}_D^h(u^h) \quad (39)$$

where  $B$  denotes the  $4N - 4$  border pixels of our image, the Raviart-Thomas TV does *worse* than the isotropic TV, while the “Condat” TV still produces sharp discontinuities, see Figure 7.

## 5 Conclusion and perspectives

In this article we developed a study of the convergence rate of the discrete towards the continuous energies of the ROF model for two discretizations of the total variation. These two discrete TV, as well as the one introduced in [19, 24] can be united under the framework of constrained Raviart-Thomas fields. Future works include estimations on

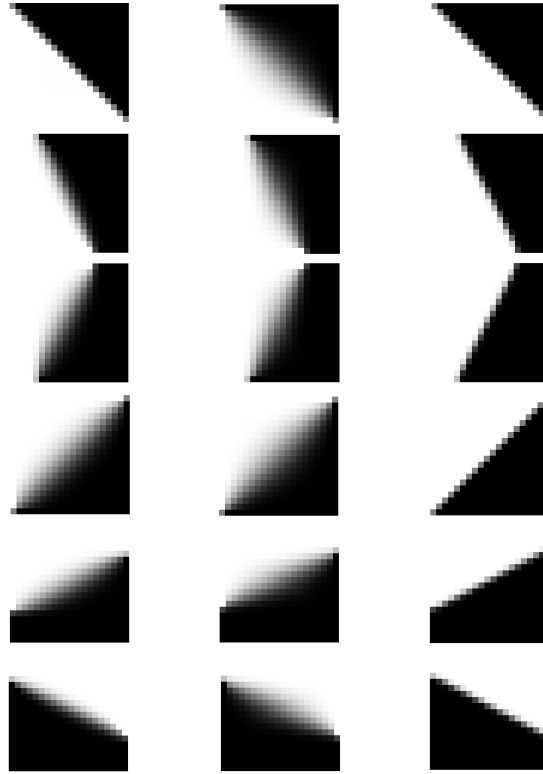


Figure 7: Inpainting with isotropic (left), RT (mid.) and Condat (right) total variations.

convergence of the minimizers  $\bar{u}^h$  towards  $\bar{u}$ , investigations on convergence rates for the inpainting problem (39), for “Condat” TV and for other directions in the isotropic TV.

## References

- [1] Luigi Ambrosio, Nicola Fusco, and Diego Pallara. *Functions of bounded variation and free discontinuity problems*. Oxford Mathematical Monographs. The Clarendon Press, Oxford University Press, New York, 2000.
- [2] Egil Bae, Jing Yuan, and Xue-Cheng Tai. Global minimization for continuous multiphase partitioning problems using a dual approach. *Int. J. Comput. Vis.*, 92(1):112–129, 2011.
- [3] Sören Bartels. Error control and adaptivity for a variational model problem defined on functions of bounded variation. *Math. Comp.*, 84(293):1217–1240, 2015.
- [4] Sören Bartels. Nonconforming discretizations of convex minimization problems and precise relations to mixed methods, 2020. preprint arXiv:2002.02359.
- [5] Sören Bartels. Error estimates for a class of discontinuous Galerkin methods for nonsmooth problems via convex duality relations. *Math. Comp.*, 90(332):2579–2602, 2021.

- [6] Sören Bartels, Ricardo H. Nochetto, and Abner J. Salgado. A total variation diminishing interpolation operator and applications. *Math. Comp.*, 84(296):2569–2587, 2015.
- [7] Sören Bartels and Zhangxian Wang. Orthogonality relations of Crouzeix-Raviart and Raviart-Thomas finite element spaces, 2020. preprint arXiv:2005.02741.
- [8] Sören Bartels, Robert Tovey, and Friedrich Wassmer. Singular solutions, graded meshes, and adaptivity for total-variation regularized minimization problems, 2021.
- [9] H. H. Bauschke and P. L. Combettes. *Convex analysis and monotone operator theory in Hilbert spaces*. CMS Books in Mathematics/Ouvrages de Mathématiques de la SMC. Springer, New York, 2011. With a foreword by Hedy Attouch.
- [10] Y. Boykov, V. Kolmogorov, D. Cremers, and A. Delong. An integral solution to surface evolution PDEs via Geo-Cuts. In A. Leonardis, H. Bischof, and A. Pinz, editors, *European Conference on Computer Vision (ECCV)*, volume 3953 of *LNCS*, pages 409–422, Graz, Austria, May 2006. Springer.
- [11] Antonin Chambolle, Vicent Caselles, Daniel Cremers, Matteo Novaga, and Thomas Pock. An introduction to total variation for image analysis. In *Theoretical foundations and numerical methods for sparse recovery*, volume 9 of *Radon Ser. Comput. Appl. Math.*, pages 263–340. Walter de Gruyter, Berlin, 2010.
- [12] Antonin Chambolle, Vicent Caselles, Daniel Cremers, Matteo Novaga, and Thomas Pock. An introduction to total variation for image analysis. *Theoretical foundations and numerical methods for sparse recovery*, 9(263-340):227, 2010.
- [13] Antonin Chambolle and Leonard Kreutz. Crystallinity of the homogenized energy density of periodic lattice systems, 2021.
- [14] Antonin Chambolle and Thomas Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *J. Math. Imaging Vision*, 40(1):120–145, 2011.
- [15] Antonin Chambolle and Thomas Pock. Crouzeix-Raviart approximation of the total variation on simplicial meshes. *J. Math. Imaging Vision*, 62(6-7):872–899, 2020.
- [16] Antonin Chambolle and Thomas Pock. Learning consistent discretizations of the total variation. *SIAM J. Imaging Sci.*, 14(2):778–813, 2021.
- [17] Antonin Chambolle and Thomas Pock. Approximating the total variation with finite differences or finite elements. In *Geometric partial differential equations. Part II*, volume 22 of *Handb. Numer. Anal.*, pages 383–418. Elsevier/North-Holland, Amsterdam, [2021] ©2021.
- [18] Antonin Chambolle, Pauline Tan, and Samuel Vaiter. Accelerated alternating descent methods for Dykstra-like problems. *J. Math. Imaging Vision*, 59(3):481–497, 2017.
- [19] Laurent Condat. Discrete total variation: new definition and minimization. *SIAM J. Imaging Sci.*, 10(3):1258–1290, 2017.
- [20] Camille Couprie, Leo Grady, Hugues Talbot, and Laurent Najman. Combinatorial continuous maximum flow. *SIAM J. Imaging Sci.*, 4(3):905–930, 2011.
- [21] P. Destuynder, M. Jaoua, and H. Sellami. A dual algorithm for denoising and preserving edges in image processing. *J. Inverse Ill-Posed Probl.*, 15(2):149–165, 2007.

- [22] Philippe Destuynder, Mohamed Jaoua, and Hela Sellami. An error estimate in image processing. *ARIMA Rev. Afr. Rech. Inform. Math. Appl.*, 15:61–81, 2012.
- [23] Marc Herrmann, Roland Herzog, Stephan Schmidt, José Vidal-Núñez, and Gerd Wachsmuth. Discrete total variation with finite elements and applications to imaging. *J. Math. Imaging Vision*, 61(4):411–431, 2019.
- [24] Michael Hintermüller, Carlos N. Rautenberg, and Jooyoung Hahn. Functional-analytic and numerical issues in splitting methods for total variation-based image reconstruction. *Inverse Problems*, 30(5):055014, 34, 2014.
- [25] Ming-Jun Lai, Bradley Lucier, and Jingyue Wang. *The Convergence of a Central-Difference Discretization of Rudin-Osher-Fatemi Model for Image Denoising*, pages 514–526. Springer Berlin Heidelberg, Berlin, Heidelberg, 2009.
- [26] Chang-Ock Lee, Eun-Hee Park, and Jongho Park. A finite element approach for the dual Rudin-Osher-Fatemi model and its nonoverlapping domain decomposition methods. *SIAM J. Sci. Comput.*, 41(2):B205–B228, 2019.
- [27] Jan Lellmann, Björn Lellmann, Florian Widmann, and Christoph Schnörr. Discrete and continuous models for partitioning problems. *Int. J. Comput. Vis.*, 104(3):241–269, 2013.
- [28] Matteo Negri. The anisotropy introduced by the mesh in the finite element approximation of the Mumford-Shah functional. *Numer. Funct. Anal. Optim.*, 20(9-10):957–982, 1999.
- [29] P. A. Raviart and J. M. Thomas. A mixed finite element method for 2-nd order elliptic problems. In Ilio Galligani and Enrico Magenes, editors, *Mathematical Aspects of Finite Element Methods*, pages 292–315, Berlin, Heidelberg, 1977. Springer Berlin Heidelberg.
- [30] R. T. Rockafellar. *Convex analysis*. Princeton Landmarks in Mathematics. Princeton University Press, Princeton, NJ, 1997. Reprint of the 1970 original, Princeton Paperbacks.
- [31] L. Rudin, S. J. Osher, and E. Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D*, 60:259–268, 1992. [also in *Experimental Mathematics: Computational Issues in Nonlinear Science* (Proc. Los Alamos Conf. 1991)].
- [32] Fabio Viola, Andrew W. Fitzgibbon, and Roberto Cipolla. A unifying resolution-independent formulation for early vision. In *2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, June 16-21, 2012*, pages 494–501, 2012.
- [33] Jingyue Wang and Bradley J. Lucier. Error bounds for finite-difference methods for Rudin-Osher-Fatemi image smoothing. *SIAM J. Numer. Anal.*, 49(2):845–868, 2011.