Monotone discretization of anisotropic differential operators using Voronoi's first reduction

Joseph Frédéric Bonnans, Guillaume Bonnet, Jean-Marie Mirebeau

Monotone discretization of anisotropic differential operators using Voronoi's first reduction

Frédéric Bonnans[*], Guillaume Bonnet[†], Jean-Marie Mirebeau[‡]

June 14, 2023

Abstract

We consider monotone discretization schemes, using adaptive finite differences on Cartesian grids, of partial differential operators depending on a strongly anisotropic symmetric positive definite matrix. For concreteness, we focus on a linear anisotropic elliptic equation, but our approach extends to divergence form or non-divergence form diffusion, and to a variety of first and second order Hamilton-Jacobi-Bellman PDEs. The design of our discretization stencils relies on a matrix decomposition technique coming from the field of lattice geometry, and related to Voronoi's reduction of positive quadratic forms. We show that it is efficiently computable numerically, in dimension up to four, and yields sparse and compact stencils. However, some of the properties of this decomposition, related with the regularity and the local connectivity of the numerical scheme stencils, are far from optimal. We thus present fixes and variants of the decomposition that address these defects, leading to stability and convergence results for the numerical schemes.

Keywords: Adaptive finite differences, Anisotropic elliptic equation, Hamilton-Jacobi equation, Selling decomposition, Voronoi's first reduction

1 Introduction

In this paper, we address a matrix decomposition problem arising in the design of monotone finite difference schemes for strongly anisotropic partial differential equations (PDEs),

[*]Paris-Saclay University, CNRS, CentraleSupélec, Inria, Laboratory of signals and systems, 91190 Gif-sur-Yvette, France

[†]Department of Mathematics, University of Maryland, College Park, Maryland 20742, USA

[‡]Paris-Saclay University, CNRS, ENS Paris-Saclay, Centre Borelli, 91190 Gif-sur-Yvette, France

The second author was partially supported by Fondo Sociale Europeo, Programma Operativo 2014/2020 – cod. reg. FP1956730001.

Mathematics Subject Classification: 65N06, 65N12, 35J70, 90C49

discretized on a Cartesian grid of dimension $d \leq 4$. As an application, we focus on a linear elliptic anisotropic PDE for concreteness, but our results are equally well suited to a variety of linear and non-linear PDEs arising in deterministic and stochastic control, or related with the Monge-Ampère equation, as discussed in Appendix A. The PDEs of interest are usually defined in terms of a field $\mathcal{D}$ of symmetric positive definite matrices, encoding the problem geometry, which is typically anisotropic and whose eigenvectors are not aligned with the discretization grid. A key step of the discretization is the pointwise decomposition of $\mathcal{D}$ as a positively weighted sum of rank one matrices with integer entries, see (6) below, from which a finite difference scheme can be devised. For that purpose, we leverage tools from lattice geometry known as Selling's decomposition and more generally Voronoi's first reduction of positive quadratic forms [30], following [6, 17, 23]. We investigate closely the properties of these matrix decompositions relevant to the implementation and analysis of the numerical schemes, including efficient computation, radius of the support, uniqueness and regularity of the coefficients, and the so-called *spanning* property which is related to the connectivity of the scheme stencils, and we propose two modified decompositions improving on these aspects.

Outline. We show in Section 1.1 how the monotone discretization of an anisotropic linear elliptic PDE can be related to a symmetric matrix decomposition problem, and how the properties of the decomposition relate with the stability and the convergence rate of the scheme solutions. We present in Section 1.2 some suitable matrix decompositions in dimension $d \leq 4$, leading to efficient numerical schemes and with a low numerical cost. The rest of this paper is devoted to proofs, whose organization is outlined in Sections 1.1 and 1.2, with the exception of Appendix A which discusses the discretization of other PDEs.

Contributions. We establish stability estimates and convergence rates for a monotone discretization of a linear anisotropic elliptic PDE, given a decomposition of the diffusion tensor field with suitable properties, see Theorems 1.3 and 1.4. Motivated by this application and others, we investigate the properties of a matrix decomposition related to Voronoi's first reduction of positive quadratic forms, including the radius of its support. Addressing some shortcomings of this construction, we present in particular a variant obeying a local connectivity property in dimension $d = 4$ in Theorem 1.6, and a variant with smooth coefficients in dimension $d = 2$ in Theorem 1.8.

Notations. Throughout this paper, we denote by $(b_1, \cdots, b_d)$ the canonical basis of $\mathbb{R}^d$, and let $\mathbb{1} := (1, \cdots, 1) \in \mathbb{R}^d$, where the dimension d is always clear from context; we also let $b_0 := -\mathbb{1}$ in such way that $b_0 + \cdots + b_d = 0$. The $d \times d$ identity matrix is denoted Id_d.

When presenting an estimate, the notation "$C = C(a, b, c, \cdots)$" means that C is a constant depending only on the specified parameters $a, b, c, \cdots$.

We denote by $\mathcal{Z}_d$ the collection of nonzero vectors of length $d \geq 1$ with integer entries, considered up to a global change of sign, and by Λ_d the collection of all non-negative maps

$\lambda : \mathcal{Z}_d \to [0, \infty[$ whose support $\mathrm{supp}(\lambda) := \{e \in \mathcal{Z}_d \mid \lambda^e \neq 0\}$ is finite:

$$\mathcal{Z}_d := (\mathbb{Z}^d \setminus \{0\})/\pm, \qquad \Lambda_d := \{\lambda = (\lambda^e)_{e\in\mathcal{Z}_d} \colon \mathcal{Z}_d \to [0,\infty[, \text{ finitely supported}\}. \tag{1}$$

In practice, we manipulate the elements $e \in \mathcal{Z}_d$ like regular vectors, but we take care to only involve them in symmetric expressions. The group of unimodular matrices of shape $d \times d$ is denoted

$$\mathrm{GL}(\mathbb{Z}^d) := \{A \in \mathbb{Z}^{d\times d} \mid |\det A| = 1\}. \tag{2}$$

We denote by $\mathcal{S}_d \supset \mathcal{S}_d^+ \supset \mathcal{S}_d^{++}$ the sets of symmetric, non-negative, and positive definite matrices respectively. Similarly $\mathbb{R}_+ := [0,\infty[$ and $\mathbb{R}_{++} :=]0,\infty[$ respectively denote non-negative and positive reals, and $\mathbb{Z}_+$ and $\mathbb{Z}_{++}$ denote non-negative and positive integers. Symmetric matrices are equipped with the Loewner order: given $M, M' \in \mathcal{S}_d$, one has $M \preceq M'$ (resp. $M \prec M'$) if $M' - M \in \mathcal{S}_d^+$ (resp. $M' - M \in \mathcal{S}_d^{++}$). For each $D \in \mathcal{S}_d^{++}$, we define a norm $\|\cdot\|_D$ on $\mathbb{R}^d$ and anisotropy ratio $\mu(D) \in [1,\infty[$ by

$$\|v\|_D := \sqrt{\langle v, Dv\rangle}, \qquad\qquad \mu(D) := \sqrt{\|D\|\|D^{-1}\|}. \tag{3}$$

Here and below we denote by $\langle v, w\rangle := v^\top w$ the Euclidean scalar product of $v, w \in \mathbb{R}^d$, by $|v|$ the Euclidean norm, and by $\|A\| := \max\{|Av| \mid |v| = 1\}$ the spectral norm of a matrix A. Any $D \in \mathcal{S}_d^{++}$ admits the eigenvalues $\lambda_{\max}(D) := \|D\|$ and $\lambda_{\min}(D) := \|D^{-1}\|^{-1}$.

1.1 Stability and convergence of an elliptic PDE discretization

We consider an anisotropic elliptic equation, with periodic boundary conditions, as a toy illustrative problem for our approach to monotone PDE discretization. A closely related numerical scheme is studied in [17], but the stability and convergence results Theorems 1.3 and 1.4 are new, see Section 5.2 and Appendix C for their respective proofs. Our discretization method equally applies, possibly more naturally, to non-linear PDEs related to deterministic or stochastic control and to the Monge-Ampère operator, but for concision this discussion is postponed to Appendix A.

Denote by $\mathbb{T} := \mathbb{R}/\mathbb{Z}$ the one dimensional torus, and by $\mathbb{T}_h := (h\mathbb{Z})/\mathbb{Z}$ its discretization where $h^{-1} \geq 2$ is an integer. Given a field of positive definite matrices $\mathcal{D} : \mathbb{T}^d \to \mathcal{S}_d^{++}$, continuously differentiable, we consider the (negated) elliptic PDE operator

$$\mathcal{L}u(x) := \mathrm{div}(\mathcal{D}(x)\nabla u(x)). \tag{4}$$

The considered discrete finite difference operator involves non-negative weights $(\lambda^e(x))_{x\in\mathbb{T}^d}^{e\in\mathcal{Z}_d}$, subject to assumptions discussed later in Definition 1.2, and is defined as

$$\mathcal{L}_h u_h(x) := \frac{1}{h} \sum_{\substack{e\in\mathcal{Z}_d \\ \sigma=\pm1}} \frac{\lambda^e(x) + \lambda^e(x+\sigma he)}{2} \frac{u_h(x+\sigma he) - u_h(x)}{h}. \tag{5}$$

We prove in Theorem 1.3 a coercivity property for the operator $\mathcal{L}_h$, and we establish in Theorem 1.4 some first and second order convergence rates of its solutions towards those of $\mathcal{L}$, under suitable assumptions. Note that similar properties can be established for more standard discretizations [19, §2.6.1], and they do not constitute the main feature of the proposed scheme $\mathcal{L}_h$, which is its non-negativity.

Remark 1.1 (Non-negativity). If $-\mathcal{L}u \geq 0$ on Ω in the weak sense[1], and $u \geq 0$ on $\partial\Omega$, for some $u \in H^1(\Omega)$ where Ω is a smooth and strict subdomain of $\mathbb{T}^d$, then $u \geq 0$ on Ω by the usual maximum principle for elliptic PDEs. Likewise, if $-\mathcal{L}_h u_h \geq 0$ on a subdomain $\Omega_h \subsetneq \mathbb{T}_h^d$, and $u_h \geq 0$ on $\mathbb{T}_h^d \setminus \Omega_h$, and Theorem 1.3 below applies, then $u_h \geq 0$ on Ω_h. Indeed, the corresponding linear system is defined by an M-matrix (by (5) this matrix has a positive and dominant diagonal, non-positive off-diagonal elements, and by the coercivity estimate (9) it is invertible). Non-negative schemes are a prerequisite in applications such as geodesic distance computation using the Varadhan formula [5, 11], since $\ln(u_h)$ is considered. Discretizations of other PDEs, based on the same principles, and enjoying similar monotonicity properties, are presented in Appendix A.

The discretization (5) is well behaved when its coefficients obey the following properties.

Definition 1.2. A family of coefficients $\lambda : X \to \Lambda_d$, denoted $\lambda = (\lambda^e(x))_{x\in X}^{e\in\mathcal{Z}_d}$ and where X is a metric space (or an open subset of $\mathbb{R}^d$ for the grad-Lipschitz property), is said to be:

- $\mathcal{D}$-Consistent, where $\mathcal{D} : X \to \mathcal{S}_d^{++}$ is a positive definite tensor field, if for all $x \in X$
$$\mathcal{D}(x) = \sum_{e\in\mathcal{Z}_d} \lambda^e(x) e e^\top. \tag{6}$$
- R-Supported, if $r(x) \leq R$ for all $x \in X$, where
$$r(x) := \max\{|e| \mid e \in \mathcal{Z}_d,\ \lambda^e(x) > 0\}.$$
- K-Lipschitz (resp. K-grad-Lipschitz), if for all $x, y \in X$ one has
$$|\lambda^e(x) - \lambda^e(y)| \leq K|x-y|, \qquad \big(resp.\ |\nabla\lambda^e(x) - \nabla\lambda^e(y)| \leq K|x-y|\big).$$
- ε-Spanning, if for all $x \in X$ there exists $e_1, \cdots, e_d \in \mathcal{Z}_d$ such that
$$|\det(e_1, \cdots, e_d)| = 1, \qquad \min\{\lambda^{e_1}(x), \cdots, \lambda^{e_d}(x)\} \geq \varepsilon. \tag{7}$$

One of the main objectives of this paper, postponed to Section 1.2 below, is to propose practical and numerically efficient methods for constructing coefficients λ obeying the above properties. For now, we discuss Definition 1.2 and contrast our approach with two-scales discretizations of PDEs, see [13, 27].

[1] In other words $\int_\Omega \langle \nabla\varphi(x), D(x)\nabla u(x)\rangle \mathrm{d}x \geq 0$ for any continuously differentiable and non-negative test function φ whose support is contained in Ω.

- $\mathcal{D}$-Consistency is a qualitative property, which ensures the second order consistency of the scheme $\mathcal{L}_h$ with $\mathcal{L}$ (assuming in addition the grad-Lipschitz property), which can be verified by a Taylor expansion as in the proof of Proposition C.15 below. Alternatively, other works consider a variant of (6) featuring a consistency error, at the price of a more complex numerical analysis. For instance two-scales discretizations [13, 27] feature such an error, depending on an intermediate scale satisfying $h \ll k \ll 1$, and vanishing as $k \to 0$. A consistency error is also unavoidable if one addresses rank deficient semi-definite diffusion tensors [24], unless their kernel is spanned by vectors with integer entries.

- R-Support is a quantitative property, controlling the effective discretization scale $k = Rh$ of the numerical scheme. The constructions proposed in this paper obey $r(x) \leq C\mu(\mathcal{D}(x))$ where $C = C(d)$, see Theorem 1.6. The radius R is therefore controlled, for our numerical scheme, by the square root of the maximal condition number (3) of the tensor field $\mathcal{D}$, and the effective scale is proportional to the grid scale. In contrast, two scales discretizations of PDEs involve an effective discretization scale k which decreases sub-linearly with h, for instance $k = h^{\frac{2}{5}}$ is optimal in [13], and for this reason they suffer from reduced convergence rates.

- K-Lipschitz and K-grad-Lipschitz are regularity properties. To the knowledge of the authors, all practical finite difference schemes based on adaptive tensor decompositions such as (6) previously proposed in the litterature, feature coefficients with Lipschitz regularity, but *not better*, see [6, 13, 17, 36]. This is often sufficient to establish the convergence of the numerical methods, yet improved convergence rates may be established if the coefficients have a higher order regularity, see for instance Theorem 1.4 and [21, 34] which are also discussed in Appendix A. In this paper, we present a matrix decomposition with Lipschitz coefficients in dimension $d \leq 4$, and a smooth (hence grad-Lipschitz) decomposition in dimension $d = 2$, see Theorems 1.6 and 1.8 below respectively.

- The ε-spanning assumption ensures that the graph underlying the numerical scheme is locally connected, see Section 5. If this was not the case, then the sub-grids of $h\mathbb{Z}^d$ corresponding (for instance) to the points with an even or odd sum of coordinates could be disconnected, leading to *chessboard artifacts*, see (36). In this paper, we establish regularity properties of the solutions of discretized PDEs under this assumption, ruling out these artifacts, see Theorem 1.3.

We further motivate Definition 1.2 by establishing stability and convergence results under these assumptions for the numerical scheme (5). For that purpose, define

$$\mathcal{Q}_h(u) := h^d \sum_{x \in \mathbb{T}_h^d} \mathcal{Q}_h^x(u), \qquad \text{where } \mathcal{Q}_h^x(u) := \frac{1}{2} \sum_{\substack{e \in \mathcal{Z}^d \\ \sigma = \pm 1}} \lambda^e(x) \Big(\frac{u(x + \sigma h e) - u(x)}{h} \Big)^2, \tag{8}$$

for any $u : \mathbb{T}_h^d \to \mathbb{R}$ and any $x \in \mathbb{T}_h^d$, in such way that $-\langle \mathcal{L}_h u, u\rangle = \mathcal{Q}_h(u)$. Our first result is a coercivity estimate relating the discrete elliptic energy Q_h to the discrete L_h^2 norm of the discrete gradient vector $\nabla_h u(x) \in \mathbb{R}^d$, where $\|u\|_{L_h^2}^2 := \langle u, u\rangle_{L_h^2}$ and

$$\nabla_h u(x) := \Big(\frac{u(x + hb_i) - u(x)}{h}\Big)_{1\le i\le d}, \qquad \langle u, v\rangle_{L_h^2} := h^d \sum_{x\in\mathbb{T}_h^d} u(x)v(x).$$

Strictly speaking, (9) is only a *semi*-coercivity estimate, since constant functions are in the kernel of $\mathcal{L}_h$, but the *semi-* prefix is dropped in the text for concision and readability.

Theorem 1.3 (Coercivity of the discrete elliptic energy)**.** *Consider weights $\lambda : \mathbb{T}^d \to \Lambda_d$ which are bounded by $\lambda_{\max}$, R-supported, K-Lipschitz, and ε-spanning, for some constants $\lambda_{\max}, R, K, \varepsilon > 0$. Then for all $0 < h \le h_0$ and all $u : \mathbb{T}_h^d \to \mathbb{R}$ one has*

$$c\|\nabla_h u\|_{L_h^2}^2 \le Q_h(u) \le C\|\nabla_h u\|_{L_h^2}^2, \tag{9}$$

where the constants C, $c > 0$ and $h_0 > 0$ only depend on $(\lambda_{\max}, R, K, \varepsilon)$.

We next establish a convergence rate for the scheme solutions, with periodic boundary conditions for simplicity. Let us acknowledge here that adapting the scheme to maintain similar convergence rates on a bounded domain, with e.g. Dirichlet or Neumann boundary conditions, is a non-trivial problem especially in the case of wide stencil schemes such as (5), which we regard as an opportunity for future work. We denote by T_h^1 the convolution operator with the indicator function of the cube $[-h/2, h/2]^d$ of width h.

Theorem 1.4 (Convergence rate, elliptic equation)**.** *Consider a diffusion tensor field $\mathcal{D} : \mathbb{T}^d \to \mathcal{S}_d^{++}$ and coefficients $\lambda : \mathbb{T}^d \to \Lambda_d$ obeying the $\mathcal{D}$-consistency, R-support, K-Lipschitz (resp. K-grad-Lipschitz) and ε-spanning properties. Consider a r.h.s. $\mathrm{f} \in L^2(\mathbb{T}^d)$, with zero mean, and let $\mathrm{f}_h := \mathrm{T}_h^1 \mathrm{T}_h^1 \mathrm{f}$. Denote by $u : \mathbb{T}^d \to \mathbb{R}$ and $u_h : \mathbb{T}_h^d \to \mathbb{R}$ a solution to $\mathcal{L}u = \mathrm{f}$ and $\mathcal{L}_h u_h = \mathrm{f}_h$ respectively. If $\|\nabla^2 u\|_{L^2}$ (resp. $\|\nabla^3 u\|_{L^2}$) is finite, then*

$$\|\nabla_h(\mathrm{T}_h^1 u - u_h)\|_{L_h^2} \le Ch\|\nabla^2 u\|_{L^2}. \qquad \big(\textit{resp. } \|\nabla_h(\mathrm{T}_h^1 u - u_h)\|_{L_h^2} \le Ch^2\|\nabla^3 u\|_{L^2},\big)$$

for all $0 < h \le h_0$, where $h_0 > 0$ and C only depend on $(\|\mathcal{D}\|_\infty, R, K, \varepsilon)$.

1.2 Voronoi's decomposition and variants

The discretization of anisotropic differential operators by finite differences naturally leads to a matrix decomposition problem, as illustrated in Section 1.1 and Appendix A, whose coefficients should obey a number of properties, summarized in Definition 1.2. In the following, we describe an efficient method for computing such decompositions, leveraging

(variants of) a tool from discrete geometry known as Voronoi's first reduction of quadratic forms [30]. For that purpose, define for all $D \in \mathcal{S}_d^{++}$

$$\Lambda(D) := \operatorname*{argmax}_{\lambda \in \Lambda_d} \Big\{ \sum_{e \in \mathcal{Z}_d} \lambda^e \mid \sum_{e \in \mathcal{Z}_d} \lambda^e e e^\top = D \Big\}. \tag{10}$$

In other words, $\Lambda(D)$ collects the decompositions of the symmetric positive definite matrix D whose offsets $e \in \mathcal{Z}_d$ have integer entries, and whose sum of weights λ^e is maximal. Note that the definition (1) of the search space Λ_d requires that the coefficients $\lambda^e \geq 0$ are non-negative (and finitely supported), so that (10) has the structure of a linear program with infinitely many unknowns and constraints.

Voronoi's first reduction [30] is classically defined as the dual linear program to (10), see Section 2. It benefits from invariances and symmetries, under the group $\mathrm{GL}(\mathbb{Z}^d)$ of unimodular changes of coordinates (2), which have been extensively studied and enable in particular the classification of the vertices of the skeletal structure of the associated polyhedron in dimension $d \leq 8$ [8, 33]. The following result describes the set (10) in dimension $d \leq 4$, and specifies an element $\lambda(D)$ within it.

Proposition 1.5. *For each $D \in S_d^{++}$, $2 \leq d \leq 4$, we have the following description of the set $\Lambda(D)$, within which we select the following element $\lambda(D)$:*

- *If $d \in \{2, 3\}$, then $\Lambda(D)$ is a singleton, and $\lambda(D)$ is defined as its unique element.*
- *If $d = 4$, then $\Lambda(D)$ is either a singleton or an equilateral triangle, and $\lambda(D)$ is defined as either its unique element or its barycenter.*

We emphasize that Proposition 1.5 is completely practical, in the sense that $\lambda(D)$ can be computed in a fast and reliable manner. The computation of $\lambda(D)$ amounts to the solution of a linear program, within a polyhedron whose vertices (known as *perfect forms*) are extensively classified in the dimensions d of interest, which is used to speed up the computation, see Remark 3.1. A pseudo-code is presented in Section 3, see Algorithms 1 and 2 and Propositions 3.5 and 3.6, and numerical codes are provided[2], extending the works [6, 17]. Recall that the anisotropy ratio $\mu(D) := \sqrt{\|D\|\|D^{-1}\|}$ is defined as the square root of the condition number of $D \in \mathcal{S}_d^{++}$.

For concreteness, we illustrate on Fig. 1 the coefficients $(\lambda(D_t))_{t \in [0,1]}$ defined by Proposition 1.5 and Theorem 1.8 below, where $D_t := (1-t)D_0 + tD_1$ interpolates between two randomly chosen $D_0, D_1 \in \mathcal{S}_d^{++}$. The main purpose of Fig. 1 is to illustrate the Lipschitz regularity of the coefficient $t \in [0,1] \mapsto \lambda^e(D_t)$ from Proposition 1.5, and the smoothness of the coefficient defined by Theorem 1.8, for any $e \in \mathcal{Z}_d$. One can also visualize on this example the set $\{e \in \mathcal{Z}_d \mid \lambda^e(D_t) > 0\}$, for any given $t \in [0,1]$. The R-support and ε-spanning properties, established in the following result, can be visually checked by observing that this set is contained in a ball of small radius and contains a basis of $\mathbb{Z}^d$.

[2] `www.github.com/Mirebeau/AdaptiveGridDiscretizations`

Theorem 1.6. *The mapping $\lambda : S_d^{++} \to \Lambda_d$, defined in Proposition 1.5 and where $2 \leq d \leq 4$, obeys the following properties. For all $D, D' \in \mathcal{S}_d^{++}$, denoting $\mu := \max\{\mu(D), \mu(D')\}$,*

- *Consistency, in the sense that $D = \sum_{e \in \mathcal{Z}_d} \lambda^e(D) e e^\top$.*
- *$R(\mu)$-Support, in the sense that $\|e\| \leq R(\mu)$ for all $e \in \mathrm{supp}(\lambda(D))$.*
- *$K(\mu)$-Lipschitz, in the sense that $|\lambda^e(D) - \lambda^e(D')| \leq K(\mu)\|D - D'\|$ for all $e \in \mathcal{Z}_d$.*
- *ε-Spanning, in the sense that in there exists $e_1, \cdots, e_d \in \mathcal{Z}_d$ such that $|\det(e_1, \cdots, e_d)| = 1$ and $\min\{\lambda^{e_1}(D), \cdots, \lambda^{e_d}(D)\} \geq \varepsilon \lambda_{\min}(D)$.*

We denoted $R(\mu) := C\mu$, $K(\mu) := C\mu^2$ with constants $C = C(d)$ and $\varepsilon = \varepsilon(d) > 0$.

The $R(\mu)$-support estimate is actually established in arbitrary dimension, see Theorem 4.3, with explicit (but not sharp) constants (33) in dimension $d \leq 4$. It was previously only known in dimension $d \in \{2, 3\}$ [22], and in arbitrary dimension d with a sub-optimal estimate $R(\mu) = C\mu^{d-1}$ [23, Proposition 1.1]. The Lipschitz and spanning properties were established in dimension $d \in \{2, 3\}$ in [5]. The definition of $\lambda(D)$ in dimension $d = 4$ from Proposition 1.5 is original to our knowledge, and ensures that those coefficients are uniquely defined and obey the Lipschitz and spanning properties of Theorem 1.6. By an immediate composition argument, we obtain the properties of Definition 1.2.

Corollary 1.7. *Let $\mathcal{D} : X \to \mathcal{S}_d^{++}$ be Lipschitz, bounded, and of bounded condition number, where $2 \leq d \leq 4$ and X is a metric space. Define $\lambda^e(x) := \lambda^e(\mathcal{D}(x))$, for all $x \in X$, $e \in \mathcal{Z}_d$. Then $\lambda : X \to \Lambda_d$ is $\mathcal{D}$-consistent, R-supported, K-Lipschitz, and ε-spanning, in the sense of Definition 1.2. The constants R, K, and $\varepsilon > 0$, only depend on $\|\mathcal{D}\|_\infty$, $\|\mu(\mathcal{D})\|_\infty$, and the Lipschitz constant of $\mathcal{D}$.*

A shortcoming of the decomposition of Proposition 1.5, is that $D \in \mathcal{S}_d^{++} \mapsto \lambda^e(D)$ has Lipschitz regularity but not better, for any $e \in \mathcal{Z}_d$. This decomposition coefficient is in fact piecewise linear w.r.t. D, a property inherited from the structure of Voronoi's first reduction which is a linear program. For this reason we introduce an alternative smooth decomposition, in dimension $d = 2$.

Theorem 1.8. *There is a computable decomposition $\tilde{\lambda} \in C^\infty(\mathcal{S}_2^{++}; \Lambda_2)$ which is consistent, $C\mu$-supported, $C\mu^2$-Lipschitz, and ε-spanning, for some $C, \varepsilon > 0$ with the notations of Theorem 1.6. It is also $C\mu^4/\lambda_{\min}$-grad-Lipschitz, in the sense that $\|\nabla^2 \tilde{\lambda}^e(D)\| \leq C\mu(D)^4/\lambda_{\min}(D)$, for all $e \in \mathcal{Z}^d$, where the hessian is defined over the open subset $\mathcal{S}_d^{++} \subset \mathcal{S}_d$.*

Remark 1.9 (Dimensions 5 and 6). For a matrix $D \in \mathcal{S}_d^{++}$ of arbitrary dimension d, the set $\Lambda(D)$ is a finite dimensional compact convex polytope. A decomposition $\lambda(D) \in \Lambda(D)$ may be computed at a reasonable numerical cost if $d \in \{5, 6\}$, using a direct extension of the techniques presented in this paper. An important caveat, however, is that no selection

principle of $\lambda(D) \in \Lambda(D)$ is able to ensure the spanning property, see Proposition 5.5. Dimension $d = 5$ may be relevant in some medical data processing techniques [15], which involve solving anisotropic PDEs on $\mathbb{R}^3 \times \mathbb{S}^2$, whereas dimension $d = 6$ may open applications to three-dimensional elasticity, whose anisotropy coefficients are gathered in a six-dimensional symmetric positive definite Hooke tensor. We regard those potential applications are opportunities for future work.

Organization. After some discussion of Voronoi's first reduction in Section 2, we prove Proposition 1.5 in Section 3. The parts of Theorem 1.6 related to Lipschitz regularity, support radius, and the spanning property, are established in Sections 3 to 5 respectively. Theorem 1.8 is proved in Section 6. Finally, we discuss in Appendix B the defects of some apparently straightforward constructions of smooth and spanning decompositions.

2 Voronoi's first reduction of positive quadratic forms

Voronoi's first reduction [35] is a tool from the field of lattice geometry [30], with applications in sphere packing, arithmetic, and PDE discretizations in this paper and [23]. In this section, we show its duality with the matrix decomposition problem (10) in Proposition 2.3, and we present a (known) structural result in dimension $d \leq 4$ in Proposition 2.8. Voronoi's first reduction was originally intended as a tool for classifying positive quadratic forms up to *arithmetical equivalence*.

Definition 2.1 (Arithmetical equivalence)**.** Two matrices M, $M' \in \mathcal{S}_d$ are said arithmetically equivalent if there exists $A \in \mathrm{GL}(\mathbb{Z}^d)$ such that $M' = A^\top M A$.

Voronoi's first reduction $\mathrm{Vor}(D)$ of $D \in \mathcal{S}_d^{++}$ is defined similarly to a linear program, although with infinitely many constraints. Its modern presentation involves an auxiliary object $\mathcal{M}_d \subset \mathcal{S}_d$, referred to as Ryskov's polyhedron:

$$\mathrm{Vor}(D) := \min_{M \in \mathcal{M}_d} \mathrm{Tr}(DM), \qquad \mathcal{M}_d := \{M \in \mathcal{S}_d \mid \forall e \in \mathcal{Z}_d,\, \langle e, Me\rangle \geq 1\}. \tag{11}$$

The optimization problem (11) is well-posed, as proved by Voronoi himself [30, 35].

Theorem 2.2 (Voronoi)**.** *Ryskov's polyhedron is a subset of $\mathcal{S}_d^{++}$, on which the determinant is positively bounded below. It is a locally finite polyhedron, in the sense that finitely many constraints are active locally in the neighborhood of any point. It has finitely many equivalence classes of vertices for the relation of arithmetical equivalence. The linear program* $\mathrm{Vor}(D)$ *is well-posed in the sense that the collection of minimizers of* (11) *is non-empty and compact for any $D \in \mathcal{S}_d^{++}$.*

For any $M \in \mathcal{S}_d^{++}$ we define the sets

$$\mathcal{S}^{++}(M) := \{D \in \mathcal{S}_d^{++} \mid \mathrm{Vor}(D) = \mathrm{Tr}(DM)\}, \quad \Xi(M) := \{e \in \mathcal{Z}_d \mid \langle e, Me\rangle \leq 1\}. \tag{12}$$

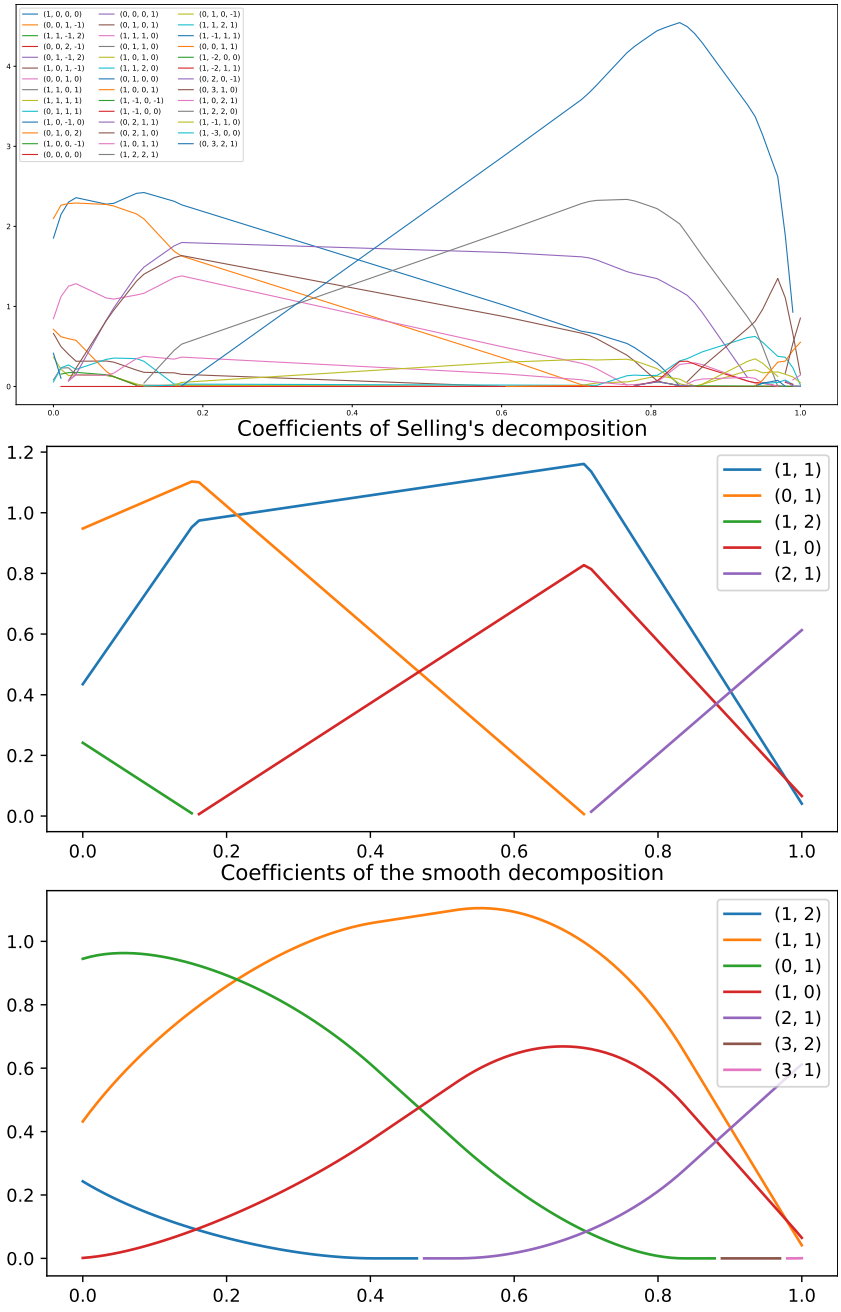

Figure 1: Illustration of the proposed decompositions of $D(t) = (1-t)D_0 + tD_1$, where $D_0, D_1 \in S_d^{++}$ are chosen arbitrarily. The curves represent coefficients $t \in [0,1] \mapsto \lambda^e(D(t))$, where the vector $e \in \mathcal{Z}_d$ is indicated in the legend. Top : $d = 4$, decomposition of Proposition 1.5. Middle : $d = 2$, Selling's decomposition, also used in Proposition 1.5. Bottom : $d = 2$, with the alternative smooth decomposition of Theorem 1.8.

If M belongs to Ryskov's polyhedron $\mathcal{M}_d$, then $\mathcal{S}^{++}(M)$ collects all matrices $D \in \mathcal{S}_d^{++}$ for which M is optimal in (11, left), whereas $\Xi(M)$ denotes the set of active constraints in (11, right). Note that $\mathcal{S}^{++}(M)$ is convex, and that $\Xi(M)$ is finite by Theorem 2.2.

We establish below some duality relations between the linear program (10) defining our discretization and Voronoi's first reduction (11).

Proposition 2.3. *Let $M \in \mathcal{M}_d$ and let $D \in \mathcal{S}^{++}(M)$. Then the set $\Lambda(D)$ of maximizers in (10) is a nonempty convex compact polytope characterized by*

$$\Lambda(D) = \Big\{ \lambda \in \Lambda_d \;\Big|\; \mathrm{supp}(\lambda) \subset \Xi(M), \sum_{e \in \Xi(M)} \lambda^e e e^\top = D \Big\}. \tag{13}$$

Proof. For now, we waive the constraint that λ is finitely supported in (10), and we define

$$\Lambda'(D) := \operatorname*{argmax}_{\lambda \in l^1_w(\mathcal{Z}_d)} \Big\{ \sum_{e \in \mathcal{Z}_d} \lambda^e \;\Big|\; \lambda \succeq 0, \sum_{e \in \mathcal{Z}_d} \lambda^e e e^\top = D \Big\} = \operatorname*{argmax}_{\lambda \in l^1_w(\mathcal{Z}_d)} (-f(\lambda) - g(A\lambda)), \tag{14}$$

where the vector space $l^1_w(\mathcal{Z}_d) := \{\lambda \colon \mathcal{Z}_d \to \mathbb{R} \mid |\lambda|_{l^1_w} < +\infty\}$ is equipped with the norm $|\cdot|_{l^1_w} : \lambda \mapsto \sum_{e \in \mathcal{Z}_d} |e|^2 |\lambda^e|$, and where

$$f(\lambda) := \chi_{\{\lambda \succeq 0\}} - \sum_{e \in \mathcal{Z}_d} \lambda^e, \qquad A\lambda := \sum_{e \in \mathcal{Z}_d} \lambda^e e e^\top, \qquad g(P) := \chi_{\{P = D\}}.$$

We denoted by χ_A the characteristic function of a set A, i.e. $\chi_A(x) = 0$ if $x \in A$ and $\chi_A(x) = \infty$ otherwise. The choice of the norm $|\cdot|_{l^1_w}$ is justified by the fact that any admissible λ in (14) satisfies

$$|\lambda|_{l^1_w} = \mathrm{Tr}\Big(\sum_{e \in \mathcal{Z}_d} \lambda^e e e^\top \Big) = \mathrm{Tr}(D) < +\infty. \tag{15}$$

The dual optimization problem to (14), in the sense of Fenchel's duality theorem [2, Theorem 1.113], is defined as $\mathrm{argmin}_{M \in \mathcal{S}_d} f^*(-A^\top M) + g^*(M)$ where

$$f^*(\mu) = \chi_{\mu \preceq -1}, \qquad A^\top M = (\langle e, Me \rangle)_{e \in \mathcal{Z}_d}, \qquad g^*(M) = \mathrm{Tr}(DM).$$

Note that the characteristic function f is defined over $l^1_w(\mathcal{Z}_d)^* = l^\infty_w(\mathcal{Z}_d)$, whose norm reads $|\mu|_{l^\infty_w} := \sup\{e \in \mathcal{Z}_d \mid |\mu(e)|/|e|^2\}$. We recognize the minimization problem (11). The duality gap is always non-negative, hence if $M \in \mathcal{M}_d$ is optimal in (11), then for any $\lambda \in l^1_w$ admissible in (14),

$$0 \leq \mathrm{Vor}(D) - \sum_{e \in \mathcal{Z}_d} \lambda^e = \mathrm{Tr}(DM) - \sum_{e \in \mathcal{Z}_d} \lambda^e = \sum_{e \in \mathcal{Z}_d} \lambda^e (\langle e, Me \rangle - 1). \tag{16}$$

Moreover, the constraint qualification condition $0 \in \mathrm{int}(\mathrm{dom}\, g - A\,\mathrm{dom}\, f)$ (equivalently $D \in \mathrm{int}(A\,\mathrm{dom}\, f)$, where $A\,\mathrm{dom}\, f = \{\sum_{e \in \mathcal{Z}^d} \lambda^e e e^\top \mid \lambda \in l^1_w(\mathcal{Z}_d),\ \lambda \succeq 0\}$) is satisfied, since

D may be approximated by symmetric positive definite matrices with rational eigenvectors. Therefore the inequality in (16) is an equality if and only if $\lambda \in \Lambda'(D)$. Using that all terms in the right-hand side of (16) are non-negative, we deduce that an admissible λ in (14) belongs to $\Lambda'(D)$ if and only if it is supported on $\Xi(M)$. In particular, any $\lambda \in \Lambda'(D)$ is finitely supported, thus $\Lambda(D) = \Lambda'(D)$ and (13) holds. The compactness of $\Lambda(D)$ follows from the fact that any λ in the finite-dimensional set $\Lambda(D)$ satisfies (15). □

Proposition 2.3 above admits a converse, Proposition 2.4, implying that for all $M \in \mathcal{M}_d$

$$\mathcal{S}^{++}(M) = \mathcal{S}_d^{++} \cap \Big\{ \sum_{e \in \mathcal{Z}_d} \lambda^e e e^\top \;\Big|\; \lambda : \mathcal{Z}_d \to [0, \infty[,\ \mathrm{supp}(\lambda) \subset \Xi(M) \Big\}. \tag{17}$$

Proposition 2.4. *Let $D \in \mathcal{S}_d^{++}$ and $M \in \mathcal{M}_d$. Assume that $D = \sum_{e \in \Xi(M)} \lambda^e e e^\top$ and that $\mathrm{supp}(\lambda) \subset \Xi(M)$, for some $\lambda \in \Lambda_d$. Then $D \in \mathcal{S}^{++}(M)$ and $\lambda \in \Lambda(D)$.*

Proof. Let $M' \in \mathcal{M}_d$. Then, for any $e \in \Xi(M)$, one has $\langle e, M'e \rangle \geq 1 = \langle e, Me \rangle$. It follows that

$$\mathrm{Tr}(DM') = \sum_{e \in \Xi(M)} \lambda^e \langle e, M'e \rangle \geq \sum_{e \in \Xi(M)} \lambda^e \langle e, Me \rangle = \mathrm{Tr}(DM).$$

Thus M is optimal in (11), and we deduce using Proposition 2.3 that $\lambda \in \Lambda(D)$. □

In order to proceed with the proof of Theorem 1.6, we need a more precise description of Ryskov's polyhedron $\mathcal{M}_d$. The vertices of $\mathcal{M}_d$ are classically called *perfect forms* [35], and we denote their set as

$$\mathrm{Perfect}(d) := \big\{ M \in \mathcal{M}_d \mid \mathrm{Span}_{\mathbb{R}} \{ e e^\top \mid e \in \Xi(M) \} = \mathcal{S}_d \big\}.$$

For any perfect form $M \in \mathrm{Perfect}(d)$, we denote by

$$\mathcal{N}(M) := \big\{ M' \in \mathrm{Perfect}(d) \mid \dim(\mathrm{Span}_{\mathbb{R}} \{ e e^\top \mid e \in \Xi(M) \cap \Xi(M') \}) = d(d+1)/2 - 1 \big\},$$

the collection of neighbor vertices of M in $\mathcal{M}_d$, where $d(d+1)/2 = \dim(\mathcal{S}_d)$. Note that

$$\mathcal{S}^{++}(M) = \{ D \in \mathcal{S}_d^{++} \mid \mathrm{Tr}(DM) \leq \mathrm{Tr}(DM'),\ \forall M' \in \mathcal{N}(M) \}, \tag{18}$$

by the usual optimality condition in linear programs. The polyhedral structure of $\mathcal{M}_d$ is compatible with the relation of arithmetical equivalence defined in Definition 2.1:

Proposition 2.5. *If $M \in \mathrm{Perfect}(d)$ and $A \in \mathrm{GL}(\mathbb{Z}^d)$, then $A^\top M A \in \mathrm{Perfect}(d)$ and*

$$\Xi(A^\top M A) = \{ A^{-1} e \mid e \in \Xi(M) \}, \qquad \mathcal{N}(A^\top M A) = \{ A^\top M' A \mid M' \in \mathcal{N}(M) \}.$$

Proof. This follows directly from the definitions of $\mathcal{M}_d$, $\Xi(M)$, and $\mathcal{N}(M)$, and from the fact that for any $A \in \mathrm{GL}(\mathbb{Z}^d)$ one has $\{ Ae \mid e \in \mathcal{Z}_d \} = \mathcal{Z}_d$. □

The classification of perfect forms up to arithmetical equivalence is a classical problem in lattice geometry [8], whose complexity explodes as dimension increases, see [33] for the latest complete classification in dimension $d = 8$. The discussion is fortunately simpler in dimension $d \leq 4$, and the only two relevant perfect forms are described in Propositions 2.6 and 2.7. There is a canonical perfect form, existing in arbitrary dimension d, and defined as follows: denoting $\mathbb{1} := (1, \dots, 1) \in \mathbb{Z}^d$,

$$\boldsymbol{A}_d := \frac{1}{2}(\mathrm{Id}_d + \mathbb{1}\mathbb{1}^\top) = \frac{1}{2}\begin{pmatrix} 2 & 1 & \dots & 1 \\ 1 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 1 \\ 1 & \dots & 1 & 2 \end{pmatrix}. \tag{19}$$

The following identity will be useful: for any $D \in S_d$ with coefficients $(D_{ij})_{i,j=1}^d$, one has

$$D = \sum_{1 \leq i \leq d} \Big(\sum_{1 \leq j \leq d} D_{ij} \Big) b_i b_i^\top - \sum_{1 \leq i < j \leq d} D_{ij} (b_i - b_j)(b_i - b_j)^\top. \tag{20}$$

The union of two sets X and Y known to be disjoint is denoted $X \sqcup Y$.

Proposition 2.6. *The matrix $\boldsymbol{A}_d$ is a perfect form, in any dimension $d \geq 1$, and*

$$\Xi(\boldsymbol{A}_d) = \{\pm b_i \mid 1 \leq i \leq d\} \sqcup \{\pm(b_i - b_j) \mid 1 \leq i < j \leq d\}. \tag{21}$$

Proof. Let $e \in \mathcal{Z}_d$ be such that $1 \geq \langle e, \boldsymbol{A}_d e \rangle = (|e|^2 + \langle e, \mathbb{1} \rangle^2)/2$. Then $|e|^2 \leq 2$, and therefore e has either one or two nonzero components, equal to ± 1. In the latter case these components have opposite sign, since $\langle e, \mathbb{1} \rangle^2 = 0$. This establishes (21) and the fact that $\boldsymbol{A}_d \in \mathcal{M}_d$. Finally (20) shows that $\mathrm{Span}_{\mathbb{R}}\{ee^\top \mid e \in \Xi(\boldsymbol{A}_d)\} = \mathcal{S}_d$, hence $\boldsymbol{A}_d \in \mathrm{Perfect}(d)$. □

In dimension $d = 4$, the following is also a perfect form, which is not arithmetically equivalent to $\boldsymbol{A}_4$ since it does not have the same determinant:

$$\boldsymbol{D}_4 := \frac{1}{2}\begin{pmatrix} 2 & 1 & 1 & 0 \\ 1 & 2 & 1 & 1 \\ 1 & 1 & 2 & 1 \\ 0 & 1 & 1 & 2 \end{pmatrix}. \tag{22}$$

The notations $\boldsymbol{A}_d$ and $\boldsymbol{D}_4$ are customary, see [8] for a more general classification and the relation with the classification of root lattices.

Proposition 2.7. *The matrix $\boldsymbol{D}_4$ is a perfect form, and*

$$\Xi(\boldsymbol{D}_4) = \{\pm b_i \mid 1 \leq i \leq 4\} \sqcup \{\pm(b_i - b_j) \mid 1 \leq i < j \leq 4,\ \{i,j\} \neq \{1,4\}\} \tag{23}$$

$$\sqcup \{\pm(b_1 - b_i + b_4) \mid 2 \leq i \leq 3\} \sqcup \{\pm(b_1 - b_2 - b_3 + b_4)\}. \tag{24}$$

Proof. We compute $\Xi(\boldsymbol{D}_4)$ using exhaustive enumeration and a computer assisted procedure, see [8, Proposition 7] for an alternative approach. If $\langle e, Me\rangle \leq 1$, for some $e \in \mathcal{Z}_d$ and $M \in \mathcal{S}_d^{++}$, then $|e|^2 \leq \lambda_{\min}(M)^{-1}$. For any of the finitely many vectors $e \in \mathcal{Z}_d$ such that $|e|^2 \leq \lambda_{\min}(\boldsymbol{D}_4)^{-1} = (5+\sqrt{17})/2 \approx 4.56$, we check that $\langle e, \boldsymbol{D}_4 e\rangle \geq 1$ and gather the cases of equality in the r.h.s. of (23). It follows that $\boldsymbol{D}_4 \in \mathcal{M}_d$, and since $\mathrm{Span}_{\mathbb{R}}\{ee^\top \mid e \in \Xi(\boldsymbol{D}_4)\} = \mathcal{S}_d$, the matrix $\boldsymbol{D}_4$ is a perfect form. □

Comparing the number $\#(\Xi(\boldsymbol{A}_4)) = 10$ and $\#(\Xi(\boldsymbol{D}_4)) = 12$ of active constraints at the perfect forms $\boldsymbol{A}_4, \boldsymbol{D}_4 \in \mathcal{M}_4$ with the dimension $\dim(\mathcal{S}_4) = 10$ of the optimization space, we find that $\boldsymbol{D}_4$ is a degenerate vertex of Ryskov's polyhedron $\mathcal{M}_4$, whereas $\boldsymbol{A}_4$ is a nondegenerate vertex.

In dimension $d \in \{2, 3\}$, there is only one equivalence class of perfect forms for the relation of arithmetical equivalence, associated with the representative $\boldsymbol{A}_d$ as established by Gauss [18]. For this reason Voronoi's first reduction (11) is particularly simple to study and to compute, using Selling's algorithm [9, 32], see Section 3.1. In contrast there is in dimension $d = 4$ one additional equivalence class of perfect forms, associated with the representative $\boldsymbol{D}_4$, see [20].

Proposition 2.8 ($d \leq 3$: [18], $d = 4$: [20])**.** *Let $d \leq 4$, and define*

$$\mathrm{Perfect}_0(d) := \{\boldsymbol{A}_d\},\ d \leq 3, \qquad \mathrm{Perfect}_0(4) := \{\boldsymbol{A}_4, \boldsymbol{D}_4\}. \tag{25}$$

Then $\mathrm{Perfect}_0(d) \subset \mathrm{Perfect}(d)$ *and each* $M \in \mathrm{Perfect}(d)$ *is arithmetically equivalent to exactly one element of* $\mathrm{Perfect}_0(d)$.

In dimension $d \geq 5$, by Theorem 2.2, there exists similarly a finite set $\mathrm{Perfect}_0(d)$ containing exactly one representative of each equivalence class of perfect forms. Voronoi's proof is non-constructive, and this set is only known *explicitly* in dimension $d \leq 8$ [33].

Elements of proof of Proposition 2.8, and discussion of Voronoi's algorithm. In addition to the historical references [18, 20], a modern proof is also presented in [8, Theorem 5]. Note that the inclusion $\mathrm{Perfect}_0(d) \subset \mathrm{Perfect}(d)$ follows from Propositions 2.6 and 2.7.

Since the set of vertices of any polyhedron is connected for the adjacency relation, and in view of the invariance properties of Ryskov's polyhedron see Proposition 2.5, it suffices to check, for any $M \in \mathrm{Perfect}_0(d)$, that any $M' \in \mathcal{N}(M)$ is arithmetically equivalent to some $M_0 \in \mathrm{Perfect}_0(d)$. This enumeration technique is known as Voronoi's algorithm. □

Remark 2.9. We implemented Voronoi's algorithm following [31], and observe that, in dimension $d = 4$, all 10 neighbors of the nondegenerate vertex $\boldsymbol{A}_4$ of $\mathcal{M}_d$ are arithmetically equivalent to $\boldsymbol{D}_4$, while the degenerate vertex $\boldsymbol{D}_4$ has 48 neighbors arithmetically equivalent to $\boldsymbol{A}_4$ and 16 neighbors arithmetically equivalent to $\boldsymbol{D}_4$. We also obtain explicitly the unimodular matrices corresponding to the above arithmetic equivalence relations. This data is needed for the numerical implementation of the decomposition presented in Algorithm 2.

3 Computing the decomposition

We explain in this section how one may compute in practice, and implement numerically, the decomposition $\lambda(D)$ of a matrix $D \in \mathcal{S}_d^{++}$ defined in Proposition 1.5. After an optional preliminary step described in Remark 3.4, the procedure is to solve the minimization problem (11) known as Voronoi's first reduction using Algorithm 2, and then to obtain the explicit coefficients as described in Propositions 3.5 and 3.6. As a side product, we establish in Proposition 3.8 the locally Lipschitz regularity of this decomposition, announced in Theorem 1.6. Finally, we discuss the special case $d \in \{2, 3\}$ in Section 3.1.

Algorithm 1 Solving Voronoi's first reduction — abstract version

Initialization: Let $M \in \mathrm{Perfect}(d)$ (for instance $M \leftarrow \boldsymbol{A}_d$).
While there exists $M' \in \mathcal{N}(M)$ such that $\mathrm{Tr}(DM') < \mathrm{Tr}(DM)$ **do** $M \leftarrow M'$.
Return M.

Algorithm 2 Solving Voronoi's first reduction — practical version

Initialization:
Let $M_0 \in \mathrm{Perfect}_0(d)$ (for instance $M \leftarrow \boldsymbol{A}_d$).
Let $A \in \mathrm{GL}(\mathbb{Z}^d)$ (for instance $A \leftarrow I_d$).
While there exists $M \in \mathcal{N}(M_0)$ such that $\mathrm{Tr}(DA^\top MA) < \mathrm{Tr}(DA^\top M_0 A)$
do
look up a decomposition $M = (A')^\top M_0' A'$ with $M_0' \in \mathrm{Perfect}_0(d)$ and $A' \in \mathrm{GL}(\mathbb{Z}^d)$
$M_0 \leftarrow M_0'$.
$A \leftarrow A'A$.
Return M_0 and A.

Since the cost minimized in (11) is linear, the minimum in attained at some vertex of Ryskov's polyhedron $\mathcal{M}_d$, which can be found by iterating over perfect forms in the manner described in Algorithm 1. This algorithm is however not directly implementable since we did not explain how the set $\mathcal{N}(M)$ is computed, for an arbitrary perfect form M. In practice, in order to benefit from the symmetries of Ryskov's polyhedron, we represent a perfect form M by a pair (M_0, A), where $M_0 \in \mathrm{Perfect}_0(d)$, $A \in \mathrm{GL}(\mathbb{Z}^d)$, and $M = A^\top M_0 A$. This yields Algorithm 2, which is equivalent to Algorithm 1 as shown by Proposition 2.5. Finally, given $M_0 \in \mathrm{Perfect}_0(d)$, we use Voronoi's algorithm as described in Remark 2.9 to express each element of $\mathcal{N}(M_0)$ in the form $(A')^\top M_0' A'$, where $M_0' \in \mathrm{Perfect}_0(d)$ and $A' \in \mathrm{GL}(\mathbb{Z}^d)$. The numerical evaluation of $\mathrm{Tr}(DA^\top MA) = \mathrm{Tr}([ADA^\top]M)$, for all $M \in \mathcal{N}(M_0)$, is made efficient by computing only once the matrix product $ADA^\top$, and recognizing the Frobenius inner product.

Remark 3.1 (Computational efficiency and on the fly computation of perfect forms). Linear programs, i.e. the minimization of a linear form over a polyhedron defined by (finitely many)

inequalities, can be addressed using a variety of numerically methods. This includes interior point methods [25], which never consider the polytope vertices, and the simplex algorithm, which discovers and computes a path through the skeleton defined by the polytope edges and vertices on the fly as the optimization progresses.

The simplex algorithm can be used to solve Voronoi's first reduction, up to minor modifications related to the fact that Ryskov's polyhedron is defined by infinitely many constraints. Algorithm 2 is a variant of the simplex algorithm enhanced by the precomputation of the skeleton of Ryskov's polyhedron on which it operates.

When solving a generic linear program, precomputing the skeleton of the optimization polytope is generally a bad idea, due to its high time and space complexity. However, in the intended applications[3] of this work, one needs to solve millions of linear programs on the *same* polytope, namely Ryskov's polyhedron, whose skeleton happens to have a particularly simple structure at least in dimension $d \leq 6$. For this reason, the precomputation of the skeleton leads to strong efficiency gains.

We report the computation times of 0.022s, 0.057s, 0.12s, 1.1s, 5.25s, in dimension $d = 2, 3, 4, 5, 6$ respectively, for computing Voronoi's decomposition of 500 000 matrices of shape $d \times d$ on a laptop equipped with an Nvidia® 2060 MaxQ GPU, using a CUDA® implementation of Algorithm 2 (with a straightforward parallelization of this embarrassingly parallel task). The processed matrices are generated as $AA^\top + \varepsilon \operatorname{Tr}(AA^\top) \operatorname{Id}$, where A is an $n \times n$ matrix whose coefficients are drawn from a normal distribution, and where the relaxation parameter $\varepsilon = 0.01$ is used to avoid excessively degenerate matrices, which are irrelevant for PDE discretizations.

The proposed approach is only effective in small dimension, since the structure of Ryskov's polyhedron becomes significantly more complex in dimension $d \geq 7$, and is not classified in dimension $d \geq 9$.

Proposition 3.2. *For any $D \in \mathcal{S}_d^{++}$, Algorithm 1 terminates and returns a perfect form M such that $D \in \mathcal{S}^{++}(M)$. Equivalently, Algorithm 2 terminates and returns a perfect form $M_0 \in \operatorname{Perfect}_0(d)$ and a matrix $A \in \operatorname{GL}(\mathbb{Z}^d)$ such that $D \in \mathcal{S}^{++}(A^\top M_0 A)$.*

Proof. By Theorem 2.2 the set $\{M \in \mathcal{M}_d \mid \operatorname{Tr}(DM) \leq \alpha\}$ is a bounded polyhedron, nonempty if α is sufficiently large, and in particular there are finitely many perfect forms M such that $\operatorname{Tr}(DM) \leq \alpha$. Thus Algorithm 1 iterates over finitely many perfect forms M. Since the cost $\operatorname{Tr}(DM)$ decreases strictly at each iteration, the algorithm terminates. The returned M satisfies $\operatorname{Tr}(DM) \leq \operatorname{Tr}(DM')$ for any $M' \in \mathcal{N}(M)$, therefore it is a minimizer in (11), see (18). □

Algorithm 2 returns the minimizer of (11) in the factorized form $A^\top M_0 A$ where $M_0 \in \operatorname{Perfect}_0(d)$ and $A \in \operatorname{GL}(\mathbb{Z}^d)$. This is useful since then, by the following proposition, we only

[3]Typically solving anisotropic PDEs on Cartesian grids, see Theorem 1.4 and Appendix A, which involves computing Voronoi's decomposition of a Riemannian metric tensor, or a diffusion tensor, or a Hooke elasticity tensor, which is given as data and is different at each discretization point.

need to know how to compute $\lambda(D)$ for matrices $D \in \mathcal{S}^{++}(M_0)$ for some $M_0 \in \mathrm{Perfect}_0(d)$.

Proposition 3.3. *Let $D \in \mathcal{S}^{++}(A^\top M_0 A)$, for some $M_0 \in \mathrm{Perfect}_0(d)$ and $A \in \mathrm{GL}(\mathbb{Z}^d)$. Then $ADA^\top \in \mathcal{S}^{++}(M_0)$, and one has (assuming $d \leq 4$ for the second identity)*

$$\Lambda(D) = \{(\lambda^{Ae})_{e\in\mathcal{Z}_d} \mid \lambda \in \Lambda(ADA^\top)\}, \qquad \lambda(D) = (\lambda^{Ae}(ADA^\top))_{e\in\mathcal{Z}_d}. \tag{26}$$

Proof. We have $A\mathbb{Z}^d = \mathbb{Z}^d$, hence $\mathcal{M}_d = \{A^\top M A \mid M \in \mathcal{M}_d\}$, which implies

$$\begin{aligned}\mathrm{Vor}(ADA^\top) &= \min_{M\in\mathcal{M}_d} \mathrm{Tr}(ADA^\top M) = \min_{M\in\mathcal{M}_d} \mathrm{Tr}(DA^\top M A) = \min_{M\in\mathcal{M}_d} \mathrm{Tr}(DM) = \mathrm{Vor}(D)\\ &= \mathrm{Tr}(DA^\top M_0 A) = \mathrm{Tr}(ADA^\top M_0).\end{aligned}$$

Therefore $ADA^\top \in \mathcal{S}^{++}(M_0)$ as announced. The equalities (26) then follow directly from Propositions 2.3 and 2.5. □

Remark 3.4 (Basis reduction as a preliminary step to Voronoi's reduction). Computing the decomposition of $D \in \mathcal{S}_d^{++}$, or of $A^\top DA$ for some $A \in \mathrm{GL}(\mathbb{Z}^d)$, are equivalent problems by Proposition 3.3. However, the number of iterations of Algorithms 1 and 2, and thus the numerical cost, may be strongly reduced in the latter case, if the anisotropy ratio $\mu(D)$ is large and if the change of coordinates A is well chosen, using typically a basis reduction method. In the special case of dimension $d \leq 3$, where Algorithm 1 reduces to Selling's algorithm presented Section 3.1, such a preprocessing is alluded to in [9, Remark 7.0.3], and it is shown in [17, Corollary 1 and Proposition 1] that Selling's algorithm terminates in a single step if the greedy lattice basis reduction algorithm [26, Fig 3] is used for preprocessing, which has complexity $\mathcal{O}(\ln \mu(D))$. Similar improvements can be expected in dimension $d = 4$. However, these concerns appear mostly relevant for applications related to number theory where $\mu(D) \gg 100$. In contrast, for the comparably mild anisotropy ratios encountered in PDE discretizations, our numerical experience suggests that the basis reduction preprocessing step is not essential.

We describe below how to compute $\lambda(D)$ when $D \in \mathcal{S}^{++}(M_0)$, for some perfect form of reference $M_0 \in \mathrm{Perfect}_0(d)$, in dimension $d \leq 4$.

Proposition 3.5. *Let $D \in \mathcal{S}^{++}(\boldsymbol{A}_d)$, with coefficients $(D_{ij})_{ij=1}^d$. Then $\Lambda(D)$ is a singleton, and*

$$\lambda^e(D) = \begin{cases} \sum_{j=1}^d D_{ij} & \text{if } e = \pm b_i,\ 1 \leq i \leq d,\\ -D_{ij} & \text{if } e = \pm(b_i - b_j),\ 1 \leq i < j \leq d,\\ 0 & \text{else.}\end{cases} \tag{27}$$

Proof. By Proposition 2.3, the set $\Lambda(D)$ is non-empty, and any $\lambda \in \Lambda(D)$ satisfies $\mathrm{supp}(\lambda) \subset \Xi(\boldsymbol{A}_d)$ and $\sum_{e\in\Xi(\boldsymbol{A}_d)} \lambda^e ee^\top = D$. On the other hand, the collection of symmetric matrices $\mathcal{B} := \{ee^\top \mid e \in \Xi(\boldsymbol{A}_d)\}$ has cardinality $\#(\mathcal{B}) := d(d+1)/2 = \dim(\mathcal{S}_d)$ and $\mathrm{Span}_{\mathbb{R}}\,\mathcal{B} = \mathcal{S}_d$, as established in Proposition 2.6, thus $\mathcal{B}$ is a basis of $\mathcal{S}_d$. It follows that there exists exactly

one $\lambda : \Xi(\boldsymbol{A}_d) \to \mathbb{R}$ such that $\sum_{e\in\Xi(\boldsymbol{A}_d)} \lambda^e ee^\top = D$, namely the one defined by (27), which concludes. □

Proposition 3.6. *Let $D \in \mathcal{S}^{++}(\boldsymbol{D}_4)$, with coefficients $(D_{ij})_{i,j=1}^d$. For any α, β, $\gamma \in \mathbb{R}$, let $\lambda_{\alpha,\beta,\gamma}(D) : \mathcal{Z}_4 \to \mathbb{R}$ be defined by*

$$\lambda^e_{\alpha,\beta,\gamma}(D) := \begin{cases} D_{i1}+D_{i2}+D_{i3}+D_{i4}+\gamma & \text{if } e = \pm b_i,\ i \in \{1,4\}, \\ D_{21}+D_{22}+D_{23}+D_{24}+\alpha & \text{if } e = \pm b_2, \\ D_{31}+D_{32}+D_{33}+D_{34}+\beta & \text{if } e = \pm b_3, \\ -D_{i2}-D_{14}+\beta & \text{if } e = \pm(b_i - b_2),\ i \in \{1,4\}, \\ -D_{i3}-D_{14}+\alpha & \text{if } e = \pm(b_i - b_3),\ i \in \{1,4\}, \\ -D_{23}+D_{14}+\gamma & \text{if } e = \pm(b_2 - b_3), \\ \alpha & \text{if } e = \pm(b_1 - b_2 + b_4), \\ \beta & \text{if } e = \pm(b_1 - b_3 + b_4), \\ D_{14}+\gamma & \text{if } e = \pm(b_1 - b_2 - b_3 + b_4), \\ 0 & \text{else.} \end{cases}$$

Then the set $\Lambda(D)$ is characterized by

$$\begin{aligned} \Lambda(D) &= \{\lambda_{\alpha,\beta,\gamma}(D) \mid \alpha \geq \alpha_*(D),\ \beta \geq \beta_*(D),\ \gamma \geq \gamma_*(D),\ \alpha+\beta+\gamma = 0\}, \qquad (28) \\ \alpha_*(D) &:= \max\{-D_{21}-D_{22}-D_{23}-D_{24},\ D_{13}+D_{14},\ D_{34}+D_{14},\ 0\}, \\ \beta_*(D) &:= \max\{-D_{31}-D_{32}-D_{33}-D_{34},\ D_{12}+D_{14},\ D_{24}+D_{14},\ 0\}, \\ \gamma_*(D) &:= \max\{-D_{11}-D_{12}-D_{13}-D_{14},\ -D_{41}-D_{42}-D_{43}-D_{44},\ D_{23}-D_{14},\ -D_{14}\}. \end{aligned}$$

Thus $\Lambda(D)$ an equilateral triangle, non-empty but possibly reduced to a single point, whose barycenter has parameters $(\alpha,\beta,\gamma) = (\alpha_*(D),\beta_*(D),\gamma_*(D)) - \frac{1}{3}(\alpha_*(D)+\beta_*(D)+\gamma_*(D))\mathbb{1}$.

Proof. Let $\Lambda_*(D) := \{\lambda\colon \Xi(\boldsymbol{D}_4) \to \mathbb{R} \mid \mathrm{supp}(\lambda) \subset \Xi(\boldsymbol{D}_4),\ \sum_{e\in\Xi(\boldsymbol{D}_4)} \lambda^e ee^\top = D\}$, so that $\Lambda(D) = \{\lambda \in \Lambda_*(D) \mid \lambda \succeq 0\}$. Recall that the elements of $\Xi(\boldsymbol{D}_4)$ are described in Proposition 2.7. Since $\#(\Xi(\boldsymbol{D}_4)) = 12 = \dim(\mathcal{S}_4)+2$, and $\mathrm{Span}_{\mathbb{R}}\{ee^\top \mid e \in \Xi(\boldsymbol{D}_4)\} = \mathcal{S}_d$ by definition of a perfect form, the set $\Lambda_*(D)$ is a two-dimensional affine space. We compute that $\Lambda_*(D) = \{\lambda_{\alpha,\beta,\gamma}(D) \mid \alpha,\beta,\gamma \in \mathbb{R},\ \alpha+\beta+\gamma = 0\}$, from which the characterization (28) follows.

The set $\Lambda(D)$ is non-empty by Proposition 2.3, hence $\alpha_*+\beta_*+\gamma_* \leq 0$ (omitting the parameter D for readability). By construction, it is the convex hull of the three points $\lambda_{\alpha,\beta,\gamma}$ whose parameters are $(\alpha_*,\beta_*,-\alpha_*-\beta_*)$, $(\alpha_*,-\alpha_*-\gamma_*,\gamma_*)$ and $(-\beta_*-\gamma_*,\beta_*,\gamma_*)$, hence $\Lambda(D)$ is a triangle with the announced barycenter. We conclude observing that the distance between any two of these vertices is $|\alpha_*+\beta_*+\gamma_*|\sqrt{6}$ in the Euclidean space $\mathbb{R}^{\Xi(\boldsymbol{D}_4)}$. □

In addition to explaining how to compute $\lambda(D)$, the above propositions also allow us to establish, in Proposition 3.8 below, the part of Theorem 1.6 about the Lipschitz regularity

of the coefficients λ of the matrix decomposition. We first obtain a lower bound on the norm of the obtained unimodular transformation.

Proposition 3.7. *Let $D \in \mathcal{S}^{++}(A^\top M_0 A)$, with $A \in \mathrm{GL}(\mathbb{Z}^d)$ and $M_0 \in \mathrm{Perfect}_0(d)$. Then $\|AD^{\frac{1}{2}}\| \leq C\lambda_{\max}(D)^{\frac{1}{2}}$, and in particular $\|A\| \leq C\mu(D)$, for some constant $C = C(d)$.*

Proof. Using successively (i) the fact that the identity matrix belongs to Ryskov's polyhedron, and (ii) the optimality of $M = A^\top M_0 A$, we obtain that

$$d\lambda_{\max}(D) \geq \mathrm{Tr}(D) \geq \mathrm{Tr}(DA^\top M_0 A) \geq \lambda_{\max}(ADA^\top)\lambda_{\min}(M_0) = \|AD^{\frac{1}{2}}\|^2 \lambda_{\min}(M_0).$$

This establishes the first estimate. We conclude noting that $\|AD^{\frac{1}{2}}\| \geq \|A\|\lambda_{\min}(D)^{\frac{1}{2}}$. □

Proposition 3.8. *Assume that $d \leq 4$, and equip Λ_d with the norm $|\cdot|_\infty \colon \lambda \mapsto \max_{e \in \mathcal{Z}_d} |\lambda^e|$. Then the mapping $D \in \mathcal{S}_d^{++} \mapsto \lambda(D)$ is locally Lipschitz continuous, with dilatation coefficient bounded by $C\mu(D)^2$ for some constant C.*

Proof. We deduce easily from Propositions 3.5 and 3.6 that the mapping λ is Lipschitz continuous on $\bigcup_{M_0 \in \mathrm{Perfect}_0(d)} \mathcal{S}^{++}(M_0)$, with some dilatation coefficient $K_0 > 0$.

Let D_1, $D_2 \in \mathcal{S}_d^{++}$, $I := \{tD_1 + (1-t)D_2 \mid t \in [0,1]\}$, and $\mu := \max\{\mu(D) \mid D \in I\} = \max\{\mu(D_1), \mu(D_2)\}$. We consider the restriction of the mapping λ to the segment I. If $M = A^\top M_0 A \in \mathrm{Perfect}(d)$, where $A \in \mathrm{GL}(\mathbb{Z}^d)$ and $M_0 \in \mathrm{Perfect}_0(d)$, is such that $I \cap \mathcal{S}^{++}(M)$ is nonempty, then $\|A\| \leq C\mu$ by Proposition 3.7. Since there are only finitely many $A \in \mathrm{GL}(\mathbb{Z}^d)$ such that $\|A\| \leq C\mu$, it follows that I is the union of finitely many closed segments $I \cap \mathcal{S}^{++}(M)$. By Proposition 3.3 and the above, λ is Lipschitz continuous on the segment $I \cap \mathcal{S}^{++}(A^\top M_0 A)$, with dilatation coefficient $K_0\|A\|^2 \leq K_0 C^2 \mu^2$. Thus λ is $K(\mu)$-Lipschitz on the whole segment I, where $K(\mu) := K_0 C^2 \mu^2$. □

3.1 Selling's algorithm and formula

Selling's algorithm [9, 32] can be regarded as a reformulation and a simplification of Algorithm 2 in dimension $d \in \{2, 3\}$, taking advantage of the fact that $\mathrm{Perfect}_0(d)$ is a singleton, see Proposition 2.8. We briefly present this approach and the related concept of superbase of a lattice, for concreteness and in order to develop a two-dimensional variant with smooth coefficients in Section 6.

Definition 3.9. A *superbase* of $\mathbb{Z}^d$ is a tuple $v := (v_0, \cdots, v_d) \in (\mathbb{Z}^d)^{d+1}$ such that $|\det(v_1, \cdots, v_d)| = 1$ and $v_0 + \cdots + v_d = 0$. We define $M_v := \frac{1}{2}\sum_{0 \leq i \leq d} v_i v_i^\top \in \mathcal{S}_d^{++}$.

Remark 3.10. The *canonical superbase* is $(b_0, \cdots, b_d)$ where $(b_i)_{i=1}^d$ is the canonical basis of $\mathbb{R}^d$, and $b_0 := -\mathbb{1}$ in such way that $b_0 + \cdots + b_d = 0$. Any superbase $(v_0, \cdots, v_d)$ can be obtained from the canonical one by a linear change of variables: $v_i = Ab_i$ for all $0 \leq i \leq d$, for some $A \in \mathrm{GL}(\mathbb{Z}^d)$. This defines a bijection between the collections of superbases and of unimodular matrices.

We show in Propositions 3.11 and 3.12 that perfect forms arithmetically equivalent to $\boldsymbol{A}_d$ can be parametrized by superbases, uniquely up to a permutation and a global change of sign. The perfect form's optimality in Voronoi's reduction (11) is characterized by a geometrical obtuseness property of the superbase in Proposition 3.13.

Proposition 3.11. *A matrix $M \in \mathcal{S}_d^{++}$ is arithmetically equivalent to $\boldsymbol{A}_d$ iff $M = M_v$ for some superbase $v = (v_0, \cdots, v_d)$. Specifically $M_v = A\boldsymbol{A}_d A^\top$ if $v_i = Ab_i$ for all $0 \leq i \leq d$, for some $A \in \mathrm{GL}(\mathbb{Z}^d)$.*

Proof. One has $\frac{1}{2}\sum_{i=0}^d b_i b_i^\top = \frac{1}{2}(\mathbb{1}\mathbb{1}^\top + \mathrm{Id}_d) = \boldsymbol{A}_d$, compare with (19). The result follows noting that a matrix A is unimodular iff the transpose $A^\top$ is unimodular. □

Proposition 3.12. *Let $v = (v_0, \cdots, v_d)$ be a superbase, and let $\tilde{M}_v := \sum_{e \in \Xi(M_v)} ee^\top$, then $\{e \in \mathcal{Z}_d \mid \langle e, \tilde{M}_v e\rangle = d\} = \{\pm v_0, \cdots, \pm v_d\}$. If two superbases v, v' are such that $M_{v'} = M_v$, then v and v' coincide up to a permutation of their elements and a global change of sign.*

Proof. Regarding the first point, we may assume up to a linear change of coordinates that v is the canonical superbase $(b_i)_{0 \leq i \leq d}$. In that case $\Xi(M_v) = \Xi(\boldsymbol{A}_d)$ is described in (21), and one readily checks that $\langle b_i, \tilde{M}_v b_i\rangle = \sum_{e \in \Xi(\boldsymbol{A}_d)} \langle e, b_i\rangle^2 = d$ for any $0 \leq i \leq d$. Conversely, let $x = (x_1, \cdots, x_d) \in \mathbb{Z}^d \setminus \{0\}$ be such that $\langle x, \tilde{M}_v x\rangle = d$. If $x_1 = \cdots = x_d$, then $\langle x, \tilde{M}_v x\rangle = x_1^2 \langle b_0, \tilde{M}_v b_0\rangle = dx_1^2$, and thus $x = \pm b_0$ as desired. Otherwise we may assume up to permuting the coordinates that $x_1 \neq x_2$, and we thus obtain using (21) that

$$d = \langle x, \tilde{M}_v x\rangle \geq \left[x_1^2 + x_2^2\right] + \left\{x_3^2 + \cdots + x_d^2\right\} + \left[(x_1 - x_2)^2\right] + \sum_{i=3}^d \left[(x_1 - x_i)^2 + (x_2 - x_i)^2\right].$$

There are d terms within square brackets, and one within curly braces. Each bracketed term is a positive integer, and the term between braces is non-negative, hence each bracketed term equals one and the term between braces vanishes. If follows that $x = \pm b_i$ for some $1 \leq i \leq d$, as announced.

Now, if $M_v = M_{v'}$, then $\tilde{M}_v = \tilde{M}_{v'}$ and therfore $\{\pm v_0, \cdots, \pm v_d\} = \{\pm v_0', \cdots, \pm v_d'\}$ by the first point. If follows that $v_i' = \varepsilon(i) v_{\sigma(i)}$ for some signs $\varepsilon : \{0, \cdots, d\} \to \{-1, 1\}$ and indices $\sigma : \{0, \cdots, d\} \to \{0, \cdots, d\}$. Recalling that v' contains a basis since $|\det(v_1', \cdots, v_d')| = 1$, and that $v_0' + \cdots + v_d' = 0$, one easily obtains that ε is constant and that σ is a permutation, which concludes. □

Proposition 3.13. *For any $D \in \mathcal{S}_d^{++}$ and any superbase $v = (v_0, \cdots, v_d)$, the following are equivalent: (i) $D \in \mathcal{S}^{++}(M_v)$, and (ii) v is D-obtuse, i.e. $\langle v_i, Dv_j\rangle \leq 0$ for all $0 \leq i < j \leq d$. In particular, each $D \in \mathcal{S}_d^{++}$ with $d \in \{2, 3\}$ admits a D-obtuse superbase.*

Proof. Up to a unimodular change of coordinates, we may assume that $v = (b_0, \cdots, b_d)$ is the canonical superbase, and thus $M_v = \boldsymbol{A}_d$. We denote by $(D_{ij})_{i,j=1}^d$ the coefficients of D, and note that $-\langle b_i, Db_j\rangle = -D_{ij}$, for all $1 \leq i < j \leq d$, and $-\langle b_0, Db_i\rangle =$

$\langle b_1+\cdots+b_d, Db_i\rangle = \sum_{j=1}^d D_{ij}$, for all $1 \le i \le d$. Assuming that $D \in \mathcal{S}^{++}(M_v) = \mathcal{S}^{++}(\boldsymbol{A}_d)$, we obtain by Proposition 3.5 that these negated scalar products are non-negative, as announced. Conversely, if the superbase $(b_i)_{i=0}^d$ is D-obtuse, we obtain by (20) a non-negative decomposition of D supported on $\Xi(\boldsymbol{A}_d)$, and thus $D \in \mathcal{S}^{++}(\boldsymbol{A}_d)$ by Proposition 2.4. This establishes the announced equivalence.

By Voronoi's theorem, for each $D \in \mathcal{S}_d^{++}$ there exists $M \in \mathrm{Perfect}(d)$ such that $D \in \mathcal{S}^{++}(M)$. In dimension $d \in \{2,3\}$, M must be arithmetically equivalent to $\boldsymbol{A}_d$ by Proposition 2.8, and thus $M = M_v$ for some superbase v by Proposition 3.11. This superbase is then D-obtuse by the first point, as announced. □

In dimension $d \in \{2,3\}$, the neighbor relation between perfect forms, on the skeleton of Ryskov's polyhedron $\mathcal{M}_d$, can be rephrased in terms of superbase transformations discovered by Selling [32], and described in the next result. This turns the solution of Voronoi's first reduction of $D \in \mathcal{S}_d^{++}$ by Algorithm 1 into a succession of superbase transformations, known as Selling's algorithm [4, Algorithm 1] (with the additional observation that $\mathrm{Tr}(M_{v'}D) < \mathrm{Tr}(M_vD)$ iff $\langle \tilde{v}_0, D\tilde{v}_1\rangle > 0$, with the notations of Proposition 3.14). The subsequent decomposition of D, as described in Propositions 3.3 and 3.5, is then known as Selling's decomposition.

Proposition 3.14 (Selling's superbase transformations)**.** *Let $d \in \{2,3\}$, and let v and v' be superbases. The following are equivalent : (i) $M_{v'} \in \mathcal{N}(M_v)$, and (ii) there are permutations $\tilde{v}$ of v, and $\tilde{v}'$ of v', and a sign $\varepsilon \in \{-1,1\}$ such that*

$$\textit{(Case } d=2\textit{)}\ \varepsilon\tilde{v}' = (\tilde{v}_0, -\tilde{v}_1, \tilde{v}_1 - \tilde{v}_0), \quad \textit{(Case } d=3\textit{)}\ \varepsilon\tilde{v}' = (\tilde{v}_0, -\tilde{v}_1, \tilde{v}_1 + \tilde{v}_2, \tilde{v}_1 + \tilde{v}_3).$$

Proof. It is proved in [4, Corollary 2.11] that $\mathcal{N}(M_v)$ contains the perfect forms associated with $(\tilde{v}_0, -\tilde{v}_1, \tilde{v}_1 - \tilde{v}_0)$ if $d = 2$ (resp. $(\tilde{v}_0, -\tilde{v}_1, \tilde{v}_1 + \tilde{v}_2, \tilde{v}_1 + \tilde{v}_3)$ if $d = 3$), where $\tilde{v}$ is any permutation of v, and no other perfect form. We conclude using the uniqueness of the superbase associated to a given perfect form, up to a permutation and a global change of sign, established in Proposition 3.12. □

4 Upper bound on the radius of the stencil

The main results of this section are Proposition 4.2 and Theorem 4.3, which imply the part of Theorem 1.6 about $R(\mu)$-supportedness (but are not restricted to dimension $d \le 4$). These results follow from the next technical lemma, which relates Voronoi's reductions of a matrix and of its inverse.

Lemma 4.1. *For any $M_0 \in \mathrm{Perfect}_0(d)$, there exists a finite subset $P(M_0) \subset \mathrm{Perfect}(d)$ such that: for all $D \in \mathcal{S}^{++}(M_0)$, one has $D^{-1} \in \mathcal{S}^{++}(M_1)$ for some $M_1 \in P(M_0)$.*

The proof is postponed, but let us immediately say that it is constructive, and that a suitable set, denoted $\mathrm{Perfect}_1(M_0)$ and obeying the conditions of $P(M_0)$ in Lemma 4.1,

is eventually obtained as the collection of vertices of the Pareto front of a multi-objective linear optimization problem posed on Ryskov's polyhedron, see (32) below. As a numerical experiment, presented in Section 4.1 below and limited to dimension $d \leq 5$, we compute this set explicitly.

The following result, together with Proposition 3.7, allows to control the largest and the smallest singular values of the unimodular transformations arising in Voronoi's reduction.

Proposition 4.2. *Let $A \in \mathrm{GL}(\mathbb{Z}^d)$, $M_0 \in \mathrm{Perfect}_0(d)$, and $D \in \mathcal{S}^{++}(A^\top M_0 A)$. Then $\|D^{-\frac{1}{2}} A^{-1}\| \leq C\lambda_{\min}(D)^{-\frac{1}{2}}$, and thus $\|A^{-1}\| \leq C\mu(D)$, for some constant $C = C(d)$.*

Proof. Since $D \in \mathcal{S}^{++}(A^\top M_0 A)$ one has $ADA^\top \in \mathcal{S}^{++}(M_0)$, see Proposition 3.3. Therefore $(ADA^\top)^{-1} \in \mathcal{S}^{++}(M_1)$ for some $M_1 \in P(M_0)$, using the notations of Lemma 4.1. Recalling that $\mathrm{Perfect}_0(d)$ contains a representative of each class of perfect forms, and that it is finite as well as $P(M_0)$, we find that $M_1 = \tilde{A}^\top M_0' \tilde{A}$ for some $M_0' \in \mathrm{Perfect}_0(d)$ and $\tilde{A} \in \mathcal{A}$, where $\mathcal{A} \subset \mathrm{GL}(\mathbb{Z}^d)$ is a fixed and finite subset.

We have obtained that $(ADA^\top)^{-1} \in \mathcal{S}^{++}(\tilde{A}^\top M_0' \tilde{A})$, equivalently $D^{-1} \in \mathcal{S}^{++}(A^{-1}\tilde{A}^\top M_0' \tilde{A} A^{-\top})$. By Proposition 3.7 one has $\|\tilde{A}A^{-\top}D^{-\frac{1}{2}}\| \leq C\lambda_{\max}(D^{-1})^{\frac{1}{2}}$, thus $\|A^{-\top}D^{-\frac{1}{2}}\| \leq CC'\lambda_{\min}(D)^{-\frac{1}{2}}$ where $C' := \max_{\tilde{A}\in\mathcal{A}} \|\tilde{A}^{-1}\|$. The first estimate follows, by transposition, and we conclude noting that $\|D^{-\frac{1}{2}}A^{-1}\| \geq \lambda_{\min}(D^{-\frac{1}{2}})\|A^{-1}\| = \|A^{-1}\|\lambda_{\max}(D)^{-\frac{1}{2}}$. □

We next estimate, as announced, the radius of the support of the decomposition of a positive quadratic form obtained from Voronoi's reduction, which is also the stencil of our finite differences scheme. To the best of our knowledge, Theorem 4.3 was previously proved only in dimension $d \leq 3$ [22, Theorem 4.11], while in higher dimension only the weaker estimate $|e| \leq C\mu(D)^{d-1}$ was known [23, Proposition 1.1]. The constant (29), when choosing $P(M_0) = \mathrm{Perfect}_1(M_0)$, is computed in Section 4.1 in dimension $d \leq 5$, see (33) and Remark 4.21.

Theorem 4.3. *For any $D \in \mathcal{S}_d^{++}$, $\lambda \in \Lambda(D)$, $e \in \mathrm{supp}(\lambda)$, one has $\|e\|_{D^{-1}} \leq C\lambda_{\min}(D)^{-\frac{1}{2}}$, and in particular $|e| \leq C\mu(D)$. A suitable (but not sharp) constant $C = C(d)$ is*

$$C(d) = \sqrt{d}\max\{\|e\|_{M_1^{-1}} \mid M_0 \in \mathrm{Perfect}_0(d),\, e \in \Xi(M_0),\, M_1 \in P(M_0)\}, \tag{29}$$

where $P(M_0)$ is a finite set obeying the conditions of Lemma 4.1.

Proof. Up to a unimodular change of coordinates, we may assume that $D \in \mathcal{S}^{++}(M_0)$ for some $M_0 \in \mathrm{Perfect}_0(d)$. Then $D^{-1} \in \mathcal{S}^{++}(M_1)$ for some $M_1 \in P(M_0)$ by Lemma 4.1, and $e \in \mathrm{supp}(\lambda(D)) \subset \Xi(M_0)$ by Proposition 2.3. Thus $\|e\|_{D^{-1}}^2 = \mathrm{Tr}(D^{-1}ee^\top) \leq (C(d)^2/d)\,\mathrm{Tr}(D^{-1}M_1) \leq (C(d)^2/d)\,\mathrm{Tr}(D^{-1}) \leq C(d)^2/\lambda_{\min}(D)$, using that $ee^\top \preceq \|e\|_{M_1^{-1}}^2 M_1$ for the second inequality, and that $\mathrm{Id}_d \in \mathcal{M}_d$ for the third. The result follows. □

Let us now turn to the proof of Lemma 4.1. We denote by $e_1 \wedge \cdots \wedge e_{d-1} \in \mathbb{R}^d$ the generalized cross product of $d-1$ vectors $e_1, \cdots, e_{d-1} \in \mathbb{R}^d$, which is characterized by the identity

$$\langle e_1 \wedge \cdots \wedge e_{d-1}, e \rangle = \det(e_1, \cdots, e_{d-1}, e) \tag{30}$$

for all $e \in \mathbb{R}^d$. In dimension $d = 3$ one recovers the usual cross product $e_1 \wedge e_2 = e_1 \times e_2$, and in dimension $d = 2$ the perpendicular to a given vector $\wedge e_1 = e_1^\perp$. For any matrix $A \in \mathbb{R}^{d \times d}$, we denote by $\mathrm{adj}(A)$ its adjugate matrix (if A is invertible, then $A^{-1} = \det(A)^{-1} \mathrm{adj}(A)$ by Cramer's rule), which is also the transposed matrix of cofactors [12].

Lemma 4.4. *Let $e_1, \cdots, e_I \in \mathbb{R}^d$, and $\lambda_1, \cdots, \lambda_I \in \mathbb{R}$, where $I \geq d-1$. Then*

$$\mathrm{adj}\Big(\sum_{1 \leq i \leq I} \lambda_i e_i e_i^\top \Big) = \sum_{1 \leq i_1 < \cdots < i_{d-1} \leq I} (\lambda_{i_1} \cdots \lambda_{i_{d-1}}) (e_{i_1} \wedge \cdots \wedge e_{i_{d-1}}) (e_{i_1} \wedge \cdots \wedge e_{i_{d-1}})^\top.$$

Proof. We can assume w.l.o.g. that $\lambda_1, \cdots, \lambda_I \geq 0$, since the identity to be proved is polynomial in these variables. Then, up to considering $\sqrt{\lambda_1} e_1, \cdots, \sqrt{\lambda_I} e_I$, we can assume that $\lambda_1 = \cdots = \lambda_I = 1$.

Denote $A_h := [e_1, \cdots, e_I, \sqrt{h} e] \in \mathbb{R}^{d \times (I+1)}$, where $h > 0$ and $e \in \mathbb{R}^d$ are arbitrary. Then

$$\begin{aligned} \det(A_h A_h^\top) &= \det\Big(\sum_{1 \leq i \leq I} e_i e_i^\top + h e e^\top \Big) = \det\Big(\sum_{1 \leq i \leq I} e_i e_i^\top \Big) + h \Big\langle e, \mathrm{adj}\big(\sum_{1 \leq i \leq I} e_i e_i^\top \big) e \Big\rangle + o(h) \\ &= \sum_{1 \leq i_1 < \cdots < i_d \leq I} \det(e_{i_1}, \cdots, e_{i_d})^2 + h \sum_{1 \leq i_1 < \cdots < i_{d-1} \leq I} \det(e_{i_1}, \cdots, e_{i_{d-1}}, e)^2. \end{aligned}$$

We used Jacobi's formula for the derivative of the determinant in the first line, and the Cauchy-Binet formula in the second line. The announced result follows by matching the first order terms w.r.t. h in these two expressions, and recalling (30) and the fact that $e \in \mathbb{R}^d$ is arbitrary. □

We introduce the following notations: for any $M_0 \in \mathrm{Perfect}_0(d)$,

$$\mathcal{E}(M_0) := \{ e_1 \wedge \cdots \wedge e_{d-1} \mid e_1, \cdots, e_{d-1} \in \Xi(M_0) \} \setminus \{0\},$$

and for any $\mu : \mathcal{E}(M_0) \to \mathbb{R}$,

$$D_\mu := \sum_{e \in \mathcal{E}(M_0)} \mu(e) e e^\top.$$

Corollary 4.5. *Given $M_0 \in \mathrm{Perfect}_0(d)$, one has $\mathcal{E}(M_0) \subset \mathcal{Z}_d$. In addition, for any $D \in \mathcal{S}^{++}(M_0)$ there exists $\mu : \mathcal{E}(M_0) \to \mathbb{R}_+$ such that $D^{-1} = D_\mu$.*

Proof. From (30) we see that $e_1 \wedge \cdots \wedge e_{d-1} \in \mathbb{Z}^d$ for all $e_1, \cdots, e_{d-1} \in \mathbb{Z}^d$, thus $\mathcal{E}(M_0) \subset \mathcal{Z}_d$ as announced. From $D \in \mathcal{S}^{++}(M_0)$, we have $\det(D) > 0$ hence $D^{-1} = \mathrm{adj}(D) / \det(D)$, and a decomposition $D = \sum_{e \in \Xi(M_0)} \lambda(e) e e^\top$ whose coefficients $\lambda : \Xi(M_0) \to \mathbb{R}_+$ are non-negative. We conclude using Lemma 4.4. □

We define a linear mapping $L_{M_0} : \mathcal{S}_d \to \mathbb{R}^{\mathcal{E}(M_0)}$, and the image $\mathcal{L}(M_0)$ of the set of perfect forms, as follows

$$L_{M_0}(M) := \big(\langle e, Me\rangle\big)_{e\in\mathcal{E}(M_0)}, \qquad \mathcal{L}(M_0) := L_{M_0}(\mathrm{Perfect}(d)). \tag{31}$$

Note that $\mathrm{Tr}(D_\mu M) = \sum_{e\in\mathcal{E}(M_0)} \mu(e)\,\mathrm{Tr}(ee^\top M) = \sum_{e\in\mathcal{E}(M_0)} \mu(e)\langle e, Me\rangle = \langle \mu, L_{M_0}(M)\rangle$.

Lemma 4.6. *The linear mapping $L_{M_0} : \mathcal{S}_d \to \mathbb{R}^{\mathcal{E}(M_0)}$ is injective, for any $M_0 \in \mathrm{Perfect}_0(d)$.*

Proof. If $L_{M_0}(M) = 0$, then $\mathrm{Tr}(D^{-1}M) = 0$ for all $D \in \mathcal{S}^{++}(M_0)$ by Corollary 4.5 and noting that $\langle e, Me\rangle = \mathrm{Tr}(Mee^\top)$ for any $e \in \mathbb{R}^d$. Thus $\mathrm{Tr}(D'M) = 0$ for all $D' \in \mathcal{S}_d$, by linearity since $\mathcal{S}^{++}(M_0)$ has non-empty interior. The result follows by recognizing the Frobenius inner product on $\mathcal{S}_d$, and choosing $D' = M$. □

Lemma 4.7. *There exists $n_0(d) \in \mathbb{Z}_{++}$ such that $\mathcal{L}(M_0) \subset \frac{1}{n_0(d)}\mathbb{Z}_+^{\mathcal{E}(M_0)}$ for all $M_0 \in \mathrm{Perfect}_0(d)$.*

Proof. The elements of $\mathrm{Perfect}_0(d)$ have rational coefficients, since they are the vertices of a polytope defined by rational inequalities, hence by finiteness there exists a positive integer such that $n_0(d)\,\mathrm{Perfect}_0(d) \subset \mathbb{Z}^{d\times d}$. By arithmetical equivalence $n_0(d)\,\mathrm{Perfect}(d) \subset \mathbb{Z}^{d\times d}$, since $\mathrm{GL}(\mathbb{Z}^d) \subset \mathbb{Z}^{d\times d}$. Recalling that $\mathcal{E}(M_0) \subset \mathcal{Z}_d$, we obtain that $n_0(d)\langle e, Me\rangle$ is an integer for all $e \in \mathcal{E}(M_0)$, which is positive since $\mathrm{Perfect}(d) \subset \mathcal{S}_d^{++}$. The result follows. □

We equip $\mathbb{R}_+^{\mathcal{E}(M_0)}$ with the componentwise partial ordering. Analogously, the tuples $a = (a_1, \cdots, a_n)$ and $b = (b_1, \cdots, b_n)$ satisfy $a \preceq b$ iff ($a_i \le b_i$ for all $1 \le i \le n$). An element a of a partially ordered set A is said to be minimal if there is no $b \in A \setminus \{a\}$ such that $b \preceq a$. There may be several minimal elements.

Corollary 4.8. *For any $M_0 \in \mathrm{Perfect}_0(d)$, the set of minimal elements of $\mathcal{L}(M_0)$, denoted $\mathcal{L}_1(M_0)$, is finite.*

Proof. It suffices to prove that the set of minimal elements of any $A \subset \mathbb{Z}_+^N$ is finite, where N is arbitrary. Hence, it suffices to prove that there is no sequence $(a_k)_{k\ge 0}$ of pairwise non-comparable elements in $\mathbb{Z}_+^N$. For contradiction, consider such a sequence, whose elements are denoted $a_k = (a_k^1, \cdots, a_k^N)$. Then for any $k \ge 0$, there exists $1 \le i \le N$ such that $a_k^i < a_0^i$. Thus, there exists a fixed $1 \le i \le N$ and a strictly increasing $\sigma : \mathbb{Z}_+ \to \mathbb{Z}_+$ such that $a^i_{\sigma(k)}$ is independent of $k \in \mathbb{Z}_+$. But this produces an infinite sequence of pairwise non-comparable elements in $\mathbb{Z}_+^{N-1}$, which by induction yields a contradiction (the case $N = 1$ being obvious), and the result is proved. □

We consider the multi-objective optimization problem consisting in minimizing L_{M_0} over a set $\mathcal{M} \subset \mathcal{S}_d$ (for instance $\mathcal{M} = \mathcal{M}_d$). The set of solutions to this problem, often referred to as the Pareto front, is defined as

$$\mathrm{Pareto}_{M_0}(\mathcal{M}) := \{M_1 \in \mathcal{M} \mid \nexists M_2 \in \mathcal{M} \setminus \{M_1\},\ L_{M_0}(M_2) \preceq L_{M_0}(M_1)\}.$$

Note that one cannot have $L_{M_0}(M_2) = L_{M_0}(M_1)$ in the above, by Lemma 4.6. In the following, we define the set $\mathrm{Perfect}_1(M_0)$ as follows:

$$\mathrm{Perfect}_1(M_0) := \mathrm{Perfect}(d) \cap \mathrm{Pareto}_{M_0}(\mathcal{M}_d). \tag{32}$$

Lemma 4.9. *For any $M_0 \in \mathrm{Perfect}_0(d)$, one has $L_{M_0}(\mathrm{Perfect}_1(M_0)) \subset \mathcal{L}_1(M_0)$.*

Proof. Let $M_1 \in \mathrm{Perfect}_1(M_0)$. Since $M_1 \in \mathrm{Perfect}(d)$, one has $L_{M_0}(M_1) \in \mathcal{L}(M_0)$. Assume that $L_{M_0}(M_1) \notin \mathcal{L}_1(M_0)$. Then there exists $a \in \mathcal{L}(M_0) \setminus \{L_{M_0}(M_1)\}$ such that $a \preceq L_{M_0}(M_1)$. By definition of $\mathcal{L}(M_0)$, there exists $M_2 \in \mathrm{Perfect}(d) \setminus \{M_1\}$ such that $a = L_{M_0}(M_2)$. Therefore $L_{M_0}(M_2) \preceq L_{M_0}(M_1)$, which is impossible since $M_1 \in \mathrm{Pareto}_{M_0}(\mathcal{M}_d)$. □

Lemma 4.10. *Let $\mathcal{M} \subset \mathcal{S}_d$, $M_0 \in \mathrm{Perfect}_0(d)$, and $\mu \in \mathbb{R}_{++}^{\mathcal{E}(M_0)}$. Then for any $M_1 \in \mathrm{argmin}_{M\in\mathcal{M}} \mathrm{Tr}(D_\mu M)$, one has $M_1 \in \mathrm{Pareto}_{M_0}(\mathcal{M})$.*

Proof. Assume that $M_1 \notin \mathrm{Pareto}_{M_0}(\mathcal{M})$. Then there exists $M_2 \in \mathcal{M} \setminus \{M_1\}$ such that $L_{M_0}(M_2) \preceq L_{M_0}(M_1)$. By Lemma 4.6, $L_{M_0}(M_2) \neq L_{M_0}(M_1)$. Therefore $\mathrm{Tr}(D_\mu M_2) = \langle \mu, L_{M_0}(M_2)\rangle < \langle \mu, L_{M_0}(M_1)\rangle = \mathrm{Tr}(D_\mu M_1)$, which contradicts the assumptions. □

Remark 4.11. When choosing $\mathcal{M} = \mathcal{M}_d$, Lemma 4.10 provides a sufficient condition for a perfect form $M_1 \in \mathrm{Perfect}(d)$ to belong to $\mathrm{Perfect}_1(M_0)$. It can be proved that this condition is also necessary, see Corollary 4.16 below.

Proof of Lemma 4.1. We choose $P(M_0) := \mathrm{Perfect}_1(M_0)$ which is defined in (32). By Corollary 4.8 and Lemma 4.9, $\mathrm{Perfect}_1(M_0)$ is a finite set. It remains to prove that for any $D \in \mathcal{S}^{++}(M_0)$, the function $M \mapsto \mathrm{Tr}(D^{-1}M)$ is minimized by some $M_1 \in \mathrm{Perfect}_1(M_0)$.

Given $D \in \mathcal{S}^{++}(M_0)$, there exist by Corollary 4.5 some weights $\mu \in \mathbb{R}_{+}^{\mathcal{E}(M_0)}$ such that $D^{-1} = D_\mu$. For $\varepsilon > 0$, we use the notation $\mu + \varepsilon\mathbb{1} := (\mu_E + \varepsilon)_{E\in\mathcal{E}(M_0)}$. The quantity $\mathrm{Tr}(D_{\mu+\varepsilon\mathbb{1}}M)$ is minimized for some $M \in \mathrm{Perfect}(d)$, which we call $M_{(\varepsilon)}$. By Lemma 4.10, one has $M_{(\varepsilon)} \in \mathrm{Pareto}_{M_0}(\mathcal{M}_d)$, hence $M_{(\varepsilon)} \in \mathrm{Perfect}_1(M_0)$. We conclude the proof by letting $M_1 := \lim_{\varepsilon\to 0} M_{(\varepsilon)}$, up to extracting a converging subsequence. □

Remark 4.12. An alternative, arguably simpler, replacement for the set $\mathrm{Perfect}_1(M_0)$ in the proof of Lemma 4.1 could be $\widetilde{\mathrm{Perfect}}_1(M_0) := \mathrm{Pareto}_{M_0}(\mathrm{Perfect}(d))$. Note that one has by definition $\mathcal{L}_1(M_0) = L_{M_0}(\mathrm{Pareto}_{M_0}(\mathrm{Perfect}(d)))$, and therefore, by Lemmas 4.6 and 4.9, $\mathrm{Perfect}_1(M_0) := \mathrm{Perfect}(d) \cap \mathrm{Pareto}_{M_0}(\mathcal{M}_d) \subset \mathrm{Pareto}_{M_0}(\mathrm{Perfect}(d)) =: \widetilde{\mathrm{Perfect}}_1(M_0)$. However, it is not obvious whether this inclusion is an equality, and Fig. 2 suggests that it may not be. Our motivation for choosing the definition (32) is that (i) it leads to a potentially sharper estimate in (29), and (ii) it allows computing $\mathrm{Perfect}_1(M_0)$ using the procedure described in Section 4.1.

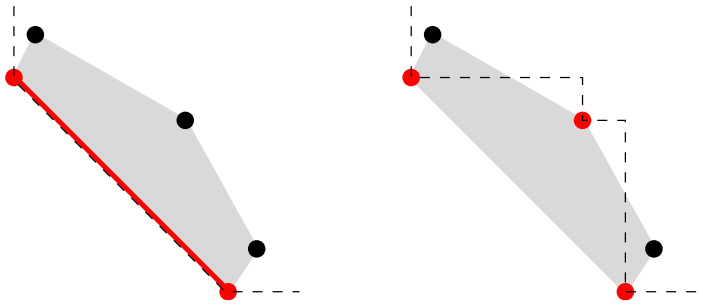

Figure 2: Multi-objective optimization over a polygon $K \subset \mathbb{R}^2$, and over its vertices. Left: Pareto front of the problem $\min_{x \in K}(x_1, x_2)$. Right: Pareto front of the problem $\min_{x \text{ vertex of } K}(x_1, x_2)$. Note that the second Pareto front is not included in the first one.

Remark 4.13 (Eutacticity). A perfect form $M \in \mathrm{Perfect}(d)$ is said eutactic if

$$M^{-1} = \sum_{e \in \Xi(M)} \mu(e) e e^\top,$$

for some positive coefficients $\mu : \Xi(M) \to \mathbb{R}_{++}$. This well-studied condition characterizes local minima of the determinant over Ryskov's polyhedron, see [30, Theorem 3.9]. Corollary 4.5 is strongly reminiscent of the eutaticity property, since it describes the decomposition of an inverse matrix as a sum of rank one matrices, but eventually we could not find a genuine connection.

4.1 Construction of the set $\mathrm{Perfect}_1(M_0)$ in dimension $d \leq 4$.

The objective of this section is to compute a constant $C(d)$ such that Theorem 4.3 holds. For that purpose, we choose $P(M_0) = \mathrm{Perfect}_1(M_0)$ in the statement of this result, for all $M_0 \in \mathrm{Perfect}_0(d)$, which fullfils the required conditions as established in the previous section. The finite set $\mathrm{Perfect}_1(M_0)$ is defined in (32) using the Pareto front for a multi-objective linear programming problem on Ryskov's polyhedron. Methods for computing Pareto fronts for multi-objective linear programming problems have been well-studied in the literature, see for instance [16]. We describe below such a method in our setting, and we apply it in dimensions $d \leq 4$, and $d = 5$ in Remark 4.21.

The only non-standard property of our setting is the fact that Ryskov's polyhedron is defined by infinitely many affine constraints. We overcome this by defining, for any $\alpha \in \mathbb{R}$,

$$\mathcal{M}_d^\alpha := \{M \in \mathcal{M}_d \mid \mathrm{Tr}(M) \leq \alpha\},$$

which is a bounded polyhedron in the usual sense, and is non-empty if α is large enough. We denote by $\mathcal{V}^\alpha(d)$ the set of vertices of $\mathcal{M}_d^\alpha$, and by $\mathcal{N}^\alpha(M) \subset \mathcal{V}^\alpha$ the set of neighbors

of any $M \in \mathcal{V}^\alpha(d)$ on the skeletal structure of $\mathcal{M}_d^\alpha$. Similarly to (32), for $M_0 \in \mathrm{Perfect}_0(d)$, we define

$$\mathcal{V}_1^\alpha(M_0) := \mathcal{V}^\alpha(d) \cap \mathrm{Pareto}_{M_0}(\mathcal{M}_d^\alpha).$$

Lemma 4.14. *Let $M \in \mathcal{S}_d$, $\alpha > \mathrm{Tr}(M)$, and $M_0 \in \mathrm{Perfect}_0(d)$. Then:*

(i) $M \in \mathrm{Perfect}(d) \iff M \in \mathcal{V}^\alpha(d)$.

(ii) $M \in \mathrm{Pareto}_{M_0}(\mathcal{M}_d) \iff M \in \mathrm{Pareto}_{M_0}(\mathcal{M}_d^\alpha)$.

(iii) $M \in \mathrm{Perfect}_1(M_0) \iff M \in \mathcal{V}_1^\alpha(M_0)$.

If moreover $M \in \mathrm{Perfect}(d)$ and $\alpha > \mathrm{Tr}(M')$ for any $M' \in \mathcal{N}(M)$, then:

(iv) $\mathcal{N}(M) = \mathcal{N}^\alpha(M)$.

Proof. Properties (i) and (iv) are true as local geometric properties of $\mathcal{M}_d$, and (iii) follows directly from (i) and (ii). It remains to prove (ii). The implication $M_1 \in \mathrm{Pareto}_{M_0}(\mathcal{M}_d) \implies M_1 \in \mathrm{Pareto}_{M_0}(\mathcal{M}_d^\alpha)$ follows directly from the definition of Pareto_{M_0}. Conversely, assume that $M_1 \notin \mathrm{Pareto}_{M_0}(\mathcal{M}_d)$. Then there exists $M_3 \in \mathcal{M}_d \setminus \{M_1\}$ such that $L_{M_0}(M_3) \preceq L_{M_0}(M_1)$. Let $M_4 := M_1 + t(M_3 - M_1)$ for $t > 0$ small enough so that $\mathrm{Tr}(M_4) \leq \alpha$. Then $M_4 \in \mathcal{M}_d^\alpha \setminus \{M_1\}$ and $L_{M_0}(M_4) \preceq L_{M_0}(M_1)$. Therefore $M_1 \notin \mathrm{Pareto}_{M_0}(\mathcal{M}_d^\alpha)$. □

We first describe a method for checking whether a perfect form $M \in \mathrm{Perfect}(d)$ belongs to $\mathrm{Perfect}_1(M_0)$.

Proposition 4.15. *Let $M_0 \in \mathrm{Perfect}_0(d)$, $\alpha \in \mathbb{R}$, and $M_1 \in \mathcal{V}^\alpha(d)$. Then $M_1 \in \mathcal{V}_1^\alpha(M_0)$ if and only if there exists $\mu \in \mathbb{R}_{++}^{\mathcal{E}(M_0)}$ satisfying one of the two following equivalent conditions:*

(i) $M_1 \in \mathrm{argmin}_{M \in \mathcal{M}_d^\alpha} \mathrm{Tr}(D_\mu M)$.

(ii) $\mathrm{Tr}(D_\mu M_1) \leq \mathrm{Tr}(D_\mu M)$, $\forall M \in \mathcal{N}^\alpha(M_1)$.

Proof. Conditions (i) and (ii) are equivalent by the usual optimality condition in linear programs. The result is a direct consequence of [16, Theorem 6.11], and of the fact that $\mathrm{Tr}(D_\mu M) = \langle \mu, L_{M_0}(M) \rangle$. □

Corollary 4.16. *Let $M_0 \in \mathrm{Perfect}_0(d)$ and $M_1 \in \mathrm{Perfect}(d)$. Then $M_1 \in \mathrm{Perfect}_1(M_0)$ if and only if there exists $\mu \in \mathbb{R}_{++}^{\mathcal{E}(M_0)}$ satisfying one of the two following equivalent conditions:*

(i) $D_\mu \in S^{++}(M_1)$.

(ii) $\mathrm{Tr}(D_\mu M_1) \leq \mathrm{Tr}(D_\mu M)$, $\forall M \in \mathcal{N}(M_1)$.

Proof. Conditions (i) and (ii) are equivalent by the usual optimality condition in linear programs, see (18). Now let $\alpha > \mathrm{Tr}(M_1)$ be large enough so that $\alpha > \mathrm{Tr}(M)$ for any $M \in \mathcal{N}(M_1)$. Then the result follows from Lemma 4.14 and Proposition 4.15. □

Checking the existence of $\mu \in \mathbb{R}_{++}^{\mathcal{E}(M_0)}$ satisfying the condition (ii) of Corollary 4.16 amounts to checking the feasibility of a linear program, which can be done algorithmically.

Let us now describe a method for checking whether a subset P of $\mathrm{Perfect}_1(M_0)$ coincides with the whole of $\mathrm{Perfect}_1(M_0)$. To this end, we use the following well-known property about the connectivity of the Pareto front of a multi-objective linear program.

Proposition 4.17. *Let $\alpha \in \mathbb{R}$, $M_0 \in \mathrm{Perfect}_0(d)$, and M_1, $M_2 \in \mathcal{V}_1^\alpha(M_0)$. Then there exists a family of matrices $(M^{(i)})_{1\leq i\leq I} \subset \mathcal{V}_1^\alpha(M_0)$, where I is a positive integer, such that $M^{(1)} = M_1$, $M^{(I)} = M_2$, and $M^{(i+1)} \in \mathcal{N}^\alpha(M^{(i)})$ for any $1 \leq i < I$.*

Proof. This follows from [16, Theorem 7.10]. □

Corollary 4.18. *Let $M_0 \in \mathrm{Perfect}_0(d)$, and let $P \subset \mathrm{Perfect}_1(M_0)$ be a nonempty set. If, for any $M \in P$, one has $\mathcal{N}(M) \cap \mathrm{Perfect}_1(M_0) \subset P$, then $P = \mathrm{Perfect}_1(M_0)$.*

Proof. Let us denote by M_1 some element of P. Let $M_2 \in \mathrm{Perfect}_1(M_0)$, and let us show that $M_2 \in P$.

Since $\mathrm{Perfect}_1(M_0)$ is a finite set, we may choose $\alpha \in \mathbb{R}$ large enough so that $\mathrm{Tr}(M) < \alpha$ and $\mathrm{Tr}(M') < \alpha$ for any $M \in \mathrm{Perfect}_1(M_0)$ and $M' \in \mathcal{N}(M)$. Then by Lemma 4.14, one has M_1, $M_2 \in \mathcal{V}_1^\alpha(M_0)$.

Let $(M^{(i)})_{1\leq i\leq I}$ be as in Proposition 4.17. Since $M^{(I)} = M_2$, it suffices to prove by induction that $M^{(i)} \in P$ for any $1 \leq i \leq I$.

We know that $M^{(1)} = M_1 \in P$. Now let $1 \leq i < I$ and assume that $M^{(i)} \in P$. Recall that $M^{(i+1)} \in \mathcal{N}^\alpha(M^{(i)})$ and $M^{(i+1)} \in \mathcal{V}_1^\alpha(M_0)$. Using Lemma 4.14, we deduce that $M^{(i+1)} \in \mathcal{N}(M^{(i)})$ and $M^{(i+1)} \in \mathrm{Perfect}_1(M_0)$. Then it follows from the assumptions that $M^{(i+1)} \in P$, which concludes the proof. □

In view of the above, we use Algorithm 3 in order to compute $\mathrm{Perfect}_1(M_0)$.

Algorithm 3 Computing $\mathrm{Perfect}_1(M_0)$

Initialization:

Let $\mu \in \mathbb{R}_{++}^{\mathcal{E}(M_0)}$ (chosen arbitrarily).

Let $M_1 \in \mathrm{Perfect}(d)$ be such that $D_\mu \in S^{++}(M_1)$ (computed using Algorithm 1).

$P \leftarrow \{M_1\}$.

Repeat

$P' \leftarrow \left(\bigcup_{M\in P} \mathcal{N}(M) \cap \mathrm{Perfect}_1(M_0)\right) \setminus P$

$P \leftarrow P \cup P'$.

while $P' \neq \emptyset$.

Return P.

Proposition 4.19. *Assuming that $M_0 \in \mathrm{Perfect}_0(d)$, Algorithm 3 terminates and returns $\mathrm{Perfect}_1(M_0)$.*

Proof. By Corollary 4.16, one has $M_1 \in \mathrm{Perfect}_1(M_0)$, see the initialization of Algorithm 3. Therefore, at each iteration, P is a nonempty subset of $\mathrm{Perfect}_1(M_0)$. Since the cardinality of P is increased at each iteration, and since $\mathrm{Perfect}_1(M_0)$ is finite by Corollary 4.8 and Lemma 4.9, the algorithm must terminate. After the termination condition is met, Corollary 4.18 guarantees that $P = \mathrm{Perfect}_1(M_0)$. □

Let us describe the results that we obtained by applying Algorithm 3 for $M_0 \in \mathrm{Perfect}_0(d)$, $2 \le d \le 4$. We obtain the cardinalities

$$\#\,\mathrm{Perfect}_1(\boldsymbol{A}_2) = 1, \quad \#\,\mathrm{Perfect}_1(\boldsymbol{A}_3) = 3, \quad \#\,\mathrm{Perfect}_1(\boldsymbol{A}_4) = 22, \quad \#\,\mathrm{Perfect}_1(\boldsymbol{D}_4) = 545.$$

For concreteness, the set $\mathrm{Perfect}_1(\boldsymbol{A}_2) = \{M_2^\times\}$ contains a single element, defined as

$$M_d^\times := \frac{1}{2}(3\,\mathrm{Id}_d - \mathbb{1}\mathbb{1}^\top) = \frac{1}{2}\begin{pmatrix} 2 & -1 & \cdots & -1 \\ -1 & 2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & -1 \\ -1 & \cdots & -1 & 2 \end{pmatrix}.$$

The elements of $\mathrm{Perfect}_1(\boldsymbol{A}_3)$ are obtained as

$$M_3^\times + b_i \otimes b_j, \quad 1 \le i < j \le 3,$$

where $b_i \otimes b_j := \frac{1}{2}(b_i b_j^\top + b_j b_i^\top)$ stands for the symmetrized outer product. The elements of $\mathrm{Perfect}_1(\boldsymbol{A}_4)$ have three possible forms:

$$M_4^\times + b_i \otimes b_j + b_j \otimes b_k + b_k \otimes b_l, \qquad M_4^\times + b_i \otimes b_j + b_j \otimes b_k + b_k \otimes b_i,$$
$$M_4^\times + b_i \otimes b_j + 2b_k \otimes b_l, \qquad \text{where } \{i,j,k,l\} = \{1,2,3,4\}.$$

We do list the elements of $\mathrm{Perfect}_1(\boldsymbol{D}_4)$, which is larger and more complex. Another observation is that $\mathrm{Perfect}_1(M_0)$ is invariant under the (transposed) unimodular isometries of M_0, as shown in Lemma 4.20 below, and we can thus consider the classes of its elements modulo the isometry group of M_0. The set $\mathrm{Perfect}_1(\boldsymbol{A}_3)$ has a single isometric equivalence class of cardinality 3. The elements of $\mathrm{Perfect}_1(\boldsymbol{A}_4)$ arithmetically equivalent to $\boldsymbol{A}_4$ (resp. $\boldsymbol{D}_4$) form a single isometric equivalence class of cardinality 12 (resp. 10). The elements of $\mathrm{Perfect}_1(\boldsymbol{D}_4)$ arithmetically equivalent to $\boldsymbol{A}_4$ (resp. $\boldsymbol{D}_4$) form three isometric equivalence classes of cardinality $48, 48, 288$ (resp. $1, 16, 144$).

Lemma 4.20. *Let $M_0 \in \mathrm{Perfect}_0(d)$ and $A \in \mathrm{GL}(\mathbb{Z}^d)$ be such that $A^\top M_0 A = M_0$. Then $AM_1A^\top \in \mathrm{Perfect}_1(M_0)$ for all $M_1 \in \mathrm{Perfect}_1(M_0)$.*

Proof. The result follows from the definition (32) of $\mathrm{Perfect}_1(M_0)$ and the observation that $\Xi(M_0) = A^{-1}\Xi(M_0)$, and $D_{A^{-1}E} = A^\top D_E A$ so that $\mathrm{Tr}(M_1 D_{A^{-1}E}) = \mathrm{Tr}(AM_1A^\top D_E)$ for any $E \subset \Xi(M_0)$, using the notations of (31). □

M	$\boldsymbol{A}_2$	$\boldsymbol{A}_3$	$\boldsymbol{D}_4$	$\boldsymbol{A}_4$	$\boldsymbol{D}_5$	$\boldsymbol{A}_5$	$\boldsymbol{A}_{5,0}$	$\boldsymbol{\phi}_2$
$\#\Xi(M)$	3	6	12	10	20	15	15	36
$\#\mathcal{N}(M)$	3	6	64	10	400	15	15	38 124
$\#\,\mathrm{Perfect}_1(M)$	1	3	545	22	11 386 192	880	333 712	?
$\#(\mathrm{Perfect}_1(M)/\mathrm{Isom}(M))$	1	1	6	2	6489	10	598	?
$r(M)^2$	8/3	2	6	4	36	6	20	?

Table 1: Properties of perfect forms related to the computation of $\mathrm{Perfect}_1$. We denote by $\mathrm{Perfect}_1(M)/\mathrm{Isom}(M)$ the collection of equivalence classes of $\mathrm{Perfect}_1(M)$ modulo unimodular isometries of M. We denote $r(M) := \max\{\|e\|_{M_1^{-1}} \mid e \in \Xi(M), M_1 \in \mathrm{Perfect}_1(M)\}$, thus $C(d) = \sqrt{d}\max\{r(M) \mid M \in \mathrm{Perfect}_0(d)\}$, see (29) with $P(M_0) = \mathrm{Perfect}_1(M_0)$.

From the sets $\mathrm{Perfect}_1(M_0)$ that we computed, we deduce the following values for the constants $C(d)$ defined in (29), when choosing $P(M_0) := \mathrm{Perfect}_1(M_0)$:

$$C(2) = 2\sqrt{2/3} \approx 1.633, \qquad C(3) = \sqrt{6} \approx 2.449, \qquad C(4) = 2\sqrt{6} \approx 4.899. \tag{33}$$

This result is new in dimension $d = 4$, and improves on [22, Theorem 4.11] when $d \in \{2, 3\}$. The constants (33) are directly related to the width of our finite difference stencils, and thus to the accuracy, parallelization potential, ease of implementing boundary conditions, of the resulting numerical schemes as discussed in Section 1.1.

Remark 4.21 (Extension to higher dimensions). The computation of $\mathrm{Perfect}_1(M_0)$ involves the numerical solution of linear optimization problems featuring $\binom{\#\Xi(M_0)}{d-1}$ unknowns and $\#\mathcal{N}(M_1)$ constraints for some $M_1 \in \mathrm{Perfect}(d)$, see Corollary 4.16, where $\#S$ denotes the cardinal of a set S. In dimension $d = 4$, the most difficult case is $\binom{12}{3} = 220$ unknowns and 64 constraints for $M_0 = M_1 = \boldsymbol{D}_4$, which is easily addressed using standard software packages.

Following the suggestion of a referee, we considered the case $d = 5$, where there are three perfect forms [8] : $\boldsymbol{D}_5$, $\boldsymbol{A}_5$ and $\boldsymbol{A}_{5,0}$, see Eqs. (19), (22) and (34) and Table 1. In dimension $d = 5$, the most difficult case is $\binom{20}{4} = 5845$ unknowns and 400 constraints for $M_0 = M_1 = \boldsymbol{D}_5$, which again is numerically tractable. Our numerical computation of $\mathrm{Perfect}_1(\boldsymbol{D}_5)$ takes advantages of the invariance property of Lemma 4.20, involves the numerical solution of more than 300 000 instances of the linear program of Corollary 4.16, and terminates in approximately 14 hours using a single core of a macbook pro laptop equipped an M1 Max processor and with 32 GB of RAM. Note that the computation terminates in seconds for $\boldsymbol{A}_2, \boldsymbol{A}_3, \boldsymbol{D}_4, \boldsymbol{A}_4, \boldsymbol{A}_5$, and in minutes for $\boldsymbol{A}_{5,0}$. A companion code is provided alongside this paper for the interested reader[4]. We obtain $C(5) = \sqrt{180} \approx 13.42$.

There are seven perfect forms in dimension $d = 6$, denoted $(\boldsymbol{\phi}_i)_{i=0}^6$ following [1, 8]. The case $M_0 = M_1 = \boldsymbol{\phi}_2$ leads to a linear program featuring $\binom{36}{5} = 376\,992$ unknowns and

[4] `https://github.com/Mirebeau/Perfect1-computation`

38 124 constraints, see Table 1. The computation of $\mathrm{Perfect}_1(\phi_2)$ likely involves solving millions of instances of such problems, but unfortunately we could not address a single one using our numerical ressources. We regard this extension as opportunity for future work.

5 Guarantees against checkerboard artifacts

In Section 5.1, we establish Theorem 5.1, which coincides with the part of Theorem 1.6 about the ϵ-spanning property. In Section 5.2, we show how this property may be used to prove the absence of checkerboard artifacts in some finite difference schemes.

5.1 The spanning property

Theorem 5.1. *Let $D \in \mathcal{S}_d^{++}$, where $d \leq 4$. Then for some constant $\varepsilon = \varepsilon(d) > 0$,*

$$\mathrm{Span}_{\mathbb{Z}}\{e \in \mathcal{Z}_d \mid \lambda^e(D) \geq \varepsilon \lambda_{\min}(D)\} = \mathbb{Z}^d.$$

Theorem 5.1 is a new result in dimension $d = 4$. In dimensions $d \in \{2, 3\}$ it was established in [5, section 4.3], for the numerical analysis of a discretization of a non-divergence form PDE with a point source singularity.

Lemma 5.2. *Let $d \in \mathcal{S}_d^{++}$ and $\lambda \in \Lambda(D)$. Then for some constant $\varepsilon = \varepsilon(d) > 0$*

$$\mathrm{Span}_{\mathbb{R}}\{e \in \mathcal{Z}_d \mid \lambda^e \geq \epsilon \lambda_{\min}(D)\} = \mathbb{R}^d.$$

Proof. The proof adapts and extends [5, Lemmas B.7 and B.8]. Let $n \geq d$, and let

$$c(d,n) := \min\Big\{ \max_{i_1 < \cdots < i_d} \lambda_{\min}\Big(\sum_{1 \leq j \leq d} e_{i_j} e_{i_j}^\top \Big) \;\Big|\; (e_1, \ldots, e_n) \in (\mathbb{R}^d)^n,\ \sum_{1 \leq i \leq n} e_i e_i^\top = \mathrm{Id}_d \Big\}.$$

Then $c(d,n)$ is positive, as the minimum of a positive continuous function over a non-empty compact set. By a simple change of variables, for any $(e_1, \ldots, e_n) \in (\mathbb{R}^d)^n$ such that $\sum_{i=1}^n e_i e_i^\top = D$, there exists $i_1 < \cdots < i_d$ such that $\sum_{j=1}^d e_{i_j} e_{i_j}^\top \succeq c(d,n) D$.

We now choose $n_d := \max\{\#(\Xi(M_0)) \mid M_0 \in \mathrm{Perfect}_0(d)\}$ and let $c_d := c(d, n_d)$. Let $M \in \mathrm{Perfect}(d)$ be minimizing in (11). By Proposition 2.3, one has $D = \sum_{e \in \Xi(M)} \lambda^e e e^\top$, where $\#(\Xi(M)) \leq n_d$. Thus there exists $\Xi \subset \Xi(M)$, $\#(\Xi) = d$, such that $\sum_{e \in \Xi} \lambda^e e e^\top \succeq c_d D$, and therefore $\mathrm{Span}_{\mathbb{R}} \Xi = \mathbb{R}^d$.

Let $e \in \Xi$, and let $v \in \mathbb{R}^d \setminus \{0\}$ be orthogonal to $\mathrm{Span}_{\mathbb{R}}(\Xi \setminus \{e\})$. Then

$$c_d \|v\|_D^2 \leq \sum_{e' \in \Xi} \lambda^{e'} \langle e', v \rangle^2 = \lambda^e \langle e, v \rangle^2 \leq \lambda^e \|e\|_{D^{-1}}^2 \|v\|_D^2 \leq C \lambda^e \|v\|_D^2 / \lambda_{\min}(D),$$

using Theorem 4.3 for the last inequality, with $C = C(d)$. Therefore $\lambda^e \geq (c_d/C)\lambda_{\min}(D)$, which concludes. □

Lemma 5.3. *Let $d \leq 4$ and $M = A^\top M_0 A \in \mathrm{Perfect}(d)$, where $A \in \mathrm{GL}(\mathbb{Z}^d)$ and $M_0 \in \mathrm{Perfect}_0(d)$. Let also $\Xi \subset \Xi(M)$ be such that $\mathrm{Span}_{\mathbb{R}} \Xi = \mathbb{R}^d$ and $\#(\Xi) = d$. If $M_0 = \boldsymbol{A}_d$, then $\mathrm{Span}_{\mathbb{Z}} \Xi = \mathbb{R}^d$. This remains true if $M_0 = \boldsymbol{D}_4$, except for the following three subsets: $\{A^{-1}e \mid e \in \Xi_i\} \subset \Xi(M)$, $1 \leq i \leq 3$, where*

$$\begin{aligned}
\Xi_1 &:= \{\pm b_1,\, \pm b_4,\, \pm(b_2 - b_3),\, \pm(b_1 - b_2 - b_3 + b_4)\},\\
\Xi_2 &:= \{\pm b_2,\, \pm(b_1 - b_3),\, \pm(b_4 - b_3),\, \pm(b_1 - b_2 + b_4)\},\\
\Xi_3 &:= \{\pm b_3,\, \pm(b_1 - b_2),\, \pm(b_4 - b_2),\, \pm(e_1 - b_3 + b_4)\}.
\end{aligned}$$

Proof, computer assisted. By Proposition 2.5, we may assume without loss of generality that $A = \mathrm{Id}_d$. Thus $\Xi = \{b_1, \cdots, b_d\}$ is a d-element subset of a known set, described in Propositions 2.6 and 2.7. In dimension $d = 4$, this gives $C^4_{\Xi(\boldsymbol{A}_4)} = C^4_{10} = 210$ or $C^4_{\Xi(\boldsymbol{D}_4)} = C^4_{12} = 495$ possibilities. We proceed by exhaustive computer enumeration, and consider all such subsets $\Xi = \{e_1, \cdots, e_d\}$ such that $\det(e_1, \cdots, e_d) \neq 0$, corresponding to the assumption that $\mathrm{Span}_{\mathbb{R}} \Xi = \mathbb{R}^d$. We check that $|\det(b_1, \cdots, b_d)| = 1$, which yields the announced result $\mathrm{Span}_{\mathbb{Z}} \Xi = \mathbb{Z}^d$, except in the case of Ξ_i, $1 \leq i \leq 3$, where $|\det(e_1, \cdots, e_d)| = 2$.

Alternatively, in the case where $M_0 = \boldsymbol{A}_d$, a formal argument which holds in arbitrary dimension d is presented in [5, Lemma B.6]. □

Before turning to the proof of Theorem 5.1, we present two results showing that this property is in a sense unexpected, and may not hold for variants of the proposed construction. Corollary 5.4 below shows that it is important to choose $\lambda(D)$ as the barycenter of $\Lambda(D)$, and not just any point of $\Lambda(D)$, in order for Theorem 5.1 to apply. Proposition 5.5 then shows that no such selection principle within $\Lambda(D)$ works in dimension $d = 5$.

Corollary 5.4. *Let $D := \sum_{e \in \Xi_1} e e^\top$, and let $\lambda \in \Lambda_d$ be defined by $\lambda^e := 1$ if $e \in \Xi_1$, $\lambda^e := 0$ otherwise. Then $\lambda \in \Lambda(D)$ and $\mathrm{Span}_{\mathbb{Z}}\{e \in \mathcal{Z}_d \mid \lambda^e > 0\} \neq \mathbb{Z}^d$.*

Proof. One has $\mathrm{Span}_{\mathbb{Z}}\{e \in \mathcal{Z}_d \mid \lambda^e > 0\} = \mathrm{Span}_{\mathbb{Z}}(\Xi_1) \neq \mathbb{Z}^d$ by Lemma 5.3. The fact that $\lambda \in \Lambda(D)$ follows from Proposition 2.4, using that $\Xi_1 \subset \Xi(\boldsymbol{D}_4)$. □

Proposition 5.5. *There exists $D_0 \in \mathcal{S}_5^{++}$ such that $\Lambda(D_0) = \{\lambda_0\}$ is a singleton, and $\mathrm{supp}(\lambda_0)$ is a basis of a sub-lattice of $\mathbb{Z}^5$ of index two. Thus, any mapping $\lambda : \mathcal{S}_5^{++} \to \Lambda_5$ such that $\lambda(D) \in \Lambda(D)$ for all $D \in \mathcal{S}_5^{++}$, fails the ε-spanning property at D_0 for all $\varepsilon > 0$.*

Proof. Let us introduce the perfect form $\boldsymbol{A}_{5,0} \in \mathrm{Perfect}(5)$, following the notations of [8], and its minimal vectors $\Xi(\boldsymbol{A}_{5,0}) = \{e_1, \cdots, e_{15}\} \subset \mathcal{Z}_5$, presented in column form:

$$\boldsymbol{A}_{5,0} = \frac{1}{4}\begin{pmatrix} 4 & 1 & 1 & -2 & -2 \\ 1 & 4 & 1 & -2 & -2 \\ 1 & 1 & 4 & -2 & -2 \\ -2 & -2 & -2 & 4 & 1 \\ -2 & -2 & -2 & 1 & 4 \end{pmatrix}, \quad \Xi(\boldsymbol{A}_{5,0}) = \begin{pmatrix} 1&0&0&0&0&1&1&0&0&0&0&1&1&0&1 \\ 0&1&0&0&0&0&0&1&1&0&0&1&0&1&1 \\ 0&0&1&0&0&0&0&0&0&1&1&0&1&1&1 \\ 0&0&0&1&0&1&0&1&0&1&0&1&1&1&1 \\ 0&0&0&0&1&0&1&0&1&0&1&1&1&1&1 \end{pmatrix}. \quad (34)$$

Let $D = \sum_{i=1}^{15} \lambda_i e_i e_i^\top$, where $\lambda_i \geq 0$ for all $1 \leq i \leq 15$ are any given non-negative coefficients. Since $\#\Xi(\boldsymbol{A}_{5,0}) = 15 = \dim(\mathcal{S}_5)$, the set of admissible decompositions $\Lambda(D)$ only contains a *single element* λ : namely the one defined by $\lambda^{e_i} = \lambda_i$, for all $1 \leq i \leq 15$, and $\lambda^e = 0$ for all $e \notin \Xi(\boldsymbol{A}_{5,0})$.

We now observe that $\det(e_1, e_2, e_{10}, e_{11}, e_{12}) = -2$, thus $(e_1, e_2, e_{10}, e_{11}, e_{12})$ is a basis of a sub-lattice of $\mathbb{Z}^5$ of index two. Letting $D_0 := \sum_{i \in \{1,2,10,11,12\}} e_i e_i^\top$, we have $D_0 \in \mathcal{S}_5^{++}$ since $(e_1, e_2, e_{10}, e_{11}, e_{12})$ is a basis of the vector space $\mathbb{R}^5$, and by the above $\Lambda(D_0) = \{\lambda_0\}$ where λ_0 is the indicator function of $\{e_1, e_2, e_{10}, e_{11}, e_{12}\}$. The result follows. □

Proof of Theorem 5.1. Let $M = A^\top M_0 A \in \mathrm{Perfect}(d)$ be minimizing in (11), where $A \in \mathrm{GL}(\mathbb{Z}^d)$ and $M_0 \in \mathrm{Perfect}_0(d)$. For now, let ϵ be as in Lemma 5.2. Then there exists $\Xi \subset \Xi(M)$, $\#(\Xi) = d$, such that $\mathrm{Span}_{\mathbb{R}} \Xi = \mathbb{R}^d$ and $\lambda^e(D) \geq \epsilon \lambda_{\min}(D)$, for any $e \in \Xi$.

We assume from now on that $M_0 = \boldsymbol{D}_4$ and $\Xi = \{A^{-1}e \mid e \in \Xi_i\}$, for some $1 \leq i \leq 3$, since otherwise Lemma 5.3 concludes the proof. Let $\kappa_1 := \min\{\lambda^e(D) \mid e \in \Xi\}$ and $\kappa_2 := \min\{\lambda^e(D) \mid e \in \Xi(M)\}$. We know that $\kappa_1 \geq \epsilon \lambda_{\min}(D)$. Let us show that $\kappa_2 \geq (\epsilon/4)\lambda_{\min}(D)$.

Note that $\Xi(\boldsymbol{D}_4) = \Xi_1 \cup \Xi_2 \cup \Xi_3$ and that $\sum_{e \in \Xi_i} e e^\top$ is independent of $i \in \{1, 2, 3\}$ (thus so is $\sum_{e \in \Xi_i} (A^{-1}e)(A^{-1}e)^\top$). We deduce that $\tilde{\lambda} \in \Lambda(D)$, where

$$\tilde{\lambda}^e := \begin{cases} \lambda^e(D) - \kappa_1 & \text{if } e \in \Xi, \\ \lambda^e(D) + \kappa_1/2 & \text{if } e \in \Xi(M) \setminus \Xi, \\ 0 & \text{else.} \end{cases}$$

Since $\Lambda(D)$ is a triangle and $\lambda(D)$ is its barycenter, see Proposition 3.6, the point $\hat{\lambda} := (3/2)\lambda(D) - (1/2)\tilde{\lambda}$ belongs to $\Lambda(D)$. By construction, there is $e_* \in \Xi(M) \setminus \Xi$ such that $\kappa_2 = \lambda^{e_*}(D)$. One has

$$0 \leq \hat{\lambda}^{e_*} = \tfrac{3}{2}\lambda^{e_*}(D) - \tfrac{1}{2}(\lambda^{e_*}(D) + \kappa_1/2) = \kappa_2 - \kappa_1/4.$$

Thus, for any $e \in \Xi(M)$, one has $\lambda^e(D) \geq \kappa_2 \geq \kappa_1/4 \geq (\epsilon/4)\lambda_{\min}(D)$. This concludes the proof, since $\mathrm{Span}_{\mathbb{Z}} \Xi(M) = \mathrm{Span}_{\mathbb{Z}} \Xi(\boldsymbol{D}_4) = \mathbb{Z}^d$. □

5.2 Absence of checkerboard artifacts

We establish the coercivity property of the discretized elliptic energy announced in Theorem 1.3. The spanning property of the scheme coefficients plays a key role, by ensuring the connectivity of the stencils of the finite differences discretization. Coercivity is used in the proof of the convergence rate Theorem 1.4, and also rules out checkerboard artifacts or similar high frequency oscillations that may arise with ill designed wide stencil finite difference schemes, such as (36) discussed below. The upper and lower estimates announced in Theorem 1.3 are established independently in Proposition 5.8 and Corollary 5.10.

Denote by $\delta_h^e u(x) := (u(x+he) - u(x))/h$ the first order finite difference of $u : \mathbb{T}_h^d \to \mathbb{R}$ in the direction $e \in \mathbb{Z}^d$. Recall that $\mathbb{T} := \mathbb{R}/\mathbb{Z}$ and $\mathbb{T}_h := (h\mathbb{Z})/\mathbb{Z}$, where the grid scale $h > 0$ is the inverse of a positive integer. The discretized elliptic energy (8) is defined as a sum of squares of finite differences. A basic ingredient of the proof is the ability to control such terms by others.

Lemma 5.6. *Let $(a_0, \dots, a_n) \in \mathbb{R}^{n+1}$. Then $(a_0 - a_n)^2 \le n\big((a_0 - a_1)^2 + \cdots + (a_{n-1} - a_n)^2\big)$.*

Proof. Apply the Cauchy-Schwarz inequality to $(a_0 - a_1, \dots, a_{n-1} - a_n)$ and $(1, \dots, 1)$. □

Corollary 5.7. *Let $e_0, \cdots, e_{n-1} \in \mathbb{Z}^d$ and $e := e_0 + \cdots + e_{n-1}$. Let $h > 0$, $x = x_0 \in h\mathbb{Z}^d$, and $x_{i+1} := x_i + he_i$ for all $0 \le i < n$. Then for any $u : h\mathbb{Z}^d \to \mathbb{R}$,*

$$\big(\delta_h^e u(x)\big)^2 \le n\Big[\big(\delta_h^{e_0} u(x_0)\big)^2 + \cdots + \big(\delta_h^{e_{n-1}} u(x_{n-1})\big)^2\Big].$$

Proof. Apply Lemma 5.6 with $a_i := u(x_i)$, $1 \le i \le n$. □

Let us mention that, in a spirit similar to Theorem 1.3, a Lipschitz estimate for solutions of the discretized eikonal equation (46) is established in [14], using closely related techniques and starting from the bound $\max\{0, \delta_h^e u(x)\} \le \max\{0, \delta_h^{e_0} u(x_0)\} + \cdots + \max\{0, \delta_h^{e_{n-1}} u(x_{n-1})\}$.

We next prove the easy part of Theorem 1.3, which is the announced upper bound. For convenience, we denote by $n(d, R) := \#\{b \in \mathbb{Z}^d \,|\, |b| \le R\}$ the number of integer points within a radius R.

Proposition 5.8 (Upper bound for the discrete elliptic energy)**.** *Consider weights $\lambda : \mathbb{T}^d \to \Lambda_d$ which are bounded by $\lambda_{\max}$ and R-supported. Then $Q_h(u) \le C\|\nabla_h u\|_{L_h^2}^2$, for all $h > 0$ and $u : h\mathbb{Z}^d \to \mathbb{R}$, where $C = C(d, \lambda_{\max}, R)$.*

Proof. Define for any $x \in \mathbb{T}_h$, recalling that $(b_1, \cdots, b_d)$ denotes the canonical basis of $\mathbb{R}^d$,

$$E_h^{x,R}(u) := \sum_{\substack{|b| \le R,\\ y := x + hb}} \sum_{1 \le i \le d} \big[(\delta_h^{b_i} u(y))^2 + (\delta_h^{-b_i} u(y))^2\big], \tag{35}$$

where implicitly the offset $b \in \mathbb{Z}^d$. Consider $e \in \mathbb{Z}^d$ with $|e| \le R$. Denote $e = (\sigma_1\alpha_1, \cdots, \sigma_d\alpha_d) \in \mathbb{Z}^d$, where $\sigma_1, \cdots, \sigma_d \in \{-1, 1\}$ and $\alpha_1, \cdots, \alpha_d \in \mathbb{Z}_+$, and observe that $e = \sigma_1 b_1 + \cdots + \sigma_1 b_1 + \sigma_2 b_2 + \cdots + \sigma_2 b_2 + \cdots$, where each term $\sigma_i b_i$ is repeated α_i times, for all $1 \le i \le d$. Then by Corollary 5.7 we successively obtain:

$$\big(\delta_h^e u(x)\big)^2 \le C_0 E_h^{x,R}(u), \quad \mathcal{E}_h^x(u) := \sum_{e \in \mathcal{Z}} \lambda^e(x)\big[\big(\delta_h^e u(x)\big)^2 + \big(\delta_h^{-e} u(x)\big)^2\big] \le C_1 E_h^{x,R}(u),$$

where $C_0 := R\sqrt{d} \ge \alpha_1 + \cdots + \alpha_d$ and $C_1 := 2\lambda_{\max} n(d, R) C_0$. Eventually, we conclude

$$\mathcal{E}_h(u) := \sum_{x \in \mathbb{T}_h^d} \mathcal{E}_h^x(u) \le C_1 \sum_{x \in \mathbb{T}_h^d} E_h^{x,R}(u) = 2n(d, R) C_1 \sum_{x \in \mathbb{T}_h^d} \|\nabla_h u(x)\|^2. \qquad \square$$

Before turning to the proof of the lower bound for the discrete elliptic energy, we would like to illustrate a situation where it may fail. The following is an example of a finite difference scheme featuring such undesirable checkerboard artifacts:

$$\frac{1}{2}\sum_{i=1}^{2}\frac{u(x+he_i)+u(x-he_i)-2u(x)}{h^2}=0 \quad \text{in } \mathbb{T}_h^2, \tag{36}$$

where $e_1 := (1,1)$, $e_2 := (1,-1)$, and $1/h \in 2\mathbb{Z}_{++}$. This scheme is a discretization of the equation $\Delta u(x) = 0$ in $\mathbb{T}^2$. Consider the two subsets of $\mathbb{T}_h^2$ defined as

$$(h\{(i,j)\in\mathbb{Z}^2 \mid i+j\in 2\mathbb{Z}\})/\mathbb{Z}^2, \qquad (h\{(i,j)\in\mathbb{Z}^2 \mid i+j\in 2\mathbb{Z}+1\})/\mathbb{Z}^2, \tag{37}$$

collecting points whose sums of coefficients are respectively even and odd after scaling by h^{-1}. Then the solutions of (36) are precisely the functions which are constant on each one of the sets (37), but are not necessarily constant on the whole of $\mathbb{T}_h^2$. The reason underlying this failure is that the offsets e_1, e_2 do not span $\mathbb{Z}^2$ with integer coefficients, although they span $\mathbb{R}^2$ with real coefficients, since $|\det(e_1,e_2)| = 2 \neq 1$. In fact $\mathrm{Span}_{\mathbb{Z}}\{e_1,e_2\} = \{(a,b)\in \mathbb{Z}^2 \mid a+b \equiv 0 \pmod 2\}$ is a subgroup of index two of $\mathbb{Z}^2$. In general, issues of similar nature can be expected to arise when the offsets of the finite difference scheme span a strict subgroup of $\mathbb{Z}^d$, of arbitrary index greater than or equal to two. The spanning property rules out this undesirable behavior.

The numerical scheme (36) can be obtained as the optimality condition of the discrete elliptic energy $\mathcal{E}_h$ defined by (5), with constant coefficients $\lambda^{\pm e_1}(x) = \lambda^{\pm e_2}(x) = 1/2$, $\lambda^e(x) = 0$ otherwise, and r.h.s. $f = 0$. These coefficients fail the spanning property, and the energy $\mathcal{E}_h$ attains its minimum (which is zero) on maps $u : \mathbb{T}_h^2 \to \mathbb{R}$ which are constant on each of the two sets (37), but are possibly not globally constant.

Lemma 5.9. *Consider weights $\lambda : \mathbb{T}^d \to \Lambda_d$ which are R-supported, K-Lipschitz, and ε-spanning. Then for any $0 < h \le h_0$ and any $x_0, x_* \in \mathbb{T}_h^d$ with $|x_0 - x_*| = h$, there exists $e_0, \cdots, e_{n-1} \in \mathbb{Z}^d \setminus \{0\}$ with $n \le n_0$ such that for all $0 \le i < n$*

$$\lambda^{\pm e_i}(x_i) \ge \varepsilon/2, \qquad \text{with } x_{i+1} := x_i + he_i$$

and in addition $x_n = x_*$. The constants $h_0 > 0$ and n_0 only depend on (d, R, K, ε).

Proof. A closely related argument appears in [14, Lemma 4.3], in the context of the discretization of a three-dimensional eikonal equation based on Selling's algorithm.

Let $h > 0$ and $x_0 \in \mathbb{T}_h^d$. By the ε-spanning property, there exist $e_1', \cdots, e_d' \in \mathbb{Z}^d$ such that $|\det(e_1', \cdots, e_d')| = 1$ and $\lambda^{\pm e_i'}(x_0) \ge \varepsilon$ for all $1 \le i \le d$. Denote by $A \in \mathrm{GL}(\mathbb{Z}^d)$ the matrix whose columns are $(e_1', \cdots, e_d')$, and note that $1 = |\det(A)| \le \|A\|^{d-1}\|A^{-1}\|^{-1}$, thus $\|A^{-1}\| \le \|A\|^{d-1} \le C_0 := (R\sqrt{d})^{d-1}$. (In the case where the weights are obtained as in Corollary 1.7, namely $\lambda(x_0) := \lambda(\mathcal{D}(x_0))$, one also has the possibly sharper estimate $\|A^{-1}\| \le C\mu(\mathcal{D}(x_0))$, see Proposition 4.2.)

Let $x_* = x_0 + hb$ where $\pm b$ is an element of the canonical basis of $\mathbb{R}^d$. Then we have $b = \alpha_1 e'_1 + \cdots + \alpha_d e'_d$, with $\alpha_1, \cdots, \alpha_d \in \mathbb{Z}$ such that $n := |\alpha_1| + \cdots + |\alpha_d| = |A^{-1}b|_{l^1} \leq \sqrt{d}\|A^{-1}\| \leq n_0$ with $n_0 := C_0\sqrt{d}$. We may assume that $\alpha_1, \cdots, \alpha_d \geq 0$, w.l.o.g. up to changing some of the $(e'_i)_{i=1}^d$ into their opposites, and we define $(e_0, \cdots, e_{n-1}) := (e'_1, \cdots, e'_1, e'_2, \cdots, e'_2, \cdots)$ where each element e'_i is repeated α_i times, for all $1 \leq i \leq d$.

Letting $x_{i+1} := x_i + he_i$, for all $0 \leq i < n$, we do have $x_n = x_0 + h(e_0 + \cdots + e_{n-1}) = x_0 + hb = x_*$. Finally, when $h \leq h_0 := \varepsilon/(2Kn_0R)$, we conclude that for all $0 \leq i < n$

$$\lambda^{\pm e_i}(x_i) \geq \lambda^{\pm e_i}(x_0) - Kh|e_0 + \cdots + e_{i-1}| \geq \varepsilon - Kn_0Rh \geq \varepsilon/2. \qquad \square$$

Corollary 5.10 (Lower bound for the elliptic energy)**.** *With the notations and assumptions of Lemma 5.9. One has $c\|\nabla_h u\|_{L^2_h}^2 \leq Q_h(u)$, for all $0 < h \leq h_0$ and $u : \mathbb{T}_h^d \to \mathbb{R}$, where the constant $c > 0$ only depends on (d, R, K, E).*

Proof. Consider $x_0 \in \mathbb{T}_h^d$, and $0 < h \leq h_0$, and let $x_* = x + hb_i$ for some $1 \leq i \leq d$ where $(b_i)_{i=1}^d$ denotes the canonical basis of $\mathbb{R}^d$. Then by Corollary 5.7 and Lemma 5.9

$$(\delta_h^{b_i} u(x_0))^2 \leq n\Big[\big(\delta_h^{e_0} u(x_0)\big)^2 + \cdots + \big(\delta_h^{e_{n-1}} u(x_{n-1})\big)^2\Big] \leq \frac{2n_0}{\varepsilon}\mathcal{E}_h^{x,R_1}(u),$$

where $\mathcal{E}_h^{x,R_1}(u) := \sum_{\substack{|b| \leq R_1 \\ y := x + hb}} \mathcal{E}_h^y(u)$ and $R_1 := n_0 R$. We then conclude

$$\|\nabla u\|_{L^2_h}^2 = h^d \sum_{x \in \mathbb{T}_h} \sum_{1 \leq i \leq d} \big(\delta^{b_i} u(x)\big)^2 \leq \frac{2dn_0}{\varepsilon} h^d \sum_{x \in \mathbb{T}_h} \mathcal{E}_h^{x,R_1}(u) = \frac{2dn_0 n(d, R_1)}{\varepsilon}\mathcal{E}_h(u). \qquad \square$$

6 Smooth decomposition

In dimensions $d \in \{2, 3\}$, the decomposition $\lambda(D)$ of a matrix $D \in \mathcal{S}_d^{++}$ defined in Proposition 1.5 coincides with *Selling's decomposition* [9, 32], see Section 3.1, and has piecewise linear coefficients. In this section, we construct a smooth variant of Selling's decomposition in dimension $d = 2$, as announced in Theorem 1.8 which gathers the results of Theorems 6.4 and 6.11 and Propositions 6.12 and 6.13 below. The extension of this construction to dimension $d = 3$ does not appear to be straightforward, and is an opportunity for future work. As a first step, we introduce two auxiliary functions, which are smooth approximations of the absolute value and of the median of three values.

Lemma 6.1 (Regularization of the absolute value)**.** *Let $\phi \in C^\infty(\mathbb{R}; [0, 1])$ be even and such that $\phi = 1$ on a neighborhood of the origin, and $\phi = 0$ on $[1, +\infty)$. Define*

$$\mathrm{sabs}(x) := \phi(x)(1 + x^2)/2 + (1 - \phi(x))|x|,$$

for all $x \in \mathbb{R}$. Then $\mathrm{sabs} \in C^\infty(\mathbb{R}; \mathbb{R})$*, and $g(x) := \mathrm{sabs}(x) - |x|$ obeys $0 \leq g \leq 1/2$ on $\mathbb{R}$, and $g = 0$ on $[1, +\infty)$.*

Proof. The function sabs is smooth as a sum of products of smooth functions, and since the coefficient $(1-\phi(x))$ vanishes in a neighborhood of the origin where the absolute value $|x|$ has a singularity.

Define $f(x) := (1+x^2)/2 - x$, so that $g(x) = \phi(x) f(|x|)$ for all $x \in \mathbb{R}$. Observing that $f(x) = (1-x)^2/2$ we obtain that $0 \le f \le 1/2$ on $[0,1]$. Since $0 \le \phi \le 1$ this implies $0 \le g \le 1/2$ on $[-1,1]$. Finally, g vanishes like ϕ outside of $[-1,1]$. □

In numerical applications, we may choose $\phi(x) = \mathbb{1}_{|x|\le 1}$, which is not smooth but nevertheless yields a regularized absolute value sabs with a Lipschitz gradient, and thus a modified Selling decomposition with a Lipschitz gradient. This convention is used in Fig. 1 (bottom). In a similar spirit, the next lemma introduces a smooth approximation of the median of three values, which in addition is expressed in terms of quantities obeying some invariance properties under Selling's superbase transformations, see the proof of Theorem 6.4.

Lemma 6.2 (Regularization of the median value)**.** *Define*

$$\operatorname{smed}(\rho_0,\rho_1,\rho_2) := S/(2\sqrt{Q+2S}), \quad \text{with } S := \rho_0\rho_1+\rho_1\rho_2+\rho_2\rho_0, \quad Q := (\rho_2-\rho_1)^2, \tag{38}$$

when $Q+2S>0$. *If* $0 \le \rho_0 \le \rho_1 \le \rho_2$ *and* $\rho_1 > 0$, *then* $\rho_1/(2\sqrt{2}) \le \operatorname{smed}(\rho_0,\rho_1,\rho_2) < \rho_1$.

Proof. Assume that $0 \le \rho_0 \le \rho_1 \le \rho_2$ and $\rho_1 > 0$. Noting that $Q+2S \ge \rho_1^2+\rho_2^2 > 0$, we obtain that $\operatorname{smed}(\rho_0,\rho_1,\rho_2)$ is well-defined. Since $f_0(t) := t/\sqrt{1+t}$ is non-decreasing on $[0,+\infty)$, and indeed $f_0'(t) = (1+t/2)(1+t)^{-3/2} \ge 0$ over that interval, we obtain that $S/\sqrt{Q+2S}$ is a non-decreasing function of S when Q is fixed, and thus that $\operatorname{smed}(\rho_0,\rho_1,\rho_2)$ is a non-decreasing function of $\rho_0 \in [0,\rho_1]$ when ρ_1 and ρ_2 are fixed. Thus it suffices to show that $\rho_1/(2\sqrt{2}) \le \operatorname{smed}(0,\rho_1,\rho_2)$ and that $\operatorname{smed}(\rho_1,\rho_1,\rho_2) < \rho_1$. By homogeneity, one may assume without loss of generality that $\rho_1 = 1$, hence it suffices to show that for $t = \rho_2 \ge 1$,

$$1/\sqrt{2} \le 2\operatorname{smed}(0,1,t) = t/\sqrt{(t-1)^2+2t} = t/\sqrt{1+t^2} =: f_1(t),$$
$$2 > 2\operatorname{smed}(1,1,t) = (1+2t)/\sqrt{(t-1)^2+2(1+2t)} = (1+2t)/\sqrt{t^2+2t+3} =: f_2(t).$$

Observing that f_1 and f_2 are strictly increasing on $[1,+\infty)$, by differentiation similarly to f_0, and that $f_1(1) = 1/\sqrt{2}$ and $f_2(t) \to 2$ as $t \to +\infty$, we conclude the proof. □

Similarly to Selling's original decomposition, the regularized decomposition that we introduce is defined using the notion of *superbase* of $\mathbb{Z}^2$, see Definition 3.9. By Proposition 3.13, any matrix $D \in \mathcal{S}_2^{++}$ admits a D-obtuse superbase $(v_0,v_1,v_2) \in (\mathbb{Z}^2)^3$. It satisfies $|\det(v_1,v_2)| = 1$, $v_0+v_1+v_2 = 0$, and $\rho_i \ge 0$ for any $i \in \{0,1,2\}$ where

$$\rho_0 := -\langle v_1, Dv_2\rangle, \qquad \rho_1 := -\langle v_0, Dv_2\rangle, \qquad \rho_2 := -\langle v_0, Dv_1\rangle. \tag{39}$$

Selling's formula, which can be deduced from Propositions 3.5 and 3.13, then asserts that

$$D = \sum_{0\le i\le 2} \rho_i e_i e_i^\top, \qquad \text{where } e_i := v_i^\perp. \tag{40}$$

Lemma 6.3. *Let $D \in \mathcal{S}_2^{++}$, and let $v = (v_0, v_1, v_2)$ be a D-obtuse superbase. Assume, up to permuting v, that the Selling weights satisfy $0 \leq \rho_0 \leq \rho_1 \leq \rho_2$. Then $\rho_2 \|v_2\|^2 \leq \lambda_{\max}(D)$ and $\rho_0 + \rho_1 \geq \|v_2\|^2 \lambda_{\min}(D)$, and in particular $\rho_2 \leq \lambda_{\max}(D)$ and $\rho_1 \geq \lambda_{\min}(D)/2 > 0$.*

Proof. The existence of a suitable permutation of v is clear. Denoting $e_i := v_i^\perp$, for all $0 \leq i \leq 2$, we obtain $\rho_2 \|e_2\|^4 \leq \langle e_2, De_2 \rangle \leq \lambda_{\max}(D)\|e_2\|^2$. On the other hand $\lambda_{\min}(D)\|v_2\|^2 \leq \langle v_2, Dv_2 \rangle = \rho_0 + \rho_1$, since $|\langle e_0, v_2 \rangle| = |\langle e_1, v_2 \rangle| = |\det(e_1, e_2)| = 1$. We conclude noting that $\|e_2\| = \|v_2\| \geq 1$, since $v_2 \in \mathbb{Z}^2 \setminus \{0\}$. □

Consider $D \in \mathcal{S}_2^{++}$ and a D-obtuse superbase $v = (v_0, v_1, v_2)$, permuted so that $\rho_0 \leq \rho_1 \leq \rho_2$. Then the *regularized Selling decomposition* $\tilde{\lambda}(D) \colon \mathcal{Z}_2 \to \mathbb{R}$ is defined as

$$\tilde{\lambda}^e(D) := \begin{cases} \rho_0 + w/2 & \text{if } e := \pm v_0^\perp, \\ \rho_1 - w & \text{if } e := \pm v_1^\perp, \\ \rho_2 - w & \text{if } e := \pm v_2^\perp, \\ w/2 & \text{if } e := \pm (v_1 - v_2)^\perp, \\ 0 & \text{else,} \end{cases} \tag{41}$$

where $w := m\,\mathrm{sabs}(\rho_0/m) - \rho_0$, with $m := \mathrm{smed}(\rho_0, \rho_1, \rho_2)$. For comparison, the coefficients $\lambda(D) \in \Lambda_2$ of Proposition 1.5, which correspond to the usual Selling decomposition (40), are obtained by choosing $w = 0$ in (41).

Theorem 6.4. *Let $D \in \mathcal{S}_2^{++}$. Then the decomposition $\tilde{\lambda}(D)$ is independent of the choice of superbase v, provided it is D-obtuse and such that $\rho_0 \leq \rho_1 \leq \rho_2$. It is consistent, in the sense that*

$$\sum_{e \in \mathcal{Z}_2} \tilde{\lambda}^e(D) e e^\top = D, \tag{42}$$

and its weights are nonnegative and have C^∞ regularity.

The next four lemmas are devoted to the proof of Theorem 6.4. We prove in Lemma 6.5 that the equality (42) holds and that the weights (41) of the regularized Selling decomposition are nonnegative. We establish in Lemma 6.7 that these weights are smooth (and in particular uniquely defined) in the neighborhood of a matrix admitting a strictly D-obtuse superbase, and in Lemma 6.8 a similar regularity result in the complementary case.

Lemma 6.5. *The regularized Selling decomposition is consistent and has non-negative weights.*

Proof. We use the notations of Theorem 6.4. Since D is nondegenerate, at most one of Selling's non-negative weights ρ_0, ρ_1, ρ_2 vanishes, see (40). Upon sorting, this yields $0 \leq \rho_0 \leq \rho_1 \leq \rho_2$ and $\rho_1 > 0$, and in particular $\mathrm{smed}(\rho_0, \rho_1, \rho_2)$ is well-defined and obeys the bounds of Lemma 6.2. It follows that $0 \leq w := mg(\rho_0/m) \leq m/2 < \rho_1/2$, by

Lemmas 6.1 and 6.2, hence the weights (41) are non-negative as announced. Using again (41) we compute

$$\sum_{e\in\mathcal{Z}_2}\tilde{\lambda}^e(D)ee^\top=\sum_{0\le i\le 2}\rho_i e_i e_i^\top+\frac{w}{2}\big(e_0e_0^\top-2e_1e_1^\top-2e_2e_2^\top+(e_1-e_2)(e_1-e_2)^\top\big),$$

where $e_i := v_i^\perp$. Observing that $e_0 = -e_1 - e_2$ and recalling the parallelogram identity $(e_1+e_2)(e_1+e_2)^\top+(e_1-e_2)(e_1-e_2)^\top = 2(e_1e_1^\top+e_2e_2^\top)$, one obtains that the second term in the r.h.s. vanishes, and consistency (42) therefore follows from Selling's formula (40). □

A superbase v is said to be *strictly* D-obtuse, where $D \in \mathcal{S}_2^{++}$, if all the weights (39) are positive.

Lemma 6.6. *Let $D \in \mathcal{S}_2^{++}$, and let $v = (v_0, v_1, v_2)$ be a D-obtuse superbase with Selling weights $\rho_0 \le \rho_1 \le \rho_2$. If v is strictly D-obtuse, i.e. $\rho_0 > 0$, then any other D-obtuse superbase coincides with v up to a permutation and a global change of sign. Otherwise if $\rho_0 = 0$ then $\tilde{v} := (v_2 - v_1, v_1, -v_2)$ is also D-obtuse, and any other D-obtuse superbase coincides with v or $\tilde{v}$ up to a permutation and a global change of sign.*

Proof. Let us recall that Selling's decomposition (40) corresponds to the coefficients $\lambda(D)$ of Proposition 1.5 in dimension $d \in \{2, 3\}$, which are uniquely defined. Let $\hat{v} = (\hat{v}_0, \hat{v}_1, \hat{v}_2)$ be an arbitrary D-obtuse superbase. If $\rho_0 > 0$ then $\{\pm v_0^\perp, \pm v_1^\perp, \pm v_2^\perp\} = \mathrm{supp}(\lambda(D)) = \{\pm\hat{v}_0^\perp, \pm\hat{v}_1^\perp, \pm\hat{v}_2^\perp\}$, from which the first point follows. If $\rho_0 = 0$, then recalling that $\rho_1 > 0$ by Lemma 6.3 we obtain that $\{\pm v_1^\perp, \pm v_2^\perp\} = \mathrm{supp}(\lambda(D)) \subset \mathcal{Z}_2$ contains two elements of $\hat{v}$. Thus $\hat{v}_1 = v_1$ and either $\hat{v}_2 = v_2$ or $\hat{v}_2 = -v_2$, up to a global change of sign and permutation of $\hat{v}$, and therefore $\hat{v} = v$ or $\hat{v} = \tilde{v}$ respectively, since $\hat{v}_0 = -\hat{v}_1 - \hat{v}_2$, as announced. Finally, we note that $\langle v_1, Dv_2\rangle = -\rho_0 = 0$, $\langle v_2 - v_1, Dv_1\rangle = -\|v_1\|_D^2 \le 0$ and $\langle v_2 - v_1, -Dv_2\rangle = -\|v_2\|_D^2 \le 0$, so that $\tilde{v}$ is D-obtuse as announced, which concludes. □

Lemma 6.7. *The weights of the regularized Selling decomposition are smooth in the neighborhood of any $D^* \in \mathcal{S}_2^{++}$ which* admits *a strictly D^*-obtuse superbase $v = (v_0, v_1, v_2)$*

Proof. Denote by $0 < \rho_0^* \le \rho_1^* \le \rho_2^*$ the Selling weights of D^* defined by (39), up to permuting the superbase v.

- *Case $\rho_0^* = \rho_1^*$.* Then $m^* = \mathrm{smed}(\rho_0^*, \rho_1^*, \rho_2^*) < \rho_1^* = \rho_0^*$ by Lemma 6.2. Thus

  $$m^*\,\mathrm{sabs}(\rho_0^*/m^*) = m^*\rho_0^*/m^* = \rho_0^*,$$

  by Lemma 6.1 since $\rho_0^*/m^* > 1$. Therefore $w^* := m^*\,\mathrm{sabs}(\rho_0^*/m^*) - \rho_0^* = 0$ and the weights (41) coincide with those of Selling's original decomposition. In particular, they are uniquely defined.

  Likewise, for D close enough to D^*, we obtain $\rho_0/m > 1$ by continuity, and thus $w = 0$, with obvious notations. As a result, the classical and the regularized Selling

decompositions have the same weights and offsets. The weights ρ_0, ρ_1, ρ_2 of Selling's decomposition (39) are linear functions of D, in a neighborhood of D^*, hence they are smooth as announced.

- *Case* $\rho_0^* < \rho_1^*$. Then for D close enough to D^* one has likewise $\rho_0 < \rho_1$. Noting that (38) is a symmetric expression of ρ_1 and ρ_2, we obtain that $m = \mathrm{smed}(\rho_0, \rho_1, \rho_2)$ depends smoothly on D in a neighborhood of D^*, even in the case where $\rho_1^* = \rho_2^*$. (This also shows that the weights (41) are uniquely defined, even when $\rho_1^* = \rho_2^*$.) Thus w and the weights (41) are also smooth by composition, which concludes. □

Lemma 6.8. *The weights of the regularized Selling decomposition are smooth in the neighborhood of any* $D^* \in \mathcal{S}_2^{++}$ *which* does not admit *a strictly* D^**-obtuse superbase.*

Proof. By Proposition 3.13 there exists a D^*-obtuse superbase $v = (v_0, v_1, v_2)$, whose weights $\rho_0^* \le \rho_1^* \le \rho_2^*$ defined by (39) satisfy $\rho_0^* = 0$ since v is *not* strictly D^*-obtuse, and $\rho_1^* > 0$ since D^* is non-degenerate. The only other D^*-obuse superbase is $\tilde{v} := (\tilde{v}_0, \tilde{v}_1, \tilde{v}_2) := (v_1 - v_2, -v_1, v_2)$ by Lemma 6.6, up to a global change of sign and a permutation. The change of sign is irrelevant, and the permutation is fixed by imposing that $0 = \rho_0^* = -\langle \tilde{v}_1, D^* \tilde{v}_2 \rangle$, $\rho_1^* = -\langle \tilde{v}_0, D^* \tilde{v}_2 \rangle$, and $\rho_2^* = -\langle \tilde{v}_0, D^* \tilde{v}_1 \rangle$ (an ambiguity remains in the special case where $\rho_1^* = \rho_2^*$, but it is harmless since (38) and (41) are symmetric expressions of ρ_1 and ρ_2).

Consider D in the neighborhood of D^*, and denote by (ρ_0, ρ_1, ρ_2) and $(\tilde{\rho}_0, \tilde{\rho}_1, \tilde{\rho}_2)$ the weights of Selling's formula associated with the superbases v and $\tilde{v}$, namely $\rho_i := -\langle v_{i-1}, D v_{i+1} \rangle$ and $\tilde{\rho}_i := -\langle v_{i-1}, D v_{i+1} \rangle$ with circular indexing (note that one of v or $\tilde{v}$ may not be D-obtuse, and thus define negative weights). Then

$$\begin{aligned}
\tilde{\rho}_0 &= -\langle -v_1, D v_2 \rangle = -\rho_0, \\
\tilde{\rho}_1 &= -\langle v_1 - v_2, D v_2 \rangle = -\langle 2v_1 + v_0, D v_2 \rangle = \rho_1 + 2\rho_0, \\
\tilde{\rho}_2 &= -\langle v_1 - v_2, -D v_1 \rangle = -\langle -2v_2 - v_0, -v_1 \rangle = \rho_2 + 2\rho_0.
\end{aligned}$$

Therefore $m := \mathrm{smed}(\rho_0, \rho_1, \rho_2) = \mathrm{smed}(\tilde{\rho}_0, \tilde{\rho}_1, \tilde{\rho}_2) =: \tilde{m}$, in view of the identities

$$\begin{aligned}
Q &:= (\rho_2 - \rho_1)^2 = (\tilde{\rho}_2 - \tilde{\rho}_1)^2 =: \tilde{Q}, \\
S &:= \rho_0\rho_1 + \rho_1\rho_2 + \rho_2\rho_0 = \det \begin{pmatrix} \rho_0 + \rho_2 & \rho_0 \\ \rho_0 & \rho_0 + \rho_1 \end{pmatrix} = \det(D) = \tilde{\rho}_0\tilde{\rho}_1 + \tilde{\rho}_1\tilde{\rho}_2 + \tilde{\rho}_2\tilde{\rho}_0 =: \tilde{S}.
\end{aligned}$$

Note that the matrix in the second line represents the quadratic form (40) in the unimodular basis (v_1, v_2). It follows that $\omega := m\,\mathrm{sabs}(\rho_0/m) = \tilde{m}\,\mathrm{sabs}(\tilde{\rho}_0/\tilde{m})$, since sabs is even, and therefore $w = \omega - \rho_0$ and $\tilde{w} = \omega + \rho_0$. The weights and offsets of Selling's regularized formula with $\tilde{v}$ are thus

$$\begin{aligned}
(\tilde{\rho}_0 + \tilde{w}/2,\, \tilde{\rho}_1 - \tilde{w},\, \tilde{\rho}_2 - \tilde{w},\, \tilde{w}/2) &= (\omega/2 - \rho_0/2,\, \rho_1 - \omega + \rho_0,\, \rho_2 - \omega + \rho_0,\, \omega/2 + \rho_0/2), \\
(\tilde{v}_0^\perp,\, \tilde{v}_1^\perp,\, \tilde{v}_2^\perp,\, \tilde{v}_1^\perp - \tilde{v}_2^\perp) &= (v_1^\perp - v_2^\perp,\, -v_1^\perp,\, v_2^\perp,\, -v_1^\perp - v_2^\perp).
\end{aligned}$$

Compare with the weights and offsets of Selling's regularized formula with v, namely

$$(\rho_0 + w/2,\, \rho_1 - w,\, \rho_2 - w, w/2) = (\omega/2 + \rho_0/2,\, \rho_1 - \omega + \rho_0,\, \rho_2 - \omega + \rho_0, \omega/2 - \rho_0/2),$$
$$(v_0^\perp,\, v_1^\perp,\, v_2^\perp,\, v_1^\perp - v_2^\perp) = (-v_1^\perp - v_2^\perp,\, v_1^\perp,\, v_2^\perp,\, v_1^\perp - v_2^\perp).$$

The decompositions agree, up to permuting the first and last weight and offset, and changing the sign of the second offset. Since they are defined by smooth expressions, the result follows. □

Quantitative regularity estimates. We quantify below the Lipschitz constant of the coefficients and of their gradients, with respect to the square root $\mu(D)$ of the condition number of the matrix $D \in \mathcal{S}_2^{++}$. From now on, all results of this section remain valid if $\phi(x) = \mathbb{1}_{|x|\leq 1}$ is chosen in Lemma 6.1 instead of a suitable $\phi \in C^\infty(\mathbb{R}; [0,1])$, except that the weights of the decomposition only have $W^{2,\infty}_{\mathrm{loc}}$ regularity (i.e. continuous first order derivatives and locally bounded second order derivatives) rather than C^∞ regularity.

The mapping considered in Lemma 6.9 (i) is known as the *perspective function* of f.

Lemma 6.9. *The following holds for any $f \in C^2(\mathbb{R}^d, \mathbb{R})$.*

(i) Define $g(x,\rho) := \rho f(x/\rho)$. Then $|\nabla g(x,\rho)| \leq 5\max\{|f(x/\rho)|, |\nabla f(x/\rho)|\}$, and $\|\nabla^2 g(x,\rho)\| \leq (16/\rho)\|\nabla^2 f(x/\rho)\|$, for all $x \in \mathbb{R}^d$ and all $\rho > 0$ such that $|x| \leq 3\rho$.

(ii) Define $h(x) := f(Bx)$, where B is a matrix of shape $d' \times d$. Then $|\nabla h(x)| \leq \|B\||\nabla f(Bx)|$ and $\|\nabla^2 h(x)\| \leq \|B\|^2\|\nabla^2 f(Bx)\|$, for all $x \in \mathbb{R}^d$.

Proof. Note that $\nabla g = (\nabla_x g,\, \partial_\rho g) \in \mathbb{R}^{d+1}$, and that the hessian $\nabla^2 g$ is built of the blocks $(\nabla_x^2 g,\, \partial_\rho \nabla_x g,\, \partial_\rho^2 g)$. The announced estimates easily follow from the exact expressions

$$\partial_\rho g(x,\rho) = f\big(\frac{x}{\rho}\big) - \frac{1}{\rho}\langle \nabla f\big(\frac{x}{\rho}\big), x\rangle, \qquad \nabla_x g(x,\rho) = \nabla f(\frac{x}{\rho}),$$
$$\partial_\rho^2 g(x,\rho) = \frac{1}{\rho^3}\langle x, \nabla f^2\big(\frac{x}{\rho}\big)x\rangle, \qquad \partial_\rho \nabla_x g(x,\rho) = -\frac{1}{\rho^2}\nabla^2 f(\frac{x}{\rho})x, \quad \nabla_x^2 g(x,\rho) = \frac{1}{\rho}\nabla^2 f(\frac{x}{\rho}),$$
$$\nabla h(x) = B^\top \nabla f(Bx), \qquad \nabla^2 h(x) = B^\top \nabla^2 f(Bx) B.$$ □

Lemma 6.10. *Define the triangular domain $T := \{(s,t) \mid 0 \leq s \leq t \leq 1\}$, and the functions $m : T \to \mathbb{R}$ and $\omega : T \setminus \{(0,0)\} \to \mathbb{R}$ by*

$$m(s,t) := \mathrm{smed}(s,t,1) = \frac{s + t + st}{2\sqrt{1 + t^2 + 2s + 2st}}, \qquad \omega(s,t) := m(s,t)\,\mathrm{sabs}\Big(\frac{s}{m(s,t)}\Big).$$

Then $|\nabla \omega(s,t)| \leq C_0$ and $\|\nabla^2 \omega(s,t)\| \leq C_1/t$, with constants C_0, C_1 depending only on ϕ.

Proof. One has $\omega = g \circ \tilde{m}$ where $g(x,\rho) := \rho\,\mathrm{sabs}(x/\rho)$ and $\tilde{m}(s,t) := (s, m(s,t))$. One has $m \in C^\infty(T)$, as a composition of smooth functions and since the denominator does not vanish, and thus ∇m and $\nabla^2 m$ are uniformly bounded over T, hence also $\nabla\tilde{m}$ and $\nabla^2\tilde{m}$.

Applying Lemma 6.9 (i) to sabs, and noting that sabs$'$ and sabs$''$ are bounded over $\mathbb{R}$, we obtain that $|\nabla g(x,\rho)| \le C$ and $\|\nabla^2 g(x,\rho)\| \le C/\rho$ for some constant C. Furthermore, over the domain of evaluation, one has $|x| = s \le t \le 3m(s,t) = 3\rho$ as required, recalling that $m(s,t) \ge t/(2\sqrt{2})$ by Lemma 6.1. The announced estimates follows by composition. □

Theorem 6.11. *For any $D_* \in \mathcal{S}_2^{++}$ and $e \in \mathcal{Z}_2$, one has for some constant $K = K(\phi)$*

$$|\nabla\tilde{\lambda}^e(D_*)| \le K\mu(D_*)^2, \qquad \|\nabla^2\tilde{\lambda}^e(D_*)\| \le K\mu(D_*)^4/\lambda_{\min}(D_*).$$

Proof. Let (v_0^*, v_1^*, v_2^*) be a D_*-obtuse superbase, sorted such that the Selling weights (39) obey $\rho_0^* \le \rho_1^* \le \rho_2^*$. Let also $A \in \mathrm{GL}(\mathbb{Z}^2)$ be such that $v_i^* = Ab_i$, for all $0 \le i \le 2$, where (b_0, b_1, b_2) is the canonical superbase see Remark 3.10, and note that $\|A\| \le C_0\mu(D)$ for some constant C_0 by Propositions 3.7, 3.11 and 3.13. For $D \in S_2^{++}$ close enough to D_*, the coefficients of Selling's smoothed decomposition (41) are obtained as the composition of

$$\begin{aligned} f_1 : D &\mapsto D' := A^\top DA, \\ f_2 : D' &\mapsto (\rho_0, \rho_1, \rho_2) := -(\langle b_1, D'b_2\rangle, \langle b_2, D'b_0\rangle, \langle b_0, D'b_1\rangle) \\ f_3 : (\rho_0, \rho_1, \rho_2) &\mapsto \boldsymbol{\omega} := \rho_2\,\omega(\rho_0/\rho_2,\ \rho_1/\rho_2), \end{aligned}$$

followed by the fixed linear mapping $(\rho_0, \rho_1, \rho_2, \boldsymbol{\omega}) \mapsto (\rho_0 + w/2,\ \rho_1 - w,\ \rho_2 - w,\ w/2)$ where $w := \boldsymbol{\omega} - \rho_0$. Therefore

$$|\nabla(f_3 \circ f_2 \circ f_1)(D_*)| \le C\|A\|^2, \quad \|\nabla^2(f_3 \circ f_2 \circ f_1)(D_*)\| \le C\|A\|^4 \frac{1}{\rho_2^*}\frac{1}{\rho_1^*/\rho_2^*} = C\frac{\|A\|^4}{\rho_1^*},$$

where we applied Lemma 6.9 (ii) to the linear mappings $f_1 : \mathcal{S}_2 \to \mathcal{S}_2$ and $f_2 : \mathcal{S}_2 \to \mathbb{R}^3$, noting that $\|f_1(D)\| = \|A^\top DA\| \le \|A\|^2\|D\|$ and that f_2 is fixed hence bounded independently of D_*. Regarding f_3 we used the estimates of Lemma 6.10, and applied Lemma 6.9 (i) with $x := (\rho_0^*, \rho_1^*)$ and $\rho := \rho_2^*$, noting that $|x| \le \rho_0^* + \rho_1^* \le 2\rho_2^*$ as required. We conclude recalling that $\|A\| \le \mu(D)$, and that $\rho_1^* \ge \lambda_{\min}(D_*)/2$ by Lemma 6.3. □

Radius and spanning property. The next results complete the proof of Theorem 1.8.

Proposition 6.12. *For any $D \in \mathcal{S}_2^{++}$ and $e \in \mathrm{supp}(\tilde{\lambda}(D))$ one has $\|e\|_{D^{-1}} \le C\lambda_{\min}(D)^{-\frac{1}{2}}$, and in particular $|e| \le C\mu(D)$, where C is an absolute constant.*

Proof. Let $v = (v_0, v_1, v_2)$ be a D-obtuse superbase, sorted such that $\rho_0 \le \rho_1 \le \rho_2$, and denote $e_i := v_i^\perp$ for all $0 \le i \le 2$. Since $0 < \rho_1 \le \rho_2$ by Lemma 6.3, the offsets $\{\pm e_1, \pm e_2\}$ are contained in the support of Selling's decomposition (40), and therefore $\max\{\|e_1\|_{D^{-1}}, \|e_2\|_{D^{-1}}\} \le C_0\lambda_{\min}(D)^{-\frac{1}{2}}$ by Theorem 4.3. Using the triangular inequality, we obtain $\|e\|_{D^{-1}} \le 2C_0\lambda_{\min}(D)^{-\frac{1}{2}}$ for all $e \in \mathrm{supp}(\tilde{\lambda}(D)) \subset \{\pm e_0, \pm e_1, \pm e_2, \pm(e_1 - e_2)\}$ since $e_0 = -e_1 - e_2$, as announced. We conclude noting that $|e| \le \|e\|_{D^{-1}}\lambda_{\max}(D)^{\frac{1}{2}}$. □

Proposition 6.13. *For any* $D \in \mathcal{S}_2^{++}$, $\mathrm{Span}_{\mathbb{Z}}\{e \in \mathcal{Z}_2 \mid \tilde{\lambda}^e(D) \geq \lambda_{\min}(D)/4\} = \mathbb{Z}^2$.

Proof. Let v be a D-obtuse superbase with Selling weights $\rho_0 \leq \rho_1 \leq \rho_2$, and denote $e_i := v_i^\perp$ for all $0 \leq i \leq 2$. Since v is a superbase, one has $|\det(e_1, e_2)| = |\det(v_1, v_2)| = 1$. Let $m := \mathrm{smed}(\rho_0, \rho_1, \rho_2)$ and $w := m\,\mathrm{sabs}(\rho_0/m) - \rho_0$. Then $w = mg(\rho_0/m) \leq m/2 \leq \rho_1/2$, using successively Lemmas 6.1 and 6.2. We conclude, using Lemma 6.3 for the last inequality

$$\tilde{\lambda}^{e_2}(D) := \rho_2 - w \geq \tilde{\lambda}^{e_1}(D) := \rho_1 - w \geq \rho_1/2 \geq \lambda_{\min}(D)/4. \qquad \square$$

References

[1] Eric Stephen Barnes. The complete enumeration of extreme senary forms. *Philosophical Transactions of the Royal Society of London. Series A, Mathematical and Physical Sciences*, 249(969):461–506, 1957.

[2] Joseph Frédéric Bonnans. *Convex and Stochastic Optimization.* Springer, 2019.

[3] Joseph Frédéric Bonnans, Guillaume Bonnet, and Jean-Marie Mirebeau. Monotone and second order consistent scheme for the two dimensional Pucci equation. In *Numerical mathematics and advanced applications ENUMATH 2019*, pages 733–742. Springer, 2021.

[4] Joseph Frédéric Bonnans, Guillaume Bonnet, and Jean-Marie Mirebeau. Second order monotone finite differences discretization of linear anisotropic differential operators. *Mathematics of computation*, 90(332):2671–2703, 2021.

[5] Joseph Frédéric Bonnans, Guillaume Bonnet, and Jean-Marie Mirebeau. A linear finite-difference scheme for approximating Randers distances on Cartesian grids. *ESAIM: Control, Optimisation and Calculus of Variations*, 28:45, 2022.

[6] Joseph Frédéric Bonnans, Elisabeth Ottenwaelter, and Hasnaa Zidani. A fast algorithm for the two dimensional HJB equation of stochastic control. *ESAIM: Mathematical Modelling and Numerical Analysis*, 38(4):723–735, 2004.

[7] Guillaume Bonnet and Jean-Marie Mirebeau. Monotone discretization of the Monge–Ampère equation of optimal transport. *ESAIM: Mathematical Modelling and Numerical Analysis*, 56(3):815–865, 2022.

[8] John Horton Conway and Neil James Alexander Sloane. Low-Dimensional Lattices. III. Perfect Forms. *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 418(1854):43–80, July 1988.

[9] John Horton Conway and Neil James Alexander Sloane. Low-Dimensional Lattices. VI. Voronoi Reduction of Three-Dimensional Lattices. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 436(1896):55–68, January 1992.

[10] Michael Grain Crandall, Hitoshi Ishii, and Pierre-Louis Lions. User's guide to viscosity solutions of second order partial differential equations. *Bulletin of the American Mathematical Society*, 27(1):1–68, January 1992.

[11] Keenan Crane, Clarisse Weischedel, and Max Wardetzky. Geodesics in heat: A new approach to computing distance based on heat flow. *ACM Transactions on Graphics (TOG)*, 32(5):152, 2013.

[12] Bernard Dacorogna. *Direct Methods in the Calculus of Variations*, volume 78 of *Applied Mathematical Sciences*. Springer, New York, 2008.

[13] Kristian Debrabant and Espen Robstad Jakobsen. Semi-Lagrangian schemes for linear and fully non-linear diffusion equations. *Mathematics of computation*, 82(283):1433–1462, 2013.

[14] François Desquilbet, Ludovic Métivier, and Jean-Marie Mirebeau. Single pass eikonal solver in tilted transversely isotropic media (preprint). 2022.

[15] Remco Duits, Stephan Meesters, Jean-Marie Mirebeau, and Jorg Portegies. Optimal paths for variants of the 2D and 3D Reeds-Shepp car with applications in image analysis. *Journal of Mathematical Imaging and Vision*, pages 1–33, 2018.

[16] Matthias Ehrgott. *Multicriteria Optimization*. Springer Berlin, Heidelberg, 2005.

[17] Jérôme Fehrenbach and Jean-Marie Mirebeau. Sparse non-negative stencils for anisotropic diffusion. *Journal of Mathematical Imaging and Vision*, 49(1):123–147, 2014.

[18] Carl Friedrich Gauß. Besprechung Des buchs von la seeber: Untersuchungen uber die eigenschaften der positiven ternaren quadratischen formen usw. *Gottingensche Gelehrte Anzeigen*, 2:188–196, 1876.

[19] Boško S Jovanović and Endre Süli. *Analysis of finite difference schemes: for linear partial differential equations with generalized solutions*, volume 46. Springer Science and Business Media, 2013.

[20] Aleksandr Korkine and Yegor Ivanovich Zolotareff. Sur les formes quadratiques positives. *Mathematische Annalen*, 11(2):242–292, 1877.

[21] Nikolai Vladimirovich Krylov. The rate of convergence of finite-difference approximations for Bellman equations with Lipschitz coefficients. *Applied Mathematics and Optimization*, 52(3):365–399, 2005.

[22] Jean-Marie Mirebeau. Fast-marching methods for curvature penalized shortest paths. *Journal of Mathematical Imaging and Vision*, pages 1–32, 2017.

[23] Jean-Marie Mirebeau. Riemannian Fast-Marching on Cartesian Grids, using Voronoi's First Reduction of Quadratic Forms. *SIAM Journal on Numerical Analysis*, 57(6):2608–2655, 2019.

[24] Theodore Samuel Motzkin and Wolfgang Wasow. On the approximation of linear elliptic differential equations by difference equations with positive coefficients. *Journal of Mathematics and Physics*, 31(1-4):253–259, 1952.

[25] Arkadi Nemirovski and Michael Jeremy Todd. Interior-point methods for optimization. *Acta Numerica*, 17, 04 2008.

[26] Phong Quang Nguyen and Damien Stehlé. Low-dimensional lattice basis reduction revisited. In Ducan Buell, editor, *ANTS*, pages 338–357. Springer, 2004.

[27] Ricardo Horacio Nochetto, Dimitrios Ntogkas, and Wujun Zhang. Two-scale method for the Monge–Ampère equation: pointwise error estimates. *IMA Journal of Numerical Analysis*, 2017.

[28] Adam Oberman. Convergent Difference Schemes for Degenerate Elliptic and Parabolic Equations: Hamilton-Jacobi Equations and Free Boundary Problems. *SIAM Journal on Numerical Analysis*, 44(2):879–895, January 2006.

[29] Elisabeth Rouy and Agnès Tourin. A Viscosity Solutions Approach to Shape-From-Shading. *SIAM Journal on Numerical Analysis*, 29(3):867–884, July 1992.

[30] Achill Schürmann. Computational geometry of positive definite quadratic forms. *University Lecture Series*, 49, 2009.

[31] Achill Schürmann. Enumerating perfect forms. *Contemporary Mathematics*, 493:359, 2009.

[32] Eduard Selling. Ueber die binären und ternären quadratischen Formen. *Journal für die Reine und Angewandte Mathematik*, 77:143–229, 1874.

[33] Mathieu Sikirić, Achill Schürmann, and Frank Vallentin. Classification of eight-dimensional perfect forms. *Electronic Research Announcements of the American Mathematical Society*, 13(3):21–32, 2007.

[34] Vidar Thomée. Discrete interior Schauder estimates for elliptic difference operators. *SIAM Journal on Numerical Analysis*, 5(3):626–645, 1968.

[35] Georgy Voronoï. Nouvelles applications des paramètres continus à la théorie des formes quadratiques. I: Sur quelques propriétés des formes quadratiques positives parfaites. *Journal für die Reine und Angewandte Mathematik*, 133:97–178, 1908.

[36] Joachim Weickert. *Anisotropic diffusion in image processing*, volume 1. Teubner Stuttgart, 1998.

A Discretization of degenerate elliptic PDEs

Degenerate ellipticity is a property of differential operators which is at the foundation of the theory of viscosity solutions of PDEs [10] and of the related comparison principles. Degenerate elliptic (DE) operators arise in a variety of contexts, such as deterministic or stochastic optimal control problems, optimal transport problems, and more generally Hamilton-Jacobi-Bellman PDEs, see the discussion below. Decompositions of symmetric positive definite matrices can be used to discretize these DE operators in a way that preserves their structure, and leads to the discrete degenerate ellipticity (DDE) property, which is a key tool in the subsequent convergence analysis [28]. In this appendix, we illustrate the relevance of Definition 1.2 and Theorems 1.6 and 1.8 for the discretization of several PDEs.

For completeness, let us recall the formal definition of the DE and DDE properties [10, 28], omitting for simplicity the discussion of boundary conditions.

Definition A.1 (Degenerate ellipticity [10])**.** Let $\mathfrak{F}$ be a differential operator on an open domain $\Omega \subset \mathbb{R}^d$, of the form

$$\mathfrak{F}u(x) := \tilde{\mathfrak{F}}(x, u(x), \nabla u(x), \nabla^2 u(x)), \tag{43}$$

for all $x \in \Omega$ and all $u \in C^2(\Omega)$. We say that $\mathfrak{F}$ is degenerate elliptic[5] if $\tilde{\mathfrak{F}} = \tilde{\mathfrak{F}}(x, v, p, M)$ is non-decreasing w.r.t. the second variable $v \in \mathbb{R}$, and non-increasing w.r.t. the last variable $M \in \mathcal{S}_d$ with respect to the Loewner order.

Definition A.2 (Discrete degenerate ellipticity [28])**.** Let X be a finite set, and let $F : \mathbb{R}^X \to \mathbb{R}^X$ be a finite differences scheme, of the form

$$Fu(x) := \tilde{F}\big(x, u(x), [u(y) - u(x)]_{y \in X \setminus \{x\}}\big). \tag{44}$$

We say that F is degenerate elliptic if $\tilde{F} = \tilde{F}(x, v, [\delta_y]_{y \in X \setminus \{x\}})$ is non-decreasing w.r.t. the second variable v, and non-increasing w.r.t. the last variable $[\delta_y]_{y \in X \setminus \{x\}}$ componentwise.

In each of the following subsections, we consider a DE operator defined via a field of symmetric positive definite matrices $\mathcal{D} : \Omega \to \mathcal{S}_d^{++}$ over a domain $\Omega \subset \mathbb{R}^d$, and a DDE scheme involving coefficients $\lambda : \Omega \to \Lambda_d$ (or a family of such fields $\mathcal{D}_\alpha$ and coefficients λ_α, indexed by some parameter $\alpha \in \mathcal{A}$). We of course recommend setting $\lambda(x) := \lambda(\mathcal{D}(x))$, following Corollary 1.7, which is a practical choice that ensures the properties ($\mathcal{D}$-consistency, R-support, K-Lipschitz, ε-spanning) of Definition 1.2 whose relevance is discussed.

A.1 The Riemannian eikonal equation.

The eikonal equation is a first order Hamilton-Jacobi PDE, non-linear and static, which characterizes geodesic distance maps. Consider a domain $\Omega \subset \mathbb{R}^d$, open and bounded for

[5] Some works separate these two monotonicity conditions, which are then referred to as properness and degenerate ellipticity [10].

simplicity, and equipped with a Riemannian metric $\mathcal{M} \in \mathrm{Lip}(\overline{\Omega}, \mathcal{S}_d^{++})$. Then the geodesic distance from the boundary $\partial\Omega$, is the unique viscosity solution [10] to the Riemannian eikonal PDE

$$\|\nabla u(x)\|_{\mathcal{D}(x)} = 1, \forall x \in \Omega, \qquad\qquad u(x) = 0, \forall x \in \partial\Omega, \tag{45}$$

where $\mathcal{D}(x) := \mathcal{M}(x)^{-1}$. The differential operator $\tilde{\mathfrak{F}}(x, v, p, M) := \|p\|_{\mathcal{D}(x)}$ is DE. Indeed it complies with the monotonicity conditions of Definition A.1 since it is independent of both v and M. The eikonal equation operator $\|\nabla u(x)\|^2_{\mathcal{D}(x)}$ may be discretized as follows [23]:

$$F_h u(x) := \sum_{e \in \mathcal{Z}_d} \lambda^e(x) \max\left\{0, -\delta_h^e u(x), -\delta_h^{-e} u(x)\right\}^2, \tag{46}$$

where $x \in \Omega_h := \Omega \cap h\mathbb{Z}^d$, and $\lambda^e(x) \geq 0$ for all $e \in \mathcal{Z}_d$. The discrete counterpart of (45) is the system of equations $F_h u = 1$ on Ω_h, where the unknown $u : \Omega_h \to \mathbb{R}$ is extended by 0 outside Ω_h. The scheme F_h is degenerate elliptic, since it is a non-increasing function of the finite differences $\delta_h^e u(x) := (u(x + he) - u(x))/h$.

The $\mathcal{D}$-*consistency* property, of the scheme coefficients λ, implies the first order consistency of the scheme: $F_h u(x) = \|\nabla u(x)\|^2_{\mathcal{D}(x)} + \mathcal{O}(h)$, for smooth u, by a Taylor expansion. The scheme (46) introduced in [23] is in fact an anisotropic generalization of the classical scheme [29], made possible by this consistency property. The stable and consistent decomposition obtained in Theorem 1.6 allows its extension to general domains of dimension $d = 4$. In contrast, previous implementations relied on Selling's decomposition and were thus limited to domains of dimension $d \leq 3$. A five dimensional Reeds-Shepp model posed on $\mathbb{R}^3 \times \mathbb{S}^2$ could nevertheless be addressed in [23], by observing that the matrices of the metric have a block diagonal structure with blocks of shape 3×3 and 2×2.

The R-*support* property leads to the error estimate $\|u_h^* - u^*\|_{L^\infty(\Omega_h)} = \mathcal{O}(\sqrt{Rh})$, between the continuous solution u^* of (45) and the discrete solution u_h^*, see [23, Theorem 1.3]. Using the decomposition of Theorem 1.6 we obtain $R \leq C\mu_{\max}$, where $\mu_{\max} := \max\{\mu(\mathcal{D}(x)) \mid x \in \Omega\}$ is an upper bound on the anisotropy ratio of the metric. In dimension $d \geq 4$, this is an improvement over the estimate $R \leq C\mu_{\max}^{d-1}$ obtained in [23, Proposition 1.1], which in addition yields improved convergence rates when one considers a relaxed sub-Riemannian model. More precisely, assume that $\mathcal{D}_\varepsilon = \mathcal{D}_0 + \varepsilon^2 \mathrm{Id}_d$ for some relaxation parameter $\varepsilon > 0$, where $\mathcal{D}_0 \in \mathrm{Lip}(\overline{\Omega}, \mathcal{S}_d^+)$ is only positive *semi*-definite pointwise. Then one has $\mu_{\max}^\varepsilon = \mathcal{O}(\varepsilon^{-1})$, thus $R_\varepsilon = \mathcal{O}(\varepsilon^{-1})$, with obvious notations. Therefore the error estimate $\|u^* - u_{h,\varepsilon}^*\|_{L^\infty(\Omega_h)} = \mathcal{O}(\varepsilon + \sqrt{R_\varepsilon h})$, between the sub-Riemannian distance u^* to the boundary and the approximation $u_{h,\varepsilon}^*$ obtained by relaxation and discretization [23, Theorem 1.8] (under suitable assumptions), boils down to $\mathcal{O}(h^{\frac{1}{3}})$ with the optimal parameter choice $\varepsilon = h^{\frac{1}{3}}$.

The K-*Lipschitz* and ε-*spanning* properties lead to a Lipschitz estimate of the discrete solution u_h^*, namely $|u_h^*(x) - u_h^*(y)| \leq C|x - y|$, for any $x, y \in \Omega_h$ and any sufficiently

small scale $h > 0$, see [14, Proposition 4.4]. This estimate rules out numerical instabilities such as checkerboards artifacts, and is also a necessary property if one wants to consider point source boundary conditions, so as to compute the geodesic distance from a single seed rather than from the domain's boundary. The proof relies on a strategy similar to the coercivity estimate obtained for elliptic PDEs in Theorem 1.3. Note that this Lipschitz estimate is established in dimension $d \in \{2, 3\}$ in [14] using Selling's decomposition for the inverse metric tensors, yet it only uses the properties of Definition 1.2, and therefore it extends in a straightforward manner to dimension $d = 4$ using Theorem 1.6.

A.2 Linear non-divergence form diffusion.

Linear non-divergence form operators arise in the study of stochastic processes, through the Feynman-Kac formula, and they are also the building blocks of the non-linear Hamilton-Jacobi-Bellman operators discussed in Appendix A.3 below. A non-divergence form diffusion operator, and its discretization, take the form

$$\mathfrak{F}u(x) = -\operatorname{Tr}(\mathcal{D}(x)\nabla^2 u(x)), \qquad F_h u(x) = -\sum_{e \in \mathcal{Z}_d} \lambda^e(x) \Delta_h^e u(x), \tag{47}$$

where $\mathcal{D} \in \operatorname{Lip}(\overline{\Omega}, \mathcal{S}_d^{++})$ is a field of diffusion tensors, and where $\lambda^e(x) \geq 0$ is a non-negative coefficient. We denoted by $\Delta_h^e u(x) := (u(x - he) - 2u(x) + h(x + he))/h^2$ the second order centered finite differences operator. The operator $\tilde{\mathfrak{F}}(x, v, p, M) := -\operatorname{Tr}(\mathcal{D}(x)M)$ is DE since the trace of a product of non-negative symmetric matrices is non-negative, and likewise F_h is DDE by observing that it is a negatively weighted linear combination of the finite differences $u(x + he) - u(x)$, $e \in \mathcal{Z}_d$.

The $\mathcal{D}$-*consistency* property, of the scheme coefficients λ, yields the second order consistency of the scheme: $F_h u(x) = \mathfrak{F}u(x) + \mathcal{O}(h^2)$ for smooth u using a straightforward Taylor expansion. In the case where $\mathcal{D}$ is only positive *semi*-definite pointwise, on may consider a relaxation procedure as discussed in Appendix A.1.

The *R-support property* keeps the size of the discretization stencil under control. This is necessary to establish convergence, and is also needed to ensure the DDE property for some modified schemes. Indeed, consider an inhomogeneous operator featuring an additional linear first order term: $\mathfrak{F}u(x) = -\operatorname{Tr}(\mathcal{D}(x)\nabla^2 u(x)) + \langle \omega(x), \nabla u(x) \rangle$. Two approaches can be envisioned to discretize the first order term, the first one being upwind finite differences, which are first order accurate and DDE. Alternatively, a second order consistent discretization using centered finite differences is proposed in [4, Definition 1.5], which is also DDE provided $|\langle \omega(x), \mathcal{D}(x)^{-1} e \rangle| \leq h^{-1}$ for any $e \in \operatorname{supp}(\lambda)$. By Theorem 4.3, this condition is met when $\|\omega(x)\|_{\mathcal{D}(x)^{-1}} \leq c\sqrt{\lambda_{\min}(\mathcal{D}(x))}/h$ for all $x \in \Omega$, where $c = c(d) > 0$.

The *K-Lipschitz* and *ε-spanning* properties are used in the analysis of a numerical scheme for the computation of geodesic distances based on the solution of linear non-divergence form diffusion equations [5] (this approach differs from Appendix A.1 where such distances are computed by solving a non-linear eikonal PDE). The method applies to

Randers metrics, a generalization of Riemannian metrics, and provides both an anisotropic generalization and a first convergence analysis for a popular minimal path computation technique in geometry processing [11]. The K-Lipschitz and ε-spanning properties are used in the convergence analysis in the case of point sources [5, Theorem 4.1], i.e. computing geodesic distances from a single seed. The proof is stated in dimension $d \leq 3$ for a scheme based on Selling's decomposition, but it extends in a straightforward manner to dimension $d = 4$ since it only relies on the properties of Definition 1.2.

The *ε-spanning* property also allows to establish discrete interior Schauder estimates for the solutions of elliptic difference operators [34], provided the scheme coefficients are *smooth*. Both properties are ensured if they are obtained from the matrix decomposition of Theorem 1.8. More precisely, [34] assumes that the principal Fourier symbol of the scheme is invertible, but as shown in the next proposition this is equivalent to the spanning property in the case of second order operators.

Proposition A.3. *Let $\lambda^e \geq 0$ for all $e \in \mathcal{Z}_d$, with only finitely many positive coefficients. Then the following properties are equivalent:*

- *(Spanning)* $\mathrm{Span}_{\mathbb{Z}}\{e \in \mathcal{Z}_d \mid \lambda^e > 0\} = \mathbb{Z}^d$.
- *(Ellipticity, in the sense of [34]) For all $\theta \in [-1/2, 1/2]^d \setminus \{0\}$, one has $p(\theta) \neq 0$, where*

$$p(\theta) := \sum_{e \in \mathcal{Z}_d} \lambda^e [\cos(2\pi \langle \theta, e \rangle) - 1]. \tag{48}$$

Proof. Assume first that the spanning property holds. Observe that $\cos(2\pi t) - 1 \leq 0$ for all $t \in \mathbb{R}$, with equality iff $t \in \mathbb{Z}$. Therefore $p(\theta) \leq 0$ for all $\theta \in \mathbb{R}^d$, with equality iff

$$\forall e \in \mathcal{Z}_d,\ \lambda^e > 0 \Rightarrow \langle \theta, e \rangle \in \mathbb{Z}. \tag{49}$$

Using the spanning property, and the linearity of the scalar product, we obtain that $\langle \theta, e \rangle \in \mathbb{Z}$ for all $e \in \mathbb{Z}^d$. This implies $\theta \in \mathbb{Z}^d$, hence $\theta \notin [-1/2, 1/2]^d \setminus \{0\}$, as announced.

Assume now that the spanning property does not hold. Denote $L := \mathrm{Span}_{\mathbb{Z}}\{e \in \mathcal{Z}_d \mid \lambda^e > 0\}$, and introduce the dual lattice $L^* := \{\theta \in \mathbb{R}^d \mid \langle \theta, e \rangle \in \mathbb{Z}, \forall e \in L\}$. Then $L \subsetneq \mathbb{Z}^d$, and therefore $L^* \supsetneq \mathbb{Z}^d$. Consider $\theta_0 \in L^* \setminus \mathbb{Z}^d$, and define $\theta := \theta_0 - \mathrm{round}(\theta_0)$, where the rounding operator returns a closest integer componentwise. Then $\theta \in L^*$, thus $p(\theta) = 0$, and by construction $\theta \in [-1/2, 1/2]^d \setminus \{0\}$, which concludes the proof. □

A.3 Non-linear second order degenerate elliptic PDEs.

A natural avenue to define and study fully non-linear differential operators, is to introduce them in the form of a maximum, or sometimes a minimum, of a family of linear operators.

Consider an arbitrary set $\mathcal{A}$ of parameters (in practice, $\mathcal{A}$ is usually a domain of $\mathbb{R}^n$), an open domain $\Omega \subset \mathbb{R}^d$, and define

$$\mathfrak{F}u(x) := \sup_{\alpha \in \mathcal{A}} \mathfrak{F}^\alpha u(x), \quad \mathfrak{F}^\alpha u(x) := a_\alpha(x) + b_\alpha(x)u(x) + \langle c_\alpha(x), \nabla u(x) \rangle - \mathrm{Tr}(\mathcal{D}_\alpha(x)\nabla^2 u(x)).$$

We assume that $b_\alpha(x) \geq 0$, $c_\alpha(x) \in \mathbb{R}^d$, and $\mathcal{D}_\alpha(x) \in \mathcal{S}_d^{++}$, for each point $x \in \Omega$ and parameter $\alpha \in \mathcal{A}$. In this way, the linear operator $\mathfrak{F}^\alpha$ is DDE, and likewise $\mathfrak{F}$ is DDE, provided the extremum over $\alpha \in \mathcal{A}$ is well defined. Extremal operators such as $\mathfrak{F}$ naturally arise in Hamilton-Jacobi-Bellman (HJB) PDEs related with stochastic control problems [21], but they are also encountered in relation with optics and optimal transport since the Monge-Ampère operator $\det(\nabla^2 u)$ can be written in this form [7].

In order to define a numerical scheme for the non-linear operator $\mathfrak{F}$, a natural first step is to discretize each linear operator $\mathfrak{F}^\alpha$. For that purpose, on a Cartesian grid $\Omega_h := \Omega \cap h\mathbb{Z}^d$ of scale $h > 0$, we proceed as in Appendix A.2 and in particular we introduce the coefficients $\lambda_\alpha^e(x)$, $e \in \mathcal{Z}_d$, of some decomposition of $\mathcal{D}_\alpha(x) \in \mathcal{S}_d^{++}$, for any $x \in \Omega$, $\alpha \in \mathcal{A}$.

The $\mathcal{D}_\alpha$-*consistency* property means that the discretization F_h^α of the linear operator $\mathfrak{F}^\alpha$ is either first or second order accurate, depending on the treatment of the first order term. From this point, one may consider a finite subset $\mathcal{A}_h \subset \mathcal{A}$, and introduce the discretization $F_h := \max_{\alpha \in \mathcal{A}_h} F_h^\alpha$ of the non-linear operator $\mathfrak{F}$. For consistency, the cardinality of $\mathcal{A}_h$ needs to grow to infinity as $h \to 0$, but for numerical efficiency, this cardinality should not be excessively large either, which leads to compromises. Alternatively the scheme $\tilde{F}_h := \max_{\alpha \in \mathcal{A}} F_h^\alpha$, where the optimization is over the full set of parameters $\mathcal{A}$, may often be computed in closed form. This leads to second order accurate DDE discretizations of the Monge-Ampère equation [7, Remark 3.4] (in the most favorable case) and of the Pucci equation [3] in dimension $d = 2$, and a similar approach is used in [14] for a first order PDE with a complex anisotropy in dimension $d = 3$. A key ingredient to computing the accurate discretization $\tilde{F}_h$ in closed form is, in each of these cases, the piecewise linear structure of the coefficients of Selling's decomposition $D \in \mathcal{S}_d^{++} \mapsto \lambda^e(D)$, where $e \in \mathcal{Z}_d$. Since the matrix decomposition introduced in Proposition 1.5 is similarly piecewise linear, these techniques may in principle be extended in dimension $d = 4$.

The *R-support* property controls the effective discretization scale of the numerical scheme. In many cases of interest, including the Monge-Ampère operator, the condition number of the diffusion tensors is unbounded over the parameter set: $\sup\{\mu(D_\alpha(x)) \mid \alpha \in \mathcal{A}\} = +\infty$. This leads to a compromise when choosing the discrete parameter set $\mathcal{A}_h \subset \mathcal{A}$, since including strongly anisotropic tensors improves the consistency with the differential operator, but also leads to large stencils and thus degraded finite difference truncation errors, see [7, Remark 3.4]. The improved support radius estimate obtained in this paper allows to conduct similar analyses in dimension $d = 4$, and is also relevant when $\mathcal{D}_\alpha$ is only positive semi-definite and a relaxation procedure is used as discussed in Appendix A.1.

The ε-*spanning* property so far has not been used in the context of extremal operators such as $\mathfrak{F}$, to our knowledge. A significant obstruction, at least to naive approaches, is that

the active parameter $\alpha(u,x) \in \mathcal{A}_h$, such that $F_h u(x) = F^{\alpha(u,x)} u(x)$ where $u : \Omega_h \to \mathbb{R}$ and $x \in \Omega_h$, varies possibly discontinuously from point to point. This lack of local consistency between the active stencils prevents simple arguments based on the concatenation of their offsets as in the proof of the coercivity property Theorem 1.3.

The *K-Lipschitz* property of the numerical scheme coefficients, or higher smoothness properties, are often ingredients of the analysis of the convergence rate of the numerical solution. For concreteness, consider the evolution PDE $\partial_t u = \mathfrak{F} u$ over the domain $[0,\infty[\times\mathbb{R}^d$. The convergence rate $\mathcal{O}(\tau^{1/4} + h^{1/2})$ is established in [21], where τ denotes the time step and h the grid scale, for finite differences discretizations similar to the one described in this section, under suitable assumptions (and with possibly time dependent coefficients). One key assumption of [21] is that the square root of the scheme coefficients $x \mapsto \sqrt{\lambda^e(x)}$ be Lipschitz, for any $e \in \mathcal{Z}_d$. This can be ensured by choosing a K-grad Lipschitz decomposition, such as the one described in Theorem 1.8 in dimension $d = 2$, as shown below.

Lemma A.4. *Let $\alpha : \mathbb{R}^d \to [0,\infty[$ be such that $\nabla\alpha$ is K-Lipschitz. Then $\sqrt{\alpha}$ is $\sqrt{K/2}$-Lipschitz.*

Proof. Assume that $d = 1$, up to restricting to a line, and that $K = 1$, up to considering α/K. Then $0 \le \alpha(x+h) \le \alpha(x) + h\alpha'(x) + h^2/2$ for any $x, h \in \mathbb{R}$. The discriminant of the r.h.s., seen as a quadratic function of h, thus obeys $\alpha'(x)^2 - 2\alpha(x) \le 0$. Therefore $|\frac{d}{dx}\sqrt{\alpha}| = |\alpha'(x)|/(2\sqrt{\alpha(x)}) \le 1/\sqrt{2}$, whenever $\alpha(x) > 0$.

Let $x_0 < x_1$, let us assume w.l.o.g. that $\alpha(x_0) < \alpha(x_1)$, and let $x_* := \max\{x \in [x_0, x_1] \mid \alpha(x) = \alpha(x_0)\}$. Then α is positive on $]x_*, x_1]$, by the intermediate value theorem. Thus $|\sqrt{\alpha(x_1)} - \sqrt{\alpha(x_0)}| = |\sqrt{\alpha(x_1)} - \sqrt{\alpha(x_*)}| \le |x_1 - x_*|/\sqrt{2} \le |x_1 - x_0|/\sqrt{2}$ by the above, which concludes the proof. □

B Alternative smooth and spanning decompositions

One of the main objectives of this paper, achieved in Theorems 1.6 and 1.8, is the design of computable matrix decompositions obeying the *spanning* property, and suitable *smoothness* properties. We anticipate in this appendix a possible objection, which is that an arbitrary given decomposition may be modified so as to obey these properties. For that purpose, we consider two such possible modifications, and show that they have undesirable side effects.

For simplicity, we assume in this section that the eigenvalues of the decomposed symmetric matrices are positively bounded below, and for that purpose we denote $\mathcal{S}_d^\varepsilon := \{D \in \mathcal{S}_d^{++} \mid D \succeq \varepsilon \,\mathrm{Id}_d\}$, for any $\varepsilon > 0$. We assume given some measurable coefficients $\lambda : \mathcal{S}_d^\varepsilon \to \Lambda_d$ (not necessarily those of Proposition 1.5), which are *consistent* in the sense that $\sum_{e\in\mathcal{Z}_d} \lambda^e(D) e e^\top = D$ for all $D \in \mathcal{S}_d^\varepsilon$, and are $R(\mu)$-*supported* in the sense that $|e| \le R(\mu(D))$ for all $e \in \mathrm{supp}(\lambda(D))$, where R is some given function.

The modified coefficients $\tilde\lambda, \hat\lambda : \mathcal{S}_d^{2\varepsilon} \to \Lambda_d$, constructed in (50) and (51) below, respectively enjoy the spanning and smoothness properties. The modifications are simple if not

trivial from the theoretical standpoint, and yet we do not recommend them in practice for the three following reasons. (1. Quantitative argument) The matrix decompositions $\tilde{\lambda}$ and $\hat{\lambda}$ are respectively $R(\mu\sqrt{2})$ and $R(\mu\sqrt{3})$ supported, hence lead to schemes with wider stencils and thus a larger truncation error than the original λ. (2. Qualitative argument) $\tilde{\lambda}$ and $\hat{\lambda}$ fail the unimodular invariance property, established in Proposition 2.5 for the constructions proposed in this paper. (3. Implementation argument) The definition of $\hat{\lambda}$ involves a $d(d+1)/2$-dimensional convolution (51) whose numerical computation is likely impractical.

Obtaining the spanning property. Define, for any $D \in \mathcal{S}_d^{2\varepsilon}$, the modified coefficients

$$\tilde{\lambda}^e(D) := \varepsilon \mathbb{1}_{|e|=1} + \lambda^e(D - \varepsilon \operatorname{Id}_d), \tag{50}$$

for all $e \in \mathcal{Z}_d$. Note that $|e| = 1$ iff $e = \pm b_i$ for some $1 \le i \le d$, where $(b_i)_{i=1}^d$ denotes the canonical basis of $\mathbb{R}^d$. The consistency of the modified coefficients follows:

$$\sum_{e\in\mathcal{Z}_d} \tilde{\lambda}^e(D) e e^\top = \varepsilon \sum_{1\le i\le d} b_i b_i^\top + \sum_{e\in\mathcal{Z}_d} \lambda^e(D - \varepsilon \operatorname{Id}_d) e e^\top = \varepsilon \operatorname{Id}_d + (D - \varepsilon \operatorname{Id}_d) = D.$$

The modified coefficients $\tilde{\lambda}$ obey the ε-spanning property (7), since $\det(b_1, \cdots, b_d) = 1$ and $\lambda^{b_i}(D) \ge \varepsilon$ for all $1 \le i \le d$. Since the defining expression (7) of $\tilde{\lambda}$ involves $\lambda(D - \varepsilon \operatorname{Id}_d)$, and since $\mu(D) \le \mu(D - \varepsilon \operatorname{Id}_d) \le \mu(D)\sqrt{2}$, we find as announced that it is $R(\mu\sqrt{2})$-supported.

Obtaining smooth coefficients. We proceed by mollification in the space of symmetric matrices. For that purpose we introduce a function $\rho \in C^\infty(\mathbb{R})$, which is even, non-negative, supported on $[-1, 1]$, and not identically zero. We also denote by $\|A\|_F := \sqrt{\operatorname{Tr}(A^\top A)} = \sqrt{\sum_{i,j=1}^n A_{ij}^2}$ the Frobenius norm of a matrix, which is related to the spectral norm by $\|A\| \le \|A\|_F \le \|A\|\sqrt{d}$. The modified coefficients are defined for each $D \in \mathcal{S}_d^{2\varepsilon}$ as

$$\hat{\lambda}^e(D) := \int_{\mathcal{S}_d} \lambda^e(D - S) \rho_\varepsilon(S) \, \mathrm{d}S, \qquad \text{with } \rho_\varepsilon(S) := \frac{1}{c(d,\varepsilon)} \rho\Big(\frac{\|S\|_F^2}{\varepsilon^2}\Big), \tag{51}$$

where $c(d, \varepsilon)$ is a normalization constant such that $\int_{\mathcal{S}_d} \rho_\varepsilon(S) \, \mathrm{d}S = 1$. The mapping $D \in \mathcal{S}_d^{2\varepsilon} \mapsto \hat{\lambda}^e(D)$ is non-negative and smooth for any $e \in \mathcal{Z}_d$, by an immediate mollification argument and in view of the local bound $0 \le \lambda^e(D) \le \|D\|/\|e\|^2 \le \|D\|$. Note that for any $S \in \mathcal{S}_d$ such that $\rho_\varepsilon(S) > 0$, one has $\|S\| \le \|S\|_F \le \varepsilon$. It follows that $D - S \in \mathcal{S}_d^\varepsilon$ for any $D \in \mathcal{S}_d^{2\varepsilon}$, and that $\mu(D - S) \le \mu(D)\sqrt{3}$, hence $\hat{\lambda}$ is $R(\mu\sqrt{3})$-supported as announced. Finally, noting that $\int_{\mathcal{S}_d} S \rho_\varepsilon(S) \, \mathrm{d}S = 0$ by anti-symmetry, we establish consistency

$$\sum_{e\in\mathcal{Z}_d} \hat{\lambda}^e(D) e e^\top = \int_{\mathcal{S}_d} \sum_{e\in\mathcal{Z}_d} \lambda^e(D - S) e e^\top \rho_\varepsilon(S) \, \mathrm{d}S = \int_{\mathcal{S}_d} (D - S) \, \rho_\varepsilon(S) \, \mathrm{d}S = D.$$

C Elliptic equations: convergence rates

In this appendix, we establish convergence rates for the monotone discretization (5) of the anisotropic elliptic PDE (4), with periodic boundary conditions, as announced in Theorem 1.4. The proof is an adaptation of [19, §2.6.1], originally addressing a classical two-dimensional scheme featuring a 7 point stencil, and lacking the non-negativity property which motivates our wide-stencil design, see Remark 1.1. The monotonicity property is not used in the proof; in fact, monotonicity offers an independent avenue for proving rates of convergence, which is briefly investigated in Appendix C.1. Let us mention that [19] presents a much wider catalog of estimates, in the L^2_h, W^1_h, and W^2_h discrete Sobolev norms, with minimal regularity assumptions, etc. We see no obstruction in principle to their adaptation to the proposed scheme (5), yet this remains outside of the scope of this paper.

Throughout this section, we assume that the scheme coefficients $\lambda : \mathbb{T}^d \to \Lambda_d$ obey the $\mathcal{D}$-consistency, R-support, K-Lipschitz, and ε-spanning properties, following the assumptions of Theorem 1.4. For convenience, we define the shorthands

$$M_{\mathcal{L}} := \max\{d, \|\mathcal{D}\|_\infty, R, \|\nabla\lambda\|_\infty, \varepsilon\}, \qquad \mathcal{Z}_d^R := \{e \in \mathcal{Z}_d \mid |e| \le R\}.$$

For any $v \in L^2(\mathbb{T}^d)$, we (abusively) consider the quantities $\|\nabla v\|_{L^2}$, $\|\nabla^2 v\|_{L^2}$, and $\|\nabla^3 v\|_{L^2}$. If v does not belong to the appropriate Sobolev space, then this quantity is defined as $+\infty$, and any estimate involving it is simply vacuous. Likewise, estimates involving the quantity $\|\nabla^2\lambda\|_\infty$ are vacuous unless we assume K-grad-Lipschitz coefficients.

Lemma C.1. *One has for $u \in L^2(\mathbb{T}^d)$ and $\lambda \in L^\infty(\mathbb{T}^d)$, with the above convention*

$$\|\nabla u\|_{L^2} \le \|\nabla^2 u\|_{L^2} \le \|\nabla^3 u\|_{L^2}, \qquad \|\nabla\lambda\|_\infty \le \|\nabla^2\lambda\|_\infty. \tag{52}$$

Proof. These estimates are easily deduced from the one-dimensional case $d = 1$, i.e. from the fact that $\|f\|_{L^p} \le \|f'\|_{L^p}$ for any $f \in W^{1,p}(\mathbb{T})$ such that $\int_{\mathbb{T}} f = 0$, where $p \in \{2, \infty\}$. When $p = 2$ this is an instance of the classical Poincaré-Wirtinger inequality, proved by e.g. considering the Fourier expansion of f, and when $p = \infty$ one can note that f has a zero and admits the Lipschitz constant $\|f'\|_\infty$. □

For any $u : \mathbb{T}^d_h \to \mathbb{R}$ one has the (semi-)coercivity estimate

$$-\langle \mathcal{L}_h u, u\rangle_{L^2_h} = \mathcal{Q}_h(u) \ge c_Q \|\nabla_h u\|^2_{L^2_h}, \tag{53}$$

where the equality holds by construction of $\mathcal{L}_h$ and $\mathcal{Q}_h$, and the inequality by Theorem 1.3 for all sufficiently small grid scales $0 < h \le h_0$ (we assume that this condition is satisfied in the following), and where $h_0 > 0$ and $c_{\mathcal{Q}} > 0$ only depend on $M_{\mathcal{L}}$. As a first step, we prove that the considered PDE and its discretization are well posed.

Lemma C.2. *For any $\mathrm{f} \in L^2(\mathbb{T}^d)$ with $E[\mathrm{f}] = 0$, there exists $u \in H^1(\mathbb{T}^d)$ such that $\mathcal{L}u = \mathrm{f}$ and $E[u] = 0$. For any $\mathrm{f}_h : \mathbb{T}^d_h \to \mathbb{R}$ with $E_h[\mathrm{f}_h] = 0$, there exists $u_h : \mathbb{T}^d_h \to \mathbb{R}$ such that $\mathcal{L}_h u_h = \mathrm{f}_h$ and $E_h[u_h] = 0$. We denoted $E[\mathrm{f}] := \int_{\mathbb{T}^d} \mathrm{f}(x)\mathrm{d}x$ and $E_h[\mathrm{f}_h] := h^d \sum_{x\in\mathbb{T}^d_h} \mathrm{f}_h(x)$.*

Proof. By construction, the operator $\mathcal{L}_h$ is self-adjoint and vanishes on constant functions, hence it leaves invariant the subspace $V_h := \{u_h : \mathbb{T}_h^d \to \mathbb{R} \mid E_h[u_h] = 0\}$. If $\mathcal{L}_h u_h = 0$ then u_h is constant by (53), and if $u_h \in V_h$ this implies $u_h = 0$. Therefore the restriction $\mathcal{L}_{h|V_h}$ is invertible, hence the existence of u_h. We only sketch the argument in the continuous case, which is analogous and extremely classical [19]. Again $\mathcal{L}$ is self-adjoint and vanishes on constant functions, hence maps $V := \{u \in H^1(\mathbb{T}^d) \mid E[u] = 0\}$ to $V^* := \{f \in H^{-1}(\mathbb{T}^d) \mid E[f] = 0\}$. Noting that $\|u\|_{L^2} \leq \|\nabla u\|_{L^2}$ for all $u \in V$ by the Poincaré-Wirtinger inequality, we obtain that V is a Hilbert space when equipped with the norm $N(u) := \|\nabla u\|_{L^2}$. By the coercivity estimate $-\langle \mathcal{L}u, u\rangle = \int_{\mathbb{T}^d} \langle \nabla u(x), \mathcal{D}(x)\nabla u(x)\rangle \mathrm{d}x \geq \lambda_* \|\nabla u\|_{L^2}^2$, and the Lax-Milgram theorem, the restriction $\mathcal{L}_{|V}$ is boundedly invertible, hence the existence of u. □

We introduce a centered finite difference operator ∂_h^e, and locally averaged coefficients λ_h^e, defined as

$$\partial_h^e u(x) := \frac{u(x + he/2) - u(x - he/2)}{h}, \quad \lambda_h^e(x) := \frac{\lambda^e(x + he/2) + \lambda^e(x - he/2)}{2}, \tag{54}$$

for any $x \in \mathbb{T}$ and $e \in \mathbb{Z}^d \setminus \{0\}$. Denoting by $\partial^e u(x) := \langle \nabla u(x), e\rangle$ the directional differentiation operator, we obtain the strikingly similar expressions

$$\mathcal{L}_h u = \sum_{e \in \mathcal{Z}_d} \partial_h^e(\lambda_h^e \partial_h^e u), \qquad \mathcal{L}u := \sum_{e \in \mathcal{Z}_d} \partial^e(\lambda^e \partial^e u), \tag{55}$$

which follow respectively from (5) and from the $\mathcal{D}$-consistency property. Note u and λ are in (55, left) only evaluated at grid points $x \in \mathbb{T}_h^d$, contrary to what (54) may suggest, because of operator composition. Define the additional operators

$$\delta_h^e u(x) := \frac{u(x + he) - u(x)}{h}, \qquad \tau_h^e u(x) := u(x + he/2).$$

Lemma C.3. *One has* $\langle \partial_h^e \eta, v\rangle_{L_h^2} = -\langle \tau_h^e \eta, \delta_h^e v\rangle_{L_h^2}$, *for any* $e \in \mathcal{Z}_d$ *and* $\tau_h^e \eta, v : \mathbb{T}_h^d \to \mathbb{R}$.

Proof. We compute, using a translation by he in the second sum of the second line

$$\begin{aligned} h^{-d}\langle \partial_h^e \eta, v\rangle_{L_h^2} &= \sum_{x \in \mathbb{T}_h^d} \eta(x + he/2)v(x) - \sum_{x \in \mathbb{T}_h^d} \eta(x - he/2)v(x) \\ &= \sum_{x \in \mathbb{T}_h^d} \eta(x + he/2)v(x) - \sum_{x \in \mathbb{T}_h^d} \eta(x + he/2)v(x + he) = -h^{-d}\langle \tau_h^e \eta, \delta_h^e v\rangle_{L_h^2}. \end{aligned}$$ □

Proposition C.4. *Assume that* $\mathcal{L}_h u = \sum_{e \in \mathcal{Z}_d^R} \partial_h^e \eta_e$, *where* $u, \tau_h^e \eta_e : \mathbb{T}_h^d \to \mathbb{R}$, *for all* $e \in \mathcal{Z}_d^R$. *Then* $\|\nabla_h u\|_{L_h^2} \leq C \sum_{e \in \mathcal{Z}_d^R} \|\tau_h^e \eta_e\|_{L_h^2}$ *for some constant* $C = C(M_{\mathcal{L}})$.

Proof. Any vector $e \in \mathbb{Z}^d$ can be written as $e = e_0 + \cdots + e_{n-1}$, where $\pm e_i$ belongs to the canonical basis, and $n = |e|_{l^1}$. Thus $n \leq n_0 := R\sqrt{d}$ if $|e| \leq R$. By Corollary 5.7, we obtain

$$\|\delta_h^e v\|_{L_h^2}^2 \leq n\big[\|\delta_h^{e_0} v\|_{L_h^2}^2 + \cdots + \|\delta_h^{e_{n-1}} v\|_{L_h^2}^2\big] \leq n_0^2 \|\nabla_h v\|_{L_h^2}^2, \tag{56}$$

for any $v : \mathbb{T}_h^d \to \mathbb{R}$. It follows that

$$|\langle \mathcal{L}_h u, v\rangle_{L_h^2}| \leq \sum_{e \in \mathcal{Z}_d} |\langle \tau_h^e \eta_e, \delta_h^e v\rangle| \leq n_0 \sum_{e \in \mathcal{Z}_d} \|\tau_h^e \eta_e\|_{L_h^2} \|\nabla_h v\|_{L_h^2}$$

where we used successively (i) the assumption, and the discrete integration by parts of Lemma C.3, and (ii) the Cauchy-Schwarz inequality and (56). We conclude the proof by choosing $v := u$ and using (53). □

We define convolution operators T_h^e, T_h^r and T_h as follows

$$\mathrm{T}_h^e\, u(x) := \int_{-\frac{1}{2}}^{\frac{1}{2}} u(x + t\, he)\, \mathrm{d}t, \qquad \mathrm{T}_h^r := \prod_{e \in \mathcal{Z}_d^r} \mathrm{T}_h^e, \qquad \mathrm{T}_h := \mathrm{T}_h^1\, \mathrm{T}_h^R, \tag{57}$$

for any $e \in \mathcal{Z}_d$ and any $r \geq 1$. Thus T_h^e denotes convolution along the segment $[-he/2, he/2]$, and T_h^1 denotes convolution with the indicator function of the unit cube $[-h/2, h/2]^d$, as already mentioned in Section 1.1. The larger convolution kernel T_h is the composition of T_h^1 with convolutions in all directions $e \in \mathcal{Z}_d^R$ potentially arising in the discretization stencils. Note that convolution operators commute with each other, with the differentiation operator ∂^e, with the finite difference ∂_h^e, and with the translation operator τ_h^e.

We also introduce a formal inverse $\hat{\mathrm{T}}_h^e$ of T_h^e, which is not an actual operator, but a *convention of notation*. Indeed the operator T_h^e is not invertible, but vanishes for instance on the function $u(x) := \sin(2\pi\langle x, f\rangle/h)$ which oscillates with high frequency, where $f \in \mathbb{Z}^d$ is arbitrary. Thus $\hat{\mathrm{T}}_h^e$ is never considered alone, but always within a product of convolution operators featuring T_h^e, from which this factor should be removed, and for clarity this grouping is emphasized by the use of square brackets. For instance given $e, e_1, \cdots, e_K \in \mathcal{Z}$ and $1 \leq k \leq K$ one has as a convention of notation

$$[\hat{\mathrm{T}}_h^e\, \mathrm{T}_h^e] := \mathrm{Id}, \qquad [\hat{\mathrm{T}}_h^{e_k}\, \mathrm{T}_h^{e_1} \cdots \mathrm{T}_h^{e_K}] := \mathrm{T}_h^{e_1} \cdots \mathrm{T}_h^{e_{k-1}}\, \mathrm{T}_h^{e_{k+1}} \cdots \mathrm{T}_h^{e_K}\,.$$

Likewise the expressions $[\hat{\mathrm{T}}_h^e\, \mathrm{T}_h^R]$ and $[\hat{\mathrm{T}}_h^{b_i}\, \mathrm{T}_h^1]$ make sense for any $e \in \mathcal{Z}_d^R$ and any $1 \leq i \leq d$.

Lemma C.5. *One has $\mathrm{T}_h^e\, \partial^e = \partial_h^e$, and thus $\mathrm{T}_h\, \partial^e = [\hat{\mathrm{T}}_h^e\, \mathrm{T}_h]\partial_h^e$, for any $e \in \mathcal{Z}_d^R$.*

Proof. The first claim is established by direct integration. Assuming w.l.o.g. that $d = 1$ and $e = 1$

$$\int_{-\frac{1}{2}}^{\frac{1}{2}} u'(x + th)\, \mathrm{d}t = \frac{u(x + h/2) - u(x - h/2)}{h}.$$

The second claim follows, since convolutions and differentiations commute. □

Proposition C.6. *Assume that* $\mathcal{L}u = \mathrm{f}$ *and* $\mathcal{L}_h u_h = \mathrm{f}_h := \mathrm{T}_h^1 \, \mathrm{T}_h^1 \, \mathrm{f}$. *Then*

$$\mathcal{L}_h(\mathrm{T}_h \, u - u_h) = \sum_{1 \le i \le d} \partial_h^{b_i} \eta_1^i + \sum_{e \in \mathcal{Z}_d^R} \partial_h^e (\eta_2^e + \eta_3^e + \eta_4^e)$$

on $\mathbb{T}_h^d$, *where denoting* $u_e := \partial^e u$ *and* $u_i := \langle b_i, D\nabla u \rangle$ *one has*

$$\begin{aligned} \eta_1^i &:= [\hat{\mathrm{T}}_h^{b_i} \, \mathrm{T}_h^1](\mathrm{T}_h^R - \mathrm{T}_h^1) u_i, & \eta_2^e &:= (\lambda_h^e - \lambda^e)[\mathrm{T}_h^e \, \mathrm{T}_h] u_e, \\ \eta_3^e &:= \lambda^e [\mathrm{T}_h^e \, \mathrm{T}_h - \hat{\mathrm{T}}_h^e \, \mathrm{T}_h] u_e, & \eta_4^e &:= \lambda^e [\hat{\mathrm{T}}_h^e \, \mathrm{T}_h] u_e - [\hat{\mathrm{T}}_h^e \, \mathrm{T}_h](\lambda^e u_e). \end{aligned}$$

Proof. We first define and compute, using Lemma C.5,

$$\eta_1 := \sum_{i=1}^d \partial_h^{b_i} \eta_1^i = (\mathrm{T}_h - \mathrm{T}_h^1 \, \mathrm{T}_h^1) \sum_{i=1}^d \partial^{b_i} (D\nabla u) = (\mathrm{T}_h - \mathrm{T}_h^1 \, \mathrm{T}_h^1) \, \mathrm{f}.$$

Then, again by direct computation, using (55) and Lemma C.5,

$$\begin{aligned} \mathcal{L}_h \, \mathrm{T}_h \, u &= \sum_{e \in \mathcal{Z}_d^R} \partial_h^e (\lambda_h^e \partial_h^e \, \mathrm{T}_h \, u) = \sum_{e \in \mathcal{Z}_d^R} \partial_h^e (\lambda_h^e [\mathrm{T}_h^e \, \mathrm{T}_h] \partial^e u), \\ \mathcal{L}_h u_h + \eta_1 &= \mathrm{T}_h \, \mathrm{f} = \sum_{e \in \mathcal{Z}_d^R} \mathrm{T}_h \, \partial^e (\lambda^e \partial^e u) = \sum_{e \in \mathcal{Z}_d^R} \partial_h^e ([\hat{\mathrm{T}}_h^e \, \mathrm{T}_h](\lambda^e \partial^e u)), \end{aligned}$$

We conclude observing that $\eta_2^e + \eta_3^e + \eta_4^e = \lambda_h^e [\mathrm{T}_h^e \, \mathrm{T}_h] \partial^e u - [\hat{\mathrm{T}}_h^e \, \mathrm{T}_h](\lambda^e \partial^e u)$. □

Combining Propositions C.4 and C.6 we obtain

$$\|\nabla(\mathrm{T}_h \, u - u_h)\|_{L_h^2} \le C \max\{\|\tau_h^{b_i} \eta_1^i\|_{L_h^2}, \|\tau_h^e \eta_2^e\|_{L_h^2}, \|\tau_h^e \eta_3^e\|_{L_h^2}, \|\tau_h^e \eta_4^e\|_{L_h^2} \mid 1 \le i \le d, \, e \in \mathcal{Z}_d^R\} \tag{58}$$

for some constant $C = C(M_{\mathcal{L}})$. In the rest of this section, we present basic estimates of the norms of convolutions, see Lemmas C.7, C.9 and C.13, followed by specializations to the members of (58, r.h.s.), see Corollaries C.8, C.11, C.12 and C.14, which together imply that $\|\nabla(\mathrm{T}_h \, u - u_h)\|_{L_h^2} \le C \min\{h \|\nabla^2 u\|_{L^2}, h^2 \|\nabla^3 u\|_{L^2}\}$. The additional estimate Corollary C.10 concludes the proof of Theorem 1.4.

Lemma C.7. *For any* $v \in L^2(\mathbb{T}^d)$, *and any* $e \in \mathcal{Z}_d$, *one has*

$$\|\, \mathrm{T}_h^e \, v\|_{L^2} \le \|v\|_{L^2} \qquad\qquad \|\, \mathrm{T}_h^1 \, v\|_{L_h^2} \le \|v\|_{L^2}. \tag{59}$$

Also $E_h[\mathrm{f}_h] = 0$ *(recall that* $\mathrm{f}_h := \mathrm{T}_h^1 \, \mathrm{T}_h^1 \, \mathrm{f}$ *and* $E[\mathrm{f}] = 0$*), with the notations* E *and* E_h *of Lemma C.2.*

Proof. The estimate (59, left) holds by convexity of the norm and since T_h^e is the convolution with a non-negative kernel of unit integral. Denote by $Z := [-1/2, 1/2]^d$ the unit cube, and observe that $(x, z) \in \mathbb{T}_h^d \times Z \mapsto x + hz \in \mathbb{T}$ is a.e. bijective with Jacobian h^d. One has $|\,\mathrm{T}_h^1\, v(x)|^2 = (\int_Z v(x+hz)\mathrm{d}z)^2 \leq \int_Z v(x+hz)^2\mathrm{d}z$, by the Cauchy-Schwartz inequality, and by summation over $x \in \mathbb{T}_h^d$ we obtain (59, right) as announced.

Similarly, one has $E[\mathrm{T}_h^e\, v] = E[v]$ and $E_h[\mathrm{T}_h^1\, v] = E[v]$, for any $e \in \mathcal{Z}_d$. Thus $E_h[\mathrm{f}_h] = E_h[\mathrm{T}_h^1\,\mathrm{T}_h^1\,\mathrm{f}] = E[\mathrm{T}_h^1\,\mathrm{f}] = E[\mathrm{f}]$ which concludes the proof. □

Corollary C.8. *One has $\|\tau_h^e \eta_2^e\|_{L_h^2} \leq Ch\|\nabla u\|_{L^2}$, for some constant $C = C(M_\mathcal{L})$ (resp. $\|\tau_h^e \eta_2^e\|_{L_h^2} \leq Ch^2\|\nabla u\|_{L^2}$ for some constant $C = C(M_\mathcal{L}, \|\nabla^2\lambda\|_\infty)$), and all $e \in \mathcal{Z}_d^R$.*

Proof. One has $\|\tau_h^e \eta_2^e\|_{L_h^2} \leq \|\lambda_h^e - \lambda^e\|_\infty \|\tau_h^e\,\mathrm{T}_h^e\,\mathrm{T}_h\, u_e\|_{L_h^2}$. Also, $\|\lambda_h^e - \lambda^e\|_\infty \leq \min\{h|e|\|\nabla\lambda^e\|_\infty, h^2|e|^2\|\nabla^2\lambda^e\|_\infty\}$ by construction (54). In addition $\|\tau_h^e\,\mathrm{T}_h^e\,\mathrm{T}_h\, u_e\|_{L_h^2} = \|\,\mathrm{T}_h^1\,\mathrm{T}_h^e\,\mathrm{T}_h^R\,\tau_h^e u_e\|_{L_h^2} \leq \|\,\mathrm{T}_h^e\,\mathrm{T}_h^R\,\tau_h^e u_e\|_{L^2} \leq \|u_e\|_{L^2} \leq |e|\|\nabla u\|_{L^2}$, using successively (59, right) and (59, left). □

Lemma C.9. *Let $e_1, \cdots, e_n \in \mathcal{Z}_d^R$ and $\mathrm{T}_h^{\mathbf{e}} := \mathrm{T}_h^{e_1} \cdots \mathrm{T}_h^{e_n}$. Then for any $v : \mathbb{T}^d \to \mathbb{R}$*

$$\|\,\mathrm{T}_h^{\mathbf{e}}\, v - v\|_{L^2} \leq nRh\|\nabla v\|_{L^2}, \qquad \|\,\mathrm{T}_h^{\mathbf{e}}\, v - v\|_{L^2} \leq nR^2h^2\|\nabla^2 v\|_{L^2}.$$

Proof. For any smooth $f : [-1/2, 1/2] \to \mathbb{R}$ and $|t| \leq 1/2$, by the Taylor integral formula

$$|f(t) - f(0)| \leq \int_{-\frac{1}{2}}^{\frac{1}{2}} |f'(s)|\mathrm{d}s, \qquad |f(t) + f(-t) - 2f(0)| \leq \int_{-\frac{1}{2}}^{\frac{1}{2}} |f''(s)|\mathrm{d}s.$$

It follows that $|\int_{-1/2}^{1/2} f(t)\mathrm{d}t - f(0)| \leq \min\{\int_{-1/2}^{1/2} |f'(s)|\mathrm{d}s, \int_{-1/2}^{1/2} |f''(s)|\mathrm{d}s\}$. Choosing $f(t) := v(x + the)$, where w.l.o.g. v is assumed to be smooth, $e \in \mathcal{Z}_d$, and $x \in \mathbb{T}^d$ is arbitrary, we obtain $|\,\mathrm{T}_h^e\, v - v| \leq \min\{h\,\mathrm{T}_h^e\,|\langle e, \nabla v\rangle|, h^2\,\mathrm{T}_h^e\,|\langle e, \nabla^2\, v\, e\rangle|\}$ pointwise (recall that the expression (57) of the operator T_h^e involves an integral over $[-1/2, 1/2]$). Thus $\|\,\mathrm{T}_h^e\, v - v\|_{L^2} \leq \min\{h\|\langle e, \nabla v\rangle\|_{L^2}, h^2\|\langle e, \nabla^2 v\, e\rangle\|_{L^2}\} \leq \min\{h|e|\|\nabla v\|_{L^2}, h^2|e|^2\|\nabla^2 v\|_{L^2}\}$, by (59, left), which establishes the case $n = 1$. Observing that $\mathrm{T}_h^{\mathbf{e}}\, v - v = \sum_{k=1}^n \mathrm{T}_h^{e_1} \cdots \mathrm{T}_h^{e_{k-1}}(\mathrm{T}_h^{e_k}\, v - v)$, we obtain the announced estimates using again (59, left). □

Corollary C.10. *$\|\nabla_h(\mathrm{T}_h^1\, u - \mathrm{T}_h\, u)\|_{L_h^2} \leq C \min\{h\|\nabla^2 u\|_{L^2}, h^2\|\nabla^3 u\|_{L^2}\}$, with $C = C(M_\mathcal{L})$.*

Proof. One has $\partial_h^{b_i}(\mathrm{T}_h^1\, u - \mathrm{T}_h\, u) = \mathrm{T}_h^1\,\mathrm{T}_h^{b_i}(\mathrm{Id} - \mathrm{T}_h^R)v_i$, for any $1 \leq i \leq d$ and with $v_i := \partial^{b_i} u$, using Lemma C.5. Then $\|\,\mathrm{T}_h^1\,\mathrm{T}_h^{b_i}(\mathrm{Id} - \mathrm{T}_h^R)v_i\|_{L_h^2} \leq Ch\|\nabla v_i\|_{L^2}$ (resp. $\leq Ch^2\|\nabla^2 v_i\|_{L^2}$) using successively (59, right), (59, left), and Lemma C.9. Summing over i we conclude. □

Corollary C.11. *One has $\|\eta_1^i\|_{L_h^2} \leq Ch\|\nabla^2 u\|_{L^2}$ for some constant $C = C(M_\mathcal{L})$ (resp. $\|\eta_1^i\|_{L_h^2} \leq Ch^2\|\nabla^3 u\|_{L^2}$ for some constant $C = C(M_\mathcal{L}, \|\nabla^2\lambda\|_\infty)$), and all $1 \leq i \leq d$.*

Proof. Define $\mathrm{T}_h^* := \prod_{e\in\mathcal{Z}_d}^{1<|e|\le R} \mathrm{T}_h^e$, in such way that $\mathrm{T}_h^R = \mathrm{T}_h^1\,\mathrm{T}_h^*$. By construction $\eta_1^i = \mathrm{T}_h^1[\hat{\mathrm{T}}_h^{b_i}\,\mathrm{T}_h^1](\mathrm{T}_h^*\,u_i - u_i)$, and therefore $\|\eta_1^i\|_{L_h^2} \le C\min\{h\|\nabla u_i\|_{L^2}, h^2\|\nabla^2 u_i\|_{L^2}\}$, using successively (59, right), (59, left), and Lemma C.9.

Recalling that $u_i := \langle b_i, D\nabla u\rangle$, we obtain $\|\nabla u_i\|_{L^2} \le C(\|\nabla\mathcal{D}\|_\infty\|\nabla u\|_{L^2} + \|\mathcal{D}\|_\infty\|\nabla^2 u\|_{L^2})$, and $\|\nabla^2 u_i\|_{L^2} \le C(\|\nabla^2\mathcal{D}\|_\infty\|\nabla u\|_{L^2} + \|\nabla\mathcal{D}\|_\infty\|\nabla^2 u\|_{L^2} + \|\mathcal{D}\|_\infty\|\nabla^3 u\|_{L^2})$, by the Leibniz rule for the differentiation of a product. We conclude using (52). □

Corollary C.12. *One has $\|\eta_3^e\|_{L_h^2} \le Ch\|\nabla^2 u\|_{L^2}$ (resp. $\|\eta_3^e\|_{L_h^2} \le Ch^2\|\nabla^3 u\|_{L^2}$), for some constant $C = C(M_{\mathcal{L}})$, and all $e \in \mathcal{Z}_d^R$.*

Proof. Observing that $\eta_3^e = \lambda^e\,\mathrm{T}_h^1[\mathrm{T}_h^R\,\hat{\mathrm{T}}_h^e](\mathrm{T}_h^e\,\mathrm{T}_h^e\,u_e - u_e)$, we obtain as announced that $\|\eta_3^e\|_{L_h^2} \le Ch\|\lambda\|_\infty\|\nabla u_e\|_{L^2}$ (resp $\|\eta_3^e\|_{L_h^2} \le Ch^2\|\lambda\|_\infty\|\nabla^2 u_e\|_{L^2}$) using successively (59, right), (59, left), and Lemma C.9. □

Lemma C.13. *Let $e_1\cdots, e_n \in \mathcal{Z}_d^R$ and $\mathrm{T}_h^{\mathbf{e}} := \mathrm{T}_h^{e_1}\cdots\mathrm{T}_h^{e_n}$. Let also $\alpha, v : \mathbb{T}^d \to \mathbb{R}$, and $\eta := \alpha\,\mathrm{T}_h^{\mathbf{e}}\,v - \mathrm{T}_h^{\mathbf{e}}(\alpha v)$. Then one has pointwise*

$$|\eta| \le nhR\|\nabla\alpha\|_\infty\,\mathrm{T}_h^{\mathbf{e}}\,|v|, \qquad |\eta| \le nh^2R^2\,\mathrm{T}_h^{\mathbf{e}}(\|\nabla^2\alpha\|_\infty|v| + \|\nabla\alpha\|_\infty|\nabla v|) \tag{60}$$

If the canonical basis is among the $(e_i)_{i=1}^n$ (e.g. $n \ge d$ and $e_i = b_i$ for all $1 \le i \le d$), then

$$\|\eta\|_{L_h^2} \le nhR\|\nabla\alpha\|_\infty\|v\|_{L^2}, \qquad \|\eta\|_{L_h^2} \le nh^2R^2(\|\nabla^2\alpha\|_\infty\|v\|_{L^2} + \|\nabla\alpha\|_\infty\|\nabla v\|_{L^2}). \tag{61}$$

Proof. Consider smooth $f, g : [-1/2, 1/2] \to \mathbb{R}$, and let $|t| \le 1/2$. Clearly one has,

$$|E(t)| \le \|f'\|_\infty|g(t)|, \qquad \text{where } E(t) := \big(f(t) - f(0)\big)g(t). \tag{62}$$

Yet, using the identity $2(ab+cd) = (a+c)(b+d) + (a-c)(b-d)$ we also note that

$$2(E(t) + E(-t)) = \big(f(t) + f(-t) - 2f(0)\big)\big(g(t) + g(-t)\big) + \big(f(t) - f(-t)\big)\big(g(t) - g(-t)\big),$$

leading to the finer estimate

$$2|E(t) + E(-t)| \le \|f''\|_\infty\big(|g(t)| + |g(-t)|\big) + \|f'\|_\infty\int_{-1/2}^{1/2}|g'(s)|\mathrm{d}s. \tag{63}$$

Define $E := \int_{-1/2}^{1/2}(f(t) - f(0))g(t)\mathrm{d}t$, so that $E = \int_{-1/2}^{1/2}E(t)\mathrm{d}t = \int_0^{1/2}(E(t) + E(-t))\mathrm{d}t$. We obtain using (62) and (63)

$$|E| \le \|f'\|_\infty\int_{-1/2}^{1/2}|g(t)|\mathrm{d}t, \qquad |E| \le \|f''\|_\infty\int_{-1/2}^{1/2}|g(t)|\mathrm{d}t + \|f'\|_\infty\int_{-1/2}^{1/2}|g'(t)|\mathrm{d}t.$$

Applying these estimates to $f(t) := \alpha(x + t\,he)$ and $g(t) := v(x + t\,he)$, where $e \in \mathcal{Z}_d^R$ and $x \in \mathbb{T}^d$ is arbitrary, we obtain (60) in the case $n = 1$. The general case follows noting that $\eta = \sum_{1\le k\le n}\mathrm{T}_h^{e_1}\cdots\mathrm{T}_h^{e_{k-1}}(\alpha\,\mathrm{T}_h^{e_k} - \mathrm{T}_h^{e_k}\,\alpha)\,\mathrm{T}_h^{e_{k+1}}\cdots\mathrm{T}_h^{e_n}\,v$. Finally, we obtain (61) by Lemma C.7, which concludes the proof. □

Corollary C.14. *One has* $\|\tau_h^e \eta_4^e\|_{L_h^2} \leq Ch\|\nabla u\|_{L^2}$ *for some constant* $C = C(M_{\mathcal{L}})$ *(resp.* $\|\tau_h^e \eta_4^e\|_{L_h^2} \leq \tilde{C}h^2\|\nabla^2 u\|_{L^2}$ *for some constant* $C = C(M_{\mathcal{L}}, \|\nabla^2\lambda\|_\infty)$*), for all* $e \in \mathcal{Z}_d^R$.

Proof. Recall that $\eta_3^e := \lambda^e[\mathrm{T}_h^1 \hat{\mathrm{T}}_h^e \mathrm{T}_h^R]u_e - [\mathrm{T}_h^1 \hat{\mathrm{T}}_h^e \mathrm{T}_h^R](\lambda^e u_e)$. The result follows from (61), applied to $\alpha := \lambda^e$ and $v := u_e = \langle e, \nabla u\rangle$, and from (52). □

C.1 Establishing a convergence rate using discrete monotonicity

The monotonicity of a finite differences scheme yields an alternative avenue for establishing convergence rates, independent of the previous arguments which rely on the Lax-Milgram theorem, exploited in Proposition C.15 to establish a convergence rate for this specific elliptic equation. See [21] for more refined techniques, and Appendix A for a general discussion. This approach is in a sense more direct, since the proof essentially only consists in checking the scheme consistency. In comparison with the previous section, we obtain convergence rates in the uniform norm, rather than the L^2 norm of the discrete gradient. For simplicity, we introduce a zeroth order term in the PDE, so as to break the invariance of the solution under the addition of a constant. Note that it is often possible to exploit the comparison principle for PDEs whose set of solutions is invariant under addition of a constant, and for their discretizations, but this leads to additional technicalities, see [7] for a discussion in the case of the Monge-Ampère equation of optimal transport.

Proposition C.15. *Assume that* $u - \mathcal{L}u = \mathrm{f}$ *on* $\mathbb{T}^d$*, for some* $u \in C^4(\mathbb{T}^d)$*,* $\mathcal{D} \in C^3(\mathbb{T}^d, \mathcal{S}_d^{++})$ *and* $\mathrm{f} \in C^2(\mathbb{T}^d)$*. Consider coefficients* $\lambda : \mathbb{T}^d \to \Lambda_d$ *which are* $\mathcal{D}$*-consistent, R-supported, and have* C^3 *regularity. Then the linear system* $u_h - \mathcal{L}_h u_h = \mathrm{f}$ *on* $\mathbb{T}_h^d$ *has a unique solution, for any* $h > 0$ *with* $h^{-1} \in \mathbb{Z}_{++}$*, and one has for some constant* $C = C(d, R, \|\nabla\mathcal{D}\|_\infty, \|\nabla^3\lambda\|_\infty)$

$$\max_{x \in \mathbb{T}_h^d} |u(x) - u_h(x)| \leq Ch^2\|\nabla^4 u\|_\infty. \tag{64}$$

Proof. We note that $-\mathcal{L}_h$ is a degenerate elliptic scheme, see its expression (5) and Definition A.2. Thus $u_h \mapsto u_h - \mathcal{L}_h u_h$ is a linear elliptic scheme, which implies the existence of a unique solution, see [5, Corollary 3.6]. Let $f \in C^3(\mathbb{R})$, $g \in C^4(\mathbb{R})$, and define $e(h) := \frac{1}{2}(f(0) + f(h))(g(h) - g(0))$ for all $h \in \mathbb{R}$. We obtain by a Taylor expansion

$$e(h) = hf_0 g_0' + \tfrac{1}{2}h^2(f_0' g_0' + f_0 g_0'') + \tfrac{1}{12}h^3(3f_0'' g_0' + 3f_0' g_0'' + 2g_0 g_0''') + \mathcal{O}(h^4),$$

with the convention $f_0 := f(0)$ and likewise for f', f'', g', g'', g''' evaluated at 0. Therefore $e(h) + e(-h) = h^2(f_0' g_0' + f_0 g_0'') + \mathcal{O}(h^4) = h^2(fg')'(0) + \mathcal{O}(h^4)$. Choosing $f(h) := \lambda(x + he)$ and $g(h) := u(x + he)$, where $e \in \mathcal{Z}_d^R$ and $x \in \mathbb{T}_h^d$ is arbitrary, this yields

$$\partial_h^e(\lambda_h^e \partial_h^e u) = \partial^e(\lambda^e \partial^e u) + \mathcal{O}(h^2), \qquad \text{hence } |\mathcal{L}_h u - \mathcal{L}u| \leq C'h^2,$$

pointwise on $\mathbb{T}_h^d$, where $C' = C\|\nabla^4 u\|_\infty$ with constant $C = C(d, R, \|\nabla \mathcal{D}\|_\infty, \|\nabla \lambda^3\|_\infty)$. (Recall that $\|\nabla^k \lambda\|_\infty \leq \|\nabla^{k+1} \lambda\|_\infty$ and $\|\nabla^k u\|_\infty \leq \|\nabla^{k+1} u\|_\infty$ on $\mathbb{T}^d$, for all $k \geq 1$, as already observed in (52, right).) We thus have have on $\mathbb{T}_h^d$

$$u - \mathcal{L}_h u - C'h^2 \leq u - \mathcal{L}u \leq u - \mathcal{L}_h u + C'h^2, \qquad u - \mathcal{L}u = \mathrm{f} = u_h - \mathcal{L}_h u_h.$$

By the discrete comparison principle [5, Lemma 3.5], and since $\mathcal{L}_h$ vanishes on constants, this implies $u - C'h^2 \leq u_h \leq u + C'h^2$, which concludes. □