Second order monotone finite differences discretization of linear anisotropic differential operators

Frédéric Bonnans, Guillaume Bonnet, Jean-Marie Mirebeau

Second order monotone finite differences discretization of linear anisotropic differential operators

Frédéric Bonnans* Guillaume Bonnet† Jean-Marie Mirebeau‡

December 20, 2020

Abstract

We design adaptive finite differences discretizations, which are degenerate elliptic and second order consistent, of linear and quasi-linear partial differential operators featuring both a first order term and an *anisotropic* second order term. Our approach requires the domain to be discretized on a Cartesian grid, and takes advantage of techniques from the field of low-dimensional lattice geometry. We prove that the stencil of our numerical scheme is optimally compact, in dimension two, and that our approach is quasi-optimal in terms of the compatibility condition required of the first and second order operators, in dimension two and three. Numerical experiments illustrate the efficiency of our method in several contexts.

1 Introduction

In this paper, we design finite difference discretizations of Degenerate Elliptic (DE) Partial Differential Equations (PDEs). This class of equations is sufficiently general to encompass a wide variety of applications, in the fields of optimal transport, game theory, differential geometry, stochastic modeling and finance, optimal control, ... Our results are limited to linear and quasi-linear operators, but could in principle be used as a building block for the discretization of fully non-linear operators, see Appendix A. On the other hand, the assumption of degenerate ellipticity yields comparison principles and stability properties [CIL92].

Discrete Degenerate Ellipticity (DDE), for numerical schemes, implies similarly strong properties [Obe06], which often turn proofs of convergence into simple verifications. A known limitation of monotone discretization schemes is their consistency order with the original PDE, which cannot exceed two for second order operators and one for first order operators [Obe06]. However, many common implementations of second order DE operators only achieve first order consistency, or sometimes less. They may also rely on excessively wide stencils [FO11, LN18], especially in the context of two-scales discretizations [LN18]. This degrades the accuracy of the numerical results, which severely constrains the practical uses of these methods. The objective of this paper is to characterize when a second order monotone discretization is feasible, and

*Inria-Saclay and CMAP, École Polytechnique, Palaiseau, France
Frédéric Bonnans acknowledges support from the Chair Finance & Sustainable Development and of the FiME Lab (Institut Europlace de Finance).
†LMO, Université Paris-Saclay, Orsay, France, and Inria-Saclay and CMAP, École Polytechnique, Palaiseau, France
‡University Paris-Saclay, ENS Paris-Saclay, CNRS, Centre Borelli, F-91190 Gif-sur-Yvette, France
This work was partly supported by ANR research grant MAGA, ANR-16-CE40-0014.

how wide the numerical scheme stencil must be, especially when the second order part of the operator is strongly anisotropic, and the first order term is non-vanishing.

We state our theoretical results in the context of linear operators with constant coefficients, defined over $\mathbb{R}^d$ where $d \in \{2, 3\}$. Because degenerate ellipticity is a local property, which is stable under a variety of transformations, they admit straightforward extensions to quasi-linear operators and some fully non-linear operators. Non-constant coefficients and bounded domains with Dirichlet boundary conditions are also easily handled. See Appendix A and the numerical experiments §4 for these extensions. Without loss of generality, our theoretical results are stated in the context of linear operators with constant coefficients.

We define the linear operator $\mathcal{L} = \mathcal{L}[\omega, D]$ on $\mathbb{R}^d$ by the expression

$$-\mathcal{L}u(x) := \langle \omega, \nabla u(x) \rangle + \mathrm{Tr}(D\nabla^2 u(x)), \tag{1.1}$$

where $\omega \in \mathbb{R}^d$, $D \in S_d^{++}$ is a symmetric positive definite matrix, and the unknown $u : \mathbb{R}^d \to \mathbb{R}$ is a smooth function. Likewise in the discrete setting we define the finite differences operator $L^h = L^h[\rho_i^h, e_i]_{1 \le |i| \le I}$, on the Cartesian grid $h\mathbb{Z}^d$ with grid scale $h > 0$, by the expression

$$-L^h u(x) := h^{-2} \sum_{1 \le |i| \le I} \rho_i^h \big(u(x + he_i) - u(x)\big), \tag{1.2}$$

where $\rho_{-I}^h, \cdots, \rho_{-1}^h, \rho_1^h, \cdots, \rho_I^h \ge 0$ are non-negative weights, and $e_1, \cdots, e_i \in \mathbb{Z}^d$ are offsets with integer entries, for some positive integer I. Here and throughout this paper, without loss of generality, we use the convention that $e_{-i} := -e_i$ for all $1 \le i \le I$.

Note that any translation invariant linear operator on $h\mathbb{Z}^d$, finitely supported and vanishing on constant functions, can be written in the form (1.2). We denote by S_d the set of symmetric $d \times d$ matrices, by S_d^+ the subset of semi-definite ones, and by S_d^{++} the positive definite ones.

Definition 1.1. The operator $\mathcal{L}[\omega, D]$ is said Degenerate Elliptic (DE) if $D \in S_d^+$. The discrete operator $L^h[\rho_i^h, e_i]_{1 \le |i| \le I}$ is said Discrete Degenerate Elliptic (DDE) if $\rho_i^h \ge 0$ for all $1 \le |i| \le I$.

In particular, the DE and DDE properties do not impose any restrictions on the first order term $\omega \in \mathbb{R}^d$, and the numerical scheme offsets $e_i \in \mathbb{Z}^d$.

Definition 1.2 (Absolute feasibility)**.** We say that the pair $(h\omega, D)$ is *absolutely feasible* if there exists $(\rho_i^h, e_i)_{1 \le |i| \le I}$ such that $\mathcal{L}[\omega, D]$ and $L^h[\rho_i^h, e_i]_{1 \le |i| \le I}$ are both degenerate elliptic, and are equal on all quadratic functions u. Equivalently, one has

$$\sum_{1 \le |i| \le I} \rho_i^h e_i = h\omega, \qquad \sum_{1 \le |i| \le I} \rho_i^h e_i e_i^{\mathrm{T}} = D, \tag{1.3}$$
$$\text{and} \quad \rho_i^h \ge 0, \ e_i \in \mathbb{Z}^d \setminus \{0\}, \ \forall 1 \le |i| \le I.$$

Definition 1.2 is stated in terms of the pair $(h\omega, D)$, and not of the triplet (h, ω, D), because it only depends eventually on the product $h\omega$, as shown by algebraic characterization (1.3). In dimension $d = 1$, one easily checks that $(h\omega, D)$ is absolutely feasible iff $h\omega \le 2D$, and in that case the standard discretization using centered finite differences obeys (1.3). Note that discretizing (1.1) using upwind finite differences for the first order term fails the consistency test (1.3). In this paper, we fully characterize when a pair $(h\omega, D)$ is absolutely feasible in dimension $d \in \{2, 3\}$, see Proposition 3.10, which is not straightforward unless D is a diagonal matrix. More practically, we provide an explicit scheme construction.

Our numerical scheme relies on a tool from a lattice geometry known as *Selling's decomposition*, described in more detail in §2.1, see also [Sel74, CS92]. It associates to each positive definite matrix $D \in S_d^{++}$, where $d \in \{2,3\}$, a specific decomposition of the following form

$$D = \sum_{1 \leq i \leq I} \sigma_i e_i e_i^{\mathrm{T}}, \qquad \text{where } \sigma_i \geq 0,\ e_i \in \mathbb{Z}^d,\ \forall 1 \leq i \leq I, \tag{1.4}$$

and $I := d(d+1)/2$. Selling's decomposition has already been used in the design of difference schemes in dimension $d \in \{2,3\}$, for (divergence form) anisotropic diffusion in [FM14], or various anisotropic eikonal equations in [Mir17, Mir19]. It is at the foundation of degenerate elliptic and second order consistent discretizations of the fully non-linear two dimensional Monge-Ampere [BCM] and Pucci [BBM21] equations. In dimension $d = 2$, an equivalent construction based on the Stern-Brocot dyadic tree of rational numbers is used in [BOZ04] for the Hamilton-Jacobi-Bellman equation of Stochastic control.

The support $(e_i)_{i=1}^I$ of Selling's decomposition, which is also the stencil of the numerical scheme proposed in this paper, tends to align with the anisotropy defined by the matrix D. This is illustrated on Figure 1, where we use the following parametrization of the set of symmetric positive definite matrices of size two and with unit determinant:

$$D(a,b) := \frac{1}{\sqrt{1-a^2-b^2}} \begin{pmatrix} 1+a & b \\ b & 1-a \end{pmatrix}, \qquad a^2 + b^2 < 1. \tag{1.5}$$

Definition 1.3 (Finite difference operators)**.** For any $e \in \mathbb{Z}^d$, $h > 0$, $u : h\mathbb{Z}^d \to \mathbb{R}$, we let

$$\delta_e^h u(x) := \frac{u(x+he) - u(x-he)}{2h}, \qquad \Delta_e^h u(x) := \frac{u(x+he) - 2u(x) + u(x-he)}{h}.$$

Given $D \in S_d^{++}$ where $d \in \{2,3\}$, with Selling decomposition $D = \sum_{1 \leq i \leq I} \sigma_i e_i e_i^{\mathrm{T}}$, we let

$$\nabla_D^h u(x) = \sum_{1 \leq i \leq I} \sigma_i\, \delta_{e_i}^h u(x)\, e_i, \qquad \Delta_D^h u(x) = \sum_{1 \leq i \leq I} \sigma_i \Delta_{e_i}^h u(x).$$

The centered finite differences $\delta_e^h u(x) = \langle e, \nabla u(x) \rangle + \mathcal{O}(h^2)$ and second order finite differences $\Delta_e^h u(x) = \langle e, \nabla^2 u(x) e \rangle + \mathcal{O}(h^2)$, are classical constructs. In combination with Selling's decomposition, they are here used to define discrete anisotropic gradient and laplacian operators, with the following consistency properties easily derived from (1.4)

$$\nabla_D^h u(x) = D \nabla u(x) + \mathcal{O}(h^2), \qquad \Delta_D^h u(x) = \mathrm{Tr}(D \nabla^2 u(x)) + \mathcal{O}(h^2). \tag{1.6}$$

For context, Selling's decomposition of the matrix $D = \mathrm{Id}$ yields up to permutation the canonical basis $(e_1, \cdots, e_d)$ with unit weights $\sigma_1 = 1, \cdots, \sigma_d = 1$, whereas the remaining weights are zero $\sigma_i = 0$, $d < i \leq I$ (and the vectors e_i, $d < i \leq I$ are not uniquely determined). As a result ∇_{Id}^h and ∇_{Id}^h are the classical finite differences discretizations of the gradient and laplacian, whose stencil only involves the immediate grid neighbors.

Definition 1.4 (Canonical discretization)**.** We say that $(h\omega, D) \in \mathbb{R}^d \times S_d^{++}$, $d \in \{2,3\}$, is *canonically feasible* if the following operator L^h is DDE

$$-L^h u(x) := \langle D^{-1}\omega, \nabla_D^h u(x) \rangle + \Delta_D^h u(x). \tag{1.7}$$

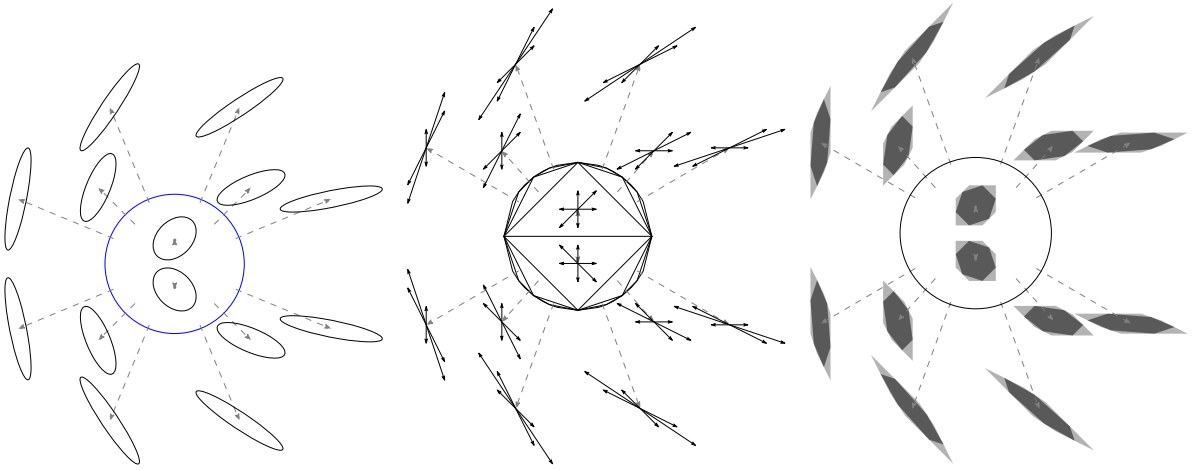

Figure 1: To each point (a, b) of the unit disk, we associate the matrix $D = D(a, b)$ defined by (1.5). *Left:* Ellipse defined by $\{\langle v, D^{-1}v\rangle \leq 1;\, v \in \mathbb{R}^2\}$. Points close to the unit disk boundary (shown blue) yield strongly anisotropic ellipses. *Center:* Support $(e_i)_{i=1}^I$ of Selling's decomposition, which is also the stencil of our finite difference scheme for the given anisotropy. *Right:* Set of vectors ω for which the pair (ω, D) is canonically feasible (dark gray), or absolutely feasible (dark and light gray), computed via Proposition 3.10. The scale of the three figures may not match.

Equivalently, but more explicitly, $L^h = L^h[\rho_i^h, e_i]_{1\leq|i|\leq I}$, involves the same offsets as Selling's decomposition $D = \sum_{1\leq i\leq I} \sigma e_i e_i^T$, with the usual convention $e_{-i} := -e_i$. The weights are obtained as

$$\rho_i^h := \sigma_i \,(1 + \tfrac{h}{2}\langle \omega, D^{-1}e_i\rangle). \tag{1.8}$$

Definition 1.4 outlines a simple and practical discretization of anisotropic linear PDE operators, often referred to as *our numerical scheme* in the paper. By construction, canonical feasibility implies absolute feasibility, but the latter can be achieved in a variety of other ways, using possibly a different number of terms I, a different support $(e_i)_{i=1}^I$, or different weights $(\rho_i^h)_{1\leq|i|\leq I}$. Note also that (1.7) is second order consistent with the PDE operator (1.1), in view of (1.6), whereas the conditions of Definition 1.2 only imply first order consistency. We next state the main result of this paper.

Theorem 1.5. *Let $(h\omega, D) \in \mathbb{R}^d \times S_d^{++}$, where $d \in \{2, 3\}$. If $(h\omega, D)$ is absolutely feasible, then $(c_d h\omega, D)$ is canonically feasible, with $c_2 := 1/2$ and $c_3 := 1/6$.*

Taking the contraposition, Theorem 1.5 shows that if the canonical discretization of Definition 1.4 does not yield a DDE scheme in some practical instance, then (up to the factor c_d) the grid scale is too coarse and there is no hope of obtaining a DDE and second order consistent finite differences scheme by any other means. The following result in contrast provides a direct criterion for canonical feasibility. We denote by $\|\cdot\|$ and $\langle\cdot,\cdot\rangle$ the Euclidean norm and scalar product. Let also $\|e\|_M := \sqrt{\langle e, Me\rangle}$ and $\|A\| := \max_{\|x\|\leq 1} \|Ax\|$ for any $e \in \mathbb{R}^d$, $M \in S_d^{++}$, and matrix A.

Theorem 1.6. *Let $(h\omega, D) \in \mathbb{R}^d \times S_d^{++}$, where $d \in \{2, 3\}$, and let $M := D^{-1}$. If one has*

$$h\|M\|^{\frac{1}{2}}\|\omega\|_M \leq c_d, \tag{1.9}$$

then $(h\omega, D)$ is canonically feasible, with $c_2 := 1$ and $c_3 := 1/(2\sqrt{3})$.

The existence of a finite differences discretization, degenerate elliptic and second order consistent, is not the only practical concern: the width of the stencil used is also of importance. Excessively wide stencils reduce the effective discretization scale of the scheme, thus also the accuracy of the numerical results. They may also raise difficulties with the treatment of boundary conditions, computer parallelization, matrix conditioning and sparsity, etc. We provide two results related to the stencil width. First, we show that the canonical discretization has the smallest support of all possible DDE and second order consistent discretizations, in dimension two, in the strong sense of convex hull inclusion. We denote by $\mathrm{Hull}(E)$ the convex hull of a subset E of a vector space.

Theorem 1.7. *Let $(h\omega, D) \in \mathbb{R}^2 \times S_2^{++}$ be canonically feasible, and let $(\rho_i^h, e_i)_{1\le|i|\le I}$ be the corresponding discretization, pruned so that $\rho_i^h \neq 0$ or $\rho_{-i}^h \neq 0$ for all $1 \le i \le I$. Let $(\rho_i'^h, e_i')_{1\le|i|\le I'}$ be another discretization, obeying* (1.3)*. Then*

$$\mathrm{Hull}\{e_i;\, 1 \le |i| \le I\} \subseteq \mathrm{Hull}\{e_i';\, 1 \le |i| \le I'\}.$$

Second, we provide explicit bounds on the stencil width in terms of the differential operator coefficients and anisotropy.

Theorem 1.8. *Let $(h\omega, D) \in \mathbb{R}^d \times S_d^{++}$ be canonically feasible, where $d \in \{2, 3\}$, and let $(\rho_i^h, e_i)_{1\le|i|\le I}$ be the corresponding discretization. Then $\|e_i\|_M \le C_d\sqrt{\|M\|}$ for all $1 \le i \le d$, where $M := D^{-1}$, and $C_2 = 2$ and $C_3 = 4\sqrt{3}$.*

Theorem 1.8 implies in particular that $\|e_i\| \le C_d\,\mathrm{Cond}(D)$, for all $1 \le i \le I$, where $\mathrm{Cond}(D) := \sqrt{\|D\|\|D^{-1}\|}$. See also [Mir16] for average case bounds in dimension $d = 2$, under random rotations $R_\theta^{\mathrm{T}} D R_\theta$ of the tensor, $\theta \in [0, 2\pi]$.

Outline

Section §2 is devoted to further discussion of the canonical discretization, and to the proofs of Theorems 1.6, 1.7 and 1.8 which follow rather directly from arguments presented in [Mir17] and [Mir16]. Section §3 establishes Theorem 1.5. Numerical experiments are presented in §4.

2 The canonical discretization

This section is devoted to a further presentation of the construction of Definition 1.4, here referred to as the *canonical discretization* of a second order linear PDE operator. We review Selling's algorithm in §2.1, finalizing the algorithmic description of our numerical scheme. We describe in §2.2 an interpretation of this algorithm as an optimization procedure, involving objects from the field of lattice geometry known as Voronoi's first reduction and Ryskov's polyhedron. Theorems 1.6, 1.7 and 1.8 are proved in §2.3.

The results presented §2.3 are new, whereas the more classical techniques described in §2.1 and §2.2 are required for completeness and as a preliminary to the proof of Theorem 1.5 in §3.

2.1 Selling's algorithm and formula

We describe Selling's algorithm [Sel74, CS92], and the related tensor decomposition formula which is invoked in Definition 1.4 of the numerical scheme considered in this paper.

Selling's algorithm

This algorithm belongs to the field of lattice geometry [NS04], which among other things studies coordinate systems in additive lattices (here $\mathbb{Z}^d$), adapted to the geometry defined by a given positive definite quadratic form (here defined by $D \in S_d^{++}$). The next definition introduces such a concept.

Definition 2.1. A superbase of $\mathbb{Z}^d$ is a $(d+1)$-tuple $b = (v_0, \cdots, v_d) \in (\mathbb{Z}^d)^{d+1}$ such that $|\det(v_1, \cdots, v_d)| = 0$ and $v_0 + \cdots + v_d = 1$. It is said D-obtuse, where $D \in S_d^{++}$, if $\langle v_i, Dv_j \rangle \leq 0$ for all $0 \leq i < j \leq d$.

Given a positive definite tensor $D \in S_d^{++}$, where $d \in \{2, 3\}$, Selling's algorithm constructs a D-obtuse superbase, see Algorithm 1. Note that the algorithm does not extend to dimension $d \geq 4$, and indeed there exists a matrix $D \in S_4^{++}$ for which no D-obtuse superbase exists [Sch09a].

Algorithm 1 Selling's algorithm

Input: A positive definite tensor $D \in S_d^{++}$, and a superbase $b = (v_0, \cdots, v_d)$, where $d \in \{2, 3\}$.
While there exists $0 \leq i < j \leq d$ such that $\langle v_i, Dv_j \rangle > 0$ **do**

If $d = 2$, $b \leftarrow (-v_i, v_j, v_i - v_j)$.
If $d = 3$, $b \leftarrow (-v_i, v_j, v_i + v_k, v_i + v_l)$ where $\{k, l\} = \{0, 1, 2, 3\} \setminus \{i, j\}$.

Output: b, which is now a D-obtuse superbase.

Proof of correctness and termination of Algorithm 1. Denote by b the current superbase at the beginning of an iteration. If the stopping criterion holds, then b is D-obtuse, as desired. Otherwise, denoting by b' the updated superbase, one easily checks that

$$\mathcal{E}_D(b') = \mathcal{E}_D(b) - C_d \langle v_i, Dv_j \rangle \qquad \text{where} \quad \mathcal{E}_D(b) := \sum_{0 \leq k \leq d} \|v_k\|_D^2, \tag{2.1}$$

and where $C_2 = 4$ and $C_3 = 2$. Thus $\mathcal{E}_D(b') < \mathcal{E}_D(b)$. Since there exists only finitely many superbases b such that $\mathcal{E}_D(b)$ is below a given constant, Selling's algorithm must terminate. □

Selling's algorithm is not the only means to produce a D-obtuse superbase. For instance Corollary 1 and Proposition 1 in [FM14] show in dimension $d \in \{2, 3\}$ how to produce a D-obtuse superbase from another type of system of coordinates referred to as D-reduced basis, resulting in a $\mathcal{O}(\ln(\|D\|\|D^{-1}\|))$ numerical complexity [NS04]. Selling's algorithm is however efficient enough for applications to PDE discretization, which usually involve moderate condition numbers, and therefore it is used in all our numerical experiments §4.

Selling's decomposition

This mathematical formula allows, once a D-obtuse superbase of $\mathbb{Z}^d$ is known, to decompose the tensor $D \in S_d^{++}$ in the form of (1.4). For that purpose, we associate to each superbase a family of vectors $(e_{ij})_{i \neq j}$ defined by duality relations.

Definition 2.2. Let $b = (v_0, \cdots, v_d)$ be a superbase of $\mathbb{Z}^d$. Then for any i, j in $\{0, \cdots, d\}$ such that $i \neq j$ we let $e_{ij} \in \mathbb{Z}^d$ be the unique vector obeying $\langle e_{ij}, v_k \rangle := \delta_{ik} - \delta_{jk}$, for all $0 \leq k \leq d$.

Note that Definition 2.2 defines $e_{ij} \in \mathbb{R}^d$ by $d+1$ linear relations. This does make sense in view of the redundancy $v_0 + \cdots + v_d = 0$ of the linear forms, and of the compatibility $(\delta_{i0} - \delta_{j0}) + \cdots + (\delta_{id} - \delta_{jd}) = 1 - 1 = 0$ of the right-hand sides. The vectors e_{ij} admit explicit expressions when $d \in \{2, 3\}$, namely (up to the sign)

$$e_{ij} = \pm v_k^\perp \text{ if } d = 2, \qquad (\text{resp. } e_{ij} = \pm v_k \times v_l \text{ if } d = 3), \tag{2.2}$$

where $\{i, j, k\} = \{0, 1, 2\}$ (resp. $\{i, j, k, l\} = \{0, 1, 2, 3\}$). For all $i, j, k, l \in \{0, \cdots, d\}$ such that $i \neq j$ and $k \neq l$ one also has the useful identity

$$\begin{aligned} \big\langle v_k, (e_{ij} e_{ij}^{\mathrm{T}}) v_l \big\rangle = \big\langle e_{ij}, (v_k \otimes v_l) e_{ij} \big\rangle &= \langle e_{ij}, v_k \rangle \langle e_{ij}, v_l \rangle \\ &= \begin{cases} -1 \text{ if } \{i, j\} = \{k, l\}, \\ 0 \text{ otherwise.} \end{cases} \end{aligned} \tag{2.3}$$

We denoted by $v \otimes w := \frac{1}{2}(vw^{\mathrm{T}} + wv^{\mathrm{T}}) \in S_d$ the symmetrized outer product of two vectors $v, w \in \mathbb{R}^d$. The next lemma shows how a superbase of $\mathbb{Z}^d$ defines a decomposition of an arbitrary tensor D, involving integer offsets. If the superbase is D-obtuse, then the weights are non-negative, and the decomposition is known as Selling's decomposition or *formula* [Sel74, CS92].

Lemma 2.3 (Selling's decomposition)**.** *Let $D \in S_d$, and let $b = (v_0, \cdots, v_d)$ be a superbase of $\mathbb{Z}^d$. Then*

$$D = - \sum_{0 \leq i < j \leq d} \langle v_i, D v_j \rangle e_{ij} e_{ij}^{\mathrm{T}}. \tag{2.4}$$

If $D \in S_d^{++}$ and b is D-obtuse, then (2.4) *is known as* Selling's decomposition *of D.*

Proof. Denote by D' the r.h.s. of (2.4). By (2.3) we obtain $\langle v_k, D v_l \rangle = \langle v_k, D' v_l \rangle$ for all $0 \leq k < l \leq d$. These $d(d+1)/2$ independent linear relations imply $D = D'$, as announced. □

We finally complete the description of our numerical scheme construction, see Definition 1.4. Given a positive definite tensor D, build a D-obtuse superbase using Selling's algorithm or another method. Then Selling's formula (2.4) yields the required tensor decomposition $D = \sum_{1 \leq i \leq I} \sigma_i e_i e_i^{\mathrm{T}}$ with $I = d(d+1)/2$, $\sigma_i \geq 0$, $e_i \in \mathbb{Z}^d \setminus \{0\}$. We emphasize that one cannot replace Selling's formula with another tensor decomposition in Definition 1.4, or Theorems 1.5, 1.6, 1.7 and 1.8 would fail. Finally, let us mention that Selling's decomposition is uniquely determined by the tensor D, and thus independent of the choice of D-obtuse superbase, see Remark 2.13.

2.2 Ryskov's polyhedron and Voronoi's first reduction

We introduce two concepts from lattice geometry, Ryskov's polyhedron and Voronoi's first reduction [Sch09a], allowing us to rephrase Selling's algorithm as a a simplex-like optimization method solving a linear program. In order to prevent any confusion, let us insist that these geometric tools are *not* connected with the classical concept of Voronoi diagram, which is instead related with Voronoi's *second* reduction [Sch09a]. Ryskov's polyhedron is an unbounded subset $\mathcal{M}_d \subseteq S_d$, defined as follows[1]

$$\mathcal{M}_d := \{M \in S_d;\, \forall e \in \mathbb{Z}^d \setminus \{0\},\, \langle e, Me \rangle \geq 1\}. \tag{2.5}$$

[1]Depending on the author, Ryskov's polyhedron (2.5) is defined via the constraints $\langle e, Me \rangle \geq \lambda$, $e \in \mathbb{Z}^d \setminus \{0\}$, where λ is one, two, or an unspecified positive constant [Sch09b]. These definitions are equivalent up to an homothety of $\mathcal{M}_d$.

Remark 2.4 (Identification of duplicate constraints). The constraints associated in (2.5) with a vector $e \in \mathbb{Z}^d \setminus \{0\}$ and with its opposite $-e$ are obviously equivalent. We regard them as a single constraint, associated with the equivalence class $\pm e$.

The main result proved in this subsection is the classification of the edges and vertices of the polyhedron $\mathcal{M}_d$ in dimension $d \in \{2, 3\}$, see Corollary 2.11. These objects are actually known in all dimensions $d \leq 8$ see [Sch09b, CS88, SSV07], hence the results presented in this subsection are not new. The proof is recalled for completeness and because its arguments are adapted in §3.1 for the proof of Theorem 1.5.

Regularity of Ryskov's polyhedron

We refer to Appendix B for some general terminology on polyhedra and linear programming. Recall that, by Minkowski's convex body theorem [Sch09a], any centrally symmetric convex body $K \subseteq \mathbb{R}^d$ of volume $\mathrm{Vol}(K) > 2^d$ contains a point of $\mathbb{Z}^d \setminus \{0\}$.

Lemma 2.5. *Each $M \in \mathcal{M}_d$ is positive definite, and $\det(M) \geq c_d$ where $c_d > 0$ is a constant.*

Proof. If $M \in \mathcal{M}_d$, then by construction M is positive semi-definite and the set $K = \{x \in \mathbb{R}^d; \langle x, Mx\rangle < 1\}$ contains no point of $\mathbb{Z}^d \setminus \{0\}$. By Minkowski's convex body theorem one has $2^d \geq \mathrm{Vol}(K) = \mathrm{Vol}(B) \det(M)^{-\frac{1}{2}}$, where B denotes the Euclidean unit ball, as announced. The announced result thus holds with (sub-optimal) constant $c_d := \mathrm{Vol}(B)^2/2^{2d}$. □

The optimal constant in Lemma 2.5 is $c_d = \gamma_d^{-d}$, where γ_d is known as Hermite's constant [Sch09a].

Corollary 2.6. *The polyhedron $\mathcal{M}_d$ is regular in the sense of Definition B.1.*

Proof. Let us check the three points of this definition. (i) The set $\mathcal{M}_d$ contains all $M \in S_d$ such that $M \succeq \mathrm{Id}$, hence it has non-empty interior, as required. (ii) The defining constraints obey $\mathrm{Span}\{ee^{\mathrm{T}}; e \in \mathbb{Z}^d \setminus \{0\}\} = \mathbb{S}_d$, as required. (iii) For any $M, M' \in S_d$ and any $e \in \mathbb{R}^d$ one has $\langle e, M'e\rangle \geq (\lambda_{\min}(M) - \|M - M'\|)\|e\|^2$, where $\lambda_{\min}(M) > 0$ denotes the smallest eigenvalue. Given $M \in S_d^{++}$, one thus has $\langle e, M'e\rangle > 1$ whenever $\|M' - M\| < \lambda_{\min}(M)/2$ and $\|e\| \geq 2/\lambda_{\min}(M)$. This shows that only finitely many constraints defining the polyhedron $\mathcal{M}_d$ active in the neighborhood of any $M \in \mathcal{M}_d$, as required. □

Vertices and edges of Ryskov's polyhedron

We describe a family of vertices of $\mathcal{M}_d$ in Lemma 2.8, the corresponding edges in Lemma 2.10, $d \in \{2, 3\}$, and show in Corollary 2.11 that this exhausts the skeleton of $\mathcal{M}_d$.

Definition 2.7. To each superbase $b = (v_0, \cdots, v_d)$ of $\mathbb{Z}^d$ one associates the matrix

$$M_b = \frac{1}{2} \sum_{0 \leq i \leq d} v_i v_i^{\mathrm{T}}. \tag{2.6}$$

Lemma 2.8. *Let $b = (v_0, \cdots, v_d)$ be a superbase of $\mathbb{Z}^d$. Then $\langle e, M_b e\rangle \geq 1$ for all $e \in \mathbb{Z}^d \setminus \{0\}$, with equality iff $e = e_{ij}$ for some $i, j \in \{0, \cdots, d\}$, $i \neq j$, see Definition 2.2.*

Proof. Let $e \in \mathbb{Z}^d \setminus \{0\}$ and $S := 2\langle e, M_b e\rangle = \sum_{0 \leq i \leq d} \langle v_i, e\rangle^2$. Then S is the sum of the squares of the integers $\langle v_i, e\rangle$, $0 \leq i \leq d$, which are not all zero, and obey $\sum_{0 \leq i \leq d} \langle v_i, e\rangle = \langle 0, e\rangle = 0$. Thus $S \geq 2$, with equality iff there exists $i \neq j$ such that $\langle v_i, e\rangle = 1$, $\langle v_j, e\rangle = -1$, and $\langle v_k, e\rangle = 0$ for all $k \notin \{i, j\}$. In other words $e = e_{ij}$, as announced. □

By Lemma 2.8, one has $M_b \in \mathcal{M}_d$ for any superbase b. Furthermore, M_b saturates the $d(d+1)/2 = \dim(S_d)$ linearly independent constraints associated with the vectors $\pm e_{ij}$, where $0 \leq i < j \leq d$, and satisfies strictly the constraints associated with any other $e \in \mathbb{Z}^d \setminus \{0\}$. This shows that M_b is a non-degenerate vertex of the polyhedron $\mathcal{M}_d$. The edges emanating from this vertex, in dimension $d \in \{2, 3\}$, are described in Lemma 2.10 below.

We introduce in the next definition an adjacency relation on the set of superbases of $\mathbb{Z}^d$, which is reminiscent of the superbase updates involved in Selling's algorithm, Algorithm 1. This similarity is not by accident, and it leads to an interpretation of Selling's algorithm as a linear program solver, see Proposition 2.12.

Definition 2.9. One defines the following adjacency relations for superbases of $\mathbb{Z}^d$, $d \in \{2, 3\}$,

$$\begin{aligned} (v_0, v_1, v_2) &\leftrightarrow (-v_0, v_1, v_0 - v_1), \\ (v_0, v_1, v_2, v_3) &\leftrightarrow (-v_0, v_1, v_2 + v_0, v_3 + v_0), \end{aligned} \tag{2.7}$$

and likewise up to a permutation and/or a global change of sign of the superbase.

Lemma 2.10. *Let $d \in \{2, 3\}$ and let b be a superbase of $\mathbb{Z}^d$. The edges of $\mathcal{M}_d$ containing M_b coincide with the segments $[M_b, M_{b'}]$, where b' is a superbase of $\mathbb{Z}^d$ adjacent to b.*

Proof. Recall that M_b is a non-degenerate vertex of $\mathcal{M}_d$. Therefore there exists $d(d+1)/2 = \dim(S_d)$ edges of $\mathcal{M}_d$ containing M_b, which are obtained by relaxing one of the constraints active at M_b (see also §B.3 on this topic). In other words, the edges of $\mathcal{M}_d$ containing M_b can be parametrized by $0 \leq \alpha < \beta \leq d$ and obtained as

$$E_{\alpha\beta} = \{M \in \mathcal{M}_d;\, \langle e_{ij}, M e_{ij} \rangle = 1,\, 0 \leq i < j \leq d,\, (i, j) \neq (\alpha, \beta)\}.$$

Let b, b' be superbases of $\mathbb{Z}^d$ as in (2.7). Then distinguishing dimensions we compute

$$(d = 2): \quad 2(M_{b'} - M_b) = (v_0 - v_1)(v_0 - v_1)^{\mathrm{T}} - v_2 v_2^{\mathrm{T}} = -4 v_0 \otimes v_1. \tag{2.8}$$

$$\begin{aligned} (d = 3): \quad & 2(M_{b'} - M_b) \\ &= (v_2 + v_0)(v_2 + v_0)^{\mathrm{T}} + (v_3 + v_0)(v_3 + v_0)^{\mathrm{T}} - v_2 v_2^{\mathrm{T}} - v_3 v_3^{\mathrm{T}} \\ &= 2 v_0 \otimes (v_0 + v_2 + v_3) = -2 v_0 \otimes v_1. \end{aligned} \tag{2.9}$$

The symmetrized outer product $\otimes$ was introduced in (2.3). Thus elements M in $[M_b, M_{b'}]$ obey the constraints $\langle e_{ij}, M e_{ij} \rangle = 1$ whenever $\{i, j\} \neq \{0, 1\}$, by (2.3). Therefore $[M_b, M_{b'}] \subseteq E_{01}$, and equality holds since M_b and $M_{b'}$ are vertices of $\mathcal{M}_d$ and $E_{01} \subseteq \partial \mathcal{M}_d$. Likewise, by permuting the indices, we obtain for all $0 \leq i < j \leq d$ an edge of $\mathcal{M}_d$ of the form $[M_b, M_{b'}]$ where b' is adjacent to b, obeying all the constraints active at the non-degenerate vertex M_b but the one associated with $\pm e_{ij}$ (previously $\pm e_{01}$). □

Corollary 2.11. *The vertices (resp. bounded edges) of Ryskov's polyhedron $\mathcal{M}_d$, where $d \in \{2, 3\}$, take the form M_b where b is a superbase of $\mathbb{Z}^d$ (resp. $[M_b, M_{b'}]$ where b' is an adjacent superbase). There are no unbounded edges in $\partial \mathcal{M}_d$.*

Proof. The result follows from Lemma 2.10, and the fact that the graph defined by the vertices and edges of a regular polyhedron is connected. □

Voronoi's first reduction

Voronoi's first reduction $\mathrm{Vor}(D)$, of a positive definite quadratic form $D \in S_d^{++}$, is defined as a linear minimization problem over Ryskov's polyhedron

$$\mathrm{Vor}(D) := \inf_{M \in \mathcal{M}_d} \mathrm{Tr}(DM). \tag{2.10}$$

This linear program, in dimension $d(d+1)/2$ and subject to infinitely many constraints, is well posed as shown by Voronoi himself [Vor08, Sch09a], in the sense that the collection of minimizers is non-empty and compact (generically it is a point) for any $D \in S_d^{++}$. The next proposition reproves this fact in dimension $d \in \{2, 3\}$.

Proposition 2.12. *Let $D \in S_d^{++}$, where $d \in \{2, 3\}$. Then Voronoi's first reduction is a well posed linear program, attaining its minimum at vertices M_b of $\mathcal{M}_d$ associated with a D-obtuse superbase b.*

Proof. By lemma 2.10, Selling's algorithm defines a walk on the graph defined by the vertices and edges of Ryskov's polyhedron. Observing that $\mathcal{E}_D(b) = 2\,\mathrm{Tr}(DM_b)$, see (2.1) and (2.6), we see that the next vertex selection reduces the linear program's objective function, whenever that is possible. Compare also (2.1, left) with (2.8) and (2.9). Since Selling's algorithm terminates, it solves the linear program (2.10), by the general results in §B.2. Furthermore, by Definition 2.1, it terminates precisely when reaching a D-obtuse superbase, which concludes the proof. □

Note that the proof of the previous proposition outlines a close relationship between Selling's algorithm and the simplex algorithm [BG15] applied to the linear program (2.10).

Remark 2.13 (Uniqueness of Selling's decomposition). Consider the decomposition (2.4) of a tensor $D \in S_d^{++}$, associated with a D-obtuse superbase b (if any exists, which is only guaranteed in dimension $d \leq 3$). By Lemma 2.8, it can be rephrased as a set of KKT relations for the linear program (2.10) at $M_b \in \mathcal{M}_d$, see Definition B.5. Since M_b is a non-degenerate vertex of $\mathcal{M}_d$, the coefficients of this KKT relation are uniquely determined, even if there is no uniqueness of the D-obtuse superbase, see Proposition B.6. In contrast, Voronoi's reduction (2.10) in dimension $d \geq 4$, or our variant $\widetilde{\mathrm{Vor}}(\omega, D)$ introduced §3.2 in dimension $d \geq 2$, involve polyhedra with degenerate vertices, at which the KKT relations are often non-uniquely determined.

2.3 Proof of Theorems 1.6, 1.7 and 1.8

Theorems 1.6 and 1.8, announced in the introduction, provide respectively a criterion for the existence of our discretization, and an estimate of the size of its support. They both follow from the next lemma, which bounds the norm of the vectors defined dually from an obtuse superbase.

Lemma 2.14 (Corollary 4.12 in [Mir17])**.** *Let $D \in S_d^{++}$ where $d \in \{2, 3\}$. Let b be a D-obtuse superbase, and let $e = e_{ij}$, for some $i, j \in \{0, \cdots, d\}$ such that $i \neq j$, see Definition 2.2. Then, denoting $C_2 := 2$ and $C_3 := 4\sqrt{3}$, one has*

$$\|e\|_M \leq C_d \|M\|^{\frac{1}{2}}, \qquad \text{where} \quad M := D^{-1}. \tag{2.11}$$

We refer to [Mir17] for the proof of Lemma 2.14, and use this result here to establish Theorems 1.6 and 1.8.

Proof of Theorem 1.8. Recall that the numerical scheme construction in Definition 1.4 relies on the decomposition of a tensor $D \in S_d^{++}$ via Lemma 2.3, and in particular it involves the offsets e_{ij}, $0 \leq i < j \leq d$ associated with a D-obtuse superbase. The announced estimate thus directly follows from Lemma 2.14. □

Proof of Theorem 1.6. Same notations as in the proof of Theorem 1.8. Denoting $M := D^{-1}$ and $e = e_{ij}$ for some $0 \leq i < j \leq d$, one has by Lemma 2.14

$$|\langle \omega, D^{-1} e \rangle| = |\langle \omega, Me \rangle| \leq \|\omega\|_M \|e\|_M \leq C_d \|M\|^{\frac{1}{2}} \|\omega\|_M.$$

Condition (1.9) thus implies that the weights (1.8) are non-negative, which as announced proves the absolute feasibility of $(h\omega, D)$. □

The rest of this section is devoted to the proof of Theorem 1.7. For that purpose, we need to introduce the geometrical concept of Voronoi vector [Sch09a].

Definition 2.15. A point $e \in \mathbb{Z}^d \setminus \{0\}$ is an M-Voronoi vector, where $M \in S_d^{++}$, if there exists $p \in \mathbb{R}^d$ (referred to as the witness) such that

$$\|p - 0\|_M = \|p - e\|_M \leq \|p - x\|_M, \text{ for all } x \in \mathbb{Z}^d. \tag{2.12}$$

One says that e is a *strict* M-Voronoi vector if the above inequality is strict for all $x \notin \{0, e\}$.

The origin 0 is introduced in (2.12, left) to emphasize the geometrical interpretation. In the language of Voronoi diagrams, e is a an M-Voronoi vector iff the Voronoi cells of 0 and e intersect, in the diagram of $\mathbb{R}^d$ associated with the sites $\mathbb{Z}^d$ and metric $\|\cdot\|_M$. The (strict) M-Voronoi vectors can be determined from an M-obtuse superbase, as shown by the next lemma in dimension $d = 2$. See Theorem 3 in [CS92] for a related argument in arbitrary dimension.

Lemma 2.16. *Let $M \in S_2^{++}$ and let e_0, e_1, e_2 be an M-obtuse superbase. Then $\pm e_0, \pm e_1, \pm e_2$ are M-Voronoi vectors. Furthermore e_0 is a strict M-Voronoi vector iff $\langle e_1, Me_2 \rangle < 0$ (likewise for $-e_0$, and likewise permuting (e_0, e_1, e_2)).*

Proof. We first show, w.l.o.g., that e_0 is an M-Voronoi vector, whose witness is $p := e_0/2$. Note that $\|p - 0\|_M = \|p - e_0\|_M (= \|e_0/2\|_M)$ as required (2.12). Let $x \in \mathbb{Z}^2$ be arbitrary. Since $\det(e_1, e_2) = 1$, there exists $a, b \in \mathbb{Z}$ such that $x = ae_1 + be_2$. From this point a direct computation yields (2.12), as announced

$$\begin{aligned}
\|p - x\|_M^2 - \|p\|_M^2 &= \|(a + 1/2)e_1 + (b + 1/2)e_2\|_M^2 - \|(e_1 + e_2)/2\|_M^2 \\
&= (a^2 + a)\|e_1\|_M^2 + (b^2 + b)\|e_2\|_M^2 + (2ab + a + b)\langle e_1, Me_2 \rangle \\
&\geq \Big((a^2 + a) + (b^2 + b) - (2ab + a + b)\Big)(-\langle e_1, Me_2 \rangle) \\
&= -(a - b)^2 \langle e_1, Me_2 \rangle \ \geq 0.
\end{aligned}$$

In the third line we used $\|e_1\|_M^2 = \langle -e_0 - e_2, Me_1 \rangle \geq -\langle e_1 M e_2 \rangle$, and likewise $\|e_2\|_M^2 \geq -\langle e_1 M e_2 \rangle$. In the rest of the proof, we show that e_0 is a strict M-Voronoi vector, under the additional assumption that $\langle e_1, Me_2 \rangle < 0$. Indeed, if $\|p\| = \|p - x\|$, then $a = b$ by the above, thus $x = -ae_0$ and therefore $\|e_0/2\|_M = \|(a+1/2)e_0\|_M$. This implies $a \in \{0, -1\}$, hence $x \in \{0, e_0\}$, and therefore e_0 is a strict M-Voronoi vector, as announced. □

Lemma 2.17. *Let $D \in S_2^{++}$ and let (e_0, e_1, e_2) be a D-obtuse superbase. Let $M := D^{-1}$ and $(v_0, v_1, v_2) = (e_0^\perp, e_1^\perp, e_2^\perp)$. Then (v_0, v_1, v_2) is an M-obtuse superbase. In addition, for any $i \neq j$ one has $\langle e_i, De_j \rangle < 0$ iff $\langle v_i, Mv_j \rangle < 0$.*

Proof. By construction one has $v_0, v_1, v_2 \in \mathbb{Z}^2$, $v_0+v_1+v_2 = (e_0+e_1+e_2)^\perp = 0$, and $\det(v_1, v_2) = \det(e_1, e_2) = \pm 1$. Thus (v_0, v_1, v_2) is a superbase of $\mathbb{Z}^2$. On the other hand, the obtuseness properties come from the following identity: for any $e, e' \in \mathbb{R}^2$, $D \in S_2^{++}$ and $M := D^{-1}$ one has

$$\langle e^\perp, Me'^\perp \rangle = \det(M)\langle e, De' \rangle.$$

(In the special case $D = \mathrm{Id}$, this identity expresses that rotation by $\pi/2$ is an isometry. In the general case $D = A^T A$ for some $A \in \mathrm{GL}_2(\mathbb{R})$, it follows from a linear change of variables and the relation $(Ae)^\perp = \mathrm{cof}(A)e^\perp$ where $\mathrm{cof}(A)$ denotes the cofactor matrix.) □

We are ready to prove Theorem 1.7, by adapting a result of [Mir16], devoted to operators without a first order term, and stated in terms of Voronoi vectors.

Lemma 2.18 (Adapted from Theorem 1.3 in [Mir16])**.** *Let $D \in S_2^{++}$, and let $D = \sum_{1\le i\le I} \sigma_i e_i e_i^{\mathrm{T}}$ be the decomposition associated with a D-obtuse superbase by Lemma 2.3, pruned so that $\sigma_i \neq 0$ for all $1 \le i \le I$. Let also $D = \sum_{1\le i\le I'} \sigma'_i e'_i e'^{\mathrm{T}}_i$ be another decomposition, with $I' > 0$, $\sigma'_i \ge 0$, $e'_i \in \mathbb{Z}^2 \setminus \{0\}$ for all $1 \le i \le I'$. Then*

$$\mathrm{Hull}\{\pm e_i; 1 \le i \le I\} \subseteq \mathrm{Hull}\{\pm e'_i; 1 \le i \le I'\}.$$

Proof. Theorem 1.3 in [Mir16] provides a similar statement, except that the vectors $(\pm e_i)_{1\le i\le I}$ are defined as the strict M-Voronoi vectors, where $M = D^{-1}$. By Lemmas 2.16 and 2.17, the tensor decomposition here considered (2.4) is also supported on the set of strict M-Voronoi vectors, and the result follows. □

Proof of Theorem 1.7. We use the notations of Theorem 1.7, and define $\sigma_i := \rho_i^h + \rho_{-i}^h$ for all $1 \le i \le I$, and $\sigma'_i := \rho'^h_i + \rho'^h_{-i}$ for all $1 \le i \le I'$. Note that $\sigma_i > 0$ since $\rho_i^h \neq 0$ or $\rho_{-i}^h \neq 0$, and both are non-negative, for all $1 \le i \le I$. Then $D = \sum_{1\le i\le I} \sigma_i e_i e_i^{\mathrm{T}} = \sum_{1\le i\le I} \sigma'_i e'_i e'^{\mathrm{T}}_i$, and by Definition 1.4 the first decomposition comes from a D-obtuse superbase as in Lemma 2.3. Applying Lemma 2.18, and recalling that $e_{-i} := -e_i$, we conclude the proof of Theorem 1.7. □

3 Proof of Theorem 1.5

We establish in this section our main result, Theorem 1.5, on a compatibility relation needed for constructing our numerical scheme. This condition relates the grid scale h (safely ignored in this section, see Remark 3.1 below), with the first order term $\omega \in \mathbb{R}^d$ and the second order term $D \in S_d^{++}$ of the discretized linear differential operator. More precisely, this result states that if (ω, D) is absolutely feasible (some discretization exists), then $(c_d\omega, D)$ is canonically feasible (our discretization exists), where $d \in \{2, 3\}$ and $c_d \in]0, 1]$ is a constant.

The guiding principle of the proof is to adapt to the pair $(\omega, D) \in \mathbb{R}^d \times \mathbb{S}_d^{++}$, of a vector and a symmetric positive definite matrix, the tools and techniques presented in §2.1 and §2.2, which originally apply to a matrix $D \in \mathbb{S}_d^{++}$ alone. The arguments are split into three parts, and proceed as follows. We define and describe in §3.1 a variant $\widetilde{\mathcal{M}}_d \subseteq \mathbb{R}^d \times \mathbb{S}_d$ of Ryskov's polyhedron $\mathcal{M}_d \subseteq \mathbb{S}_d$, see (2.5), involving an asymmetric perturbation of the constraints. The corresponding generalization $\widetilde{\mathrm{Vor}}(\omega, D)$ of Voronoi's first reduction $\widetilde{\mathrm{Vor}}(D)$, see (2.10), is discussed in §3.2. We conclude the proof of Theorem 1.5 in §3.3, by studying a low dimensional linear feasibility problem.

Remark 3.1 (ω vs $h\omega$). In this section, we speak of the canonical (resp. absolute) feasibility of a pair denoted $(\omega, D) \in S_d^{++} \times \mathbb{R}^d$, instead of $(h\omega, D)$ as in §1. Indeed, the presence of the

gridscale $h > 0$ is superflous in Definitions 1.2 and 1.4 or Theorems 1.5 - 1.8. It only appears for consistency with the Taylor expansions, and so as to outline the different homogeneity properties of the vector ω and the second order tensor D. These results do not depend on h and ω separately, but only on the product $h\omega$, which can be treated as a block and is here simply renamed ω.

3.1 A variant of Ryskov's polyhedron

We study of a variant of Ryskov's polyhedron (2.5). Denoted $\widetilde{\mathcal{M}}_d \subseteq \mathbb{R}^d \times S_d$, it is defined as follows

$$\widetilde{\mathcal{M}}_d := \{(\eta, M) \in \mathbb{R}^d \times S_d;\, \forall e \in \mathbb{Z}^d \setminus \{0\}, \langle \eta, e\rangle + \langle e, Me\rangle \geq 1\}. \tag{3.1}$$

This subsection is devoted to description of the vertices and edges of $\widetilde{\mathcal{M}}_d$, when $d \in \{2, 3\}$, see Theorem 3.2 below (no other result from this section is used in the following ones). Surprisingly enough, this structure is only barely richer than that of Ryskov's original polyhedron, see Corollary 2.11, despite the higher dimension.

The concepts of superbase b of $\mathbb{Z}^d$, the associated matrix $M_b \in S_d^{++}$, and the notion of adjacent superbases (b, b'), were introduced in Definitions 2.1, 2.7, and 2.9 respectively. Regular polyhedra and their edges are introduced in Definitions B.1 and B.2 of Appendix B.

Theorem 3.2. *Let $d \in \{2, 3\}$. Then $\widetilde{\mathcal{M}}_d$ is a regular polyhedron, with:*

(a) Vertices: $(0, M_b)$, for all superbases b of $\mathbb{Z}^d$.

(b) Bounded edges: $[(0, M_b), (0, M_{b'})]$, for all adjacent superbases b and b' of $\mathbb{Z}^d$.

(c) Unbounded edges: $\{(0, M_b) + \lambda(v_I, v_I v_I^{\mathrm{T}});\, \lambda \geq 0\}$, for all superbases b of $\mathbb{Z}^d$ and all $I \subsetneq \{0, \cdots, d\}$, $I \neq \emptyset$, where $b = (v_0, \cdots, v_d)$ and $v_I := \sum_{i \in I} v_i$.

The rest of this section is devoted to the proof of Theorem 3.2, following a line of arguments similar to the proof of Corollary 2.11. For commodity, we introduce a scalar product on $\mathbb{R}^d \times S_d$, as well as a family of elements $l_e \in \mathbb{R}^d \times S_d$, $e \in \mathbb{Z}^d \setminus \{0\}$, defined as follows:

$$\langle\!\langle (\eta, M), (\omega, D) \rangle\!\rangle := \langle \eta, \omega\rangle + \mathrm{Tr}(MD), \qquad\qquad l_e := (e, ee^{\mathrm{T}}). \tag{3.2}$$

By construction $\langle\!\langle l_e, (\eta, M) \rangle\!\rangle = \langle e, \eta\rangle + \langle e, Me\rangle$, which is convenient in view of (3.1). Observe that for any $\lambda_1, \cdots, \lambda_I, \mu_1, \cdots, \mu_I \in \mathbb{R}$ and $e_1, \cdots e_I \in \mathbb{Z}^d$, one has

$$\begin{aligned} &\sum_{1 \leq i \leq I} \frac{\lambda_i + \mu_i}{2} \Big(e_i, e_i e_i^{\mathrm{T}}\Big) + \frac{\lambda_i - \mu_i}{2} \Big(- e_i, (-e_i)(-e_i)^{\mathrm{T}}\Big) \\ &= \left(\sum_{1 \leq i \leq I} \mu_i e_i, \sum_{1 \leq i \leq I} \lambda_i e_i e_i^{\mathrm{T}} \right). \end{aligned} \tag{3.3}$$

Remark 3.3 (Erdahl's cone of quadratic functions). The set (3.1) is reminiscent of Erdahl's cone [Erd92, DSSV12], another inhomogeneous generalization of Voronoi's constructions, defined as follows:

$$\mathcal{E}_d := \{f \text{ quadratic function on } \mathbb{R}^d;\, \forall e \in \mathbb{Z}^d, f(e) \geq 0\}$$

Recall that a quadratic function on is a map of the form $x \in \mathbb{R}^d \mapsto \alpha + \langle \eta, x\rangle + \langle x, Mx\rangle$. Thus for any $f \in \mathcal{E}_d$, the normalized function $f/f(0)$ (assuming $f(0) \neq 0$) can be identified with an element of

$$\{(\eta, M) \in \mathbb{R}^d \times S_d;\, \forall e \in \mathbb{Z}^d \setminus \{0\},\, \langle \eta, e\rangle + \langle e, Me\rangle \geq -1\}. \tag{3.4}$$

Despite the apparent similarity between (3.4) and (3.1), the set $\widetilde{\mathcal{M}}_d$ only resembles Erdahl's cone superficially. The set $\widetilde{\mathcal{M}}_d$ is more closely related with Ryskov's original polyhedron $\mathcal{M}_d$, as shown by Theorem 3.2 and Corollary 2.11.

Lemma 3.4. *For all $(\eta, M) \in \widetilde{\mathcal{M}}_d$ one has $M \in \mathcal{M}_d$.*

Proof. One has $\langle e, Me\rangle = \frac{1}{2}(\langle \eta, e\rangle + \langle e, Me\rangle) + \frac{1}{2}(\langle \eta, -e\rangle + \langle -e, M(-e)\rangle) \geq 1$, $\forall e \in \mathbb{Z}^d \setminus \{0\}$. □

Lemma 3.5. *The polyhedron $\widetilde{\mathcal{M}}_d$ is regular, in the sense of Definition B.1.*

Proof. (i) Let $(\eta, M) \in \mathbb{R}^d \times S_d$ be such that $\|\eta\| \leq 1$ and $M \succeq 2\,\mathrm{Id}$. Then for any $e \in \mathbb{Z}^d \setminus \{0\}$ one has $\langle \eta, e\rangle + \langle e, Me\rangle \geq -\|e\| + 2\|e\|^2 \geq 1$ since $\|e\| \geq 1$. Thus $(\eta, M) \in \widetilde{\mathcal{M}}_d$, and therefore $\widetilde{\mathcal{M}}_d$ has a non-empty interior. (ii) Recalling that $\mathrm{Span}\{ee^{\mathrm{T}}; e \in \mathbb{Z}^d \setminus \{0\}\} = S_d$, see Lemma 2.3, and using (3.3) one obtains $\mathrm{Span}\{(e, ee^{\mathrm{T}}); e \in \mathbb{Z}^d \setminus \{0\}\} = \mathbb{R}^d \times \mathbb{S}_d$, as required. (iii) Let $(\eta, M) \in \widetilde{\mathcal{M}}_d$. Then $M \in \mathcal{M}_d$, by Lemma 3.4, and therefore M is a symmetric positive definite matrix whose smallest eigenvalue is denoted $\lambda_{\min}(M) > 0$. Then for any (η', M') such that $\|\eta - \eta'\| \leq 1$ and $\|M - M'\| \leq \lambda_{\min}(M)/2$ one has for all $e \in \mathbb{Z}^d \setminus \{0\}$

$$\begin{aligned}\langle \eta', e\rangle + \langle e, M'e\rangle &\geq -\|\eta'\|\|e\| + (\lambda_{\min}(M) - \|M - M'\|)\|e\|^2 \\ &\geq (\lambda_{\min}(M)\|e\|/2 - \|\eta\| - 1)\|e\| \geq 2,\end{aligned}$$

assuming $\|e\| \geq 2(\|\eta\| + 3)/\lambda_{\min}(M)$ for the last inequality. This shows that only finitely many of the constraints defining the polyhedron $\widetilde{\mathcal{M}}_d$ are active in the neighborhood of $(\eta, M) \in \widetilde{\mathcal{M}}_d$, as required. □

The next lemma describes a family of vertices of $\widetilde{\mathcal{M}}_d$.

Lemma 3.6. *For any vertex M of $\mathcal{M}_d$, the pair $(0, M)$ is a vertex of $\widetilde{\mathcal{M}}_d$. In addition, the active constraints at a vertex $M \in \mathcal{M}_d$, and at the corresponding vertex $(0, M) \in \widetilde{\mathcal{M}}_d$, are associated with the same vectors $e \in \mathbb{Z}^d \setminus \{0\}$.*

Proof. We first check that $(0, M) \in \widetilde{\mathcal{M}}_d$. Indeed, for any $e \in \mathbb{Z}^d \setminus \{0\}$, one has $\langle\!\langle l_e, (0, M)\rangle\!\rangle = \langle 0, e\rangle + \langle e, Me\rangle = \langle e, Me\rangle \geq 1$, since $M \in \mathcal{M}_d$.

We next prove that $(0, M)$ is a vertex of $\widetilde{\mathcal{M}}_d$, relying on the characterization of Remark B.3. By assumption, since M is a vertex of $\mathcal{M}_d$, there exists $e_1, \cdots, e_I$ in $\mathbb{Z}^d \setminus \{0\}$ such that $\langle e_i, Me_i\rangle = 1$ for all $1 \leq i \leq I$, and $\mathrm{Span}\{e_i e_i^{\mathrm{T}}\}_{1 \leq i \leq I} = S_d$. The latter property implies that $\{e_i\}_{i=1}^I$ spans $\mathbb{R}^d$, hence using (3.3) we obtain $\mathrm{Span}\{l_{e_i}\}_{1 \leq |i| \leq I} = \mathbb{R}^d \times \mathbb{S}^d$, with the usual convention $e_{-i} = -e_i$. Since $\langle\!\langle l_{\pm e_i}, (0, M)\rangle\!\rangle = \langle e_i, Me_i\rangle = 1$, we conclude that $(0, M)$ is a vertex of $\widetilde{\mathcal{M}}_d$. The additional point is straightforward, since the vectors $e \in \mathbb{Z}^d$ associated to active constraints at $(0, M) \in \widetilde{\mathcal{M}}_d$ are characterized by the identity $1 = \langle e, 0\rangle + \langle e, Me\rangle = \langle e, Me\rangle$. □

It the rest of this section, we compute the edges emanating from vertex $(0, M)$ in $\widetilde{\mathcal{M}}_d$. We apply the strategy of §B.3 to compute the outgoing direction of each edge, and eventually only encounter the following two cases:

(i) The computed edge direction has the form $\nu = (0, N)$ for some $N \in S_d$, hence the corresponding edge is internal to $\widetilde{\mathcal{M}}_d \cap (\{0\} \times S_d) = \{0\} \times \mathcal{M}_d$. Since the edges of $\mathcal{M}_d$ are known, see Corollary 2.11, this must be a bounded edge in the form of Theorem 3.2 (b).

(ii) The computed edge direction has the form $\nu = (v, vv^{\mathrm{T}})$, where $v \in \mathbb{Z}^d$ (more precisely, v has the form indicated in Theorem 3.2 (c)). Thus for any e in $\mathbb{Z}^d \setminus \{0\}$,

$$\langle\!\langle l_e, \nu \rangle\!\rangle = \langle\!\langle (e, ee^{\mathrm{T}}), (v, vv^{\mathrm{T}}) \rangle\!\rangle = \langle e, v\rangle + \mathrm{Tr}(ee^{\mathrm{T}}vv^{\mathrm{T}}) = \langle e, v\rangle + \langle e, v\rangle^2.$$

Since e and v have integer coordinates, the scalar product $\langle e, v\rangle$ is an integer, and therefore $\langle\!\langle l_e, \nu \rangle\!\rangle \geq 0$ (with equality iff $\langle e, v\rangle \in \{0, -1\}$). Thus ν yields an unbounded edge in $\widetilde{\mathcal{M}}_d$ starting from $(0, M)$, in the form of Theorem 3.2 (c).

The graph defined by the edges and vertices of a regular polyhedron is connected, see Appendix B. Once the above dichotomy is established, it follows that $\widetilde{\mathcal{M}}_d$ has no other vertices than those already found in Lemma 3.6, which concludes the proof of Theorem 3.2.

Notation (i-ii) and (A-D) in §3.1.1 and 3.1.2.

We establish the above dichotomy (i-ii) in §3.1.1 and 3.1.2, in dimension two and three respectively. For that purpose, we rely on the algorithm presented in §B.3 for enumerating the outgoing edges from a vertex in a polyhedron, and explicitly refer to its steps (A-D).

3.1.1 Edges of $\widetilde{\mathcal{M}}_2$

Let $b = (v_0, v_1, v_2)$ be a superbase of $\mathbb{Z}^2$, and let M_b be the corresponding vertex of $\mathcal{M}_2$, see (2.6). By Lemma 2.8, the active constraints at the vertex $M_b \in \mathcal{M}_2$ correspond to the set of vectors $E := \{e_{ij};\ i, j \in \{0, 1, 2\},\ i \neq j\}$ associated with the superbase b, see Definition 2.2. By Lemma 3.6, $(0, M_b)$ is a vertex of the polyhedron $\widetilde{\mathcal{M}}_2$, at which the constraints associated with the same vectors $e \in E$ are active. Since the number $\#(E) = 6$ of active constraints at $(0, M_b) \in \widetilde{\mathcal{M}}_2$ exceeds the dimension $\dim(\mathbb{R}^2 \times S_2) = 2 + 3 = 5$ of the embedding vector space, the vertex is degenerate. The edges containing $(0, M_b) \in \widetilde{\mathcal{M}}_2$ are obtained by selecting 4 out of the six active constraints, in other words by removing two elements from the set E. The following cases can be distinguished:

- Removing e_{12} and e_{21}. The corresponding direction is $\nu = (0, v_1 \otimes v_2)$, which lies within $\{0\} \times S_d$, and thus falls in case (i). Validation of the direction: one has $\langle\!\langle l_{e_{01}}, \nu \rangle\!\rangle = \langle e_{01}, v_1 \otimes v_2\, e_{01}\rangle = \langle e_{01}, v_1\rangle\langle e_{01}, v_2\rangle = 0$, since $\langle e_{01}, v_2\rangle = 0$. Likewise $\langle\!\langle l_e, \nu \rangle\!\rangle = 0$ for all $e \in \{\pm e_{01}, \pm e_{02}\}$, hence ν obeys the conditions of (B) of Algorithm §B.3.

- Removing e_{01} and e_{02}. The corresponding direction is $\nu = (v_0, v_0 \otimes v_0)$, which falls in the case (ii) of an unbounded edge. Validation of the direction: one has $\langle\!\langle l_{e_{12}}, \nu \rangle\!\rangle = \langle e_{12}, v_0\rangle^2 + \langle e_{12}, v_0\rangle = 0^2 + 0 = 0$, and $\langle\!\langle l_{e_{10}}, \nu \rangle\!\rangle = \langle e_{10}, v_0\rangle^2 + \langle e_{10}, v_0\rangle = (-1)^2 + (-1) = 0$. Likewise for $e \in \{e_{21}, e_{20}\}$.

- Removing e_{01} and e_{20}. The corresponding direction is $\nu = (v_0, v_1 \otimes v_1 - v_2 \otimes v_2)$, but it does not correspond to an edge, since it is eliminated in step (C) of Algorithm §B.3. Indeed, noting that $\langle l_e, \nu\rangle = \langle e, v_1\rangle^2 - \langle e, v_2\rangle^2 + \langle e, v_0\rangle$ we obtain

$$\begin{aligned}
\langle l_{e_{10}}, \nu\rangle &= 1^2 - 0^2 - 1 = 0,\\
\langle l_{e_{02}}, \nu\rangle &= 0^2 - (-1)^2 + 1 = 0,\\
\langle l_{\pm e_{12}}, \nu\rangle &= (\pm 1)^2 - (\mp 1)^2 + 0,
\end{aligned}$$

showing that the direction ν is correct. However since

$$\langle l_{e_{01}}, \nu\rangle = (-1)^2 - 0^2 + 1 = 2, \qquad \langle l_{e_{20}}, \nu\rangle = 0^2 - 1^2 - 1 = -2,$$

have opposite signs, the direction ν does not yield an edge of positive length.

There are 15 distinct two element subsets of $E := \{e_{ij};\, i,j \in \{0,1,2\},\, i \neq j\}$, and we have considered just three. However by permuting indices, the above considered cases respectively cover 3, 6, and again 6, distinct two element subsets E. Thus our enumeration is complete, and Theorem 3.2 is proved in dimension $d = 2$.

3.1.2 Edges of $\widetilde{\mathcal{M}}_3$

Let $b = (v_0, v_1, v_2, v_3)$ be a superbase of $\mathbb{Z}^3$, and let M_b be the corresponding vertex of $\mathcal{M}_3$, see (2.6). By Lemma 2.8, the active constraints at the vertex $M_b \in \mathcal{M}_3$ correspond to the set of vectors $E := \{e_{ij};\ i,j \in \{0,1,2,3\},\ i \neq j\}$ associated with the superbase b, see Definition 2.2. By Lemma 3.6, $(0, M_b)$ is a vertex of the polyhedron $\widetilde{\mathcal{M}}_3$, at which the constraints associated with the same vectors $e \in E$ are active. Since the number $\#(E) = 12$ of active constraints at $(0, M_b) \in \widetilde{\mathcal{M}}_2$ exceeds the dimension $\dim(\mathbb{R}^3 \times S_3) = 3 + 6 = 9$ of the embedding vector space, the vertex is degenerate. The edges containing $(0, M_b) \in \widetilde{\mathcal{M}}_3$ are obtained by selecting 8 out of the twelve active constraints, in other words by removing four elements from the set E. The following cases can be distinguished:

- Removing $\pm e_{01}$ and two other unspecified elements of E. If the subset is not rejected in step (B), then the corresponding direction is $\nu = (0, v_0 \otimes v_1)$, which lies within $\{0\} \times S_d$ and thus falls into case (i). Validation of the direction: one has $\langle\!\langle l_{e_{ij}}, \nu\rangle\!\rangle = \langle e_{ij}, v_0\rangle\langle e_{ij}, v_1\rangle = 0$ as soon as $\{i,j\} \neq \{0,1\}$, hence ν obeys the conditions of (B).

- Removing $\alpha_{01}e_{01}$, $\alpha_{02}e_{02}$, $\alpha_{03}e_{03}$ and another unspecified element of E, where $\alpha_{01}, \alpha_{02}, \alpha_{03} \in \{-1, 1\}$. The corresponding direction is, up to a global sign change,

  $$\nu = (-v_0,\ \alpha_{01}v_0 \otimes v_1 + \alpha_{02}v_0 \otimes v_2 + \alpha_{03}v_0 \otimes v_3).$$

  It is rejected in step (B) or (C), unless $\alpha_{01} = \alpha_{02} = \alpha_{03}$ in which case $\nu = (-\alpha_{01}v_0,\ v_0 \otimes v_0)$ (here with the correct sign) falls in case (ii) and defines an unbounded edge. (Note that $v_0 \otimes v_1 + v_0 \otimes v_2 + v_0 \otimes v_3 = -v_0 \otimes v_0$ since $v_1 + v_2 + v_3 = -v_0$.) Indeed, for any $i, j \in \{1,2,3\}$ such that $i \neq j$ one computes

  $$\langle l_{\pm e_{0i}}, \nu\rangle = -\alpha_{0i} \mp 1, \qquad \langle l_{e_{ij}}, \nu\rangle = 0. \tag{3.5}$$

- Removing $\alpha_{01}e_{01}$, $-\alpha_{12}e_{12}$, $\alpha_{23}e_{23}$, $-\alpha_{30}e_{30}$, where $\alpha_{01}, \alpha_{12}, \alpha_{23}, \alpha_{30}$ belong to $\{-1, 1\}$. Then the corresponding direction is, up to a global sign change,

  $$\nu = (v_1 + v_3,\ \alpha_{01}v_0 \otimes v_1 + \alpha_{12}v_1 \otimes v_2 + \alpha_{23}v_2 \otimes v_3 + \alpha_{30}v_3 \otimes v_0).$$

  It is rejected in step C, unless $\alpha_{01} = \alpha_{12} = \alpha_{23} = \alpha_{30}$, in which case $\nu = (v, v \otimes v)$ (here with the correct sign) with $v = -\alpha_{01}(v_1 + v_3) = \alpha_{01}(v_0 + v_2)$ falls in case (ii) and thus defines an unbounded edge. (Note that $-(v_1 + v_3) \otimes (v_1 + v_3) = (v_0 + v_2) \otimes (v_1 + v_3) = v_0 \otimes v_1 + v_1 \otimes v_2 + v_2 \otimes v_3 + v_3 \otimes v_0$ since $v_0 + v_2 = -(v_1 + v_3)$.) Indeed, we check that

  $$\langle l_{e_{02}}, \nu\rangle = 0 + 0 + 0 + 0 + 0, \quad \text{likewise for } e \in \{\pm e_{02}, \pm e_{13}\}$$
  $$\langle l_{\pm e_{01}}, \nu\rangle = -\alpha_{01} \pm 1, \quad \text{likewise for } e \in \{\pm e_{01}, \pm e_{12}, \pm e_{23}, \pm e_{30}\}.$$

Finally, we need to show that all the possible 4 element subsets $S \subseteq \{e_{ij};\, i \neq j\}$ correspond to one of the considered cases, up to a permutation of the superbase. We refer to $\{i, j\}$ as the indices of vector e_{ij}. If two elements of S share the same two indices, a.k.a. $e_{ij}, e_{ji} \in S$ for some $i \neq j$, then we fall in the first case. Otherwise, if (at least) three elements of S share one index, then we fall in the second case. Otherwise, each index $i \in \{0, \cdots, 3\}$ appears in at most two elements of S, thus exactly two since $\#(S) = 4 = \#\{0, \cdots, 3\}$. It follows that the indices of S define a cycle, and we fall in the last case.

3.2 A variant of Voronoi's first reduction

We introduce and study a variant of Voronoi's first reduction, applying to pairs (ω, D) of a vector $\omega \in \mathbb{R}^d$ and a positive definite symmetric tensor $D \in S_d^{++}$, instead of the matrix D alone in the original formulation (2.10). It is defined as follows:

$$\widetilde{\mathrm{Vor}}(\omega, D) := \inf\{\langle \omega, \eta\rangle + \mathrm{Tr}(DM);\, (\eta, M) \in \widetilde{\mathcal{M}}_d\}. \tag{3.6}$$

Somewhat surprisingly, our generalization of Voronoi's first reduction reduces to the original one, subject to a compatibility condition.

Theorem 3.7. *Let $d \leq 3$. For any $(D, \omega) \in S_d^{++} \times \mathbb{R}^d$ one has*

$$\widetilde{\mathrm{Vor}}(D, \omega) = \begin{cases} -\infty & \text{if } \exists v \in \mathbb{Z}^d \setminus \{0\},\ \langle v, Dv\rangle + \langle \omega, v\rangle < 0, \\ \mathrm{Vor}(D) & \text{otherwise.} \end{cases}$$

Proof. The result follows from the description of the vertices and unbounded edges of the polytope $\widetilde{\mathcal{M}_d}$ in Theorem 3.2, and from the general expression (B.2) of the value of a linear program. Note also that any $v_1 \in \mathbb{Z}^d \setminus \{0\}$ with co-prime coordinates can be completed into a basis $v_1, \cdots, v_d$ of $\mathbb{Z}$, hence also into a superbase with $v_0 := -(v_1 + \cdots + v_d)$. Hence the set of directions of all unbounded edges of $\widetilde{\mathcal{M}}_d$, see Theorem 3.2 (c), is $\mathbb{Z}^d \setminus \{0\}$. □

Proposition 3.8. *Let $d \leq 3$, and let $(\omega, D) \in \mathbb{R}^d \times S_d^{++}$. The following are equivalent:*

(i) The pair (ω, D) is absolutely feasible.

(ii) The linear program $\widetilde{\mathrm{Vor}}(\omega, D)$ is bounded.

In case (ii), any set of KKT relations for $\widetilde{\mathrm{Vor}}(\omega, D)$ yields a simultaneous decomposition of (ω, D), showing (i) explicitly.

Proof. Proof that (i) ⇒ (ii). Assume that (ω, D) is absolutely feasible, and denote by $\rho_i \geq 0$ the weights, and $e_i \in \mathbb{Z}^d$ the offsets of the corresponding decomposition, so that

$$\omega = \sum_{1 \leq i \leq I} \rho_i e_i, \qquad\qquad D = \sum_{1 \leq i \leq I} \rho_i e_i e_i^T. \tag{3.7}$$

Then for any $(M, \eta) \in \widetilde{\mathcal{M}}_d$, one obtains using the identity $\langle e, Me\rangle = \mathrm{Tr}(Mee^{\mathrm{T}})$

$$\langle \omega, \eta\rangle + \mathrm{Tr}(DM) = \sum_{1 \leq i \leq I} \rho_i \left(\langle \omega, e_i\rangle + \langle e_i, Me_i\rangle\right) \geq \sum_{1 \leq i \leq I} \rho_i \geq 0.$$

Therefore $\widetilde{\mathrm{Vor}}(D, \omega) \geq 0 > -\infty$ is bounded.

Proof that (ii) ⇒ (i). By Proposition 2.12 there exists a vertex M_b of $\mathcal{M}_d$, where b is a superbase of $\mathbb{Z}^d$, such that $\mathrm{Vor}(D) = \mathrm{Tr}(DM_b)$. By Lemmas 2.8 and 3.6, $(0, M_b)$ is a vertex of $\widetilde{\mathcal{M}}_d$ at which finitely many constraints $(e_i)_{i=1}^I$ are active. By Theorem 3.7, $\widetilde{\mathrm{Vor}}(D) = \mathrm{Vor}(D) = \mathrm{Tr}(DM_b) = \langle \omega, 0\rangle + \mathrm{Tr}(DM_b)$ and this minimum is attained at the vertex $(0, M_b)$. The KKT relations express that there exists non-negative weights $(\rho_i)_{i=1}^I$ (possibly non-unique) such that the objective function and the weighted sum of the constraints are equal: one has $\langle \omega, \eta\rangle + \mathrm{Tr}(DM) = \sum_{1\leq i\leq I} \rho_i(\langle e_i, \eta\rangle + \langle e_i, Me_i\rangle)$ for all $(\eta, M) \in \mathbb{R}^d \times S_d$. From this point, the simultaneous decomposition (3.7) holds by identification, as announced. □

Remark 3.9 (Degeneracy of the vertices of $\widetilde{\mathcal{M}}_d$). The vertices of $\widetilde{\mathcal{M}}_d$ are degenerate, in dimension $d \in \{2, 3\}$, in the sense that exactly $d(d+1)$ constraints are active, which is strictly greater than $\dim(\mathbb{R}^d \times S_d) = d(d+1)/2 + d$. As a result, the KKT relations for the linear program $\mathrm{Vor}(\omega, D)$ in general do not *uniquely* determine the decomposition (3.7) of the pair (ω, D). This is in constrast with Voronoi's first reduction in dimension $d \leq 3$, see Remark 2.13.

3.3 Local study of feasibility

In this section, we compare the conditions of canonical and absolute feasibility of a pair (ω, D), in dimension $d \leq 3$, concluding the proof of Theorem 1.5. For that purpose, we fix a symmetric positive definite matrix $D \in S_d^{++}$, denote by $b = (v_0, \cdots, v_d)$ a D-obtuse superbase, and recall Selling's decomposition (2.4)

$$D = \sum_{0\leq i<j\leq d} \sigma_{ij} e_{ij} e_{ij}^{\mathrm{T}}, \tag{3.8}$$

where $\sigma_{ij} := -\langle v_i, Dv_j\rangle \geq 0$ and where $e_{ij} \in \mathbb{Z}^d \setminus \{0\}$ for all $0 \leq i < j \leq d$ is introduced in Definition 2.2. In this subsection, for notational convenience, the indices i and j, are always implicitly constrained to lie in the set $\{0, \cdots, d\}$.

We characterize, in the next proposition, the canonical and absolute feasibility of a pair (ω, D) in terms of Selling's decomposition of D. The argument, in the case of absolute feasibility, heavily relies on the results established in §3.2.

Proposition 3.10. *Assume $d \leq 3$. Let $\omega \in \mathbb{R}^d$ and $D \in S_d^{++}$. We use the notations b, $(\sigma_{ij}, e_{ij})_{i<j}$ of Selling's decomposition* (3.8)*. Then*

- *(ω, D) is absolutely feasible iff there exists $\mu_{ij} \in [-1, 1]$, for all $0 \leq i < j \leq d$, such that $\omega = \sum_{i<j} \mu_{ij}\sigma_{ij}e_{ij}$*
- *(ω, D) is canonically feasible iff $|\langle e_{ij}, D^{-1}\omega\rangle| \leq 1$ for all $0 \leq i < j \leq d$ such that $\sigma_{ij} > 0$.*

Proof. First equivalence. If the pair (ω, D) is absolutely feasible, then by Proposition 3.8 the linear program $\widetilde{\mathrm{Vor}}(\omega, D)$ is bounded, and attains its minimum at the vertex $(0, M_b)$, at which the active constraints are associated with the vectors e_{ij}, $i \neq j$. By the KKT relations, there exists non-negative weights ρ_{ij}, $i \neq j$, such that

$$\omega = \sum_{i\neq j} \rho_{ij} e_{ij} \qquad\qquad D = \sum_{i\neq j} \rho_{ij} e_{ij} e_{ij}^T.$$

Recalling that $e_{ji} = -e_{ij}$ for all $i \neq j$, we obtain

$$\omega = \sum_{i<j} (\rho_{ij} - \rho_{ij}) e_{ij} \qquad\qquad D = \sum_{i<j} (\rho_{ij} + \rho_{ji}) e_{ij} e_{ij}^T.$$

By uniqueness of Selling's decomposition, see Remark 2.13, one has $\sigma_{ij} = \rho_{ij} + \rho_{ji}$ for all $i < j$. Denoting $\mu_{ij} := (\rho_{ij} - \rho_{ji})/\sigma_{ij} \in [-1, 1]$ when $\sigma_{ij} > 0$ (and e.g. $\mu_{ij} = 0$ if $\sigma_{ij} = 0$), we obtain $\omega = \sum_{i<j} \mu_{ij}\sigma_{ij}e_{ij}$ as announced. The reverse implication is trivial, by defining $\rho_{ij} = \sigma_{ij}(1 + \mu_{ij})/2$ and $\rho_{ji} = \sigma_{ij}(1 - \mu_{ij})/2$, for all $i < j$.

Second equivalence. By construction, see Definition 1.4, the pair (ω, D) obeys is canonically feasible iff $\sigma_{ij}(1 + \varepsilon\langle e_{ij}, \eta\rangle) \geq 0$ for all $i < j$ and all $\varepsilon \in \{-1, 1\}$, where $\eta := D^{-1}\omega$. This is indeed equivalent to $|\langle e_{ij}, D^{-1}\omega\rangle| \leq 1$, for all $i < j$ such that $\sigma_{ij} > 0$, as announced. □

We next state two technical lemmas which, combined with Proposition 3.10 above, let us conclude the proof of Theorem 1.5. The proof of Lemma 3.11 is postponed to §3.3.1.

Lemma 3.11. *Let $D \in S_d^{++}$, $d \leq 3$. We use the notations b, $(\sigma_{ij}, e_{ij})_{i<j}$ of Selling's decomposition* (3.8)*. Then $|\langle e_{ij}, Me_{kl}\rangle| \leq \|e_{ij}\|_M^2$ for all $i < j$ and all $k < l$, where $M := D^{-1}$.*

Lemma 3.12. *Let $D = \sum_{r=1}^R \sigma_r e_r e_r^T$, where $\sigma_r \geq 0$, $e_r \in \mathbb{R}^d$ for all $1 \leq r \leq R$, and R is a positive integer. If D is positive definite, then $\sigma_r\langle e_r, D^{-1}e_r\rangle \leq 1$ for all $1 \leq r \leq R$.*

Proof. For any $1 \leq r \leq R$, one has $D \succeq \sigma_r e_r e_r^{\mathrm{T}}$, in the sense of symmetric matrices. Therefore, letting $v_r := D^{-1}e_r$, we obtain $\langle e_r, D^{-1}e_r\rangle = \langle v_r, Dv_r\rangle \geq \sigma_r\langle v_r, e_r\rangle^2 = \sigma_r\langle e_r, D^{-1}e_r\rangle^2$. This implies $1 \geq \sigma_r\langle e_r, D^{-1}e_r\rangle$, as announced. □

Proof of Theorem 1.5. Assume that (D, ω) is absolutely feasible. Then for any $i < j$, using the notations of Proposition 3.10, we obtain

$$\begin{aligned}
|\langle e_{ij}, D^{-1}\omega\rangle| &\leq \sum_{k<l} \sigma_{kl}|\mu_{kl}|\,|\langle e_{ij}, D^{-1}e_{kl}\rangle| \\
&\leq \sum_{k<l} \sigma_{kl}\langle e_{kl}, D^{-1}e_{kl}\rangle \\
&\leq \sum_{k<l} 1 = d(d+1)/2.
\end{aligned}$$

The three inequalities follows, successively, from Proposition 3.10 (first point), Lemma 3.11, and Lemma 3.12. If follows from Proposition 3.10 (second point) that $(D, \omega/C)$ is canonically feasible, with $C := d(d+1)/2$, as announced. □

3.3.1 Proof of Lemma 3.11

Throughout this subsection, we use for convenience the notation $\langle v, w\rangle_M := \langle v, Mw\rangle$, for any $v, w \in \mathbb{R}^d$, $M \in S_d^{++}$. We use the notations of Lemma 3.11. In particular $D \in S_d^{++}$, $M := D^{-1}$, $b = (v_0, \cdots, v_d)$ is a D-obtuse superbase, and $(\sigma_{ij}, e_{ij})_{i<j}$ are the coefficients and offsets of Selling's decomposition (3.8) of D. As before, the indices i, j implicitly lie in $\{0, \cdots, d\}$.

Proof in dimension $d = 2$. Assume that the superbase $b = (v_0, v_1, v_2)$ satisfies

$$\det(v_1, v_2) = 1,$$

without loss of generality and up to exchanging v_1 and v_2. Then $(e_{12}, e_{20}, e_{01}) = (v_0^\perp, v_1^\perp, v_2^\perp)$ by (2.2), and this triplet is an M-obtuse superbase by Lemma 2.17. Denoting $(w_0, w_1, w_2) := (e_{12}, e_{20}, e_{01})$ one obtains

$$-\langle w_0, w_1\rangle_M - \langle w_0, w_2\rangle_M = \langle w_0, -w_1 - w_2\rangle_M = \|w_0\|_M^2,$$

and therefore $0 \leq -\langle w_0, w_1\rangle_M \leq \|w_0\|_M^2$. Likewise $0 \leq -\langle w_i, w_j\rangle_M \leq \|w_i\|_M^2$ for all $i \neq j$, which is the announced result. □

Proof in dimension $d = 3$. Denote $w_{ij} := v_i \times v_j$ for all $i \neq j$. In the following, $\{i, j, k, l\}$ denotes an arbitrary permutation of $\{0, 1, 2, 3\}$, thus for instance $w_{ij} = \pm e_{kl}$ by (2.2). Note also that

$$w_{ij} = -w_{ji}, \qquad \text{and} \qquad w_{ij} + w_{ik} + w_{il} = v_i \times (v_j + v_k + v_l) = -v_i \times v_i = 0.$$

The scalar products defined by $D \in S_3^{++}$ and its inverse $M := D^{-1}$ are related by the following identity, where $u, v, w \in \mathbb{R}^3$

$$\det(D)\langle u \times v,\, u \times w\rangle_M = \|u\|_D^2 \langle v, w\rangle_D - \langle u, v\rangle_D \langle u, w\rangle_D.$$

(In the case $D = \mathrm{Id}$ this is known as the Binet-Cauchy identity. In the general case where $D = A^T A$ for some $A \in \mathrm{GL}_3(\mathbb{R})$ it follows from a linear change of variables and the relation $(Au) \times (Av) = \mathrm{cof}(A)(u \times v)$ where $\mathrm{cof}(A)$ denotes the cofactor matrix.)

Choosing $u = v_i$, $v = v_j$ and $w = v_k$, we obtain that $\langle w_{ij}, w_{ik}\rangle_M \leq 0$. On the other hand

$$\begin{aligned} -\langle w_{ij}, w_{ik}\rangle_M - \langle w_{ij}, w_{il}\rangle_M &= \langle w_{ij}, v_i \times (-v_k - v_l)\rangle_M \\ &= \langle w_{ij}, v_i \times (v_i + v_j)\rangle_M \\ &= \|w_{ij}\|_M^2, \end{aligned}$$

thus $0 \leq -\langle w_{ij}, w_{ik}\rangle_M \leq \|w_{ij}\|_M^2$. Finally, since $-w_{kl} = w_{ki} + w_{kj}$, we obtain that

$$-\langle w_{ij}, w_{kl}\rangle_M = \langle w_{ij}, w_{ki} + w_{kj}\rangle_M = -\langle w_{ij}, w_{ik}\rangle_M + \langle w_{ji}, w_{jk}\rangle_M,$$

and therefore, by the previous estimate, $-\|w_{ij}\|_M^2 \leq \langle w_{ij}, w_{kl}\rangle_M \leq \|w_{ij}\|_M^2$. This concludes the proof of Lemma 3.11. □

4 Numerical experiments

We illustrate the PDE discretization introduced in this paper with synthetic numerical experiments, in dimension $d \in \{2, 3\}$, involving linear and quasi-linear operators, and using Dirichlet boundary conditions on a non-square and non-smooth domain. Let us mention that a close variant of the proposed scheme, involving the divergence form operator $\mathrm{div}(D(x)(\nabla u(x) - \omega(x)u(x)))$ featuring both a first and second order term, is used in [PCC+19] for image impainting purposes in dimension $d = 2$, in collaboration with one of the authors. See also [FM14] for applications to image denoising in dimension $d \in \{2, 3\}$, with an operator lacking the first order term however. Additional concrete applications of the proposed scheme will be the object of future work.

The PDEs addressed numerically in this section take the form

$$\mathcal{L}u(x) = f(x),\ \forall x \in \Omega, \qquad\qquad u(x) = g(x),\ \forall x \in \partial\Omega, \tag{4.1}$$

where $\Omega := \{x \in \mathbb{R}^d;\, \|x\| < 1\} \cup]0, 1[^d$ is the union of the d-dimensional unit ball and of the d-dimensional unit cube. The PDE operator $-\mathcal{L}u(x)$ is chosen as the following linear (resp. quasi-linear) expression

$$\begin{aligned} &\langle \omega(x), \nabla u(x)\rangle + \mathrm{Tr}(D(x)\nabla^2 u(x)) \\ &\Big(\text{resp. } \frac{1}{2}\langle \omega(x), \nabla u(x)\rangle^2 + \mathrm{Tr}(D(x)\nabla^2 u(x))\Big) \end{aligned} \tag{4.2}$$

whose coefficients $\omega \colon \overline{\Omega} \to \mathbb{R}^d$ and $D \colon \overline{\Omega} \to S_d^{++}$ are defined for any $x = (x_1, \cdots, x_d)$ in $\mathbb{R}^d$ by

$$\begin{aligned} \omega(x) &:= \frac{2 - \cos(\pi x_1)}{3}\omega_0(x), \\ D(x) &:= \mu\frac{2 + \cos(\pi x_1)}{3}\big(\nu I_d + (1 - \nu)\omega_0(x/2)\omega_0(x/2)^T\big), \end{aligned}$$

where the parameters $\mu, \nu > 0$ are specified in Figures 3 to 6, and where

$$\omega_0(x) := \begin{cases} (\cos(\pi x_2), \sin(\pi x_2)) & \text{if } d = 2, \\ (\cos(\pi x_2), \sin(\pi x_2)\cos(\pi x_3), \sin(\pi x_2)\sin(\pi x_3)) & \text{if } d = 3. \end{cases}$$

This particular choice of operator and coefficients is only meant to be reasonably simple and explicit, and to feature substantial anisotropy for the second order term — with $\nu = 1/10$ in the experiments. It also allows for a direct analytic verification of the assumptions of Theorem 1.6 ensuring the DDE property, see the last paragraph of this section.

For any discretization step $h > 0$, we let $\Omega_h := \Omega \cap h\mathbb{Z}^d$ and consider the finite differences scheme

$$L^h u(x) = f(x) \quad \text{in } \Omega_h, \tag{4.3}$$

where one has, denoting $g(x,p) := \langle \omega(x), p \rangle$ in the linear case (resp. $g(x,p) := \frac{1}{2}\langle \omega(x), p \rangle^2$ in the quasi-linear case)

$$-L^h u(x) := g(x, D(x)^{-1}\nabla^h_{D(x)} u(x)) + \Delta^h_{D(x)} u(x).$$

The Dirichlet boundary condition from (4.1) does not appear in (4.3) because it is implicitly implemented via the finite differences operators, defined as (A.3) and (A.4) when the point x is near $\partial\Omega$. See Appendix A for more discussion on the extension of the scheme of Definition 1.4 to non-constant coefficients, Dirichlet boundary conditions, and non-linear operators.

As announced, we present synthetic tests of our numerical scheme. For that purpose, a function $u \colon \overline{\Omega} \to \mathbb{R}$ is chosen with a closed form expression, and the right hand side $f \colon \Omega \to \mathbb{R}$ is generated by symbolic differentiation and evaluation of $\mathcal{L}u$, so that u obeys (4.1) with boundary condition $g := u_{|\partial\Omega}$. The discretized PDE (4.3) is then solved for a range of grid scales $h > 0$, and the resulting $l^1(\Omega_h)$ and $l^\infty(\Omega_h)$ reconstruction errors are reported in Figures 3 to 6.

The chosen exact solutions are a smooth function $\mathbf{u}_1$, a $C^{2,0.5}$ function $\mathbf{u}_2$, and a singular function $\mathbf{u}_3$, inspired by [FJ17] for $\mathbf{u}_1$ and by [FO13] for $\mathbf{u}_2$ and $\mathbf{u}_3$, and defined in $\overline{\Omega}$ by

$$\mathbf{u}_1(x) := \frac{1}{4}\|x\|^4, \qquad \mathbf{u}_2(x) := \max(0, \|x\| - 0.4)^{2.5}, \qquad \mathbf{u}_3(x) := \sqrt{d - \|x\|^2}. \tag{4.4}$$

The multiplicative coefficient $1/4$ in the definition of $\mathbf{u}_1$ is choosen so that the range of values taken by $\|\nabla \mathbf{u}_1\|$ in Ω remains close to the one of values taken by $\|\nabla \mathbf{u}_2\|$, since those values influence the DDE property of the scheme (4.3) in the quasi-linear case, see §4.1. In numerical experiments, we also adjust the parameter μ in the definition of the tensor field $D \colon \overline{\Omega} \to S_d^{++}$ so that DDE holds at reasonable grid scales.

Empirically we observe second order convergence $\|\mathbf{u} - u_h\|_1 = \mathcal{O}(h^{-2})$ and $\|\mathbf{u} - u_h\|_\infty = \mathcal{O}(h^{-2})$, where $\mathbf{u}$ is among the two test functions $\mathbf{u}_1$ and $\mathbf{u}_2$ defined in (4.4) and u_h is the numerical solution of (4.3) with the corresponding r.h.s. for both the linear and quasi-linear operators (4.2), in both dimension two and three, see Figures 3 to 6. For the test function $\mathbf{u}_3$, first order convergence $\|\mathbf{u}_3 - u_h\|_1 = \mathcal{O}(h^{-1})$ and $\|\mathbf{u}_3 - u_h\|_\infty = \mathcal{O}(h^{-1})$ is observed instead. From a theoretical standpoint, convergence was not expected for $\mathbf{u}_3$ and the quasi-linear scheme, since the DDE property is not guaranteed in this case, even for small h.

For the quasi-linear equations, a Newton method is used, converging in at most 12 iterations in our experiments with tolerance 10^{-8} on the max-norm of residual of the discretized PDE.

Remark 4.1 (Dominant source of numerical error). The curves of convergence associated to the linear and quasi-linear equations are conspicuously similar for the function $\mathbf{u}_2$ in dimension two,

see Figures 3 and 4, and $\mathbf{u}_1$, $\mathbf{u}_2$, and $\mathbf{u}_3$ in dimension three, see Figures 5 and 6. This suggests that the discretization of the first order term in (4.2) is not the dominant source of error in these cases.

For $\mathbf{u}_1$ and $\mathbf{u}_3$ in dimension three, we obtained a different convergence curve when changing the tensor field $D : \overline{\Omega} \to S_d^{++}$, suggesting that the discretization of $\mathrm{Tr}(D\nabla^2 u)$ is the dominant source of numerical error. For the $C^{2,0.5}$ function $\mathbf{u}_2$, we did *not* observe a significative difference in the curves of convergence when changing the tensor field D, but we did observe one when replacing the radius $r = 0.4$ with 0.5 in its definition, suggesting that the dominant source of error is related to the configuration of the grid points Ω_h in the vicinity of the sphere of radius r across which $\mathbf{u}_2$ is non-smooth.

4.1 Theoretical guarantees of Discrete Degenerate Ellipticity

An a-priori analysis allows to guarantee the DDE property of the numerical schemes used in our numerical experiments (except in one case where it fails), thanks to the explicit and reasonably simple expression of the PDE coefficients (4.2) (and, in the quasi-linear case, of the PDE solution (4.4)). In practical applications, such an analysis may not be possible, but alternatively the DDE property can be checked numerically by looking at the sign of the coefficients of the Jacobian matrix of the discretized operator L^h.

Letting $M(x) := D(x)^{-1}$, one easily obtains

$$\|M(x)\| = \mu^{-1}(3/(2+\cos(\pi x_1)))\nu^{-1} \leq 3\mu^{-1}\nu^{-1},$$

and therefore

$$\|M(x)\|^{1/2}\|\omega(x)\|_{M(x)} \leq \|M(x)\|\|\omega(x)\| \leq 3\mu^{-1}\nu^{-1}.$$

It follows that the pair $(h\omega(x), D(x))$ is canonically feasible as soon as $h \leq c_d\mu\nu/3$, where the absolute constant c_d is specified in Theorem 1.6. The discretization of the linear operator (4.2, left) is thus DDE under these conditions.

We now check whether the discretization of the *quasi*-linear operator (4.2, right) is DDE in a neighborhood of the solutions (4.4), by linearizing the operator. For any $x \in \overline{\Omega}$ and $p \in \mathbb{R}^d$ one has $\|\nabla_p g(x, \nabla u(x))\| = \|\langle\omega(x), \nabla u(x)\rangle\omega(x)\| \leq \|\omega(x)\|^2\|\nabla u(x)\| \leq \|\nabla u(x)\|$. By the same resoning as above, if $\mathbf{u}$ denotes either one of the functions $\mathbf{u}_1$ and $\mathbf{u}_2$ in (4.4), then the pair $(h\nabla_p g(x, \nabla\mathbf{u}(x)), D(x))$ is canonically feasible for all $x \in \Omega$, and thus the scheme (4.3) is DDE in the neighborhood of $\mathbf{u}$, as soon as

$$h < \frac{c_d\mu\nu}{3\sup_{x\in\Omega}\|\nabla\mathbf{u}(x)\|},$$

where we used that $\|\nabla\mathbf{u}_1(x)\|$ and $\|\nabla\mathbf{u}_2(x)\|$ are bounded on Ω. In contrast $\|\nabla\mathbf{u}_3(x)\|$ is unbounded when $x \to (1, \cdots, 1) \in \partial\Omega$. Thus DDE fails in the neighborhood of $\mathbf{u}_3$, but as noted above we do still observe convergence empirically in this particular case.

5 Conclusion and perspectives

In this paper, we answer whether one can discretize linear PDE operator, of order at most two and in dimension $d \leq 3$, using a second order consistent finite difference scheme obeying the degenerate ellipticity property. The question is basic and of broad interest, and in dimension $d = 1$ the answer is indeed simple, well known, and taught at a basic level. In dimension $d \in \{2, 3\}$ however the anisotropy of the second order part of the operator comes into play, and

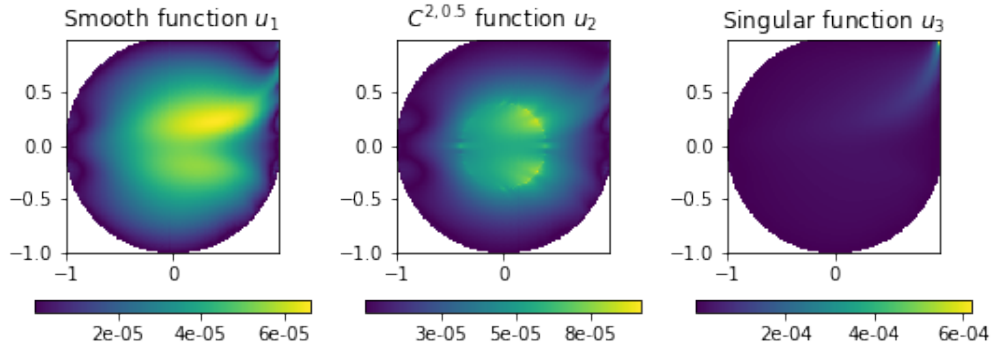

Figure 2: Errors in numerical solutions to the linear equation in dimension $d = 2$, with parameters $\mu = 1$, $\nu = 1/10$, $h = 1/100$, and with exact solutions $\mathbf{u}_1$, $\mathbf{u}_2$, and $\mathbf{u}_3$.

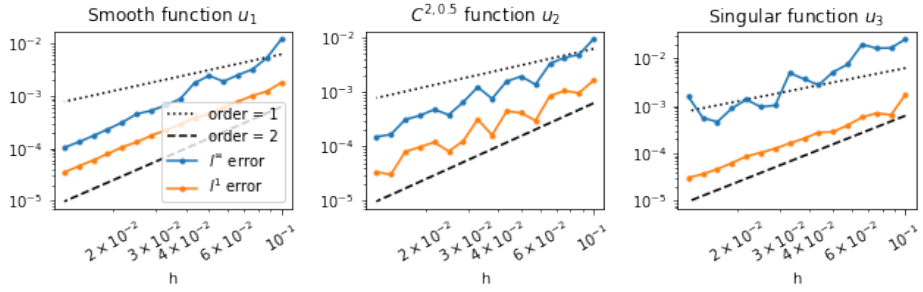

Figure 3: Convergence of the numerical scheme for the linear equation in dimension $d = 2$, with parameters $\mu = 1$ and $\nu = 1/10$, and with exact solutions $\mathbf{u}_1$, $\mathbf{u}_2$, and $\mathbf{u}_3$. Degenerate ellipticity is guaranteed by §4.1 for $h \leq 1/30 \approx 0.0333$ and empirically observed up to $h \approx 0.0660$.

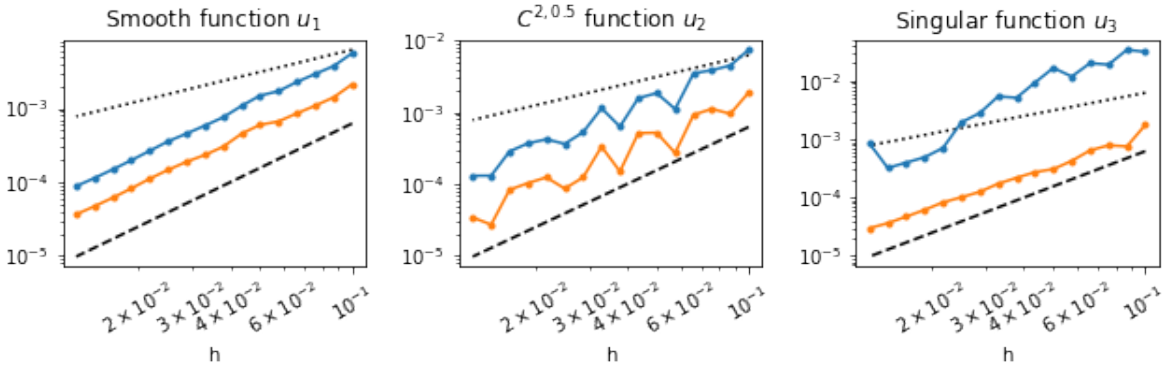

Figure 4: Convergence of the numerical scheme for the quasi-linear equation in dimension $d = 2$, with parameters $\mu = 2$ and $\nu = 1/10$, and with exact solutions $\mathbf{u}_1$, $\mathbf{u}_2$, and $\mathbf{u}_3$. The legend is as in Figure 3. In the neighborhood of functions $\mathbf{u}_1$ and $\mathbf{u}_2$, degenerate ellipticity is guaranteed by §4.1 respectively for $h < 1/(30\sqrt{2}) \approx 0.0236$ and for $h < 1/(75(\sqrt{2} - 0.4)^{1.5}) \approx 0.0131$. It is observed empirically in the last iteration of the Newton method respectively up to $h \approx 0.0379$ and up to $h \approx 0.0435$. In the case of the singular function $\mathbf{u}_3$, degenerate ellipticity is not theoretically guaranteed, but it is nevertheless observed empirically in the last iteration of the Newton method up to $h \approx 0.0574$.

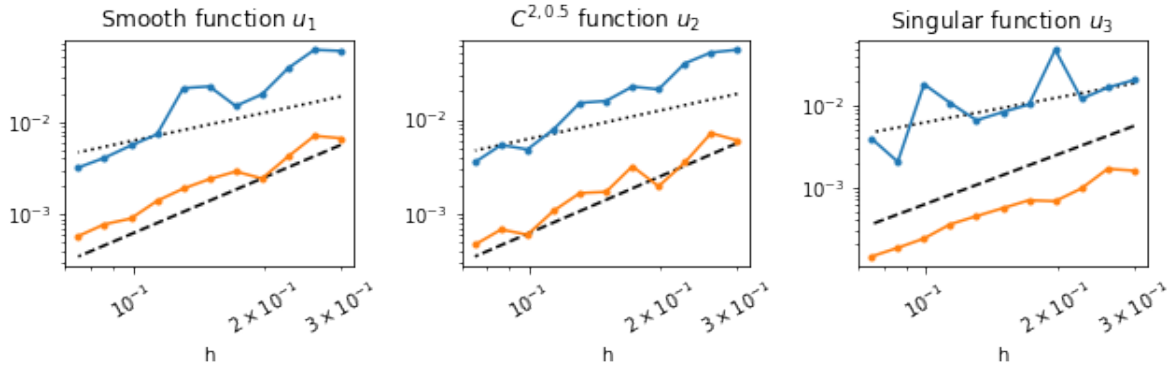

Figure 5: Convergence of the numerical scheme for the linear equation in dimension $d = 3$, with parameters $\mu = 4$ and $\nu = 1/10$, and with exact solutions $\mathbf{u}_1$, $\mathbf{u}_2$, and $\mathbf{u}_3$. The legend is as in Figure 3. Degenerate ellipticity is guaranteed by §4.1 for $h \leq 1/(5\sqrt{3}) \approx 0.115$ and empirically observed up to $h \approx 0.198$.

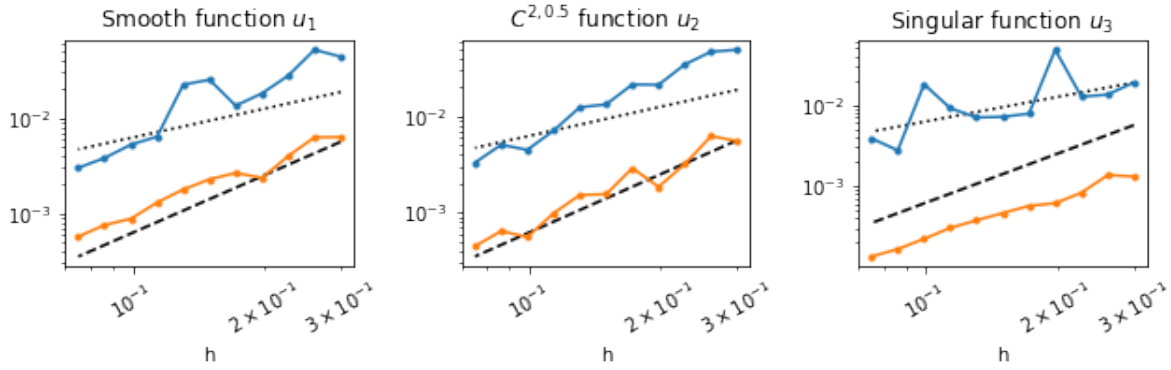

Figure 6: Convergence of the numerical scheme for the quasi-linear equation in dimension $d = 3$, with parameters $\mu = 8$ and $\nu = 1/10$, and with exact solutions $\mathbf{u}_1$, $\mathbf{u}_2$, and $\mathbf{u}_3$. The legend is as in Figure 3. In the neighborhood of functions $\mathbf{u}_1$ and $\mathbf{u}_2$, degenerate ellipticity is guaranteed by §4.1 respectively for $h < 2/135 \approx 0.0148$ and for $h < 1/(75\sqrt{3}(\sqrt{3} - 0.4)^{1.5}) \approx 0.00501$. It is observed empirically in the last iteration of the Newton method respectively up to $h \approx 0.131$ and up to $h \approx 0.261$. In the case of the singular function $\mathbf{u}_3$, degenerate ellipticity is not theoretically guaranteed, but it is nevertheless observed empirically in the last iteration of the Newton method up to $h = 0.3$, that is, for all values of h we tested in the graphs above.

a subtler analysis is required. Leveraging tools from the field of Euclidean lattice geometry, we could characterize whether a discretization exists, and provide an explicit (quasi-)optimal construction. Numerical experiments illustrate the efficiency of the method in dimension $d \in \{2, 3\}$, on linear and quasi-linear problems.

Several research directions are open, both practical and theoretical, including (i) applications to PDEs arising from concrete problems, especially those whose first order term is large, e.g. depending on a relaxation parameter, (ii) extensions to fully non-linear HJB PDEs, and (iii) a theoretical analysis of the convergence rates. Another interesting open problem is the extension of our results for a dimension $d > 3$, which is not obvious since, as already mentioned after Definition 2.1, then D-obtuse bases do not necessarily exist.

A Adaptation to quasi-linear and fully non-linear PDEs

The numerical scheme presented in the introduction of this paper applies to *linear schemes* with *constant coefficients*, defined over the *full space* $\mathbb{R}^d$, $d \in \{2, 3\}$. We illustrate in this appendix how the three restrictions in emphasis can be relaxed. For that purpose let us recall the definition of a degenerate elliptic scheme, in a general setting.

Definition A.1. Let X be a discrete set, and for each $x \in X$ let $V(x) \subseteq X \setminus \{x\}$ be a finite set (the neighbors, or stencil of x). Let also $\mathbb{U} := \mathbb{R}^X$. A numerical scheme on X, with stencil V, is a mapping $F : \mathbb{U} \to \mathbb{U}$ of the form

$$Fu(x) := \mathcal{F}(x, u(x), [u(x) - u(y)]_{y \in V(x)}).$$

It is said discrete degenerate elliptic (DDE) iff $\mathcal{F}$ is non-decreasing w.r.t. the second and third arguments (coordinate wise).

Definition 1.1, from the introduction, is a special case of Definition A.1, adapted to linear schemes with constant coefficients, and choosing $X = h\mathbb{Z}^d$ and $V(x) := \{x + he_i;\ 1 \leq |i| \leq I\}$. In the rest of this appendix, we show how various natural extensions of our numerical scheme fit into the general framework of Definition A.1.

Non constant coefficients

Discrete Degenerate Ellipticity is a local property, which only needs to be verified pointwise, independently at each point $x \in X$ of the discretization domain, see Definition A.1. As a result, the numerical scheme presented in this paper trivially extends to non-constant coefficients. More precisely, let ω and D be a field of vectors and of symmetric positive definite matrices, and let $h > 0$ be a grid scale. Then we can define the counterparts with variable coefficients of the linear PDE operator (1.1) and of its canonical discretization (1.7)

$$-\mathcal{L}u(x) := \langle \omega(x), \nabla u(x) \rangle + \mathrm{Tr}(D(x)\nabla^2 u(x)), \tag{A.1}$$

$$-L^h u(x) := \langle D(x)^{-1}\omega(x), \nabla^h_{D(x)} u(x) \rangle + \Delta^h_{D(x)} u(x). \tag{A.2}$$

The scheme L^h is DDE under the same conditions, pointwise, as in the constant coefficient case. It is not hard to show that the coefficients $x \mapsto \rho^h_i(x) \geq 0$ of L^h expressed as in (1.1) are Lipschitz, provided ω and D are Lipschitz. Interestingly, convergence rates have been established in a similar setting [Kry05] but under the slightly stronger assumption that $x \mapsto \sqrt{\rho^h_i(x)}$ is Lipchitz.

Dirichlet boundary conditions

Consider a bounded open domain $\Omega \subseteq \mathbb{R}^d$, $d \in \{2, 3\}$, equipped with Dirichlet boundary conditions $f : \partial\Omega \to \mathbb{R}$, and let $\Omega_h := \Omega \cap h\mathbb{Z}^d$, where the grid scale $h > 0$ is fixed in the following. Forall $x \in \Omega_h$, $e \in \mathbb{Z}^d \setminus \{0\}$, define $h_x^e := \min\{k > 0;\ x + ke \in \Omega_h\}$, and note that $0 < h_x^e \leq h$. Introduce the first and second finite difference operators, where for convenience we denote $h^\pm := h_x^{\pm e}$, and where $u : \Omega_h \to \mathbb{R}$ is extended to $\partial\Omega$ using the provided Dirichlet boundary condition

$$\delta_e^h u(x) := \frac{1}{2}\Big(\frac{u(x + h^+ e) - u(x)}{h^+} - \frac{u(x - h^- e) - u(x)}{h^-}\Big), \tag{A.3}$$

$$\Delta_e^h u(x) := \frac{2}{h^+ + h^-}\Big(\frac{u(x + h^+ e) - u(x)}{h^+} + \frac{u(x - h^- e) - u(x)}{h^-}\Big). \tag{A.4}$$

Note that this construction coincides with Definition 1.3 when x is sufficiently far from $\partial\Omega$. For smooth u, one has $\delta_e^h u(x) = \langle \nabla u(x), e\rangle + \mathcal{O}(h^r)$ and $\Delta_e^h u(x) = \mathcal{O}(h^r)$ where $r = 1$ if x is close to $\partial\Omega_h$, and $r = 2$ otherwise. In addition the discrete operator defined by

$$-L^h u(x) := \lambda\, \delta_e^h u(x) + \Delta_e^h u(x)$$

is DDE provided $h\lambda \leq 2$, similarly to the constant coefficient case, since $0 < h_x^{\pm e} \leq h$. Therefore (A.3) and (A.4) can be used as a drop in replacement for the finite difference operators of Definition 1.3 when Dirichlet boundary conditions are used, the resulting scheme is DDE under the same conditions. More complex boundary conditions may require ad-hoc treatment.

Quasi-linear operators

Let $D \in S_d^{++}$, $d \in \{2, 3\}$, and let $g : \mathbb{R}^d \to \mathbb{R}$ be a smooth function. Consider the quasi-linear operator $\mathcal{L}$ and its discretization L^h defined by

$$-\mathcal{L}u(x) := g(\nabla u(x)) + \mathrm{Tr}(D\nabla^2 u(x)), \qquad -L^h u(x) := g(D^{-1}\nabla_D^h u(x)) + \Delta_D^h u(x).$$

The operator $\mathcal{L}$ is degenerate elliptic, since in the continuous setting this property is independent of the first order term of the PDE. On the other hand, the scheme L^h is DDE provided the linear scheme $\tilde{L}^h$ defined by $-\tilde{L}^h u(x) := \langle D^{-1}\omega, \nabla_D^h u(x)\rangle + \Delta_D^h u(x)$ is DDE for all $\omega \in \nabla g(\mathbb{R}^d) = \{\nabla g(x); x \in \mathbb{R}^d\}$. (This is a severe restriction if g is e.g. a quadratic function, but for such applications it can be enough to check that the scheme is DDE in a neighborhood of the solution.)

Fully non-linear operators

Fully non-linear HJB operators can be expressed, under mild regularity assumptions, in Isaacs form

$$\mathcal{L}u(x) := \sup_{\alpha\in A}\inf_{\beta\in B} \mathcal{L}_{\alpha\beta} u(x), \tag{A.5}$$

where A, B are known as the control sets. In addition $\mathcal{L}_{\alpha\beta}$ is a linear DE operator, for all $\alpha \in A$, $\beta \in B$,

$$\mathcal{L}_{\alpha\beta}u(x) := \mu_{\alpha\beta}(x) + \lambda_{\alpha\beta}(x)u(x) + \langle \omega_{\alpha\beta}(x), \nabla u(x)\rangle - \mathrm{Tr}(D_{\alpha\beta}(x)\nabla^2 u(x)),$$

where $\mu_{\alpha\beta}(x) \in \mathbb{R}$, $\lambda_{\alpha\beta}(x) \geq 0$, $\omega_{\alpha\beta}(x) \in \mathbb{R}^d$, and $D_{\alpha\beta}(x) \in S_d^+$. In the special case where the set A or B is a singleton, which is common (consider the Monge-Ampere [BCM] or Pucci [BBM21] equations), then (A.5) is known as the Bellman form of the operator.

It is in principle possible to introduce samples $A_h \subseteq A$ and $B_h \subseteq B$ of the control sets, and construct a discretization $L^h_{\alpha\beta}$ of each linear operator $\mathcal{L}_{\alpha\beta}$ following the approach presented in this paper. This produces a DDE approximation of the operator $\mathcal{L}$

$$L^h u(x) := \sup_{\alpha \in A_h} \inf_{\beta \in B_h} L^h_{\alpha\beta} u(x).$$

Let us acknowledge, however, that this construction is far from straightforward to put in practice, especially if the sets A and B are non-compact, and if the condition number of the matrices $D_{\alpha\beta}(x)$ is not uniformly bounded.

B Terminology and elementary properties of polyhedra

In this section, we recall some of the terminology and elementary properties related with polyhedra, limiting our attention to those which are immediately useful in the study of Ryskov's polyhedron and its variant §2.2 and §3.1. See [BG15] for a more complete reference.

B.1 Regularity and skeleton

Definition B.1. A *polyhedron* in $\mathbb{R}^n$ is a set of the form

$$\mathcal{M} := \{x \in \mathbb{R}^n;\, \forall i \in I,\, \langle l_i, x\rangle \geq \alpha_i\}, \tag{B.1}$$

where $l_i \in \mathbb{R}^n$, $\alpha_i \in \mathbb{R}$, and I is a finite or countable set. The polyhedron $\mathcal{M}$ is said *regular* iff it (i) has a non-empty interior, (ii) does not contain any affine line, and (iii) can be locally described by the constraints corresponding to a finite subset of I.

By definition, a polyhedron is thus a convex set. Condition (ii) can be reformulated as $\mathrm{Span}\{l_i; i \in I\} = \mathbb{R}^n$. Condition (iii) can be reformulated as follows: for all $x \in \mathcal{M}$ there exists a positive radius $r > 0$ and a finite subset $I_0 \subseteq I$ such that

$$\langle l_i, y\rangle > \alpha_i,\ \forall i \in I \setminus I_0,\ \forall y \in B(x, r).$$

Definition B.2. Let $\mathcal{M}$ be a regular polyhedron, defined as in (B.1). A k-facet of $\mathcal{M}$, where $1 \leq k \leq n$, is a *non-empty* subset of $\mathcal{M}$ of the form

$$\{x \in \mathcal{M};\, \forall i \in J,\, \langle l_i, x\rangle = \alpha_i\}, \qquad \text{where} \quad \dim \mathrm{Span}\{l_i;\, i \in J\} = n - k,$$

and where $J \subseteq I$ denotes a subset of the constraint indices.

By construction, a k-facet is a convex subset of $\partial\mathcal{M}$ of affine dimension k. If a k-facet satisfies $\#(J) > n - k$, where $J \subseteq I$ is chosen maximal for inclusion, then it is said degenerate. By construction 0-facets are singletons, and their single point is called a *vertex*. On the other hand 1-facets are known as *edges* and come in two flavors

- *Bounded edges*, of the form $[x_1, x_2] := \{(1-t)x_1 + tx_2;\, 0 \leq t \leq 1\}$, where x_1 and x_2 are vertices.
- *Unbounded edges*, of the form $\{x + \lambda v;\, \lambda \geq 0\}$, where x is a vertex, and $v \in \mathbb{R}^n \setminus \{0\}$ is called the unbounded edge direction (unique up to multiplication by a positive constant).

Note that *doubly unbounded edges*, of the form $\{x + \lambda v;\, \lambda \in \mathbb{R}\}$, are affine lines and are thus excluded by Definition B.1.

Remark B.3. Let $\mathcal{M}$ be a regular polyhedron, in the sense of Definition B.1. An element $x \in \mathcal{M}$ is a vertex iff $\mathbb{R}^n = \mathrm{Span}\{l_i;\, i \in I, \langle l_i, x\rangle = \alpha_i\}$.

B.2 Linear programs

Linear programs are defined as the optimization of a linear functional over a polytope. A fundamental result of operational research, is that such problems can under suitable assumptions be solved by a greedy search over the graph defined by the edges of the polytope, such as the simplex algorithm [BG15]. Since Definition B.1 allows for infinitely many constraints, which is slightly more general than the common setting, we establish in Proposition B.4 a basic result on such programs, used in §3.2. Note that the infima in (B.2) may not be attained.

Proposition B.4. *Let $\mathcal{M}$ be a regular polyhedron. Then for any $l \in \mathbb{R}^n$*

$$\begin{aligned} &\inf\{\langle l, x\rangle;\, x \in \mathcal{M}\} \\ &= \begin{cases} -\infty \text{ if } \langle l, v\rangle < 0 \text{ for some unbounded edge direction } v, \\ \inf\{\langle l, x\rangle;\ x \text{ vertex of } \mathcal{M}\} \text{ otherwise.} \end{cases} \end{aligned} \tag{B.2}$$

Proof. By point (i) of Definition B.1, there exists $x_* \in \mathrm{int}(\mathcal{M})$. By point (ii) of Definition B.1, one has $\mathrm{Span}\{l_i; i \in I\} = \mathbb{R}^n$, otherwise $x_* + \mathbb{R}v$ is an affine line contained in $\mathcal{M}$ for any non-zero $v \in \mathrm{Span}\{l_i; i \in I\}^\perp$, hence there exists $I_* \subseteq I$ with $\#(I_*) = n$ and such that $(l_i)_{i\in I_*}$ is a basis of $\mathbb{R}^n$.

Define $l_* := \sum_{i\in I_*} l_i$, and consider for each $\alpha > l_*(x_*)$ the set $\mathcal{M}_\alpha := \{x \in \mathcal{M};\, \langle l_*, x\rangle \leq \alpha\}$. Note that for each $x \in \mathcal{M}_\alpha$ and $i \in I_*$ one has $0 \leq l_i(x) - \alpha_i \leq \alpha - \sum_{i\in I_*} \alpha_i$, hence $\mathcal{M}_\alpha$ is bounded. Thus $\mathcal{M}_\alpha$ is a compact polyhedron with non-empty interior, which by Definition B.1 (iii) is characterized by finitely many linear constraints. By Carathéodory's theorem, $\min\{\langle l, x\rangle;\, x \in \mathcal{M}_\alpha\}$ is attained at a vertex of $\mathcal{M}_\alpha$, which by construction is either a vertex of $\mathcal{M}$ or the intersection of an edge of $\mathcal{M}$ (bounded or not) with the hyperplane $\{x \in \mathbb{R}^n;\, \langle l_*, x\rangle = \alpha\}$. From this point, and noting that $\mathcal{M} = \cup_{\alpha\in\mathbb{R}} \mathcal{M}_\alpha$, the announced result easily follows. □

Definition B.5 (Karush-Kuhn-Tucker relations)**.** A set of KKT relations for l in $\mathbb{R}^n$ and x in $\mathcal{M}$ is a finitely supported family of non-negative coefficients $(\lambda_i)_{i\in I}$ such that

$$l = \sum_{i\in I} \lambda_i l_i, \qquad \text{and } \ \forall i \in I,\ \lambda_i = 0 \ \text{ or } \ \langle l_i, x\rangle = \alpha_i.$$

It is known [BG15] that a linear form $l \in \mathbb{R}^n$ attains its minimum at a given point x of a regular polyhedron $\mathcal{M}$, *if and only if* there exists KKT relations for l and x. The next result establishes a uniqueness property of the KKT relations.

Proposition B.6. *Let $\mathcal{M} \subseteq \mathbb{R}^n$ be a regular polyhedron, in the sense of Definition B.1. Assume that one has a set of KKT relations $(\lambda_i)_{i\in I}$ for some $l \in \mathbb{R}^d$ at a* non-degenerate *vertex $x \in \mathcal{M}$. Then any other KKT relations $(\lambda'_i)_{i\in I}$ at some $x' \in \mathcal{M}$ (possibly distinct from x), for the same l, obey $\lambda_i = \lambda'_i$ for all $i \in I$.*

Proof. For all $i \in I$ such that $\lambda'_i > 0$ one has $\langle l_i, x'\rangle = \alpha_i$, thus $\langle l_i, x - x'\rangle \geq 0$. On the other hand one has $\langle l, x\rangle = \langle l, x'\rangle = \inf_{\mathcal{M}}$, and therefore $0 = \langle l, x - x'\rangle = \sum_{i\in I} \lambda'_i \langle l_i, x - x'\rangle$. Combining these two arguments we obtain that for all $i \in I$ such that $\lambda'_i > 0$ one has $\langle l_i, x - x'\rangle = 0$, and therefore $\langle l_i, x\rangle = \alpha_i$. Since x is a non-degenerate vertex, the family $\{l_i;\, i \in I, \langle l_i, x\rangle = \alpha_i\}$ is a basis of $\mathbb{R}^n$, which implies the announced result. □

B.3 Edges originating from a vertex

In this section, we present a constructive enumeration of all the edges of a regular polyhedron $\mathcal{M}$ containing a given vertex x. This description follows from Definition B.2 of k-facets, here with $k = 1$. We use the notations of Definition B.1.

Let $J := \{i \in I;\ \langle l_i, x\rangle = \alpha_i\}$ denote the indices of all the active constraints at the vertex x of $\mathcal{M}$. In order to enumerate all the edges of $\mathcal{M}$ containing x, bounded or unbounded, the steps are the following:

(A) Consider successively all subsets S of J with cardinality $n-1$.

(B) If $\dim \mathrm{Span}\{l_i;\ i \in S\} < n-1$, then skip this subset. Otherwise denote by $\nu \in \mathbb{R}^n \setminus \{0\}$ the vector, which is unique up to a scalar multiplication, such that $\langle l_i, \nu\rangle = 0$ for all $i \in S$.

(C) Replace ν with its opposite $-\nu$, if necessary, in such way that $\langle l_i, \nu\rangle \geq 0$ for all $i \in J \setminus S$. If that is not possible, then skip this subset.

(D) Compute $\Lambda := \sup\{\lambda \in \mathbb{R}; x + \lambda\nu \in \mathcal{M}\}$. If $\Lambda = +\infty$, then there is an unbounded edge at x in the direction of ν. Otherwise, x and $x + \Lambda\nu$ are the vertices of a bounded edge of $\mathcal{M}$.

References

[BBM21] J. F. Bonnans, G. Bonnet, and J.-M. Mirebeau, *Monotone and second order consistent scheme for the two dimensional Pucci equation*, Numerical Mathematics and Advanced Applications ENUMATH 2019 (F. J. Vermolen and C. Vuik, eds.), Springer International Publishing, to appear, 2021.

[BCM] J.-D. Benamou, F. Collino, and J.-M. Mirebeau, *Monotone and consistent discretization of the Monge-Ampère operator*.

[BG15] J. F. Bonnans and S. Gaubert, *Recherche opérationnelle*, Aspects mathématiques et applications, Les Éditions de l'École polytechnique, 2015.

[BOZ04] J. F. Bonnans, É. Ottenwaelter, and H. Zidani, *A fast algorithm for the two dimensional HJB equation of stochastic control*, ESAIM: Mathematical Modelling and Numerical Analysis **38** (2004), no. 4, 723–735.

[CIL92] M. G. Crandall, H. Ishii, and P.-L. Lions, *User's guide to viscosity solutions of second order partial differential equations*, Bulletin of the American Mathematical Society **27** (1992), no. 1, 1–67.

[CS88] J. H. Conway and N. J. A. Sloane, *Low-dimensional lattices. III. Perfect forms*, Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences **418** (1988), no. 1854, 43–80.

[CS92] ———, *Low-dimensional lattices. VI. Voronoi reduction of three-dimensional lattices*, Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences **436** (1992), no. 1896, 55–68.

[DSSV12] M. Dutour Sikirić, A. Schürmann, and F. Vallentin, *Inhomogeneous extreme forms*, Annales de l'institut Fourier, 2012, pp. 2227–2255.

[Erd92] R. Erdahl, *A cone of inhomogeneous second-order polynomials*, Discrete and Computational Geometry **8** (1992), no. 4, 387–416.

[FJ17] X. Feng and M. Jensen, *Convergent semi-Lagrangian methods for the Monge–Ampère equation on unstructured grids*, SIAM Journal on Numerical Analysis **55** (2017), no. 2, 691–712.

[FM14] J. Fehrenbach and J.-M. Mirebeau, *Sparse non-negative stencils for anisotropic diffusion*, Journal of Mathematical Imaging and Vision **49** (2014), no. 1, 123–147.

[FO11] B. D. Froese and A. M. Oberman, *Convergent finite difference solvers for viscosity solutions of the elliptic Monge–Ampère equation in dimensions two and higher*, SIAM Journal on Numerical Analysis **49** (2011), no. 4, 1692–1714.

[FO13] ———, *Convergent filtered schemes for the Monge-Ampère partial differential equation*, SIAM Journal on Numerical Analysis **51** (2013), no. 1, 423–444.

[Kry05] N. V. Krylov, *The rate of convergence of finite-difference approximations for Bellman equations with Lipschitz coefficients*, Applied Mathematics and Optimization **52** (2005), no. 3, 365–399.

[LN18] W. Li and R. H. Nochetto, *Optimal pointwise error estimates for two-scale methods for the Monge–Ampère equation*, SIAM Journal on Numerical Analysis **56** (2018), no. 3, 1915–1941.

[Mir16] J.-M. Mirebeau, *Minimal stencils for discretizations of anisotropic PDEs preserving causality or the maximum principle*, SIAM Journal on Numerical Analysis **54** (2016), no. 3, 1582–1611.

[Mir17] ———, *Fast-marching methods for curvature penalized shortest paths*, Journal of Mathematical Imaging and Vision (2017), 1–32.

[Mir19] ———, *Riemannian fast-marching on cartesian grids, using Voronoi's first reduction of quadratic forms*, SIAM Journal on Numerical Analysis **57** (2019), no. 6, 2608–2655.

[NS04] P. Q. Nguyen and D. Stehlé, *Low-dimensional lattice basis reduction revisited*, ANTS (Ducan Buell, ed.), Springer, 2004, pp. 338–357.

[Obe06] A. M. Oberman, *Convergent difference schemes for degenerate elliptic and parabolic equations: Hamilton-Jacobi equations and free boundary problems*, SIAM Journal on Numerical Analysis **44** (2006), no. 2, 879–895.

[PCC+19] S. Parisotto, L. Calatroni, M. Caliari, C.-B. Schönlieb, and J. Weickert, *Anisotropic osmosis filtering for shadow removal in images*, Inverse Problems **35** (2019), no. 5, 054001.

[Sch09a] A. Schürmann, *Computational geometry of positive definite quadratic forms*, University Lecture Series **49** (2009).

[Sch09b] ———, *Enumerating perfect forms*, Contemporary Mathematics **493** (2009), 359.

[Sel74] E. Selling, *Ueber die binären und ternären quadratischen Formen*, Journal fur die Reine und Angewandte Mathematik **77** (1874), 143–229.

[SSV07] M. Sikirić, A. Schürmann, and F. Vallentin, *Classification of eight-dimensional perfect forms*, Electronic Research Announcements of the American Mathematical Society **13** (2007), no. 3, 21–32.

[Vor08] G. Voronoi, *Sur quelques propriétés des formes quadratiques positives parfaites*, J. reine angew. Math, 1908.