Monotone discretization of the Monge-Ampère equation of optimal transport

Guillaume Bonnet, Jean-Marie Mirebeau

Monotone discretization of the Monge-Ampère equation of optimal transport

Guillaume Bonnet* Jean-Marie Mirebeau†

March 12, 2022

Abstract

We design a monotone finite difference discretization of the second boundary value problem for the Monge-Ampère equation, whose main application is optimal transport. We prove the existence of solutions to a class of monotone numerical schemes for degenerate elliptic equations whose sets of solutions are stable by addition of a constant, and we show that the scheme that we introduce for the Monge-Ampère equation belongs to this class. We prove the convergence of this scheme, although only in the setting of quadratic optimal transport. The scheme is based on a reformulation of the Monge-Ampère operator as a maximum of semilinear operators. In dimension two, we recommend to use Selling's formula, a tool originating from low-dimensional lattice geometry, in order to choose the parameters of the discretization. We show that this approach yields a closed-form formula for the maximum that appears in the discretized operator, which allows the scheme to be solved particularly efficiently. We present some numerical results that we obtained by applying the scheme to quadratic optimal transport problems as well as to the far field refractor problem in nonimaging optics.

1 Introduction

The problem of *optimal transport* [45] is strongly related to the *Monge-Ampère equation* [30]: under suitable assumptions, the potential function which solves an optimal transport problem is also solution to the Monge-Ampère equation associated with this problem, equipped with the relevant boundary condition [20]. Some problems in nonimaging optics are also described by Monge-Ampère equations, among which some fit in the framework of optimal transport [12, 30] and some do not [31, 34].

Let us outline some approaches to the numerical resolution of optimal transport problems. One may solve an entropic regularization of a discrete optimal transport problem using Sinkhorn's iterations [17]. The Benamou-Brenier method [2] is based on an extension of the optimal transport problem, with an added time variable. Some methods were also developed to solve semi-discrete optimal transport problems [33], and applied to problems in nonimaging optics [18]. Finally, one may solve numerically the Monge-Ampère equation associated with the considered optimal transport problem, as suggested in this paper and previously in [4, 26]. One benefit of this last approach is that, as illustrated by our numerical experiments in the setting of nonimaging optics in section 6.5, it can be applied to optimal transport problems with various cost functions, provided

*LMO, Université Paris-Saclay, Orsay, France, and Inria-Saclay and CMAP, École Polytechnique, Palaiseau, France

†Université Paris-Saclay, ENS Paris-Saclay, CNRS, Centre Borelli, Gif-sur-Yvette, France

that one uses a numerical scheme capable of handling Monge-Ampère equations with arbitrary coefficients. Another benefit is that convergence results may be established using the theory of monotone schemes for degenerate elliptic partial differential equations [1]. Note however that establishing theoretical guarantees is more complicated when considering general cost functions, and in this paper our convergence proof only applies in the setting of a quadratic transport cost.

We design a monotone finite difference discretization of the Monge-Ampère equation

$$\det{}_+ \left(D^2 u(x) - A(x, Du(x))\right) = B(x, Du(x)) \quad \text{in } X, \tag{1}$$

where X is an open bounded subset of $\mathbb{R}^d$ containing the origin and A and B are bounded functions on $\overline{X} \times \mathbb{R}^d$, whose values are respectively symmetric matrices and nonnegative numbers, A and $B^{1/d}$ being Lipschitz continuous with respect to their second variables uniformly with respect to their first variables, and A being continuous with respect to both its variables. For any symmetric matrix M of size d, we denoted

$$\det{}_+ M := \begin{cases} \det M & \text{if } M \succeq 0, \\ -\infty & \text{else.} \end{cases}$$

(We use the Loewner order on the space of symmetric matrices: $M_1 \succeq M_2$ if $M_1 - M_2$ is positive semidefinite. From now on, we denote respectively by $\mathcal{S}_d$, $\mathcal{S}_d^+$, and $\mathcal{S}_d^{++}$ the sets of symmetric, symmetric positive semidefinite, and symmetric positive definite matrices of size d.)

Since we consider Monge-Ampère equations which are related to the problem of optimal transport, see section 5.1 and Remark 5.1, we also have to discretize the relevant boundary condition, described in section 1.2. We prove the *existence* of solutions, under suitable assumptions, to the proposed finite difference scheme. We also prove the *convergence* of solutions to this scheme, but only in the setting of *quadratic optimal transport*, where the function A is identically zero and the function B is separable in the form $B(x, p) = f(x)/g(p)$.

The Monge-Ampère equation is *degenerate elliptic*, meaning that it may be written in the form

$$F_{\mathrm{MA}}(x, Du(x), D^2 u(x)) = 0 \quad \text{in } X, \tag{2}$$

where the operator $F_{\mathrm{MA}} \colon \overline{X} \times \mathbb{R}^d \times \mathcal{S}_d \to \overline{\mathbb{R}}$ is *degenerate elliptic*, that is, nondecreasing with respect to its last variable: $F_{\mathrm{MA}}(x, p, M_1) \leq F_{\mathrm{MA}}(x, p, M_2)$ if $M_1 \succeq M_2$. The degenerate ellipticity property has a discrete counterpart which we call monotonicity, see Definition 2.5. Convergence of monotone schemes for degenerate elliptic equations may often be proved using a general argument, which was introduced in [1]. We use the fundamental part of this argument, see Theorem 2.7. As we discuss below Theorem 2.7, the full convergence result stated in [1] requires the approximated equation to satisfy a *strong comparison principle* which does not hold for the Monge-Ampère equation equipped with the boundary condition (24). Therefore, in order to prove Theorem 5.25, our convergence result in the setting of quadratic optimal transport, we need to establish an appropriate substitute to this comparison principle, in the form of Theorems 5.11 and 5.12.

One way to define the operator $F_{\mathrm{MA}}(x, p, M)$ so that it is both degenerate elliptic and consistent with (1) would be as

$$B(x, p) - \det{}_+(M - A(x, p)). \tag{3}$$

This is not the definition we use, however. The reason is that there is no obvious way to build a monotone scheme by directly discretizing (3).

Instead, we use strategies described in [35, 36] to reformulate the Monge-Ampère equation in the form (2), where F_{MA} is a supremum of semilinear operators (see also Proposition 5.8

for a more detailed description of what follows). First, note that formally, solutions to the Monge-Ampère equation satisfy the *admissibility* constraint

$$D^2u(x) \succeq A(x, Du(x)) \quad \text{in } X, \tag{4}$$

since otherwise the left-hand side in (1) would be equal to $-\infty$. For any symmetric positive semidefinite matrix M, it holds that

$$d(\det M)^{1/d} = \inf_{\substack{\mathcal{D}\in\mathcal{S}_d^{++}\\ \det\mathcal{D}=1}} \langle \mathcal{D}, M\rangle = \inf_{\substack{\mathcal{D}\in\mathcal{S}_d^{++}\\ \mathrm{Tr}(\mathcal{D})=1}} (\det \mathcal{D})^{-1/d}\langle \mathcal{D}, M\rangle, \tag{5}$$

where $\langle \mathcal{D}, M\rangle := \mathrm{Tr}(\mathcal{D}M)$. Choosing $M = D^2u(x) - A(x, Du(x))$ yields the two following reformulations of the Monge-Ampère equation (1):

$$B(x, Du(x)) - \inf_{\substack{\mathcal{D}\in\mathcal{S}_d^{++}\\ \det\mathcal{D}=1}} \left(\frac{\langle \mathcal{D}, D^2u(x) - A(x, Du(x))\rangle}{d}\right)^d = 0 \tag{6}$$

and alternatively, following [23],

$$\max_{\substack{\mathcal{D}\in\mathcal{S}_d^{+}\\ \mathrm{Tr}(\mathcal{D})=1}} L_{\mathcal{D}}\left(B(x, Du(x)), D^2u(x) - A(x, Du(x))\right) = 0 \quad \text{in } X, \tag{7}$$

where for any symmetric matrices $\mathcal{D}$ and M and nonnegative number b,

$$L_{\mathcal{D}}(b, M) := db^{1/d}(\det \mathcal{D})^{1/d} - \langle \mathcal{D}, M\rangle.$$

Note that the maximum in (7) is attained, as the maximum over a compact set of the continuous function $\mathcal{D} \mapsto L_{\mathcal{D}}(b, M)$ (this function is also concave, by the Minkowski determinant inequality). On the contrary, the parameter set of the infimum in (6) is not compact. Both reformulations enforce the admissibility constraint (4): for instance in (7), for any unit vector $e \in \mathbb{R}^d$, choosing $\mathcal{D} = e \otimes e$ in the maximum yields the inequality

$$\left\langle e, \left(D^2u(x) - A(x, Du(x))\right) e\right\rangle \geq 0,$$

from which it follows that $D^2u(x) \succeq A(x, Du(x))$.

The numerical scheme that we study in this paper is a discretization of (7). Hence we define the operator F_{MA} in (2) by

$$F_{\mathrm{MA}}(x, p, M) := \max_{\substack{\mathcal{D}\in\mathcal{S}_d^{+}\\ \mathrm{Tr}(\mathcal{D})=1}} L_{\mathcal{D}}(B(x, p), M - A(x, p)). \tag{8}$$

1.1 Discretization of the Monge-Ampère equation

For any discretization step $h > 0$, we discretize the operator F_{MA} on a grid $\mathcal{G}_h \subset X \cap h\mathbb{Z}^d$. Denoting by d_H the Hausdorff distance between compact subsets of $\mathbb{R}^d$, which we recall is defined by

$$d_H(K_1, K_2) := \max\left\{\max_{x\in K_1}\min_{y\in K_2}|x-y|, \max_{x\in K_2}\min_{y\in K_1}|x-y|\right\}, \tag{9}$$

we will assume that

$$\lim_{h\to 0} d_H\left(\partial X \cup ((X \cap h\mathbb{Z}^d) \setminus \mathcal{G}_h), \partial X\right) = 0, \tag{10}$$

or equivalently that if $K \subset X$ is compact, then for sufficiently small $h > 0$ one has $K \cap h\mathbb{Z}^2 \subset \mathcal{G}_h$. We will also need the technical assumption (42) of uniform connectedness of the grid $\mathcal{G}_h$.

Before introducing the discretization of F_{MA}, we need to define some finite difference operators. For any function $u\colon \mathcal{G}_h \to \mathbb{R}$, point $x \in \mathcal{G}_h$, and vector $e \in \mathbb{Z}^d$, we define

$$T_h^e u[x] := \begin{cases} u[x+he] & \text{if } x+he \in \mathcal{G}_h, \\ +\infty & \text{else,} \end{cases} \tag{11}$$

$$\delta_h^e u[x] := \frac{T_h^e u[x] - u[x]}{h}, \qquad \Delta_h^e u[x] := \frac{T_h^e u[x] + T_h^{-e} u[x] - 2u[x]}{h^2}. \tag{12}$$

The constant $+\infty$ in the definition of T_h^e is related to the way we recommend discretizing the optimal transport boundary condition, discussed in section 1.2.

In the whole paper, we denote by $(e_1, \dots, e_d)$ the canonical basis of $\mathbb{Z}^d$. For any function $u\colon \mathcal{G}_h \to \mathbb{R}$ and point $x \in \mathcal{G}_h$, we define the Laplacian approximation and, whenever it makes sense, the centered gradient approximation

$$\Delta_h u[x] := \sum_{i=1}^d \Delta_h^{e_i} u[x], \qquad D_h u[x] := \left(\frac{\delta_h^{e_i} u[x] - \delta_h^{-e_i} u[x]}{2}\right)_{1\le i\le d}. \tag{13}$$

We use Lax-Friedrichs approximations of the gradient of u in $A(x, Du(x))$ and $B(x, Du(x))$. To this end, we let $a_{\min} \le 0$, $a_{\mathrm{LF}} \ge 0$, and $b_{\mathrm{LF}} \ge 0$ be three constants independent of h. We will assume that for any $x \in \overline{X}$ and p, $p' \in \mathbb{R}^d$,

$$A(x,p) \succeq a_{\min} I_d, \tag{14}$$

$$|A(x,p) - A(x,p')|_2 \le a_{\mathrm{LF}} |p - p'|_1, \tag{15}$$

$$|B(x,p)^{1/d} - B(x,p')^{1/d}| \le b_{\mathrm{LF}} |p - p'|_1. \tag{16}$$

For any function $u\colon \mathcal{G}_h \to \mathbb{R}$, point $x \in \mathcal{G}_h$, and vector $e \in \mathbb{Z}^d$, we define

$$A_h^e u[x] := \begin{cases} a_{\min}|e|^2 \vee \left(\langle e, A(x, D_h u[x]) e\rangle - \frac{h}{2} a_{\mathrm{LF}} |e|^2 \Delta_h u[x]\right) & \text{if } \Delta_h u[x] < +\infty, \\ a_{\min}|e|^2 & \text{else,} \end{cases} \tag{17}$$

$$B_h u[x] := \begin{cases} 0 \vee \left(B(x, D_h u[x])^{1/d} - \frac{h}{2} b_{\mathrm{LF}} \Delta_h u[x]\right)^d & \text{if } \Delta_h u[x] < +\infty, \\ 0 & \text{else.} \end{cases} \tag{18}$$

(In the whole paper, we denote respectively by $a \vee b$ and $a \wedge b$ the maximum and the minimum of two real numbers a and b.) For any family $v = (v_i)_{1\le i\le I}$ of vectors of $\mathbb{Z}^d$ and any $\gamma \in \mathbb{R}^I$, we define

$$\mathcal{D}_v(\gamma) := \sum_{i=1}^I \gamma_i v_i \otimes v_i.$$

Finally, for any function $u\colon \mathcal{G}_h \to \mathbb{R}$, point $x \in \mathcal{G}_h$, and family v of vectors of $\mathbb{Z}^d$, we define

$$\Delta_h^v u[x] := (\Delta_h^e u[x])_{e\in v}, \qquad A_h^v u[x] := (A_h^e u[x])_{e\in v}. \tag{19}$$

For any $h > 0$, let V_h be a set of families of size $d(d+1)/2$ of vectors of $\mathbb{Z}^d$ such that

$$\lim_{h\to 0} d_H\left(\{\mathcal{D}_v(\gamma) \mid v \in V_h,\, \gamma \in \mathbb{R}_+^{d(d+1)/2},\, \mathrm{Tr}(\mathcal{D}_v(\gamma)) = 1\}, \{\mathcal{D} \in \mathcal{S}_d^+ \mid \mathrm{Tr}(\mathcal{D}) = 1\}\right) = 0. \quad (20)$$

Equivalently, if $K \subset \mathcal{S}_d^{++}$ is compact, then for sufficiently small $h > 0$ each element of K can be written as $\mathcal{D}_v(\gamma)$ where $v \in V_h$ and $\gamma \in \mathbb{R}_+^{d(d+1)/2}$. We will also need to assume that

$$\lim_{h\to 0} h \max_{v\in V_h} \max_{e\in v} |e| = 0, \quad (21)$$

and that for any $h > 0$,

$$e_1 \in \bigcup_{v\in V_h} \bigcup_{e\in v} \{\pm e\}, \quad (22)$$

where we recall that e_1 denotes the first vector of the canonical basis of $\mathbb{R}^d$. We discretize F_{MA} by the operator $S_{\mathrm{MA}}^h : \mathbb{R}^{\mathcal{G}_h} \to \overline{\mathbb{R}}^{\mathcal{G}_h}$ defined by

$$S_{\mathrm{MA}}^h u[x] := \max_{v\in V_h} \max_{\substack{\gamma\in\mathbb{R}_+^{d(d+1)/2} \\ \mathrm{Tr}(\mathcal{D}_v(\gamma))=1}} L_{v,\gamma}(B_h u[x], \Delta_h^v u[x] - A_h^v u[x]), \quad (23)$$

where for any family $v = (v_i)_{1\le i\le I}$ of vectors of $\mathbb{Z}^d$, $\gamma \in \mathbb{R}_+^I$, $b \ge 0$, and $m \in (\mathbb{R} \cup \{+\infty\})^I$,

$$L_{v,\gamma}(b, m) := db^{1/d}(\det \mathcal{D}_v(\gamma))^{1/d} - \langle \gamma, m\rangle.$$

Coefficients of γ are required to be nonnegative in order for the discretization to result in a numerical scheme which satisfies the monotonicity property (defined rigorously in Definition 2.12). Note that the constraint $\mathrm{Tr}(\mathcal{D}_v(\gamma)) = 1$ may be rewritten as $\sum_{i=1}^{d(d+1)/2} \gamma_i |v_i|^2 = 1$.

In dimension $d = 2$, we recommend choosing V_h as a set of superbases of $\mathbb{Z}^2$:

Definition 1.1. A pair $v = (v_1, v_2)$ of vectors of $\mathbb{Z}^2$ is a *basis* of $\mathbb{Z}^2$ if $\det(v_1, v_2) = \pm 1$. A triple $v = (v_1, v_2, v_3)$ of vectors of $\mathbb{Z}^2$ is a *superbase* of $\mathbb{Z}^2$ if $v_1 + v_2 + v_3 = 0$ and $\det(v_1, v_2) = \pm 1$.

Note that in the definition above, the constraint $\det(v_1, v_2) = \pm 1$ is equivalent to $\det(v_2, v_3) = \pm 1$ or $\det(v_1, v_3) = \pm 1$. We explain in Appendix B how a set V_h of superbases of $\mathbb{Z}^2$ satisfying the above assumptions may be constructed, using tools from the fields of lattice geometry and arithmetic known as the Selling's decomposition [43] and the Stern-Brocot tree [9]. We prove in section 4 that when choosing V_h in this way, the second maximum in (23) admits a closed-form expression, at least when no infinite values are involved (infinite values may stem from the handling of the boundary condition, see (11), and a simple modification of the formula of Theorem 1.2 allows to compute the maximum in this case, by excluding finite differences whose value is infinite):

Theorem 1.2. *If $v = (v_1, v_2)$ is a basis of $\mathbb{Z}^2$, then for any $b \ge 0$ and $m \in \mathbb{R}^2$,*

$$\max_{\substack{\gamma\in\mathbb{R}_+^2 \\ \mathrm{Tr}(\mathcal{D}_v(\gamma))=1}} L_{v,\gamma}(b, m) = \tilde{H}_v(b, m),$$

where

$$\tilde{H}_v(b, m) := \left(\frac{b}{|v_1|^2|v_2|^2} + \left(\frac{m_1}{2|v_1|^2} - \frac{m_2}{2|v_2|^2}\right)^2\right)^{1/2} - \frac{m_1}{2|v_1|^2} - \frac{m_2}{2|v_2|^2}.$$

If $v = (v_1, v_2, v_3)$ is a superbase of $\mathbb{Z}^2$, then for any $b \geq 0$ and $m \in \mathbb{R}^3$,

$$\max_{\substack{\gamma \in \mathbb{R}^3_+ \\ \mathrm{Tr}(\mathcal{D}_v(\gamma))=1}} L_{v,\gamma}(b,m) = H_v(b,m) \vee \max_{1 \leq i < j \leq 3} \tilde{H}_{(v_i,v_j)}(b,m),$$

where

$$H_v(b,m) := \begin{cases} (b + \langle m, Q_v m \rangle)^{1/2} + \langle w_v, m \rangle & \text{if } Q_v m + (b + \langle m, Q_v m \rangle)^{1/2} w_v <_{\mathrm{vec}} 0, \\ -\infty & \text{else,} \end{cases}$$

$$Q_v := \frac{1}{4} \begin{pmatrix} |v_2|^2 |v_3|^2 & \langle v_1, v_2 \rangle |v_3|^2 & \langle v_1, v_3 \rangle |v_2|^2 \\ \langle v_1, v_2 \rangle |v_3|^2 & |v_1|^2 |v_3|^2 & \langle v_2, v_3 \rangle |v_1|^2 \\ \langle v_1, v_3 \rangle |v_2|^2 & \langle v_2, v_3 \rangle |v_1|^2 & |v_1|^2 |v_2|^2 \end{pmatrix}, \qquad w_v := \frac{1}{2} \begin{pmatrix} \langle v_2, v_3 \rangle \\ \langle v_1, v_3 \rangle \\ \langle v_1, v_2 \rangle \end{pmatrix},$$

and, for $a \in \mathbb{R}^d$, we write $a <_{\mathrm{vec}} 0$ (respectively $a >_{\mathrm{vec}} 0$) if all components of a are negative (respectively positive).

1.2 Discretization of the boundary condition

In the setting of optimal transport, the relevant problem for the Monge-Ampère equation (1) is the *second boundary value problem*, which involves the *optimal transport boundary condition*

$$Du(x) \in \overline{P(x)}, \quad \forall x \in X, \tag{24}$$

where for any $x \in \overline{X}$, $P(x)$ is an open bounded convex nonempty subset of $\mathbb{R}^d$. We assume that $\overline{P(x)}$ depends continuously on x, for the Hausdorff distance d_H over compact subsets of $\mathbb{R}^d$ whose definition we recalled in (9). In the particular setting of quadratic optimal transport, in which we will prove convergence of the proposed numerical scheme, the set $P(x)$ does not depend on the variable x.

Note that despite being called a boundary condition, the constraint (24) involves the whole domain X. Some numerical approaches for solving the second boundary value problem, although not the one that we describe in this paper, rely on the fact that, in some cases, the constraint (24) can be reformulated in a way that only involves the boundary ∂X of the domain X, see for instance [4].

For now, let us consider the class of numerical schemes for equations (1) and (24) that are defined, for any discretization step $h > 0$, by an operator $S^h_{\mathrm{MABV2}} \colon \mathbb{R}^{\mathcal{G}_h} \to \overline{\mathbb{R}}^{\mathcal{G}_h}$, and may be written as

$$S^h_{\mathrm{MABV2}} u[x] = 0 \quad \text{in } \mathcal{G}_h. \tag{25}$$

One property of equations (1) and (24) is that their expressions depend only on derivatives of the function u and not on u itself, and therefore that the set of solutions is stable by addition of a constant. Accordingly, we say that the operator S^h_{MABV2} and the scheme (25) are *additively invariant* if for any function $u \colon \mathcal{G}_h \to \mathbb{R}$ and real number ξ, $S^h_{\mathrm{MABV2}}(u + \xi) = S^h_{\mathrm{MABV2}} u$.

We adapt the approach introduced in [26] to build an operator S^h_{MABV2} suitable for (25). The idea is to build S^h_{MABV2} as a maximum of S^h_{MA} and of a monotone discretization $S^h_{\mathrm{BV2}} \colon \mathbb{R}^{\mathcal{G}_h} \to \overline{\mathbb{R}}^{\mathcal{G}_h}$ of the left-hand side in a degenerate elliptic formulation of (24).

We use the following formulation of (24), initially introduced in [5]:

$$F_{\mathrm{BV2}}(x, Du(x)) \leq 0 \quad \text{in } X, \tag{26}$$

where $F_{\mathrm{BV2}}\colon \overline{X} \times \mathbb{R}^d \to \mathbb{R}$ is defined by

$$F_{\mathrm{BV2}}(x,p) := \max_{|e|=1} (\langle e,p\rangle - \sigma_{P(x)}(e)). \tag{27}$$

(We denote by $\sigma_{P(x)}$ the support function of the convex set $P(x)$: for any $e \in \mathbb{R}^d$, $\sigma_{P(x)}(e) := \sup_{p\in P(x)}\langle e,p\rangle$. Formally, if p belongs to the boundary $\partial P(x)$ of $P(x)$, then the maximum in the definition of F_{BV2} is attained when e is the unit outer normal of $\partial P(x)$ at the point p.)

For any function $u\colon \mathcal{G}_h \to \mathbb{R}$, point $x \in \mathcal{G}_h$, and vector $e \in \mathbb{R}^d$, we define the upwind finite difference

$$D_h^e u[x] := \sum_{i=1}^d ((0 \wedge \langle e,e_i\rangle)\delta_h^{e_i} u[x] - (0 \vee \langle e,e_i\rangle)\delta_h^{-e_i} u[x]),$$

using the convention $0 \times (+\infty) = 0$ (this convention is only needed in the immediate neighborhood of ∂X, where $\delta_h^{\pm e_i} u[x]$ may take infinite values). Then we define S^h_{BV2} and S^h_{MABV2} as

$$\begin{aligned} S^h_{\mathrm{BV2}} u[x] &:= \max_{|e|=1} (D_h^e u[x] - \sigma_{P(x)}(e)), \\ S^h_{\mathrm{MABV2}} u[x] &:= S^h_{\mathrm{MA}} u[x] \vee S^h_{\mathrm{BV2}} u[x]. \end{aligned} \tag{28}$$

In this setting, the scheme (25) is additively invariant.

Additively invariant schemes of the form (25) are not well-posed: their sets of solutions are stable by addition of a constant, thus not a singleton. Moreover they often have no solutions. One way to see this formally is that a well-posed scheme would need an additional equality to guarantee uniqueness of solutions, for instance $u[0] = 0$, but that then there would be one more equality than unknowns in the scheme. In the continuous setting, equations whose sets of solutions are stable by addition of a constant often admit solutions if and only if their coefficients satisfy some nonlocal condition, such as the mass balance condition (56) in the case of the Monge-Ampère equation of optimal transport; however, there may be no obvious discrete counterpart to this condition. See section 2 for further discussion of this issue.

In order to get around this difficulty, we solve an altered form of the scheme (25), following the approach used in the numerical experiments of [4] (note that we present a fully detailed mathematical analysis of this alteration, in contrast with [4] where it was introduced essentially as a numerical trick). We add an unknown α to the scheme, which must be a real number. For fixed α, we define the operators $S^{h,\alpha}_{\mathrm{MA}}\colon \mathbb{R}^{\mathcal{G}_h} \to \overline{\mathbb{R}}^{\mathcal{G}_h}$ and $S^{h,\alpha}_{\mathrm{MABV2}}\colon \mathbb{R}^{\mathcal{G}_h} \to \overline{\mathbb{R}}^{\mathcal{G}_h}$ as

$$S^{h,\alpha}_{\mathrm{MA}} u[x] := S^h_{\mathrm{MA}} u[x] + \alpha, \qquad S^{h,\alpha}_{\mathrm{MABV2}} u[x] := S^{h,\alpha}_{\mathrm{MA}} u[x] \vee S^h_{\mathrm{BV2}} u[x]. \tag{29}$$

The scheme that we actually solve, with respect to the extended unknown (α, u), is

$$S^{h,\alpha}_{\mathrm{MABV2}} u[x] = 0 \quad \text{in } \mathcal{G}_h. \tag{30}$$

1.3 Main contributions and relation to previous works

We introduce the numerical scheme (30) for the Monge-Ampère equation (1), equipped with the boundary condition (24). We prove the existence of solutions to a class of monotone additively invariant numerical schemes featuring an additional unknown $\alpha \in \mathbb{R}$ as in (30), see section 2, and we show, in section 3, that the scheme (30) belongs to this class. This scheme is based on a discretization of the reformulation (7) of the Monge-Ampère equation. We prove in section 4 that this discretization admits a closed-form expression, as stated in Theorem 1.2. We prove convergence of the scheme in the setting of quadratic optimal transport, see section 5; convergence

in the setting of more general optimal transport problems remains an open problem. We present in section 6 some numerical experiments, including an application to the far field refractor problem in nonimaging optics.

The closed-form expression obtained in Theorem 1.2 makes the implementation of the scheme particularly efficient, since no discretization of the parameter set of the maximum in (7) is needed. While to our knowledge the proposed discretization is the first one to admit such a closed-form expression among those that are based on the reformulation (7) of the Monge-Ampère equation, it is to be related to the MA-LBR scheme, introduced in [3] in the setting of the Dirichlet problem for the Monge-Ampère equation when the function A is identically zero, and to the scheme we introduced in [8] for the Pucci equation. Both of the above-mentioned schemes involve the notion of superbases of $\mathbb{Z}^2$. We prove in Appendix A that the MA-LBR scheme is a discretization of (6), although it was not introduced as such in [3].

As opposed to (6), the reformulation (7) has the benefit that its left-hand side remains finite even when (4) is not satisfied. Thus schemes based on it are more stable numerically than those based on (6), and can handle the degenerate case of functions $B\colon \overline{X} \times \mathbb{R}^d \to \mathbb{R}_+$ which are not everywhere positive, in which case the solutions to the Monge-Ampère problem typically satisfy (4) but not its strict variant. On the contrary, the MA-LBR scheme only applies in the case $B > 0$, and in addition solving it using the damped Newton requires using extremely small steps so that the constraint (4) remains satisfied along the iterations. This behavior of the Newton method is illustrated numerically in section 6.4, and does not occur with the scheme introduced in this paper.

Numerical schemes based on (7) were previously introduced in [23], and then in [14], although only in the setting of the Dirichlet problem for the Monge-Ampère equation when $A = 0$. In those papers, no counterpart of Theorem 1.2 was proved, hence the parameter set of the maximum in (7) had to be discretized.

Convergence of schemes for the second boundary value problem was previously studied in [4] and in [26] in the setting of the quadratic optimal transport problem. Schemes considered in those two papers were based on the MA-LBR scheme introduced in [3], and adapted in order to discretize the boundary condition (24).

In [4], convergence of a scheme of the form (25) was proved, but existence of solutions to this scheme was not. It turns out that solutions typically do not exist, due to the scheme being additively invariant. The approach used to solve the scheme in the numerical experiments was equivalent to adding an unknown $\alpha \in \mathbb{R}$ as in (30), but the proof of convergence was not extended to this setting.

Remark 1.3 (Applicability of Theorem 2.15 to the scheme in [4]). The work [4] establishes the convergence of the solutions to a discretization of the optimal transport problem, under the assumption that they exist. The latter point is dubious, as discussed above and acknowledged by the authors of [4], and for that reason an altered variant of the scheme is considered in the numerical experiments section, featuring an additional unknown which is analogous to the parameter $\alpha \in \mathbb{R}$ in (29); the existence of solutions to this variant is observed numerically in [4], but left as an open problem from a theoretical standpoint.

While the detailed analysis of the scheme in [4] is out of the scope of this paper, let us discuss the applicability to this scheme of the assumptions of our existence result, Theorem 2.15. Those assumptions are continuity, monotonicity and stability, in the sense of Definition 2.12. The continuity of the scheme in [4] is easy to prove, since no infinite values are involved in the definition of the scheme operator, contrary to the scheme that we introduce in this paper. The monotonicity property is not satisfied by the scheme recommended by default in [4] due to the centered discretization of the gradient of the unknown u; however, a monotone Lax-Friedrichs discretization was described as an alternative in [4, section 4.4]. The remaining open question

is the stability of the scheme in [4], again in the sense of Definition 2.12; we could not see an immediate proof of this property, but one could expect to develop one based on the sketches of the proofs of Proposition 3.6 and also of [4, Proposition 4.3, item (5)], which is a result about the Lipschitz continuity of the solutions to this scheme.

Another scheme of the form (25) is studied in [26]. In that work, a Dirichlet boundary condition is enforced on ∂X, which in our setting would amount to replacing $+\infty$ with some fixed constant $C \in \mathbb{R}$ in (11). Therefore the scheme considered in [26] is not additively invariant. The Dirichlet boundary condition is to be understood in a weak sense (the one of viscosity solutions, see Definition 2.3). It may formally be simplified to $u(x) \leq C$ on ∂X, with equality at some point $x_* \in \partial X$. An important assumption in [26] is that the scheme satisfies a property of *underestimation*, discussed in Remark 1.4. Under this assumption, the existence and convergence of solutions is proved. The property of underestimation is satisfied in the case of quadratic optimal transport at the cost of a careful handling of the constraint (24), but it does not seem obvious that it is satisfied for similar schemes in the case of more general optimal transport problems, with $A \neq 0$ in (1). No numerical experiments were performed in [26]. In our experience, the scheme introduced in that paper has the drawback that the numerical error of its solutions tends to be unevenly distributed. This effect is related to the particular role played in the discretization by the point $x_* \in \partial X$ where the Dirichlet condition is satisfied in the classical sense, which leads to numerical artifacts and tends to decrease the accuracy of the scheme.

In our proof of convergence of the scheme (30), we use the arguments introduced in [26] when appropriate. However, the property of underestimation is not required in our setting.

Remark 1.4 (Role of the property of underestimation in [4,26], and substitutes used in this paper). In each of the papers [4,26], the theoretical analysis of the considered scheme uses the fact that this scheme satisfies some property of *underestimation*. The schemes considered in those papers are both based on the MA-LBR discretization [3], represented in Appendix A by the operator Λ_h in (76), of the Monge-Ampère operator $u \mapsto \det D^2 u(\cdot)$.

The property of underestimation used in [4] is formulated as the fact that the operator Λ_h overestimates the Lebesgue measure of the subgradient, in the sense that $\Lambda_h u[x] \geq h^{-d} |\partial \tilde{u}(x)|$ for any function $u \colon \mathcal{G}_h \to \mathbb{R}$ and for any suitable point $x \in \mathcal{G}_h$, where $\tilde{u} \colon \mathbb{R}^d \to \overline{\mathbb{R}}$ denotes the convex envelope of u. This property is described in more detail in [4, Lemma 4.2]. Whether this property could be extended to the setting of the scheme (30) considered in this paper is not clear, since this scheme is based on a discretization of the reformulation (2) of the Monge-Ampère equation, which does not feature directly the Monge-Ampère operator $u \mapsto \det D^2 u(\cdot)$. The arguments in the convergence analysis in [4] that use the property of underestimation are based on the construction of solutions to semi-discrete optimal transport problems, and are completely different from the arguments in this paper, which are based on the theory of convergence of monotone schemes to viscosity solutions to degenerate elliptic equations.

In [26], two distinct definitions of the property of underestimation are given. The first one [26, Definition 3.8] is similar to the definition in [4]. The second one [26, Remark 3.9] asks that $S^h_{\mathrm{MABV2}} u[x] \leq F_{\mathrm{MABV2}} u(x)$ in $\mathcal{G}_h$, for all smooth convex functions u, where S^h_{MABV2} is the discrete operator describing the whole scheme and F_{MABV2} is its continuous counterpart. The second definition is claimed to correctly approximate the first one at small grid scales.

The property of underestimation is not only used in the convergence analysis in [26], but it is also crucial for the proof [26, Lemma 3.12], which guarantees the existence of a subsolution to the scheme in [26] and is an intermediary step for proving the existence of solutions. One technique that is often used to build a subsolution to a scheme is to consider a strict subsolution to the continuous problem, and to use the consistency of the scheme to show that it is also a subsolution to the scheme. However in the setting of [26] no strict subsolutions to the continuous problem may exist, as shown by some counterpart to Theorem 5.11 in this setting (with $\alpha = 0$,

see [26, Theorem 2.1]). Formally, the property of underestimation allows one to consider a solution to the continuous problem instead of a strict subsolution in the argument above (in [26], a slight variation of the solution, constructed by solving a semi-discrete optimal transport problem, is considered instead for technical reasons). In our setting, Theorem 5.11 does not prevent us to build a strict subsolution to the continuous problem for some value $\alpha < 0$, which is sufficient in order to apply our existence result Theorem 2.15, hence we have no need for the underestimation property at this stage.

Let us finally discuss why the property of underestimation is needed in the proof of the convergence result [26, Theorem 3.11]. In the standard theory of convergence of monotone schemes [1], it is shown that some appropriately defined upper and lower limits $\overline{u}$ and $\underline{u}$ of sequences of solutions to the scheme are respectively a subsolution and supersolution to the continuous problem (see Theorem 2.7). In [26], the counterpart [26, Theorem 2.1] to Theorem 5.11 is used to deduce that the subsolution $\overline{u}$ is actually a solution to the Monge-Ampère problem. From this point, it remains to prove that $\underline{u} = \overline{u}$. In the general setting of monotone schemes, this is often done by using a comparison principle for the continuous problem, but such a comparison principle is lacking in the setting of [26]. Another strategy is to prove that the solutions to the scheme are sufficiently regular so that the limits $\overline{u}$ and $\underline{u}$ coincide by definition, at least up to extraction of a subsequence: this is what we do in this paper, see our stability result Proposition 3.6. In [26], no such stability result is proved, and the arguments used instead to show that $\underline{u} = \overline{u}$ rely on the assumption that the scheme satisfies the property of underestimation, indirectly through the use of the subsolutions to the scheme built in the previous paragraph.

Note that the scheme (30), and its continuous counterpart (43) below, which both feature an additional unknown or parameter $\alpha \in \mathbb{R}$, fit in the framework of eigenvalue problems recently studied in [27]. Although our proof of convergence only applies to Monge-Ampère equation in the setting of quadratic optimal transport, our existence result, Theorem 2.15, is applicable to other such eigenvalue problems, as illustrated by the examples in section 2.

2 Monotone additively invariant schemes

2.1 Degenerate elliptic additively invariant equations

In this section, we study numerical schemes for a general degenerate elliptic equation of the form

$$F(x, Du(x), D^2u(x)) = 0 \quad \text{in } \overline{X}. \tag{31}$$

Typically, F is discontinuous and $F(x, p, M)$ is defined differently depending on whether x belongs to X or to ∂X, in order to take into account the boundary condition in equation (31). The equation without the boundary condition would then be

$$F(x, Du(x), D^2u(x)) = 0 \quad \text{in } X. \tag{32}$$

Let us recall the definition of degenerate ellipticity:

Definition 2.1 (Degenerate ellipticity)**.** The operator $F\colon \overline{X} \times \mathbb{R}^d \times \mathcal{S}_d \to \overline{\mathbb{R}}$, and the equations (31) and (32), are *degenerate elliptic* if F is nonincreasing with respect to its last variable for the Loewner order: $F(x, p, M_1) \leq F(x, p, M_2)$ if $M_1 \succeq M_2$.

We say that equations (31) and (32) are *additively invariant* since, for reasonable notions of solutions, their sets of solutions are stable by addition of a constant, due to the fact that at any point x, the left-hand sides of those equations depend only on the derivatives $Du(x)$ and

$D^2u(x)$ of the function u, and not on its value $u(x)$. This is not a standard property, and we will show that it is a source of difficulty in the design of monotone numerical schemes. Typically, an additively invariant equation only has solutions if its coefficients are well-chosen and satisfy a particular nonlocal property.

Example 2.2. Throughout this section, we illustrate our definitions and results with Poisson's equation on the one-dimensional domain $X = (-1, 1)$, with the zero Neumann boundary condition:

$$\begin{cases} u''(x) = \psi(x) & \text{in } (-1,1), \\ u'(-1) = u'(1) = 0, \end{cases}$$

where $\psi\colon [-1,1] \to \mathbb{R}$ is an integrable function. We write this equation in the form

$$F_{\mathrm{ex}}(x, u'(x), u''(x)) = 0 \quad \text{in } [-1,1], \tag{33}$$

where the degenerate elliptic operator $F_{\mathrm{ex}}\colon [-1,1] \times \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ is defined by

$$F_{\mathrm{ex}}(x,p,m) := \begin{cases} -p & \text{if } x = -1, \\ p & \text{if } x = 1, \\ \psi(x) - m & \text{else.} \end{cases}$$

(The choice $F_{\mathrm{ex}}(x,p,m) = \psi(x) - m$, rather than $F_{\mathrm{ex}}(x,p,m) = m - \psi(x)$, is required in order for the equation (33) to be degenerate elliptic. The choice $F_{\mathrm{ex}}(-1,p,m) = -p$ and $F_{\mathrm{ex}}(1,p,m) = p$, rather than $F_{\mathrm{ex}}(-1,p,m) = p$ and $F_{\mathrm{ex}}(1,p,m) = -p$, is not dictated by the degenerate ellipticity property, but is the standard formulation of Neumann boundary conditions for degenerate elliptic equations, and the one which allows a comparison principle to hold in the context of more favorable equations such as $\psi - u'' + u = 0$, see [16, section 7.B].) The equation (33) only has solutions (respectively subsolutions, supersolutions) if $\int_{-1}^{1} \psi(x)\,dx = 0$ (respectively ≤ 0, ≥ 0), which we assume. Notice the similarity with the mass balance condition (56) which occurs in the setting of optimal transport.

An appropriate notion of solutions for degenerate elliptic equations, and for the study of discretizations of such equations, is the one of *viscosity solutions*. Before defining them, let us recall the definitions of the upper semicontinuous envelope F^* and lower semicontinuous envelope F_* of a function $F\colon E \to \overline{\mathbb{R}}$, E being a subset of $\mathbb{R}^n$: for any $x \in \overline{E}$,

$$F^*(x) := \limsup_{x' \to x} F(x), \qquad\qquad F_*(x) := \liminf_{x' \to x} F(x).$$

Definition 2.3 (Viscosity solution)**.** A function $u\colon \overline{X} \to \mathbb{R}$ is a *viscosity subsolution* to (31) if (i) it is upper semicontinuous and (ii) for any function φ in $C^2(\overline{X})$ and local maximum x of $u - \varphi$ in $\overline{X}$,

$$F_*(x, D\varphi(x), D^2\varphi(x)) \le 0.$$

It is a *viscosity supersolution* if (i) it is lower semicontinuous and (ii) for any function φ in $C^2(\overline{X})$ and local minimum x of $u - \varphi$ in $\overline{X}$,

$$F^*(x, D\varphi(x), D^2\varphi(x)) \ge 0.$$

It is a *viscosity solution* if it is both a viscosity subsolution and a viscosity supersolution. The same definitions, with $\overline{X}$ replaced by X, apply to equation (32).

Note that if a viscosity subsolution (respectively supersolution) u to (31) is twice differentiable at some point $x \in \overline{X}$ and if $F_*(x, Du(x), D^2u(x)) = F^*(x, Du(x), D^2u(x))$, then u is a classical subsolution (respectively supersolution) to (31) at the point x.

2.2 Discretization

For any discretization step $h > 0$, let $\mathcal{G}_h$ be a finite nonempty subset of $\overline{X}$ containing the origin. In the rest of this paper, it is required that $\mathcal{G}_h$ be a subset of the Cartesian grid $X \cap h\mathbb{Z}^d$; however, this is not necessary in this section. What will be required in our definition of consistency is that

$$\lim_{h \to 0} d_H(\mathcal{G}_h, X) = 0. \tag{34}$$

Note that *in the case* that $\mathcal{G}_h$ is included in $X \cap h\mathbb{Z}^d$, then (34) is implied by (10).

We represent discretizations of the operator F by operators $S\colon \mathbb{R}^{\mathcal{G}_h} \to \overline{\mathbb{R}}^{\mathcal{G}_h}$ that are additively invariant, according to the following definition:

Definition 2.4. An operator $S\colon \mathbb{R}^{\mathcal{G}_h} \to \overline{\mathbb{R}}^{\mathcal{G}_h}$ is *additively invariant* if for any $u\colon \mathcal{G}_h \to \mathbb{R}$, $\xi \in \mathbb{R}$, and $x \in \mathcal{G}_h$, it holds that

$$S(u + \xi)[x] = Su[x].$$

For now, we let $S^h\colon \mathbb{R}^{\mathcal{G}_h} \to \overline{\mathbb{R}}^{\mathcal{G}_h}$ be an additively invariant operator, for any $h > 0$, and we consider a numerical scheme of the form

$$S^h u[x] = 0 \quad \text{in } \mathcal{G}_h. \tag{35}$$

Definition 2.5. The scheme (35) is:

- *Monotone* if for any $h > 0$, $x \in \mathcal{G}_h$, and $\overline{u}$, $\underline{u}\colon \mathcal{G}_h \to \mathbb{R}$ such that $\overline{u}[x] = \underline{u}[x]$ and $\overline{u} \geq \underline{u}$ in $\mathcal{G}_h$, it holds that $S^h\overline{u}[x] \leq S^h\underline{u}[x]$.
- *Consistent* with equation (31) if (34) holds and for any $\varphi \in C^\infty(\overline{X})$ and $x \in \overline{X}$,

$$\limsup_{\substack{h>0,\, h\to 0 \\ x' \in \mathcal{G}_h,\, x' \to x}} S^h\varphi[x'] \leq F^*(x, D\varphi(x), D^2\varphi(x)),$$

$$\liminf_{\substack{h>0,\, h\to 0 \\ x' \in \mathcal{G}_h,\, x' \to x}} S^h\varphi[x'] \geq F_*(x, D\varphi(x), D^2\varphi(x)).$$

Remark 2.6. Schemes of the form (35) are typically called *degenerate elliptic* if for any $h > 0$, $x \in \mathcal{G}_h$, and $\overline{u}$, $\underline{u}\colon \mathcal{G}_h \to \mathbb{R}$ such that $\overline{u}[x] \leq \underline{u}[x]$ (rather than $\overline{u}[x] = \underline{u}[x]$ in Definition 2.5) and $\overline{u} \geq \underline{u}$ in $\mathcal{G}_h \setminus \{x\}$, it holds that $S^h\overline{u}[x] \leq S^h\underline{u}[x]$. In our setting, monotonicity and degenerate ellipticity are equivalent, since the operators S^h are additively invariant.

A framework is outlined in [1] for the proof of convergence of monotone schemes. The following fundamental result follows directly from the proof of [1, Theorem 2.1]:

Theorem 2.7. *Assume that there exist a sequence $(h_n)_{n\in\mathbb{N}}$ of discretization steps $h_n > 0$ converging to zero and a sequence $(u_n)_{n\in\mathbb{N}}$ of solutions $u_n\colon \mathcal{G}_{h_n} \to \mathbb{R}$ to (35) with $h = h_n$ such that $u_n[x]$ is bounded, uniformly over $n \in \mathbb{N}$ and $x \in \mathcal{G}_{h_n}$. If (35) is monotone and consistent with equation (31), then functions $\overline{u}$, $\underline{u}\colon \overline{X} \to \mathbb{R}$ defined by*

$$\overline{u}(x) := \limsup_{\substack{n\in\mathbb{N},\, n\to+\infty \\ x' \in \mathcal{G}_{h_n},\, x' \to x}} u_n[x'], \qquad \underline{u}(x) := \liminf_{\substack{n\in\mathbb{N},\, n\to+\infty \\ x' \in \mathcal{G}_{h_n},\, x' \to x}} u_n[x'], \tag{36}$$

are respectively a viscosity subsolution and supersolution to (31).

The definition of consistency in Definition 2.5 is slightly simpler than the one in [1], due to the assumption that operators S^h are additively invariant. In the framework of [1], in which the left-hand side in (31) may also depend on $u(x)$, a *strong comparison principle*, that is, a result stating that viscosity subsolutions to (31) are always less than viscosity supersolutions, is used after applying Theorem 2.7 to prove that $\overline{u} \leq \underline{u}$, which allows to conclude that $\overline{u} = \underline{u}$, since $\overline{u} \geq \underline{u}$ by definition. Obviously, no strong comparison principle may hold if the set of viscosity solutions is nonempty and stable by addition of a constant. In our proof of convergence in the setting of quadratic optimal transport, we use Theorems 5.11 and 5.12 as a substitute to this comparison principle.

An important difficulty that we encounter is that numerical schemes of the form (35) typically have no solutions.

Example 2.8. Let $X = [-1, 1]$. For any $h > 0$, we let $\tilde{h} := \lceil h^{-1} \rceil^{-1}$, $\mathcal{G}_h := [-1, 1] \cap \tilde{h}\mathbb{Z}$, and we define the additively invariant operator $S^h_{\mathrm{ex}} \colon \mathbb{R}^{\mathcal{G}_h} \to \mathbb{R}^{\mathcal{G}_h}$ by

$$S^h_{\mathrm{ex}} u[x] := \begin{cases} (u[-1] - u[-1 + \tilde{h}])/\tilde{h} & \text{if } x = -1, \\ (u[1] - u[1 - \tilde{h}])/\tilde{h} & \text{if } x = 1, \\ \psi(x) - (u[x + \tilde{h}] + u[x - \tilde{h}] - 2u[x])/\tilde{h}^2 & \text{else.} \end{cases}$$

Then the scheme

$$S^h_{\mathrm{ex}} u[x] = 0 \quad \text{in } \mathcal{G}_h$$

is monotone and consistent with equation (33). Solving this scheme is equivalent to solving a square linear system, since the scheme operator $S^h_{\mathrm{ex}} \colon \mathbb{R}^{\mathcal{G}_h} \to \mathbb{R}^{\mathcal{G}_h}$ is an affine map. However, this linear system is noninvertible, since all constant functions belong to the kernel of the associated linear operator.

To get around this difficulty, we add a parameter $\alpha \in \mathbb{R}$ to the equation (31), yielding a new equation

$$F^\alpha(x, Du(x), D^2 u(x)) = 0 \quad \text{in } \overline{X}, \tag{37}$$

where for any $\alpha \in \mathbb{R}$, $F^\alpha \colon \overline{X} \times \mathbb{R}^d \times \mathcal{S}_d \to \overline{\mathbb{R}}$ is a given operator, typically degenerate elliptic. The idea is to choose F^α so that $F^0 = F$ and (37) has no viscosity subsolutions when $\alpha > 0$ and no viscosity supersolutions when $\alpha < 0$.

Example 2.9. For any $\alpha \in \mathbb{R}$, we define $F^\alpha_{\mathrm{ex}} \colon [-1, 1] \times \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ by

$$F^\alpha_{\mathrm{ex}}(x, p, m) := \begin{cases} -p & \text{if } x = -1, \\ p & \text{if } x = 1, \\ \psi(x) - m + \alpha & \text{else.} \end{cases}$$

Then equation

$$F^\alpha_{\mathrm{ex}}(x, u'(x), u''(x)) = 0 \quad \text{in } \overline{X}$$

coincides with (33) when $\alpha = 0$, and only has solutions (respectively subsolutions, supersolutions) if $\int_{-1}^1 \psi(x)\, dx = -2\alpha$ (respectively $\leq -2\alpha$, $\geq -2\alpha$). Recall that we assumed that $\int_{-1}^1 \psi(x)\, dx = 0$.

Accordingly, we add an unknown $\alpha \in \mathbb{R}$ to the numerical scheme. For any $h > 0$ and $\alpha \in \mathbb{R}$, we let $S^{h,\alpha} \colon \mathbb{R}^{\mathcal{G}_h} \to \overline{\mathbb{R}}^{\mathcal{G}_h}$ be an additively invariant operator, and we consider the scheme

$$S^{h,\alpha} u[x] = 0 \quad \text{in } \mathcal{G}_h. \tag{38}$$

Example 2.10. In the setting of Example 2.8, for any $h > 0$ and $\alpha \in \mathbb{R}$, we define $S_{\mathrm{ex}}^{h,\alpha} \colon \mathbb{R}^{\mathcal{G}_h} \to \mathbb{R}^{\mathcal{G}_h}$ by

$$S_{\mathrm{ex}}^{h,\alpha} u[x] := \begin{cases} (u[-1] - u[-1 + \tilde{h}])/\tilde{h} & \text{if } x = -1, \\ (u[1] - u[1 - \tilde{h}])/\tilde{h} & \text{if } x = 1, \\ \psi(x) - (u[x + \tilde{h}] + u[x - \tilde{h}] - 2u[x])/\tilde{h}^2 + \alpha & \text{else} \end{cases}$$

(recall that $\tilde{h} := \lceil h^{-1} \rceil^{-1}$). Then a solution $(\alpha, u) \in \mathbb{R} \times \mathbb{R}^{\mathcal{G}_h}$ to the scheme

$$S_{\mathrm{ex}}^{h,\alpha} u[x] = 0 \quad \text{in } \mathcal{G}_h$$

may easily be constructed explicitly.

The definition of solutions $(\alpha, u) \in \mathbb{R} \times \mathbb{R}^{\mathcal{G}_h}$ to (38) is obvious, but we will also need a notion of subsolutions (we could define supersolutions similarly, but this will not be needed):

Definition 2.11 (Subsolution)**.** Let $h > 0$. A pair $(\alpha, u) \in \mathbb{R} \times \mathbb{R}^{\mathcal{G}_h}$ is a *subsolution* to (38) if $S^{h,\alpha} u[x] \leq 0$ in $\mathcal{G}_h$.

Since α is an unknown of the scheme, and not simply a fixed parameter, Definition 2.5 needs to be adapted to this new setting. We also define some other properties that the scheme (38) may satisfy. Conceptually, the following definition is intended for schemes such that $S^{h,\alpha} u[x]$ is nondecreasing with respect to α.

Definition 2.12. The scheme (38) is:

- *Monotone* if for any $\alpha \in \mathbb{R}$, the scheme (35) with $S^h = S^{h,\alpha}$ is monotone in the sense of Definition 2.5.
- *Consistent* with the parametrized equation (37) if for any family of real numbers $(\alpha_h)_{h>0}$ converging to some $\alpha \in \mathbb{R}$ as h approaches zero, the scheme (35) with $S^h = S^{h,\alpha_h}$ is consistent with equation (37) in the sense of Definition 2.5.
- *Continuous* if for any small $h > 0$, the map $\mathbb{R} \times \mathbb{R}^{\mathcal{G}_h} \to \overline{\mathbb{R}}^{\mathcal{G}_h}$, $(\alpha, u) \mapsto S^{h,\alpha} u$ takes finite values and is continuous.
- *Stable* if the following properties hold:
  - (i) For any small $h > 0$, there exists a subsolution $(\alpha, u) \in \mathbb{R} \times \mathbb{R}^{\mathcal{G}_h}$ to (38).
  - (ii) There exists a nonincreasing function $\omega \colon \mathbb{R} \to \mathbb{R}_+$ such that for any small $h > 0$, any subsolution $(\alpha, u) \in \mathbb{R} \times \mathbb{R}^{\mathcal{G}_h}$ to (38), and any x_1, $x_2 \in \mathcal{G}_h$, one has
  $$|u[x_1] - u[x_2]| \leq \omega(\alpha).$$
  - (iii) There exists $\alpha_0 \in \mathbb{R}$ such that for any small $h > 0$ and any subsolution $(\alpha, u) \in \mathbb{R} \times \mathbb{R}^{\mathcal{G}_h}$ to (38), one has $\alpha \leq \alpha_0$.
  - (iv) There exists $\alpha_1 \in \mathbb{R}$ such that for any small $h > 0$ and any solution $(\alpha, u) \in \mathbb{R} \times \mathbb{R}^{\mathcal{G}_h}$ to (38), one has $\alpha \geq \alpha_1$.
- *Equicontinuously stable* if it satisfies all items in the definition of stability above, with (ii) replaced by the following:
  - (ii') There exists a function $\omega \colon \mathbb{R} \times \mathbb{R}_+ \to \mathbb{R}_+$, nonincreasing with respect to its first variable and satisfying $\lim_{t \to 0} \omega(\alpha, t) = 0$ for any $\alpha \in \mathbb{R}$, such that for any small $h > 0$, any subsolution $(\alpha, u) \in \mathbb{R} \times \mathbb{R}^{\mathcal{G}_h}$ to (38), and any x_1, $x_2 \in \mathcal{G}_h$, one has
  $$|u[x_1] - u[x_2]| \leq \omega(\alpha, |x_1 - x_2|).$$

Note that the value $\alpha = 0$ does not play a special role in Definition 2.12. The role of the functions ω in the definitions of stability and equicontinuous stability is to allow schemes to become unstable when $\alpha \to -\infty$.

Obviously, if (38) is equicontinuously stable, then it is stable. In the case of the scheme considered in this paper for the Monge-Ampère equation, subsolutions will be established to be uniformly Lipschitz continuous, which is stronger than equicontinuity, see the proof of Proposition 3.6. In particular, the boundary condition $u(x) - \infty = 0$ on ∂X (to be understood in the viscosity sense, as mentioned in section 1) does not induce a boundary layer.

Theorem 2.7 is easily adapted to the scheme (38):

Corollary 2.13. *Assume that there exist a sequence $(h_n)_{n\in\mathbb{N}}$ of discretization steps $h_n > 0$ converging to zero, a sequence $(\alpha_n)_{n\in\mathbb{N}}$ of real numbers α_n converging to some $\alpha \in \mathbb{R}$, and a sequence $(u_n)_{n\in\mathbb{N}}$ of functions $u_n \colon \mathcal{G}_{h_n} \to \mathbb{R}$ such that (α_n, u_n) is solution to (38) with $h = h_n$ and $u_n[x]$ is bounded, uniformly over $n \in \mathbb{N}$ and $x \in \mathcal{G}_{h_n}$. If (38) is monotone and consistent with (37), then limits superior and inferior $\overline{u}$, $\underline{u} \colon \overline{X} \to \mathbb{R}$ defined as in (36) are respectively a viscosity subsolution and supersolution to (37) in $\overline{X}$.*

If (38) is equicontinuously stable, then Corollary 2.13 is simplified by the fact that, by the Arzelà-Ascoli theorem, sequences $(\alpha_n)_{n\in\mathbb{N}}$ and $(u_n)_{n\in\mathbb{N}}$ converge uniformly, up to extracting a subsequence, to some $\alpha \in \mathbb{R}$, and to some continuous function $u \colon \overline{X} \to \mathbb{R}$, which coincides with the limits superior and inferior $\overline{u}$ and $\underline{u}$ for this subsequence.

2.3 Existence

Our main result in this section concerns existence of solutions to the scheme (38). The proof is an adaptation of discrete Perron's method to our setting.

Remark 2.14 (Discrete Perron's method). In the context of this remark, let us consider a scheme of the form (35) where the operator $S^h \colon \mathbb{R}^{\mathcal{G}_h} \to \overline{\mathbb{R}}^{\mathcal{G}_h}$ is *not necessarily additively invariant*. Discrete Perron's method states that, if this scheme is monotone in the sense of Definition 2.5 and if, at some fixed step size $h > 0$, the map $S^h \colon \mathbb{R}^{\mathcal{G}_h} \to \overline{\mathbb{R}}^{\mathcal{G}_h}$ takes finite values and is continuous, then the function $\tilde{u} \colon \mathcal{G}_h \to \overline{\mathbb{R}}$ defined by

$$\tilde{u}[x] := \sup\{\overline{u}[x] \mid \overline{u} \in \mathbb{R}^{\mathcal{G}_h},\, \forall x' \in \mathcal{G}_h,\, S^h\overline{u}[x'] \leq 0\} \tag{39}$$

is a solution to the scheme, *provided that this function takes finite values*. While the discrete version of Perron's method is not as extensively described in the literature as its continuous part [16, section 4], some of its variants are stated and proved in [40, Theorem 2.3] and [42, Theorem 3.5]. Discrete Perron's method is not directly applicable when the operator S^h is additively invariant, since in this case one has $\tilde{u}[x] = +\infty$ for all $x \in \mathcal{G}_h$ (unless the scheme does not admit subsolutions, in which case $\tilde{u}[x] = -\infty$ for all $x \in \mathcal{G}_h$). In the proof of Theorem 2.15 below, we get around this difficulty by further restricting the set of admissible subsolutions $\overline{u}$ considered in the supremum (39).

While we assume in Theorem 2.15 that the scheme (38) is stable in the sense of Definition 2.12, this assumption may be relaxed, see Remark 2.16 below.

Theorem 2.15 (Existence)**.** *Assume that* (38) *is monotone, continuous, and stable. Then for small $h > 0$, there exists a solution to* (38)*.*

Proof. We define the set

$$U := \{(\alpha, u) \in \mathbb{R} \times \mathbb{R}^{\mathcal{G}_h} \mid S^{h,\alpha}u[x] \leq 0 \text{ in } \mathcal{G}_h\}$$

of subsolutions to (38). Since we assumed that (38) is stable, U is nonempty and there exists $\alpha \in \mathbb{R}$ defined by

$$\alpha := \sup_{(\overline{\alpha},\overline{u})\in U} \overline{\alpha}. \tag{40}$$

Let us show that there exists $u\colon \mathcal{G}_h \to \mathbb{R}$ such that (α, u) is a subsolution to (38). Let $((\alpha_n, u_n))_{n\in\mathbb{N}}$ be a maximizing sequence in the definition of α, and let $\alpha_* := \min_{n\in\mathbb{N}} \alpha_n$. We may assume, up to adding a constant to u_n, that $u_n[0] = 0$ for any $n \in \mathbb{N}$. Then by stability, $|u_n[x]| = |u_n[x] - u_n[0]| \le \omega(\alpha_n) \le \omega(\alpha_*)$, for any $n \in \mathbb{N}$ and $x \in \mathbb{R}^{\mathcal{G}_h}$. This means that the sequence $(u_n)_{n\in\mathbb{N}}$ is bounded in $\mathbb{R}^{\mathcal{G}_h}$ and thus that it converges, up to extracting a subsequence, to some function $\hat{u}\colon \mathcal{G}_h \to \mathbb{R}$. By continuity of the scheme, $(\alpha, \hat{u})$, as the limit of subsolutions $((\alpha_n, u_n))_{n\in\mathbb{N}}$, is a subsolution to (38).

Among all functions $u\colon \mathcal{G}_h \to \mathbb{R}$ such that (α, u) is a subsolution to (38), we choose one which maximizes the cardinality of the set $\mathcal{G}_* := \{x \in \mathcal{G}_h \mid S^{h,\alpha}u[x] < 0\}$. We will show that such a function u may be turned into another function $\tilde{u}\colon \mathcal{G}_h \to \mathbb{R}$ such that $(\alpha, \tilde{u})$ is a solution to (38), by a maximization procedure similar to the construction of the function $\tilde{u}$ in Remark 2.14.

One important difference between our proof and the classical proof of Perron's method is that in the classical proof, no specific subsolution u has to be used as a basis for constructing the solution $\tilde{u}$. In particular the assumption about the maximal cardinality of the set $\mathcal{G}_*$ is specific to our setting.

Note that $\mathcal{G}_*$ cannot be equal to $\mathcal{G}_h$, since in this case, by continuity of the scheme, there would exist $\alpha' > \alpha$ such that $(\alpha', u) \in U$ (choose α' close enough to α), and this would contradict (40).

Knowing that $\mathcal{G}_* \neq \mathcal{G}_h$, and using stability, we may define, for small $\varepsilon > 0$, the function $\tilde{u}_\varepsilon\colon \mathcal{G}_h \to \mathbb{R}$ by

$$\tilde{u}_\varepsilon[x] := \sup\{\overline{u}[x] \mid (\alpha, \overline{u}) \in U,\ \overline{u} = u \text{ in } \mathcal{G}_h \setminus \mathcal{G}_*,\ S^{h,\alpha}\overline{u}[x] \le -\varepsilon \text{ in } \mathcal{G}_*\}. \tag{41}$$

To ensure that the supremum above is the one of a nonempty set, we choose ε small enough so that u itself is suitable choice of function $\overline{u}$.

In the following two paragraphs, we show, using arguments from the classical proof of Perron's method, that $(\alpha, \tilde{u}_\varepsilon)$ is a subsolution to (38) and that $S^{h,\alpha}\tilde{u}_\varepsilon[x] = -\varepsilon$ in $\mathcal{G}_*$.

By continuity of the scheme, we may pass to the limit in maximizing sequences and deduce that for any $x \in \mathcal{G}_h$, there exists $\overline{u}\colon \mathcal{G}_h \to \mathbb{R}$ such that $(\alpha, \overline{u}) \in U$, $S^{h,\alpha}\overline{u}[x] \le -\varepsilon$ in $\mathcal{G}_*$, $\tilde{u}_\varepsilon \ge \overline{u}$ in $\mathcal{G}_h$, and $\tilde{u}_\varepsilon[x] = \overline{u}[x]$. Then by monotonicity, $S^{h,\alpha}\tilde{u}_\varepsilon[x] \le S^{h,\alpha}\overline{u}[x]$. It follows that $(\alpha, \tilde{u}_\varepsilon)$ is a subsolution to (38) and that $S^{h,\alpha}\tilde{u}_\varepsilon[x] \le -\varepsilon$ in $\mathcal{G}_*$.

Let us show that $S^{h,\alpha}\tilde{u}_\varepsilon[x] = -\varepsilon$ in $\mathcal{G}_*$. Assume that there exists $x_* \in \mathcal{G}_*$ so that $S^{h,\alpha}\tilde{u}_\varepsilon[x_*] < -\varepsilon$. For any $\delta > 0$, we define $\tilde{u}_{\varepsilon,\delta}\colon \mathcal{G}_h \to \mathbb{R}$ by

$$\tilde{u}_{\varepsilon,\delta}[x] := \begin{cases} \tilde{u}_\varepsilon[x] + \delta & \text{if } x = x_*, \\ \tilde{u}_\varepsilon[x] & \text{else.} \end{cases}$$

By monotonicity, $S^{h,\alpha}\tilde{u}_{\varepsilon,\delta}[x] \le S^{h,\alpha}\tilde{u}_\varepsilon[x]$ for any $x \in \mathcal{G}_h \setminus \{x_*\}$, and by continuity, we may choose δ small enough so that $S^{h,\alpha}\tilde{u}_{\varepsilon,\delta}[x_*] \le -\varepsilon$. This contradicts (41), since $\tilde{u}_{\varepsilon,\delta}$ is a suitable choice for $\overline{u}$ and $\tilde{u}_{\varepsilon,\delta}[x_*] > \tilde{u}_\varepsilon[x_*]$.

We now define $\tilde{u}\colon \mathcal{G}_h \to \mathbb{R}$ by

$$\tilde{u}[x] := \lim_{\varepsilon\to 0} \tilde{u}_\varepsilon[x].$$

Note that the right-hand side is the limit of a bounded nondecreasing sequence. By continuity, $S^{h,\alpha}\tilde{u}[x] = 0$ in $\mathcal{G}_*$ and $(\alpha, \tilde{u})$ is a subsolution to (38). Let us show that it is a solution. If it is not the case, then there exists $x_* \in \mathcal{G}_h \setminus \mathcal{G}_*$ such that $S^{h,\alpha}\tilde{u}[x_*] < 0$. By continuity, there exists $\varepsilon > 0$

such that $S^{h,\alpha}\tilde{u}_\varepsilon[x_*] < 0$. Since $(\alpha, \tilde{u}_\varepsilon)$ is a subsolution to (38) and $S^{h,\alpha}\tilde{u}_\varepsilon[x] < 0$ in $\mathcal{G}_*$, this contradicts the assumption that $\mathcal{G}_*$ is of maximal cardinal. Thus $(\alpha, \tilde{u})$ is necessarily a solution to (38). □

Remark 2.16. Since $h > 0$ is fixed in Theorem 2.15, the subsolution, the function ω, and the number α_0 in (i), (ii), and (iii) in the definition of stability of the scheme (Definition 2.12) only need to exist for this fixed value of h. Also, (iv) is not needed.

3 Properties of the proposed scheme

In this section, we show that the scheme (30) satisfies the properties we defined in section 2. First note that for any $h > 0$ and $\alpha \in \mathbb{R}$, the operator $S^{h,\alpha}_{\mathrm{MABV2}} \colon \mathbb{R}^{\mathcal{G}_h} \to \overline{\mathbb{R}}^{\mathcal{G}_h}$ is additively invariant.

Proposition 3.1 (Monotonicity)**.** *Assume the Lipschitz regularity properties* (15) *and* (16)*. Then the scheme* (30) *is monotone, in the sense of Definition 2.12.*

Proof. Let $h > 0$, $\alpha \in \mathbb{R}$, $x \in \mathcal{G}_h$, and $\overline{u}$, $\underline{u} \colon \mathcal{G}_h \to \mathbb{R}$ be such that $\overline{u}[x] = \underline{u}[x]$ and $\overline{u} \geq \underline{u}$ in $\mathcal{G}_h$. We need to show that

$$S^{h,\alpha}_{\mathrm{MABV2}}\overline{u}[x] \leq S^{h,\alpha}_{\mathrm{MABV2}}\underline{u}[x].$$

By the definition (29) of the operator $S^{h,\alpha}_{\mathrm{MABV2}}$, it suffices to prove that both $S^h_{\mathrm{MA}}\overline{u}[x] \leq S^h_{\mathrm{MA}}\underline{u}[x]$ and $S^h_{\mathrm{BV2}}\overline{u}[x] \leq S^h_{\mathrm{BV2}}\underline{u}[x]$. The second inequality follows directly from the definition (28) of S^h_{BV2}, so let us prove the first one.

By the definition (23) of S^h_{MA}, it suffices to prove that for any family $v = (v_1, \ldots, v_I)$ of vectors of $\mathbb{Z}^d$ and any $\gamma \in \mathbb{R}^I_+$,

$$L_{v,\gamma}(B_h\overline{u}[x], \Delta^v_h\overline{u}[x] - A^v_h\overline{u}[x]) \leq L_{v,\gamma}(B_h\underline{u}[x], \Delta^v_h\underline{u}[x] - A^v_h\underline{u}[x]).$$

First note that the operator Δ^v_h was defined so that $\Delta^v_h\overline{u}[x] \geq \Delta^v_h\underline{u}[x]$ elementwise. If $B_h\overline{u}[x] = 0$, then $B_h\overline{u}[x]^{1/d} \leq B_h\underline{u}[x]^{1/d}$, since B^h is a nonnegative operator. If $B_h\overline{u}[x] > 0$ (which, by definition of B_h, implies that $x \pm he_i \in \mathcal{G}_h$ for any $i \in \{1, \ldots, d\}$), then, using (16) for the second inequality,

$$\begin{aligned}
B_h\overline{u}[x]^{1/d} - B_h\underline{u}[x]^{1/d} &\leq B(x, D_h\overline{u}[x])^{1/d} - B(x, D_h\underline{u}[x])^{1/d} - \frac{h}{2}b_{\mathrm{LF}}\Delta_h(\overline{u} - \underline{u})[x] \\
&\leq b_{\mathrm{LF}}\left(|D_h\overline{u}[x] - D_h\underline{u}[x]|_1 - \frac{h}{2}\Delta_h(\overline{u} - \underline{u})[x]\right) \\
&= \frac{b_{\mathrm{LF}}}{2h}\sum_{i=1}^{d}\Big(|(\overline{u} - \underline{u})[x + he_i] - (\overline{u} - \underline{u})[x - he_i]| \\
&\qquad\qquad - (\overline{u} - \underline{u})[x + he_i] - (\overline{u} - \underline{u})[x - he_i]\Big) \\
&\leq 0,
\end{aligned}$$

and thus $B_h\overline{u}[x]^{1/d} \leq B_h\underline{u}[x]^{1/d}$. Similarly, for any $e \in v$, if $A^e_h\overline{u}[x] = a_{\min}|e|^2$, then $A^e_h\overline{u}[x] \leq A^e_h\underline{u}[x]$, and otherwise, using (15),

$$\begin{aligned}
A^e_h\overline{u}[x] - A^e_h\underline{u}[x] &\leq \big\langle e, \big(A(x, D_h\overline{u}[x]) - A(x, D_h\underline{u}[x])\big)e\big\rangle - \frac{h}{2}a_{\mathrm{LF}}|e|^2\Delta_h(\overline{u} - \underline{u})[x] \\
&\leq a_{\mathrm{LF}}|e|^2\left(|D_h\overline{u}[x] - D_h\underline{u}[x]|_1 - \frac{h}{2}\Delta_h(\overline{u} - \underline{u})[x]\right) \leq 0,
\end{aligned}$$

hence $A^e_h\overline{u}[x] \leq A^e_h\underline{u}[x]$. We easily conclude that $S^h_{\mathrm{MA}}\overline{u}[x] \leq S^h_{\mathrm{MA}}\underline{u}[x]$. □

From the grid $\mathcal{G}_h$, we may build a graph whose nodes are the points of $\mathcal{G}_h$ and whose edges are pairs of points that are neighbors on the grid, that is, between whom the Euclidean distance is equal to h. To prove other properties of the scheme, we need the technical assumption that the distance on this graph, multiplied by h, is equivalent to the Euclidean distance, uniformly over small $h > 0$. Equivalently, we require that there exists some positive constant $C_{\mathcal{G}}$, such that for any small $h > 0$ and any function $\varphi\colon \mathcal{G}_h \to \mathbb{R}$,

$$\max_{\substack{x_1,x_2\in\mathcal{G}_h\\ x_1\neq x_2}} \frac{|\varphi[x_1]-\varphi[x_2]|}{|x_1-x_2|} \leq C_{\mathcal{G}} \max_{\substack{x_1,x_2\in\mathcal{G}_h\\ |x_1-x_2|=h}} \frac{|\varphi[x_1]-\varphi[x_2]|}{h}. \tag{42}$$

Proposition 3.2 (Continuity)**.** *Assume* (42)*. Then the scheme* (30) *is continuous, in the sense of Definition 2.12.*

Proof. For any $x \in \mathcal{G}_h$, the function $\mathbb{R}^{\mathcal{G}_h} \to \overline{\mathbb{R}}$, $u \mapsto S^{h,\alpha}_{\mathrm{MABV2}}u[x]$ is a maximum over a compact set of continuous functions with values in $\mathbb{R}\cup\{-\infty\}$, see (23), (28), and (29). Hence it is a continuous function with values in $\mathbb{R}\cup\{-\infty\}$. It remains to prove that $S^{h,\alpha}_{\mathrm{MABV2}}u[x] > -\infty$.

By (42), there exists $e = \pm e_i$, $i \in \{1,\dots,d\}$, such that $x - he \in \mathcal{G}_h$. Therefore

$$S^{h,\alpha}_{\mathrm{MABV2}}u[x] \geq S^{h}_{\mathrm{BV2}}u[x] \geq D^e_h u[x] - \sigma_{P(x)}(e) = -\delta^{-e}_h u[x] - \sigma_{P(x)}(e) > -\infty,$$

which concludes the proof. □

Let us now study the consistency of the scheme (30) with the degenerate elliptic equation

$$F^{\alpha}_{\mathrm{MABV2}}(x, Du(x), D^2u(x)) = 0 \quad \text{in } \overline{X}, \tag{43}$$

where for any $\alpha \in \mathbb{R}$, $x \in \overline{X}$, $p \in \mathbb{R}^d$, and $M \in \mathcal{S}_d$,

$$F^{\alpha}_{\mathrm{MABV2}}(x,p,M) := \begin{cases} (F_{\mathrm{MA}}(x,p,M)+\alpha) \vee F_{\mathrm{BV2}}(x,p) & \text{if } x \in X,\\ -\infty & \text{else,}\end{cases}$$

and $F_{\mathrm{MA}}(x,p,M)$ and $F_{\mathrm{BV2}}(x,p)$ are defined respectively in (8) and (27). We first prove a consistency property that is stronger to the one we introduced in Definition 2.12, and that will be useful in the study of stability of the scheme.

Proposition 3.3 (Consistency)**.** *Assume* (10)*,* (14)*,* (20)*, and* (21)*. Let $\varphi \in C^\infty(\overline{X})$ and $(\alpha_h)_{h>0}$ be a family of real numbers converging to some $\alpha \in \mathbb{R}$ as h approaches zero. Then*

$$S^{h,\alpha_h}_{\mathrm{MABV2}}\varphi[x] \leq F^{\alpha}_{\mathrm{MABV2}}(x, D\varphi(x), D^2\varphi(x)) + o_{h\to 0}(1), \tag{44}$$

uniformly over $x \in \mathcal{G}_h$ and $\alpha \in \mathbb{R}$. Moreover, for any compact subset K of X,

$$S^{h,\alpha_h}_{\mathrm{MABV2}}\varphi[x] \geq F^{\alpha}_{\mathrm{MABV2}}(x, D\varphi(x), D^2\varphi(x)) + o_{h\to 0}(1), \tag{45}$$

uniformly over $x \in K \cap \mathcal{G}_h$ and $\alpha \in \mathbb{R}$.

Proof. Let K be a compact subset of X. For convenience, when $a_h(x)$ and $b_h(x)$ are real numbers depending on $h > 0$ and on $x \in \mathcal{G}_h$, we write $a_h(x) \leqq_K b_h(x)$ if $a_h(x) \leq b_h(x)$ for any $h > 0$ and $x \in \mathcal{G}_h$, with equality if $x \in K$. Then it suffices to show that

$$S^{h}_{\mathrm{MA}}\varphi[x] \leqq_K F_{\mathrm{MA}}(x, D\varphi(x), D^2\varphi(x)) + o_{h\to 0}(1), \tag{46}$$

$$S^{h}_{\mathrm{BV2}}\varphi[x] \leqq_K F_{\mathrm{BV2}}(x, D\varphi(x)) + o_{h\to 0}(1), \tag{47}$$

uniformly over $x \in \mathcal{G}_h$.

For any $x \in \mathcal{G}_h$ and $i \in \{1, \dots, d\}$, it holds that $T_h^{\pm e_i}\varphi[x] \geq \varphi(x \pm he_i)$, and using (10), we may assume that h is small enough so that the equality $T_h^{\pm e_i}\varphi[x] = \varphi(x \pm he_i)$ holds whenever $x \in K$. Then inserting first-order Taylor expansions of φ in the definition of S^h_{BV2} yields (47).

If $x \in \mathcal{G}_h$ is such that $\Delta_h\varphi[x] < +\infty$, then $x \pm he_i \in \mathcal{G}_h$ for any $i \in \{1, \dots, d\}$, and thus $D_h\varphi[x] = D\varphi(x) + O(h^2)$ and $\Delta_h\varphi[x] = \Delta\varphi(x) + O(h^2)$. In particular, $\Delta_h\varphi[x]$ is bounded. Therefore, using that $B^{1/d}$ is Lipschitz continuous with respect to its last variable, uniformly with respect to its first variable,

$$B(x, D_h\varphi[x])^{1/d} - \frac{h}{2} b_{\mathrm{LF}} \Delta_h\varphi[x] = B(x, D\varphi(x))^{1/d} + O(h).$$

Since $B \geq 0$ and using the definition (18) of B_h, it follows that

$$B_h\varphi[x]^{1/d} = B(x, D\varphi(x))^{1/d} + O(h).$$

Now if $\Delta_h\varphi[x] = +\infty$ (by (10), for h small, this may only happen if $x \notin K$), it holds that $B_h\varphi[x] = 0 \leq B(x, D\varphi(x))$. We deduce that

$$B_h\varphi[x]^{1/d} \leqq_K B(x, D\varphi(x))^{1/d} + O(h)$$

uniformly over $x \in \mathcal{G}_h$. Similarly, for any $v \in V_h$ and $e \in v$, we may assume, using (10) and (21), that h is small enough so that $x \pm he \in \mathcal{G}_h$ whenever $x \in K \cap \mathcal{G}_h$, and then, using (14) and the same reasoning as above,

$$\begin{aligned} A_h^e\varphi[x] &\leqq_K \langle e, A(x, D\varphi(x))e\rangle + O(h|e|^2), \\ -\Delta_h^e\varphi[x] &\leqq_K -\langle e, D^2\varphi(x)e\rangle + O(h^2|e|^4), \end{aligned}$$

uniformly over $x \in \mathcal{G}_h$. Then for any $v \in V_h$ and $\gamma \in \mathbb{R}_+^{d(d+1)/2}$ such that $\mathrm{Tr}(\mathcal{D}_v(\gamma)) = \sum_{i=1}^{d(d+1)/2} \gamma_i |v_i|^2 = 1$, using (21) for the last equality,

$$\begin{aligned} -\langle \gamma, \Delta_h^v\varphi[x] - A_h^v\varphi[x]\rangle &= -\sum_{i=1}^{d(d+1)/2} \gamma_i (\Delta_h^{v_i}\varphi[x] - A_h^{v_i}\varphi[x]) \\ &\leqq_K -\sum_{i=1}^{d(d+1)/2} \gamma_i \left\langle v_i, \left(D^2\varphi(x) - A(x, D\varphi(x))\right) v_i \right\rangle \\ &\quad + \sum_{i=1}^{d(d+1)/2} \gamma_i O(h|v_i|^2 + h^2|v_i|^4) \\ &= -\langle \mathcal{D}_v(\gamma), D^2\varphi(x) - A(x, D\varphi(x))\rangle + O(h + h^2|v_i|^2) \\ &= -\langle \mathcal{D}_v(\gamma), D^2\varphi[x] - A(x, D\varphi(x))\rangle + o_{h\to 0}(1), \end{aligned} \tag{48}$$

uniformly over $x \in \mathcal{G}_h$, v, and γ. Thus

$$S^h_{\mathrm{MA}}\varphi[x] \leqq_K \max_{v \in V_h} \max_{\substack{\gamma \in \mathbb{R}_+^{d(d+1)/2} \\ \mathrm{Tr}(\mathcal{D}_v(\gamma)) = 1}} L_{\mathcal{D}_v(\gamma)} \left(B(x, D\varphi(x)), D^2\varphi(x) - A(x, D\varphi(x))\right) + o_{h\to 0}(1).$$

We deduce (46) using (20) and that the affine map

$$\{\mathcal{D} \in \mathcal{S}_d^+ \mid \mathrm{Tr}(\mathcal{D}) = 1\} \to \mathbb{R}, \quad \mathcal{D} \mapsto L_{\mathcal{D}}(b, M) \tag{49}$$

is continuous, uniformly over b and M belonging to compact sets. □

Remark 3.4 (Order of consistency). Under appropriate assumptions, the order of consistency of the scheme (30) is easily deduced from the proof of Proposition 3.3. Let $\varphi \in C^\infty(X)$, and let $K \subset X$ be compact. Then, for small $h > 0$ and uniformly over $x \in K \cap \mathcal{G}_h$,

$$S^h_{\mathrm{BV2}}\varphi[x] = F_{\mathrm{BV2}}(x, D\varphi(x)) + O(h).$$

For the operator S^h_{MA}, we distinguish two cases:

(*General case*) If there exist $r_1 > 0$ and $r_2 \in (0,1)$ such that the following refinements of (20) and (21) hold:

$$d_H\left(\{\mathcal{D}_v(\gamma) \mid v \in V_h,\, \gamma \in \mathbb{R}_+^{d(d+1)/2},\, \mathrm{Tr}(\mathcal{D}_v(\gamma)) = 1\}, \{\mathcal{D} \in \mathcal{S}_d^+ \mid \mathrm{Tr}(\mathcal{D}) = 1\}\right) = O(h^{r_1}),$$
$$\max_{v \in V_h} \max_{e \in v} |e| = O(h^{-r_2}),$$

then, refining the last equality in (48) and using that the map (49) is $1/d$-Hölder continuous, one has, for small $h > 0$ and uniformly over $x \in K \cap \mathcal{G}_h$,

$$S^h_{\mathrm{MA}}\varphi[x] = F_{\mathrm{MA}}(x, D\varphi(x), D^2\varphi(x)) + O(h^{1 \wedge (2-2r_2) \wedge (r_1/d)}).$$

In dimension $d = 2$, when choosing V_h as in Remark B.9, one has $r_1 = 2r$ and $r_2 = r$, hence S^h_{MA} is consistent with F_{MA} to the order $1 \wedge (2-2r) \wedge r$, and the optimal choice for r is $r = 2/3$, yielding consistency to the order $2/3$.

(*Smooth case*) The consistency is improved if (2) admits a solution $u \in C^2(\overline{X})$ such that, uniformly over K, $D^2u(x) - A(x, Du(x)) \in \mathcal{S}_d^{++}$ has condition number less than some constant $c > 1$. In this setting, the maximum in (8) is attained for $\mathcal{D} = (D^2u(x) - A(x, Du(x)))^{-1} / \mathrm{Tr}((D^2u(x) - A(x, Du(x)))^{-1})$, which has condition number less than c for all $x \in K$. We thus recommend choosing the set V_h independently of h, but such that any $\mathcal{D} \in \mathcal{S}_d^{++}$ with condition number less than c is of the form $\mathcal{D} = \mathcal{D}_v(\gamma)$ for some $v \in V_h$ and $\gamma \in \mathbb{R}_+^{d(d+1)/2}$ (see Appendix B for a suitable construction of V_h in dimension $d = 2$). Then (20) is not satisfied, but in a neighborhood of the solution u, the operator S^h_{MA} is still consistent with F_{MA}, to the order one, uniformly over $x \in K$.

In practice, one may choose to implement the scheme with Lax-Friedrichs relaxation parameters $a_{\mathrm{LF}} = b_{\mathrm{LF}} = 0$, as we do in section 6. The drawback of doing this is that (15) and (16), and thus Proposition 3.1, do not hold anymore unless $A(x,p)$ and $B(x,p)$ do not depend on p. The benefit is that consistency is improved. In the setting of the smooth case described above, if $a_{\mathrm{LF}} = b_{\mathrm{LF}} = 0$, then, in a neighborhood of u and uniformly over $x \in K$, S^h_{MA} is consistent with F_{MA} to the order two.

Note that the order of consistency of the whole scheme (30) is the minimum of the ones of S^h_{BV2} and S^h_{MA}, but for a fixed point x, the order is the one of the operator for which the maximum is reached in (29), which in practice is $S^{h,\alpha}_{\mathrm{MA}} = S^h_{\mathrm{MA}} + \alpha$ at most points of the grid.

Corollary 3.5 (Consistency)**.** *Assume* (10)*,* (14)*,* (20)*, and* (21)*. Then the scheme* (30) *is consistent with equation* (43)*, in the sense of Definition 2.12.*

Proof. We have to show that if φ, $(\alpha_h)_{h>0}$, and α are as in Proposition 3.3, then for any $x \in \overline{X}$,

$$\limsup_{\substack{h>0,\, h\to 0 \\ x' \in \mathcal{G}_h,\, x' \to x}} S^{h,\alpha_h}_{\mathrm{MABV2}}\varphi[x'] \le (F^\alpha_{\mathrm{MABV2}})^*(x, D\varphi(x), D^2\varphi(x)), \tag{50}$$

$$\liminf_{\substack{h>0,\, h\to 0 \\ x' \in \mathcal{G}_h,\, x' \to x}} S^{h,\alpha_h}_{\mathrm{MABV2}}\varphi[x'] \ge (F^\alpha_{\mathrm{MABV2}})_*(x, D\varphi(x), D^2\varphi(x)). \tag{51}$$

If $x \in X$, then (50) and (51) follow respectively from (44) and (45), taking first the limit over h and then the limit over x'. If $x \in \partial X$, then (50) follows from (44) and (51) is always true, since $(F^\alpha_{\mathrm{MABV2}})_*(x, D\varphi(x), D^2\varphi(x)) = -\infty$. □

Finally, we establish stability of the proposed scheme.

Proposition 3.6 (Equicontinuous stability)**.** *Assume* (10), (14) *to* (16), (20) *to* (22), *and* (42). *If there exists a function* $\varphi \in C^\infty(\overline{X})$ *such that for any* $x \in \overline{X}$, $D\varphi(x) \in P(x)$, *then the scheme* (30) *is equicontinuously stable, in the sense of Definition 2.12.*

Proof. Let us check all items in the definition of equicontinuous stability.

(i) The function φ was chosen so that $F_{\mathrm{BV2}}(x, D\varphi(x)) < 0$, uniformly over $x \in \overline{X}$. Also, since A and B are bounded, there exists $\alpha_1 \leq 0$ such that $F_{\mathrm{MA}}(x, D\varphi(x), D^2\varphi(x)) < -\alpha_1$, uniformly over $x \in \overline{X}$. It follows that $(F^{\alpha_1}_{\mathrm{MABV2}})^*(x, D\varphi(x), D^2\varphi(x)) < 0$, uniformly over $x \in \overline{X}$. Then by Proposition 3.3, for any small $h > 0$ and any $x \in \mathcal{G}_h$, $S^{h,\alpha_1}_{\mathrm{MABV2}}\varphi[x] < 0$. Hence (α_1, φ) is a subsolution to (30) for small $h > 0$.

(ii') Let $h > 0$ be small and let $(\alpha, u) \in \mathbb{R} \times \mathbb{R}^{\mathcal{G}_h}$ be a subsolution to (30). Then for any $x \in \mathcal{G}_h$, $S^h_{\mathrm{BV2}}u[x] \leq 0$. Choosing $e = \pm e_i$, $i \in \{1, \dots, d\}$ in the definition of S^h_{BV2}, it follows that $-\delta^{\pm e_i}_h u[x] \leq \sigma_{P(x)}(\mp e_i)$. Since the compact set $\overline{P(x)}$ is continuous with respect to $x \in \overline{X}$ for the Hausdorff distance, there exists $C_P \geq 0$ such that for any $x \in \overline{X}$ and $i \in \{1, \dots, d\}$, $\sigma_{P(x)}(\pm e_i) \leq C_P$. Hence $-\delta^{\pm e_i}_h u[x] \leq C_P$. Using (42), we easily deduce that

$$\max_{\substack{x_1, x_2 \in \mathcal{G}_h \\ x_1 \neq x_2}} \frac{|u[x_1] - u[x_2]|}{|x_1 - x_2|} \leq C_{\mathcal{G}} C_P.$$

Hence (ii') holds with $\omega(\alpha, t) := C_{\mathcal{G}} C_P t$.

(iii) Let $h > 0$ be small and $(\alpha, u) \in \mathbb{R} \times \mathbb{R}^{\mathcal{G}_h}$ be a subsolution to (30). Then for any $x \in \mathcal{G}_h$, $S^h_{\mathrm{MA}}u[x] \leq -\alpha$. By (22), there exists $v \in V_h$ and $\gamma \in \mathbb{R}_+^{d(d+1)/2}$ such that $\mathcal{D}_v(\gamma) = e_1 \otimes e_1$ (and thus $\mathrm{Tr}(\mathcal{D}_v(\gamma)) = 1$). Choosing v and γ as parameters in the definition of S^h_{MA} yields $A^{e_1}_h u[x] - \Delta^{e_1}_h u[x] \leq -\alpha$. Since $A^{e_1}_h u[x] \geq a_{\min}$, it follows that $\Delta^{e_1}_h u[x] \geq a_{\min} + \alpha$.

Let $\ell > 0$, independent of h, be such that the segment $[0, \ell e_1]$ belongs to X (recall that $0 \in X$ by assumption), and let $n_h := \lceil \ell/h \rceil$. By (10), we may assume that h is small enough so that $ihe_1 \in X$, for any $i \in \{0, \dots, n_h + 1\}$. Then for any $i \in \{1, \dots, n_h\}$, $h\Delta^{e_1}_h u[ihe_1] = \delta^{e_1}_h u[ihe_1] + \delta^{-e_1}_h u[ihe_1] = \delta^{e_1}_h u[ihe_1] - \delta^{e_1}_h u[(i-1)he_1]$, hence $\delta^{e_1}_h u[ihe_1] = \delta^{e_1}_h u[(i-1)he_1] + h\Delta^{e_1}_h u[ihe_1]$ and

$$\delta^{e_1}_h u[n_h h e_1] = \delta^{e_1}_h u[0] + h \sum_{i=1}^{n_h} \Delta^{e_1}_h u[ihe_1] \geq \delta^{e_1}_h u[0] + n_h h(a_{\min} + \alpha).$$

Since $n_h h \geq \ell$, if $\alpha \geq -a_{\min}$, then

$$\delta^{e_1}_h u[n_h h e_1] \geq \delta^{e_1}_h u[0] + \ell(a_{\min} + \alpha) = -\delta^{-e_1}_h u[he_1] + \ell(a_{\min} + \alpha).$$

We proved in (ii) that $\delta^{e_1}_h u[n_h h e_1] \leq C_P$ and $\delta^{-e_1}_h u[he_1] \leq C_P$. Therefore

$$\alpha \leq \frac{2C_P}{\ell} - a_{\min}.$$

(iv) Let $h > 0$ be small and $(\alpha, u) \in \mathbb{R} \times \mathbb{R}^{\mathcal{G}_h}$ be a solution to (30) (note that in the proof, we only use that it is a supersolution). Let $\alpha_1 \geq 0$ be as in (i). Up to adding a constant to u, we may assume that there exists $x \in \mathcal{G}_h$ such that $u[x] = \varphi[x]$ and $u \geq \varphi$ in $\mathcal{G}_h$. Then by Proposition 3.1, $S^{h,\alpha_1}_{\mathrm{MABV2}}u[x] \leq S^{h,\alpha_1}_{\mathrm{MABV2}}\varphi[x]$. We proved in (i) that $S^{h,\alpha_1}_{\mathrm{MABV2}}\varphi[x] < 0$. Thus $S^{h,\alpha_1}_{\mathrm{MABV2}}u[x] < 0$, and by definition of $S^{h,\alpha_1}_{\mathrm{MABV2}}$, it holds that $S^h_{\mathrm{BV2}}u[x] < 0$ and $S^h_{\mathrm{MA}}u[x] < -\alpha_1$. On the other hand, the equality $S^{h,\alpha}_{\mathrm{MABV2}}u[x] = 0$ may be expanded as

$$S^h_{\mathrm{BV2}}u[x] \vee (S^h_{\mathrm{MA}}u[x] + \alpha) = 0.$$

Since $S^h_{\mathrm{BV2}}u[x] < 0$, we deduce that $\alpha = -S^h_{\mathrm{MA}}u[x] > \alpha_1$. □

Note that in the proof of item (ii'), we actually proved that solutions to the scheme are Lipschitz continuous uniformly over small $h > 0$.

The existence of a suitable function φ in Proposition 3.6 is a natural assumption in the setting of optimal transport. We defer discussion of this assumption to section 5.1, and in particular to Remark 5.1.

4 Closed-form formula in dimension two

This section is devoted to the proof of Theorem 1.2, whose motivation is to improve the numerical efficiency of the scheme. Recall that the scheme residues are defined as the value (23) of a maximization problem. In Remark 4.1, we contrast the numerical cost of computing this maximal value using the explicit formula of Theorem 1.2 with a more traditional approach based on a grid search in the parameter space. In practice, and in the numerical experiments section 6, the objective is to solve the scheme using a Newton method, which requires the following additional ingredients: (i) generating the sparse Jacobian matrix of the scheme, (ii) solving the linearized scheme, and (iii) iterating the previous two steps until the residues fall below a given threshold. Point (i) is addressed using a custom automatic differentiation library[1], combining sparse and dense forward differentiation, and which takes advantage of the envelope theorem [13, Section 6.1] so as to efficiently differentiate the maximal value (23). Point (ii) relies on the standard SuperLU sparse direct solver. Point (iii) usually terminates in less than a dozen steps in practice, and the proposed scheme compares favorably to alternatives in this regard, see section 6.4. Eventually, the evaluation of the scheme residues nevertheless accounts for a substantial part of the complexity of the proposed numerical method, and is also its most specific ingredient.

Remark 4.1 (Numerical complexity of the scheme numerical evaluation). Consider a two-dimensional Cartesian grid $\mathcal{G}_h$ with $O(N^2)$ points. Assume that at any point $x \in \mathcal{G}_h$, one has to perform respectively M_{MA} and M_{BV2} operations in order to compute $S^h_{\mathrm{MA}}u[x]$ and $S^h_{\mathrm{BV2}}u[x]$. Then the overall numerical complexity of the scheme on the grid $\mathcal{G}_h$ is $O(N^2(M_{\mathrm{MA}} + M_{\mathrm{BV2}}))$.

When using Theorem 1.2 in the implementation of the scheme, M_{MA} is proportional to the number of superbases in the set V_h. As in Remark 3.4, we distinguish between the *smooth case* and the *general case*. In the smooth case, V_h does not depend on N, hence $M_{\mathrm{MA}} = O(1)$. In the general case, if V_h is built as in Remark B.9, with $r = 2/3$ as suggested by Remark 3.4, then by Proposition B.10, $M_{\mathrm{MA}} = O(N^{2/3} \log N)$.

For comparison, one could choose to discretize the parameter set of the maximum in the definition (8) of the operator S^h_{MA} instead of using Theorem 1.2, and in this case M_{MA} would be proportional to the number of points in this discretization. Since the set of symmetric positive semidefinite matrices of size two and of unit trace has dimension two, in order to guarantee consistency of the scheme to some order $r > 0$, one should choose at least $M_{\mathrm{MA}} = O(N^{2r})$. This is more costly than using Theorem 1.2, both in the smooth case (in which the desired order, according to Remark 3.4, is $r = 1$, or even $r = 2$ if $a_{\mathrm{LF}} = b_{\mathrm{LF}} = 0$) and in the general case (in which the desired order is $r = 2/3$).

There is also a maximum in the definition (28) of S^h_{BV2} which, depending on the expression of the set-valued function P in (24), either admits a closed-form formula or needs to be discretized. If it admits a closed-form formula, then M_{BV2} does not depend on N. If it needs to be discretized, then M_{BV2} is proportional to the number of points in the discretization and, in order to guarantee consistency of the operator S^h_{BV2} with F_{BV2} at some order $r > 0$, one should choose $M_{\mathrm{BV2}} = O(N^r)$, since the parameter set is one-dimensional. The numerical cost of this discretization is negligible in the general case, but not in the smooth case. In practice, in many

[1] See `https://github.com/Mirebeau/AdaptiveGridDiscretizations`.

applications, the maximum in (29) is only attained by $S^h_{\mathrm{BV2}}u[x]$ at points $x \in \mathcal{G}_h$ that are close to ∂X. A perspective for future research would be to prove that one may use a variant of the scheme (30) which would only require computing $S^h_{\mathrm{BV2}}u[x]$ at such points, reducing the numerical cost of handling the boundary condition (24).

In dimension $d = 2$, choosing V^h as a family of superbases of $\mathbb{Z}^2$ (see Definition 1.1) is motivated by *Selling's formula* [43]: for any family $v = (v_1, v_2, v_3)$ of vectors of $\mathbb{Z}^2$, recall that we defined $\gamma\colon \mathbb{R}^3 \to \mathcal{S}_2^+$ by

$$\mathcal{D}_v(\gamma) := \sum_{i=1}^{3} \gamma_i v_i \otimes v_i,$$

and let us also define $\gamma_v\colon \mathcal{S}_2 \to \mathbb{R}^3$ by

$$\gamma_v(\mathcal{D}) := (-\langle v_{i+1}^\perp, \mathcal{D} v_{i+2}^\perp\rangle)_{1\le i\le 3}, \tag{52}$$

where we consider the indices of the elements of v modulo three, and where if $e = (a, b) \in \mathbb{R}^2$, we denote $e^\perp := (-b, a)$.

Proposition 4.2 (Selling's formula)**.** *If $v = (v_1, v_2, v_3)$ is a superbase of $\mathbb{Z}^2$, then γ_v is the inverse bijection of $\mathcal{D}_v$: for any $\mathcal{D} \in \mathcal{S}_2$, $\mathcal{D} = \mathcal{D}_v(\gamma_v(\mathcal{D}))$.*

Proof. It suffices to show that for any $1 \le i \le j \le 2$,

$$\langle v_i^\perp, \mathcal{D} v_j^\perp\rangle = \langle v_i^\perp, \mathcal{D}_v(\gamma_v(\mathcal{D}))v_j^\perp\rangle.$$

This is easily verified using the properties of superbases of $\mathbb{Z}^2$ and the fact that for any $\{i, j\} \subset \{1, 2, 3\}$, $\langle v_i^\perp, v_j\rangle = \det(v_i, v_j)$. □

Proof of Theorem 1.2. We prove separately the two statements of the theorem.

Case of bases. Let $v = (v_1, v_2)$ be a basis of $\mathbb{Z}^2$, $b \ge 0$, and $m = (m_1, m_2) \in \mathbb{R}^2$. Note that

$$\{\gamma \in \mathbb{R}^+_2 \mid \mathrm{Tr}(\mathcal{D}_v(\gamma)) = 1\} = \left\{ \left(\frac{1+t}{2|v_1|^2}, \frac{1-t}{2|v_2|^2} \right) \mid t \in [-1, 1] \right\},$$

is the segment of endpoints $(1/|v_1|^2, 0)$ and $(0, 1/|v_2|^2)$. Then

$$\begin{aligned}
&\max_{\substack{\gamma\in\mathbb{R}^2_+ \\ \mathrm{Tr}(\mathcal{D}_v(\gamma))=1}} L_{v,\gamma}(b, m) \\
&= \max_{t\in[-1,1]} \left(2b^{1/2} \left(\det \mathcal{D}_v \left(\left(\frac{1+t}{2|v_1|^2}, \frac{1-t}{2|v_2|^2} \right) \right) \right)^{1/2} - \frac{1+t}{2|v_1|^2} m_1 - \frac{1-t}{2|v_2|^2} m_2 \right).
\end{aligned}$$

We compute that for any $t \in [-1, 1]$,

$$\begin{aligned}
\det \mathcal{D}_v \left(\left(\frac{1+t}{2|v_1|^2}, \frac{1-t}{2|v_2|^2} \right) \right) &= \det \left(\frac{(1+t)}{2|v_1|^2} v_1 \otimes v_1 + \frac{(1-t)}{2|v_2|^2} v_2 \otimes v_2 \right) \\
&= \frac{1}{4}(1-t^2) \frac{\det(v_1, v_2)^2}{|v_1|^2|v_2|^2} = \frac{(1-t^2)}{4|v_1|^2|v_2|^2},
\end{aligned}$$

using the definition of $\mathcal{D}_v$ for the first equality, that $\det(a \otimes a + b \otimes b) = \det(a, b)^2$ for any a, $b \in \mathbb{R}^2$ for the second equality, and that $\det(v_1, v_2) = \pm 1$ for the third equality. After defining $\omega_v^{(0)} \in \mathbb{R}$ and $\omega_v^{(1)}$, $\omega_v^{(2)} \in \mathbb{R}^2$ by

$$\omega_v^{(0)} := \frac{1}{|v_1|^2|v_2|^2}, \qquad \omega_v^{(1)} := \frac{1}{2} \begin{pmatrix} 1/|v_1|^2 \\ -1/|v_2|^2 \end{pmatrix}, \qquad \omega_v^{(2)} := \frac{1}{2} \begin{pmatrix} 1/|v_1|^2 \\ 1/|v_2|^2 \end{pmatrix},$$

it follows that

$$\max_{\substack{\gamma\in\mathbb{R}^2_+\\ \mathrm{Tr}(\mathcal{D}_v(\gamma))=1}} L_{v,\gamma}(b,m) = \max_{t\in[-1,1]} \left((\omega_v^{(0)})^{1/2} b^{1/2}(1-t^2)^{1/2} - \langle \omega_v^{(1)}, m\rangle t - \langle \omega_v^{(2)}, m\rangle\right).$$

This is the maximum of a concave function over $[-1,1]$. Writing the first order optimality condition yields that the optimal t must satisfy

$$t^2 = \frac{\langle \omega_v^{(1)}, m\rangle^2}{\omega_v^{(0)} b + \langle \omega_v^{(1)}, m\rangle^2},$$

from which we deduce the expected formula

$$\max_{\substack{\gamma\in\mathbb{R}^2_+\\ \mathrm{Tr}(\mathcal{D}_v(\gamma))=1}} L_{v,\gamma}(b,m) = (\omega_v^{(0)} b + \langle \omega_v^{(1)}, m\rangle^2)^{1/2} - \langle \omega_v^{(2)}, m\rangle = \tilde{H}_v(b,m).$$

Case of superbases. We use that in the space of symmetric matrices size two equipped with the Frobenius norm, the set of symmetric positive semidefinite matrices of unit trace is a disk. More precisely, let us define the affine map $\mathfrak{D}\colon \mathbb{R}^2 \to \mathcal{S}_2$ by

$$\mathfrak{D}(\rho) = \frac{1}{2}\begin{pmatrix} 1+\rho_1 & \rho_2 \\ \rho_2 & 1-\rho_1 \end{pmatrix}. \tag{53}$$

Note that the above definition is closely related to Pauli matrices in quantum mechanics. It is easily proved that

$$\{\mathcal{D}\in\mathcal{S}_2^+ \mid \mathrm{Tr}(\mathcal{D})=1\} = \{\mathfrak{D}(\rho) \mid |\rho|\le 1\}. \tag{54}$$

Moreover, for any $\rho\in\mathbb{R}^d$ such that $|\rho|\le 1$,

$$\det \mathfrak{D}(\rho) = \frac{1}{4}(1-|\rho|^2), \qquad \mathrm{Cond}(\mathfrak{D}(\rho)) = \frac{1+|\rho|}{1-|\rho|}. \tag{55}$$

Let $v=(v_1,v_2,v_3)$ be a superbase of $\mathbb{Z}^2$, $b\ge 0$, and $m\in\mathbb{R}^3$. The Minkowski determinant inequality states, in any dimension $d\in\mathbb{N}$, the function $\det(\cdot)^{1/d}$ is concave over $\mathcal{S}_d^+$. Hence the function

$$\{\gamma\in\mathbb{R}^3 \mid \mathcal{D}_v(\gamma)\succeq 0,\ \mathrm{Tr}(\mathcal{D}_v(\gamma))=1\} \to \mathbb{R}, \quad \gamma\mapsto L_{v,\gamma}(b,m)$$

is concave too. Recall that $\mathcal{D}_v(\gamma)\succeq 0$ whenever $\gamma\in\mathbb{R}^3_+$. Let

$$\gamma_v^*(b,m) \in \operatorname*{argmax}_{\substack{\gamma\in\mathbb{R}^3\\ \mathcal{D}_v(\gamma)\succeq 0\\ \mathrm{Tr}(\mathcal{D}_v(\gamma))=1}} L_{v,\gamma}(b,m).$$

If the strict elementwise inequality $\gamma_v^*(b,m) >_{\mathrm{vec}} 0$ is not satisfied, then

$$\max_{\substack{\gamma\in\mathbb{R}^3_+\\ \mathrm{Tr}(\mathcal{D}_v(\gamma))=1}} L_{v,\gamma}(b,m) = \max_{1\le i<j\le 3} \max_{\substack{\gamma\in\mathbb{R}^2_+\\ \mathrm{Tr}(\mathcal{D}_v(\gamma))=1}} L_{(v_i,v_j),\gamma}(b,m) = \max_{1\le i<j\le 3} \tilde{H}_{(v_i,v_j)}(b,m),$$

since the maximum in the left-hand side is attained on the boundary of the parameter set. Thus it suffices to prove that

$$H_v(b,m) = \begin{cases} L_{v,\gamma_v^*(b,m)}(b,m) & \text{if } \gamma_v^*(b,m) >_{\mathrm{vec}} 0,\\ -\infty & \text{else.}\end{cases}$$

Let us prove the above. If $\gamma_v \colon \mathcal{S}_2 \to \mathbb{R}^3$ and $\mathfrak{D} \colon \mathbb{R}^2 \to \mathcal{S}_2$ are functions defined respectively by (52) and (53), then, by (54) and Selling's Formula (Proposition 4.2), it holds that

$$\max_{\substack{\gamma \in \mathbb{R}^3 \\ \mathcal{D}_v(\gamma) \succeq 0 \\ \mathrm{Tr}(\mathcal{D}_v(\gamma))=1}} L_{v,\gamma}(b,m) = \max_{|\rho| \leq 1} L_{v,\gamma_v(\mathfrak{D}(\rho))}(b,m),$$

and there exists

$$\rho_v^*(b,m) \in \operatorname*{argmax}_{|\rho| \leq 1} L_{v,\gamma_v(\mathfrak{D}(\rho))}(b,m)$$

such that

$$\gamma_v^*(b,m) = \gamma_v(\mathfrak{D}(\rho_v^*(b,m))).$$

Let

$$W_v := \frac{1}{2} \begin{pmatrix} v_{2,1}v_{3,1} - v_{2,2}v_{3,2} & v_{2,1}v_{3,2} + v_{2,2}v_{3,1} \\ v_{1,1}v_{3,1} - v_{1,2}v_{3,2} & v_{1,1}v_{3,2} + v_{1,2}v_{3,1} \\ v_{1,1}v_{2,1} - v_{1,2}v_{2,2} & v_{1,1}v_{2,2} + v_{1,2}v_{2,1} \end{pmatrix}.$$

Recall that $Q_v \in \mathcal{S}_3$ and $w_v \in \mathbb{R}^3$ were defined in the statement of the theorem, and note that $Q_v = W_v W_v^\top$. It is easily computed that for any $\rho \in \mathbb{R}^2$,

$$\gamma_v(\mathfrak{D}(\rho)) = W_v \rho - w_v,$$

and thus, using also (55), that

$$L_{v,\gamma_v(\mathfrak{D}(\rho))}(b,m) = b^{1/2}(1-|\rho|^2)^{1/2} - \langle W_v\rho - w_v, m\rangle.$$

Therefore, $\rho_v^*(b,m)$ is the argmax of a concave function over the unit disk, and writing the first-order optimality condition yields

$$\rho_v^*(v,m) = -\frac{W_v^\top m}{(b+|W_v^\top m|^2)^{1/2}} = -\frac{W_v^\top m}{(b+\langle m, Q_v m\rangle)^{1/2}}.$$

Thus

$$\gamma_v^*(b,m) = \gamma_v(\mathfrak{D}(\rho_v^*(b,m))) = -\frac{Q_v m}{(b+\langle m, Q_v m\rangle)^{1/2}} - w_v$$

and

$$L_{v,\gamma_v^*(b,m)}(b,m) = L_{v,\gamma_v(\mathfrak{D}(\rho_v^*(b,m)))}(b,m) = (b+\langle m, Q_v m\rangle)^{1/2} + \langle w_v, m\rangle,$$

which concludes the proof. □

5 Application to quadratic optimal transport

We specialize in this section the proposed scheme to the quadratic optimal transport problem and provide a convergence proof, taking advantage of specific tools in this setting such as Aleksandrov solutions and the mass balance equation, in addition to the generic tools introduced in section 3.

5.1 The quadratic optimal transport problem

Let Y be an open bounded convex nonempty subset of $\mathbb{R}^d$ and $f\colon \overline{X} \to \mathbb{R}_+$ and $g\colon \overline{Y} \to \mathbb{R}_+^*$ be two densities satisfying the mass balance condition

$$\int_X f(x)\,dx = \int_Y g(y)\,dy, \tag{56}$$

f being continuous almost everywhere and bounded and g being Lipschitz continuous. For convenience, in this paper we extend the function g to the whole domain $\mathbb{R}^d$ in such a manner that $g^{-1/d}\colon \mathbb{R}^d \to \mathbb{R}_+^*$ is bounded and Lipschitz continuous.

In the *quadratic optimal transport problem* between f and g, one aims to solve the minimization problem

$$\inf_{T_\# f = g} \int_X |x - T(x)|^2 f(x)\,dx, \tag{57}$$

where the unknown is a Borel map $T\colon X \to \overline{Y}$ and the constraint $T_\# f = g$ means that for any Borel subset E of Y,

$$\int_{T^{-1}(E)} f(x)\,dx = \int_E g(y)\,dy. \tag{58}$$

In the literature, it is typically assumed that:

$$\text{the set } X \text{ is convex.} \tag{59}$$

For simplicity, we will sometimes assume instead that:

$$\text{the set } X \text{ is strongly convex.} \tag{60}$$

It was proved in [10] (see also [44, Theorem 2.12]) that, under assumption (59), the optimal transport problem (57) admits a solution T which is the gradient of a convex function $u\colon X \to \mathbb{R}$, called the *potential function* of the problem. Then, if u is smooth enough, it may be deduced by performing the change of variables $y = T(x)$ in the right-hand side of (58) that u is solution to the Monge-Ampère equation (1), where

$$A(x,p) = 0, \qquad B(x,p) = \frac{f(x)}{g(p)}. \tag{61}$$

Additionally, the constraint that $T(x) = Du(x) \in \overline{Y}$, for any $x \in X$, may be written as (24), where for any $x \in \overline{X}$,

$$P(x) = Y. \tag{62}$$

Note that in this setting, a possible choice of function φ in Proposition 3.6 is given by $\varphi(x) := \langle x, y_0 \rangle$, for some $y_0 \in Y$.

Remark 5.1 (General optimal transport). In the general optimal transport problem, a cost function $c \in C^2(\mathbb{R}^d \times \mathbb{R}^d)$ is given, and one aims to solve

$$\inf_{T_\# f = g} \int_X c(x, T(x)) f(x)\,dx. \tag{63}$$

If c is defined by $c(x,y) = |x-y|^2$, this problem reduces to (57). It is also equivalent to (57) when $c(x,y) = -\langle x, y \rangle$, as follows directly from the equality $|x-y|^2 = |x|^2 + |y|^2 - 2\langle x, y \rangle$.

Under suitable assumptions (see [20, 37]), there exists a solution $T\colon X \to \overline{Y}$ to (63) of the form $T(x) = c\text{-exp}_x(Du(x))$, where for any $x \in X$ and p, $y \in \mathbb{R}^d$, the function $c\text{-exp}_x\colon \mathbb{R}^d \to \mathbb{R}^d$ is such that

$$y = c\text{-exp}_x(p) \iff p = -D_x c(x, y), \tag{64}$$

and where the function u (called the *potential function*) is c-convex, in the sense that for any $x_0 \in X$, there exists $y_0 \in \mathbb{R}^d$ and $z_0 \in \mathbb{R}$ such that

$$u(x_0) = -c(x_0, y_0) - z_0, \qquad u(x) \geq -c(x, y_0) - z_0 \quad \text{in } X.$$

If $c(x, y) = -\langle x, y\rangle$, c-convexity coincides with the usual notion of convexity. In the general setting, if u is smooth enough then it may be shown to be a solution to the Monge-Ampère equation (1), with

$$A(x, p) = -D_{xx} c(x, c\text{-exp}_x(p)), \tag{65}$$

$$B(x, p) = \frac{f(x)}{g(c\text{-exp}_x(p))} |\det D_{xy} c(x, c\text{-exp}_x(p))|, \tag{66}$$

and the constraint that $T(x) = c\text{-exp}_x(Du(x)) \in \overline{Y}$, for any $x \in X$, may be written as (24), where for any $x \in \overline{X}$,

$$P(x) = -D_x c(x, Y). \tag{67}$$

Then a suitable choice of function φ in Proposition 3.6 would be $\varphi(x) := -D_x c(x, y_0)$ (or a mollification of it), for some $y_0 \in Y$.

5.2 Weak solutions to the Monge-Ampère equation

If the open set X is convex, and if $u\colon X \to \mathbb{R}$ is a convex function and E is a subset of X, then we denote by $\partial u(E)$ the union $\bigcup_{x \in E} \partial u(x)$, where $\partial u(x)$ is the subgradient of u at point x:

$$\partial u(x) := \{p \in \mathbb{R}^d \mid \forall x' \in X,\, u(x') \geq u(x) + \langle p, x' - x\rangle\}.$$

A notion of weak solutions to the Monge-Ampère equation that is directly related to the optimal transport problem (57) is the one of *Brenier solutions*.

Definition 5.2 (Brenier solution)**.** Assume (59), (61), and (62). A function $u\colon X \to \mathbb{R}$ is a *Brenier solution* to (1) and (24) if (i) it is convex and (ii) $(Du)_{\#} f = g$, in the sense that (58) holds for $T = Du$. It is a *minimal Brenier solution* if moreover $\partial u(X)$ is included in $\overline{Y}$.

Brenier solutions are a standard notion. Note that their definition allows that $Du(x) \notin \overline{Y}$, typically at points where $f(x) = 0$. Minimal Brenier solutions were introduced in [4] to prevent this and to guarantee uniqueness of solutions up to addition of a constant, as explained in the proof of [4, Proposition 3.1] (the proof uses the assumptions that Y is convex and g is nonnegative in Y):

Theorem 5.3 (Adapted from [4, Proposition 3.1])**.** *Assume* (59), (61), *and* (62). *If u, $v\colon X \to \mathbb{R}$ are two minimal Brenier solutions to* (1) *and* (24), *then there exists $\xi \in \mathbb{R}$ such that $v = u + \xi$.*

For any function $u\colon \mathbb{R}^d \to \overline{\mathbb{R}}$, let us denote by $u^c\colon \mathbb{R}^d \to \overline{\mathbb{R}}$ its Legendre-Fenchel transform, which we recall is defined by

$$u^c(y) := \sup_{x \in \mathbb{R}^d} (\langle x, y\rangle - u(x)).$$

If u is only defined in X (respectively $\overline{X}$), we define u^c in the same manner after having extended u with value $+\infty$ outside X (respectively $\overline{X}$). In addition to the convex envelope $u^{cc}\colon \mathbb{R}^d \times \overline{\mathbb{R}}$ of u, let us define the function $u_Y^{cc}\colon \mathbb{R}^d \to \overline{\mathbb{R}}$ by

$$u_Y^{cc}(x) := \sup_{y\in Y}(\langle x, y\rangle - u^c(y)).$$

One motivation for the definition of u_Y^{cc} (which is similar to the definition of the function $\tilde{u}_n$ in [4, section 5.1]) is that under the assumptions (59), (61), and (62), if $u\colon X \to \mathbb{R}$ is a Brenier solution to (1) and (24), then u_Y^{cc} is a minimal Brenier solution to (1) and (24).

Another standard notion of solutions to (1) and (24) is the one of *Aleksandrov solutions*:

Definition 5.4 (Aleksandrov solution)**.** Assume (59), (61), and (62). A function $u\colon X \to \mathbb{R}$ is an *Aleksandrov solution* to (1) and (24) if (i) it is convex and (ii) for any Borel subset E of X,

$$\int_E f(x)\,dx = \int_{Y\cap\partial u(E)} g(y)\,dy.$$

It is a *minimal Aleksandrov solution* to (1) and (24) if moreover $\partial u(X) \subset \overline{Y}$.

In our setting, Brenier and Aleksandrov solutions coincide, see for instance [24] (noting that the relevant part [24, Section 1] is not specific to the dimension two):

Proposition 5.5. *Assume* (59), (61), *and* (62). *Then* $u\colon X \to \mathbb{R}$ *is a Brenier solution (respectively minimal Brenier solution) to* (1) *and* (24) *if and only if it is an Aleksandrov solution (respectively minimal Aleksandrov solution) to* (1) *and* (24).

This is related to the fact that Y is convex and g is nonnegative in Y, and that this does not remain true in more general settings.

We will also need to use the notion of Aleksandrov solution to the Monge-Ampère equation equipped with the Dirichlet boundary condition

$$u(x) = \psi(x) \quad \text{on } \partial X. \tag{68}$$

Definition 5.6 (Aleksandrov solution to the Dirichlet problem)**.** Assume (59) and (61). A function $u\colon \overline{X} \to \mathbb{R}$ is an *Aleksandrov solution* to (1) and (68) if (i) it is convex continuous with $u(x) = \psi(x)$ on ∂X and (ii) for any Borel subset E of X,

$$\int_E f(x)\,dx = \int_{\partial u(E)} g(y)\,dy.$$

If $u\colon \overline{X} \to \mathbb{R}$ is continuous and is a minimal Aleksandrov solution to (1) and (24), then it is an Aleksandrov solution to (1) and (68) with $\psi = u_{|\partial X}$; however, this does not remain true if u is not minimal.

Below is the adaptation of [29, Theorem 1.6.2] to our setting. For simplicity, it is assumed that $g(p) = 1$ for any $p \in \mathbb{R}^d$, which turns (1) into the basic Monge-Ampère equation $\det_+(D^2u(x)) = f(x)$. Note however that we only use Theorem 5.7 as an intermediary result and that our convergence result, Theorem 5.25, is not limited to the case $g(p) = 1$.

Theorem 5.7 (Adapted from [29, Theorem 1.6.2])**.** *Assume* (61), *that* X *is strictly convex,* $g(p) = 1$ *for any* $p \in \mathbb{R}^d$, *and* $\psi\colon \partial X \to \mathbb{R}$ *is continuous. Then there exists a unique Aleksandrov solution* $u\colon \overline{X} \to \mathbb{R}$ *to* (1) *and* (68).

5.3 Reformulation of the Monge-Ampère equation

Let us now study the reformulation of the Monge-Ampère equation (1) in the form (2), in the setting of quadratic optimal transport. We sum up the idea of the reformulation in the following proposition:

Proposition 5.8. *Let $b \geq 0$ and $M \in \mathcal{S}_d^+$. Then*

$$\max_{\substack{\mathcal{D}\in\mathcal{S}_d^+ \\ \mathrm{Tr}(\mathcal{D})=1}} L_{\mathcal{D}}(b,M) \leq 0 \iff b \leq \det M, \tag{69}$$

$$\max_{\substack{\mathcal{D}\in\mathcal{S}_d^+ \\ \mathrm{Tr}(\mathcal{D})=1}} L_{\mathcal{D}}(b,M) \geq 0 \iff b \geq \det M. \tag{70}$$

Proof. We refer to [35, Lemma 3.2.2] for the proof of the equivalence

$$\max_{\substack{\mathcal{D}\in\mathcal{S}_d^+ \\ \mathrm{Tr}(\mathcal{D})=1}} L_{\mathcal{D}}(b,M) = 0 \iff b = \det M. \tag{71}$$

Also, the first equality in (5) is proved in [35, Lemma 3.2.1] (it is related to the inequality of arithmetic and geometric means applied to eigenvalues of the product $\mathcal{D}^{1/2}M\mathcal{D}^{1/2}$), while the second one follows from the identity

$$\{\mathcal{D} \in \mathcal{S}_d^{++} \mid \det \mathcal{D} = 1\} = \{(\det \mathcal{D})^{-1/d}\mathcal{D} \mid \mathcal{D} \in \mathcal{S}_d^{++},\, \mathrm{Tr}(\mathcal{D}) = 1\}.$$

From (5), we deduce that

$$\begin{aligned}
b \leq \det M &\iff db^{1/d} - d(\det M)^{1/d} \leq 0 \\
&\iff \sup_{\substack{\mathcal{D}\in\mathcal{S}_d^{++} \\ \mathrm{Tr}(\mathcal{D})=1}} (db^{1/d} - (\det \mathcal{D})^{-1/d}\langle \mathcal{D}, M\rangle) \leq 0 \\
&\iff \sup_{\substack{\mathcal{D}\in\mathcal{S}_d^{++} \\ \mathrm{Tr}(\mathcal{D})=1}} (db^{1/d}(\det \mathcal{D})^{1/d} - \langle \mathcal{D}, M\rangle) \leq 0 \\
&\iff \sup_{\substack{\mathcal{D}\in\mathcal{S}_d^{++} \\ \mathrm{Tr}(\mathcal{D})=1}} L_{\mathcal{D}}(b,M) \leq 0.
\end{aligned}$$

Then (69) follows from the continuity of $L_{\mathcal{D}}(b,M)$ with respect to $\mathcal{D} \in \mathcal{S}_d^+$, and (70) follows from (69) and (71). □

First we prove that Aleksandrov solutions to the Monge-Ampère equation are viscosity solutions to its reformulation.

Proposition 5.9. *Assume* (59) *and* (61). *If, for some function $\psi \in C(\partial X)$, $u\colon \overline{X} \to \mathbb{R}$ is an Aleksandrov solution to* (1) *and* (68), *then u is a viscosity solution to* (2).

The proof is an adaptation of the one of [29, Proposition 1.3.4]. It uses [29, Lemma 1.4.1], which we recall below in our setting:

Lemma 5.10 (Adapted from [29, Lemma 1.4.1])**.** *Assume* (59). *Let u, $v\colon X \to \mathbb{R}$ be convex and let E be an open set such that $\overline{E} \subset X$. If $u \leq v$ in E and $u = v$ on ∂E, then $\partial v(E) \subset \partial u(E)$.*

Proof of Proposition 5.9. We adapt the proof of [29, Proposition 1.3.4], which is a particular case of this proposition.

First let us show that u is a viscosity subsolution to (2). Let $\varphi \in C^2(X)$, and let $x_0 \in X$ be a local maximum of $u - \varphi$. Since u is convex, $D^2\varphi(x)$ must be positive semidefinite. We may assume without loss of generality that φ is convex, that $\varphi(x_0) = u(x_0)$, and that x_0 is a strict local maximum. For any small $\varepsilon > 0$, there exists an open set S_ε such that $\overline{S}_\varepsilon \subset X$, $\varphi \leq u + \varepsilon$ in S_ε, $\varphi = u + \varepsilon$ on ∂S_ε, and $\lim_{\varepsilon \to 0} d_H(S_\varepsilon, \{x_0\}) = 0$ (see [29] for detail). By Lemma 5.10, $\partial u(S_\varepsilon) = \partial(u + \varepsilon)(S_\varepsilon) \subset \partial\varphi(S_\varepsilon)$. Thus, since u is an Aleksandrov solution,

$$\int_{S_\varepsilon} f(x)\,dx = \int_{\partial u(S_\varepsilon)} g(y)\,dy \leq \int_{\partial\varphi(S_\varepsilon)} g(y)\,dy = \int_{S_\varepsilon} g(D\varphi(x)) \det D^2\varphi(x)\,dx.$$

Passing to the limit in ε, we deduce that $f_*(x_0) \leq g(D\varphi(x_0)) \det D^2\varphi(x_0)$. By Proposition 5.8, it follows that $(F_{\mathrm{MA}})_*(x_0, D\varphi(x_0), D^2\varphi(x_0)) \leq 0$, and thus that u is a viscosity subsolution to (2).

Now let us show that u is a viscosity supersolution to (2). Let $\varphi \in C^2(X)$, and let $x_0 \in X$ be a local minimum of $u - \varphi$. If there exists a unit vector $e \in \mathbb{R}^d$ such that $\langle e, D^2\varphi(x_0)e\rangle \leq 0$, then choosing $\mathcal{D} = e \otimes e$ in the maximum in the definition (8) of the operator F_{MA} yields

$$(F_{\mathrm{MA}})^*(x_0, D\varphi(x_0), D^2\varphi(x_0)) \geq -\langle e, D^2\varphi(x_0)e\rangle \geq 0.$$

If on the contrary $D^2\varphi(x_0)$ is positive definite, then we may assume without loss of generality that φ is convex, that $\varphi(x_0) = u(x_0)$, and that x_0 is a strict local minimum. By the same reasoning as above, we prove that $f^*(x_0) \geq g(D\varphi(x_0)) \det D^2\varphi(x_0)$, and we deduce using Proposition 5.8 that $(F_{\mathrm{MA}})^*(x_0, D\varphi(x_0), D^2\varphi(x_0)) \geq 0$. Therefore u is a viscosity supersolution to (2). □

In order to prove convergence of a family of monotone numerical schemes for the Monge-Ampère equation, we need to study under which conditions viscosity solutions to (43) are minimal Aleksandrov solutions to (1) and (24), and in particular what happens when $\alpha \neq 0$. Thus the remaining part of section 5.3 is devoted to the proof of the two following theorems:

Theorem 5.11. *Assume* (59), (61), *and* (62). *If* $u\colon \overline{X} \to \mathbb{R}$ *is a viscosity subsolution to* (43) *with* $\alpha \geq 0$, *then* $\alpha = 0$ *and* u *is a minimal Aleksandrov solution to* (1) *and* (24).

Theorem 5.12. *Assume* (60) *to* (62). *If* $u\colon \overline{X} \to \mathbb{R}$ *is a viscosity supersolution to* (43) *with* $\alpha \leq 0$, *then* $\alpha = 0$ *and* u_Y^{cc} *is a minimal Aleksandrov solution to* (1) *and* (24).

Note also those two theorems are particularly strong results, since they apply to viscosity subsolutions and supersolutions and not only to viscosity solutions as one would have expected. In the particular case of viscosity solutions, one has the following immediate corollary, which is also particularly strong since it does not assume that $\alpha = 0$, but instead proves this equality:

Corollary 5.13. *Assume* (60) *to* (62). *If* $u\colon \overline{X} \to \mathbb{R}$ *is a viscosity solution to* (43) *for some* $\alpha \in \mathbb{R}$, *then* $\alpha = 0$ *and* u *is a minimal Aleksandrov solution to* (1) *and* (24).

Proof. Since u is a viscosity solution, it is both a viscosity subsolution and supersolution. By Theorem 5.12, if $\alpha \leq 0$, then $\alpha = 0$. This means that in any case $\alpha \geq 0$. Therefore Theorem 5.11 applies, and concludes the proof. □

Corollary 5.13 is the main original argument that we use in the proof of our convergence result Theorem 5.25, in combination with standard arguments [1] about the convergence of monotone schemes for degenerate elliptic equations.

Note that in the proof of Corollary 5.13, we did not use the part of Theorem 5.12 about u_Y^{cc} being a minimal Aleksandrov solution to (1) and (24). We mention this fact nevertheless in the statement of Theorem 5.12 since it is a direct consequence of our proof that $\alpha = 0$.

Remark 5.14 (Sketch of proof of Theorems 5.11 and 5.12). The rigorous proofs of Theorems 5.11 and 5.12 are delayed to the end of section 5.3, but let us first explain the main arguments that we use in those proofs. In this remark we will only discuss the proof that u (respectively u_Y^{cc}) is a minimal Aleksandrov solution to (1) and (24); in order to prove that $\alpha = 0$ one just needs to sufficiently refine the arguments below.

Theorem 5.11 is very close to [26, Theorem 2.1] (although the considered reformulation of the Monge-Ampère equation is not the same) and thus we follow the same sketch of proof. If u is a viscosity subsolution to (43) with $\alpha \geq 0$, then it is a viscosity subsolution to both (2) and (26). From the fact that u is a subsolution to (2), we deduce that it is a convex function, see Lemma 5.16. The fact that u is a subsolution to (24) means that the optimal transport boundary condition $\partial u(X) \subset \overline{Y}$ is satisfied, see Lemma 5.17. We deduce from the optimal transport boundary condition the inequality $\int_{\partial u(X)} g(y)\,dy \leq \int_Y g(y)\,dy = \int_X f(x)\,dx$. On the other hand, we are able to deduce from the fact that u is a viscosity subsolution to the (reformulated) Monge-Ampère equation (2) that for any Borel set $E \subset X$, one has the inequality $\int_{\partial u(E)} g(y)\,dy \geq \int_E f(x)\,dx$, which is the inequality variant of the equality in the definition of Aleksandrov solutions. Observe that the two previous inequalities are in competition with each other. Thus we are able to show that they are actually equalities and that u is therefore a minimal Aleksandrov solution to (1) and (24).

Contrary to Theorem 5.11, no counterpart to Theorem 5.12 is established in [26]. The proof of Theorem 5.11 does not translate directly to the setting of Theorem 5.12. This is because, for an arbitrary viscosity supersolution u to (43) with $\alpha \leq 0$, on the one hand one cannot expect u to be a viscosity supersolution to (2) (contrary to the case of subsolutions), and on the other hand u is not even guaranteed to be convex, so for instance the optimal transport boundary condition $\partial u(X) \subset \overline{Y}$ does not make sense. We get around those difficulties by considering the modified convex envelope u_Y^{cc} instead of the function u itself. By construction, u_Y^{cc} is guaranteed to be convex and to satisfy the optimal transport boundary condition $\partial u_Y^{cc}(X) \subset \overline{Y}$. By analyzing the meaning of the Dirichlet boundary condition $u - \infty \geq 0$ in the viscosity sense, we are able to prove the converse inclusion $Y \subset \partial u_Y^{cc}(X)$, see Lemma 5.19. We are also able to show that, contrary to u, u_Y^{cc} is guaranteed to be a viscosity supersolution to (2), see Lemma 5.21. From this point the proof of Theorem 5.12 is similar to the one of Theorem 5.11, although the rigorous proof of the inequality $\int_{\partial u(E)} g(y)\,dy \leq \int_E f(x)\,dx$, for Borel sets $E \subset X$, is a bit more technical than its counterpart in the setting of Theorem 5.11 and involves Lemma 5.22 in addition to Proposition 5.8 and Lemma 5.24.

Let us now turn to the intermediary results needed in the proofs of Theorems 5.11 and 5.12

We will need the following comparison principle for equation (2). The assumptions that we make on the function B are more restrictive than in the rest of the paper, but this does not affect the generality of our main results since we will only need to apply this comparison principle to the case of a constant function B, see Lemma 5.22.

Proposition 5.15 (Comparison principle)**.** *Assume that $B^{1/d}$ is continuous, in addition to being Lipschitz continuous with respect to its second variable, uniformly with respect to its first variable. Then there exists $r > 0$ such that the following holds: for any open subset E of X such that $\operatorname{diam}(E) \leq r$ and for any respectively upper and lower semicontinuous functions $\overline{u}$, $\underline{u}\colon \overline{E} \to \mathbb{R}$, if $\overline{u}$ and $\underline{u}$ are respectively a viscosity subsolution and a viscosity supersolution to*

$$F_{\mathrm{MA}}(x, Du(x), D^2u(x)) = 0 \quad \text{in } E,$$

and if $\overline{u} \leq \underline{u}$ on ∂E, then $\overline{u} \leq \underline{u}$ in E.

Proof. Let $x_0 \in E$. For any $\varepsilon > 0$, let $\overline{u}_\varepsilon \colon \overline{E} \to \mathbb{R}$ be defined by

$$\overline{u}_\varepsilon(x) := \overline{u}(x) + \frac{\varepsilon}{2}|x - x_0|^2 - \frac{\varepsilon}{2}\operatorname{diam}(E)^2,$$

so that $\overline{u}_\varepsilon \le \overline{u} \le \underline{u}$ on ∂E. Let $x_1 \in E$, $\varphi \in C^2(E)$, and $\varphi_\varepsilon := \varphi + (\varepsilon/2)|\cdot - x_0|^2$. Then x_1 is a local maximum of $\overline{u}_\varepsilon - \varphi_\varepsilon$ if and only if it is a local maximum of $\overline{u} - \varphi$. For some constant $C > 0$ and for $r = 1/(2C)$, using that $|D\varphi_\varepsilon(x_1) - D\varphi(x_1)| \le r\varepsilon$ and $D^2\varphi_\varepsilon(x_1) = D^2\varphi(x_1) + \varepsilon I_d$, it holds for any $\mathcal{D} \in \mathcal{S}_d^+$ satisfying $\operatorname{Tr}(\mathcal{D}) = 1$ that

$$\begin{aligned}
&L_{\mathcal{D}}\left(B(x, D\varphi_\varepsilon(x)), D^2\varphi_\varepsilon(x) - A(x, D\varphi_\varepsilon(x))\right)\\
&= dB(x, D\varphi_\varepsilon(x))^{1/d}(\det \mathcal{D})^{1/d} - \langle \mathcal{D}, D^2\varphi_\varepsilon(x) - A(x, D\varphi_\varepsilon(x))\rangle\\
&\le dB(x, D\varphi(x))^{1/d}(\det \mathcal{D})^{1/d} - \langle \mathcal{D}, D^2\varphi(x) - A(x, D\varphi(x))\rangle + Cr\varepsilon - \varepsilon\\
&= L_{\mathcal{D}}\left(B(x, D\varphi(x)), D^2\varphi(x) - A(x, D\varphi(x))\right) + Cr\varepsilon - \varepsilon\\
&\le L_{\mathcal{D}}\left(B(x, D\varphi(x)), D^2\varphi(x) - A(x, D\varphi(x))\right) - \varepsilon/2.
\end{aligned}$$

Thus if x_1 is a local maximum of $\overline{u}_\varepsilon - \varphi_\varepsilon$,

$$F_{\mathrm{MA}}(x_1, D\varphi_\varepsilon(x_1), D^2\varphi_\varepsilon(x_1)) \le F_{\mathrm{MA}}(x_1, D\varphi(x_1), D^2\varphi(x_1)) - \varepsilon/2 \le -\varepsilon/2.$$

Then by [16, Theorem 3.3 and section 5.C], $\overline{u}_\varepsilon \le \underline{u}$ in E, and we conclude by letting ε approach zero. □

Notice that we did not need to assume (61); however, if (61) holds, it may be shown that the assumption that $\operatorname{diam}(E) \le r$ is not necessary, see [32, Theorem V.2] for the argument.

We will also need the following lemmas.

Lemma 5.16. *Assume* (59) *and* (61). *If* $u \colon X \to \mathbb{R}$ *is a viscosity subsolution to* (2), *then it is convex.*

Proof. Let $\varphi \in C^2(X)$ and x_0 be a local maximum of $u - \varphi$ in X. Then, using that u is a viscosity subsolution and choosing $\mathcal{D} = e \otimes e$ in the maximum in the definition of F_{MA},

$$0 \ge (F_{\mathrm{MA}})_*(x_0, D\varphi(x_0), D^2\varphi(x_0)) \ge -\min_{|e|=1}\langle e, D^2\varphi(x_0)e\rangle.$$

Thus u is a viscosity subsolution to

$$-\min_{|e|=1}\langle e, D^2u(x_0)e\rangle = 0 \quad \text{in } X.$$

By [41, Theorem 1], it follows that u is convex. □

Lemma 5.17. *Assume* (59) *and* (62). *If* $u \colon X \to \mathbb{R}$ *is a convex viscosity subsolution to* (26), *then* $\partial u(X) \subset \overline{Y}$.

The proof of Lemma 5.17 is a direct transposition to our setting to the one of [26, Lemma 2.5], so we do not reproduce it here.

Lemma 5.18. *Assume* (60), *i.e. that* X *is strongly convex. Then for any* $x_0 \in \partial X$ *and* C, $\varepsilon > 0$, *there exists a convex function* $\psi \in C^2(\overline{X})$ *such that* x_0 *is a local maximum of* ψ *and* $|D\psi(x_0)| \le \varepsilon$, $\det D^2\psi(x_0) \ge C$.

Proof. Since X is strongly convex, there exists $r > 0$ and a unit vector $e \in \mathbb{R}^d$, $|e| = 1$, such that $X \subset B_d(x_0 - re, r)$. Then for any $x \in \overline{X}$, one has $|x - (x_0 - re)|^2 \leq r^2$. Since $|x - (x_0 - re)|^2 = |x - x_0 + re|^2 = |x - x_0|^2 + 2r\langle e, x - x_0\rangle + r^2$, we deduce that $|x - x_0|^2 + 2r\langle e, x - x_0\rangle \leq 0$. Thus x_0 is a local maximum of $|\cdot - x_0|^2 + 2r\langle e, \cdot - x_0\rangle$ in $\overline{X}$. Therefore, using that $\langle e, x - x_0\rangle \leq -|x - x_0|^2/(2r) < 0$ for any $x \in X$, x_0 is also a local maximum in $\overline{X}$ of the convex function $\psi \in C^2(\overline{X})$ defined by

$$\begin{aligned}\psi(x) &:= \frac{\varepsilon}{4r}|x - x_0|^2 + \varepsilon\langle e, x - x_0\rangle + \tilde{C}\langle e, x - x_0\rangle^2 \\ &= \frac{\varepsilon}{4r}\left(|x - x_0|^2 + 2r\langle e, x - x_0\rangle\right) + \frac{\varepsilon}{2}\langle e, x - x_0\rangle + \tilde{C}\langle e, x - x_0\rangle^2,\end{aligned}$$

where $\tilde{C} \in \mathbb{R}$ is an arbitrary constant. We compute that $|D\psi(x_0)| = \varepsilon$ and $\det D^2\psi(x_0) = (\varepsilon/(2r))^{d-1}(\varepsilon/(2r) + 2\tilde{C})$. Choosing $\tilde{C} := (C/2)(2r/\varepsilon)^{d-1}$, we conclude the proof. □

Lemma 5.19. *Assume* (60) *to* (62). *If $u\colon \overline{X} \to \mathbb{R}$ is a viscosity supersolution to* (43) *with $\alpha \leq 0$, then $Y \subset \partial u^{cc}_Y(X) \subset \overline{Y}$.*

Proof. By definition of u^{cc}_Y, one has $\partial u^{cc}_Y(X) \subset \overline{Y}$, and more precisely $\partial u^{cc}_Y(X) = \partial u^{cc}(X) \cap \overline{Y}$. Therefore, it suffices to prove that $Y \subset \partial u^{cc}(X)$.

Let $y_0 \in Y$. Since u is lower semicontinuous, there exists $x_0 \in \overline{X}$ such that $y_0 \in \partial u^{cc}(x_0)$ (meaning that x_0 is a local minimum of $u^{cc} - \langle \cdot, y_0\rangle$) and $u^{cc}(x_0) = u(x_0)$. Let us show that $x_0 \in X$.

Since $u^{cc} \leq u$ in $\overline{X}$, x_0 is a local minimum of $u - \langle \cdot, y_0\rangle$. If $x_0 \in \partial X$, then for any $\varepsilon > 0$, we may build using Lemma 5.18 a convex function $\varphi_\varepsilon \in C^2(\overline{X})$ such that x_0 is a local minimum of $u - \varphi_\varepsilon$ and

$$|D\varphi_\varepsilon(x_0) - y_0| \leq \varepsilon, \qquad \det D^2\varphi_\varepsilon(x_0) > \sup_{y \in \mathbb{R}^d} \frac{f^*(x_0)}{g(y)} \geq \frac{f^*(x_0)}{g(D\varphi_\varepsilon(x_0))}$$

(choose $\varphi_\varepsilon = \langle \cdot, y_0\rangle + \psi$ where ψ is from Lemma 5.18). Then by Proposition 5.8,

$$(F_{\mathrm{MA}})^*(x_0, D\varphi_\varepsilon(x_0), D^2\varphi_\varepsilon(x_0)) < 0.$$

We may choose ε small enough so that $D\varphi_\varepsilon(x_0) \in Y$, and thus $F_{\mathrm{BV2}}(x_0, D\varphi_\varepsilon(x_0)) < 0$. Then

$$(F^\alpha_{\mathrm{MABV2}})^*(x_0, D\varphi_\varepsilon(x_0), D^2\varphi_\varepsilon(x_0)) < 0,$$

which is impossible since u is a viscosity supersolution to (43). Therefore x_0 may not belong to ∂X. □

Lemma 5.20. *Assume* (59). *Let $u\colon X \to \mathbb{R}$ be a convex function satisfying $\partial u(X) \subset \overline{Y}$, and let $\varphi \in C^2(X)$. If x_0 is a local minimum of $u - \varphi$ in X and if $D^2\varphi(x_0)$ is positive definite, then $D\varphi(x_0) \in Y$.*

Proof. Let $e \in \mathbb{R}^d$ be a unit vector and let $t > 0$ be small enough so that $u(x_0) - \varphi(x_0) \leq u(x_0 + te) - \varphi(x_0 + te)$ and $\varphi(x_0 + te) > \varphi(x_0) + t\langle e, D\varphi(x_0)\rangle$. Combining those two inequalities, we get $u(x_0) - \varphi(x_0) < u(x_0 + te) - \varphi(x_0) - t\langle e, D\varphi(x_0)\rangle$, which simplifies to $\langle e, D\varphi(x_0)\rangle < (u(x_0 + te) - u(x_0))/t$.

Let $y \in \partial u(x_0 + te)$. By definition of $\partial u(x_0 + te)$, one has $u(x_0) \geq u(x_0 + te) - t\langle e, y\rangle$. Therefore $\langle e, y\rangle \geq (u(x_0 + te) - u(x_0))/t > \langle e, D\varphi(x_0)\rangle$.

Since $\partial u(X) \subset \overline{Y}$, one has $y \in \overline{Y}$. Thus we showed that for any unit vector $e \in \mathbb{R}^d$, there exists $y \in \overline{Y}$ such that $\langle e, y\rangle > \langle e, D\varphi(x_0)\rangle$. Since Y is convex, it follows that $D\varphi(x_0) \in Y$. □

Lemma 5.21. *Assume* (59), (61), *and* (62). *If* $u\colon \overline{X} \to \mathbb{R}$ *is a viscosity supersolution to* (43) *with* $\alpha \leq 0$, *then* u_Y^{cc} *is a viscosity supersolution to* (2). *Moreover, if* $\alpha < 0$, $\varphi \in C^2(X)$, x_0 *is a local minimum of* $u_Y^{cc} - \varphi$ *in* X, *and* $f^*(x_0) > 0$, *then*

$$(F_{\mathrm{MA}})^*(x_0, D\varphi(x_0), D^2\varphi(x_0)) > 0.$$

Proof. Let $\varphi \in C^2(X)$, and let x_0 be a local minimum of $u_Y^{cc} - \varphi$ in X. Note that $D\varphi(x_0) \in \partial u_Y^{cc}(x_0) \subset \overline{Y}$.

First we consider the case where $u_Y^{cc}(x_0) = u(x_0)$ and $D\varphi(x_0) \in Y$. Since $u_Y^{cc} \leq u$ in X, x_0 is a local minimum of $u - \varphi$ in X. Thus

$$(F^{\alpha}_{\mathrm{MABV2}})^*(x_0, D\varphi(x_0), D^2\varphi(x_0)) \geq 0.$$

Since $D\varphi(x_0) \in Y$, one has

$$F_{\mathrm{BV2}}(x_0, D\varphi(x_0)) < 0.$$

It follows that

$$(F_{\mathrm{MA}})^*(x_0, D\varphi(x_0), D^2\varphi(x_0)) \geq 0,$$

with a strict inequality if $\alpha < 0$.

Now we consider the case where either $u_Y^{cc}(x_0) < u(x_0)$ or $D\varphi(x_0) \in \partial Y$. In this case, there exists a unit vector $e \in \mathbb{R}^d$ such that $\langle e, D^2\varphi(x)e\rangle \leq 0$ (using Lemma 5.20 for the case $D\varphi(x_0) \in \partial Y$). Choosing $\mathcal{D} = (1-\varepsilon)e \otimes e + (\varepsilon/d)I_d$ in the definition of F_{MA} yields

$$(F_{\mathrm{MA}})^*(x_0, D\varphi(x_0), D^2\varphi(x_0)) \geq d\frac{f^*(x_0)^{1/d}}{g(D\varphi(x_0))^{1/d}}\left(1 - \frac{d-1}{d}\varepsilon\right)^{1/d}\varepsilon^{(d-1)/d} - \frac{\varepsilon}{d}\operatorname{Tr}(D^2\varphi(x_0)).$$

If $f^*(x_0) > 0$, we conclude by choosing ε small enough so that the right-hand side is positive. If $f^*(x_0) = 0$, we conclude by letting ε approach zero. □

Lemma 5.22. *Assume* (59) *and* (61). *If* $u\colon X \to \mathbb{R}$ *is a convex viscosity supersolution to* (2), *then for any Borel subset* E *of* X *of Lebesgue measure zero,* $\partial u(E)$ *has Lebesgue measure zero.*

Proof. Let $K > 0$, and let E be a subset of X of Lebesgue measure zero. Then for any $\varepsilon > 0$, there exists an open set $G \subset X$ such that $E \subset G$ and $\mathcal{L}^d(G) \leq \varepsilon$. For any $x \in G$, let $r(x) > 0$ and $S(x) := B_d(x, r(x))$, choosing $r(x)$ small enough so that $\overline{S(x)} \subset G$. By Theorem 5.7, there exists an Aleksandrov solution $v \in C(\overline{S(x)})$ to

$$\begin{cases} \det{}_+ D^2 v(x) = K & \text{in } S(x), \\ v(x) = u(x) & \text{on } \partial S(x). \end{cases}$$

By Proposition 5.9, v is a viscosity solution to (2) with $A(x,p)$ replaced by zero, $B(x,p)$ replaced by K, and X replaced by E. Choosing K large enough, it is easily verified that u is a viscosity supersolution to (2) with the same parameters. Then by Proposition 5.15, up to choosing $r(x)$ smaller, $v \leq u$ in $S(x)$. Since $u = v$ on $\partial S(x)$, Lemma 5.10 shows that $\partial u(S(x)) \subset \partial v(S(x))$. Thus

$$\mathcal{L}^d(\partial u(S(x))) \leq \mathcal{L}^d(\partial v(S(x))) = K\mathcal{L}^d(S(x)).$$

Let $\delta < 1/5$ (for instance $\delta = 1/6$) and for any $x \in G$, let $S_\delta(x) \subset G$ be defined by $S_\delta(x) := B_d(x, \delta r(x))$. Then by Vitali's covering theorem [22, Theorem 1.5.1], there exists a countable family

$(x_i)_{i\in\mathbb{N}}$ of points of G such that $\bigcup_{x\in G} S_\delta(x) \subset \bigcup_{i\in\mathbb{N}} S(x_i)$ and balls of the family $(S_\delta(x_i))_{i\in\mathbb{N}}$ are all disjoint. Since $E \subset G = \bigcup_{x\in G} S_\delta(x)$, we deduce that $\partial u(E) \subset \bigcup_{i\in\mathbb{N}} \partial u(S(x_i))$ and thus

$$\begin{aligned}\mathcal{L}^d(\partial u(E)) &\le \sum_{i\in\mathbb{N}} \mathcal{L}^d(\partial u(S(x_i))) \le K \sum_{i\in\mathbb{N}} \mathcal{L}^d(S(x_i)) = K\delta^{-d} \sum_{i\in\mathbb{N}} \mathcal{L}^d(S_\delta(x_i)) \le K\delta^{-d}\mathcal{L}^d(G)\\ &\le K\delta^{-d}\varepsilon.\end{aligned}$$

We conclude by letting ε approach zero that $\mathcal{L}^d(\partial u(E)) = 0$. □

Lemma 5.23. *Assume* (59). *If $u\colon X \to \mathbb{R}$ is convex, then the set*

$$\{y \in \mathbb{R}^d \mid \exists x_1, x_2 \in X,\ x_1 \neq x_2 \text{ and } y \in \partial u(x_1) \cap \partial u(x_2)\}$$

has Lebesgue measure zero.

Proof. This standard result follows directly from the facts that u^c is not twice differentiable at points of this set (since $\{x_1, x_2\} \subset \partial u^c(y)$) and that u^c, as a convex, hence locally Lipschitz function, is differentiable almost everywhere, by Rademacher's theorem [22, Theorem 3.1.2]. □

In the lemma below, the right-hand side in (72) is to be understood as the integral of function which coincides almost everywhere with $g(Du(\cdot)) \det D^2u(\cdot)$. Indeed, the convex function u is twice differentiable almost everywhere by Aleksandrov's theorem [22, Theorem 6.4.1]. In particular, points where u is not twice differentiable do not contribute to the integral in the right-hand side, while they do contribute to the one in the left-hand side.

Lemma 5.24. *Assume* (59). *If $u\colon X \to \mathbb{R}$ is convex, then for any Borel subset E of X,*

$$\int_{\partial u(E)} g(y)\,dy \ge \int_E g(Du(x)) \det D^2u(x)\,dx. \tag{72}$$

If moreover $\partial u(E')$ has Lebesgue measure zero for any subset E' of X of Lebesgue measure zero, then the above inequality is an equality.

Proof. Since u is convex, its gradient Du belongs to $BV_{\mathrm{loc}}(X;\mathbb{R}^d)$, see [22, Theorem 6.3.3]. By [22, Theorem 6.6.2], for any $k \in \mathbb{N}^*$, there exists a subset E_k of E such that Du is Lipschitz continuous in E_k and $\mathcal{L}^d(E \setminus E_k) \le 1/k$. We define $\tilde{E} := \bigcup_{k=1}^\infty E_k$ and, for any $k \in \mathbb{N}^*$, $\tilde{E}_k := E_k \setminus (\bigcup_{i=1}^{k-1} E_i)$.

Using Lemma 5.23,

$$\begin{aligned}\int_{\partial u(E)} g(y)\,dy \ge \int_{\partial u(\tilde{E})} g(y)\,dy &= \sum_{k=1}^\infty \int_{\partial u(\tilde{E}_k)} g(y)\,dy = \sum_{k=1}^\infty \int_{Du(\tilde{E}_k)} g(y)\,dy\\ &= \sum_{k=1}^\infty \int_{\mathbb{R}^d} \left[\sum_{x\in(Du)^{-1}(\{y\})} \mathbb{1}_{\tilde{E}_k(x)} g(Du(x))\right] dy\end{aligned}$$

(here $(Du)^{-1}(\{y\})$ is a singleton for almost every y), with equality if $\partial u(E \setminus \tilde{E})$ has Lebesgue measure zero (note that $E \setminus \tilde{E}$ always has Lebesgue measure zero).

By the change of variables formula [22, Theorem 3.3.2], which is a corollary of the area formula of geometric measure theory, for any $k \in \mathbb{N}^*$,

$$\int_{\mathbb{R}^d} \left[\sum_{x\in(Du)^{-1}(\{y\})} \mathbb{1}_{\tilde{E}_k(x)} g(Du(x))\right] dy = \int_{\tilde{E}_k} g(Du(x)) \det D^2u(x)\,dx.$$

It follows that

$$\int_{\partial u(\tilde{E})} g(y)\,dy = \sum_{k=1}^{\infty} \int_{\tilde{E}_k} g(Du(x)) \det D^2u(x)\,dx = \int_E g(Du(x)) \det D^2u(x)\,dx,$$

which concludes the proof. □

Let us now prove the main Theorem 5.11 and Theorem 5.12.

Proof of Theorem 5.11. If $u\colon \overline{X} \to \mathbb{R}$ is a viscosity subsolution to (43) with $\alpha \geq 0$, it is both a viscosity subsolution to (2) and (26). Thus by Lemma 5.16 and Lemma 5.17, it is convex in X and $\partial u(X) \subset \overline{Y}$.

By Aleksandrov's theorem [22, Theorem 6.4.1], u is twice differentiable almost everywhere. Thus it is almost everywhere a classical subsolution to (43). It follows that for almost every $x \in X$, $F_{\mathrm{MA}}(x, Du(x), D^2u(x)) \leq 0$, with a strict inequality if $\alpha > 0$. Then, using Proposition 5.8, for any Borel subset E of X,

$$\int_E f(x)\,dx \leq \int_E g(Du(x)) \det D^2u(x)\,dx,$$

with a strict inequality if $\alpha > 0$ and E has positive Lebesgue measure.

By Lemma 5.24, we deduce that

$$\int_E f(x)\,dx \leq \int_{\partial u(E)} g(y)\,dy,$$

with a strict inequality if $\alpha > 0$ and E has positive Lebesgue measure. The same is true when replacing E by $X \setminus E$, and by Lemma 5.23, $\partial u(E) \cap \partial u(X \setminus E)$ has Lebesgue measure zero. Thus

$$\int_X f(x)\,dx = \int_E f(x)\,dx + \int_{X\setminus E} f(x)\,dX \leq \int_{\partial u(E)} g(y)\,dy + \int_{\partial u(X\setminus E)} g(y)\,dy = \int_{\partial u(X)} g(y)\,dy,$$

with a strict inequality if $\alpha > 0$, since at least one of the sets E and $X \setminus E$ has positive Lebesgue measure. On the other hand, since $\partial u(X) \subset \overline{Y}$, one has the converse inequality

$$\int_{\partial u(X)} g(y)\,dy \leq \int_Y g(y)\,dy = \int_X f(x)\,dx.$$

Therefore the case $\alpha > 0$ cannot happen, and moreover

$$\int_E f(x)\,dx = \int_{\partial u(E)} g(y)\,dy, \qquad \int_{X\setminus E} f(x)\,dx = \int_{\partial u(X\setminus E)} g(y)\,dy,$$

from which it follows that u is a minimal Aleksandrov solution to (1) and (24). □

Proof of Theorem 5.12. When applicable, we follow the same sketch of proof as for Theorem 5.11. Let $u\colon \overline{X} \to \mathbb{R}$ be a viscosity supersolution to (43) with $\alpha \leq 0$. By Aleksandrov's theorem [22, Theorem 6.4.1], u_Y^{cc} is twice differentiable almost everywhere. Then by Lemma 5.21, for almost every $x \in X$, one has $F_{\mathrm{MA}}(x, Du_Y^{cc}(x), D^2u_Y^{cc}(x)) \geq 0$, with a strict inequality if $\alpha < 0$ and $f(x) > 0$. Using Proposition 5.8, for any Borel subset E of X,

$$\int_E f(x)\,dx \geq \int_E g(Du_Y^{cc}(x)) \det D^2u_Y^{cc}(x)\,dx,$$

with a strict inequality if $\alpha < 0$ and $\mathcal{L}^d(\{x \in E \mid f(x) > 0\}) > 0$.

By Lemma 5.22 and Lemma 5.24, we deduce that

$$\int_E f(x)\,dx \geq \int_{\partial u_Y^{cc}(E)} g(y)\,dy,$$

with a strict inequality if $\alpha < 0$ and $\mathcal{L}^d(\{x \in E \mid f(x) > 0\}) > 0$. The same is true when replacing E by $X \setminus E$, thus

$$\int_X f(x)\,dx = \int_E f(x)\,dx + \int_{X\setminus E} f(x)\,dX \geq \int_{\partial u_Y^{cc}(E)} g(y)\,dy + \int_{\partial u_Y^{cc}(X\setminus E)} g(y)\,dy \geq \int_{\partial u_Y^{cc}(X)} g(y)\,dy.$$

Also note that at least one of the two conditions $\mathcal{L}^d(\{x \in E \mid f(x) > 0\}) > 0$ and $\mathcal{L}^d(\{x \in X \setminus E \mid f(x) > 0\}) > 0$ is satisfied, thus one has a strict inequality if $\alpha < 0$. On the other hand, since by Lemma 5.19 $Y \subset \partial u_Y^{cc}(X) \subset \overline{Y}$, one has the equality

$$\int_{\partial u_Y^{cc}(X)} g(y)\,dy = \int_Y g(y)\,dy = \int_X f(x)dx.$$

Therefore the case $\alpha < 0$ cannot happen, and moreover

$$\int_E f(x)\,dx = \int_{\partial u_Y^{cc}(E)} g(y)\,dy, \qquad \int_{X\setminus E} f(x)\,dx = \int_{\partial u_Y^{cc}(X\setminus E)} g(y)\,dy,$$

from which it follows that u_Y^{cc} is a minimal Aleksandrov solution to (1) and (24). □

5.4 Convergence

We are now able to prove convergence of a family of numerical schemes (which includes the scheme (30), see section 3) for the Monge-Ampère equation, in the setting of quadratic optimal transport.

Theorem 5.25 (Convergence)**.** *Assume* (60) *to* (62)*. If the scheme* (38) *is monotone, consistent with equation* (43)*, and equicontinuously stable (in the sense of Definition 2.12), and if for any small $h > 0$, there exists a solution $(\alpha_h, u_h) \in \mathbb{R} \times \mathbb{R}^{\mathcal{G}_h}$ to* (38) *satisfying $u_h[0] = 0$, then as h approaches zero, α_h converges to zero and u_h converges uniformly to the unique minimal Aleksandrov solution (or equivalently minimal Brenier solution) $u\colon X \to \mathbb{R}$ to* (1) *and* (24) *satisfying $u(0) = 0$.*

Proof. Let $(h_n)_{n\in\mathbb{N}}$ be a sequence of small discretization steps $h_n > 0$ converging to zero. Since (38) is equicontinuously stable, the sequence $(\alpha_{h_n})_{n\in\mathbb{N}}$ is bounded, and $(u_{h_n})_{n\in\mathbb{N}}$ is uniformly bounded and uniformly equicontinuous. Then by the Arzelà-Ascoli theorem, up to extracting a subsequence, α_{h_n} converges to some $\alpha \in \mathbb{R}$ and u_{h_n} converges uniformly to some continuous function $u\colon \overline{X} \to \mathbb{R}$, satisfying $u(0) = 0$. By Corollary 2.13, u is a viscosity solution to (43). By Corollary 5.13, $\alpha = 0$ and u is the minimal Aleksandrov solution to (1) and (24), which concludes the proof. □

6 Numerical experiments

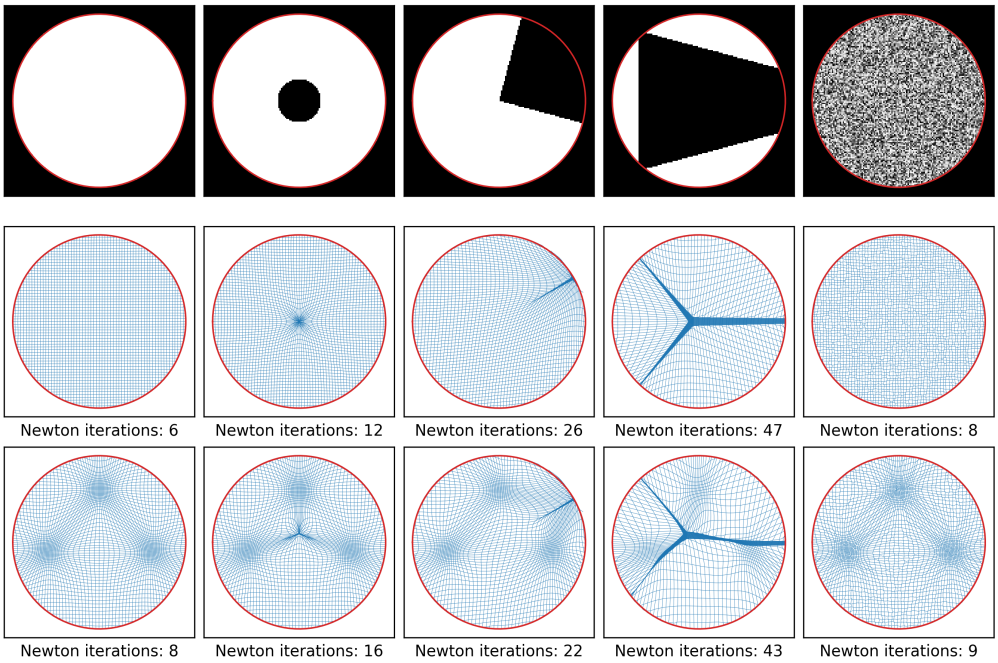

Figure 1: Image of a Cartesian grid by the approximated optimal transport maps, for the quadratic optimal transport problem with some given source and target densities.

6.1 Approximation of optimal transport maps for some quadratic optimal transport problems

We apply the scheme (30) to the numerical resolution of some problems of the form (57), see Figure 1. The problems considered are inspired by the numerical experiments in [4]. The source and target domains X and Y are chosen as the unit disk $B_2(0,1)$. The source density $f\colon \overline{B_2(0,1)} \to \mathbb{R}_+$ is chosen among the ones depicted in the top row of Figure 1 and may have a non-convex or non-connected support, while the target density $g\colon \overline{B_2(0,1)} \to \mathbb{R}_+^*$ (extended to all of $\mathbb{R}^2$ for numerical purposes, as explained in section 5.1) is either the uniform density $g\colon y \mapsto 1/\pi$ or the following combination of Gaussian densities and of a small uniform density:

$$g\colon y \mapsto \frac{\rho + \sum_{i=1}^3 \exp\left(-\frac{|y-\hat{y}_i|^2}{2\sigma^2}\right)}{\int_{B_2(0,1)} \rho + \sum_{i=1}^3 \exp\left(-\frac{|y-\hat{y}_i|^2}{2\sigma^2}\right) dy}, \tag{73}$$

where $\rho := 0.1$, $\sigma := 0.1$, $\hat{y}_1 := (0, 0.6)$, $\hat{y}_2 := (-0.6, -0.1)$, and $\hat{y}_3 := (0.6, -0.1)$.

In all our numerical experiments, we choose the discretization step $h > 0$ as $h = 2/N$, for some $N \in 2\mathbb{N}^*$, so that the Cartesian grid $[-1,1]^2 \cap h\mathbb{Z}^2$ contains exactly $(N+1)^2$ points. We define the Cartesian grid $\mathcal{G}_h := X \cap h\mathbb{Z}^2$, as well as the smaller grid

$$\tilde{\mathcal{G}}_h := \{x \in \mathcal{G}_h \mid \forall i \in \{1,2\},\ x + he_i \in \mathcal{G}_h \text{ and } x - he_i \in \mathcal{G}_h\}.$$

In Figure 1, we choose $N = 128$. Following Appendix B and in particular Table 1, we choose $V_h = V^\mu$ with $\mu = 2 + \sqrt{5} \approx 4.236$. We solve the numerical scheme (30) on the grid $\mathcal{G}_h$, in the setting described by (61) and (62), and we denote by $(\alpha_h, u_h) \in \mathbb{R} \times \mathbb{R}^{\mathcal{G}_h}$ the solution to the

scheme. We approximate the optimal transport map $T\colon X \to \overline{Y}$ by $D_h u_h\colon \tilde{\mathcal{G}}_h \to \mathbb{R}^d$, where D_h is the centered finite difference operator defined in (13). The grids displayed in Figure 1 are the image of $\tilde{\mathcal{G}}_h$ (coarsened for readability) by the approximate transport map $D_h u_h$ in each of the settings considered.

Remark 6.1 (Difference between the theoretical and experimental settings). In the definitions (17) and (18) of the discrete operators A_h^e and B_h, the scheme (30) involves some parameters $a_{\min} \leq 0$ and a_{LF}, $b_{\mathrm{LF}} \geq 0$. Although they do not fit in the theoretical setting, we use in all our experiments the values $a_{\min} = -\infty$ and $a_{\mathrm{LF}} = b_{\mathrm{LF}} = 0$, which simplify the expression of the scheme and improve its consistency, see Remark 3.4. We did not observe any practical difficulties related to this choice in the experiments that we considered.

Remark 6.2 (Solving the numerical scheme). We solve the scheme (30) using the Newton method. In practice, in all our numerical experiments instead of solving

$$S_{\mathrm{MA}}^{h,\alpha} u[x] \vee S_{\mathrm{BV2}}^{h} u[x] = 0 \quad \text{in } \mathcal{G}_h,$$

we solve the equivalent scheme

$$S_{\mathrm{MA}}^{h,\alpha} u[x] \vee \kappa S_{\mathrm{BV2}}^{h} u[x] = 0 \quad \text{in } \mathcal{G}_h$$

with $\kappa = 20$, since we observe that better convergence of the Newton method tends to be achieved when rescaling the contribution of the discretization of the optimal transport boundary condition with respect to the one of the discretization of the Monge-Ampère equation. The Newton method is applied to finding a zero of the function $(\alpha, u) \mapsto S_{\mathrm{MA}}^{h,\alpha} u[x] \vee \kappa S_{\mathrm{BV2}}^{h} u[x]$ over the hyperplane $\{(\alpha, u) \in \mathbb{R} \times \mathbb{R}^{\mathcal{G}_h} \mid u[0] = 0\}$. In Figure 1, we display the number of iterations required in order to achieve convergence of the Newton method for each of the problems considered, with initialization $u[x] = |x|^2$ and with the stopping criterion

$$\max_{x \in \mathcal{G}_h} |S_{\mathrm{MA}}^{h,\alpha} u[x] \vee \kappa S_{\mathrm{BV2}}^{h} u[x]| < 10^{-8}.$$

We observe that more iterations seem to be required when the size of the support of the source density f is small comparatively to the source domain X.

6.2 Numerical convergence analysis

In Figures 2 and 3, we display the approximation errors obtained when using the recommended scheme in order to solve some Monge-Ampère problems whose solution $u\colon \overline{X} \to \mathbb{R}$ is known exactly.

According to Remark 3.4, the discretization of the Monge-Ampère operator by the operator S_{MA}^{h} defined in (23) is expected to achieve second-order consistency in favorable cases. However, consistency to an order higher than one cannot be expected for the whole scheme (30) due to the fact that the discretization (28) of the optimal transport boundary condition is only first-order consistent. In order to study separately the errors stemming from the discretizations of the Monge-Ampère operator and of the optimal transport boundary condition, we consider both the second and the first boundary value problems for the Monge-Ampère equation. We use the scheme (30) in order to approximate the solution to the second boundary value problem. The first boundary value problem involves a Dirichlet boundary condition; the variant (82) of our numerical scheme devoted to this setting is described in Appendix C.

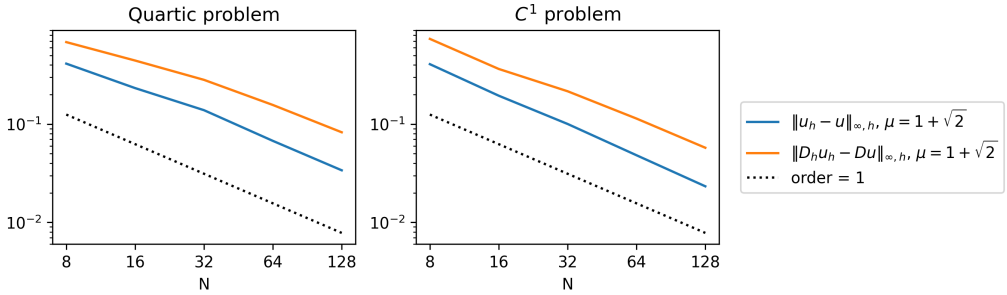

Figure 2: Numerical approximation error with respect to the grid size, with the optimal transport boundary condition.

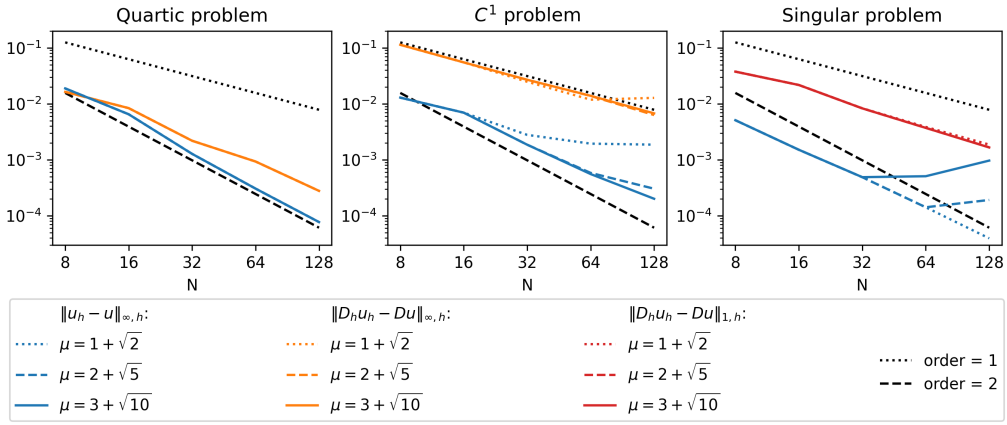

Figure 3: Numerical approximation error with respect to the grid size, with the Dirichlet boundary condition.

We design our test cases by first choosing the domain X and the exact solution u as follows:

$$\begin{aligned}
&\textit{(Quartic problem)} && X := B_2(0,1), && u(x) := \frac{|x|^4}{4},\\
&\textit{(C^1 problem)} && X := B_2(0,1), && u(x) := (0 \vee (|x| - 1/2))^2,\\
&\textit{(Singular problem)} && X := R_{\pi/3}]-1,1[^2, && u(x) := -\sqrt{2 - |x|^2},
\end{aligned}$$

where $R_\theta := \left(\begin{smallmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{smallmatrix}\right)$. The quartic problem is inspired by the numerical experiments in [23], while the C^1 and singular problems are inspired by the ones in [25]. The role of the rotation $R_{\pi/3}$ is to prevent the domain of the singular problem from being axis-aligned, which otherwise provides an unfair advantage to schemes defined on Cartesian grids such as ours.

We only consider Monge-Ampère equations whose coefficients are in the form (61), as in the quadratic optimal transport problem. We choose the target density $g\colon \mathbb{R}^2 \to \mathbb{R}_+^*$ according to (73), and we choose the other parameters of each of the problems considered (the source density $f\colon X \to \mathbb{R}_+$ and the function $\psi\colon \partial X \to \mathbb{R}$ in the case of the Dirichlet boundary condition; the source density $f\colon X \to \mathbb{R}_+$ and the target domain $Y \subset \mathbb{R}^2$ in (62) in the case of the optimal transport boundary condition) appropriately so that the function u is the solution to the problem. For the quartic and C^1 problem, the target domain is $Y = B_2(0,1)$. For the singular problem, we only consider the Dirichlet boundary condition, since with the optimal transport boundary condition the target domain Y would be unbounded and non-convex, which does not fit in our framework.

We define the l^∞ approximation error between the exact solution u and the numerical solution u_h as

$$\|u_h - u\|_{\infty,h} := \max_{x \in \mathcal{G}_h} |u_h[x] - u(x)|.$$

We also display the error

$$\|D_h u_h - Du\|_{\infty,h} := \max_{x \in \tilde{\mathcal{G}}_h} |D_h u_h[x] - Du(x)|$$

between Du and its approximation obtained by applying to u_h the centered finite difference operator D_h defined in (13).

It may be of practical interest to approximate Du by $D_h u_h$ since, at least in the setting of the second boundary value problem, Du coincides with the optimal transport map for the associated optimal transport problem, see section 5.1. Note however that theoretical guarantees on the convergence of $D_h u_h$ towards Du are unreachable using the techniques developed in this paper.

In the case of the singular problem, convergence of $D_h u_h$ towards Du is not observed in the l^∞ norm, which is expected since Du is unbounded in X. For this reason, we display instead the l^1 error

$$\|D_h u_h - Du\|_{1,h} := h^2 \sum_{x \in \tilde{\mathcal{G}}_h} |D_h u_h[x] - Du(x)|.$$

According to Appendix B and Table 1, we choose the set of superbases V_h in (23) as $V_h = V^\mu$ for some $\mu > 1$. In Figure 2, we use the value $\mu = 1 + \sqrt{2}$. According to Proposition B.8 larger values of μ may need to be used for small discretization steps h in order to observe convergence. This is illustrated in Figure 3, where the values $\mu = 2 + \sqrt{5}$ and $\mu = 3 + \sqrt{10}$ are also considered.

6.3 Effect of the set of superbases on the pointwise approximation error

In order to describe more visually the effect of the choice of the parameter μ on the numerical solution to the Monge-Ampère problem, we display in Figure 4 the solution to the scheme (82),

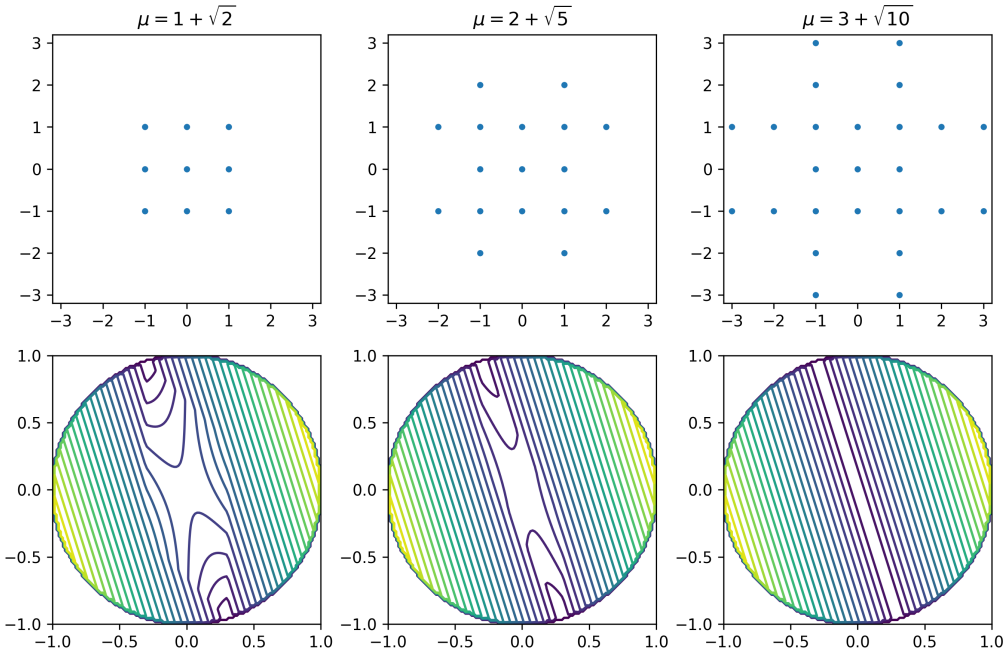

Figure 4: Effect of the choice of the parameter μ on the reconstruction of $u(x) = |\langle e, x\rangle|$, where $e := (1, 1/\sqrt{10})$. Top: finite difference stencil. Bottom: numerical solution.

used with $N = 128$ and $V_h = V^\mu$, for several choices of μ, in order to approximate the solution $u\colon x \mapsto |\langle e, x\rangle|$ to the Dirichlet problem

$$\begin{cases} \det{}_+ D^2 u(x) = 0 & \text{in } B_2(0,1), \\ u(x) = |\langle e, x\rangle| & \text{on } \partial B_2(0,1), \end{cases}$$

with $e := (1, 1/\sqrt{10})$. We observe as expected that the solution is better reconstructed, especially near its singularity, for larger values of μ, which correspond to wider finite difference stencils.

6.4 Comparison between the recommended scheme and the MA-LBR scheme

We study the behavior of the Newton method applied to the resolution of the recommended scheme (82) and the MA-LBR scheme (83), in the setting of the Dirichlet problem

$$\begin{cases} \det{}_+ D^2 u(x) = 1 & \text{in } X, \\ u(x) = 0 & \text{on } \partial X, \end{cases}$$

on domains $X = B_2(0,1) \cup]-1,1[^2$ and $X = B_2(0,1) \setminus [-1,1]^2$. While the second of those domains does not fit in standard theoretical frameworks for the Monge-Ampère equation due to being non-convex, this choice has to be considered as a stress test for the considered numerical methods.

Let us denote by $u_0\colon \mathcal{G}_h \to \mathbb{R}$ our initial guess in the Newton method. The iterates of the Newton method are defined as $u_n := u_{n-1} + 2^{-k_n} d_n$, where d_n is the Newton descent direction

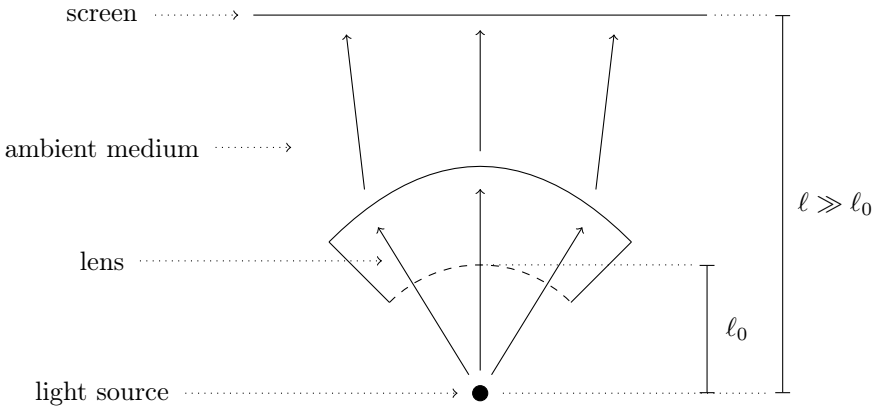

Figure 5: The far field refractor problem. In practice only the shape of the upper interface of the lens has to be approximated numerically. The lower interface, represented by the dashed curve, can be chosen as a portion of a sphere so that it does not refract light rays emanating from the origin.

and $k_n \in \mathbb{N}$ is a damping parameter. In the case of the recommended scheme (82), no damping is required, so we always choose $k_n = 0$. In the case of the MA-LBR scheme (83), one has to use damping so that the constraint (84) remains satisfied along the iterates, as discussed in section 1.3. More precisely, let us introduce the following quantitative variant of the constraint (84):

$$\tilde{\Lambda}_h u_n[x] \geq \tilde{B}_h u_n[x]/2, \quad \forall x \in \mathcal{G}_h \tag{74}$$

(in the setting of our experiments, one has $\tilde{B}_h u[x]/2 = 1/2$). Following [3, 39], and in the spirit of [33], we assume that (74) is satisfied for u_0 and at each iteration we let k_n be the smallest natural number such that (74) holds.

We use the grid size $N = 120$ and the set of superbases $V_h = V^\mu$, $\mu = 2 + \sqrt{5}$. We initialize the Newton method with $u_0[x] := |x|^2 - 2$ (since the simpler initialization $u_0[x] := |x|^2$ does not satisfy (74) close to ∂X, in view of the boundary condition $u = 0$ on ∂X) and we use the stopping criterion

$$\max_{x \in \mathcal{G}_h} |S^h u[x]| < 10^{-8},$$

where either $S^h = S^h_{\text{MABV1}}$ or $S_h = S^h_{\text{MA-LBR-BV1}}$ as appropriate.

On the domain $X = B_2(0,1) \cup]-1,1[^2$, the Newton method for the recommended scheme (82) converges in 9 iterations without damping. The Newton method for the MA-LBR scheme converges in 47 iterations and the largest observed value for k_n is $k_n = 4$, corresponding to a damping step $2^{-k_n} = 0.0625$.

On the domain $X = B_2(0,1) \setminus [-1,1]^2$, the Newton method for the recommended scheme converges in 7 iterations without damping. The Newton method for the MA-LBR scheme converges in 52 iterations and the largest observed value for k_n is $k_n = 5$, corresponding to a damping step $2^{-k_n} = 0.03125$.

6.5 Application of the scheme to the far field refractor problem in nonimaging optics

We apply the scheme (30) to the *far field refractor problem* [30] in nonimaging optics. Various numerical methods for solving this problem, and some of its variants such as the reflector problem,

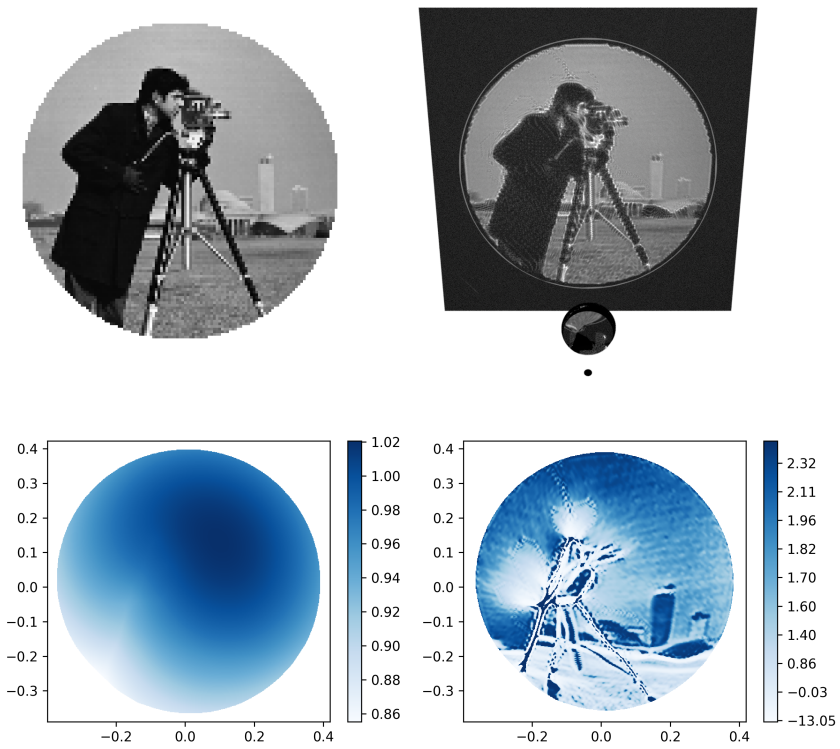

Figure 6: Top left: target image in the far field refractor problem. Bottom left: approximation of the height field v. Bottom right: approximation of the pointwise Gaussian curvature of v. Top right: simulation of the scene (for ℓ large but finite), using the appleseed® rendering engine and the shape computed numerically for the lens.

have been previously studied in the literature [6,11,18,19,28]. We illustrate the refractor problem in Figure 5. Light rays emanate from a point source of light located at the origin, in directions belonging to some subset of the upper hemisphere $\{x \in S^2 \mid x_3 > 0\}$ of the unit sphere S^2. In our experiments we assume that the intensity of those light rays is constant with respect to their direction. The rays propagate in the ambient medium, whose index of refraction we denote by $n_1 > 0$, until they hit a lens, which is located at distance ℓ_0 from the origin and whose index of refraction we denote by $n_2 > 0$. The rays are refracted by the lens, then continue to propagate in the ambient medium until they hit a screen, represented by the plane $\mathbb{R}^2 \times \{\ell\}$, where $\ell > \ell_0$ denotes the distance from the screen to the origin. The aim is to find a suitable shape for the lens that refracts the light rays so that a given target image is reconstructed on the screen.

The far field refractor problem is studied in the limit $\ell \to \infty$. In this limit, it has been shown [30] to be equivalent to an optimal transport problem with a specific, non-quadratic cost, and thus to reduce to the second boundary value problem for the Monge-Ampère equation (1) with coefficients of the form (65) to (67) (as opposed to (61) and (62)). In the above-mentioned optimal transport problem, the source and target densities, which may be exchanged up to an appropriate transformation of the transport map, are the density describing the image that has to be reconstructed in the refractor problem and the one describing the intensity of light rays depending on their initial direction of emission.

Remark 6.3. The problem with finite ℓ is called the *near field refractor problem* and has been shown [31] to reduce to a *generated Jacobian equation* of the form

$$\det{}_+ \left(D^2 u(x) - A(x, u(x), Du(x))\right) = B(x, u(x), Du(x)) \quad \text{in } X. \tag{75}$$

The difference between (75) and (1) is that in (75) the coefficients A and B depend on the values of the function u and not only on its derivatives. The study of the extension of the scheme (30) to generated Jacobian equations is outside the scope of this paper and is an opportunity for future research.

We approximate the solution to the far field refractor problem by solving the scheme (30) on the unit disk $X = B_2(0,1)$, choosing as the source density f the one describing the target image in the problem. We use the grid size $N = 120$ and the set of superbases $V_h = V^\mu$, $\mu = 2 + \sqrt{5}$ (see Table 1). We choose the indices of refraction $n_1 = 1$ and $n_2 = 1.5$, corresponding to a glass-air interface. Eleven iterations of the Newton method are needed in order to solve the scheme. Up to an appropriate postprocessing of the solution to the scheme (30), we obtain an approximation of the height map $v\colon B_2(0,r) \to \mathbb{R}_+$ describing the upper interface $\{(x, v(x)) \mid x \in B_2(0,r)\}$ of the lens, where $r > 0$ denotes the radius of the lens.

We display our numerical results in Figure 6. On the representation of the approximation of the pointwise curvature of v, we observe that the parts of the refractor corresponding to dark areas of the image have a small area, compared to the ones corresponding to bright areas. This is consistent with the fact that the total intensity of the light traversing them should be low, in order for the image to be properly reconstructed. In order to validate our results, we inject the shape that we obtain for the lens in a simulation of the scene that we perform using the appleseed®[2] rendering engine.

7 Conclusion and perspectives

We were able to adapt Perron's method in order to prove the existence of solutions to a class of monotone numerical schemes whose sets of solutions are stable by addition of a constant. We

[2] `https://appleseedhq.net/`

designed a finite difference scheme for the Monge-Ampère equation that belongs to this class, and proved convergence of the scheme in the setting of quadratic optimal transport. We showed that in dimension two, the discretization of the Monge-Ampère operator admits a closed-form formulation, and thus yields a particularly efficient numerical method, when carefully choosing its parameters using Selling's formula. We validated the method by numerical experiments in the context of the far field refractor problem in nonimaging optics.

A natural perspective is the adaptation of the proof of convergence of the scheme to the setting of more general optimal transport problems. The extension of the scheme to generated Jacobian equations such as (75) could also be studied. This would require adapting both the proof of convergence and the one of existence of solutions to the scheme, since the invariance in the set of solutions would not be the same in this case.

Another perspective is the extension of this work to higher dimensions. While our existence and convergence results are valid in any dimension, the closed-form formula that we obtain for the maximum in the discretized Monge-Ampère operator is specific to the dimension two. In higher dimensions, this maximum could be approximated either by sampling the parameter set or by resorting to a numerical optimization procedure, since (23) is an instance of a semidefinite program. We expect the second approach to be more efficient, but developing such an optimization procedure is still an open research direction. The design of this procedure could be made easier by an appropriate choice of the set V_h in (23). In dimension three, one could choose it as a set of superbases of $\mathbb{Z}^3$, benefiting from the fact that Selling's formula (described in Proposition 4.2 in dimension two) also holds in dimension three; in this case, which superbases exactly the set V_h should contain is an open question, since the selection criterion based on the Stern-Brocot tree and presented in Appendix B is two-dimensional only. In dimensions four and higher, one could resort to Voronoi's first reduction of quadratic forms [15], which generalizes Selling's formula to those dimensions.

References

[1] G. Barles and P. E. Souganidis. Convergence of approximation schemes for fully nonlinear second order equations. *Asymptotic Anal.*, 4(3):271–283, 1991.

[2] J.-D. Benamou and Y. Brenier. A computational fluid mechanics solution to the Monge-Kantorovich mass transfer problem. *Numer. Math.*, 84(3):375–393, 2000.

[3] J.-D. Benamou, F. Collino, and J.-M. Mirebeau. Monotone and consistent discretization of the Monge-Ampère operator. *Math. Comp.*, 85(302):2743–2775, 2016.

[4] J.-D. Benamou and V. Duval. Minimal convex extensions and finite difference discretization of the quadratic Monge-Kantorovich problem. *Eur. J. Appl. Math.*, 30(6):1041–1078, 2019.

[5] J.-D. Benamou, B. D. Froese, and A. M. Oberman. Numerical solution of the Optimal Transportation problem using the Monge-Ampère equation. *J. Comput. Phys.*, 260:107–126, 2014.

[6] J.-D. Benamou, W. Ijzerman, and G. Rukhaia. An entropic optimal transport numerical approach to the reflector problem, 2020. HAL preprint hal-02539799.

[7] J. F. Bonnans, G. Bonnet, and J.-M. Mirebeau. A linear finite-difference scheme for approximating Randers distances on Cartesian grids, 2021. HAL preprint hal-03125879.

[8] J. F. Bonnans, G. Bonnet, and J.-M. Mirebeau. Monotone and second order consistent scheme for the two dimensional Pucci equation. In F. J. Vermolen and C. Vuik, editors, *Numerical Mathematics and Advanced Applications ENUMATH 2019*, pages 733–742. Springer, Cham, 2021.

[9] J. F. Bonnans, É. Ottenwaelter, and H. Zidani. A fast algorithm for the two dimensional HJB equation of stochastic control. *ESAIM Math. Model. Numer. Anal.*, 38(4):723–735, 2004.

[10] Y. Brenier. Polar factorization and monotone rearrangement of vector-valued functions. *Comm. Pure Appl. Math.*, 44(4):375–417, 1991.

[11] K. Brix, Y. Hafizogullari, and A. Platen. Solving the Monge-Ampère equations for the inverse reflector problem. *Math. Models Methods Appl. Sci.*, 25(05):803–837, 2015.

[12] L. A. Caffarelli and V. I. Oliker. Weak solution of one inverse problem in geometric optics. *J. Math. Sci.*, 154(1):39–49, 2008.

[13] M. Carter. *Foundations of Mathematical Economics*. MIT Press, Cambridge, MA, 2001.

[14] Y. Chen, J. W. L. Wan, and J. Lin. Monotone mixed finite difference scheme for Monge-Ampère equation. *J. Sci. Comput.*, 76(3):1839–1867, 2018.

[15] J. H. Conway and N. J. A. Sloane. Low-dimensional lattices. III. Perfect forms. *Proc. Roy. Soc. London Ser. A*, 418(1854):43–80, 1988.

[16] M. G. Crandall, H. Ishii, and P.-L. Lions. User's guide to viscosity solutions of second order partial differential equations. *Bull. Amer. Math. Soc.*, 27(1):1–67, 1992.

[17] M. Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In C. J. C. Burges, L. Bottou, M. Welling, and Z. Ghahramani, editors, *NIPS'13: Proceedings of the 26th International Conference on Neural Information Processing Systems — Volume 2*, pages 2292–2300. Curran Associates Inc., Red Hook, NY, 2013.

[18] P. M. M. De Castro, Q. Mérigot, and B. Thibert. Far-field reflector problem and intersection of paraboloids. *Numer. Math.*, 134(2):389–411, 2016.

[19] R. De Leo, C. E. Gutiérrez, and H. Mawi. On the numerical solution of the far field refractor problem. *Nonlinear Anal.*, 157:123–145, 2017.

[20] G. De Philippis and A. Figalli. The Monge-Ampère equation and its link to optimal transportation. *Bull. Amer. Math. Soc.*, 51(4):527–580, 2014.

[21] F. Desquilbet, J. Cao, P. Cupillard, L. Métivier, and J.-M. Mirebeau. Single pass computation of first seismic wave travel time in three dimensional heterogeneous media with general anisotropy. *J. Sci. Comput.*, 89(1):1–37, 2021.

[22] L. C. Evans and R. F. Gariepy. *Measure Theory and Fine Properties of Functions*. Studies in Advanced Mathematics. CRC Press, Boca Raton, FL, 1992.

[23] X. Feng and M. Jensen. Convergent semi-Lagrangian methods for the Monge-Ampère equation on unstructured grids. *SIAM J. Numer. Anal.*, 55(2):691–712, 2017.

[24] A. Figalli and G. Loeper. C^1 regularity of solutions of the Monge-Ampère equation for optimal transport in dimension two. *Calc. Var. Partial Differential Equations*, 35(4):537–550, 2009.

[25] B. D. Froese and A. M. Oberman. Convergent filtered schemes for the Monge-Ampère partial differential equation. *SIAM J. Numer. Anal.*, 51(1):423–444, 2013.

[26] B. Froese Hamfeldt. Convergence framework for the second boundary value problem for the Monge-Ampère equation. *SIAM J. Numer. Anal.*, 57(2):945–971, 2019.

[27] B. Froese Hamfeldt and J. Lesniewski. A convergent finite difference method for computing minimal Lagrangian graphs, 2021. arXiv preprint arXiv:2102.10159.

[28] B. Froese Hamfeldt and A. G. R. Turnquist. Convergent numerical method for the reflector antenna problem via optimal transport on the sphere. *J. Optical Soc. Amer. A*, 38(11):1704–1713, 2021.

[29] C. E. Gutiérrez. *The Monge-Ampère Equation*, volume 89 of *Progress in Nonlinear Differential Equations and Their Applications*. Birkhäuser, Basel, 2016.

[30] C. E. Gutiérrez and Q. Huang. The refractor problem in reshaping light beams. *Arch. Ration. Mech. Anal.*, 193(2):423–443, 2009.

[31] C. E. Gutiérrez and Q. Huang. The near field refractor. *Ann. Inst. H. Poincaré Anal. Non Linéaire*, 31(4):655–684, 2014.

[32] H. Ishii and P.-L. Lions. Viscosity solutions of fully nonlinear second-order elliptic partial differential equations. *J. Differential Equations*, 83(1):26–78, 1990.

[33] J. Kitagawa, Q. Mérigot, and B. Thibert. Convergence of a Newton algorithm for semi-discrete optimal transport. *J. Eur. Math. Soc.*, 21(9):2603–2651, 2019.

[34] S. A. Kochengin and V. I. Oliker. Determination of reflector surfaces from near-field scattering data. *Inverse Problems*, 13(2):363–373, 1997.

[35] N. V. Krylov. *Nonlinear Elliptic and Parabolic Equations of Second Order*, volume 7 of *Mathematics and its Applications*. Springer Netherlands, 1987.

[36] P.-L. Lions. Two remarks on Monge-Ampere equations. *Ann. Mat. Pura Appl.*, 142(1):263–275, 1985.

[37] X.-N. Ma, N. S. Trudinger, and X.-J. Wang. Regularity of potential functions of the optimal transportation problem. *Arch. Ration. Mech. Anal.*, 177(2):151–183, 2005.

[38] J.-M. Mirebeau. Efficient fast marching with Finsler metrics. *Numer. Math.*, 126(3):515–557, 2014.

[39] J.-M. Mirebeau. Discretization of the 3d Monge-Ampère operator, between wide stencils and power diagrams. *ESAIM Math. Model. Numer. Anal.*, 49(5):1511–1523, 2015.

[40] J.-M. Mirebeau. Riemannian fast-marching on cartesian grids, using voronoi's first reduction of quadratic forms. *SIAM J. Numer. Anal.*, 57(6):2608–2655, 2019.

[41] A. Oberman. The convex envelope is the solution of a nonlinear obstacle problem. *Proc. Amer. Math. Soc.*, 135(6):1689–1694, 2007.

[42] A. J. Salgado and W. Zhang. Finite element approximation of the Isaacs equation. *ESAIM Math. Model. Numer. Anal.*, 53(2):351–374, 2019.

[43] E. Selling. Uber die binären und ternären quadratischen Formen. *J. Reine Angew. Math.*, 77:143–229, 1874.

[44] C. Villani. *Topics in Optimal Transportation*, volume 58 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2003.

[45] C. Villani. *Optimal Transport*, volume 338 of *Grundlehren der mathematischen Wissenschaften*. Springer, Berlin, 2009.

A Relation to the MA-LBR scheme

The MA-LBR scheme, introduced in [3], is a discretization of the two-dimensional Monge-Ampère equation. Its natural extension to the generalized equation (1) may be described by a discrete operator $S^h_{\text{MA-LBR}} \colon \mathbb{R}^{\mathcal{G}_h} \to \overline{\mathbb{R}}^{\mathcal{G}_h}$, which is an alternative to the operator S^h_{MA} defined in (23). The operator $S^h_{\text{MA-LBR}}$ is defined as follows:

$$S^h_{\text{MA-LBR}} u[x] := B_h u[x] - \Lambda_h u[x], \tag{76}$$

where

$$\Lambda_h u[x] := \min_{v \in V_h} G(\Delta^v_h u[x] - A^v_h u[x]). \tag{77}$$

We denoted by V_h a given set of superbases of $\mathbb{Z}^2$, by Δ^v_h the second order finite differences defined in (13) and (19), and by A^v_h and B_h the first order finite difference operators defined in (17) to (19). Finally, for any $m \in (\mathbb{R} \cup \{+\infty\})^3$,

$$\begin{aligned} G(m) &:= G_0((0 \vee m_1, 0 \vee m_2, 0 \vee m_3)), \\ G_0(\tilde{m}) &:= \begin{cases} \tilde{m}_2 \tilde{m}_3 & \text{if } \tilde{m}_1 \geq \tilde{m}_2 + \tilde{m}_3, \\ \tilde{m}_1 \tilde{m}_3 & \text{if } \tilde{m}_2 \geq \tilde{m}_1 + \tilde{m}_3, \\ \tilde{m}_1 \tilde{m}_2 & \text{if } \tilde{m}_3 \geq \tilde{m}_1 + \tilde{m}_2, \\ \frac{1}{2}(\tilde{m}_1 \tilde{m}_2 + \tilde{m}_1 \tilde{m}_3 + \tilde{m}_2 \tilde{m}_3) & \\ \quad -\frac{1}{4}(\tilde{m}_1^2 + \tilde{m}_2^2 + \tilde{m}_3^2) & \text{else.} \end{cases} \end{aligned} \tag{78}$$

Contrary to the operator S^h_{MA}, the operator $S^h_{\text{MA-LBR}}$ is only intended to be applied to functions $u \colon \mathcal{G}_h \to \mathbb{R}$ satisfying the constraint

$$\Lambda_h u[x] > 0, \quad \forall x \in \mathcal{G}_h, \tag{79}$$

which is a discrete counterpart to the strict variant of the admissibility constraint (4). If this condition fails, then the Jacobian matrix of the scheme is not invertible [39], which breaks the Newton method relied upon.

Lemma A.1. *Let $u \colon \mathcal{G}_h \to \mathbb{R}$ be a function satisfying (79). Then for any $x \in \mathcal{G}_h$ and $v \in V_h$, letting $m := \Delta^v_h u[x] - A^v_h u[x]$, one has $G(m) = G_0(m) > 0$ and $m > 0$ elementwise.*

Proof. Let $\tilde{m} := (0 \vee m_1, 0 \vee m_2, 0 \vee m_3)$, so that $G(m) = G_0(\tilde{m})$. By (79), one has $G(m) > 0$. Thus $G_0(\tilde{m}) = G(m) > 0$, from which it is easy to deduce that $\tilde{m} > 0$ elementwise. Therefore $m = \tilde{m}$, hence $m > 0$ elementwise and $G(m) = G_0(m)$. □

Recall that, for any superbase $v \in V_h$ and any $\gamma \in \mathbb{R}^3$, one has $\mathcal{D}_v(\gamma) := \sum_{i=1}^3 \gamma_i v_i \otimes v_i$. The following proposition shows that the MA-LBR scheme may be seen as a discretization of the reformulation (6) of the Monge-Ampère equation:

Proposition A.2. *Let $u\colon \mathcal{G}_h \to \mathbb{R}$ be a function satisfying* (79)*. Then for any $x \in \mathcal{G}_h$,*

$$S^h_{\text{MA-LBR}}u[x] = B_h u[x] - \min_{v\in V_h} \inf_{\substack{\gamma\in\mathbb{R}^3_+ \\ \det \mathcal{D}_v(\gamma)=1}} \frac{\langle \gamma, \Delta^v_h u[x] - A^v_h u[x]\rangle^2}{4}.$$

Proof. It suffices to show that for any superbase $v \in V_h$, if $m := \langle \gamma, \Delta^v_h u[x] - A^v_h u[x]\rangle$, then

$$G(m) = \inf_{\substack{\gamma\in\mathbb{R}^3_+ \\ \det \mathcal{D}_v(\gamma)=1}} \frac{\langle \gamma, m\rangle^2}{4}.$$

By Lemma A.1, one has $G(m) = G_0(m)$ and $m > 0$ elementwise. Using that v is a superbase of $\mathbb{Z}^2$, and that therefore $\det(v_i, v_j) = \pm 1$ for any $1 \le i < j \le 3$, one can compute that for any $\gamma \in \mathbb{R}^3_+$,

$$\begin{aligned}
\det \mathcal{D}_v(\gamma) &= \left(\sum_{i=1}^3 \gamma_i v_{i,1}^2\right)\left(\sum_{i=1}^3 \gamma_i v_{i,2}^2\right) - \left(\sum_{i=1}^3 \gamma_i v_{i,1} v_{i,2}\right)^2 \\
&= \sum_{i=1}^3\sum_{j=1}^3 \gamma_i\gamma_j v_{i,1}^2 v_{j,2}^2 - \sum_{i=1}^3\sum_{j=1}^3 \gamma_i\gamma_j v_{i,1} v_{i,2} v_{j,1} v_{j,2} \\
&= \sum_{i=1}^3\sum_{j=1}^3 \gamma_i\gamma_j v_{i,1} v_{j,2} \det(v_i, v_j) \\
&= \sum_{1\le i<j\le 3} \gamma_i\gamma_j v_{i,1} v_{j,2} \det(v_i, v_j) + \sum_{1\le i<j\le 3} \gamma_j\gamma_i v_{j,1} v_{i,2} \det(v_j, v_i) \\
&= \sum_{1\le i<j\le 3} \gamma_i\gamma_j \det(v_i, v_j)^2 \\
&= \sum_{1\le i<j\le 3} \gamma_i\gamma_j.
\end{aligned}$$

Thus it remains to prove that

$$G_0(m) = \inf_{\substack{\gamma\in\mathbb{R}^3_+ \\ \gamma_1\gamma_2+\gamma_1\gamma_3+\gamma_2\gamma_3=1}} \frac{\langle \gamma, m\rangle^2}{4},$$

or equivalently that

$$\inf_{\substack{\gamma\in\mathbb{R}^3_+ \\ \gamma_1\gamma_2+\gamma_1\gamma_3+\gamma_2\gamma_3=1}} \langle \gamma, m\rangle = 2\sqrt{G_0(m)}. \tag{80}$$

The infimum in (80) is attained at some $\gamma \in \mathbb{R}^3_+$, fixed in the following, since the objective function $\gamma \mapsto \langle \gamma, m\rangle$ is coercive in $\mathbb{R}^3_+$ and since the constraint $\gamma_1\gamma_2 + \gamma_1\gamma_3 + \gamma_2\gamma_3 = 1$ is closed.

If $\gamma_3 = 0$, then by elimination of $\gamma_2 = 1/\gamma_1$, the optimization problem becomes $\min\{m_1\gamma_1 + m_2/\gamma_1 \mid \gamma_1 \ge 0\}$, whose solution is $2\sqrt{m_1 m_2}$ attained for $\gamma_1 = \sqrt{m_2/m_1}$, consistently with (78) and (80). Likewise if $\gamma_1 = 0$ or $\gamma_2 = 0$ instead of $\gamma_3 = 0$.

Consider now the case where $\gamma > 0$ elementwise. Then by the KKT conditions for the optimization problem in the left-hand side of (80), there exists a Lagrange multiplier $\lambda \ge 0$ such that $m = (\lambda/2)(\gamma_2 + \gamma_3, \gamma_1 + \gamma_3, \gamma_1 + \gamma_2)$. Equivalently,

$$\lambda\gamma = (m_2 + m_3 - m_1, m_1 + m_3 - m_2, m_1 + m_2 - m_3). \tag{81}$$

In particular, we obtain that the elementwise positiveness of γ cannot hold if $m_1 \geq m_2 + m_3$, $m_2 \geq m_1 + m_3$, or $m_3 \geq m_1 + m_2$, as announced in the expression (78) of G_0. Replacing the elements of γ with their expressions in terms of λ and of the elements of m, see (81), and performing straightforward simplifications, we obtain new expressions of the objective and the constraint of the optimization problem in (80):

$$\langle \gamma, m \rangle = \Delta/\lambda, \qquad\qquad 1 = \gamma_1\gamma_2 + \gamma_1\gamma_3 + \gamma_2\gamma_3 = \Delta/\lambda^2,$$

where $\Delta := 2(m_1m_2 + m_1m_3 + m_2m_3) - (m_1^2 + m_2^2 + m_3^2)$. The constraint yields $\lambda = \sqrt{\Delta}$, and the objective value is thus $\Delta/\lambda = \sqrt{\Delta} = 2\sqrt{G_0(m)}$ as announced. □

B Choosing the set of superbases in dimension two

In this appendix, we explain how one may choose, in dimension $d = 2$ and for any $h > 0$, a finite set V_h of superbases of $\mathbb{Z}^2$ satisfying (20) to (22). The motivation is to use this set V_h in (23). The construction of V_h is based on the Stern-Brocot tree of bases of $\mathbb{Z}^2$ (see [9] for a similar approach in the setting of Hamilton-Jacobi-Bellman equations):

Definition B.1. A pair (u, v) of vectors of $\mathbb{Z}^2$ is a *direct basis* of $\mathbb{Z}^2$ if $\det(u, v) = 1$.

Definition B.2. The *Stern-Brocot tree* $\mathcal{T}$ is the collection of direct bases of $\mathbb{Z}^2$ defined inductively as follows: (i) the canonical basis belongs to $\mathcal{T}$, and (ii) for any $(u, v) \in \mathcal{T}$, one has $(u, u + v) \in \mathcal{T}$ and $(u + v, v) \in \mathcal{T}$.

Remark B.3. In classical descriptions of the Stern-Brocot tree, the vector $u = (p, q)$ is often identified with the ratio p/q, which is a nonnegative rational, or with $+\infty$, and likewise for $v = (r, s)$ (note that p and q are nonnegative and coprime by construction).

For any $(u, v) \in \mathcal{T}$, the scalar product $\langle u, v \rangle$ is a nonnegative integer, as follows from an immediate induction. The set $\mathcal{T}_s := \{(u, v) \in \mathcal{T}; \langle u, v \rangle < s\}$ is a finite subtree which can be generated by exploration with the obvious stopping criterion, since $\min\{\langle u, u + v \rangle, \langle u + v, v \rangle\} = \langle u, v \rangle + \min\{|u^2|, |v|^2\} \geq \langle u, v \rangle + 1$.

Lemma B.4. *Let $\mu > 1$ and $(u, v) \in \mathcal{T}_{(\mu-\mu^{-1})/2}$. Then*

$$\max\{|u|, |v|\} < \frac{\mu + \mu^{-1}}{2} < \mu.$$

Proof. It holds that

$$|u|^2 \leq |u|^2|v|^2 = \det(u, v)^2 + \langle u, v \rangle^2 < 1 + \left(\frac{\mu - \mu^{-1}}{2}\right)^2 = \frac{\mu^2 + \mu^{-2} + 2}{4} = \left(\frac{\mu + \mu^{-1}}{2}\right)^2,$$

and similarly for v. □

For any $\mathcal{D} \in \mathcal{S}_2^{++}$, we define

$$\mu(\mathcal{D}) := \sqrt{|\mathcal{D}||\mathcal{D}^{-1}|}, \qquad\qquad s(\mathcal{D}) := \frac{1}{2}(\mu(\mathcal{D}) - \mu(\mathcal{D})^{-1}).$$

Note that $\mu(\mathcal{D})$ is the square root of the condition number of $\mathcal{D}$.

Lemma B.5. *Let $(u, v) \in \mathcal{T}$ and $\mathcal{D} \in \mathcal{S}_2^{++}$. If $\langle u, v \rangle \geq s(\mathcal{D})$, then $\langle u, \mathcal{D}v \rangle \geq 0$.*

Proof. Denote by $\sphericalangle(u,v) \in [0,\pi]$ the unoriented angle between two vectors, defined by

$$\cos \sphericalangle(u,v) := \frac{\langle u,v\rangle}{|u||v|}.$$

On the one hand one has

$$\sin \sphericalangle(u,v) = \frac{\det(u,v)}{\sqrt{\langle u,v\rangle^2 + \det(u,v)^2}} = (1+\langle u,v\rangle^2)^{-1/2}.$$

On the other hand on can show [21, Corollary B.4] that for any vector v,

$$(\mu(\mathcal{D}) + \mu(\mathcal{D})^{-1}) \cos \sphericalangle(v, \mathcal{D}v) \geq 2.$$

If $\langle u,v\rangle \geq (\mu(\mathcal{D}) - \mu(\mathcal{D})^{-1})/2$, then one obtains $\sin \sphericalangle(u,v) \leq \cos \sphericalangle(v,\mathcal{D}v)$, and therefore $\sphericalangle(u,v) + \sphericalangle(v,\mathcal{D}v) \leq \pi/2$. By subadditivity of angles, $\sphericalangle(u,\mathcal{D}v) \leq \pi/2$, which is the announced result. □

Definition B.6. Let $\mathcal{D} \in \mathcal{S}_2^+$. A superbase $v = (v_1,v_2,v_3)$ of $\mathbb{Z}^2$ is *$\mathcal{D}$-obtuse* if $\langle v_i, \mathcal{D}v_j\rangle \leq 0$, for any $1 \leq i < j \leq 3$.

Corollary B.7. *For any $\mathcal{D} \in \mathcal{S}_2^{++}$, there exists $(u,v) \in \mathcal{T}$ such that $\langle u,v\rangle \leq s(\mathcal{D})$ and, denoting $\tilde{u} := (u_1,-u_2)$ and $\tilde{v} := (v_1,-v_2)$, either $(u,v,-u-v)$ or $(\tilde{u},\tilde{v},-\tilde{u}-\tilde{v})$ is a $\mathcal{D}$-obtuse superbase.*

Proof. We can assume that the nondiagonal coefficient of $\mathcal{D}$ is negative, up to reversing the orientation of one axis, and removing the trivial case of diagonal matrices. Let $(u,v) \in \mathcal{T}$ be such that $\langle u,\mathcal{D}v\rangle < 0$ and $\langle u,v\rangle$ is maximal. Such an element exists since the canonical basis obeys the condition $\langle u,\mathcal{D}v\rangle < 0$, since $\langle u,v\rangle$ is a nonnegative integer, and since $\langle u,\mathcal{D}v\rangle \geq 0$ when $\langle u,v\rangle \geq s(\mathcal{D})$, by Lemma B.5. Then, by construction, $\langle u,\mathcal{D}(u+v)\rangle \geq 0$ and $\langle u+v,\mathcal{D}v\rangle \geq 0$, which shows that $(u,v,-u-v)$ is a $\mathcal{D}$-obtuse superbase. □

For any $\mu > 1$, we define

$$V^\mu := \bigcup_{(u,v)\in\mathcal{T}_{(\mu-\mu^{-1})/2}} \{(-u^\perp,-v^\perp,u^\perp+v^\perp),(-\tilde{u}^\perp,-\tilde{v}^\perp,\tilde{u}^\perp+\tilde{v}^\perp)\},$$

where $\tilde{u} := (u_1,-u_2)$ and $\tilde{v} := (v_1,-v_2)$. The construction of the set V^μ is motivated by the following observation: if $\mathcal{D} \in \mathcal{S}_d^{++}$ obeys $\mu(\mathcal{D}) < \mu$, then, using Corollary B.7 and that $s(\mathcal{D}) < (\mu-\mu^{-1})/2$, there exists a superbase $v = (v_1,v_2,v_3) \in V^\mu$ such that $(v_1^\perp, v_2^\perp, v_3^\perp)$ is $\mathcal{D}$-obtuse.

Note that by construction, there exist countably many values $1 = \mu_0 < \mu_1 < \mu_2 < \cdots$ such that the map $\mu \mapsto V^\mu$, defined on $(1,+\infty)$, is constant on each interval $(\mu_n,\mu_{n+1}]$, $n \in \mathbb{N}$, and satisfies $V^{\mu_n} \subsetneq V^{\mu_{n+1}}$, for any $n \in \mathbb{N}^*$. We display in Table 1 the values of μ_n for small $n \in \mathbb{N}^*$, as well as some properties of the associated sets of superbases V^{μ_n}.

One may choose a sequence $(\mu_h)_{h>0}$ of parameters $\mu_h > 1$, and let $V_h = V^{\mu_h}$.

Proposition B.8. *For any $h > 0$, let $\mu_h > 1$ be such that*

$$\lim_{h\to 0} \mu_h = +\infty, \qquad\qquad \lim_{h\to 0} h\mu_h = 0,$$

and let $V_h = V^{\mu_h}$. Then (20) to (22) are satisfied.

n	μ_n	μ_n^2	$\#(V^{\mu_n})$	$\max_{v\in V^{\mu_n}} \max_{e\in v} \lvert e\rvert$	additional superbases (up to transformations)
1	$1+\sqrt{2}\approx 2.414$	≈ 5.828	2	$\sqrt{2}\approx 1.414$	$\left(\binom{0}{-1}, \binom{-1}{0}, \binom{1}{1}\right)$
2	$2+\sqrt{5}\approx 4.236$	≈ 17.944	6	$\sqrt{5}\approx 2.236$	$\left(\binom{0}{-1}, \binom{-1}{-1}, \binom{1}{2}\right)$
3	$3+\sqrt{10}\approx 6.162$	≈ 37.974	10	$\sqrt{10}\approx 3.162$	$\left(\binom{0}{-1}, \binom{-1}{-2}, \binom{1}{3}\right)$
4	$4+\sqrt{17}\approx 8.123$	≈ 65.985	18	$\sqrt{17}\approx 4.123$	$\left(\binom{0}{-1}, \binom{-1}{-3}, \binom{1}{4}\right)$, $\left(\binom{-1}{-2}, \binom{-1}{-1}, \binom{2}{3}\right)$
5	$5+\sqrt{26}\approx 10.099$	≈ 101.99	22	$\sqrt{26}\approx 5.099$	$\left(\binom{0}{-1}, \binom{-1}{-4}, \binom{1}{5}\right)$

Table 1: Properties of V^μ for $\mu \le 5+\sqrt{26}$. In the rightmost column, we display for each n the elements of a set of superbases $\hat{V}_n$ such that $V^{\mu_n} = \bigcup_{1\le i\le n} \bigcup_{v\in\hat{V}_i} \{(v_1,v_2,v_3),\ (v_2^\perp, v_1^\perp, v_3^\perp),\ (\tilde{v}_2^\perp, \tilde{v}_1^\perp, \tilde{v}_3^\perp),\ (-\tilde{v}_1, -\tilde{v}_2, -\tilde{v}_3)\}$, where $\tilde{e} := (e_1, -e_2)$ denotes the reflection with respect to the horizontal axis, as in Corollary B.7. For the first three values of n, the points of the finite difference stencils $\bigcup_{v\in V^{\mu_n}} \bigcup_{e\in v} \{\pm e\}$ are displayed in Figure 4.

Proof. For fixed $h>0$, let $\mathcal{D}\in\mathcal{S}_2^{++}$ be such that $\mu(\mathcal{D}) < \mu_h$. Then there exists a superbase $v = (v_1,v_2,v_3) \in V_h = V^{\mu_h}$ such that $(v_1^\perp, v_2^\perp, v_3^\perp)$ is $\mathcal{D}$-obtuse. By Selling's formula Proposition 4.2, there exists $\gamma\in\mathbb{R}_+^3$ such that $\mathcal{D} = \mathcal{D}_v(\gamma)$ (choose $\gamma = \gamma_v(\mathcal{D})$). It follows that

$$\{\mathcal{D}\in\mathcal{S}_2^{++} \mid \mathrm{Tr}(\mathcal{D})=1,\ \mu(\mathcal{D})<\mu_h\} \subset \{\mathcal{D}_v(\gamma) \mid v\in V_h,\ \gamma\in\mathbb{R}_+^3,\ \mathrm{Tr}(\mathcal{D}_v(\gamma))=1\}.$$

Therefore

$$\begin{aligned}
&\lim_{h\to 0} d_H(\{\mathcal{D}_v(\gamma) \mid v\in V_h,\ \gamma\in\mathbb{R}_+^3,\ \mathrm{Tr}(\mathcal{D}_v(\gamma))=1\}, \{\mathcal{D}\in\mathcal{S}_2^+ \mid \mathrm{Tr}(\mathcal{D})=1\})\\
&\le \lim_{h\to 0} d_H(\{\mathcal{D}\in\mathcal{S}_2^{++} \mid \mathrm{Tr}(\mathcal{D})=1,\ \mu(\mathcal{D})\le\mu_h\}, \{\mathcal{D}\in\mathcal{S}_2^+ \mid \mathrm{Tr}(\mathcal{D})=1\})\\
&= 0,
\end{aligned}$$

which proves (20).

Let $v=(v_1,v_2,v_3)$ be a superbase belonging to V_h. By Lemma B.4, $\max_{1\le i\le 3}|v_i| \le 2\mu_h$, and (21) follows.

Finally, (22) is satisfied since the subtree $\mathcal{T}_{(\mu_h-\mu_h^{-1})/2}$ always contains the canonical basis (e_1,e_2), hence $(-e_2, e_1, e_2-e_1) = (-e_1^\perp, -e_2^\perp, e_1^\perp + e_2^\perp) \in V_h$. □

Remark B.9. Let $c>0$, $r\in(0,1)$, and, for sufficiently small $h>0$, choose $V_h = V^{\mu_h}$ where $\mu_h := ch^{-r}$. Then the proof of Proposition B.8 yields the following refinements of (20) and (21):

$$\begin{aligned}
&d_H\left(\{\mathcal{D}_v(\gamma) \mid v\in V_h,\ \gamma\in\mathbb{R}_+^3,\ \mathrm{Tr}(\mathcal{D}_v(\gamma))=1\}, \{\mathcal{D}\in\mathcal{S}_2^+ \mid \mathrm{Tr}(\mathcal{D})=1\}\right) = O(h^{2r}),\\
&\max_{v\in V_h}\max_{e\in v}|e| = O(h^{-r}),
\end{aligned}$$

where the exponent in the first formula may be obtained by rewriting the relevant part of (55) as $1-|\rho| = 2/(\mathrm{Cond}(\mathfrak{D}(\rho))-1) = 2/(\mu(\mathfrak{D}(\rho))^2-1) = O(\mu(\mathfrak{D}(\rho)))^{-2}$.

Let us give the following upper bound on the cardinal of the set V^μ:

Proposition B.10. *There exists $C>0$ such that for any $\mu>1$, one has $\#(V^\mu) \le C\mu(1+\log\mu)$.*

Proof. By [38, Lemma 2.7], there exists $C>0$ such that for any $s>1$, one has $\#(\mathcal{T}_s) \le Cs(1+\log s)$. The stated result follows, since $\#(V^\mu) = 2\#(\mathcal{T}_{(\mu-\mu^{-1})/2})$ and $\mathcal{T}_{(\mu-\mu^{-1})/2} \subset \mathcal{T}_\mu$. □

C Scheme for the Dirichlet problem

In some numerical experiments of sections 6.2 to 6.4, we consider the Monge-Ampère problem equipped with the Dirichlet boundary condition (68), instead of the optimal transport boundary condition. Let us describe how we adapt the scheme (30), or at least the discretization (23) of the Monge-Ampère operator, to this setting.

The function $\psi\colon \partial X \to \mathbb{R}$ defining the Dirichlet boundary condition is assumed to be given. For any $x \in \mathcal{G}_h$ and $e \in \mathbb{Z}^d \setminus \{0\}$, we define

$$h^e(x) := \min\{h' > 0 \mid x + h'e \in \mathcal{G}_h \cap \partial X\}.$$

Similarly to (11), we define the translation operator $\tilde{T}_h^e\colon \mathbb{R}^{\mathcal{G}_h} \to \mathbb{R}^{\mathcal{G}_h}$, applied to a function $u\colon \mathcal{G}_h \to \mathbb{R}$, by

$$\tilde{T}_h^e u[x] := \begin{cases} u[x + he] & \text{if } x + he \in \mathcal{G}_h \text{ and } h^e(x) = h, \\ \psi(x + h^e(x)) & \text{else.} \end{cases}$$

We then define the first- and second-order finite difference operators

$$\tilde{\delta}_h^e u[x] := \frac{\tilde{T}_h^e u[x] - u[x]}{h^e(x)}, \qquad \tilde{\Delta}_h^e u[x] := \frac{2}{h^e(x) + h^{-e}(x)}(\tilde{\delta}_h^e u[x] + \tilde{\delta}_h^{-e} u[x]),$$

as well as the approximations of the Laplacian and of the gradient

$$\tilde{\Delta}_h u[x] := \sum_{i=1}^d \tilde{\Delta}_h^{e_i} u[x], \qquad \tilde{D}_h u[x] := \left(\frac{\tilde{\delta}_h^{e_i} u[x] - \tilde{\delta}_h^{-e_i} u[x]}{2} \right)_{1 \le i \le d}.$$

Note that, under the assumption (10), the operators $\tilde{T}_h^e$, $\tilde{\delta}_h^e$, $\tilde{\Delta}_h^e$, $\tilde{\Delta}_h$, and $\tilde{D}_h$ reduce respectively to the operators T_h^e, δ_h^e, Δ_h^e, Δ_h, and D_h defined in (11) to (13) at all points x that are far enough from ∂X. We previously used the same construction for finite difference operators near the boundary of the considered domain in [7, 8].

Similarly to (17), (18), and (19), we define

$$\tilde{A}_h^e u[x] := a_{\min}|e|^2 \vee \left(\langle e, A(x, \tilde{D}_h u[x])e \rangle - \frac{h}{2} a_{\mathrm{LF}} |e|^2 \tilde{\Delta}_h u[x] \right),$$

$$\tilde{B}_h u[x] := 0 \vee \left(B(x, \tilde{D}_h u[x])^{1/d} - \frac{h}{2} b_{\mathrm{LF}} \tilde{\Delta}_h u[x] \right)^d,$$

and, for any family v of vectors of $\mathbb{Z}^d \setminus \{0\}$,

$$\tilde{\Delta}_h^v u[x] := (\tilde{\Delta}_h^e u[x])_{e \in v}, \qquad \tilde{A}_h^v u[x] := (\tilde{A}_h^e u[x])_{e \in v}.$$

Then the scheme that we use for the Dirichlet problem may be written as

$$S_{\mathrm{MABV1}}^h u[x] = 0 \quad \text{in } \mathcal{G}_h, \tag{82}$$

where

$$S_{\mathrm{MABV1}}^h u[x] := \max_{v \in V_h} \max_{\substack{\gamma \in \mathbb{R}_+^{d(d+1)/2} \\ \mathrm{Tr}(\mathcal{D}_v(\gamma)) = 1}} L_{v,\gamma}(\tilde{B}_h u[x], \tilde{\Delta}_h^v u[x] - \tilde{A}_h^v u[x]).$$

The complete study of the theoretical properties of this scheme, such as monotonicity and convergence, is outside the scope of this paper.

In section 6.4, we also apply the MA-LBR scheme, see Appendix A, to the Dirichlet problem. The scheme The MA-LBR scheme in this setting may be written as

$$S^h_{\text{MA-LBR-BV1}} u[x] = 0 \quad \text{in } \mathcal{G}_h, \tag{83}$$

where we defined, similarly to (76) and (77),

$$S^h_{\text{MA-LBR-BV1}} u[x] := \tilde{B}_h u[x] - \tilde{\Lambda}_h u[x], \qquad \tilde{\Lambda}_h u[x] := \min_{v \in V_h} G(\tilde{\Delta}^v_h u[x] - \tilde{A}^v_h u[x]).$$

The operator $S^h_{\text{MA-LBR-BV1}}$ is intended to be applied to functions $u \colon \mathcal{G}_h \to \mathbb{R}$ satisfying the admissibility constraint

$$\tilde{\Lambda}_h u[x] > 0, \quad \forall x \in \mathcal{G}_h, \tag{84}$$

which is the natural counterpart to (79) in the Dirichlet setting.