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†

June 9, 2021

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 quasilinear operators. In dimension two, we show how using Selling's formula, a tool originating from low-dimensional lattice geometry, in order to choose the parameters of the discretization yields a closed-form formula for the maximum that appears at the discrete level, allowing the scheme to be solved particularly efficiently. We present the numerical results that we obtained when applying the scheme to the far field refractor problem in nonimaging optics.

1 Introduction

The problem of *optimal transport* [35] is strongly related to the *Monge-Ampère equation* [22]: under suitable assumptions, the potential function which solves an optimal transport problem is also solution to the Monge-Ampère equation associated to this problem, equipped with the relevant boundary condition [14]. Some problems in nonimaging optics are also described by Monge-Ampère equations, among which some fit in the framework of optimal transport [8, 22] and some do not [23, 26].

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 [12]. 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 [25], and applied to problems in nonimaging optics [13]. Finally, one may solve numerically the Monge-Ampère equation associated to the considered optimal transport problem, as suggested in this paper and previously in [4, 19]. Benefits of this last approach include that it is easily adapted to various optimal transport problems by simply changing the coefficients of the approximated Monge-Ampère equation, and that one may use the theory of numerical

*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

schemes for degenerate elliptic partial differential equations [1] in order to establish convergence results.

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, 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 (21). Therefore, in order to prove Theorem 5.21, 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 [27, 28] to reformulate the Monge-Ampère equation in the form (2), where F_{MA} is a supremum of quasilinear 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

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

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 [17],

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

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 $\mathbb{R}^d$, we will assume that

$$\lim_{h\to 0} d_H\left(\partial X \cup ((X \cap h\mathbb{Z}^d) \setminus \mathcal{G}_h), \partial X\right) = 0, \tag{9}$$

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 (38) 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

$$T_h^e u[x] := \begin{cases} u[x + he] & \text{if } x + he \in \mathcal{G}_h,\\ +\infty & \text{else,}\end{cases} \tag{10}$$

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

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

$$\Delta_h u[x] := \sum_{i=1}^{d} \Delta_h^{e_i} u[x], \qquad D_h u[x] := \left(\frac{\delta_h^{e_i} u[x] - \delta_h^{-e_i} u[x]}{2} \right)_{1 \le i \le d}. \tag{11}$$

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, $p' \in \mathbb{R}^d$,

$$A(x,p) \succeq a_{\min} I_d, \tag{12}$$

$$|A(x,p) - A(x,p')|_2 \le a_{\mathrm{LF}} |p - p'|_1, \tag{13}$$

$$|B(x,p)^{1/d} - B(x,p')^{1/d}| \le b_{\mathrm{LF}} |p - p'|_1. \tag{14}$$

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

$$A_h^e u[x] := \begin{cases} a_{\min}|e|^2 \vee (\langle e, A(x, D_h u[x]) e \rangle - h a_{\mathrm{LF}} |e|^2 \Delta_h u[x]) & \text{if } \Delta_h u[x] < +\infty, \\ a_{\min}|e|^2 & \text{else,} \end{cases} \tag{15}$$

$$B_h u[x] := \begin{cases} 0 \vee (B(x, D_h u[x])^{1/d} - h b_{\mathrm{LF}} \Delta_h u[x])^d & \text{if } \Delta_h u[x] < +\infty, \\ 0 & \text{else.} \end{cases} \tag{16}$$

(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

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

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

$$\lim_{h \to 0} d_H \left(\{ \mathcal{D}_v(\gamma) \mid v \in V_h,\, \gamma \in \mathbb{R}_+^{d(d+1)/2},\, \mathrm{Tr}(\mathcal{D}_v(\gamma)) = 1 \}, \{ \mathcal{D} \in \mathcal{S}_d^+ \mid \mathrm{Tr}(\mathcal{D}) = 1 \} \right) = 0. \tag{17}$$

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

$$\lim_{h \to 0} h \max_{v \in V_h} \max_{e \in v} |e| = 0, \tag{18}$$

and that for any $h > 0$,

$$e_1 \in \bigcup_{v \in V_h} \bigcup_{e \in v} \{ \pm e \}, \tag{19}$$

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 $S^h_{\mathrm{MA}} : \mathbb{R}^{\mathcal{G}_h} \to \overline{\mathbb{R}}^{\mathcal{G}_h}$ defined by

$$S^h_{\mathrm{MA}} u[x] := \max_{v \in V_h} \max_{\substack{\gamma \in \mathbb{R}_+^{d(d+1)/2} \\ \mathrm{Tr}(\mathcal{D}_v(\gamma))=1}} L_{v,\gamma}(B_h u[x], \Delta^v_h u[x] - A^v_h u[x]), \tag{20}$$

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 $m \in (\mathbb{R} \cup \{+\infty\})^I$,

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

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. We prove in section 4 that when choosing V_h in this way, the second maximum in (20) 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 (10), 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 $b \ge 0$ and $m \in \mathbb{R}^3$,

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

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 a constraint of the form

$$Du(x) \in \overline{P(x)}, \quad \forall x \in X, \tag{21}$$

where for any $x \in \overline{X}$, $P(x)$ is an open bounded convex nonempty subset of $\mathbb{R}^d$, such that $\overline{P(x)}$ depends continuously on x for the Hausdorff distance over compact subsets of $\mathbb{R}^d$. 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 x. The constraint (21) is called a boundary condition, since under some assumptions it suffices that it is satisfied on the boundary of X.

For now, let us consider the class of numerical schemes for equations (1) and (21) 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

$$S^h_{\mathrm{MABV2}} u[x] = 0 \quad \text{in } \mathcal{G}_h. \tag{22}$$

One property of equations (1) and (21) 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 (22) 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 [19] to build an operator S^h_{MABV2} suitable for (22). 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 (21).

We use the following formulation of (21):

$$F_{\mathrm{BV2}}(x, Du(x)) \leq 0 \quad \text{in } X, \tag{23}$$

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

$$F_{\mathrm{BV2}}(x, p) := \max_{|e|=1} (\langle e, p \rangle - \sigma_{P(x)}(p)). \tag{24}$$

(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 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

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

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

$$\begin{aligned} S^h_{\mathrm{BV2}} u[x] &:= \max_{|e|=1} (D^e_h u[x] - \sigma_{P(x)}(e)), \\ S^h_{\mathrm{MABV2}} u[x] &:= S^h_{\mathrm{MA}} u[x] \vee S^h_{\mathrm{BV2}} u[x]. \end{aligned} \tag{25}$$

In this setting, the scheme (22) is additively invariant.

Additively invariant schemes of the form (22) 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 (52) 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 (22), following the approach used in the numerical experiments in [4]. 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

$$S^{h,\alpha}_{\mathrm{MA}} u[x] := S^h_{\mathrm{MA}} u[x] + \alpha, \qquad S^{h,\alpha}_{\mathrm{MABV2}} u[x] := S^{h,\alpha}_{\mathrm{MA}} u[x] \vee S^h_{\mathrm{BV2}} u[x]. \tag{26}$$

The scheme we actually solve is

$$S^{h,\alpha}_{\mathrm{MABV2}} u[x] = 0 \quad \text{in } \mathcal{G}_h. \tag{27}$$

1.3 Main contributions and relation to previous works

We introduce the numerical scheme (27) for the Monge-Ampère equation (1), equipped with the boundary condition (21). 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 (27), see section 2, and we show, in section 3, that the scheme (27) 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 apply the scheme to the far field refractor problem in nonimaging optics, see section 6.

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 [5] 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, and thus it is more stable numerically than (6). When solving schemes based on (6) using the damped Newton method, extremely small steps are typically required to ensure that the constraint (4) remains satisfied along the iterations; this is not the case with (7). Numerical schemes based on (7) were previously introduced in [17], and then in [9], 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 [19] 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 (21).

In [4], convergence of a scheme of the form (22) 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 (27), but the proof of convergence was not extended to this setting.

In [19], convergence of another scheme of the form (22) was proved. A Dirichlet boundary condition was enforced on ∂X, which in our setting would translate to replacing $+\infty$ by some constant $C \in \mathbb{R}$ in (10). Therefore the scheme considered in that paper is not additively invariant and does admit solutions. The Dirichlet boundary condition is to be understood in a weak sense (the one of viscosity solutions, see Definition 2.3), and may formally be simplified to $u(x) \leq C$ on ∂X, with equality at some point x_* of the boundary, provided that the scheme satisfied a property of *underestimation*, which is an assumption of the proof of convergence. This property is satisfied in the case of quadratic optimal transport at the cost of a careful handling of the constraint (21), 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 [19]. 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 $x_* \in \partial X$ were 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 (27), we use the arguments introduced in [19] when appropriate. However, the property of underestimation is not required in our setting.

Note that the scheme (27), and its continuous counterpart (39) below, which both feature an additional unknown or parameter $\alpha \in \mathbb{R}$, fit in the framework of eigenvalue problems recently studied in [20]. Although our proof of convergence only applies to Monge-Ampère equation in the setting of quadratic optimal transport, our existence result, Theorem 2.14, 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

$$F(x, Du(x), D^2u(x)) = 0 \quad \text{in } \overline{X}. \tag{28}$$

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 (28). The equation without the boundary condition would then be

$$F(x, Du(x), D^2u(x)) = 0 \quad \text{in } X. \tag{29}$$

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 (28) and (29), 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 (28) and (29) 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 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

$$F_{\text{ex}}(x, u'(x), u''(x)) = 0 \quad \text{in } [-1,1], \tag{30}$$

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

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

The equation 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 (52) 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 (28) 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 (29).

Note that if a viscosity subsolution (respectively supersolution) u to (28) 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 (28) at 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

$$\lim_{h \to 0} d_H(\mathcal{G}_h, X) = 0. \tag{31}$$

Note that *in the case* that $\mathcal{G}_h$ is included in $X \cap h\mathbb{Z}^d$, then (31) is implied by (9).

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

$$S^h u[x] = 0 \quad \text{in } \mathcal{G}_h. \tag{32}$$

Definition 2.5. The scheme (32) 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 (28) if (31) 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 (32) 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 $\overline{u}[x] \leq \underline{u}[x]$ 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 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 (32) 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 (32) is monotone and consistent with equation (28), then functions $\overline{u}$, $\underline{u}\colon \overline{X} \to \mathbb{R}$ defined by*

$$\overline{u}(x) := \limsup_{\substack{n\in\mathbb{N},\, n\to+\infty \\ x' \in \mathcal{G}_{h_n},\, x' \to x}} u_n[x'], \qquad \underline{u}(x) := \liminf_{\substack{n\in\mathbb{N},\, n\to+\infty \\ x' \in \mathcal{G}_{h_n},\, x' \to x}} u_n[x'], \tag{33}$$

are respectively a viscosity subsolution and supersolution to (28).

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 (28) may also depend on $u(x)$, a *strong comparison principle*, that is, a result stating that viscosity subsolutions to (28) 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 (32) 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 (30). However, solving this scheme is equivalent to solving a square, noninvertible linear system.

To get around this difficulty, we add a parameter $\alpha \in \mathbb{R}$ to the equation (28), yielding a new equation

$$F^\alpha(x, Du(x), D^2 u(x)) = 0 \quad \text{in } \overline{X}, \tag{34}$$

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 (34) 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 (30) 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

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

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

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

(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

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

may easily be constructed explicitly.

The definition of solutions $(\alpha, u) \in \mathbb{R} \times \mathbb{R}^{\mathcal{G}_h}$ to (35) 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 (35) 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 (35) may satisfy. Conceptually, the following definition is intended for schemes such that $S^{h,\alpha} u[x]$ is non-decreasing with respect to α.

Definition 2.12. The scheme (35) is:

- *Monotone* if for any $\alpha \in \mathbb{R}$, the scheme (32) with $S^h = S^{h,\alpha}$ is monotone in the sense of Definition 2.5.
- *Consistent* with the parametrized equation (34) if for any family of real numbers $(\alpha_h)_{h>0}$ converging to some $\alpha \in \mathbb{R}$ as h approaches zero, the scheme (32) with $S^h = S^{h,\alpha_h}$ is consistent with equation (34) 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 (35).
  - (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 (35), 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 (35), 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 (35), 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 (35), and any x_1, $x_2 \in \mathcal{G}_h$, one has
  $$|u[x_1] - u[x_2]| \leq \omega(\alpha, |x_1 - x_2|).$$

Obviously, if (35) 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 (35):

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* (35) *with $h = h_n$ and $u_n[x]$ is bounded, uniformly over $n \in \mathbb{N}$ and $x \in \mathcal{G}_{h_n}$. If* (35) *is monotone and consistent with* (34)*, then limits superior and inferior $\overline{u}$, $\underline{u} \colon \overline{X} \to \mathbb{R}$ defined as in* (33) *are respectively a viscosity subsolution and supersolution to* (34) *in $\overline{X}$.*

If (35) 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 (35). The proof is an adaptation of Perron's method to this setting. While we assume that (35) is stable in the sense of Definition 2.12, this assumption may be relaxed, see Remark 2.15 below.

Theorem 2.14 (Existence)**.** *Assume that* (35) *is monotone, continuous, and stable. Then for small $h > 0$, there exists a solution to* (35)*.*

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 (35). Since we assumed that (35) is stable, U is nonempty and there exists $\alpha \in \mathbb{R}$ defined by

$$\alpha := \sup_{(\overline{\alpha},\overline{u})\in U} \overline{\alpha}. \tag{36}$$

Let us show that there exists $u \colon \mathcal{G}_h \to \mathbb{R}$ such that (α, u) is a subsolution to (35). Let $((\alpha_n, u_n))_{n\in\mathbb{N}}$ be a maximizing sequence in the definition of α. 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, the sequence $(u_n)_{n\in\mathbb{N}}$ is bounded in $\mathbb{R}^{\mathcal{G}_h}$, and thus 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 (35).

Among all functions $u \colon \mathcal{G}_h \to \mathbb{R}$ such that (α, u) is a subsolution to (35), we choose one which maximizes the cardinal of the set $\mathcal{G}_* := \{x \in \mathcal{G}_h \mid S^{h,\alpha}u[x] < 0\}$. Let us show how such a function u may be transformed into a solution to the scheme.

First note that $\mathcal{G}_*$ may not be equal to $\mathcal{G}_h$, since in this case, by continuity of the scheme, there would exist $\alpha' > \alpha$ close enough to α so that $(\alpha', u) \in U$, contradicting (36).

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

$$\tilde{u}_\varepsilon[x] := \sup\{\overline{u}[x] \mid (\alpha, \overline{u}) \in U,\ \overline{u} = u \text{ in } \mathcal{G}_h \setminus \mathcal{G}_*,\ S^{h,\alpha}\overline{u}[x] \leq -\varepsilon \text{ in } \mathcal{G}_*\}. \tag{37}$$

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 $\overline{u}$. 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 $(\alpha, \overline{u}) \in U$, $S^{h,\alpha}\overline{u}[x] \leq -\varepsilon$ in $\mathcal{G}_*$, $\tilde{u}_\varepsilon \geq \overline{u}$ in $\mathcal{G}_h$, and $\tilde{u}_\varepsilon[x] = \overline{u}[x]$. Then by monotonicity, $S^{h,\alpha}\tilde{u}_\varepsilon[x] \leq S^{h,\alpha}\overline{u}[x]$. It follows that $(\alpha, \tilde{u}_\varepsilon)$ is a subsolution to (35) and that $S^{h,\alpha}\tilde{u}_\varepsilon[x] \leq -\varepsilon$ 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 (37), 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 (35). 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 (35) 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 (35). □

Remark 2.15. Since $h > 0$ is fixed in Theorem 2.14, 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 (27) 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* (13) *and* (14)*. Then the scheme* (27) *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 $\overline{u} \ge \underline{u}$ in $\mathcal{G}_h$. We need to show that

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

By the definition (26) of the operator $S^{h,\alpha}_{\mathrm{MABV2}}$, it suffices to prove that both $S^h_{\mathrm{MA}}\overline{u}[x] \le S^h_{\mathrm{MA}}\underline{u}[x]$ and $S^h_{\mathrm{BV2}}\overline{u}[x] \le S^h_{\mathrm{BV2}}\underline{u}[x]$. The second inequality follows directly from the definition (25) of S^h_{BV2}, so let us prove the first one.

By the definition (20) of S^h_{MA}, it suffices to prove that for any family $v = (v_1, \dots, v_I)$ of vectors of $\mathbb{Z}^d$ and any $\gamma \in \mathbb{R}^I_+$,

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

First note that the operator Δ^v_h was defined so that $\Delta^v_h\overline{u}[x] \ge \Delta^v_h\underline{u}[x]$ elementwise. If $B_h\overline{u}[x] = 0$, then $B_h\overline{u}[x]^{1/d} \le B_h\underline{u}[x]^{1/d}$, 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 $i \in \{1, \dots, d\}$), then, using (14) for the second

inequality,

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

and thus $B_h\overline{u}[x]^{1/d} \le B_h\underline{u}[x]^{1/d}$. Similarly, for any $e \in v$, if $A_h^e\overline{u}[x] = a_{\min}|e|^2$, then $A_h^e\overline{u}[x] \le A_h^e\underline{u}[x]$, and otherwise, using (13),

$$\begin{aligned} A_h^e\overline{u}[x] - A_h^e\underline{u}[x] &\le \langle e, (A(x, D_h\overline{u}[x]) - A(x, D_h\underline{u}[x]))e\rangle - ha_{\mathrm{LF}}|e|^2\Delta_h(\overline{u}-\underline{u})[x] \\ &\le a_{\mathrm{LF}}|e|^2(|D_h\overline{u}[x] - D_h\underline{u}[x]|_1 - \Delta_h(\overline{u}-\underline{u})[x]) \le 0, \end{aligned}$$

hence $A_h^e\overline{u}[x] \le A_h^e\underline{u}[x]$. We easily conclude that $S_{\mathrm{MA}}^h\overline{u}[x] \le S_{\mathrm{MA}}^h\underline{u}[x]$. □

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 $\varphi\colon \mathcal{G}_h \to \mathbb{R}$,

$$\max_{\substack{x_1,x_2\in\mathcal{G}_h \\ x_1\neq x_2}} \frac{|\varphi[x_1] - \varphi[x_2]|}{|x_1 - x_2|} \le C_{\mathcal{G}} \max_{\substack{x_1,x_2\in\mathcal{G}_h \\ |x_1-x_2|=h}} \frac{|\varphi[x_1] - \varphi[x_2]|}{h}. \tag{38}$$

Proposition 3.2 (Continuity)**.** *Assume* (38)*. Then the scheme* (27) *is continuous, in the sense of Definition 2.12.*

Proof. For any $x \in \mathcal{G}_h$, the function $\mathbb{R}^{\mathcal{G}_h} \to \overline{\mathbb{R}}$, $u \mapsto S_{\mathrm{MABV2}}^{h,\alpha}u[x]$ is a maximum over a compact set of continuous functions with values in $\mathbb{R} \cup \{-\infty\}$, see (20), (25), and (26). Hence it is a continuous function with values in $\mathbb{R} \cup \{-\infty\}$. It remains to prove that $S_{\mathrm{MABV2}}^{h,\alpha}u[x] > -\infty$.

By (38), there exists $e = \pm e_i$, $i \in \{1, \dots, d\}$, such that $x - he \in \mathcal{G}_h$. Therefore

$$S_{\mathrm{MABV2}}^{h,\alpha}u[x] \ge S_{\mathrm{BV2}}^{h}u[x] \ge D_h^e u[x] - \sigma_{P(x)}(e) = -\delta_h^{-e}u[x] - \sigma_{P(x)}(e) > -\infty,$$

which concludes the proof. □

Let us now study consistency of the scheme (27) with the degenerate elliptic equation

$$F_{\mathrm{MABV2}}^{\alpha}(x, Du(x), D^2u(x)) = 0 \quad \text{in } \overline{X}, \tag{39}$$

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

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

and $F_{\mathrm{MA}}(x, p, M)$ and $F_{\mathrm{BV2}}(x, p)$ are defined respectively in (8) and (24). 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* (9), (12), (17), *and* (18). *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*

$$S^{h,\alpha_h}_{\mathrm{MABV2}}\varphi[x] \le F^{\alpha}_{\mathrm{MABV2}}(x, D\varphi(x), D^2\varphi(x)) + o_{h\to 0}(1), \tag{40}$$

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

$$S^{h,\alpha_h}_{\mathrm{MABV2}}\varphi[x] \ge F^{\alpha}_{\mathrm{MABV2}}(x, D\varphi(x), D^2\varphi(x)) + o_{h\to 0}(1), \tag{41}$$

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 $a_h(x) \le b_h(x)$ for any $h > 0$ and $x \in \mathcal{G}_h$, with equality if $x \in K$. Then it suffices to show that

$$S^h_{\mathrm{MA}}\varphi[x] \leqq_K F_{\mathrm{MA}}(x, D\varphi(x), D^2\varphi(x)) + o_{h\to 0}(1), \tag{42}$$

$$S^h_{\mathrm{BV2}}\varphi[x] \leqq_K F_{\mathrm{BV2}}(x, D\varphi(x)) + o_{h\to 0}(1), \tag{43}$$

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

For any $x \in \mathcal{G}_h$ and $i \in \{1, \dots, d\}$, it holds that $T_h^{\pm e_i}\varphi[x] \ge \varphi(x \pm he_i)$, and using (9), 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 injecting first-order Taylor expansions of φ in the definition of S^h_{BV2} yields (43).

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 is Lipschitz continuous with respect to its last variable, uniformly with respect to its first variable,

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

Since $B \ge 0$ and using the definition (16) 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 (9), for h small, this may only happen if $x \notin K$), it holds that $B_h\varphi[x] = 0 \le B(x, D\varphi(x))$. 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 (9) and (18), 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 (12) and the same reasoning as above,

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

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 (18) for the last equality,

$$\begin{aligned}
-\langle \gamma, \Delta_h^v \varphi[x] - A_h^v \varphi[x] \rangle &= - \sum_{i=1}^{d(d+1)/2} \gamma_i (\Delta_h^{v_i} \varphi[x] - A_h^{v_i} \varphi[x]) \\
&\leqq_K - \sum_{i=1}^{d(d+1)/2} \gamma_i \left\langle v_i, \left(D^2\varphi(x) - A(x, D\varphi(x))\right) v_i \right\rangle \\
&\quad + \sum_{i=1}^{d(d+1)/2} \gamma_i O(h|v_i|^2 + h^2 |v_i|^4) \\
&= -\langle \mathcal{D}_v(\gamma), D^2\varphi(x) - A(x, D\varphi(x)) \rangle + O(h + h^2|v_i|^2) \\
&= -\langle \mathcal{D}_v(\gamma), D^2\varphi[x] - A(x, D\varphi(x)) \rangle + o_{h\to 0}(1),
\end{aligned} \tag{44}$$

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

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

We deduce (42) using (17) and that the affine map

$$\{\mathcal{D} \in \mathcal{S}_d^+ \mid \mathrm{Tr}(\mathcal{D}) = 1\} \to \mathbb{R}, \quad \mathcal{D} \mapsto L_{\mathcal{D}}(b, M) \tag{45}$$

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 (27) 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 $x \in K \cap \mathcal{G}_h$,

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

For the operator S_{MA}^h, we distinguish two cases:

(*General case*) If there exist $r_1 > 0$ and $r_2 \in (0, 1)$ such that the following refinements of (17) and (18) hold:

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

then, refining the last equality in (44) and using that the map (45) is $1/d$-Hölder continuous, one has, for small $h > 0$ and uniformly over $x \in K \cap \mathcal{G}_h$,

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

In dimension $d = 2$, when choosing V_h as in Remark B.9, one has $r_1 = 2r$ and $r_2 = r$, hence S_{MA}^h 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 $x \in K$. We 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 (17) 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 (13) and (14), 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 (27) 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 (26), 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* (9), (12), (17), *and* (18). *Then the scheme* (27) *is consistent with equation* (39)*, 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 $x \in \overline{X}$,

$$\limsup_{\substack{h>0,\, h\to 0 \\ x'\in\mathcal{G}_h,\, x'\to x}} S^{h,\alpha_h}_{\mathrm{MABV2}}\varphi[x'] \le (F^{\alpha}_{\mathrm{MABV2}})^*(x, D\varphi(x), D^2\varphi(x)), \tag{46}$$

$$\liminf_{\substack{h>0,\, h\to 0 \\ x'\in\mathcal{G}_h,\, x'\to x}} S^{h,\alpha_h}_{\mathrm{MABV2}}\varphi[x'] \ge (F^{\alpha}_{\mathrm{MABV2}})_*(x, D\varphi(x), D^2\varphi(x)). \tag{47}$$

If $x \in X$, then (46) and (47) follow respectively from (40) and (41), taking first the limit over h and then the limit over x'. If $x \in \partial X$, then (46) follows from (40) and (47) 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* (9), (12) *to* (14), (17) *to* (19), *and* (38). *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* (27) *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 $\alpha_1 \le 0$ 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 (27) 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 (27). Then for any $x \in \mathcal{G}_h$, $S^h_{\mathrm{BV2}}u[x] \le 0$. Choosing $e = \pm e_i$, $i \in \{1, \dots, d\}$ in the definition of S^h_{BV2}, it follows that $-\delta_h^{\pm e_i}u[x] \le \sigma_{P(x)}(\mp e_i)$. Since the compact set $\overline{P(x)}$ is continuous with respect to $x \in \overline{X}$ for the Hausdorff distance, there exists $C_P \ge 0$ such that for any $x \in \overline{X}$ and $i \in \{1, \dots, d\}$, $\sigma_{P(x)}(\pm e_i) \le C_P$. Hence $-\delta_h^{\pm e_i}u[x] \le C_P$. Using (38), we easily deduce that

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

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 (27). Then for any $x \in \mathcal{G}_h$, $S^h_{\mathrm{MA}} u[x] \leq -\alpha$. By (19), 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 (9), 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 (27) (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.

Remark 4.1 (Numerical complexity of the scheme). The motivation for Theorem 1.2 is to improve the numerical efficiency of the scheme.

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 (25) of S^h_{BV2} which, depending on the expression of the set-valued function P in (21), 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 applications, the maximum in (26) 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 (27) which would only require computing $S^h_{\mathrm{BV2}}u[x]$ at such points, reducing the numerical cost of handling the boundary condition (21).

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* [32]: 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

$$\gamma_v(\mathcal{D}) := (-\langle v_{i+1}^\perp, \mathcal{D} v_{i+2}^\perp \rangle)_{1 \le i \le 3}, \tag{48}$$

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

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

as the segment whose endpoints are $(1/|v_1|^2, 0)$ 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

$$\mathfrak{D}(\rho) = \frac{1}{2}\begin{pmatrix} 1+\rho_1 & \rho_2\\ \rho_2 & 1-\rho_1\end{pmatrix}. \tag{49}$$

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

$$\{\mathcal{D}\in\mathcal{S}_2^+ \mid \mathrm{Tr}(\mathcal{D}) = 1\} = \{\mathfrak{D}(\rho) \mid |\rho| \le 1\}. \tag{50}$$

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

$$\det \mathfrak{D}(\rho) = \frac{1}{4}(1-|\rho|^2), \qquad \mathrm{Cond}(\mathfrak{D}(\rho)) = \frac{1+|\rho|}{1-|\rho|}. \tag{51}$$

Let $v = (v_1, v_2, v_3)$ be a superbase of $\mathbb{Z}^2$, $b \geq 0$, 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

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

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 (48) and (49), then, by (50) 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 (51), 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

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

$$\int_X f(x)\,dx = \int_Y g(y)\,dy, \tag{52}$$

f being bounded and continuous almost everywhere and $1/g$ 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 $1/g\colon \mathbb{R}^d \to \mathbb{R}_+^*$ is bounded and Lipschitz continuous.

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

$$\inf_{T_\# f = g} \int_X |x - T(x)|^2 f(x)\,dx, \tag{53}$$

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,

$$\int_{T^{-1}(E)} f(x)\,dx = \int_E g(y)\,dy. \tag{54}$$

In the literature, it is typically assumed that:

$$\text{the set } X \text{ is convex.} \tag{55}$$

For simplicity, we will sometimes assume instead that:

$$\text{the set } X \text{ is strongly convex.} \tag{56}$$

It was proved in [7] (see also [34, Theorem 2.12]) that, under assumption (55), the optimal transport problem (53) 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 (54) that u is solution to the Monge-Ampère equation (1), where

$$A(x,p) = 0, \qquad B(x,p) = \frac{f(x)}{g(p)}. \tag{57}$$

Additionally, the constraint that $T(x) = Du(x) \in \overline{Y}$, for any $x \in X$, may be written as (21), where for any $x \in \overline{X}$,

$$P(x) = Y. \tag{58}$$

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

$$\inf_{T_\# f = g} \int_X c(x, T(x)) f(x)\, dx. \tag{59}$$

If c is defined by $c(x,y) = |x-y|^2$, this problem reduces to (53). It is also equivalent to (53) 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 [14, 29]), there exists a solution $T\colon X \to \overline{Y}$ to (59) 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

$$y = c\text{-exp}_x(p) \iff p = -D_x c(x,y), \tag{60}$$

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

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

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

$$A(x,p) = -D_{xx} c(x, c\text{-exp}_x(p)), \tag{61}$$

$$B(x,p) = \frac{f(x)}{g(c\text{-exp}_x(p))} |\det D_{xy} c(x, c\text{-exp}_x(p))|, \tag{62}$$

and the constraint that $T(x) = c\text{-exp}_x(Du(x)) \in \overline{Y}$, for any $x \in X$, may be written as (21), where for any $x \in \overline{X}$,

$$P(x) = -D_x c(x, Y). \tag{63}$$

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.

A notion of weak solutions to the Monge-Ampère equation that is directly related to the optimal transport problem (53) is the one of *Brenier solutions*.

Definition 5.2 (Brenier solution)**.** Assume (55), (57), and (58). A function $u\colon X \to \mathbb{R}$ is a *Brenier solution* to (1) and (21) if (i) it is convex and (ii) $(Du)_\# f = g$, in the sense that (54) 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* (55), (57), *and* (58). *If u, $v\colon X \to \mathbb{R}$ are two minimal Brenier solutions to* (1) *and* (21), *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 $\overline{X}$). If $\tilde{Y}$ is a subset of $\mathbb{R}^d$, let us also define $u^{cc}_{\tilde{Y}}\colon \mathbb{R}^d \to \overline{\mathbb{R}}$ by

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

so that $u^{cc} = u^{cc}_{\mathbb{R}^d}$. The motivation for the last definition is that under assumptions (55), (57), and (58), if $u\colon X \to \mathbb{R}$ is a Brenier solution to (1) and (21), then u^{cc}_Y is a minimal Brenier solution to (1) and (21).

Another standard notion of solutions to (1) and (21) is the one of *Aleksandrov solutions*:

Definition 5.4 (Aleksandrov solution)**.** Assume (55), (57), and (58). A function $u\colon X \to \mathbb{R}$ is an *Aleksandrov solution* to (1) and (21) 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 (21) if moreover $\partial u(X) \subset \overline{Y}$.

In our setting, Brenier and Aleksandrov solutions coincide, see for instance [18] (noting that the relevant part of [18] is not specific to the dimension two):

Proposition 5.5. *Assume* (55)*,* (57)*, and* (58)*. Then $u\colon X \to \mathbb{R}$ is a Brenier solution (respectively minimal Brenier solution) to* (1) *and* (21) *if and only if it is an Aleksandrov solution (respectively minimal Aleksandrov solution) to* (1) *and* (21)*.*

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

$$u(x) = \psi(x) \quad \text{on } \partial X. \tag{64}$$

Definition 5.6 (Aleksandrov solution to the Dirichlet problem)**.** Assume (55) and (57). A function $u\colon \overline{X} \to \mathbb{R}$ is an *Aleksandrov solution* to (1) and (64) 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 (21), then it is an Aleksandrov solution to (1) and (64) with $\psi = u_{|\partial X}$; however, this does not remain true if u is not minimal.

Below is the adaptation of [21, Theorem 1.6.2] to our setting. For simplicity, it is assumed that $g(p) = 1$ for any $p \in \mathbb{R}^d$, but note that we only use Theorem 5.7 as an intermediary result and that our convergence result, Theorem 5.21, is not limited to the case $g(p) = 1$.

Theorem 5.7 (Adapted from [21, Theorem 1.6.2])**.** *Assume* (57)*, 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* (64)*.*

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*

$$\max_{\substack{\mathcal{D}\in\mathcal{S}_d^+ \\ \mathrm{Tr}(\mathcal{D})=1}} L_{\mathcal{D}}(b,M) \leq 0 \iff b \leq \det M, \tag{65}$$

$$\max_{\substack{\mathcal{D}\in\mathcal{S}_d^+ \\ \mathrm{Tr}(\mathcal{D})=1}} L_{\mathcal{D}}(b,M) \geq 0 \iff b \geq \det M. \tag{66}$$

Proof. We refer to [27, Lemma 3.2.2] for the proof of the equivalence

$$\max_{\substack{\mathcal{D}\in\mathcal{S}_d^+ \\ \mathrm{Tr}(\mathcal{D})=1}} L_{\mathcal{D}}(b,M) = 0 \iff b = \det M. \tag{67}$$

Also, the first equality in (5) is proved in [27, 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 (65) follows from the continuity of $L_{\mathcal{D}}(b,M)$ with respect to $\mathcal{D} \in \mathcal{S}_d^+$, and (66) follows from (65) and (67). □

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

Proposition 5.9. *Assume* (55) *and* (57). *If, for some function $\psi \in C(\partial X)$, $u\colon \overline{X} \to \mathbb{R}$ is an Aleksandrov solution to* (1) *and* (64), *then u is a viscosity solution to* (2).

The proof is an adaptation of the one of [21, Proposition 1.3.4]. It uses [21, Lemma 1.4.1], which we recall below in our setting:

Lemma 5.10. *Assume* (55). *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 [21, 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 [21] 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 subsolutions and supersolutions to (39) are minimal Aleksandrov solutions to (2) and (21). Thus the remaining part of this subsection is devoted to the proof of the two following theorems:

Theorem 5.11. *Assume* (55), (57), *and* (58). *If* $u\colon \overline{X} \to \mathbb{R}$ *is a viscosity subsolution to* (39) *with* $\alpha \geq 0$, *then* $\alpha = 0$ *and* u *is a minimal Aleksandrov solution to* (1) *and* (21).

Theorem 5.12. *Assume* (56) *to* (58). *If* $u\colon \overline{X} \to \mathbb{R}$ *is a viscosity supersolution to* (39) *with* $\alpha \leq 0$, *then* $\alpha = 0$ *and* u_Y^{cc} *is a minimal Aleksandrov solution to* (1) *and* (21).

The result in the case of viscosity subsolutions is very close to [19, Theorem 2.1] (although the reformulation of the Monge-Ampère equation is not the same) and thus we follow the same sketch of proof. The case of viscosity supersolutions was not studied in [19] since it was not necessary for the proof of convergence of the scheme considered in that paper.

We will need the following comparison principle for equation (2):

Proposition 5.13 (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 $\underline{u}_\varepsilon \colon \overline{E} \to \mathbb{R}$ 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 [11, Theorem 3.3 and section 5.C], $\overline{u}_\varepsilon \le \underline{u}$ in E, and we conclude letting ε approach zero. □

Notice that we did not need to assume (57); however, if (57) holds, it may be shown that the assumption that $\operatorname{diam}(E) \le r$ is not necessary, see [24, Theorem V.2] for the argument.

We will also need the following lemmas.

Lemma 5.14. *Assume* (55) *and* (57). *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 [31, Theorem 1], it follows that u is convex. □

Lemma 5.15. *Assume* (55) *and* (58). *If* $u \colon X \to \mathbb{R}$ *is a convex viscosity subsolution to* (23), *then* $\partial u(X) \subset \overline{Y}$.

The proof of Lemma 5.15 is a direct transposition to our setting to the one of [19, Lemma 2.5], so we do not reproduce it here.

Lemma 5.16. *Assume* (56) *to* (58). *If* $u \colon \overline{X} \to \mathbb{R}$ *is a viscosity supersolution to* (39) *with* $\alpha \le 0$, *then* $Y \subset \partial u^{cc}(X)$.

Proof. 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 since X is strongly convex, for any $\varepsilon > 0$, there exists $\varphi_\varepsilon \in C^2(\overline{X})$ such that x_0 is a local minimum of $u - \varphi$ and

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

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 (39). Therefore x_0 may not belong to ∂X. □

Lemma 5.17. *Assume* (55), (57), *and* (58). *If $u\colon \overline{X} \to \mathbb{R}$ is a viscosity supersolution to* (39) *with $\alpha \leq 0$, then u^{cc}_Y is a viscosity supersolution to* (2). *Moreover, if $\alpha < 0$, $\varphi \in C^2(X)$, x_0 is a local minimum of $u^{cc}_Y - \varphi$ 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 $u^{cc}_Y - \varphi$ in X.

First we consider the case where $u^{cc}_Y(x_0) = u(x_0)$ and $\partial u^{cc}_Y(x_0) \subset Y$. Since $u^{cc}_Y \leq u$ 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 $\partial u^{cc}_Y(x_0) \subset Y$, $D\varphi(x_0)$ belongs to Y. Therefore

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

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 $u^{cc}_Y(x_0) < u(x_0)$ or $\partial u^{cc}_Y(x_0) \cap \partial Y \neq \emptyset$. In this case, there exists a unit vector $e \in \mathbb{R}^d$ such that $\langle e, D^2\varphi(x)e \rangle \leq 0$. 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 5.18. *Assume* (55) *and* (57). *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 5.13, up to choosing $r(x)$ smaller, $v \le u$ 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

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

For any $x \in G$, let $\hat{S}(x) := B_d(x, r(x)/5)$. By Vitali's covering theorem [16, Theorem 1.5.1], there exists a countable family $(x_i)_{i\in\mathbb{N}}$ of points of G such that $G = \bigcup_{i\in\mathbb{N}} S(x_i)$ and balls of the family $(\hat{S}(x_i))_{i\in\mathbb{N}}$ are disjoint. Thus

$$\mathcal{L}^d(\partial u(E)) \le \sum_{i\in\mathbb{N}} \mathcal{L}^d(\partial u(S(x_i))) \le K \sum_{i\in\mathbb{N}} \mathcal{L}^d(S(x_i)) = 5^d K \sum_{i\in\mathbb{N}} \mathcal{L}^d(\hat{S}(x_i)) \le 5^d K\varepsilon.$$

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

In the lemma below, the right-hand side in (68) is to be understood as the integral of function which coincides almost everywhere with $g(Du(\cdot)) \det D^2u(\cdot)$, the convex function u being twice differentiable almost everywhere by Aleksandrov's theorem [16, 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 5.19. *Assume* (55). *If $u\colon X \to \mathbb{R}$ is convex, then the set*

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

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 [16, Theorem 3.1.2]. □

Lemma 5.20. *Assume* (55). *If $u\colon X \to \mathbb{R}$ is convex, then for any Borel subset E of X,*

$$\int_{\partial u(E)} g(y)\,dy \ge \int_E g(Du(x)) \det D^2u(x)\,dx. \tag{68}$$

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 [16, Theorem 6.3.3]. By [16, 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 5.19,

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

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

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

It follows that

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

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 solution to (39) with $\alpha \geq 0$, it is both a viscosity subsolution to (2) and (23). Thus by Lemma 5.14 and Lemma 5.15, it is convex in X and $\partial u(X) \subset \overline{Y}$.

By Aleksandrov's theorem [16, Theorem 6.4.1], u is twice differentiable almost everywhere. Thus it is almost everywhere a classical subsolution to (39). It follows that for almost every $x \in X$, $F_{\mathrm{MA}}(x, Du(x), D^2u(x)) \geq 0$, with a strict inequality if $\alpha > 0$. Then, using Proposition 5.8, for any Borel subset E of X,

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

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

By Lemma 5.20,

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

The same is true when replacing E by $X \setminus E$, and by Lemma 5.19, $\partial u(E) \cap \partial u(X \setminus E)$ has Lebesgue measure zero. But since $\partial u(X) \subset \overline{Y}$,

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

It follows that

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

Thus $\alpha = 0$ and u is a minimal Aleksandrov solution to (1) and (21). □

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 (39) with $\alpha \leq 0$. By Aleksandrov's theorem [16, Theorem 6.4.1], u_Y^{cc} is twice differentiable almost everywhere. Then by Lemma 5.17, for almost every $x \in X$, $F_{\mathrm{MA}}(x, Du_Y^{cc}(x), D^2 u_Y^{cc}(x)) \geq 0$, with a strict inequality if $\alpha < 0$. Using Proposition 5.8, for any Borel subset E of X,

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

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

By Lemma 5.18 and Lemma 5.20,

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

The same is true when replacing E by $X \setminus E$. But by Lemma 5.16, $Y \subset \partial u_Y^{cc}(X)$ and thus

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

It follows that

$$\int_E f(x)\,dx = \int_E g(Du_Y^{cc}(x)) \det D^2 u_Y^{cc}(x)\,dx = \int_{\partial u_Y^{cc}(E)} g(y)\,dy.$$

Thus $\alpha = 0$ and u_Y^{cc} is a minimal Aleksandrov solution to (1) and (21). □

5.4 Convergence

We are now able to prove convergence of a family of numerical schemes (which includes the scheme (27), see section 3) for the Monge-Ampère equation, in the setting of quadratic optimal transport.

Theorem 5.21 (Convergence)**.** *Assume* (56) *to* (58)*. If the scheme* (35) *is monotone, consistent with equation* (39)*, 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* (35) *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* (21) *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 (35) 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 Lipschitz continuous function $u\colon \overline{X} \to \mathbb{R}$, satisfying $u(0) = 0$.

Let us show that $\alpha = 0$ and that u is a minimal Aleksandrov solution to (1) and (21). By Corollary 2.13, u is a viscosity solution (hence both a viscosity subsolution and supersolution) to (39). If $\alpha \leq 0$, then Theorem 5.12 implies that $\alpha = 0$. Thus we proved that $\alpha \geq 0$, and we may conclude by applying Theorem 5.11. □

6 Numerical application to nonimaging optics

We apply the finite difference scheme (27) to the far field refractor problem [22] in nonimaging optics. In this problem, and its variant, the near field refractor problem [23], light rays emanate from a light source located at the origin, in directions belonging to some subset $\hat{Y}$ of the unit sphere S^2, and with intensity described by a density $\hat{g}\colon \hat{Y} \to \mathbb{R}_+$. They propagate in an isotropic medium with index of refraction $n_1 > 0$, called medium I, until they hit a refractor, represented by a surface $\mathcal{R} \subset \mathbb{R}^3$. We will impose that $\mathcal{R}$ contains the point $e_3 = (0, 0, 1)$. The refracted rays then propagate in another isotropic medium, called medium II, with index of refraction $n_2 = \kappa n_1$, $0 < \kappa < 1$ (the case $\kappa > 1$, also studied in [22], has a different mathematical structure and is not

addressed here). The refracted rays continue to propagate until they hit a screen, represented by the plane $\mathbb{R}^2 \times \{\ell\}$, for some $\ell > 0$. The aim is to find a suitable shape for the refractor $\mathcal{R}$ so that refracted rays hit the screen $\mathbb{R}^2 \times \{\ell\}$ at points belonging to $\ell(X \times \{1\})$, for some given subset X of $\mathbb{R}^2$, with intensity described by $\ell^{-2} f(\cdot/\ell)$, for some density $f \colon X \to \mathbb{R}_+$. Here we consider the far field problem, that is, the limit problem as ℓ approaches $+\infty$, while in the near field problem ℓ is a fixed finite number. We illustrate the problem in Figure 1.

Let us define $\psi \colon \mathbb{R}^2 \to S^2$ by

$$\psi(x) := \mathfrak{n}(x)(x, 1), \qquad \mathfrak{n}(x) := (1 + |x|^2)^{-1/2}, \tag{69}$$

so that $\psi(x)$ is the orthogonal projection of the point $(x, 1)$ onto the unit sphere (thus $\psi(x)$ is a unit vector, while $\mathfrak{n}(x)$ is a normalization factor). Then the far field refractor problem is equivalent to prescribing that light rays be refracted in directions belonging to the set $\hat{X} := \psi(X)$, with intensity described by a density $\hat{f} \colon \hat{X} \to \mathbb{R}_+$ such that

$$f(x) = J\psi(x)\hat{f}(\psi(x)) = \mathfrak{n}(x)^3 \hat{f}(\psi(x)),$$

where $J\psi$ is the Jacobian of ψ, in the sense of [16, section 3.2.2]. We will assume that there exists $Y \subset \mathbb{R}^2$ such that $\hat{Y} = \psi(Y)$, and define $g \colon Y \to \mathbb{R}_+$ by

$$g(y) := J\psi(y)\hat{g}(\psi(y)) = \mathfrak{n}(y)^3 \hat{g}(\psi(y)). \tag{70}$$

It was shown in [22] that under suitable assumptions, including the mass balance condition (52) and the inequality

$$\inf_{\hat{x} \in \hat{X},\, \hat{y} \in \hat{Y}} \langle \hat{x}, \hat{y} \rangle \geq \kappa, \tag{71}$$

there exists an admissible refractor shape $\mathcal{R}$ to the far field refractor problem, of the form

$$\mathcal{R} = \{\exp(\kappa v(y))\psi(y) \mid y \in Y\}, \tag{72}$$

where $v \colon Y \to \mathbb{R}$ and $u \colon X \to \mathbb{R}$ are functions satisfying

$$v(y) = \sup_{x \in X}(-c(x, y) - u(x)), \qquad u(x) = \sup_{y \in Y}(-c(x, y) - v(y)), \tag{73}$$

with

$$c(x, y) := \frac{1}{\kappa} \log(1 - \kappa \langle \psi(x), \psi(y) \rangle), \tag{74}$$

and u is solution, in a generalized sense, to the Monge-Ampère equation (1), with the boundary condition (21) and with coefficients that we derive from [22] in Appendix C. The function v is solution to the same Monge-Ampère equation after reversing the roles of X and Y and of f and g, but numerically it is practical to discretize the equation satisfied u and not the one satisfied by v, since, in the setting considered below, the density g is Lipschitz continuous and uniformly nonzero on its domain, while f is not. As a remark, note that the solution to the near field refractor problem, that we do not approximate here, is described by the solution to a Monge-Ampère equation [23], but that this equation is of the form

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

where in comparison with (1) the functions A and B feature an additional dependency with respect to u, and its set of solutions is stable by an invariance that is not the addition of a constant (this equation, as well as some other ones of the form (75), fit in the framework of *generated Jacobian equations* [33]).

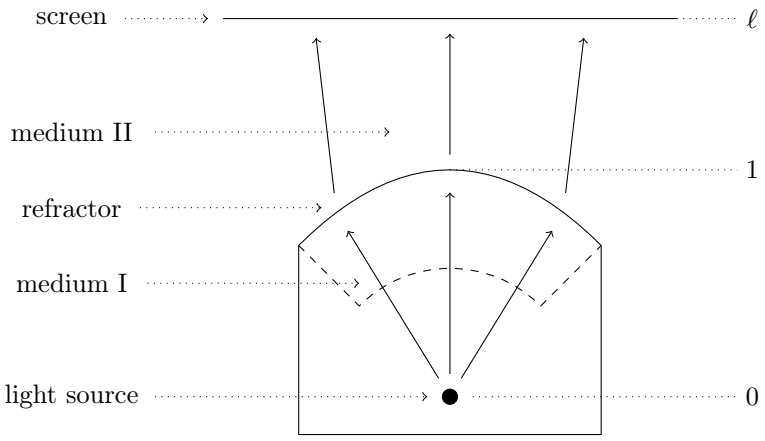

Figure 1: The far field refractor problem. Note that while in theory the light source belongs to medium I, in practice a second, spherical interface, represented by the dashed line, may be added between media I and II, since it would not refract light rays.

We consider the far field refractor problem with $\kappa = 2/3$, which is a typical value for a glass-air interface, source and target sets

$$\hat{Y} := \{\hat{y} \in S^2 \mid \hat{y}_3 > \delta_y\}, \qquad \hat{X} := \{\hat{x} \in S^2 \mid \hat{x}_3 > \delta_x\}, \qquad \delta_x = \delta_y = \cos(\pi/8),$$

corresponding to

$$Y = \{y \in \mathbb{R}^2 \mid |y|^2 < \delta_y^{-2} - 1\}, \qquad X = \{x \in \mathbb{R}^2 \mid |x|^2 < \delta_x^{-2} - 1\},$$

and with a uniform source density $\hat{g}(\hat{y}) = 1$, $\hat{y} \in \hat{Y}$, and a discontinuous target density f describing the image depicted in Figure 2, normalized so that (52) holds.

We approximate the pair $(0, u)$, where u is solution to (1) and (21) with the coefficients mentioned above, by a solution (α_h, u_h) to the numerical scheme (27) and to $u_h[0] = 0$ on the intersection $\mathcal{G}_h$ of the set X and of an $N \times N$ square Cartesian grid, where $N = 120$. More precisely, $\mathcal{G}_h := X \cap h\mathbb{Z}^d$, where

$$h := \frac{\operatorname{diam}(X)}{N} = \frac{2\sqrt{\delta_x^{-2} - 1}}{N} \approx 0.0069.$$

In (15) and (16), we choose $a_{\min} = -\infty$ and $a_{\mathrm{LF}} = b_{\mathrm{LF}} = 0$. These parameters do not fit in the theoretical framework, but this does not seem to be a problem in practice for our application (recall that a_{LF} and b_{LF} are Lax-Friedrichs relaxation parameters and that choosing them as zero improves consistency of the scheme but fails to guarantee its monotonicity, see Proposition 3.1 and Remark 3.4; recall also that the finiteness of $a_{\min}$ is used in the proof of Proposition 3.6).

We let $\mu := 2 + \sqrt{5} \approx 4.24$, and we choose $V_h = V^\mu$, where the set V^μ, defined in Appendix B, contains the following superbases:

$$\left(\begin{pmatrix}0\\-1\end{pmatrix}, \begin{pmatrix}1\\0\end{pmatrix}, \begin{pmatrix}-1\\1\end{pmatrix}\right), \qquad \left(\begin{pmatrix}0\\-1\end{pmatrix}, \begin{pmatrix}-1\\0\end{pmatrix}, \begin{pmatrix}1\\1\end{pmatrix}\right),$$
$$\left(\begin{pmatrix}0\\-1\end{pmatrix}, \begin{pmatrix}1\\-1\end{pmatrix}, \begin{pmatrix}-1\\2\end{pmatrix}\right), \qquad \left(\begin{pmatrix}0\\-1\end{pmatrix}, \begin{pmatrix}-1\\-1\end{pmatrix}, \begin{pmatrix}1\\2\end{pmatrix}\right),$$
$$\left(\begin{pmatrix}1\\-1\end{pmatrix}, \begin{pmatrix}1\\0\end{pmatrix}, \begin{pmatrix}-2\\1\end{pmatrix}\right), \qquad \left(\begin{pmatrix}-1\\-1\end{pmatrix}, \begin{pmatrix}-1\\0\end{pmatrix}, \begin{pmatrix}2\\1\end{pmatrix}\right).$$

Figure 2: Left: target image. Right: simulation of the scene using the appleseed® rendering engine; the small black disk at the bottom represents the light source.

We compute the second maximum in (20) using Theorem 1.2. We discretize the maximum in (25) over the set $\{(\cos(k\pi/50), \sin(k\pi/50) \mid k \in \{-50, \dots, 50\}\}$; however, for computing $\sigma_{P(x)}(e)$, which is itself defined as a supremum, we use a closed-form formula, see Appendix C.

We use a Newton method to solve the scheme (27), together with the additional constraint $u[0] = 0$. More precisely, we look for a zero of the function $(\alpha, u) \mapsto S_h^\alpha u[x]$ over the hyperplane $\mathbb{R} \times \{u \in \mathbb{R}^{\mathcal{G}_h} \mid u[0] = 0\}$. We use the initialization

$$u_h^{(0)}(x) := -c(x, 0) = -\frac{1}{\kappa} \log(1 - \kappa \mathfrak{n}(x)),$$

which describes a refractor with the uniform refraction property, see [22]. The Newton method converges in 12 iterations, with the stopping criterion

$$\max_{x \in \mathcal{G}_h} |S_h^\alpha u[x]| < 10^{-11}.$$

Let us now explain how we approximate the refractor $\mathcal{R}$ itself. Formally, if x is optimal in the first supremum in (73), then $-D_x c(x, y) - Du(x) = 0$, which we rewrite as $y = c\text{-exp}_x(Du(x))$, using the notation introduced in (60). This yields the formula

$$v(c\text{-exp}_x(Du(x))) = -c(x, c\text{-exp}_x(Du(x))) - u(x).$$

This motivates us to approximate the graph of the function v by the set

$$\{(y_h(x), v_h(x)) \mid x \in \tilde{\mathcal{G}}_h\},$$

where

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

$$y_h(x) := c\text{-exp}_x(D_h u_h(x)), \qquad v_h(x) := -c(x, c\text{-exp}_x(D_h u_h(x))) - u_h(x),$$

and the operator D_h is defined in (11). We then define the set

$$\mathcal{R}_h := \{\exp(\kappa v_h(x))\psi(y_h(x)) \mid x \in \tilde{\mathcal{G}}_h\},$$

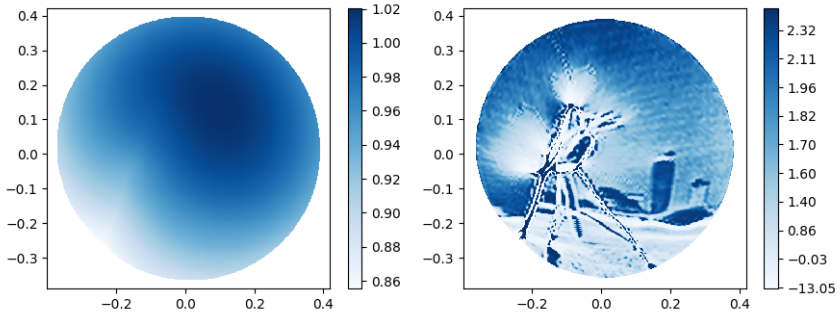

Figure 3: Approximation of the refractor (left) and of its pointwise curvature (right). Note that in the case of the curvature, the scale on the right is not linear.

up to adding a constant to the function v_h so that $\mathcal{R}_h$ contains a point close to $e_3 := (0, 0, 1)$, and we approximate the refractor $\mathcal{R}$ by the graph of a function

$$\tilde{v}\colon \operatorname{Conv}(\{(\hat{y}_1, \hat{y}_2) \mid \hat{y} \in \mathcal{R}_h\}) \to \mathbb{R}$$

which is a cubic (Clough-Tocher) interpolation of the points of $\mathcal{R}_h$.

In order to validate the numerical results, we simulated the scene of the far field refractor problem using the appleseed®[1] rendering engine. The chosen refractor is a triangle mesh finely approximating the graph of $\tilde{v}$, and the screen is at distance $\ell = 10$ from the light source at the origin. We present the result of the simulation in Figure 2. The bright circle around the reconstructed image corresponds to light rays near the boundary that do not hit the refractor.

In Figure 3, we display the graph of the function $\tilde{v}$, as well as a finite difference approximation of its pointwise Gaussian curvature, defined formally by the map

$$\tilde{y} \mapsto \frac{\det D^2\tilde{v}(\tilde{y})}{(1 + |D\tilde{v}(\tilde{y})|^2)^2}.$$

Here the finite differences do not need to be monotone, thus we simply approximate separately all elements of the Hessian of $\tilde{v}$, and we use the standard formula for computing the determinant. Since $\tilde{v}$ is not expected to be twice differentiable, the finite difference approximation is not necessarily convergent, but it is informative nevertheless. 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.

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 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

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

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 studying how parameters of the discretization may be chosen to make the evaluation of the scheme efficient in dimensions higher than two, possibly using Selling's formula in dimension three or its counterpart, Voronoi's first reduction of quadratic forms [10], in dimensions four and higher.

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. 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*. Springer International Publishing, 2021.

[6] 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.

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

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

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

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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, 2021. HAL preprint hal-03244537.

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

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

[18] 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.

[19] 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.

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

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

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

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

[24] 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.

[25] 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.

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

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

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

[29] 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.

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

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

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

[33] N. S. Trudinger. On the local theory of prescribed Jacobian equations. *Discrete Contin. Dyn. Syst.*, 34(4):1663–1681, 2014.

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

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

A Relation to the MA-LBR scheme

MA-LBR is a numerical scheme for the two-dimensional Monge-Ampère equation, which was introduced in [3]. Its natural extension to our setting would amount to replacing the definition (20) of the operator $S^h_{\mathrm{MA}}\colon \mathbb{R}^{\mathcal{G}_h} \to \overline{\mathbb{R}}^{\mathcal{G}_h}$ by

$$S^h_{\mathrm{MA}}u[x] := S^h_{\mathrm{adm}}u[x] \vee (B_h u[x] - \min_{v\in V_h} G(\Delta^v_h u[x] - A^v_h u[x])),$$

where V_h is a finite set of superbases of $\mathbb{Z}^2$, $S^h_{\mathrm{adm}}u\colon \mathbb{R}^{\mathcal{G}_h} \to \overline{\mathbb{R}}^{\mathcal{G}_h}$ enforces a discrete version of the admissibility constraint (4), and the function $G\colon (\mathbb{R}\cup\{+\infty\})^3 \to \mathbb{R}\cup\{+\infty\}$ is defined by

$$G(m) := \begin{cases} m_2 m_3 & \text{if } m_1 \geq m_2 + m_3, \\ m_1 m_3 & \text{if } m_2 \geq m_1 + m_3, \\ m_1 m_2 & \text{if } m_3 \geq m_1 + m_2, \\ \frac{1}{2}(m_1m_2 + m_1m_3 + m_2m_3) & \\ \qquad -\frac{1}{4}(m_1^2 + m_2^2 + m_3^2) & \text{else.} \end{cases}$$

The operator S^h_{adm} is typically chosen as

$$S^h_{\mathrm{adm}}u[x] := -\min_{v\in E_h} (\Delta^e_h u[x] - A^e_h u[x]),$$

where E_h is a finite subset of $\mathbb{Z}^2$. Then any solution u to $S^h_{\mathrm{MA}}u[x] = 0$ in $\mathcal{G}_h$ satisfies $\Delta^e_h u[x] - A^e_h u[x] \geq 0$ in $\mathcal{G}_h$, for any $e \in E_h$.

Recall that for any $v \in (\mathbb{Z}^2)^3$, $\gamma \in \mathbb{R}^3$, and sufficiently smooth function u,

$$\langle \mathcal{D}_v(\gamma), D^2u(x)\rangle = \sum_{i=1}^3 \gamma_i \langle v_i, D^2u(x)v_i\rangle \approx \sum_{i=1}^3 \gamma_i \Delta^{v_i}_h u[x] = \langle \gamma, \Delta^v_h u[x]\rangle.$$

Thus 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.1. *If v is a superbase of $\mathbb{Z}^2$ and $m \in \overline{\mathbb{R}}^3_+$, then*

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

Proof. Using that v is a superbase of $\mathbb{Z}^2$, and thus for any $1 \le i < j \le 3$, $\det(v_i, v_j) = \pm 1$, we compute that for any $\gamma \in \mathbb{R}^3_+$,

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

We conclude by noticing that

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

which is easily proved. □

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 (17) to (19). The motivation is to use this set V_h in (20). The construction of V_h is based on the Stern-Brocot tree of bases of $\mathbb{Z}^2$ (see [6] 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 non-negative 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 non-negative 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 $\min\{\langle u, u + v \rangle, \langle u + v, v \rangle\} = \langle u, v \rangle + \min\{|u^2|, |v|^2\} \ge \langle u, v \rangle + 1$.

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

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

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 $\cos\sphericalangle(u,v) = \langle u, v\rangle/(|u||v|)$. 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 [15, 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 non-diagonal 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 non-negative 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.

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*

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

and let $V_h = V^{\mu_h}$. Then (17) to (19) are satisfied.

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 (17).

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 (18) follows.

Finally, (19) 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 (17) and (18):

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

where the exponent in the first formula may be obtained by rewriting the relevant part of (51) 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 [30, 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 Coefficients of the Monge-Ampère equation in the far field refractor problem

In this appendix, we compute the coefficients of the Monge-Ampère equation associated to the far field refractor problem that we solve numerically in section 6. It was shown in [22] that under suitable assumptions, the far field refractor problem admits a solution of the form (72) and (73), where u is the potential function of the optimal transport problem (59) with a cost function c defined by (74) in the domain $\overline{X} \times \overline{Y}$. This means that u is solution, in a generalized sense, to the Monge-Ampère equation (1), with the boundary condition (21) and with coefficients defined by (61) to (63).

Recall that X, $Y \subset \mathbb{R}^2$, $\hat{X}$, $\hat{Y} \subset S^2$, and $\hat{X} = \psi(X)$, $\hat{Y} = \psi(Y)$, where ψ is defined in (69). Let us denote by $\hat{c}$ the real-valued function defined in a neighborhood of $\hat{X} \times \hat{Y}$ by

$$\hat{c}(\hat{x}, \hat{y}) := \frac{1}{\kappa} \log \left(1 - \kappa \frac{\langle \hat{x}, \hat{y} \rangle}{|\hat{x}||\hat{y}|} \right).$$

Then, for any $x \in \overline{X}$ and $y \in \overline{Y}$,

$$c(x, y) = \hat{c}(\psi(x), \psi(y)). \tag{76}$$

In [22], the optimal transport problem was expressed on domains $\hat{X}$ and $\hat{Y}$ with the cost function $\hat{c}$, rather than on X and Y with the cost function c. We consider the problem on X and Y, since those domains may be discretized as two-dimensional Cartesian grids while $\hat{X}$ and $\hat{Y}$ may not. This requires adapting to our setting the formulae that were obtained in [22] for the coefficients of the Monge-Ampère equation.

Proposition C.1. *Assume* (71) *and let $x \in \overline{X}$. Then the map $\overline{Y} \to -D_x c(x, \overline{Y})$, $y \mapsto -D_x c(x, y)$ is a bijection. If $c\text{-exp}_x \colon -D_x c(x, \overline{Y}) \to \overline{Y}$ denotes its inverse bijection, and if A and B are functions defined respectively by* (61) *and* (62)*, then for any $p \in -D_x c(x, \overline{Y})$,*

$$c\text{-exp}_x(p) = \frac{\lambda \mathfrak{n}(x) x + (1 - \kappa\lambda)\mathfrak{n}(x)^{-1} p}{\lambda \mathfrak{n}(x) - (1 - \kappa\lambda)\mathfrak{n}(x)^{-1} \langle p, x \rangle}, \tag{77}$$

$$A(x, p) = \kappa p \otimes p + \frac{\lambda}{1 - \kappa\lambda} (\mathfrak{n}(x)^4 x \otimes x - \mathfrak{n}(x)^2 I_2) - \mathfrak{n}(x)^2 (p \otimes x + x \otimes p), \tag{78}$$

$$B(x, p) = \frac{f(x)}{\mathfrak{m} \hat{g}(\hat{y})} \left| \frac{\mathfrak{n}(x)^4 - (\lambda + \kappa - \kappa\lambda^2)\mathfrak{n}(x)^2 \langle p, x \rangle}{(1 - \kappa\lambda)^2} - \frac{|p|^2}{1 - \kappa\lambda} + \mathfrak{n}(x)^2 \det(p, x)^2 \right|, \tag{79}$$

where $\lambda = \lambda(x, p)$, $\mathfrak{m} = \mathfrak{m}(x, p)$, and $\hat{y} = \hat{y}(x, p)$ are defined by

$$\lambda = \frac{\kappa \mathfrak{n}(x)^{-2} (|p|^2 + \langle p, x \rangle^2) + \left(1 - (1 - \kappa^2)\mathfrak{n}(x)^{-2} (|p|^2 + \langle p, x \rangle^2)\right)^{1/2}}{1 + \kappa^2 \mathfrak{n}(x)^{-2} (|p|^2 + \langle p, x \rangle^2)}, \tag{80}$$

$$\mathfrak{m} = \lambda \mathfrak{n}(x) - (1 - \kappa\lambda)\mathfrak{n}(x)^{-1} \langle p, x \rangle, \tag{81}$$

$$\hat{y} = \lambda \mathfrak{n}(x) \begin{pmatrix} x \\ 1 \end{pmatrix} + (1 - \kappa\lambda)\mathfrak{n}(x)^{-1} \begin{pmatrix} p \\ -\langle p, x \rangle \end{pmatrix}. \tag{82}$$

Proposition C.2. *Assume* (71) *and that Y is the centered Euclidean ball of radius $(\delta_y^{-2} - 1)^{1/2}$, for some $\delta_y \in (0, 1)$. Let P be the set-valued function defined by* (63)*. Then for any $x \in \overline{X}$ and any $e \in \mathbb{R}^2$ of unit norm,*

$$\sigma_{P(x)}(e) = \frac{\mathfrak{n}(x)\delta_y \langle e, y_* \rangle - \mathfrak{n}(x)^3 \delta_y \langle x, y_* \rangle \langle e, x \rangle - \mathfrak{n}(x)^3 \delta_y \langle e, x \rangle}{1 - \kappa \mathfrak{n}(x) \delta_y \langle x, y_* \rangle - \kappa \mathfrak{n}(x) \delta_y}, \tag{83}$$

where $y_* = y_*(x, e)$ and $\mathfrak{f} = \mathfrak{f}(x, e)$ are defined by

$$y_* = \frac{((\delta_y^{-2} - 1)|\mathfrak{f}|^2 - \kappa^2 \mathfrak{n}(x)^4 (1 - \delta_y^2)^2 \det(e, x)^2)^{1/2} \mathfrak{f}}{|\mathfrak{f}|^2} + \frac{\kappa \mathfrak{n}(x)^2 (1 - \delta_y^2) \det(e, x) \mathfrak{f}^\perp}{|\mathfrak{f}|^2}, \tag{84}$$

$$\mathfrak{f} = (1 - \kappa \mathfrak{n}(x) \delta_y) \mathfrak{n}(x) \delta_y e - \mathfrak{n}(x)^3 \delta_y \langle e, x \rangle x. \tag{85}$$

The motivation for Proposition C.2 is that $\sigma_{P(x)}(e)$ is part of the definition (25) of the operator S^h_{BV2}, and that this is the only occurrence of the function P in the definition of the scheme.

The rest of this section is devoted to the proofs of Propositions C.1 and C.2. Those proofs are based on the chain rule for differentiating composite functions, some simplifications based on identities from linear algebra such as (104), and the study of a constrained optimization problem in Lemma C.3. A natural objective, if the proposed numerical scheme is adapted to other settings in optimal transport or optics, is to automatize part of the construction of the coefficients of the Monge-Ampère equation by taking advantage of machine symbolic computation and automatic differentiation.

Proof of Proposition C.1. Let $x \in \overline{X}$, $y \in \overline{Y}$, $p := -D_x c(x,y)$, $\hat{x} := \psi(x)$, $\hat{y} := \psi(y)$, and $\hat{p} := -D_{\hat{x}}\hat{c}(\hat{x},\hat{y})$.

By (76) and the chain rule, using implicit summation on repeated indices,

$$p = D\psi_i(x)\hat{p}_i. \tag{86}$$

On the other hand, it is easily verified from the definition of $\hat{c}$ that

$$\langle \hat{p}, \hat{x} \rangle = 0. \tag{87}$$

Therefore it holds that

$$\begin{pmatrix} p \\ 0 \end{pmatrix} = \begin{pmatrix} D\psi_i(x) \\ -\mathfrak{n}(x)^2 \hat{x}_i \end{pmatrix} \hat{p}_i = \mathfrak{n}(x) \begin{pmatrix} I_2 - \mathfrak{n}(x)^2 x \otimes x & -\mathfrak{n}(x)^2 x \\ -\mathfrak{n}(x)^2 x^\top & -\mathfrak{n}(x)^2 \end{pmatrix} \hat{p}.$$

Inverting this system yields

$$\hat{p} = \mathfrak{n}(x)^{-1} \begin{pmatrix} I_2 & -x \\ -x^\top & -1 \end{pmatrix} \begin{pmatrix} p \\ 0 \end{pmatrix} = \mathfrak{n}(x)^{-1} \begin{pmatrix} p \\ -\langle p, x \rangle \end{pmatrix}. \tag{88}$$

It was proved in [22] that

$$\hat{y} = \lambda \hat{x} + (1 - \kappa\lambda)\hat{p}, \tag{89}$$

$$\langle \hat{x}, \hat{y} \rangle = \lambda, \tag{90}$$

$$-D_{\hat{x}\hat{x}}\hat{c}(\hat{x},\hat{y}) = \kappa \left(\hat{p} + \frac{\lambda}{1-\kappa\lambda}\hat{x} \right) \otimes \left(\hat{p} + \frac{\lambda}{1-\kappa\lambda}\hat{x} \right) - \frac{\lambda}{1-\kappa\lambda} I_3, \tag{91}$$

$$-D_{\hat{x}\hat{y}}\hat{c}(\hat{x},\hat{y}) = \frac{I_3}{1-\kappa\lambda} + \frac{\kappa\hat{y} \otimes \hat{x} - \hat{x} \otimes \hat{x} - \hat{y} \otimes \hat{y} + \lambda \hat{x} \otimes \hat{y}}{(1-\kappa\lambda)^2}, \tag{92}$$

where

$$\lambda := \frac{\kappa|\hat{p}|^2 + \mathfrak{h}^{1/2}}{1 + \kappa^2|\hat{p}|^2}, \qquad \mathfrak{h} := 1 - (1-\kappa^2)|\hat{p}|^2. \tag{93}$$

Note that (90) follows directly from (87) and (89).

We deduce (80) from (88) and (93). We deduce (82) from the definition of $\hat{x}$ and from (88) and (89).

Since $\hat{y} = (\mathfrak{n}(y)y, \mathfrak{n}(y))$, it follows from (82) that

$$\mathfrak{n}(y)y = \lambda \mathfrak{n}(x)x + (1-\kappa\lambda)\mathfrak{n}(x)^{-1}p, \tag{94}$$

$$\mathfrak{n}(y) = \lambda \mathfrak{n}(x) - (1-\kappa\lambda)\mathfrak{n}(x)^{-1}\langle p, x \rangle. \tag{95}$$

Dividing (94) by (95), we deduce that $y = c\text{-}\exp_x(p)$ satisfies (77), and thus that it is uniquely determined by x and p.

We may rewrite (90) as

$$\mathfrak{n}(x)\mathfrak{n}(y)(\langle x,y\rangle+1)=\lambda. \tag{96}$$

We compute that

$$D\psi(x)=\begin{pmatrix}\mathfrak{n}(x)I_2-\mathfrak{n}(x)^3x\otimes x\\ -\mathfrak{n}(x)^3x^\top\end{pmatrix},\qquad D\psi(y)=\begin{pmatrix}\mathfrak{n}(y)I_2-\mathfrak{n}(y)^3y\otimes y\\ -\mathfrak{n}(y)^3y^\top\end{pmatrix}.$$

Therefore, using (96) for (99), (100), and (102),

$$\begin{aligned}
D\psi_i(x)\hat{x}_i&=\mathfrak{n}(x)^2x-\mathfrak{n}(x)^4(|x|^2+1)x=0, &(97)\\
D\psi_i(y)\hat{y}_i&=\mathfrak{n}(y)^2y-\mathfrak{n}(y)^4(|y|^2+1)y=0, &(98)\\
D\psi_i(y)\hat{x}_i&=\mathfrak{n}(x)\mathfrak{n}(y)x-\mathfrak{n}(x)\mathfrak{n}(y)^3(\langle x,y\rangle+1)y\\
&=\mathfrak{n}(x)\mathfrak{n}(y)x-\lambda\mathfrak{n}(y)^2y, &(99)\\
D\psi_i(x)\hat{y}_i&=\mathfrak{n}(x)\mathfrak{n}(y)y-\mathfrak{n}(x)^3\mathfrak{n}(y)(\langle x,y\rangle+1)x\\
&=\mathfrak{n}(x)\mathfrak{n}(y)y-\lambda\mathfrak{n}(x)^2x, &(100)\\
D\psi_i(x)\otimes D\psi_i(x)&=\mathfrak{n}(x)^2I_2-2\mathfrak{n}(x)^4x\otimes x+\mathfrak{n}(x)^6(|x|^2+1)x\otimes x\\
&=\mathfrak{n}(x)^2I_2-\mathfrak{n}(x)^4x\otimes x, &(101)\\
D\psi_i(x)\otimes D\psi_i(y)&=\mathfrak{n}(x)\mathfrak{n}(y)I_2-\mathfrak{n}(x)^3\mathfrak{n}(y)x\otimes x-\mathfrak{n}(x)\mathfrak{n}(y)^3y\otimes y\\
&\quad+\mathfrak{n}(x)^3\mathfrak{n}(y)^3(\langle x,y\rangle+1)x\otimes y\\
&=\mathfrak{n}(x)\mathfrak{n}(y)I_2-\mathfrak{n}(x)^3\mathfrak{n}(y)x\otimes x-\mathfrak{n}(x)\mathfrak{n}(y)^3y\otimes y\\
&\quad+\lambda\mathfrak{n}(x)^2\mathfrak{n}(y)^2x\otimes y. &(102)
\end{aligned}$$

We also compute that

$$D_{x_ix_j}\psi(x)=\begin{pmatrix}3\mathfrak{n}(x)^5x_ix_jx-\mathfrak{n}(x)^3(\delta_{ij}x+x_ie_j+x_je_i)\\ 3\mathfrak{n}(x)^5x_ix_j-\mathfrak{n}(x)^3\delta_{ij}\end{pmatrix}.$$

Therefore, using (88),

$$\begin{aligned}
D^2\psi_i(x)\hat{p}_i&=3\mathfrak{n}(x)^4(\langle p,x\rangle-\langle p,x\rangle)x\otimes x-\mathfrak{n}(x)^2(\langle p,x\rangle-\langle p,x\rangle)I_2\\
&\quad-\mathfrak{n}(x)^2(p\otimes x+x\otimes p)\\
&=-\mathfrak{n}(x)^2(p\otimes x+x\otimes p). &(103)
\end{aligned}$$

Formula of $A(x,p)$. Using (61) for the first equality, (76) and the chain rule for the second one, ant the definition of $\hat{p}$ for the third one, we compute that

$$\begin{aligned}
A(x,p)&=-D_{xx}c(x,y)\\
&=-D\psi_i(x)\otimes D\psi_j(x)D_{\hat{x}_i\hat{x}_j}\hat{c}(\hat{x},\hat{y})-D^2\psi_i(x)D_{\hat{x}_i}\hat{c}(\hat{x},\hat{y})\\
&=-D\psi_i(x)\otimes D\psi_j(x)D_{\hat{x}_i\hat{x}_j}\hat{c}(\hat{x},\hat{y})+D^2\psi_i(x)\hat{p}_i.
\end{aligned}$$

We deduce (78) using (86), (91), (97), (101), and (103).

Formula of $B(x,p)$. Using (76) and the chain rule for the first equality, (92), (97) to (100),

and (102) for the second one, and (94) for the fourth one, we compute that

$$\begin{aligned}
-D_{xy}c(x,y) &= -D\psi_i(x)\otimes D_{\psi_j}(y)D_{\hat{x}_i\hat{y}_j}\hat{c}(\hat{x},\hat{y})\\
&= \frac{1}{1-\kappa\lambda}(\mathfrak{n}(x)\mathfrak{n}(y)I_2 - \mathfrak{n}(x)^3\mathfrak{n}(y)x\otimes x - \mathfrak{n}(x)\mathfrak{n}(y)^3 y\otimes y + \lambda\mathfrak{n}(x)^2\mathfrak{n}(y)^2 x\otimes y)\\
&\quad + \frac{\kappa}{(1-\kappa\lambda)^2}(\mathfrak{n}(x)^2\mathfrak{n}(y)^2 x\otimes y - \lambda\mathfrak{n}(x)^3\mathfrak{n}(y)x\otimes x - \lambda\mathfrak{n}(x)\mathfrak{n}(y)^3 y\otimes y\\
&\qquad\qquad + \lambda^2\mathfrak{n}(x)^2\mathfrak{n}(y)^2 y\otimes x)\\
&= \frac{\mathfrak{n}(x)\mathfrak{n}(y)}{(1-\kappa\lambda)^2}((1-\kappa\lambda)I_2 - \mathfrak{n}(x)^2 x\otimes x - \mathfrak{n}(y)^2 y\otimes y + (\lambda+\kappa-\kappa\lambda^2)\mathfrak{n}(x)\mathfrak{n}(y)x\otimes y\\
&\qquad\qquad + \kappa\lambda^2\mathfrak{n}(x)\mathfrak{n}(y)y\otimes x)\\
&= \frac{\mathfrak{n}(x)\mathfrak{n}(y)}{1-\kappa\lambda}I_2 - \frac{\mathfrak{n}(x)^3\mathfrak{n}(y)}{1-\kappa\lambda}x\otimes x - \mathfrak{n}(x)^{-1}\mathfrak{n}(y)p\otimes p - \lambda\mathfrak{n}(x)\mathfrak{n}(y)p\otimes x\\
&\quad - \frac{\kappa\mathfrak{n}(x)\mathfrak{n}(y)}{1-\kappa\lambda}x\otimes p
\end{aligned}$$

Using the formula

$$\begin{aligned}
&\det(aI_2 - bx\otimes x - cp\otimes p - dp\otimes x - ex\otimes p)\\
&= a^2 - ab|x|^2 - ac|p|^2 - a(d+e)\langle p,x\rangle + (bc-de)\det(p,x)^2
\end{aligned} \tag{104}$$

with suitable coefficients a, b, c, d, $e\in\mathbb{R}$ for the first equality, and using that $\mathfrak{n}(x)^{-2} - |x|^2 = 1$ for the second one,

$$\begin{aligned}
\det D_{xy}c(x,y) &= \frac{\mathfrak{n}(x)^2\mathfrak{n}(y)^2}{(1-\kappa\lambda)^2} - \frac{\mathfrak{n}(x)^4\mathfrak{n}(y)^2|x|^2}{(1-\kappa\lambda)^2} - \frac{\mathfrak{n}(y)^2|p|^2}{1-\kappa\lambda} - \frac{(\lambda+\kappa-\kappa\lambda^2)\mathfrak{n}(x)^2\mathfrak{n}(y)^2\langle p,x\rangle}{(1-\kappa\lambda)^2}\\
&\quad + \mathfrak{n}(x)^2\mathfrak{n}(y)^2\det(p,x)^2\\
&= \frac{\mathfrak{n}(x)^4\mathfrak{n}(y)^2 - (\lambda+\kappa-\kappa\lambda^2)\mathfrak{n}(x)^2\mathfrak{n}(y)^2\langle p,x\rangle}{(1-\kappa\lambda)^2} - \frac{\mathfrak{n}(y)^2|p|^2}{1-\kappa\lambda}\\
&\quad + \mathfrak{n}(x)^2\mathfrak{n}(y)^2\det(p,x)^2.
\end{aligned}$$

Thus, using (62) for the first equality and (70) for the second one,

$$\begin{aligned}
B(x,p) &= \frac{f(x)}{g(y)}|\det D_{xy}c(x,y)| = \frac{f(x)}{\mathfrak{n}(y)^3\hat{g}(\hat{y})}|\det D_{xy}c(x,y)|\\
&= \frac{f(x)}{\mathfrak{n}(y)\hat{g}(\hat{y})}\left|\frac{\mathfrak{n}(x)^4 - (\lambda+\kappa-\kappa\lambda^2)\mathfrak{n}(x)^2\langle p,x\rangle}{(1-\kappa\lambda)^2} - \frac{|p|^2}{1-\kappa\lambda} + \mathfrak{n}(x)^2\det(p,x)^2\right|.
\end{aligned}$$

We define $\mathfrak{m} := \mathfrak{n}(y)$, so that the above immediately yields (79), and (81) is a simple rewriting of (95). □

In order to prove Proposition C.2, we will need the following lemma.

Lemma C.3. *Let $\alpha\in\mathbb{R}$, β, $\eta>0$, $\mathfrak{a}$, $\mathfrak{b}\in\mathbb{R}^2$, and $\mathfrak{f} := \beta\mathfrak{a} - \alpha\mathfrak{b}$ be such that $\mathfrak{f}\neq 0$ and $\eta|\mathfrak{b}| < \beta$. Then*

$$\operatorname*{argmax}_{|y|=\eta}\frac{\alpha+\langle\mathfrak{a},y\rangle}{\beta+\langle\mathfrak{b},y\rangle} = \frac{(\eta^2|\mathfrak{f}|^2 - \eta^4\det(\mathfrak{a},\mathfrak{b})^2)^{1/2}\mathfrak{f} - \eta^2\det(\mathfrak{a},\mathfrak{b})\mathfrak{f}^\perp}{|\mathfrak{f}|^2}.$$

Proof. Let $y_* \in \mathbb{R}^2$, $|y_*| = \eta$ belong to the argmax. Let us write the first order optimality condition:

$$(\beta + \langle \mathfrak{b}, y_* \rangle)\mathfrak{a} - (\alpha + \langle \mathfrak{a}, y_* \rangle)\mathfrak{b} = \lambda y_*, \quad \lambda \in \mathbb{R}.$$

Then

$$\begin{aligned} 0 &= (\beta + \langle \mathfrak{b}, y_* \rangle)\langle \mathfrak{a}^\perp, y_* \rangle - (\alpha + \langle \mathfrak{a}, y_* \rangle)\langle \mathfrak{b}^\perp, y_* \rangle \\ &= \langle \beta\mathfrak{a}^\perp - \alpha\mathfrak{b}^\perp, y_* \rangle + \langle \mathfrak{b}, y_* \rangle\langle \mathfrak{a}^\perp, y_* \rangle - \langle \mathfrak{a}, y_* \rangle\langle \mathfrak{b}^\perp, y_* \rangle \\ &= \langle \mathfrak{f}^\perp, y_* \rangle + \det(\mathfrak{a}, \mathfrak{b})|y_*|^2 = \langle \mathfrak{f}^\perp, y_* \rangle + \eta^2 \det(\mathfrak{a}, \mathfrak{b}). \end{aligned}$$

Therefore $\langle \mathfrak{f}^\perp, y_* \rangle = -\eta^2 \det(\mathfrak{a}, \mathfrak{b})$, and thus

$$\langle \mathfrak{f}, y_* \rangle = \pm(|\mathfrak{f}|^2|y_*|^2 - \langle \mathfrak{f}^\perp, y_* \rangle^2)^{1/2} = \pm(\eta^2|\mathfrak{f}|^2 - \eta^4 \det(\mathfrak{a}, \mathfrak{b})^2)^{1/2}.$$

Using that

$$y_* = \frac{\langle \mathfrak{f}, y_* \rangle \mathfrak{f} + \langle \mathfrak{f}^\perp, y_* \rangle \mathfrak{f}^\perp}{|\mathfrak{f}|^2},$$

we deduce that $y_* \in \{y_1, y_{-1}\}$, where for $\varepsilon \in \{-1, 1\}$,

$$y_\varepsilon := \frac{\varepsilon(\eta^2|\mathfrak{f}|^2 - \eta^4 \det(\mathfrak{a}, \mathfrak{b})^2)^{1/2}\mathfrak{f} - \eta^2 \det(\mathfrak{a}, \mathfrak{b})\mathfrak{f}^\perp}{|\mathfrak{f}|^2}.$$

Using that $\langle \mathfrak{a}, \mathfrak{f}^\perp \rangle = \alpha \det(\mathfrak{a}, \mathfrak{b})$ and $\langle \mathfrak{b}, \mathfrak{f}^\perp \rangle = \beta \det(\mathfrak{a}, \mathfrak{b})$, we compute that

$$\frac{\alpha + \langle \mathfrak{a}, y_\varepsilon \rangle}{\beta + \langle \mathfrak{b}, y_\varepsilon \rangle} = \frac{(1 - \eta^2 \det(\mathfrak{a}, \mathfrak{b})/|\mathfrak{f}|^2)\alpha + \varepsilon(\eta^2|\mathfrak{f}|^2 - \eta^4 \det(\mathfrak{a}, \mathfrak{b})^2)^{1/2}\langle \mathfrak{a}, \mathfrak{f} \rangle}{(1 - \eta^2 \det(\mathfrak{a}, \mathfrak{b})/|\mathfrak{f}|^2)\beta + \varepsilon(\eta^2|\mathfrak{f}|^2 - \eta^4 \det(\mathfrak{a}, \mathfrak{b})^2)^{1/2}\langle \mathfrak{b}, \mathfrak{f} \rangle}.$$

Note that the denominator is always positive, since $|\mathfrak{b}||y_\varepsilon| = \eta|\mathfrak{b}| < \beta$. We deduce from $0 < |\mathfrak{f}|^2 = \langle \mathfrak{f}, \mathfrak{f} \rangle = \langle \beta\mathfrak{a} - \alpha\mathfrak{b}, \mathfrak{f} \rangle$ that $\beta\langle \mathfrak{a}, \mathfrak{f} \rangle > \alpha\langle \mathfrak{b}, \mathfrak{f} \rangle$, and thus that

$$\frac{\alpha + \langle \mathfrak{a}, y_1 \rangle}{\beta + \langle \mathfrak{b}, y_1 \rangle} > \frac{\alpha + \langle \mathfrak{a}, y_{-1} \rangle}{\beta + \langle \mathfrak{b}, y_{-1} \rangle},$$

which implies that $y_* = y_1$. □

Proof of Proposition C.2. Recall that

$$\sigma_{P(x)}(e) := \sup_{p \in P(x)} \langle e, p \rangle = \sup_{p \in -D_x c(x,Y)} \langle e, p \rangle.$$

By Proposition C.1, the map $\overline{Y} \to -D_x c(x, \overline{Y})$, $y \mapsto -D_x c(x, y)$ is a continuous bijection, hence

$$\sigma_{P(x)}(e) = \sup_{p \in -D_x c(x,Y)} \langle e, p \rangle = \max_{p \in \partial(-D_x c(x,Y))} \langle e, p \rangle = \max_{y \in \partial Y} -\langle e, D_x c(x, y) \rangle.$$

If $y \in \overline{Y}$, we compute that

$$\begin{aligned}
-D_x c(x,y) &= -D\psi_i(x) D_{\hat{x}_i} \hat{c}(\psi(x), \psi(y)) \\
&= \big(\mathfrak{n}(x) I_2 - \mathfrak{n}(x)^3 x \otimes x, -\mathfrak{n}(x)^3 x\big) \frac{\psi(y) - \langle \psi(x), \psi(y)\rangle \psi(x)}{1 - \kappa \langle \psi(x), \psi(y)\rangle} \\
&= \frac{1}{1 - \kappa \langle \psi(x), \psi(y)\rangle} (\mathfrak{n}(x)\mathfrak{n}(y) y - \mathfrak{n}(x)^2 \langle \psi(x), \psi(y)\rangle x - \mathfrak{n}(x)^3 \mathfrak{n}(y)(\langle x, y\rangle + 1) x \\
&\qquad + \mathfrak{n}(x)^4 (|x|^2 + 1) \langle \psi(x), \psi(y)\rangle x) \\
&= \frac{1}{1 - \kappa \langle \psi(x), \psi(y)\rangle} (\mathfrak{n}(x)\mathfrak{n}(y) y - \mathfrak{n}(x)^2 \langle \psi(x), \psi(y)\rangle x - \mathfrak{n}(x)^2 \langle \psi(x), \psi(y)\rangle x \\
&\qquad + \mathfrak{n}(x)^2 \langle \psi(x), \psi(y)\rangle x) \\
&= \frac{\mathfrak{n}(x)\mathfrak{n}(y) y - \mathfrak{n}(x)^2 \langle \psi(x), \psi(y)\rangle x}{1 - \kappa \langle \psi(x), \psi(y)\rangle} \\
&= \frac{\mathfrak{n}(x)\mathfrak{n}(y) y - \mathfrak{n}(x)^3 \mathfrak{n}(y) \langle x, y\rangle x - \mathfrak{n}(x)^3 \mathfrak{n}(y) x}{1 - \kappa \mathfrak{n}(x)\mathfrak{n}(y) \langle x, y\rangle - \kappa \mathfrak{n}(x) \mathfrak{n}(y)}.
\end{aligned}$$

If $y \in \partial Y$, then $|y|^2 = \delta_y^{-2} - 1$ and thus $\mathfrak{n}(y) = (|y|^2 + 1)^{-1/2} = \delta_y$. Therefore

$$\begin{aligned}
\sigma_{P(x)}(e) &= \max_{y \in \partial Y} -\langle e, D_x c(x,y)\rangle \\
&= \max_{y \in \partial Y} \frac{\mathfrak{n}(x)\delta_y \langle e, y\rangle - \mathfrak{n}(x)^3 \delta_y \langle x, y\rangle \langle e, x\rangle - \mathfrak{n}(x)^3 \delta_y \langle e, x\rangle}{1 - \kappa \mathfrak{n}(x) \delta_y \langle x, y\rangle - \kappa \mathfrak{n}(x) \delta_y} \\
&= \max_{y \in \partial Y} \frac{\alpha + \langle \mathfrak{a}, y\rangle}{\beta + \langle \mathfrak{b}, y\rangle} = \max_{|y| = (\delta_y^{-2} - 1)^{1/2}} \frac{\alpha + \langle \mathfrak{a}, y\rangle}{\beta + \langle \mathfrak{b}, y\rangle},
\end{aligned}$$

where

$$\begin{aligned}
&\alpha := -\mathfrak{n}(x)^3 \delta_y \langle e, x\rangle, && \mathfrak{a} := \mathfrak{n}(x)\delta_y e - \mathfrak{n}(x)^3 \delta_y \langle e, x\rangle x, \\
&\beta := 1 - \kappa \mathfrak{n}(x) \delta_y, && \mathfrak{b} := -\kappa \mathfrak{n}(x) \delta_y x.
\end{aligned}$$

We let y_* denote a solution to the above maximum, so that (83) holds.

Let $\mathfrak{f} := \beta \mathfrak{a} - \alpha \mathfrak{b}$, so that (85) holds. The vector $\mathfrak{f}$ is always nonzero, since otherwise $x = \pm |x| e$ and then, using that $1 - \mathfrak{n}(x)^2 |x|^2 = \mathfrak{n}(x)^2$ and that, by (71), $\mathfrak{n}(x) = \langle \psi(x), \psi(0) \geq \kappa$,

$$\mathfrak{f} = (\mathfrak{n}(x)\delta_y - \kappa \mathfrak{n}(x)^2 \delta_y^2 - \mathfrak{n}(x)^3 \delta_y |x|^2) e = (\mathfrak{n}(x)^3 \delta_y - \kappa \mathfrak{n}(x)^2 \delta_y^2) e \neq 0.$$

Since $|x| = (\mathfrak{n}(x)^{-2} - 1)^{1/2}$, it always holds that

$$(\delta_y^{-2} - 1)^{1/2} |\mathfrak{b}| = \kappa \mathfrak{n}(x)(1 - \delta_y^2)^{1/2} |x| = \kappa (1 - \mathfrak{n}(x)^2)^{1/2} (1 - \delta_y^2)^{1/2} < 1 - \kappa \mathfrak{n}(x) \delta_y = \beta.$$

We prove (82) by applying Lemma C.3, using that

$$(\delta_y^{-2} - 1) \det(\mathfrak{a}, \mathfrak{b}) = -\kappa \mathfrak{n}(x)^2 (1 - \delta_y^2) \det(e, x).$$

This concludes the proof. □