Chapter 2

The Monge–Ampère equation

Michael Neilan[a,*], Abner J. Salgado[b] and Wujun Zhang[c]
[a]Department of Mathematics, University of Pittsburgh, Pittsburgh, PA, United States
[b]Department of Mathematics, University of Tennessee, Knoxville, TN, United States
[c]Department of Mathematics, Rutgers University, Piscataway, NJ, United States
[*]Corresponding author: e-mail: neilan@pitt.edu

Chapter Outline

Abstract
We review recent advances in the numerical analysis of the Monge–Ampère equation. Various computational techniques are discussed including wide stencil finite difference schemes, two-scaled methods, finite element methods, and methods based on geometric

Handbook of Numerical Analysis, Vol. 21. https://doi.org/10.1016/bs.hna.2019.05.003
© 2020 Elsevier B.V. All rights reserved.

considerations. Particular focus is the development of appropriate stability and consistency estimates which lead to rates of convergence of the discrete approximations. Finally we present numerical experiments which highlight each method for a variety of test problem with different levels of regularity.

Keywords: Monge–Ampère, Convergence analysis, Error estimates, Comparison principle, Fully nonlinear equations

AMS Classification Codes: 65N12, 65N15, 35B51, 35D40, 35J96

1 Introduction

> All exact science is dominated by the idea of approximation. When a man tells you that he knows the exact truth about anything, you are safe in inferring that he is an inexact man.
>
> Russell (1931)

In this chapter we review recent progress in the numerical treatment of Monge–Ampère type equations. In its simplest form, and assuming Dirichlet boundary conditions, the problem we consider is to seek a scalar function u satisfying the partial differential equation (PDE)

$$\det D^2 u(x) = f(x) \quad x \in \Omega, \tag{1a}$$

$$u(x) = g(x) \quad x \in \partial\Omega. \tag{1b}$$

Here, $D^2 u$ denotes the Hessian matrix of u, $f \geq 0$, and g are given functions, and $\Omega \subset \mathbb{R}^d$ is a bounded, convex domain. Problem (1) is a prototypical second order, fully nonlinear PDE, and it arises in several broad applications in differential geometry, meteorology, cosmology, economics, and optimal mass transportation problems. Some of these applications are briefly described below.

Despite its growing list of applications, and in contrast to its extensive and mature PDE theory, the construction and analysis of computational methods for (1) is still a relatively new and emerging field in numerical analysis. Numerical algorithms, based on geometric considerations, for the two-dimensional problem ($d = 2$) first appeared in 1988 in Oliker and Prussner (1988), and the extension to practical three-dimensional schemes were not introduced until some 20 years later (Brenner and Neilan, 2012; Feng and Neilan, 2009; Froese and Oberman, 2011a,b). Other early attempts that deserve mention are the least squares and augmented Lagrangian approaches of Dean and Glowinski (2003, 2004, 2005, 2006a,b), and we refer the reader to Feng et al. (2013) for more details on these schemes.

The reasons for this delayed development in numerical methods are plentiful. The most evident obstacle is the full nonlinearity of the problem. However, this is arguably a secondary difficulty, as black-box nonlinear solvers can, at least heuristically, be applied to algebraic systems resulting from

discretizations of (1). A rather fundamental difficulty to construct, and especially to analyze, computational methods for Monge–Ampère type equations is the variety of solution concepts and, correspondingly, the low regularity solutions generically possess. As we explain below, weak solutions are not based on variational principles, but rather on either geometric considerations or by monotonicity conditions of test functions that touch the graph of the solution from above or below. These solution concepts are difficult to mimic at the discrete level, and as a result, the construction of convergent schemes is an arduous task. Finally, as if these complications were not enough, the Monge–Ampère equation (1) is usually supplemented by the constraint that the solution u is convex. This is not only because of geometric applications, but in many cases a necessary condition for uniqueness, and for the existence of a well-developed PDE theory. As convexity is a global constraint, it is very difficult to enforce it in a discrete setting.

Nonetheless, an explosion of results and new techniques to develop them in computational methods for (1) have occurred during the last 10 years. These include the construction of monotone, wide stencil finite difference schemes, semi-Lagrangian methods, and finite element methods. Within only the past few years, significant progress has been made in the convergence analysis with an emphasis on the rates of convergence for various discretization schemes.

The main goal of this chapter is to highlight these recent advances in the numerical analysis of the Monge–Ampère problem (1). To this end, we organize the chapter as follows. After stating some geometric applications and a brief PDE theory of the Monge–Ampère problem in this section, we discuss wide stencil finite difference schemes in Section 2. There we introduce the monotone finite difference schemes (Froese and Oberman, 2011a,b; Oberman, 2008b) and the corresponding filtered schemes (Froese and Oberman, 2013), lattice reduction schemes (Benamou et al., 2016), methods based on power diagrams (Mirebeau, 2015), and the so-called two scale methods (Nochetto and Ntogkas, 2018; Nochetto et al., 2019a,b). Of particular focus will be the rates of convergence of these schemes if available. Next, in Section 3, we review the original method of Oliker and Prussner (1988), which in honour of its proponents henceforth we shall call the Oliker–Prussner scheme. This method is based on geometric interpretations of the Monge–Ampère operator and the notion of Alexandrov solutions. Again, the emphasis of the discussion is on consistency error and pointwise rates of convergence recently established in Nochetto and Zhang (2019). Section 4 discusses finite element methods for both smooth and singular solutions. Finally in Section 5 we perform some numerical experiments using some of the methods we discuss in this review for a variety of test problems with different levels of regularity.

We remark that, by design, this review has several major omissions. We intend to minimize the overlap between two other existing, and rather recent, reviews on fully nonlinear problems in general and the Monge–Ampère

equation in particular. Namely, the overview of Feng et al. (2013), which is dedicated to the Monge–Ampère equation exclusively, and Neilan et al. (2017) which contains a chapter on the Monge–Ampère equation, and where the reader can find much more details, for instance, on the semi-Lagrangian schemes described in Section 2.8.1.

1.1 Geometric applications

To draw connections with the theme of the current volume in the *Handbook of Numerical Analysis*, and to further emphasize the prevalence of the Monge–Ampère problem, in this section we briefly summarize some applications with a geometric flavour where the Monge–Ampère problem plays an essential role.

1.1.1 Gauss curvature problem

The classic Gauss curvature problem (cf. Bakelman, 1994; Guan and Spruck, 1993; Oliker, 1984) seeks a manifold $\mathcal{M} \subset \mathbb{R}^{n+1}$ with prescribed boundary and Gauss curvature $\mathcal{K}$. We recall that Gauss curvature is the product of the principal curvatures, which themselves are the eigenvalues of the shape operator (or Weingarten map). One may reduce this problem to a PDE problem of Monge–Ampère type if one assumes that the manifold is the graph of a function, i.e.,

$$\mathcal{M} = \{(x, u(x)) : \quad u : \Omega \to \mathbb{R}\}.$$

The shape operator is given by $s = \mathrm{I}^{-1}\,\mathrm{II}$, where I and II denote, respectively, the first and second fundamental forms. In the case that $\mathcal{M}$ is the graph of the function u, we have $\mathrm{I} = I + \nabla u \otimes \nabla u$ and $\mathrm{II} = \frac{D^2 u}{\sqrt{1+|\nabla u|^2}}$, where I denotes the $d \times d$ identity matrix. Therefore the Gauss curvature is given by

$$\mathcal{K} = \det(s) = \frac{\det(\mathrm{II})}{\det(\mathrm{I})} = \frac{\det D^2 u}{(1+|\nabla u|^2)^{(d+2)/2}}.$$

Thus, the problem is to find a scalar function $u : \bar{\Omega} \to \mathbb{R}$ satisfying

$$\det D^2 u(x) = \mathcal{K}(x)(1+|\nabla u(x)|^2)^{(d+2)/2} \quad \text{in } \Omega, \tag{2a}$$

$$u(x) = g(x) \quad \text{on } \partial\Omega. \tag{2b}$$

In conclusion the Gauss curvature problem, in this setting, seeks solutions of a Monge–Ampère type problem with lower-order terms.

1.1.2 Reflector design problem

The reflector design problem (Norris and Westcott, 1976; Oliker, 1987; Oliker and Waltman, 1987; Wang, 1996) can be posed as follows: Let S^2 be the unit

sphere in $\mathbb{R}^3$ centred at the origin, and let $\Omega, \mathcal{O}$ be two disjoint domains on S^2. Let f be a positive function defined on $\mathcal{O}$, and suppose that rays originate from the origin with density ρ. We then seek a surface, called Γ, whose radial projection onto S^2 equals Ω, such that the directions of the reflected rays cover $\mathcal{O}$ with distributed density equal to f.

To formulate a PDE model for this problem, we set $\Gamma = \{xm(x) : x \in \Omega\}$, so that if a ray radiates from the origin with direction x, then it is reflected at the point $xm(x)$. This will create a reflected ray in the direction $T(x) \in \mathcal{O}$. Now if we denote by $\boldsymbol{n}$ the unit normal of Γ at xm (x), then we have $T(x) - x = -2(x \cdot \boldsymbol{n})\boldsymbol{n}$, and calculations show that $\boldsymbol{n} = (\nabla m - mx)/\sqrt{m^2 + |\nabla m|^2}$. Here, $\nabla = e^{ij}\partial_i x \partial_j$, where x is a smooth parametrization of S^2, $e = e_{ij}\mathrm{d}t^i\mathrm{d}t^j$ is the first fundamental form of S^2, $e^{ij} = (e_{ij})^{-1}$, and $\partial_j = \partial/\partial t^j$. Combining these two identities we find that the direction T is related to m via

$$T(x) = \frac{2m\nabla m + (|\nabla m|^2 - m^2)x}{m^2 + |\nabla m|^2}. \tag{3}$$

Next, if the directions of the reflected light do not overlap and if no loss of energy occurs in the reflection, then we have the energy conservation property

$$\int_E \rho(x)\mathrm{d}x = \int_{T(E)} f(y)\mathrm{d}y = \int_E f(T(x))\frac{|\partial_1 T(x) \times \partial_2 T(x)|}{\det(e_{ij})}\mathrm{d}x$$

for all Borel sets $E \subset \Omega$. Thus we have, at least formally,

$$\frac{|\partial_1 T(x) \times \partial_2 T(x)|}{\det(e_{ij})} = \frac{\rho(x)}{f(T(x))}.$$

Finally, we set $u(x) = 1/m(x)$, and substitute (3) into this last equation to get the following problem of Monge–Ampère type (see Oliker and Newman, 1993; Wang, 1996 for details)

$$\frac{\det(D^2 u + (u - \eta)e_{ij})}{\eta^2 \det(e_{ij})} = \frac{\rho(x)}{f(T(x))} \qquad x \in \Omega,$$

where T is given by (3) and $\eta = (|\nabla u|^2 + u^2)/(2u)$.

1.1.3 Affine plateau problem

Following Trudinger and Wang (2005, 2008) and Calabi (1990), we consider the following problem. Let $\mathcal{M}_0 \subset \mathbb{R}^{d+1}$ be a bounded and connected hypersurface with smooth boundary that is locally uniformly convex We denote by $S[\mathcal{M}_0]$ the set of locally uniformly convex hypersurfaces that can be smoothly deformed from $\mathcal{M}_0$ within the family of locally uniformly convex hypersurfaces and whose Gauss map images lie in that of $\mathcal{M}_0$. As in Section 1.1.1, for a manifold $\mathcal{M}$ we denote by II its second fundamental

form and by $\mathcal{K}$ its Gauss curvature. Associated with $\mathcal{M}$ is the *Berwald–Blaschke metric*

$$g = \mathcal{K}^{-1/(d+2)}\mathrm{II},$$

which is an affine invariant Riemannian metric on the surface. The *affine Plateau problem* is then to determine the maximizer of the *affine area functional*

$$A(\mathcal{M}) = \int_{\mathcal{M}} \mathcal{K}^{1/(d+2)} \mathrm{d}\mathcal{M}$$

over $S[\mathcal{M}_0]$.

Recall that if $\mathcal{M} = \mathcal{M}_u$ is the graph of a function $u : \Omega \to \mathbb{R}$, with $\Omega \subset \mathbb{R}^n$, then the Gauss curvature is $\mathcal{K} = \det(D^2 u)/(1 + |\nabla u|^2)^{(d+2)/2}$, and so, we have by a change of variables,

$$A(\mathcal{M}_u) = \int_{\Omega} (\det D^2 u(x))^{1/(d+2)} \mathrm{d}x.$$

Thus if $\mathcal{M}_0$ is the graph of a locally uniformly convex g, then in the graph case, $S[\mathcal{M}_0]$ consists of the graphs of locally uniformly convex functions $v \in C^2(\Omega) \cap C^0(\bar{\Omega})$ satisfying $v = g$ on $\partial\Omega$ and $\nabla v(\Omega) \subset \nabla g(\Omega)$. In this setting the affine Plateau problem seeks u such that

$$A(\mathcal{M}_u) = \sup\{A(\mathcal{M}_v) : \ \mathcal{M}_v \in S[\mathcal{M}_0]\}.$$

Formally taking the Euler–Lagrange equation yields the *affine maximal surface equation*

$$\operatorname{cof} D^2 u : D^2 w = 0, \qquad w = (\det D^2 u)^{-(d+1)/(d+2)}.$$

1.1.4 Optimal mass transport problem

> *This problem appeared as a generalization of an earlier considered practical problem of assigning production locations on a railway network to consumption locations with minimum total transportation expenses.*
>
> Kantorovich (2004)

The optimal mass transport problem was originally proposed by G. Monge in the 18th century to find the optimal way to move oil to an excavation with minimal transportation cost. In general, the mass transport problem seeks, for two given sets and densities, the optimal mass-preserving mapping between them.

In further detail, given bounded $\Omega, \mathcal{O} \subset \mathbb{R}^d$ and measures $\rho_\Omega : \Omega \to \mathbb{R}$, $\rho_\mathcal{O} : \mathcal{O} \to \mathbb{R}$, the optimal transport problem with quadratic cost seeks a map $T : \Omega \to \mathcal{O}$ such that $T_{\#}\rho_\Omega = \rho_\mathcal{O}$ that minimizes the functional

$$\frac{1}{2}\int_{\Omega} |x - T(x)|^2 \mathrm{d}\rho_\Omega(x) \tag{4}$$

over all mass-preserving maps. Here, we assume that the measures are absolutely continuous with respect to Lebesgue measure, with $\mathrm{d}\rho_\Omega = f_\Omega \mathrm{d}x$ and $\mathrm{d}\rho_\mathcal{O} = f_\mathcal{O}\mathrm{d}x$, and that the measures satisfy the mass balance condition

$$\int_\Omega f_\Omega(x)\mathrm{d}x = \int_\mathcal{O} f_\mathcal{O}(x)\mathrm{d}x.$$

Above, we denoted by $T_{\#}\rho_\Omega$ the pushforward of the measure ρ_Ω under the mapping T, i.e., under the given assumptions, we have

$$\int_E f_\mathcal{O}(x)\mathrm{d}x = \int_{T^{-1}(E)} f_\Omega(x)\mathrm{d}x.$$

Thus, by a change of variables, we have, at least formally,

$$\det(\nabla T(x)) f_\mathcal{O}(T(x)) = f_\Omega(x) \quad x \in \Omega, \tag{5}$$

with $T(\Omega) \subset \mathcal{O}$. Thus in summary, we seek a mapping T that minimizes (4) with the constraint (5). One of the fundamental results in the theory of optimal transport (Brenier, 1991; Cuesta and Matrán, 1989; Rüschendorf and Rachev, 1990a,b) is that there exists a unique solution to this problem and that this optimal mapping is characterized as the gradient of some convex function u:

$$T(x) = \nabla u(x).$$

Hence, by substituting this relation into (5), we see that the problem reduces to a Monge–Ampère type PDE

$$f_\mathcal{O}(\nabla u(x)) \det D^2 u(x) = f_\Omega(x) \quad x \in \Omega \tag{6}$$

with the constraint $\nabla u(\bar{\Omega}) \subset \bar{\mathcal{O}}$. Thus we find that, with quadratic cost, the optimal mass transport problem reduces to a Monge–Ampère equation with transport boundary conditions.

1.2 Solution concepts for the Monge–Ampère equation

It is impossible to understand an unmotivated definition […]

Arnol'd (1998)

In order to properly analyze the numerical schemes that we present below, it is important to understand in which sense a function $u : \bar{\Omega} \to \mathbb{R}$ must satisfy the equation and boundary conditions in (1) to be a solution. It is not our intention here to give a survey of the PDE theory regarding the Monge–Ampère equation, and we refer the reader to Gutiérrez (2001), Figalli (2017), and Bakelman (1994) for an in-depth presentation.

1.2.1 Classical solutions

The first definition of a solution to (1) is that of a classical solution. Essentially we require that (1) holds at every point of $\bar{\Omega}$.

Definition 1 (classical solution)**.**
A function $u \in C^2(\Omega) \cap C(\bar{\Omega})$ is called a classical solution of (1) if these identities hold for every $x \in \bar{\Omega}$.

Notice that this necessarily implies that the right-hand side $f : \Omega \to \mathbb{R}$ is continuous. Regarding the existence of classical solutions we have the following result; see Figalli (2017, Section 3.1) for a detailed presentation.
Theorem 1 (existence of classical solutions)**.**
Let $\alpha \in (0, 1)$. Assume that Ω is a bounded and uniformly convex domain, whose boundary is of class C^3, $f \in C^\alpha(\bar{\Omega})$ with $f \geq f_0 > 0$, and $g \in C^3(\partial\Omega)$. Then problem (1) has a unique solution $u \in C^{2,\alpha}(\bar{\Omega})$.

It is important to notice that classical solutions may not always exist, see for instance the counterexample given in Figalli (2017, Section 3.2). This motivates us to introduce weaker notions of solutions.

1.2.2 Viscosity solutions

The Monge–Ampère operator $w \mapsto \det D^2 w$ is a fully nonlinear second order operator, that is it depends nonlinearly on the highest (in this case second) order derivatives that appear in the expression. For this reason, the theory regarding fully nonlinear operators can guide us to develop a notion of solution (viscosity solution) that is weaker than classical. We refer the reader to Gilbarg and Trudinger (2001, Chapter 17), Caffarelli and Cabré (1995), Crandall et al. (1992), and Neilan et al. (2017, Section 2) for additional details.

We begin with a definition that encodes the type of admissible nonlinearities that will allow for the development of the theory of viscosity solutions. Here and in what follows we denote by $\mathbb{S}^d$ the collection of symmetric $d \times d$ matrices. The set $\mathbb{S}^d$ is endowed with a partial order: if $M, N \in \mathbb{S}^d$ then we say that $M \leq N$ if $\boldsymbol{v} \cdot M\boldsymbol{v} \leq \boldsymbol{v} \cdot N\boldsymbol{v}$ for every $\boldsymbol{v} \in \mathbb{R}^d$.
Definition 2 (elliptic operator)**.**
Let $F : \bar{\Omega} \times \mathbb{R} \times \mathbb{S}^d \to \mathbb{R}$ be locally bounded. We say that F is *elliptic* if it satisfies the following *monotonicity condition*: For all $x \in \bar{\Omega}$, $r, s \in \mathbb{R}$ and $M, N \in \mathbb{S}^d$ with $r \geq s$ and $M \leq N$ then

$$F(x, r, M) \leq F(x, s, N).$$

Moreover, we say F is *uniformly elliptic* if for all $r, s \in \mathbb{R}$ and $M \in \mathbb{S}^d$ with $r \geq s$ we have

$$F(x, r, M) \leq F(x, s, M),$$

and, in addition, there are constants $0 < \lambda \leq \Lambda$ such that for all $M \in \mathbb{S}^d$ we have

$$\lambda \| N \|_2 \leq F(x, r, M + N) - F(x, s, M) \leq \Lambda \| N \|_2, \quad \forall N \geq 0.$$

Letting $F:\bar{\Omega}\times\mathbb{R}\times\mathbb{S}^d\to\mathbb{R}$ be an elliptic operator as defined above, we consider the fully nonlinear elliptic problem

$$F(x,u(x),D^2u(x))=0 \quad \text{in}\,\bar{\Omega}. \tag{7}$$

To be able to properly describe the notion of viscosity solutions we need to recall the following.

Definition 3 (upper and lower semicontinuous envelopes)**.**
Let $w:\bar{\Omega}\to\mathbb{R}$. By $w^\star\in USC(\bar{\Omega})$ and $w_\star\in LSC(\bar{\Omega})$, we denote the upper and lower semicontinuous envelopes, respectively, of the function w. In other words

$$w^\star(x)=\limsup_{y\to x} w(x), \qquad w_\star(x)=\liminf_{y\to x} w(x).$$

Finally, by $USC(\bar{\Omega})$ and $LSC(\bar{\Omega})$ we denote, respectively, the sets of upper and lower semicontinuous functions.

We are now ready to introduce the notion of viscosity solution.

Definition 4 (viscosity solution)**.**
Let F be elliptic in the sense of Definition 2. We say that the locally bounded function $u:\bar{\Omega}\to\mathbb{R}$ is:

1. A *viscosity subsolution* of (7) if whenever $x_0\in\bar{\Omega}$, $\varphi\in C^2(\bar{\Omega})$ and $u^\star-\varphi$ has a local maximum at x_0 we have that

$$F_\star(x_0,\varphi(x_0),D^2\varphi(x_0))\geq 0.$$

2. A *viscosity supersolution* of (7) if whenever $x_0\in\bar{\Omega}$, $\varphi\in C^2(\bar{\Omega})$ and $u_\star-\varphi$ has a local minimum at x_0 we have that

$$F^\star(x_0,\varphi(x_0),D^2\varphi(x_0))\leq 0.$$

3. A *viscosity solution* if it is a sub- and supersolution.

The condition "$u^\star-\varphi$ has a local maximum at x_0" is usually phrased as "*φ touches the graph of u from above at x_0*". The reader is encouraged to draw a picture to see why these two have the same meaning. Similarly, "$u_\star-\varphi$ has a local minimum at x_0" is: "*φ touches the graph of u from below at x_0*".

One of the main technical tools in asserting existence and uniqueness of viscosity solutions is a comparison principle.

Definition 5 (comparison principle)**.**
We say that problem (7) satisfies a comparison principle if whenever $\overline{w}\in USC(\bar{\Omega})$ and $\underline{w}\in LSC(\bar{\Omega})$ are sub- and supersolutions, respectively, then we must have

$$\overline{w}\leq\underline{w}.$$

Notice now that if we define

$$F_{MA}(x,r,M) = \begin{cases} \det M - f(x), & x \in \Omega, \\ g(x) - r, & x \in \partial\Omega, \end{cases} \tag{8}$$

this operator satisfies the monotonicity conditions given in Definition 2 *only* if we restrict the third argument to the set of positive semidefinite matrices which we denote by $\mathbb{S}^d_+$. Consequently, we need to restrict the class of admissible functions, that define a viscosity solution to (1) to the set of convex functions.

Definition 6 (viscosity solution)**.**
Let $u \in C(\bar{\Omega})$ be a convex function. We say that u is:

1. A *viscosity subsolution* of (1) on the set of convex functions if $u \leq g$ on $\partial\Omega$ and, whenever $x_0 \in \Omega$, $\varphi \in C^2(\Omega)$, and $u - \varphi$ has a local maximum at x_0 we have that
$$\det D^2\varphi(x_0) \geq f(x_0).$$
2. A *viscosity supersolution* of (1) on the set of convex functions if $u \geq g$ on $\partial\Omega$ and, whenever $x_0 \in \Omega$, $\varphi \in C^2(\Omega)$ is convex, and $u - \varphi$ has a local minimum at x_0 we have that
$$\det D^2\varphi(x_0) \leq f(x_0).$$
3. A *viscosity solution* if it is a sub- and supersolution on the set of convex functions.

The reader may wonder why these definitions are asymmetric. The concept of supersolution requires convexity of the test functions, whereas subsolutions do not. This is due to the fact that, as noted in Gutiérrez (2001, Remark 1.3.2), if u is convex and $u - \varphi$ has a local maximum at x_0, then φ is (locally) convex.

The existence and uniqueness of viscosity solutions will be a consequence of Theorems 2 and 3. Here we mention a remarkable property of viscosity solutions, namely their stability. The following result can be found, for instance, in Nochetto et al. (2019a, Lemma 5.3).

Proposition 1 (continuous dependence)**.**
Let $f_1, f_2 \in C(\bar{\Omega})$ with $f_1, f_2 \geq 0$ and $g_1, g_2 \in C(\partial\Omega)$ and denote by $u_1, u_2 \in C(\bar{\Omega})$ the corresponding convex viscosity solutions to (1). Then we have

$$\| u_1 - u_2 \|_{L^\infty(\Omega)} \leq C \, \| f_1 - f_2 \|^{1/d}_{L^\infty(\Omega)} + \| g_1 - g_2 \|_{L^\infty(\partial\Omega)}.$$

In addition, if $f_1 \geq f_2 \geq 0$ and $g_1 \leq g_2$ we have that $u_1 \leq u_2$.

Finally we comment that viscosity solutions can be approximated by classical ones over larger, but smooth, domains; see Nochetto et al. (2019a, Lemma 5.4).

Proposition 2 (smooth approximation)**.**
Let Ω be uniformly convex, $f, g \in C(\bar{\Omega})$ with $f \geq 0$, and u the convex viscosity solution to (1). There exists:

1. *A decreasing (in the sense of inclusion) sequence of uniformly convex smooth domains Ω_n such that*

$$\mathrm{dist}_H(\Omega_n, \Omega) \to 0, \quad n \to \infty,$$

where by $dist_H(A, B)$ we mean the d-dimensional Hausdorff distance between the sets A and B.

2. *A decreasing sequence of smooth functions $f_n : \bar{\Omega}_n \to \mathbb{R}$ with $f_n > 0$ such that*

$$\|f_n - f\|_{L^\infty(\Omega)} \to 0, \quad n \to \infty.$$

3. *A sequence of smooth functions $g_n : \bar{\Omega}_n \to \mathbb{R}$ such that*

$$\| g_n - g\|_{L^\infty(\Omega)} \to 0, \quad n \to \infty.$$

Moreover, if $u_n \in C(\bar{\Omega}_n)$ denotes the convex viscosity solution to (1) over the domain Ω_n and with data f_n and g_n, then

$$\| u_n - u\|_{L^\infty(\Omega)} \to 0, \quad n \to \infty.$$

1.2.3 Alexandrov solutions

Besides the concept of solution in the viscosity sense, another type of weak solution to the Monge–Ampère equation is the Alexandrov solution, which is based on a geometric interpretation. To motivate it, let $w \in C^2(\Omega)$ be convex so that the gradient map $\nabla w : \Omega \to \mathbb{R}^d$ is well defined and monotone. In this case, an interesting observation is that $\det D^2 w$ is actually the determinant of the Jacobian of the gradient map. Therefore, for any open (or Borel) subset $E \subset \Omega$, we have

$$\int_E \det D^2 w(x) \mathrm{d}x = \int_{\nabla w(E)} \mathrm{d}y = |\nabla w(E)|,$$

where $|\cdot|$ denotes the d-dimensional Lebesgue measure.

What is remarkable about this simple observation is that to make sense of $\det D^2 u$, we only require $\nabla w(E)$ to be well defined for any Borel set E. This enables us to make sense of the previous identity even if $w \notin C^2(\Omega)$. To define the weak (Alexandrov) solution, we first introduce the subdifferential of a convex function.

Definition 7 (subdifferential)**.**
Let Ω be convex and $w : \Omega \to \mathbb{R}$ be a convex function. The subdifferential of w at point $x \in \Omega$ is the set

$$\partial w(x) := \{\boldsymbol{p} \in \mathbb{R}^d, w(x) + \boldsymbol{p} \cdot (y - x) \leq w(y) \quad \forall y \in \Omega\}.$$

For any Borel set $E \subset \Omega$, we define

$$\partial w(E) = \cup_{x \in E} \partial w(x).$$

In other words, the subdifferential is the collection of slopes of all affine functions that touch the graph of w at $(x, w(x))$ and bound the graph from below. From this observation, it is easy to see that if w is strictly convex and smooth, then $\partial w(x) = \{\nabla w(x)\}$. Here we give an example of subdifferential of a convex (but not strictly convex) function.

Example 1 (subdifferential)**.**
Let $\Omega = B_1(0) \subset \mathbb{R}^2$ and

$$w(x) = |x|.$$

Then at the origin $x = 0$, we note that

$$w(0) + \boldsymbol{p} \cdot y \leq w(y) \quad \forall y \in \Omega$$

provided that the norm of the vector $|\boldsymbol{p}| \leq 1$. Hence, by definition, the subdifferential of w at $x = 0$ is the closed unit ball centred at 0, i.e.

$$\partial w(0) = \overline{B_1(0)}.$$

At any other point $x \in \Omega$, since the function w is differentiable, we note that the inequality

$$w(x) + \boldsymbol{p} \cdot (y - x) \leq w(y) \quad \forall y \in \Omega$$

holds if and only if $\boldsymbol{p} = \nabla w(x)$. Hence, for all $x \in \Omega \backslash \{0\}$,

$$\partial w(x) = \{\nabla w(x)\}.$$

With this motivation at hand we can introduce the so-called Monge–Ampère measure, which will be essential in defining Alexandrov solutions.

Definition 8 (Monge–Ampère measure)**.**
Let $\Omega \subset \mathbb{R}^d$ be convex and $w : \Omega \to \mathbb{R}$ be a convex function. The *Monge–Ampère measure* associated to w is

$$\mu_w(E) = |\partial w(E)|.$$

It can be shown, see Figalli (2017, Theorem 2.3) that this is indeed a locally finite Borel measure on Ω. With this, we are ready to define Alexandrov solutions.

Definition 9 (Alexandrov solution)**.**
Let f be a Borel measure defined in Ω. A convex function $u \in C(\bar{\Omega})$ is an *Alexandrov solution* to the Monge–Ampère equation (1) if $u = g$ on $\partial\Omega$ and $\mu_u = f$, that is,

$$|\partial u(E)| = f(E). \tag{9}$$

for all Borel sets $E \subset \Omega$.

To illustrate the definition of the Alexandrov solution, we consider Example 1. Let $E \subset \Omega$ be Borel, if the set contains the origin, we have the subdifferential

$$\partial u(E) = \cup_{x \in E} \partial u(x) = \overline{B_1(0)},$$

which yields

$$|\partial u(E)| = |B_1(0)| = \pi \quad \text{if } x \in E.$$

On the other hand, if the set does not contain the origin, then the subdifferential

$$\partial u(E) = \cup_{x \in E} \{\nabla u(x)\} \subset \partial B_1(0)$$

Hence, we get $|\partial u(E)| = 0$ if $0 \notin E$. Finally, we conclude that u is an Alexandrov solution of Monge–Ampère equation

$$\det D^2 u(x) = \pi \delta_{\{x=0\}},$$

where $\delta_{\{x=0\}}$ is the Dirac measure at the origin. It is worth mentioning that u is not a viscosity solution because the right-hand side is not a (continuous) function. Also note that the continuity of the source term f is no longer required for (9) to be well defined.

The existence and uniqueness of Alexandrov solutions is summarized in the next theorem, see Gutiérrez (2001, Theorem 1.6.2) and Figalli (2017, Theorem 2.14).

Theorem 2 (existence and uniqueness)**.**
Let $\Omega \subset \mathbb{R}^d$ be a strictly convex domain, let $g \in C(\partial\Omega)$ and f be a nonnegative Borel measure on Ω with $f(\Omega) < \infty$. Then there exists a unique convex function $u \in C(\bar{\Omega})$ that is a solution of (1) in the sense of Definition 9.

An important property of Alexandrov solutions is their stability with respect to weak convergence. We refer the reader to Gutiérrez (2001, Lemma 1.2.3) for a proof of the following result.

Lemma 1 (weak convergence)**.**
Let $\{w_k\}_{k=1}^{\infty}, w$ be convex functions on Ω and assume that, as $k \to \infty$, we have $w_k \to w$ uniformly over compact subsets of Ω. Then, the associated Monge–Ampère measures μ_{w_k} tend to μ_w weakly, that is,

$$\int_{\Omega} \phi(x) \mathrm{d}\mu_{w_k}(x) \to \int_{\Omega} \phi(x) \mathrm{d}\mu_{w}(x),$$

for every ϕ continuous with compact support in Ω.

The relation between viscosity and Alexandrov solutions is given in the following result (Gutiérrez, 2001, Propositions 1.3.4 and 1.7.1). Notice that this result not only shows, as we have already pointed out, that the notion of Alexandrov solution is strictly weaker than that of viscosity solutions but, on the basis of Theorem 2, shows existence and uniqueness of viscosity solutions.

Theorem 3 (equivalence)**.**
Let $u \in C(\bar{\Omega})$ be an Alexandrov solution of (1). If $f \in C(\Omega)$, then u is also a viscosity solution in the sense of Definition 6. Conversely, if u is a viscosity solution of (1) and $f \in C(\bar{\Omega})$ with $f > 0$, then u is an Alexandrov solution.

Since it will be useful in the sequel, we introduce here the convex envelope of a function, which is the largest convex function that is bounded above by the given one.

Definition 10 (convex envelope)**.**
Let $\Omega \subset \mathbb{R}^d$ be convex and $w : \bar{\Omega} \to \mathbb{R}$. The *convex envelope* of w, denoted by Γw, is the largest convex function whose graph lies below the graph of w. It can be computed by

$$\Gamma w(x) = \sup\{L(x) : L \text{ affine function and } L(y) \leq w(y) \quad \forall y \in \bar{\Omega}\}.$$

To conclude our preliminary discussion we recall the Brunn–Minkowski inequality, a celebrated result in convex geometry. Given two compact sets A, B of $\mathbb{R}^d$, we define their *Minkowski sum*

$$A + B := \{v + w \in \mathbb{R}^d : v \in A \text{ and } w \in B\}. \tag{10}$$

The Brunn–Minkowski inequality relates the Lebesgue measures of compact subsets A, B of Euclidean space $\mathbb{R}^d$ with that of their Minkowski sum $A + B$.

Lemma 2 (Brunn–Minkowski inequality)**.**
Let A and B be two nonempty compact subsets of $\mathbb{R}^d$ for $d \geq 1$. Then the following inequality holds:

$$|A + B|^{1/d} \geq |A|^{1/d} + |B|^{1/d}.$$

2 Wide stencil finite differences

> *Problems involving the classical linear partial differential equations of mathematical physics can be reduced to algebraic ones of a very much simpler structure by replacing the differentials by difference quotients on some (say rectilinear) mesh.*
>
> Courant et al. (1967)

In this section we will study finite difference schemes that aim to approximate the viscosity solution, in the sense of Definition 6, of (1).

2.1 A general framework for approximation schemes

Let us describe a general framework under which convergence of approximation schemes can be shown. Let $F:\bar{\Omega}\times\mathbb{R}\times\mathbb{S}^d\to\mathbb{R}$ be elliptic in the sense of Definition 2 and assume we wish to approximate the viscosity solution to (7). To do so, we introduce a family of approximation schemes, which are described by the collection of maps $\{F_\varepsilon\}_{\varepsilon>0}$, where $F_\varepsilon:\bar{\Omega}\times\mathbb{R}\times B(\bar{\Omega})\to\mathbb{R}$, and $B(\bar{\Omega})$ denotes the space of bounded functions on $\bar{\Omega}$. The parameter ε can be understood as a discretization parameter. With this family at hand, we seek for $u_\varepsilon\in B(\bar{\Omega})$ such that

$$F_\varepsilon(x,u_\varepsilon(x),u_\varepsilon)=0,\quad \text{in}\,\bar{\Omega}. \tag{11}$$

We assume that the approximation schemes satisfy the following assumptions:

1. *Monotonicity*: For all $\varepsilon>0$, $x\in\bar{\Omega}$, $t\in\mathbb{R}$, and $u,v\in B(\bar{\Omega})$ such that $u\geq v$ we have that
$$F_\varepsilon(x,t,u)\geq F_\varepsilon(x,t,v). \tag{12}$$
2. *Stability*: There is $\varepsilon_0>0$ such that if $\varepsilon<\varepsilon_0$, the scheme (11) has a unique solution and there is a constant, independent of ε, such that
$$\|u_\varepsilon\|_{L^\infty(\Omega)}\leq C. \tag{13}$$
3. *Consistency*: For all $x_0\in\bar{\Omega}$ and $\varphi\in C^2(\bar{\Omega})$ we have
$$\limsup_{\varepsilon\downarrow 0,\,y\to x_0,\,\xi\to 0} F_\varepsilon(y,\varphi(y)+\xi,\varphi+\xi)\leq F_\star(x_0,\varphi(x_0),D^2\varphi(x_0)) \tag{14a}$$
$$\liminf_{\varepsilon\downarrow 0,\,y\to x_0,\,\xi\to 0} F_\varepsilon(y,\varphi(y)+\xi,\varphi+\xi)\geq F^\star(x_0,\varphi(x_0),D^2\varphi(x_0)). \tag{14b}$$

The main convergence result in this framework is the following; see Barles and Souganidis (1991, Theorem 2.1).

Theorem 4 (Barles–Souganidis)**.**
Assume that the family of approximation schemes (11) *is monotone, stable and consistent, in the sense of* (12), (13), *and* (14), *respectively. Assume, in addition, that problem* (7) *has a comparison principle in the sense of Definition* 5. *Then, as* $\varepsilon\downarrow 0$, *the functions* u_ε, *solution of* (11) *converge locally uniformly to* u, *solution of* (7).

Proof. Define $\overline{u},\underline{u}\in B(\bar{\Omega})$ by

$$\overline{u}(x)=\limsup_{y\to x,\,\varepsilon\downarrow 0} u_\varepsilon(y),\qquad \underline{u}(x)=\liminf_{y\to x,\,\varepsilon\downarrow 0} u_\varepsilon(y).$$

Notice that, by stability, we obtain that these functions are well defined and bounded. In addition, we have that $\overline{u},\underline{u}$ are upper and lower semicontinuous, respectively.

The idea now is to show that $\overline{u}$ is a subsolution and $\underline{u}$ is a supersolution of (7), for if that is the case we can invoke the comparison principle to see that $\overline{u} \leq \underline{u}$, and so that these must coincide with the viscosity solution of (7). This, in turn, implies the local uniform convergence of u_ε to u.

Let us then show that $\overline{u}$ is a subsolution. Let $\varphi \in C^2(\bar{\Omega})$ and assume that $\overline{u} - \varphi$ has a local maximum at $x_0 \in \bar{\Omega}$ with $\overline{u}(x_0) = \varphi(x_0)$. It can be shown then that there are sequences $\{\varepsilon_n\}_{n=1}^{\infty} \subset \mathbb{R}^+$ and $\{y_n\}_{n=1}^{\infty} \subset \bar{\Omega}$ such that $\varepsilon_n \downarrow 0$, $y_n \to x_0$, $u_{\varepsilon_n}(y_n) \to \overline{u}(x_0)$ and the sequence of functions $u_{\varepsilon_n} - \varphi$ attains its maximum at y_n.

Notice now that, upon denoting $\xi_n = u_{\varepsilon_n}(y_n) - \varphi(y_n)$, we get that $\xi_n \to 0$ and $u_{\varepsilon_n}(x) - \varphi(x) \leq \xi_n$ locally. Monotonicity then implies that

$$\begin{aligned} 0 = F_{\varepsilon_n}(y_n, u_{\varepsilon_n}(y_n), u_{\varepsilon_n}) &= F_{\varepsilon_n}(y_n, \varphi(y_n) + \xi_n, \varphi + (u_{\varepsilon_n} - \varphi)) \\ &\leq F_{\varepsilon_n}(y_n, \varphi(y_n) + \xi_n, \varphi + \xi_n), \end{aligned}$$

which by the consistency condition (14a) yields

$$F_\star(x_0, \varphi(x_0), D^2\varphi(x_0)) \geq 0,$$

so that $\overline{u}$ is a subsolution. □

Remark 1 (limitations).
We must remark that, although Theorem 4 seems sufficiently general:

1. It only provides sufficient conditions for convergence. There is no guideline towards the construction of monotone, consistent and stable finite difference schemes.
2. This result, as is, *cannot* be applied to approximate viscosity solutions of the Monge–Ampère equation (1) directly. This is because, as pointed out in Section 1.2.2, the Monge–Ampère operator is only elliptic over $\bar{\Omega} \times \mathbb{R} \times \mathbb{S}^d_+$.
3. The existence of a comparison principle in the sense of Definition 5 is assumed. Notice that, in Jensen and Smears (2018, Proposition 2.1) it is shown that, for a reformulation of the Monge–Ampère problem as a Hamilton Jacobi Bellman equation (which will be discussed in Section 2.8.1), if $f \equiv 0$, there cannot be a comparison principle for this problem. In other words, this is a highly nontrivial assumption.

■

Although not applicable to the Monge–Ampère equation (1), one of the messages of Theorem 4 is that monotonicity of a numerical scheme is a highly desirable property. Thus, it is necessary to explore how to construct monotone approximation schemes. In the context of finite difference schemes it was realized as early as in Motzkin and Wasow (1953) that, even for linear problems, monotonicity of a numerical scheme requires the use of wide stencils, which is rather problematic at points near the boundary. We refer the reader to Neilan et al. (2017, Section 3.2) for more details, and to Mirebeau (2016) for the construction of minimal stencils in two dimensions. For this reason, in the remaining of this section, we will consider wide stencil finite difference schemes to approximate the viscosity solution of (1).

2.2 A variational characterization of the determinant

Let us provide a variational characterization of the determinant that will motivate most of the constructions which will come below. This was originally shown in Froese and Oberman (2011a, Lemma 2).

Lemma 3 (characterization of the determinant)**.**
Let A be a symmetric positive definite $d \times d$ matrix and let

$$\mathcal{V} = \left\{ \{\boldsymbol{w}_i\}_{i=1}^d \subset \mathbb{R}^d : \boldsymbol{w}_i \cdot \boldsymbol{w}_j = \delta_{i,j} \right\},$$

be the set of all orthonormal bases of $\mathbb{R}^d$. Then we have that

$$\det A = \min_{\{\boldsymbol{w}_i\}_{i=1}^d \in \mathcal{V}} \prod_{i=1}^d \boldsymbol{w}_i \cdot A\boldsymbol{w}_i.$$

Proof. To shorten notation, let $M = \min_{\{\boldsymbol{w}_i\}_{i=1}^d \in \mathcal{V}} \prod_{i=1}^d \boldsymbol{w}_i \cdot A\boldsymbol{w}_i$. Then let $\{\boldsymbol{v}_j\}_{j=1}^d$ be an orthonormal set of eigenvectors of A so that

$$\det A = \prod_{i=1}^d \boldsymbol{v}_i \cdot A\boldsymbol{v}_i \geq M.$$

On the other hand, for $\{\boldsymbol{w}_i\}_{i=1}^d \in \mathcal{V}$, we can represent them in the basis of eigenvectors $\boldsymbol{w}_i = \sum_{k=1}^d (\boldsymbol{w}_i \cdot \boldsymbol{v}_k)\boldsymbol{v}_k$. We have

$$\begin{aligned} -\log \prod_{i=1}^d \boldsymbol{w}_i \cdot A\boldsymbol{w}_i &= -\sum_{i=1}^d \log (\boldsymbol{w}_i \cdot A\boldsymbol{w}_i) \\ &= -\sum_{i=1}^d \log \left(\sum_{m=1}^d (\boldsymbol{w}_i \cdot \boldsymbol{v}_m)\boldsymbol{v}_m \cdot \sum_{k=1}^d (\boldsymbol{w}_i \cdot \boldsymbol{v}_k) A\boldsymbol{v}_k \right) \\ &= -\sum_{i=1}^d \log \left(\sum_{k=1}^d \lambda_k (\boldsymbol{w}_i \cdot \boldsymbol{v}_k)^2 \right), \end{aligned}$$

where $\sigma(A) = \{\lambda_k\}_{k=1}^d$ is the spectrum of A. Since $|\boldsymbol{w}_i| = 1$ the term $\sum_{k=1}^d \lambda_k (\boldsymbol{w}_i \cdot \boldsymbol{v}_k)^2$ is a convex combination of the elements of $\sigma(A)$. Owing to the convexity of $t \mapsto -\log t$ we can apply Jensen's inequality to obtain that

$$-\log \prod_{i=1}^d \boldsymbol{w}_i \cdot A\boldsymbol{w}_i \leq -\sum_{k=1}^d \log \lambda_k \sum_{i=1}^d (\boldsymbol{w}_i \cdot \boldsymbol{v}_k)^2 = -\sum_{k=1}^d \log \lambda_k = -\log \prod_{i=1}^d \lambda_i.$$

As the function $t \mapsto -\log t$ is decreasing, we conclude that

$$\det A \leq \prod_{i=1}^d \boldsymbol{w}_i \cdot A\boldsymbol{w}_i,$$

which since $\{\boldsymbol{w}_i\}_{i=1}^d \in \mathcal{V}$ was arbitrary implies $\det A \leq M$ and this concludes the proof. □

The previous result allows us to conclude that, if $\varphi \in C^2(\Omega)$ is convex, we can express the determinant of its Hessian at a point in terms of second directional derivatives, that is, if $x_0 \in \Omega$ we have

$$\det D^2\varphi(x_0) = \min_{\{\boldsymbol{w}_i\}_{i=1}^d \in \mathcal{V}} \prod_{i=1}^d \boldsymbol{w}_i \cdot D^2\varphi(x_0)\boldsymbol{w}_i = \min_{\{\boldsymbol{w}_i\}_{i=1}^d \in \mathcal{V}} \prod_{i=1}^d \frac{\partial^2 \varphi}{\partial \boldsymbol{w}_i^2}(x_0).$$

Recall, in addition, that a solution to (1) must be convex. To enforce convexity we then introduce the following operator

$$\mathrm{MA}[\varphi](x_0) = \min_{\{\boldsymbol{w}_i\}_{i=1}^d \in \mathcal{V}} \left[\prod_{i=1}^d \left(\frac{\partial^2 \varphi}{\partial \boldsymbol{w}_i^2}(x_0) \right)^+ - \sum_{i=1}^d \left(\frac{\partial^2 \varphi}{\partial \boldsymbol{w}_i^2}(x_0) \right)^- \right], \tag{15}$$

where $x^+ = \max\{x, 0\}$ and $x^- = (-x)^+$ denote the positive and negative parts of x, respectively. Notice that, if $\varphi \in C^2(\bar{\Omega})$ is convex, $\mathrm{MA}[\varphi] = \det D^2\varphi$. The idea behind (15) is that, if $D^2\varphi(x_0)$ has a negative eigenvalue, then there is $V \in \mathcal{V}$ and $\boldsymbol{w} \in V$ for which $\boldsymbol{w} \cdot D^2\varphi(x_0)\boldsymbol{w} < 0$. Thus,

$$\mathrm{MA}[\varphi](x_0) \leq 0 - (\boldsymbol{w} \cdot D^2\varphi(x_0)\boldsymbol{w})^- < 0.$$

Consequently, φ cannot be a solution to (1) since, at x_0 we have

$$\det D^2\varphi(x_0) = f(x_0) \geq 0.$$

These ideas are made rigorous in Nochetto et al. (2019a, Lemma 5.6).

Proposition 3 (equivalence of operators)**.**
Let $f \in C(\Omega)$ with $f \geq 0$. The function $u \in C(\bar{\Omega})$ is a convex viscosity solution of (1) in the sense of Definition 6 if and only if it is a viscosity solution, in the sense of Definition 4, of the following problem

$$F_{vMA}(x, u(x), D^2u(x)) = 0 \tag{16}$$

with

$$F_{vMA}(x, u(x), D^2u(x)) = \begin{cases} \mathrm{MA}[u](x) - f(x), & x \in \Omega, \\ g(x) - u(x), & x \in \partial\Omega. \end{cases}$$

One of the advantages of formulation (16) is that it has a comparison principle.

Proposition 4 (comparison principle for the F_{vMA} operator)**.**
The operator F_{vMA}, defined in (16) has a comparison principle in the sense of Definition 5.

Proof. It follows from the fact that the operator F_{vMA} satisfies the structural assumptions given, for instance, in Crandall et al. (1992, Theorem 3.3). □

The characterization of the determinant given in Lemma 3 will be the basis of many of the wide stencil schemes we will describe below.

2.3 Wide stencil finite difference schemes

Let us describe the first class of methods that exploit the characterization described in Lemma 3 via the operator introduced in (15) as originally proposed in Froese and Oberman (2011a). Let $h > 0$ be a (spatial) discretization parameter and assume that, up to a linear change of variables, our domain Ω is discretized on a Cartesian grid. In other words, we let

$$\bar{\Omega}_h = \bar{\Omega} \cap \mathbb{Z}_h^d, \quad \mathbb{Z}_h^d = \{h\boldsymbol{e} : \boldsymbol{e} \in \mathbb{Z}^d\}, \quad \partial\Omega_h = \partial\Omega \cap \mathbb{Z}_h^d, \quad \Omega_h = \bar{\Omega}_h \setminus \partial\Omega_h.$$

We set X_h as the space of *grid functions*, that is the collection of functions $w_h : \bar{\Omega}_h \to \mathbb{R}$.

Given $\boldsymbol{e} \in \mathbb{Z}^d$ we call the point $x_h \in \Omega_h$ *interior* with respect to $\boldsymbol{e}$ if $x_h \pm h\boldsymbol{e} \in \bar{\Omega}_h$. We will also say that a point is interior with respect to a subset of $S \subset \mathbb{Z}^d$ if it is interior with respect to all elements of S.

Given $\boldsymbol{e} \in \mathbb{Z}^d$ and an interior point x_h, we define the *second difference* in the direction $\boldsymbol{e}$ to be the operator

$$\Delta_{\boldsymbol{e}} w_h(x_h) = \frac{1}{|e|^2 h^2}(w_h(x_h + h\boldsymbol{e}) - 2w_h(x_h) + w_h(x_h - h\boldsymbol{e})). \tag{17}$$

When x_h is not interior with respect to $\boldsymbol{e}$, it essentially means that x_h is close to $\partial\Omega$. Owing to the convexity of Ω, there are unique $\rho_\pm \in (0, 1]$ such that $x_h \pm \rho_\pm h\boldsymbol{e} \in \partial\Omega$. Thus, we can use the boundary condition (1b) to extend this definition as

$$\Delta_e w_h(x_h) = \frac{2}{(\rho_+ + \rho_-)|\boldsymbol{e}|^2 h^2}\left(\frac{\tilde{g}(x_h + \rho_+ h\boldsymbol{e}) - w_h(x_h)}{\rho_+} - \frac{w_h(x_h) - \tilde{g}(x_h + \rho_- h\boldsymbol{e})}{\rho_-}\right), \tag{18}$$

where $\tilde{g}$ is either the boundary condition, or an interpolant of w_h based on neighbouring nodes. With these notions at hand, we would like to define the discretization of the operator MA $[\cdot]$, introduced in (15), as

$$\mathrm{MA}_h^{\mathrm{WS}}[w_h](x_h) = \min_{\{\boldsymbol{w}_i\}_{i=1}^d \in \mathcal{V}} \prod_{i=1}^d (\Delta_{\boldsymbol{w}_i} w_h(x_h))^+.$$

Notice, however, that the given expressions may not be defined for all $\mathcal{V}$, as the points $x_h \pm h\boldsymbol{w}_i$ may not belong to $\bar{\Omega}_h$. Even if they did, it may be very computationally expensive to compute these directional differences at all the nodes. For these reasons, we also need to introduce a discretization of $\mathcal{V}$. To this end we introduce a finite subset $\mathcal{G}_\theta \subset (\mathbb{Z}^d)^d$ such that, if $\{\boldsymbol{\nu}_i\}_{i=1}^d \in \mathcal{G}_\theta$ then the vectors $\boldsymbol{\nu}_i$ are pairwise orthogonal. We call this the *directional* discretization of the Monge–Ampère operator and parametrize it by $\theta > 0$. Thus we define the operator

$$\mathrm{MA}^{\mathrm{WS}}_{h,\theta}[w_h](x_h) = \min_{\{\boldsymbol{\nu}_i\}_{i=1}^d \in \mathcal{G}_\theta} \prod_{i=1}^d (\Delta_{\boldsymbol{\nu}_i} w_h(x_h))^+. \tag{19}$$

With this notation at hand, we define the *wide stencil* finite difference approximation scheme of (1) as: Find $u_h \in X_h$ such that

$$\mathrm{MA}^{\mathrm{WS}}_{h,\theta}[u_h](x_h) = f(x_h), \quad \forall x_h \in \Omega_h, \tag{20a}$$

$$u_h(x_h) = g(x_h), \quad \forall x_h \in \partial\Omega_h. \tag{20b}$$

Remark 2 (variant).
We could have also introduced another wide stencil operator via

$$\mathrm{MA}^{\mathrm{WS}}_{h,\theta}[w_h](x_h) = \min_{\{\boldsymbol{\nu}_i\}_{i=1}^d \in \mathcal{G}_\theta} \left[\prod_{i=1}^d (\Delta_{\boldsymbol{\nu}_i} w_h(x_h))^+ - \sum_{i=1}^d (\Delta_{\boldsymbol{\nu}_i} w_h(x_h))^- \right],$$

see (15). ■

Remark 3 (a regularized version).
Notice that, owing to the presence of the min and max operator in the definition of (19), this operator is not differentiable. This may make it difficult to efficiently solve the ensuing nonlinear systems, since Newton methods are not directly applicable. One could, instead, use semismooth Newton methods (Hintermüller et al., 2002) since these operators are slant differentiable; see Hintermüller et al. (2002, Lemma 3.1). However, if we insist in dealing with smooth operators, Froese and Oberman (2011a, Section 3.5) introduces a regularized version of $\mathrm{MA}^{\mathrm{WS}}_{h,\theta}[\,\cdot\,]$ given by

$$\mathrm{MA}^{\mathrm{WS}}_{h,\theta,\delta}[w_h](x_h) = \min^{\delta}_{\{\boldsymbol{\nu}_i\}_{i=1}^d \in \mathcal{G}_\theta} \prod_{i=1}^d (\Delta_{\boldsymbol{\nu}_i} w_h(x_h))^{+,\delta},$$

where

$$\max^{\delta}\{x,y\} = \frac{1}{2}\left(x + y + \sqrt{(x-y)^2 + \delta^2}\right),$$

$$\min^{\delta}\{x,y\} = \frac{1}{2}\left(x + y - \sqrt{(x-y)^2 + \delta^2}\right),$$

$$\min^{\delta}\{x_1,\ldots,x_n\} = \min^{\delta}\{\min^{\delta}\{x_1,\ldots,x_{n-1}\},x_n\},$$

and $x^{+,\delta} = \max^{\delta}\{x,0\}$. The properties of this operator are similar to those of $\mathrm{MA}^{\mathrm{WS}}_{h,\theta}[\,\cdot\,]$. ■

Remark 4 (two dimensions).
Given $A \in \mathbb{S}^d$ we have the classical Rayleigh–Ritz relations

$$\lambda_m(A) = \min_{\boldsymbol{w}\in\mathbb{R}^d} \frac{\boldsymbol{w}\cdot A\boldsymbol{w}}{|\boldsymbol{w}|^2} = \min\sigma(A), \quad \lambda_M(A) = \max_{\boldsymbol{w}\in\mathbb{R}^d} \frac{\boldsymbol{w}\cdot A\boldsymbol{w}}{|\boldsymbol{w}|^2} = \max\sigma(A),$$

so that, if $d = 2$, we have that

$$\det A = \min_{\boldsymbol{w}\in\mathbb{R}^2} \frac{\boldsymbol{w}\cdot A\boldsymbol{w}}{|\boldsymbol{w}|^2} \max_{\boldsymbol{w}\in\mathbb{R}^2} \frac{\boldsymbol{w}\cdot A\boldsymbol{w}}{|\boldsymbol{w}|^2}.$$

This relation was used in Oberman (2008b) to introduce a two-dimensional scheme via

$$\mathrm{MA}^{\mathrm{WS,2d}}_{h,\theta}[w_h](x_h) = \min_{\boldsymbol{\nu}_i\in\{\boldsymbol{\nu}_j\}_{j=1}^d\in\mathcal{G}_\theta} \left(\Delta_{\boldsymbol{\nu}_i} w_h(x_h)\right)^+ \max_{\boldsymbol{\nu}_i\in\{\boldsymbol{\nu}_j\}_{j=1}^d\in\mathcal{G}_\theta} \left(\Delta_{\boldsymbol{\nu}_i} w_h(x_h)\right)^+.$$

Note that, although similar to (20), these operators are different. This was illustrated in Froese and Oberman (2011a, Section 3.4) with the following example: Let

$$w(x_1,x_2) = x_1^2 + x_2^2 + x_1^2x_2^2,$$

which is convex in a neighbourhood of the origin, and

$$\mathcal{G}_\theta = \left\{\left\{\begin{pmatrix}1\\0\end{pmatrix},\begin{pmatrix}0\\1\end{pmatrix}\right\},\left\{\begin{pmatrix}1\\1\end{pmatrix},\begin{pmatrix}-1\\1\end{pmatrix}\right\}\right\}.$$

Computing each of the operators over these directions yields

$$\mathrm{MA}^{\mathrm{WS,2d}}_{h,\theta}[w](0,0) = 4 + 2h^2, \quad \mathrm{MA}^{\mathrm{WS}}_{h,\theta}[w](0,0) = 4.$$

Notice however, that since both operators are consistent with order $\mathcal{O}(h^2)$ we have that, for a convex function v,

$$\left|\mathrm{MA}^{\mathrm{WS,2d}}_{h,\theta}[v](x_h) - \mathrm{MA}^{\mathrm{WS}}_{h,\theta}[v](x_h)\right| = \mathcal{O}(h^2), \quad \forall x_h.$$

■

The analysis of method (20) will be a particular case of the methods and analyses presented in Section 2.7. We just comment that, even for smooth solutions, wide stencils are required in this scheme to assert consistency. Let us illustrate this in a simple case where there is no boundary conditions and in two dimensions ($d = 2$). In other words, given $x_0 \in \Omega$ we assume that it is an interior point for any $\boldsymbol{e} \in \mathbb{Z}^2$. Let now $\varphi(x) = \frac{1}{2}x\cdot Mx$ be a convex quadratic, so that

$$\Delta_{\boldsymbol{e}}\varphi(x_0) = \frac{1}{|\boldsymbol{e}|^2}\boldsymbol{e}\cdot M\boldsymbol{e},$$

and therefore

$$\mathrm{MA}^{\mathrm{WS}}_{h,\theta}[\varphi](x_0) = \min_{\{\boldsymbol{\nu}_1,\boldsymbol{\nu}_2\}\in\mathcal{G}_\theta} \frac{1}{|\boldsymbol{\nu}_1|^2|\boldsymbol{\nu}_2|^2}(\boldsymbol{\nu}_1\cdot M\boldsymbol{\nu}_1)(\boldsymbol{\nu}_1\cdot M\boldsymbol{\nu}_2),$$

independently of x_0 and the mesh size. At this point we need to recall that there is $\{\boldsymbol{w}_1, \boldsymbol{w}_2\} \in \mathcal{V}$, namely the normalized eigenvectors of M, for which

$$\det D^2\varphi = \det M = (\boldsymbol{w}_1 \cdot M\boldsymbol{w}_1)(\boldsymbol{w}_2 \cdot M\boldsymbol{w}_2).$$

Notice finally, that once $\boldsymbol{w}_1$ is determined, $\boldsymbol{w}_2 = \boldsymbol{w}_1^{\perp}$ is obtained by a rotation. In conclusion, to assert consistency, given a $\boldsymbol{w} \in \mathbb{R}^2$ in the unit sphere, for every $\delta > 0$ we must be able to find $\boldsymbol{e} \subset \mathbb{Z}^2$ such that

$$\left| \boldsymbol{w} - \frac{1}{|\boldsymbol{e}|}\boldsymbol{e} \right| < \delta. \tag{21}$$

Indeed, if we denote by $\boldsymbol{e}_1$ the vector that satisfies this property with respect to $\boldsymbol{w}_1$, then $\boldsymbol{e}_2 = \boldsymbol{e}_1^{\perp}$ does so for $\boldsymbol{w}_2$. Let now $\boldsymbol{\nu}_i = \frac{1}{|\boldsymbol{e}_i|}\boldsymbol{e}_i$ for $i = 1, 2$. Then we have that

$$|\det M - (\boldsymbol{\nu}_1 \cdot M\boldsymbol{\nu}_1)(\boldsymbol{\nu}_2 \cdot M\boldsymbol{\nu}_2)| \leq C(\Lambda)\delta,$$

where $C(\Lambda)$ is a constant that depends polynomially on Λ, the maximal eigenvalue of M.

Notice that, since $\boldsymbol{e}_i \in \mathbb{Z}^2$, then $\boldsymbol{\nu}_i \in \mathbb{Q}^2$, so finding points that satisfy (21) is the problem of *rational approximation in the sphere*. While how to actually find such points is beyond our discussion here, what we are interested in is the size of $|\boldsymbol{e}|$, which would serve as an estimate of the stencil size that guarantees convergence. The following result is a specialization of Schmutz (2008, Lemma 2.1) to the two-dimensional case; we refer the reader to this reference a proof, its generalization to $d > 2$, and to the case of rational approximation orthogonal matrices which is of interest when finding elements of $\mathcal{G}_\theta$.

Proposition 5 (rational approximation)**.**
Let $\boldsymbol{w} \in \mathbb{R}^2$ be such that $|\boldsymbol{w}| = 1$. Then, for every $\delta > 0$, there exists $\boldsymbol{\nu} \in \mathbb{Q}^2$ such that $|\boldsymbol{\nu}| = 1$ and

$$|\boldsymbol{w} - \boldsymbol{\nu}| < \delta.$$

Moreover, if $\boldsymbol{\nu} = (p_1/q_1, p_2/q_2)^{\mathsf{T}}$ with $p_1, p_2 \in \mathbb{Z}$ and $q_1, q_2 \in \mathbb{N}$ then we have that

$$0 < q_i \leq \frac{64}{\delta^2}.$$

Now, for a given $\boldsymbol{w} \in \mathbb{R}^2$, let $\boldsymbol{\nu}$ be as in Proposition 5. This means that $\boldsymbol{e} = \mathrm{hcf}(q_1, q_2)\boldsymbol{\nu} \in \mathbb{Z}^2$ is the smallest vector parallel to $\boldsymbol{\nu}$ that satisfies (21) (here, hcf(q_1, q_2) denotes the highest common factor of q_1 and q_2). Consequently, we have that, generically

$$|\boldsymbol{e}| \leq C \;\; \mathrm{hcf}(q_1, q_2) \leq C \max\{q_1, q_2\} \leq \frac{C}{\delta^2}.$$

In conclusion, the size of the stencil must grow unboundedly if we restrict ourselves to Cartesian meshes.

2.4 Filtered schemes

The estimates on the stencil size of the previous section are rather pessimistic. This is because they are not assuming anything but convexity of the solution. On the other hand, say in the two-dimensional case ($d = 2$), a standard nine point stencil finite difference approximation can be proposed

$$\mathrm{MA}_h^{\mathrm{FD}}[w_h](x_h) = \Delta_{(1,0)} w_h(x_h) \Delta_{(0,1)} w_h(x_h) - \left(\mathring{\Delta}_{(1,1)} w_h(x_h)\right)^2, \tag{22}$$

where, if $z_h = (x_1, x_2)^{\mathsf{T}}$, then

$$\mathring{\Delta}_{(1,1)} w_h(z_h) = \frac{1}{2h}\left(\frac{w_h(x_1+h, x_2+h) - w_h(x_1-h, x_2+h)}{2h} - \frac{w_h(x_1+h, x_2-h) - w_h(x_1-h, x_2-h)}{2h}\right).$$

This formula easily extends to higher dimensions.

It is not difficult to see that $\mathrm{MA}_h^{\mathrm{FD}}[\,\cdot\,]$ has second-order consistency, even for nonconvex functions. However, it is not monotone, even if one forgets about boundary conditions. Thus, it does not perform well when used to discretize problems that have singular solutions.

Froese and Oberman (2011b) takes advantage of the simplicity of (22) and the robustness of a wide stencil scheme by proposing a *hybrid* scheme. Locally, it is a convex combination of each one of these schemes, where the weighting is chosen depending on the expected behaviour of the solution. At points where the solution should be smooth the simple scheme (22) is used, whereas if the solution is expected to be singular the robustness of (19) is better suited to capture this behaviour. Summing up, the following discretization is used

$$\begin{aligned}\mathrm{MA}_h^{\mathrm{H}}[w_h](x_h) = {} & \omega(x_h)\mathrm{MA}_h^{\mathrm{FD}}[w_h](x_h) \\ & + (1-\omega(x_h))\mathrm{MA}_{h,\theta}^{\mathrm{WS}}[w_h](x_h).\end{aligned} \tag{23}$$

Here $\omega \in C(\bar{\Omega}, [0,1])$ is a weighting function defined a priori from the data as follows: For $\epsilon > 0$ we let Ω_ϵ be a neighbourhood of the set where the solution u may be singular, that is,

$$\Omega_\epsilon = \{x \in \Omega : 0 \leq f(x) < \epsilon\} \cup \{x \in \partial\Omega : g \notin C^{2,\alpha}(U_x), \text{ or } U_x \cap \partial\Omega \text{ is flat}\},$$

where U_x is a neighbourhood of the point x. We then set $\omega \equiv 0$ in Ω_ϵ and one away from it. This scheme was tested in Froese and Oberman (2011b) for a

series of cases, ranging from smooth to singular solutions, and computational experiments suggested that this method is robust and accurate.

This method, however, has a major drawback. The tunable function ω must be described by the user, and its values depend on the behaviour of the problem data. For this reason in Froese and Oberman (2013) it was proposed that instead the difference

$$\left|\mathrm{MA}_{h,\theta}^{\mathrm{WS}}[w_h](x_h) - \mathrm{MA}_h^{\mathrm{FD}}[w_h](x_h)\right|,$$

be used as an a posteriori indicator of accuracy. In regions where this difference is small, it is expected that the solution is smooth, whereas when this is large one expects singularities. On the basis of this, we can choose which scheme to apply. The way to measure this difference is by introducing a filter.

Definition 11 (filter)**.**
A *filter* is a function $S \in C_0(\mathbb{R})$ such that $S(t) = t$ in a neighbourhood of the origin.

For instance, the function

$$S(t) = \begin{cases} x, & |x| \leq 1 \\ 0, & |x| \geq 2, \\ 2 - x, & 1 < x < 2, \\ -x - 2, & -2 < x < -1 \end{cases} \tag{24}$$

depicted in Fig. 1 is a possible filter, see Froese and Oberman (2013, Fig. 1.1 and (1.3)). With this at hand, a *filtered* operator can be defined via

$$\begin{aligned} \mathrm{MA}_h^{\mathrm{F}}[w_h](x_h) &= \mathrm{MA}_{h,\theta}^{\mathrm{WS}}[w_h](x_h) \\ &\quad + h^\alpha S\left(\frac{\mathrm{MA}_h^{\mathrm{FD}}[w_h](x_h) - \mathrm{MA}_{h,\theta}^{\mathrm{WS}}[w_h](x_h)}{h^\alpha}\right), \end{aligned} \tag{25}$$

where $\alpha \in (0, 2]$ is to be chosen by the user. A filtered scheme seeks $u_h \in X_h$ such that

$$\mathrm{MA}_h^{\mathrm{F}}[u_h](x_h) = f(x_h), \quad \forall x_h \in \Omega_h, \tag{26a}$$

$$u_h(x_h) = g(x_h), \quad \forall x_h \in \partial\Omega_h. \tag{26b}$$

Remark 5 (consistency).
Recall that (Kossaczký et al., 2016; Oberman, 2006) a monotone scheme cannot be more than second-order accurate. Notice, in addition, that by construction we have

$$\left|\mathrm{MA}_h^{\mathrm{F}}[w_h](x_h) - \mathrm{MA}_{h,\theta}^{\mathrm{WS}}[w_h](x_h)\right| \leq h^\alpha,$$

so that a filtered scheme is also consistent, up to second order. Moreover, if the parameter α is chosen smaller than the consistency order of both the

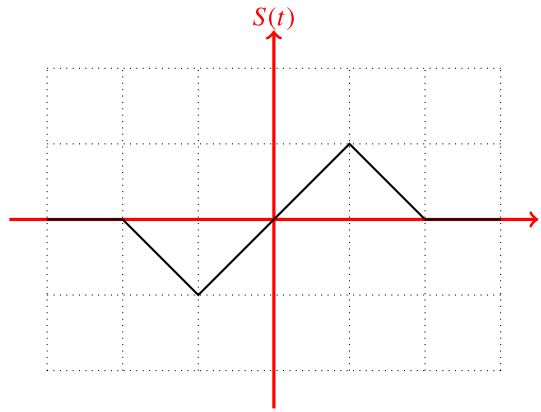

FIG. 1 The function S defined in (24) is a filter.

wide stencil, and the finite difference scheme, and the mesh size h is sufficiently small, it can be shown that

$$\mathrm{MA}_h^{\mathrm{F}}[\varphi](x) = \mathrm{MA}_h^{\mathrm{FD}}[\varphi](x),$$

whenever φ is sufficiently smooth. These two observations serve as a guideline for the choice of α. ■

Remark 6 (motivation).
The construction of a filtered scheme seems to be motivated by similar constructions for conservation laws and first order Hamilton Jacobi equations. For instance, Lions and Souganidis (1995) shows the convergence of filtered finite difference schemes (constructed in a similar way), for Hamilton Jacobi equations. In the realm of hyperbolic conservation laws, several types of limiters or artificial viscosity methods (Bonito et al., 2014; Guermond and Pasquetti, 2011; Guermond et al., 2011, 2018) have been derived from these ideas. ■

As a step towards the analysis of schemes like (26), Froese and Oberman (2013) introduced a class of schemes called *nearly monotone*, and showed that the theory of Section 2.1 also applies to them. To show this, we begin with a definition.

Definition 12 (nearly monotone)**.**
The family of approximation schemes $\{F_\varepsilon\}_{\varepsilon>0}$ where $F_\varepsilon : \bar{\Omega} \times \mathbb{R} \times B(\bar{\Omega})$ is called *nearly monotone*, if every F_ε can be written as

$$F_\varepsilon = F_\varepsilon^M + F_\varepsilon^P,$$

where F_ε^M is monotone in the sense of (12), and the function F_ε^P, called a perturbation, satisfies

$$\lim_{\varepsilon \downarrow 0} |F_\varepsilon^P(x,t,v)| = 0,$$

uniformly on bounded subsets of $\bar{\Omega} \times \mathbb{R} \times B(\bar{\Omega})$.

The convergence of nearly monotone schemes closely follows that of monotone schemes.

Corollary 1 (convergence)**.**
Let $\{F_\varepsilon\}_\varepsilon$ be a family of approximation schemes, that is nearly monotone, in the sense of Definition 12; consistent, in the sense of (14); and stable in the sense of (11). Assume, in addition, that problem (7) has a strong comparison principle. In this setting we have that, as $\varepsilon \downarrow 0$, the functions u_ε, solutions of $F_\varepsilon(x, u_\varepsilon(x), u_\varepsilon) = 0$ converge locally uniformly to u, solution of (7).

Proof. The proof is a small variation on the proof of Theorem 4. Indeed, with the notation of this proof, we have

$$\begin{aligned} 0 &= F_{\varepsilon_n}(y_n, u_{\varepsilon_n}(y_n), u_{\varepsilon_n}) \\ &= F^M_{\varepsilon_n}(y_n, \varphi(y_n) + \xi_n, \varphi + (u_{\varepsilon_n} - \varphi)) + F^P_{\varepsilon_n}(y_n, u_{\varepsilon_n}(y_n), u_{\varepsilon_n}) \\ &\leq F^M_{\varepsilon_n}(y_n, \varphi(y_n) + \xi_n, \varphi + \xi_n) + F^P_{\varepsilon_n}(y_n, u_{\varepsilon_n}(y_n), u_{\varepsilon_n}). \end{aligned}$$

The stability of the scheme allows us to invoke the fact that the perturbation vanishes in the limit. Consequently, we still have that $\overline{u}$ is a subsolution. □

Notice that the same considerations made in Remark 1 apply in this setting.

2.5 Lattice basis reduction scheme

Let us now discuss a two-dimensional method, which was introduced in Benamou et al. (2016) and is termed the *lattice basis reduction* scheme. The aim of this scheme is, for a given stencil, to obtain a different way to compute the determinant, so that the scheme is more accurate. We begin with a definition.

Definition 13 (superbasis)**.**
We will say that a *basis* of $\mathbb{Z}^2$ is a pair of vectors $(\boldsymbol{e}_1, \boldsymbol{e}_2) \in (\mathbb{Z}^2)^2$ that satisfy $|\det(\boldsymbol{e}_1, \boldsymbol{e}_2)| = 1$. A *superbasis* of $\mathbb{Z}^2$ is a triple $(\boldsymbol{e}_0, \boldsymbol{e}_1, \boldsymbol{e}_2) \in (\mathbb{Z}^2)^3$ such that $(\boldsymbol{e}_1, \boldsymbol{e}_2)$ is a basis and $\boldsymbol{e}_0 + \boldsymbol{e}_1 + \boldsymbol{e}_2 = 0$.

We will call a *stencil* a finite subset of $\mathbb{Z}^2 \backslash \{0\}$ that is symmetric around the origin. To a stencil S we associate the set of superbases

$$Y(S) = \left\{ (\boldsymbol{e}_0, \boldsymbol{e}_1, \boldsymbol{e}_2) \in S^3 : |\det(\boldsymbol{e}_1, \boldsymbol{e}_2)| = 1, \quad \boldsymbol{e}_0 + \boldsymbol{e}_1 + \boldsymbol{e}_2 = 0 \right\}.$$

With these notations at hand, we define the *lattice basis reduction* Monge–Ampère operator

$$\mathrm{MA}_{h,S}^{\mathrm{LBR}}[w_h](x_h) = \min_{(\boldsymbol{e}_0,\boldsymbol{e}_1,\boldsymbol{e}_2)\in Y(S)} \gamma\big((\Delta_{\boldsymbol{e}_0} w_h(x_h))^+, (\Delta_{\boldsymbol{e}_1} w_h(x_h))^+, (\Delta_{\boldsymbol{e}_2} w_h(x_h))^+\big), \tag{27}$$

where

$$\gamma(\delta_0,\delta_1,\delta_2) = \begin{cases} \delta_{i+1}\delta_{i+2}, & \delta_i \geq \delta_{i+1} + \delta_{i+2}, \\ \frac{1}{2}(\delta_0\delta_1 + \delta_1\delta_2 + \delta_0\delta_2) - \frac{1}{4}(\delta_0^2 + \delta_1^2 + \delta_2^2), & \text{otherwise.} \end{cases}$$

This allows us to introduce the following scheme: Find $u_h \in X_h$ such that

$$\mathrm{MA}_{h,S}^{\mathrm{LBR}}[u_h](x_h) = f(x_h), \quad \forall x_h \in \Omega_h, \tag{28a}$$

$$u_h(x_h) = g(x_h), \quad \forall x_h \in \partial\Omega_h. \tag{28b}$$

The motivation for this, at first glance obscure, definition of the operator $\mathrm{MA}_{h,S}^{\mathrm{LBR}}[\,\cdot\,]$ is given in Benamou et al. (2016, Remark 1.10). Let $Y = (\boldsymbol{e}_0, \boldsymbol{e}_1, \boldsymbol{e}_2) \in Y(S)$ and notice that for any point x_h that is interior with respect to Y, we have that the convex hull of $\{x_h \pm h\boldsymbol{e}_i\}_{i=0}^2$ is a hexagon. Given a function $w_h \in X_h$ we can associate to it its *local convex envelope*, that is the maximal convex function $\Gamma_{x_h,Y} w_h$ that is bounded from above by w_h at the points $\{x_h \pm h\boldsymbol{e}_i\}_{i=0}^2$. It is then possible to show that $\Gamma_{x_h,Y} w_h$ is a piecewise linear function over a particular triangulation of the aforementioned hexagon. Then we have that

$$\gamma\big((\Delta_{\boldsymbol{e}_0} w_h(x_h))^+, (\Delta_{\boldsymbol{e}_1} w_h(x_h))^+, (\Delta_{\boldsymbol{e}_2} w_h(x_h))^+\big) = |\partial\Gamma_{x_h,Y} w_h(x_h)|, \tag{29}$$

which is consistent with the definition of the Monge–Ampère operator in the sense of Alexandrov given in Definition 9 and hints at the consistency of this scheme.

The consistency analysis of the operator (27) hinges on the following definition.

Definition 14 (M–obtuseness)**.**
Let $M \in \mathbb{S}_+^2$. We say that the superbasis $(\boldsymbol{e}_0, \boldsymbol{e}_1, \boldsymbol{e}_2)$ of $\mathbb{Z}^2$ is M–obtuse if and only if

$$\boldsymbol{e}_j \cdot M\boldsymbol{e}_i \leq 0, \quad \forall 0 \leq i < j \leq 2.$$

From this definition, a necessary and sufficient condition for consistency follows (Benamou et al., 2016, Theorem 1.9).

Theorem 5 (consistency)**.**
Let $\varphi = \frac{1}{2} x \cdot Mx$ be a convex quadratic polynomial. We have that

$$\mathrm{MA}_{h,S}^{\mathrm{LBR}}[\varphi](x) = \det M, \quad \forall x$$

if and only if $Y(S)$ contains an M-obtuse superbais.

Proof. We will follow Benamou et al. (2016, Section 2.1). To simplify the discussion, we set

$$\mathcal{D} = \{(a_0,a_1,a_2) \in \mathbb{R}^3 : a_i \le a_{i+1} + a_{i+2}, \quad i = 0,1,2, \mod 3\},$$
$$\gamma_1(a_0,a_1,a_2) = \frac{1}{2}(a_0a_1 + a_1a_2 + a_0a_2) - \frac{1}{4}(a_0^2 + a_1^2 + a_2^2).$$

Notice that $\gamma(a_0, a_1, a_2) = \gamma_1(a_0, a_1, a_2)$ if and only if $(a_0,a_1,a_2) \in \mathcal{D}$, and that if that is not the case, then $\gamma(a_0,a_1,a_2) - \gamma_1(a_0,a_1,a_2) = \frac{1}{4}(a_0 - a_1 - a_2)^2 > 0$. In conclusion, we have that

$$\begin{cases} \gamma(a_0,a_1,a_2) \ge \gamma_1(a_0,a_1,a_2), \\ \gamma(a_0,a_1,a_2) = \gamma_1(a_0,a_1,a_2) \Leftrightarrow (a_0,a_1,a_2) \in \mathcal{D}. \end{cases} \tag{30}$$

Given a superbasis $(\boldsymbol{e}_0, \boldsymbol{e}_1, \boldsymbol{e}_2)$ define $\delta_i = \boldsymbol{e}_i \cdot M\boldsymbol{e}_i = (\Delta_{\boldsymbol{e}_i}\varphi(x_h))^+$. For a permutation (i, j, k) of (0, 1, 2) we have

$$\delta_i - \delta_j - \delta_k = (\boldsymbol{e}_j + \boldsymbol{e}_k) \cdot M(\boldsymbol{e}_j + \boldsymbol{e}_k) - \boldsymbol{e}_j \cdot M\boldsymbol{e}_j - \boldsymbol{e}_k M\boldsymbol{e}_k = 2\boldsymbol{e}_j \cdot M\boldsymbol{e}_k.$$

Consequently, $(\delta_0,\delta_1,\delta_2) \in \mathcal{D}$ if and only if the superbasis $(\boldsymbol{e}_0, \boldsymbol{e}_1, \boldsymbol{e}_2)$ is M–obtuse.

Let A be the linear transformation that maps $\boldsymbol{e}_1$ and $\boldsymbol{e}_2$ to $\boldsymbol{f}_1 = (1,0)^{\mathsf{T}}$ and $\boldsymbol{f}_2 = (0,1)^{\mathsf{T}}$, respectively. Then we must have that $\boldsymbol{f}_0 = A\boldsymbol{e}_0 = (-1,-1)^{\mathsf{T}}$. Thus, $\delta_i = \boldsymbol{e}_i \cdot M\boldsymbol{e}_i = A^{-1}\boldsymbol{f}_i \cdot MA^{-1}\boldsymbol{f}_i$, and so

$$\gamma_1(\mu_0,\mu_1,\mu_2) = \det(A^{-\mathsf{T}}MA^{-1}).$$

However, $\det A = |\det(\boldsymbol{e}_1,\boldsymbol{e}_2)/\det(\boldsymbol{f}_1,\boldsymbol{f}_2)| = 1$. Combining this with (30) we obtain the claim. □

Essentially, the previous result shows that the operator $\mathrm{MA}^{\mathrm{LBR}}_{h,S}[\,\cdot\,]$ systematically overestimates the determinant of the Hessian for quadratic functions, and that we have equality if and only if the stencil S contains a M-obtuse superbasis. For this reason, it is of interest to obtain conditions on the size of the stencil that guarantee that such a superbasis can be found. The following result is a restatement of Benamou et al. (2016, Proposition 1.12).

Proposition 6 (stencil size estimate)**.**
The stencil

$$S = \{\boldsymbol{e} \in \mathbb{Z}^2 : \gcd(\boldsymbol{e}) = 1, |\boldsymbol{e}| \le 2\kappa\},$$

contains a M–obtuse superbasis for every matrix $M \in \mathbb{S}^2_+$ *that satisfies*

$$\|M\|_2 \, \|M^{-1}\|_2 \le \kappa^2.$$

Notice that the cardinality of the stencil stated in Proposition 6 is quite large, approximately κ^2, and that if the solution degenerates, that is

$\det D^2 u(x_0) = 0$ at some point, then the stencil size must again grow unboundedly to maintain consistency.

Remark 7. The recent paper (Benamou and Duval, 2018) shows convergence of the lattice basis reduction scheme (28) applied to the optimal transport problem.

2.6 Discretization based on power diagrams

In Mirebeau (2015) the following discretization of the Monge–Ampère operator is proposed and analyzed. Let S be a stencil such that $\operatorname{span} S = \mathbb{R}^d$ and such that its elements have coprime coordinates, that is, if $\boldsymbol{e} = (e_1, \dots, e_d)^{\mathsf{T}} \in S$, then $\gcd(\boldsymbol{e}) = \gcd(e_1, \dots, e_d) = 1$. We define

$$\mathrm{MA}^{\mathrm{PD}}_{h,S}[w_h](x_h) = \left| \left\{ \boldsymbol{g} \in \mathbb{R}^d : \forall \boldsymbol{e} \in S : 2\boldsymbol{g} \cdot \boldsymbol{e} \leq |\boldsymbol{e}|^2 \Delta_{\boldsymbol{e}} w_h(x_h) \right\} \right|. \tag{31}$$

Here, we denote the Lebesgue measure by $|\cdot|$. With this operator at hand, we define the problem: find $u_h \in X_h$ such that

$$\mathrm{MA}^{\mathrm{PD}}_{h,S}[u_h](x_h) = f(x_h), \quad \forall x_h \in \Omega_h, \tag{32a}$$

$$u_h(x_h) = g(x_h), \quad \forall x_h \in \partial\Omega_h. \tag{32b}$$

Notice that the set entering the definition (31) is a polytope. Efficient ways to compute the volume of a polytope are available. For instance, if the dimension is not too high (and recall that we are mostly interested in the cases $d = 2$ or $d = 3$), one can first triangulate this polytope to then easily compute its volume.

Let us study the consistency of this scheme. To do so, we must introduce a definition.

Definition 15 (Voronoi cells and facets)**.**
Let $M \in \mathbb{S}^d_+$. The Voronoi cell and facet are

$$\mathrm{Vor}(M) = \left\{ \boldsymbol{g} \in \mathbb{R}^d : \forall \boldsymbol{e} \in \mathbb{Z}^d, \ \ \boldsymbol{g} \cdot M\boldsymbol{g} \leq (\boldsymbol{g} - \boldsymbol{e}) \cdot M(\boldsymbol{g} - \boldsymbol{e}) \right\},$$
$$\mathrm{Vor}(M, \boldsymbol{e}) = \left\{ \boldsymbol{g} \in \mathrm{Vor}(M) : \boldsymbol{g} \cdot M\boldsymbol{g} = (\boldsymbol{g} - \boldsymbol{e}) \cdot M(\boldsymbol{g} - \boldsymbol{e}) \right\}.$$

A M-Voronoi vector is an element $\boldsymbol{e} \in \mathbb{Z}^d \setminus \{0\}$ such that $\mathrm{Vor}(M, \boldsymbol{e}) \neq \emptyset$. It is a strict M-Voronoi vector if the facet $\mathrm{Vor}(M, \boldsymbol{e})$ is $(d-1)$-dimensional.

Now, the consistency of the operator defined in (31) is as follows.

Proposition 7 (consistency)**.**
Let $\varphi(x) = \frac{1}{2} x \cdot Mx$ be a convex quadratic. Then we have that

$$\mathrm{MA}^{\mathrm{PD}}_{h,S}[\varphi](x) = \det M, \quad \forall x$$

if and only if the stencil S contains all the strict M-Voronoi vectors.

Proof. Let $\kappa = \sqrt{\|M\|_2 \, \|M^{-1}\|_2}$. We divide the proof in several steps.

1. Any point $\boldsymbol{g} \in \mathrm{Vor}(M)$ must satisfy $|\boldsymbol{g}| \leq \frac{1}{2}\kappa\sqrt{d}$. Any M-Voronoi vector $\boldsymbol{e}$ satisfies $|\boldsymbol{e}| \leq \kappa\sqrt{d}$ and has coprime coordinates:

Indeed, if $\boldsymbol{g} \in \mathrm{Vor}(M)$, then let $\boldsymbol{e}_{\boldsymbol{g}} \in \mathbb{Z}^d$ be obtained by rounding its coordinates to the nearest integer, so that $|\boldsymbol{g} - \boldsymbol{e}_{\boldsymbol{g}}| \leq \frac{1}{2}\sqrt{d}$. The estimate

$$\frac{1}{\|M^{-1}\|_2}|\boldsymbol{g}|^2 \leq \boldsymbol{g} \cdot M\boldsymbol{g} \leq (\boldsymbol{g} - \boldsymbol{e}_{\boldsymbol{g}}) \cdot M(\boldsymbol{g} - \boldsymbol{e}_{\boldsymbol{g}}) \leq \|M\|_2 |\boldsymbol{g} - \boldsymbol{e}_{\boldsymbol{g}}|^2 \leq \frac{d}{4}\|M\|_2$$

yields the desired estimate. In addition, if $\boldsymbol{e}$ is a M-Voronoi vector, there is $\boldsymbol{g} \in \mathrm{Vor}(M)$ for which $|\boldsymbol{g}| = |\boldsymbol{e} - \boldsymbol{g}|$ so that

$$|\boldsymbol{e}| \leq 2|\boldsymbol{g}| \leq \kappa\sqrt{d}.$$

Finally, to show that the coordinates must be coprime consider $k\boldsymbol{e} \in \mathbb{Z}^d$ with $k \geq 2$ and notice that, for every $\boldsymbol{g} \in \mathbb{R}^d$ we have

$$(k\boldsymbol{e} - \boldsymbol{g}) \cdot M(k\boldsymbol{e} - \boldsymbol{g}) + (k-1)\boldsymbol{g} \cdot M\boldsymbol{g} = k(\boldsymbol{e} - \boldsymbol{g}) \cdot M(\boldsymbol{e} - \boldsymbol{g}) + (k^2 - k)\boldsymbol{e} \cdot M\boldsymbol{e}.$$

Consequently,

$$(\boldsymbol{e} - \boldsymbol{g}) \cdot M(\boldsymbol{e} - \boldsymbol{g}) < \max\{(k\boldsymbol{e} - \boldsymbol{g}) \cdot M(k\boldsymbol{e} - \boldsymbol{g}), \boldsymbol{g} \cdot M\boldsymbol{g}\},$$

and $k\boldsymbol{e}$ cannot be a M-Voronoi vector.

2. Let E be the set of strict M-Voronoi vectors, then

$$\mathrm{Vor}(M) \subset \left\{\boldsymbol{g} \in \mathbb{R}^d : \forall \boldsymbol{e} \in S : 2\boldsymbol{g} \cdot M\boldsymbol{e} \leq \boldsymbol{e} \cdot M\boldsymbol{e}\right\},$$

with equality if and only if $E \subset S$:

Notice that $\boldsymbol{g} \cdot M\boldsymbol{g} \leq (\boldsymbol{g} - \boldsymbol{e}) \cdot M(\boldsymbol{g} - \boldsymbol{e})$ is equivalent to saying that $2\boldsymbol{g} \cdot M\boldsymbol{e} \leq \boldsymbol{e} \cdot M\boldsymbol{e}$. This shows that $\mathrm{Vor}(M)$ is a convex polytope, defined by inequalities of this type where $\boldsymbol{e}$ runs over the set of strict M-Voronoi vectors. The bound established in the previous step shows that there can only be a finite number of them.

3. $|\mathrm{Vor}(M)| = 1$:

It follows from the observation that $\mathrm{Vor}(M)$ collects all elements $\boldsymbol{g} \in \mathbb{R}^d$ that are closer to the origin (in the metric induced by the matrix M) than to any other point $\boldsymbol{e} \in \mathbb{Z}^d \setminus \{0\}$.

4. Consistency:

Recall that, for any $\boldsymbol{e} \in S$ we have that $|\boldsymbol{e}|^2 \Delta_{\boldsymbol{e}} \varphi(x) = \boldsymbol{e} \cdot M\boldsymbol{e}$. Consequently,

$$\begin{aligned} \mathrm{MA}^{\mathrm{PD}}_{h,S}[\varphi](x) &= \left|\left\{\boldsymbol{g} \in \mathbb{R}^d : \forall \boldsymbol{e} \in S : 2\boldsymbol{g} \cdot \boldsymbol{e} \leq \boldsymbol{e} \cdot M\boldsymbol{e}\right\}\right| \\ &= \left|M\left\{\boldsymbol{g} \in \mathbb{R}^d : \forall \boldsymbol{e} \in S : 2\boldsymbol{g} \cdot M\boldsymbol{e} \leq \boldsymbol{e} \cdot M\boldsymbol{e}\right\}\right|. \end{aligned}$$

A combination of the second and third steps then yields

$$\mathrm{MA}^{\mathrm{PD}}_{h,S}[\varphi](x) \geq \det M |\mathrm{Vor}(M)| = \det M,$$

with equality if $\mathrm{Vor}(M) = \{\boldsymbol{g} \in \mathbb{R}^d : \forall \boldsymbol{e} \in S : 2\boldsymbol{g} \cdot \boldsymbol{e} \leq \boldsymbol{e} \cdot M\boldsymbol{e}\}$ with equality if $\mathrm{Vor}(M)$ contains all strict M-Voronoi vectors.

This concludes the proof. □

Since the consistency of the operator $\mathrm{MA}^{\mathrm{PD}}_{h,S}[\,\cdot\,]$ requires the stencil to contain all strict Voronoi vectors, it is necessary to provide sufficient conditions for this to happen.

Corollary 2 (stencil size estimate)**.**
Let $\kappa > 0$ and define

$$S = \left\{\boldsymbol{e} \in \mathbb{Z}^d : |\boldsymbol{e}| \leq \sqrt{d}\kappa, \quad \gcd(\boldsymbol{e}) = 1\right\}.$$

Let $\varphi(x) = \frac{1}{2} x \cdot M \cdot x$, then we have that

$$\mathrm{MA}^{\mathrm{PD}}_{h,S}[\varphi](x) = \det M, \quad \forall x$$

provided $\|M\|_2 \|M^{-1}\|_2 \leq \kappa^2$.

Proof. It immediately follows from the norm estimates given in Step 1 in the proof of Proposition 7. □

Let us now provide a convergence analysis of scheme (32), which will follow from the framework provided in Section 2.1. To do so, we introduce the operator $F_{h,S} : \bar{\Omega}_h \times \mathbb{R} \times X_h \to \mathbb{R}$ via

$$F_{h,S}(x_h, t, w) = \begin{cases} \mathrm{MA}^{\mathrm{PD}}_{h,S}[w](x_h) - f(x_h), & x_h \in \Omega_h, \\ g(x_h) - t, & x_h \in \partial\Omega_h, \end{cases} \tag{33}$$

and notice that (32) can be compactly written as

$$F_{h,S}(x_h, u_h(x_h), u_h) = 0, \quad \forall x_h \in \bar{\Omega}_h.$$

Let us also define the operator $F_S : \bar{\Omega} \times \mathbb{R} \times \mathbb{S}^d_+ \to \mathbb{R}$

$$F_S(x, t, M) = \begin{cases} |K(M)| - f(x), & x \in \Omega, \\ g(x) - t, & x \in \partial\Omega, \end{cases} \tag{34}$$

where

$$K(M) = \{\boldsymbol{v} \in \mathbb{R}^d : \forall \boldsymbol{e} \in S, \quad 2\boldsymbol{v} \cdot \boldsymbol{e} \leq \boldsymbol{e} \cdot M\boldsymbol{e}\}.$$

Notice that, if $D^2u(x_0)$ exists for all $x_0 \in \Omega$ and its eigenvalues are properly bounded, see Corollary 2, we have that

$$\det D^2u(x_0) - f(x_0) = F_S(x_0, u(x_0), D^2u(x_0)).$$

For this reason, we will consider the problem: find u that is a viscosity solution of

$$F_S(x,u(x),D^2u(x))=0,\quad x\in\bar{\Omega}. \tag{35}$$

Following Mirebeau (2015, Section 2.3) we will now show the convergence of scheme (32) via Theorem 4. To do so, we must show that scheme (32) is monotone, consistent, and stable in the sense of (12), (14), and (13), respectively. We have shown consistency in Proposition 7. For stability, we refer the reader to Mirebeau (2015, Section 2.2), where stability is shown by proving global convergence of a damped Newton algorithm. We will focus then on the monotonicity of the scheme.

Proposition 8 (monotonicity)**.**
The operator $F_{h,S}$, defined in (33) is monotone in the sense of (12).

Proof. Notice that, if $x_h \in \partial\Omega_h$, then there is nothing to show. On the other hand, if $x_h \in \Omega_h$, then $\mathrm{MA}^{\mathrm{PD}}_{h,S}[w](x_h)$ is an increasing function of the second differences $\Delta_{\boldsymbol{e}} w_h(x_h)$. Indeed, increasing this difference makes the polytope larger. Notice also that $\Delta_{\boldsymbol{e}} w_h(x_h)$ is a linear combination, with positive coefficients, of $w_h(x_h + \boldsymbol{e}h) - w_h(x_h)$ and $w_h(x_h + \boldsymbol{e}h) - w_h(x_h)$, with the obvious modification for points that are not interior with respect to $\boldsymbol{e}$. Thus, we can invoke (Neilan et al., 2017, Lemma 3.11) to conclude the monotonicity. □

Next to be able to apply Theorem 4 we must make sure that the operator F_S satisfies a comparison principle. To establish this we begin with an auxiliary result.

Lemma 4 (polytope comparison)**.**
Let $M_1, M_2 \in \mathbb{S}^d_+$ and $x \in \Omega$. If $M_1 \le M_2$ then, for every $t \in \mathbb{R}$ we have that $F_S(x, t, M_1) \le F_S(x, t, M_2)$. In addition,

$$\begin{aligned}(F_S(x,t,M_1+M_2)+f(x))^{1/d} \ge (F_S(x,t,M_1)+f(x))^{1/d}\\ +(F_S(x,t,M_2)+f(x))^{1/d}.\end{aligned}$$

Proof. Notice that, since $x \in \Omega$ we have, independently of t,

$$F(x,t,M)+f(x)=|K(M)|,\quad K(M)=\left\{\boldsymbol{v}\in\mathbb{R}^d:\forall \boldsymbol{e}\in S,\ \ 2\boldsymbol{v}\cdot\boldsymbol{e}\le \boldsymbol{e}\cdot M\boldsymbol{e}\right\}.$$

Notice, in addition, that $M_1 \le M_2$ implies that $\boldsymbol{e} \cdot M_1\boldsymbol{e} \le \boldsymbol{e} \cdot M_2\boldsymbol{e}$ for every $\boldsymbol{e} \in \mathbb{R}^d$. Consequently, $M_1 \le M_2$ implies $K(M_1) \subset K(M_2)$ from which the first statement follows.

Now, since $\boldsymbol{e} \cdot (M_1 + M_2)\boldsymbol{e} = \boldsymbol{e} \cdot M_1\boldsymbol{e} + \boldsymbol{e} \cdot M_2\boldsymbol{e}$ we have that $K(M_1 + M_2)$ contains $K(M_1) + K(M_2)$. The Brunn–Minkowski inequality given in Lemma 2 allows us to conclude. □

Now we can establish a comparison principle for F_S.

Proposition 9 (comparison)**.**
Let $\overline{u} \in USC(\bar{\Omega})$ and $\underline{u} \in LSC(\bar{\Omega})$ be a sub- and supersolution, respectively, of (35). Then we have that $\overline{u} \le \underline{u}$.

Proof. We begin by noticing that, since $F_{S,\star}(x,t,M) \leq F_S(x,t,M)$ for all $x \in \bar{\Omega}$ we obtain that, if $x_0 \in \partial\Omega$ we must have that

$$0 \leq F_{S,\star}(x_0,\varphi(x_0),D^2\varphi(x_0)) \leq F_S(x_0,\varphi(x_0),D^2\varphi(x_0)) = g(x_0) - \varphi(x_0),$$

for every φ sufficiently smooth that satisfies the conditions given in Definition 4. As a consequence, In this case, the condition defining a viscosity subsolution at boundary points reduces to $\overline{u} \leq g$ on $\partial\Omega$. Similarly we can show that for a supersolution we must have $g \leq \underline{u}$ on $\partial\Omega$. In conclusion, at the boundary $\partial\Omega$ we have $\overline{u} \leq \underline{u}$.

By the semicontinuity assumption we can also define $\delta = \sup_{\bar{\Omega}}(\overline{u} - \underline{u}) \in \mathbb{R}$. Additionally, since Ω is bounded, there is $R > 0$ such that $\Omega \subset B_R$. Assume now, for the sake of contradiction, that $\delta > 0$.

Let us define, for $\varepsilon > 0$, the operator $F_{S,\varepsilon} : \bar{\Omega} \times \mathbb{R} \times \mathbb{S}^d \to \mathbb{R}$ by

$$F_{S,\varepsilon}(x,t,M) = \begin{cases} F_S(x,t,M) - \varepsilon(t - \underline{u}(x)), & x \in \Omega, \\ g(x) - t, & x \in \partial\Omega, \end{cases}$$

and notice that this operator satisfies all the conditions of the comparison principle given in Crandall et al. (1992, Theorem 3.3). Moreover, since for all $x \in \bar{\Omega}$ we have that $F_{S,\varepsilon}(x,\underline{u}(x),D^2\underline{u}(x)) = F_S(x,\underline{u}(x),D^2\underline{u}(x))$ we conclude that $\underline{u}$ is a supersolution for the operator $F_{S,\varepsilon}$.

We now construct a subsolution. Define

$$v(x) = \frac{(\varepsilon\delta)^{1/d}}{2}\left(|x|^2 - R^2\right), \quad u_\varepsilon(x) = \overline{u}(x) + v(x)$$

and notice that $u_\varepsilon \in USC(\bar{\Omega})$ and, moreover, $u_\varepsilon \leq \overline{u} \leq g \leq \underline{u}$ on $\partial\Omega$. In addition, we have that, for $x \in \Omega$

$$D^2 v(x) = (\varepsilon\delta)^{1/d} I, \quad F_S(x,t,D^2 v(x)) + f(x) = \varepsilon\delta,$$

see the proof of Proposition 7. Let now $x \in \Omega$ and, to shorten notation, denote

$$F_S[w] = F_S(x,w(x),D^2 w(x)) + f(x).$$

If this is the case we have that, in the viscosity sense

$$\begin{aligned} F_{S,\varepsilon}(x,u_\varepsilon(x),D^2 u_\varepsilon(x)) &= F_S[\overline{u} + v] - f(x) - \varepsilon(\overline{u}(x) - \underline{u}(x)) - \varepsilon v(x) \\ &\geq \left(F_S[\overline{u}]^{1/d} + F_S[v]^{1/d}\right)^d - f(x) - \varepsilon(\overline{u}(x) - \underline{u}(x)) \\ &\geq \left(f(x)^{1/d} + F_S[v]^{1/d}\right)^d - f(x) - \varepsilon(\overline{u}(x) - \underline{u}(x)) \\ &\geq F_S[v] - \varepsilon(\overline{u}(x) - \underline{u}(x)) \\ &= \varepsilon(\delta - (\overline{u}(x) - \underline{u}(x))) \geq 0. \end{aligned}$$

where we used Lemma 4, the fact that $v(x) \leq 0$ for all $x \in \bar{\Omega}$, that $\overline{u}$ is a subsolution for the operator F_S, the elementary identity

$$(x+y)^\theta \le x^\theta + y^\theta, \quad \forall x,y \in \mathbb{R}_+, \quad \forall \theta \in (0,1],$$

and the definition of δ. In conclusion, u_ε is a subsolution for the operator $F_{S,\varepsilon}$. The comparison principle of Crandall et al. (1992, Theorem 3.3) then yields that

$$u_\varepsilon(x) = \overline{u}(x) + v(x) \le \underline{u}(x), \quad \forall x \in \bar{\Omega}$$

or that

$$\underline{u}(x) - \overline{u}(x) \ge -\frac{(\varepsilon\delta)^{1/d}}{2} R^2, \quad \forall x \in \bar{\Omega}.$$

Letting $\varepsilon \downarrow 0$ we obtain $\overline{u}(x) \le \underline{u}(x)$, contradicting that $\delta > 0$. □

As a consequence, we have convergence.

Corollary 3 (convergence)**.**
Let $\{u_h\}_{h>0} \subset X_h$ be the solutions to (32). Then, as $h \downarrow 0$, we have that $u_h \to u$ locally uniformly, where u is the (unique) viscosity solution of (35).

Proof. Apply Theorem 4. It is only relevant to mention that owing to the comparison principle showed in Proposition 9, u must necessarily be unique. □

2.7 Two scale methods

We will now present and analyze the so-called *two scale method*, which can be understood as a generalization of the wide stencil schemes presented in Section 2.3 to unstructured meshes (see also Froese (2018)). Here and in what follows we will implicitly assume that Ω is uniformly convex. Additional assumptions will be explicitly stated. Next, for $h > 0$, we introduce a quasiuniform (in the sense of Ciarlet (2002)) simplicial triangulation $\mathcal{T}_h$ of our domain Ω. We denote by Ω_h^i and Ω_h^b the set of interior and boundary nodes, respectively, of $\mathcal{T}_h$. We define X_h to be the set of piecewise linear and continuous functions subject to this triangulation. The mesh size h will constitute the fine scale of discretization. The large scale, denoted by δ, will be the one at which second-order differences will be evaluated. Notice that, since now we are dealing with continuous functions, these can be evaluated at any point. Indeed, given $x_h \in \Omega_h^i$ and $\boldsymbol{w} \in \mathbb{R}^d$ with $|\boldsymbol{w}| = 1$ we define, for $w_h \in X_h$

$$\nabla^2_{\delta \boldsymbol{w}} w_h(x_h) = \frac{w_h(x_h + \rho\delta\boldsymbol{w}) - 2w_h(x_h) + w_h(x_h - \rho\delta\boldsymbol{w})}{\rho^2\delta^2}, \tag{36}$$

where $\rho \in (0, 1]$ is the largest number so that $x_h \pm \rho\delta\boldsymbol{w} \in \bar{\Omega}$; compare with (17) and (18). As a final discretization ingredient, as in the case of the wide stencil schemes of Section 2.3, we need a directional discretization. That is if, as

before, $\mathcal{V}$ denotes the set of all orthonormal bases of $\mathbb{R}^d$ we must construct, for $\theta > 0$, a set $\mathcal{V}_\theta$ of collections of d unit vectors such that if $\{\boldsymbol{w}_i\}_{i=1}^d \in \mathcal{V}$, then there is $\{\boldsymbol{w}_i^\theta\}_{i=1}^d \in \mathcal{V}_\theta$ such that

$$\max_{i=1,\dots,d} |\boldsymbol{w}_i - \boldsymbol{w}_i^\theta| \leq \theta. \tag{37}$$

It is important to notice that the elements of $\mathcal{V}_\theta$ are not required to be orthonormal collections of vectors.

Having defined all the discretization ingredients, which are parametrized by the triple $\varepsilon = (h, \delta, \theta)$, following Nochetto et al. (2019a) we introduce the two scale discrete Monge–Ampère operator by defining, for $w_h \in X_h$, and $x_h \in \Omega_h^i$,

$$\begin{aligned} \mathrm{MA}^{2\mathrm{S}}_{h,\delta,\theta}[w_h](x_h) = \min_{\{\boldsymbol{w}_i\}_{i=1}^d \in \mathcal{V}_\theta} \Bigg[& \prod_{i=1}^d \left(\nabla^2_{\delta \boldsymbol{w}_i} w_h(x_h) \right)^+ \\ & - \sum_{i=1}^d \left(\nabla^2_{\delta \boldsymbol{w}_i} w_h(x_h) \right)^- \Bigg], \end{aligned} \tag{38}$$

compare with the scheme discussed in Remark 2. With these ingredients at hand, the two scale method seeks a function $u_h^\varepsilon \in X_h$ such that

$$\mathrm{MA}^{2\mathrm{S}}_{h,\delta,\theta}[u_h^\varepsilon](x_h) = f(x_h), \quad \forall x_h \in \Omega_h^i, \tag{39a}$$

$$u_h^\varepsilon(x_h) = g(x_h), \quad \forall x_h \in \Omega_h^b. \tag{39b}$$

Remark 8 (generalization).
Starting from the Cartesian mesh Ω_h used to define the wide stencil schemes (20) it is possible to construct a simplicial triangulation of Ω without introducing new vertices: in two dimensions this is accomplished by subdividing each square by its diagonal, and a similar construction is possible in three dimensions. Once this is done, it can be seen that scheme (39) is, after little modifications, a generalization of the wide stencil scheme (20). ■

Remark 9 (domain approximation).
Notice that, since the domain Ω is assumed to be uniformly convex, it is not possible to triangulate it exactly. If we denote $\bar{\Omega}_{\mathcal{T}_h} = \cup_{T \in \mathcal{T}_h} \overline{T}$, then we have $\bar{\Omega}_{\mathcal{T}_h} \subsetneq \bar{\Omega}$. In our discussion we will ignore this fact. This is because we can either replace Ω by $\Omega_{\mathcal{T}_h}$ in all the statements that we shall make, or we can consider all functions in X_h as defined in Ω by extending them to $\Omega \backslash \Omega_{\mathcal{T}_h}$ by a constant in the normal direction to faces. This is a standard construction and we shall not delve into it further. ■

Let us now provide, following Li and Nochetto (2018a); Nochetto et al. (2019a,b), an analysis of (39). We will first introduce a discrete notion of convexity, based on the positivity of the second differences defined in (36).

The operator (38) turns out to have a comparison principle, and acts in a particular way on discretely convex functions. This will allow us to establish existence, uniqueness, and stability of solutions to (39). In addition, since the size large scale δ is reduced near the boundary, the consistency can only hold sufficiently far away from it. For this reason, appropriate barrier functions need to be constructed. All these ingredients will allow us to assert convergence of the method. Finally, using the comparison principle and suitable barriers, we will establish rates of convergence for classical solutions.

2.7.1 Discrete convexity

The second-order differences defined in (36) and the set of directions $\mathcal{V}_\theta$ give a discrete notion of convexity.

Definition 16 (discrete convexity)**.**
We say that the function $w_h \in X_h$ is discretely convex if

$$\nabla^2_{\delta \boldsymbol{w}_j} w_h(x_h) \geq 0, \quad \forall x_h \in \Omega_h^i, \quad \forall \boldsymbol{w}_j \in \{\boldsymbol{w}_i\}_{i=1}^d \in \mathcal{V}_\theta.$$

It is well known that if a function is convex, then its second-order differences are nonnegative. On the other hand, discrete convexity does not imply convexity. This is due, for instance, to the fact that convexity and interpolation are not easily compatible. In other words, if $w \in C(\bar{\Omega})$ is convex, then its Lagrange interpolant $\mathcal{I}_h w \in X_h$ satisfies $\mathcal{I}_h w \geq w$ so that it is discretely convex, but $\mathcal{I}_h w$ is not necessarily convex.

On the other hand, discrete convexity implies nonnegativity of the two scale discrete Monge–Ampère operator; see Nochetto et al. (2019a, Lemma 2.2).

Lemma 5 (discrete convexity)**.**
A function $w_h \in X_h$ is discretely convex if and only if

$$\mathrm{MA}^{2\mathrm{S}}_{h,\delta,\theta}[w_h](x_h) \geq 0, \quad \forall x_h \in \Omega_h^i.$$

Moreover, for a discretely convex function we have that

$$\mathrm{MA}^{2\mathrm{S}}_{h,\delta,\theta}[w_h](x_h) = \min_{\{\boldsymbol{w}_i\}_{i=1}^d \in \mathcal{V}_\theta} \prod_{i=1}^d \nabla^2_{\delta \boldsymbol{w}_i} w_h(x_h).$$

2.7.2 A comparison principle

Let us now show that the operator defined in (38) is monotone and has a comparison principle. From this we will obtain uniqueness of solutions to (39).

Lemma 6 (monotonicity)**.**
Let $v_h, w_h \in X_h$ be such that $v_h - w_h$ attains its maximum at the interior node $x_h \in \Omega_h^i$. Then we have

$$\mathrm{MA}^{2\mathrm{S}}_{h,\delta,\theta}[w_h](x_h) \geq \mathrm{MA}^{2\mathrm{S}}_{h,\delta,\theta}[v_h](x_h).$$

Proof. Since x_h is the maximum, for suitable $\rho > 0$ and any unit vector $\boldsymbol{w}$ we have

$$v_h(x_h) - w_h(x_h) \geq v_h(x_h \pm \rho\delta\boldsymbol{w}) - w_h(x_h \pm \rho\delta\boldsymbol{w}),$$

which implies that

$$\nabla^2_{\delta\boldsymbol{w}} v_h(x_h) \leq \nabla^2_{\delta\boldsymbol{w}} w_h(x_h).$$

multiplying this inequality as $\boldsymbol{w}$ runs over all elements of $\mathcal{V}_\theta$ allows us to conclude. □

The previous result gives us a comparison principle for the operator (38).

Proposition 10 (comparison)**.**
Let v_h, $w_h \in X_h$ be such that $v_h \leq w_h$ on $\partial\Omega$, and

$$\mathrm{MA}^{2\mathrm{S}}_{h,\delta,\theta}[v_h](x_h) \geq \mathrm{MA}^{2\mathrm{S}}_{h,\delta,\theta}[w_h](x_h), \quad \forall x_h \in \Omega_h^i,$$

then we must have that $v_h \leq w_h$ in $\bar{\Omega}$.

Proof. We consider two cases for the inequality between the operators:

1. *The inequality is strict.* Let us assume, for the sake of contradiction, $v_h - w_h$ attains a maximum at an interior node. Lemma 6 then gives a contradiction.
2. *The inequality is not strict.* Since Ω is bounded, there is $R > 0$ such that the convex quadratic $q(x) = \frac{1}{2}(|x|^2 - R)$ is nonpositive on $\bar{\Omega}$. Let $q_h = \mathcal{I}_h q \in X_h$. This function is strictly convex and satisfies

$$\nabla^2_{\delta\boldsymbol{w}} q_h(x_h) \geq \nabla^2_{\delta\boldsymbol{w}} q(x_h) = \frac{\partial^2 q(x_h)}{\partial \boldsymbol{w}^2} = 1.$$

We claim now that, for all $\alpha > 0$ and $x_h \in \Omega_h^i$, we have that

$$\mathrm{MA}^{2\mathrm{S}}_{h,\delta,\theta}[v_h + \alpha q_h](x_h) \geq \mathrm{MA}^{2\mathrm{S}}_{h,\delta,\theta}[v_h](x_h) + \min\left\{\frac{\alpha^d}{2^d} + \frac{\alpha}{2}\right\}. \tag{40}$$

Indeed, fix $\{\boldsymbol{w}_i\} \in \mathcal{V}_\theta$ and assume first that $\nabla^2_{\delta\boldsymbol{w}_i}(v_h(x_h) + \frac{\alpha}{2} q_h(x_h)) \geq 0$ for all i. In this case

$$\begin{aligned}
\prod_{i=1}^{d}\left(\nabla^2_{\delta\boldsymbol{w}_i} v_h(x_h) + \alpha q_h(x_h)\right) &\geq \prod_{i=1}^{d}\left(\nabla^2_{\delta\boldsymbol{w}_i}(v_h(x_h) + \frac{\alpha}{2} q_h(x_h)) + \frac{\alpha}{2}\right) \\
&\geq \min_{\{\boldsymbol{w}_i\}\in\mathcal{V}_\theta} \prod_{i=1}^{d}\left(\nabla^2_{\delta\boldsymbol{w}_i}(v_h(x_h) + \frac{\alpha}{2} q_h(x_h))\right) + \frac{\alpha^d}{2^d} \\
&\geq \left(\prod_{i=1}^{d}\left(\nabla^2_{\delta\boldsymbol{w}_i} v_h(x_h)\right)^+ - \sum_{i=1}^{d}\left(\nabla^2_{\delta\boldsymbol{w}_i} v_h(x_h)\right)^-\right) + \frac{\alpha^d}{2^d}.
\end{aligned}$$

On the other hand, if there is $i \in \{1, \ldots, d\}$ for which $\nabla^2_{\delta \boldsymbol{w}_i}(v_h(x_h) + \alpha q_h(x_h)) < 0$, then this implies that $\nabla^2_{\delta \boldsymbol{w}_i} v_h(x_h) < 0$. Thus,

$$\prod_{i=1}^{d} \left(\nabla^2_{\delta \boldsymbol{w}_i} v_h(x_h)\right)^+ = 0,$$

and

$$\begin{aligned} -\sum_{i=1}^{d} \left(\nabla^2_{\delta \boldsymbol{w}_i}(v_h(x_h) + \alpha q_h(x_h))\right)^- &\geq -\sum_{i=1}^{d} \left(\nabla^2_{\delta \boldsymbol{w}_i} v_h(x_h)\right)^- + \frac{\alpha}{2} \\ &= \left(\prod_{i=1}^{d} \left(\nabla^2_{\delta \boldsymbol{w}_i} v_h(x_h)\right)^+ - \sum_{i=1}^{d} \left(\nabla^2_{\delta \boldsymbol{w}_i} v_h(x_h)\right)^-\right) + \frac{\alpha}{2}. \end{aligned}$$

A combination of these two cases, since $\{\boldsymbol{w}_i\}_{i=1}^d \in \mathcal{V}_\theta$ was arbitrary, implies (40).

Finally, since, $v_h + \alpha q_h \leq v_h \leq w_h$ on $\partial\Omega$ and, on the basis of (40), we have

$$\mathrm{MA}^{2\mathrm{S}}_{h,\delta,\theta}[v_h + \alpha q_h](x_h) > \mathrm{MA}^{2\mathrm{S}}_{h,\delta,\theta}[v_h](x_h) \geq \mathrm{MA}^{2\mathrm{S}}_{h,\delta,\theta}[w_h](x_h), \quad \forall x_h \in \Omega_h^i,$$

the previous step then implies that $v_h + \alpha q_h \leq w_h$. Letting $\alpha \downarrow 0$ we can conclude. □

Remark 10 (discrete interior barrier).
Notice, that, in the course of the second case of the proof of this result we effectively constructed a discrete interior barrier. If $q(x) = \frac{1}{2}(|x|^2 - R)$ with $R > 0$ sufficiently large, then we have that

$$\mathcal{I}_h q \leq 0, \quad \text{on } \partial\Omega, \qquad \mathrm{MA}^{2\mathrm{S}}_{h,\delta,\theta}[\mathcal{I}_h q_h](x_h) \geq 1, \quad \forall x_h \in \Omega_h^i.$$

■

As an immediate consequence, we also have uniqueness of solutions to (39).

Corollary 4 (uniqueness)**.**
Scheme (39) *cannot have more than one solution.*

As a final application of the comparison principle, let us now show existence and uniform bounds on the solution to (39).

Theorem 6 (existence and stability)**.**
For all $\varepsilon = (h, \delta, \theta) > 0$ *scheme* (39) *has a solution* $u_h^\varepsilon \in X_h$*. Moreover, this solution is stable in the sense that* $\| u_h^\varepsilon \|_{L^\infty(\Omega)}$ *is bounded independently of* ε*.*

Proof. The existence proceeds via Perron's method. For this reason, we will only indicate how to construct a discrete subsolution, that is a function $u_h^0 \in X_h$ such that $u_h^0 = \mathcal{I}_h g$ on $\partial\Omega$ and

$$\mathrm{MA}^{2\mathrm{S}}_{h,\delta,\theta}[u_h^0](x_h) \geq f(x_h), \quad \forall x_h \in \Omega_h^i.$$

To construct this function, we define

$$s(x) = \sum_{x_h \in \Omega_h^i} s_{x_h}(x), \qquad s_{x_h}(x) = \frac{\delta \rho_{x_h}}{2} f(x_h)^{1/d} |x - x_h|,$$

where $\rho_{x_h} \in (0,1]$ is the largest number such that, for all $\boldsymbol{w} \in \mathbb{R}^d$ with $|\boldsymbol{w}| = 1$ we have $x_h \pm \rho_{x_h} \boldsymbol{w} \in \bar{\Omega}$. Notice that $\nabla^2_{\delta \boldsymbol{w}} s_{x_h}(y_h) \geq 0$ for all $y_h \in \Omega_h^i$, and that

$$\nabla^2_{\delta \boldsymbol{w}} s_{x_h}(x_h) = f(x_h)^{1/d}, \quad \forall \boldsymbol{w} \in \mathbb{R}^d, \quad |\boldsymbol{w}| = 1.$$

Consequently, for $y_h \in \Omega_h^i$

$$\nabla^2_{\delta \boldsymbol{w}} \mathcal{I}_h s(y_h) \geq \nabla^2_{\delta \boldsymbol{w}} s(y_h) \geq f(y_h)^{1/d} \geq 0,$$

which, by Lemma 5 implies

$$\mathrm{MA}^{2\mathrm{S}}_{h,\delta,\theta}[\mathcal{I}_h s](x_h) = \min_{\{\boldsymbol{w}_i\}_{i=1}^d \in \mathcal{V}_\theta} \prod_{i=1}^d \nabla^2_{\delta \boldsymbol{w}_i} \mathcal{I}_h s(x_h) \geq f(x_h), \quad \forall x_h \in \Omega_h^i.$$

Let now $w \in C(\bar{\Omega})$ be the convex envelope of $(\mathcal{I}_h(g - s))|_{\partial\Omega}$, and set $w_h = \mathcal{I}_h w$. By convexity of w we have that

$$\mathrm{MA}^{2\mathrm{S}}_{h,\delta,\theta}[w_h](x_h) \geq 0, \quad \forall x_h \in \Omega_h^i.$$

Thus, we define

$$u_h^0 = w_h + \mathcal{I}_h s.$$

This function, by construction, is discretely convex and $u_h^0 = \mathcal{I}_h g$ on $\partial\Omega$. Since the second differences of w_h are nonnegative, then we have that

$$\begin{aligned} \mathrm{MA}^{2\mathrm{S}}_{h,\delta,\theta}[u_h^0](x_h) &= \min_{\{\boldsymbol{w}_i\}_{i=1}^d \in \mathcal{V}_\theta} \prod_{i=1}^d \left[\nabla^2_{\delta \boldsymbol{w}_i} w_h(x_h) + \nabla^2_{\delta \boldsymbol{w}_i} \mathcal{I}_h s(x_h) \right] \\ &\geq \min_{\{\boldsymbol{w}_i\}_{i=1}^d \in \mathcal{V}_\theta} \prod_{i=1}^d \nabla^2_{\delta \boldsymbol{w}_i} \mathcal{I}_h s(x_h) \geq f(x_h), \end{aligned}$$

and so u_h^0 is a discrete subsolution.

It remains to show the uniform boundedness. To achieve this we will show that every discrete subsolution is uniformly bounded. Let then $w_h \in X_h$ be a discrete subsolution and $b_h = \max_{x \in \partial\Omega} g(x) \in X_h$. We have then that

$$\mathrm{MA}^{2\mathrm{S}}_{h,\delta,\theta}[b_h](x_h) = 0 \leq f(x_h) \leq \mathrm{MA}^{2\mathrm{S}}_{h,\delta,\theta}[w_h](x_h), \quad \forall x_h \in \Omega_h^i.$$

Since, in addition, we have that $b_h \geq w_h$ on $\partial\Omega$, the comparison principle of Proposition 10 implies that

$$w_h \leq b_h.$$

This is enough since Perron's method shows existence of a solution by constructing an increasing sequence of subsolutions. Thus, u_h^0 is a lower bound for the solution and, evidently, $\|u_h^0\|_{L^\infty(\Omega)}$ is independent of ε. □

2.7.3 Consistency and discrete barriers

Let us now examine the consistency of the operator (38). As we have stated above, the operator can only be consistent at points sufficiently far away from the boundary. For this reason, we define the δ-interior and δ-boundary layer of Ω via

$$\Omega_\delta = \bigcup_{T\in\mathcal{T}_h:\operatorname{dist}(T,\partial\Omega)>\delta} T, \quad (\partial\Omega)_\delta = \bar{\Omega}\backslash\Omega_\delta.$$

For an interior node $x_h \in \Omega_h^i$ its interior patch is

$$\omega_{x_h} = \bigcup_{T\in\mathcal{T}_h:\operatorname{dist}(x_h,T)<\rho\delta} \overline{T},$$

where, as before, $\rho \in (0, 1]$ is the largest number such that, for any $\boldsymbol{w} \in \mathbb{R}^d$ with $|\boldsymbol{w}| = 1$ we have $x_h \pm \rho\delta\boldsymbol{w} \in \bar{\Omega}$.

The following result follows, essentially, by a Taylor expansion argument.

Lemma 7 (consistency of second differences)**.**
Let $x_h \in \Omega_h^i$ and assume that $\varphi \in C^{1,1}(\omega_{x_h})$, then for all $\boldsymbol{w} \in \mathbb{R}^d$ with $|\boldsymbol{w}| = 1$ we have

$$|\nabla^2_{\delta\boldsymbol{w}}\mathcal{I}_h\varphi(x_h)| \leq C|\varphi|_{C^{1,1}(\omega_{x_h})}.$$

If, in addition, we have that $x_h \in \Omega_\delta$ and $\varphi \in C^{k+2,\alpha}(\omega_{x_h})$ for $k = 0, 1$, and $\alpha \in (0, 1]$ then we also have that

$$\left|\nabla^2_{\delta\boldsymbol{w}}\mathcal{I}_h\varphi(x_h) - \frac{\partial^2\varphi(x_h)}{\partial\boldsymbol{w}^2}\right| \leq C\left(|\varphi|_{C^{k+2,\alpha}(\omega_{x_h})}\delta^{k+\alpha} + \frac{h^2}{\delta^2}|\varphi|_{C^{1,1}(\omega_{x_h})}\right).$$

Finally, if φ is, in addition, convex then we have

$$\frac{\partial^2\varphi(x_h)}{\partial\boldsymbol{w}^2} - \nabla^2_{\delta\boldsymbol{w}}\mathcal{I}_h\varphi(x_h) \leq C|\varphi|_{C^{k+2,\alpha}(\omega_{x_h})}\delta^{k+\alpha}.$$

The previous result can be applied to obtain interior consistency of (38). The following result was first obtained in Nochetto et al. (2019a, Lemma 4.2) under the assumption that $\mathcal{V}_\theta \subset \mathcal{V}$. This assumption was later removed in Li and Nochetto (2018a, Lemma 2.4).

Theorem 7 (interior consistency)**.**
Let $x_h \in \Omega_h^i$ and $\varphi \in C^{k+2,\alpha}(\omega_{x_h})$ with $k = 0, 1$ and $\alpha \in (0, 1]$ be convex. In this setting the following estimates are valid:

1. *If $\theta \leq \frac{1}{4d}$ then, for any $\{\boldsymbol{w}_i\}_{i=1}^d \in \mathcal{V}_\theta$, we have*

$$\det D^2\varphi(x_h) \leq \prod_{i=1}^{d} \frac{\partial^2 \varphi(x_h)}{\partial \boldsymbol{w}_i^2} \left(1 + 16\theta^2 (d-1)^2\right).$$

2. *If $\{\boldsymbol{v}_i\}_{i=1}^d \in \mathcal{V}$ realizes the minimum in the variational characterization of the determinant given in Lemma 3, then for any $\{\boldsymbol{v}_i^\theta\}_{i=1}^d \in \mathcal{V}_\theta$ that satisfies (37) we have*

$$\left| \frac{\partial^2 \varphi(x_h)}{\partial \boldsymbol{v}_i^2} - \frac{\partial^2 \varphi(x_h)}{\partial \boldsymbol{v}_i^{\theta 2}} \right| \leq C |\varphi|_{C^{1,1}(\omega_{x_h})} \theta^2.$$

3. *Finally, if $x_h \in \Omega_h^i \cap \Omega_\delta$, then*

$$\left| \det D^2\varphi(x_h) - \mathrm{MA}^{2\mathrm{S}}_{h,\delta,\theta}[\mathcal{I}_h\varphi](x_h) \right| \leq C_1 \delta^{k+\alpha} + C_2 \left(\frac{h^2}{\delta^2} + \theta^2 \right),$$

where the constants C_1 and C_2 depend only on the smoothness of φ, the domain Ω, and the dimension d.

Proof. We prove each statement separately.

1. Let $W_\theta = (\boldsymbol{w}_1, \ldots, \boldsymbol{w}_d)$. We have

$$\det (W_\theta^\mathsf{T} W_\theta) \det D^2\varphi(x_h) = \det (W_\theta^\mathsf{T} D^2\varphi(x_h) W_\theta) \leq \prod_{i=1}^{d} \frac{\partial^2 \varphi(x_h)}{\partial \boldsymbol{w}_i^2},$$

where, in the last step, we used that $W_\theta^\mathsf{T} D^2\varphi(x_h) W_\theta$ is positive semidefinite and Hadamard's inequality. We now need to estimate the determinant of $W = W_\theta^\mathsf{T} W_\theta$ from below. Write

$$W = \begin{pmatrix} W_0 & \boldsymbol{w} \\ \boldsymbol{w}^\mathsf{T} & 1 \end{pmatrix} = \begin{pmatrix} I & 0 \\ \boldsymbol{w}^\mathsf{T} W_0^{-1} & 1 \end{pmatrix} \begin{pmatrix} W_0 & \boldsymbol{w} \\ 0 & 1 - \boldsymbol{w} \cdot W_0 \boldsymbol{w} \end{pmatrix}$$

implying that $\det W = (1 - \boldsymbol{w} \cdot W_0 \boldsymbol{w}) \det W_0$, which holds if the submatrix W_0 is nonsingular. Notice, however, that $W_{i,\,i} = 1$ and $|W_{i,\,j}| \leq 2\theta$ as the columns of W_θ form an element of $\mathcal{V}_\theta$. This implies, for $\theta \leq \frac{1}{4d}$, that $W_0 \geq \frac{1}{2} I$ and $|\boldsymbol{w}| \leq 2\theta\sqrt{d-1}$. Thus, $W_0^{-1} \geq 2I$ and

$$|\boldsymbol{w} \cdot W_0 \boldsymbol{w}| \leq 8\theta^2 (d-1) \qquad \det W \geq (1 - 8\theta^2 (d-1)) \det W_0,$$

which by repeating this process yields

$$\det W \geq (1 - 8\theta^2 (d-1))^{d-1} \geq 1 - 8\theta^2 (d-1)^2,$$

and using, again the bound on θ

$$\frac{1}{\det W} \leq 1 + 16\theta^2 (d-1)^2.$$

2. We begin by noticing that, given the minimization assumption, $\{\boldsymbol{v}_i\}_{i=1}^d$ must be the normalized eigenvectors of $D^2\varphi(x_h)$. Set $\boldsymbol{v}_i^\theta = \boldsymbol{v}_i + \boldsymbol{w}_i$ and write

$$\frac{\partial^2\varphi(x_h)}{\partial \boldsymbol{v}_i^{\theta 2}} = \boldsymbol{v}_i^\theta \cdot D^2\varphi(x_h)\boldsymbol{v}_i^\theta = \frac{\partial^2\varphi(x_h)}{\partial \boldsymbol{v}_i^2} + 2\boldsymbol{w}_i \cdot D^2\varphi(x_h)\boldsymbol{v}_i + \boldsymbol{w}_i \cdot D^2\varphi(x_h)\boldsymbol{w}_i.$$

Since $\{\boldsymbol{v}_i\}_{i=1}^d$ are eigenvectors, $|\boldsymbol{w}_i| \leq \theta$, and $|\boldsymbol{v}_i \cdot \boldsymbol{w}_i| \leq \frac{1}{2}\theta^2$ we then have

$$\left|\frac{\partial^2\varphi(x_h)}{\partial \boldsymbol{v}_i^{\theta 2}} - \frac{\partial^2\varphi(x_h)}{\partial \boldsymbol{v}_i^2}\right| \leq C\theta^2.$$

3. By Lemma 5, since $\mathcal{I}_h\varphi$ is discretely convex, we have that

$$\mathrm{MA}^{2\mathrm{S}}_{h,\delta,\theta}[\mathcal{I}_h\varphi](x_h) = \min_{\{\boldsymbol{w}_i\}_{i=1}^d \in \mathcal{V}_\theta} \prod_{i=1}^d \nabla^2_{\delta \boldsymbol{w}_i} \mathcal{I}_h\varphi(x_h).$$

Let $\{\boldsymbol{w}_i\}_{i=1}^d \in \mathcal{V}_\theta$ be the set that realizes the minimum in this expression. Using Lemma 3 we can write that

$$\begin{aligned}\det D^2\varphi(x_h) - \mathrm{MA}^{2\mathrm{S}}_{h,\delta,\theta}[\mathcal{I}_h\varphi](x_h) &\leq \prod_{i=1}^d \frac{\partial^2\varphi(x_h)}{\partial \boldsymbol{w}_i^2} - \prod_{i=1}^d \nabla^2_{\delta \boldsymbol{w}_i}\mathcal{I}_h\varphi(x_h) \\ &\leq C\delta^{k+\alpha},\end{aligned}$$

where, in the last step, we used repeatedly Lemma 7.

Let now $\{\boldsymbol{v}_i\}_{i=1}^d \in \mathcal{V}$ be the normalized eigenpairs of $D^2\varphi(x_h)$, and $\{\boldsymbol{v}_i^\theta\}_{i=1}^d \in \mathcal{V}_\theta$ the collection that realizes (37). Then we have

$$\begin{aligned}&\mathrm{MA}^{2\mathrm{S}}_{h,\delta,\theta}[\mathcal{I}_h\varphi](x_h) - \det D^2\varphi(x_h) \\ &\leq \left(\prod_{i=1}^d \nabla^2_{\delta \boldsymbol{v}_i^\theta}\mathcal{I}_h\varphi(x_h) - \prod_{i=1}^d \frac{\partial^2\varphi(x_h)}{\partial \boldsymbol{v}_i^{\theta 2}}\right) + \left(\prod_{i=1}^d \frac{\partial^2\varphi(x_h)}{\partial \boldsymbol{v}_i^{\theta 2}} - \prod_{i=1}^d \frac{\partial^2\varphi(x_h)}{\partial \boldsymbol{v}_i{}^2}\right).\end{aligned}$$

The first term can be handled by repeatedly applying Lemma 7, while the second by applying the previous step.

All the estimates have been proved and the interior consistency is thus obtained. □

As mentioned before, the operator is not consistent near the boundary. For this reason we will, instead, construct discrete barriers which will allow us to control the behaviour of the solution near the boundary.

Proposition 11 (discrete barrier I)**.**
Let $E > 0$ be arbitrary and $x_h \in \Omega_h^i$ be such that dist$(x_h, \partial\Omega) \leq \delta$. Then, there is $p_h \in X_h$ such that

$$p_h \leq 0, \text{ on } \partial\Omega, \quad \mathrm{MA}^{2\mathrm{S}}_{h,\delta,\theta}[p_h](y_h) \geq E, \quad \forall y_h \in \Omega_h^i, \quad |p_h(x_h)| \leq CE^{1/d}\delta,$$

where the constant C depends only on the domain Ω.

Proof. Without loss of generality, we can assume that $x_h = (0,\dots,0,z)^{\mathsf{T}}$ with $z > 0$ so that $0 \in \partial\Omega$ and $z = \operatorname{dist}(x_h, \partial\Omega)$. The uniform convexity of Ω shows that there is $R > 0$ such that, in this system of coordinates,

$$\Omega \subset \left\{ x \in \mathbb{R}^d : \sum_{i=1}^{d-1} x_i^2 - (x_d - R)^2 \le R^2 \right\}.$$

Let

$$p(x) = \frac{E^{1/d}}{2}\left(\sum_{i=1}^{d-1} x_i^2 - (x_d - R)^2 - R^2 \right).$$

We claim that $p_h = \mathcal{I}_h p$ is the desired barrier. Indeed, by construction $p_h \le 0$ on the boundary $\partial\Omega$ and, since $z \le \delta$ we have that $|p_h(x_h)| \le CE^{1/d}\delta$. Finally, since p_h is discretely convex, for any interior node y_h we have

$$\mathrm{MA}^{2\mathrm{S}}_{h,\delta,\theta}[p_h](y_h) \ge \mathrm{MA}^{2\mathrm{S}}_{h,\delta,\theta}[p](y_h) = \prod_{i=1}^{d} E^{1/d} = E,$$

as claimed. □

To obtain rates of convergence we shall also require another discrete barrier that was originally introduced in Nochetto and Zhang (2018, Section 6.2). We define

$$\zeta : [0,\infty) \to (-\infty, 0], \qquad \zeta(t) = \begin{cases} (t-2\delta)^2 - (2\delta)^2, & t \in [0, 2\delta], \\ -(2\delta)^2, & t \in (2\delta, \infty). \end{cases}$$

The graph of this function is illustrated in Fig. 2. With this function at hand, we define

$$b(x) = \zeta(\operatorname{dist}(x, \partial\Omega)),$$

and $b_h = \mathcal{I}_h b$. The properties of this barrier are as follows.

Proposition 12 (discrete barrier II)**.**
For $\theta \le \frac{1}{2\sqrt{d}}$ the barrier function b_h satisfies:

1. *For all $x_h \in \Omega_h^i$ and any $\boldsymbol{w} \in \mathbb{R}^d$ with $|\boldsymbol{w}| = 1$,*

$$\nabla^2_{\delta \boldsymbol{w}} b_h(x_h) \ge 0.$$

2. *For all $x_h \in \Omega_h^i \backslash \Omega_\delta$ and $\{\boldsymbol{w}_i^\theta\}_{i=1}^d \in \mathcal{V}_\theta$,*

$$\max_{i=1,\dots,d} \nabla^2_{\delta \boldsymbol{w}_i^\theta} b_h(x_h) \ge \frac{1}{2d}.$$

3. *For all $x \in \bar{\Omega}$*

$$-4\delta^2 \le b_h(x) \le 0.$$

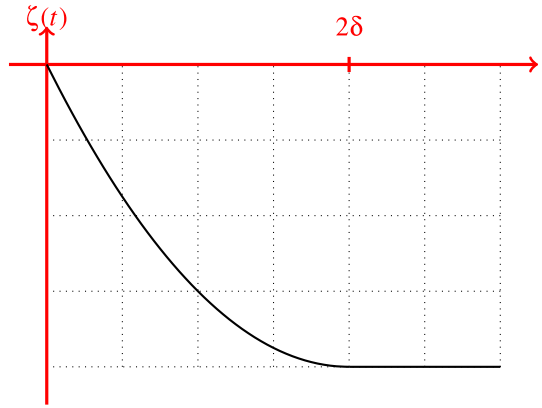

FIG. 2 The function ζ used to define the discrete barrier of Proposition 12.

Proof. We consider each property separately.

1. Let x_+, $x_-\in\Omega$ with $x_+ \neq x_-$. The convexity of Ω ensures that $x_0 = \frac{1}{2}(x_+ + x_-) \in \Omega$. Denote by $y \in \partial\Omega$ the closest point to x_0. Since Ω is convex, there is a supporting hyperplane P at y, whose normal is $\boldsymbol{n} = \frac{1}{|x_0 - y|}(x_0 - y)$. Let now $\boldsymbol{v} = \pm(x_+ - x_-)$, where the sign is chosen so that $\boldsymbol{n} \cdot \boldsymbol{v} \geq 0$. Consequently, see Fig. 3,

$$\operatorname{dist}(x_\pm, \partial\Omega) \leq \operatorname{dist}(x_\pm, P) = \operatorname{dist}(x_0, \partial\Omega) \pm \boldsymbol{n} \cdot \boldsymbol{v}.$$

With this estimate, and using that ζ is nonincreasing, we can compute

$$\begin{aligned} b(x_+) + b(x_-) &\geq \zeta(\operatorname{dist}(x_0, \partial\Omega) + \boldsymbol{v} \cdot \boldsymbol{n}) + \zeta(\operatorname{dist}(x_0, \partial\Omega) - \boldsymbol{v} \cdot \boldsymbol{n}) \\ &\geq 2\zeta(\operatorname{dist}(x_0, \partial\Omega)) = 2b(x_0), \end{aligned}$$

where the second inequality follows directly from the definition of ζ. We then conclude (cf. Krasnosel'skiĭ and Rutickiĭ, 1961, Pages 1–2) that the function b is convex and the stated property of b_h follows.

2. With the notation of the previous step, if we take a node $x_h \in \Omega_h^i \backslash \Omega_\delta$, and $\boldsymbol{v} \in \mathbb{R}^d$ with $|\boldsymbol{v}| = 1$, then $\operatorname{dist}(x_h, \partial\Omega) \pm \rho\delta\boldsymbol{w} \cdot \boldsymbol{n} \in [0, 2\delta]$. Since ζ is nonincreasing and quadratic on that interval

$$\nabla^2_{\delta\boldsymbol{v}} b_h(x_h) \geq \nabla^2_{\delta\boldsymbol{v}} b(x_h) \geq 2\frac{\rho^2\delta^2|\boldsymbol{v} \cdot \boldsymbol{n}|^2}{\rho^2\delta^2} = 2|\boldsymbol{v} \cdot \boldsymbol{n}|^2.$$

Now, if we let $\boldsymbol{v}$ run over $\{\boldsymbol{w}_i^\theta\}_{i=1}^d \in \mathcal{V}_\theta$ we have obtained that

$$\max_{i=1,\ldots,d} \nabla^2_{\delta\boldsymbol{w}_i^\theta} b_h(x_h) \geq 2 \max_{i=1,\ldots,d} |\boldsymbol{w}_i^\theta \cdot \boldsymbol{n}|^2.$$

Let now $\{\boldsymbol{w}_i\}_{i=1}^d \in \mathcal{V}$ be such that it satisfies (37). Since $|\boldsymbol{n}| = 1$ we must have that

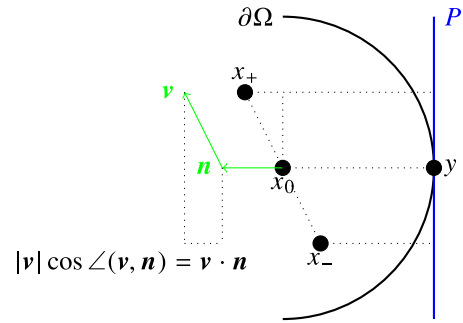

FIG. 3 The construction Proposition 12 that shows that the function b is convex. The distance between x_+ and the supporting hyperplane P equals the sum of the distance from x_0 to the boundary $\partial\Omega$ and the inner product between $\boldsymbol{n}$ and $\boldsymbol{v}$.

$$\sum_{i=1}^{d} |\boldsymbol{n}\cdot\boldsymbol{w}_i|^2 = 1, \quad \Longrightarrow \max_{i=1,\dots,d} |\boldsymbol{n}\cdot\boldsymbol{w}_i| \geq \frac{1}{\sqrt{d}}.$$

Therefore,

$$|\boldsymbol{w}_i^\theta\cdot\boldsymbol{n}| \geq |\boldsymbol{w}_i\cdot\boldsymbol{n}| - |(\boldsymbol{w}_i-\boldsymbol{w}_i^\theta)\cdot\boldsymbol{n}| \geq |\boldsymbol{w}_i\cdot\boldsymbol{n}| - \theta \geq \frac{1}{2\sqrt{d}},$$

where we used that $\theta \leq \frac{1}{2\sqrt{d}}$. This implies the estimate.

3. The last property follows directly from the definition of the function ζ. □

2.7.4 Convergence

Let us now show convergence. We will do so by adapting the arguments developed in Section 2.1 to take into account that test functions must be convex. We will rely on Proposition 3.

Theorem 8 (convergence)**.**
Let Ω be uniformly convex, $f \in C(\bar{\Omega})$ such that $f \geq 0$, and $g \in C(\partial\Omega)$. As $\varepsilon = (h, \delta, \theta) \to 0$ with $h\delta^{-1} \to 0$ we have that the family $\{u_h^\varepsilon\}_\varepsilon$ of solutions of (39) converges uniformly to $u \in C(\bar{\Omega})$, the solution of (1).

Proof. In a similar way to Theorem 7 we have that, for all $x_0 \in \Omega$, $x_h \in \Omega_h^i \cap \Omega_\delta$ and all $\varphi \in C^{2,\alpha}(\omega_{x_h})$, it holds that

$$|\mathrm{MA}[\varphi](x_0) - \mathrm{MA}^{2S}_{h,\delta,\theta}[\mathcal{I}_h\varphi](x_h)| \leq C_1(\delta^\alpha + |x_0 - x_h|^\alpha) + C_2\left(\frac{h^2}{\delta^2} + \theta^2\right). \tag{41}$$

Indeed, we only need to use that the operations $t \mapsto t^\pm$ are Lipschitz and with Lipschitz constant equal one.

We now extend the ideas of Theorem 4. As there we define

$$\overline{u}(x) = \limsup_{\varepsilon\to 0, \frac{h}{\delta}\to 0, y\to x} u_h^\varepsilon(y), \qquad \underline{u}(x) = \liminf_{\varepsilon\to 0, \frac{h}{\delta}\to 0, y\to x} u_h^\varepsilon(y)$$

and we will show that $\overline{u}$ is a subsolution. For that we assume that $\overline{u} - \varphi$, with $\varphi \in C^{2,\alpha}(\bar{\Omega})$ attains a maximum at $x_0 \in \Omega$. Let $\{x_h\}$ be the sequence of nodes such that $x_h \to x_0$ and $u_h^\varepsilon - \mathcal{I}_h\varphi$ attains a maximum at x_h. By the monotonicity result of Lemma 6 we obtain then that

$$\mathrm{MA}^{2S}_{h,\delta,\theta}[\mathcal{I}_h\varphi](x_h) \geq \mathrm{MA}^{2S}_{h,\delta,\theta}[u_h^\varepsilon](x_h) \geq f(x_h),$$

the consistency, as expressed in (41), implies by passing to the limit that

$$\mathrm{MA}[\varphi](x_0) \geq f(x_0).$$

It remains to understand the boundary behaviour of $\overline{u}$. We will show that the boundary condition is attained in a classical sense, that is $\overline{u} = g$. Let $x \in \partial\Omega$ and p_k be the continuous quadratic constructed during the proof of existence of the boundary barrier function in Proposition 11 with constant $E = k$. As k can be taken arbitrarily large, the sequence of points where $g \pm p_k$ attains a maximum (minimum) over $\partial\Omega$, converges to x.

We now observe that the monotonicity of Lemma 6 implies that if $v_h \in X_h$ is such that $\mathrm{MA}^{2S}_{h,\delta,\theta}[v_h](x_h) > 0$ for all $x_h \in \Omega_h^i$, then v_h attains its maximum on $\partial\Omega$. Since

$$\mathrm{MA}^{2S}_{h,\delta,\theta}[u_h^\varepsilon + \mathcal{I}_h p_k](x_h) > 0, \quad \forall x_h \in \Omega_h^i,$$

we can apply this observation to $u_h^\varepsilon + \mathcal{I}_h p_k$ to obtain that, for $x \in \partial\Omega$,

$$\begin{aligned}\overline{u}(x) &\leq \limsup_{\varepsilon\to 0, \frac{h}{\delta}\to 0, y\to x} \left(u_h^\varepsilon(y) + \mathcal{I}_h p_k(y)\right) - \liminf_{\varepsilon\to 0, \frac{h}{\delta}\to 0, y\to x} \mathcal{I}_h p_k(y) \\ &\leq \limsup_{\varepsilon\to 0, \frac{h}{\delta}\to 0, y\to x} \max_{z\in\partial\Omega} \mathcal{I}_h(g(z) + p_k(z)) - p_k(x) \leq g(x_k) + p_k(x_k) - p_k(x),\end{aligned}$$

where x_k is the point where $g + p_k$ attains its maximum over $\partial\Omega$. Letting $k \to \infty$ we conclude $\overline{u} \leq g$. Similarly $\underline{u} \geq g$.

Finally we invoke the comparison principle of Proposition 4 to conclude. □

Remark 11 (convergence by regularization).
It is interesting to note that by invoking the continuous dependence result given in Proposition 1, and the approximation result of Proposition 2, another proof of convergence can be developed. See Nochetto et al. (2019a, Section 5.3) for details. ■

2.7.5 Rates of convergence

The ingredients used to assert the convergence of the two scale method (39) were employed in Nochetto et al. (2019b) to obtain rates of convergence.

The techniques used in this reference were very similar to those that we will describe in Section 3 and so, to avoid repetition, we shall not elaborate on them here. This is further justified by that fact that, although Nochetto et al. (2019b) was the first work to provide rates of convergence for wide stencil-type methods, the rates of convergence obtained in this work were suboptimal.

Let us here, instead, present the results obtained in Li and Nochetto (2018a), where optimal rates of convergence have been obtained. The main tools in this are the comparison principle of Proposition 10 and the discrete barriers constructed in Section 2.7.3.

We begin by noticing that we shall only assume

$$f \geq 0,$$

so that the Monge–Ampère equation (1) may be degenerate. The main result about rates of convergence for classical solutions is the following.

Theorem 9 (error estimate)**.**
Let $u \in C^{2,\alpha}(\bar{\Omega})$, with $\alpha \in (0,1]$, solve (1) and $u_h^\varepsilon \in X_h$ solve (39). If $\theta \leq \frac{1}{4d}$ then we have

$$\|u - u_h^\varepsilon\|_{L^\infty(\Omega)} \leq C\left[h^2\left(1+\delta^{-2}\right)|u|_{C^{1,1}(\bar{\Omega})} + \delta^\alpha |u|_{C^{2,\alpha}(\bar{\Omega})}\right],$$

where the constant C depends on the domain Ω, the dimension d, and the shape regularity of the mesh $\mathcal{T}_h$, but is independent of h, and the solution u.

Proof. Recall that a standard interpolation estimate yields

$$\| u - \mathcal{I}_h u\|_{L^\infty(\Omega)} \leq Ch^2 |u|_{C^{1,1}(\bar{\Omega})},$$

so that we only need to bound the difference $u_h - \mathcal{I}_h u$. To do so, we will construct a suitable discrete subsolution u_h^- and supersolution u_h^+ and use the comparison principle of Proposition 10.

Let $u_h^- = \mathcal{I}_h u + K_1 q_h \in X_h$, where q_h is the interior barrier of Remark 10 and $K_1 > 0$ is to be chosen later. Notice that, by construction

$$u_h^- \leq \mathcal{I}_h u = \mathcal{I}_h g, \quad \text{on } \partial\Omega.$$

Thus, to guarantee that this is a subsolution we must show that

$$\mathrm{MA}^{2S}_{h,\delta,\theta}[u_h^-](x_h) \geq f(x_h) = \det D^2 u(x_h), \quad \forall x_h \in \Omega_h^i.$$

However, since u_h^- is discretely convex, showing this inequality reduces to showing that, for all $\{\boldsymbol{w}_i\}_{i=1}^d \in \mathcal{V}_\theta$ we have

$$\prod_{i=1}^d \nabla^2_{\delta \boldsymbol{w}_i} u_h^-(x_h) \geq \det D^2 u(x_h), \quad \forall x_h \in \Omega_h^i,$$

see Lemma 5. Using the convexity of u, we have, according to Lemma 7, that

$$\nabla^2_{\delta \boldsymbol{w}_i} \mathcal{I}_h u(x_h) \geq \frac{\partial^2 u(x_h)}{\partial \boldsymbol{w}_i^2} - C|u|_{C^{2,\alpha}(\bar{\Omega})}\delta^\alpha,$$

so that, upon choosing

$$K_1 = C\left[\delta^\alpha |u|_{C^{2,\alpha}(\bar{\Omega})} + \left(\frac{h^2}{\delta^2} + \theta^2\right)|u|_{C^{1,1}(\bar{\Omega})}\right],$$

where C is sufficiently large, we have

$$\begin{aligned}\nabla^2_{\delta \boldsymbol{w}_i} u_h^-(x_h) &\geq \frac{\partial^2 u(x_h)}{\partial \boldsymbol{w}_i^2} - C|u|_{C^{2,\alpha}(\bar{\Omega})}\delta^\alpha + K_1 \geq \frac{\partial^2 u(x_h)}{\partial \boldsymbol{w}_i^2} + C\theta^2 |u|_{C^{1,1}(\bar{\Omega})} \\ &\geq \left(1 + 16\theta^2(d-1)^2\right)\frac{\partial^2 u(x_h)}{\partial \boldsymbol{w}_i^2} \geq \left(1 + 16\theta^2(d-1)^2\right)^{1/d}\frac{\partial^2 u(x_h)}{\partial \boldsymbol{w}_i^2}.\end{aligned}$$

Finally, since $\theta \leq \frac{1}{4d}$, we multiply this inequality over $i = 1, \ldots, d$ and invoke Theorem 7 item 1 to conclude that u_h^- is a subsolution. The comparison principle of Proposition 10 then yields that

$$\begin{aligned}u_h^\varepsilon \geq u_h^- &= \mathcal{I}_h u + C\left(\delta^\alpha |u|_{C^{1,1}(\bar{\Omega})} + \left(\frac{h^2}{\delta^2} + \theta^2\right)|u|_{C^{1,1}(\bar{\Omega})}\right) q_h \\ &\geq \mathcal{I}_h u - C\left(\delta^\alpha |u|_{C^{1,1}(\bar{\Omega})} + \theta^2 |u|_{C^{1,1}(\bar{\Omega})}\right).\end{aligned}$$

We now define

$$u_h^+ = \mathcal{I}_h u - K_1 q_h - K_2 b_h,$$

where q_h and K_1 are as before, b_h is the barrier of Proposition 12 and $K_2 > 0$ is to be chosen. We show that u_h^+ is a supersolution. First of all, because of the choice of signs

$$u_h^+ \geq \mathcal{I}_h u = \mathcal{I}_h g, \quad \text{on } \partial\Omega.$$

Now, to show the inequality between operators we must consider in Ω_δ and outside of it separately. Let $x_h \in \Omega_h^i \cap \Omega_\delta$ and $\{\boldsymbol{v}_i\}_{i=1}^d \in \mathcal{V}$ such that

$$f(x_h) = \det D^2 u(x_h) = \prod_{i=1}^d \frac{\partial^2 u(x_h)}{\partial \boldsymbol{v}_i^2}.$$

Let now $\{\boldsymbol{v}_i^\theta\}_{i=1}^d \in \mathcal{V}_\theta$ satisfy (37). The interior consistency of second differences of Lemma 7, together with the estimate of Theorem 7 item 2 gives us that

$$\left|\nabla^2_{\delta \boldsymbol{v}_i^\theta} \mathcal{I}_h u(x_h) - \frac{\partial^2 u(x_h)}{\partial \boldsymbol{v}_i^2}\right| \leq C\left[\delta^\alpha |u|_{C^{1,1}(\bar{\Omega})} + \left(\frac{h^2}{\delta^2} + \theta^2\right)|u|_{C^{1,1}(\bar{\Omega})}\right],$$

which, using that $\nabla^2_{\delta \boldsymbol{v}_i^\theta} q_h(x_h) \geq 1$, $\nabla^2_{\delta \boldsymbol{v}_i^\theta} b_h(x_h) \geq 0$, and the definition of K_1 immediately implies that

$$\nabla^2_{\delta \boldsymbol{v}_i^\theta} u_h^+(x_h) \leq \frac{\partial^2 u(x_h)}{\partial \boldsymbol{v}_i^2}.$$

Notice now that u_h^+ might not be discretely convex, so that $\nabla^2_{\delta \boldsymbol{v}_i^\theta} u_h^+(x_h)$ might be negative. To deal with this we define the function

$$G : \mathbb{R}^d \to \mathbb{R}, \quad G(\boldsymbol{z}) = \prod_{i=1}^d (\boldsymbol{z} \cdot \boldsymbol{e}_i)^+ - \sum_{i=1}^d (\boldsymbol{z} \cdot \boldsymbol{e}_i)^-,$$

where $\{\boldsymbol{e}_i\}_{i=1}^d$ is the canonical basis of $\mathbb{R}^d$. Notice that this function is monotone in each coordinate of $\boldsymbol{z}$. Moreover if, for $\{\boldsymbol{w}_i\}_{i=1}^d \in \mathcal{V}_\theta$ and $w_h \in X_h$, we define the vectors

$$\boldsymbol{\chi}(w_h, \{\boldsymbol{w}_i\}) = \left(\nabla^2_{\delta \boldsymbol{w}_1} w_h(x_h), \ldots, \nabla^2_{\delta \boldsymbol{w}_d} w_h(x_h) \right)^\mathsf{T},$$

$$\boldsymbol{\gamma} = \left(\frac{\partial^2 u(x_h)}{\partial \boldsymbol{v}_1^2}, \ldots, \frac{\partial^2 u(x_h)}{\partial \boldsymbol{v}_d^2} \right)^\mathsf{T}.$$

Then we have that

$$\mathrm{MA}^{2\mathrm{S}}_{h,\delta,\theta}[w_h](x_h) = \min_{\{\boldsymbol{w}_i\}_{i=1}^d \in \mathcal{V}_\theta} G(\boldsymbol{\chi}(w_h, \{\boldsymbol{w}_i\})).$$

Therefore

$$\mathrm{MA}^{2\mathrm{S}}_{h,\delta,\theta}[u_h^+](x_h) \leq G(\boldsymbol{\chi}(u_h^+, \{\boldsymbol{v}_i^\theta\})) \leq G(\boldsymbol{\gamma}) = \prod_{i=1}^d \frac{\partial^2 u(x_h)}{\partial \boldsymbol{v}_i^2} = f(x_h).$$

Consider now a node close to the boundary, that is $x_h \in \Omega_h^i \setminus \Omega_\delta$, and let $\{\boldsymbol{w}_i^\theta\}_{i=1}^d \in \mathcal{V}_\theta$. Using Proposition 12 item 2 we have that

$$\max_{i=1,\ldots,d} \nabla^2_{\delta \boldsymbol{w}_i^\theta} b_h(x_h) \geq \frac{1}{2d}.$$

Assume that this maximum is attained for index k. Using Lemma 7 we can conclude that

$$\begin{aligned} \nabla^2_{\delta \boldsymbol{w}_k^\theta} u_h^+(x_h) &\leq \nabla^2_{\delta \boldsymbol{w}_k^\theta} \mathcal{I}_h u(x_h) - K_2 \nabla^2_{\delta \boldsymbol{w}_k^\theta} b_h(x_h) \\ &\leq \nabla^2_{\delta \boldsymbol{w}_k^\theta} \mathcal{I}_h u(x_h) - \frac{1}{2d} K_2 \leq C |u|_{C^{1,1}(\bar{\Omega})} - \frac{1}{2d} K_2 \leq 0, \end{aligned}$$

where the last step holds upon choosing K_2 sufficiently large. This shows that

$$\min_{i=1,\dots,d} \nabla^2_{\delta \boldsymbol{w}_i^\theta} u_h^+(x_h) = 0 \implies \mathrm{MA}^{2\mathrm{S}}_{h,\delta,\theta}[u_h^+](x_h) = 0 \le f(x_h).$$

We have shown that, for all $x_h \in \Omega_h^i$, we have $\mathrm{MA}^{2\mathrm{S}}_{h,\delta,\theta}[u_h^+](x_h) \le f(x_h)$, so that u_h^+ is a supersolution. The discrete comparison principle of Proposition 10 then allows us to conclude that

$$u_h \le u_h^+ = \mathcal{I}_h u - K_1 q_h - K_2 b_h \le \mathcal{I}_h u + C_1 K_1 + C_2 \delta^2 K_2,$$

where we used the lower bounds on q_h and b_h. Recalling the choices of K_1 and K_2 allows us to conclude. □

Choosing relations between the discretization parameters h, δ, and θ we can obtain explicit rates of convergence.

Corollary 5 (rates of convergence)**.**

In the setting of Theorem 9, if $\delta = C_1 h^{\frac{2}{2+\alpha}}$ *and* $\theta = C_2 h^{\frac{2}{2+\alpha}}$*, we have*

$$\|u - u_h^\varepsilon\|_{L^\infty(\Omega)} \le C h^{\frac{2\alpha}{2+\alpha}}.$$

On the other hand, choosing $\delta = h^{2/3}$ *and* $\theta = h^{1/3}$*, then we have*

$$\|u - u_h^\varepsilon\|_{L^\infty(\Omega)} \le C h^{\frac{2\alpha}{3}}.$$

In both estimates the hidden constant is independent of h*.*

Notice that both choices of relations between the coarse parameters and the mesh size h in Corollary 5 have its benefits and drawbacks. While the first choice yields a faster rate of convergence, it requires knowledge of the regularity of u. On the other hand, the second choice yields a slower convergence rate, but does not require a priori knowledge of the smoothness of u.

Remark 12 (error estimates under different assumptions).

The results of Theorem 9 have been extended in Li and Nochetto (2018a) in several directions:

1. *Smoother solutions*: If $u \in C^{3,\alpha}(\bar{\Omega})$*mutatis mutandis* the proof of Theorem 9 it follows a rate of convergence. The discretization parameters can be related to each other in such a way that the error is $\mathcal{O}(h)$, and numerical experiments indicate that this is sharp.
2. *Estimates for solutions with Sobolev regularity*: Assuming that $u \in W^{s,p}(\Omega)$ with $s \le 3$ and $s - d/p > 2$, and that $D^2 u(x) \ge \lambda I$, it has been shown (Li and Nochetto, 2018a, Theorem 5.7) that we have

$$\|u-u_h^\varepsilon\|_{L^\infty(\Omega)} \le C\left(\frac{h^2}{\delta^2}+\theta^2+\delta^2+\frac{\delta^{s-2}}{\lambda}\right),$$

where the constant depends on the smoothness of u. Once again, the discretization parameters can be optimized to obtain a rate $\mathcal{O}(h^{2-4/s})$.

■

2.8 Extensions, generalizations, and applications

We conclude the discussion on finite difference schemes and its variants by briefly describing some connections, extensions, generalizations, and applications of the schemes discussed here.

2.8.1 Hamilton Jacobi Bellman formulation and semi-Lagrangian schemes

Let

$$\Lambda = \left\{ \boldsymbol{\lambda} \in \mathbb{R}^d : \boldsymbol{\lambda}_i \ge 0, \quad i=1,\dots,d, \quad \sum_{i=1}^d \boldsymbol{\lambda}_i = 1 \right\}.$$

Define the function $h : \mathbb{S}^d \times \mathbb{R}_+ \to \mathbb{R}$ by

$$h(M,t) = \sup_{\substack{\{w_i\}_{i=1}^d \in \nu \\ \boldsymbol{\lambda}\in\Lambda}} \left[-\frac{1}{d}\sum_{i=1}^d \boldsymbol{\lambda}_i \boldsymbol{w}_i \cdot M\boldsymbol{w}_i + t^{1/d}\left(\prod_{i=1}^d \boldsymbol{\lambda}_i\right)^{1/d} \right].$$

The following result is from Krylov (1987), see also Neilan et al. (2017, Proposition 6.13).

Proposition 13 (determinant)**.**
For $M \in \mathbb{S}^d$ and $\delta \in \mathbb{R}_+$ we have that

$$h(M,\delta) = 0,$$

if and only if $M \in \mathbb{S}^d_+$ and $\det M = \delta$.

This motivates to define the function $F_{HJB} : \bar{\Omega} \times \mathbb{R} \times \mathbb{S}^d \to \mathbb{R}$ by

$$F_{HJB}(x,r,M) = \begin{cases} h(M, f(x)), & x \in \Omega, \\ g(x) - r, & x \in \partial\Omega, \end{cases}$$

and consider the problem: find $u \in C(\bar{\Omega})$ that is a viscosity solution of

$$F_{HJB}(x, u(x), D^2u(x)) = 0, \quad x \in \bar{\Omega}. \tag{42}$$

It turns out that this problem has an intimate connection with (1), as shown in Feng and Jensen (2017, Theorems 3.3 and 3.5).

Theorem 10 (equivalence)**.**
Let $f \in C(\Omega)$ be nonnegative. The function $u \in C(\Omega) \cap B(\bar{\Omega})$ is a viscosity solution of (42), in the sense of Definition 4, if and only if it is a viscosity solution on the set of convex functions of (1), in the sense of Definition 6.

It is remarkable that the convexity assumption on the solution is not enforced in (42), it is rather a consequence of the formulation. This motivated Feng and Jensen (2017) to use (42) for numerical purposes. They proposed a so-called semi-Lagrangian scheme which we now describe. Over a triangulation $\mathcal{T}_h$ we introduce X_h as the space of piecewise linear and continuous functions. On the basis of (42) we introduce over X_h the operator

$$\mathrm{MA}^{\mathrm{SL}}_{h,k}[w_h](x_h) = \sup_{\substack{\{w_i\}_{i=1}^d \in \nu \\ \boldsymbol{\lambda} \in \Lambda}} \left[-\frac{1}{d} \sum_{i=1}^{d} \boldsymbol{\lambda}_i \nabla^2_{k\boldsymbol{w}_i} w_h(x_h) + f(x_h)^{1/d} \left(\prod_{i=1}^{d} \boldsymbol{\lambda}_i \right)^{1/d} \right],$$

where $x_h \in \Omega_h^i$ and $k > 0$ is a discretization parameter. The semi-Lagrangian scheme then seeks for $u_h \in X_h$ such that

$$\mathrm{MA}^{\mathrm{SL}}_{h,k}[u_h](x_h) = 0, \quad \forall x_h \in \Omega_h^i, \tag{43a}$$

$$u_h(x_h) = g(x_h), \quad \forall x_h \in \Omega_h^b. \tag{43b}$$

Feng and Jensen (2017) showed existence and uniqueness of solutions to (43) as well as, provided $(h, k) \to 0$ with $\frac{h}{k} \to 0$, convergence to the viscosity solution of (42) and, as a consequence of Theorem 10, to the viscosity solution of (1) over the set of convex functions. Rates of convergence, however, were not provided.

Although rates of convergence for general semi-Lagrangian schemes were given in Debrabant and Jakobsen (2013, Corollary 7.3) let us here explore a connection between the solutions of the scheme (43) and the two scale method of Section 2.7 as described in Li and Nochetto (2018a, Section 6). For that one needs to notice, first, that the scheme given in (43) is not fully practical. This is because in the operator $\mathrm{MA}^{\mathrm{SL}}_{h,k}[\,\cdot\,]$ the supremum runs over all of $\mathcal{V}$. We need to introduce a directional discretization by, as before, using $\mathcal{V}_\theta$ whose elements satisfy (37). With this we introduce the new operator

$$\mathrm{MA}^{\mathrm{SL}}_{h,k,\theta}[w_h](x_h) = \sup_{\substack{\{w_i\}_{i=1}^d \in \nu_\theta \\ \boldsymbol{\lambda} \in \Lambda}} \left[-\frac{1}{d} \sum_{i=1}^{d} \boldsymbol{\lambda}_i \nabla^2_{k\boldsymbol{w}_i} w_h(x_h) + f(x_h)^{1/d} \left(\prod_{i=1}^{d} \boldsymbol{\lambda}_i \right)^{1/d} \right],$$

and denote by $u_h^{(k,\theta)} \in X_h$ the solution to (43) but with this new operator. The following is a rather surprising fact. For a proof see Li and Nochetto (2018a, Proposition 6.2).

Proposition 14 (equivalence)**.**

Let $u_h^\varepsilon \in X_h$ denote the solution to the two scale method (39) and $u_h^{(k,\theta)} \in X_h$ the solution to the modified semi-Lagrangian scheme with the operator $\mathrm{MA}^{\mathrm{SL}}_{h,k,\theta}[\cdot]$. In this case, we have $u_h^\varepsilon = u_h^{(k,\theta)}$.

From Proposition 14 and the results of Section 2.7.5, rates of convergence for (43) can be deduced.

Remark 13 (nonconvex domains).

Notice that convexity of the solution is not a constraint in (42) but rather a consequence of it. This has motivated (Jensen, 2018) to explore the possibility of using (42) as an extension of the Monge–Ampère equation to nonconvex domains, or cases with nonconvex data. ■

2.8.2 Filtered two scale schemes

In Nochetto and Ntogkas (2018) the ideas of two scale methods of Section 2.7 and those of filtered schemes of Section 2.4 were extended to construct a *filtered two scale scheme*. Let $\mathcal{T}^2_{2h}$ be a quasiuniform triangulation of $\bar\Omega$ of size $2h > 0$. The superscript in this triangulation indicates that we are doing a quadratic approximation of the boundary. This can be accomplished, for instance, by the use of isoparametric approximation of the boundary; see Brenner and Scott (2008, Section 10.4) and Ciarlet (2002, Section 4.3). Over this mesh we construct X^2_{2h}, the space of piecewise quadratic and continuous functions. For $w_{2h} \in X^2_{2h}$ and $x_{2h} \in \Omega^i_{2h}$ we define

$$\mathrm{MA}^{2\mathrm{Sq}}_{2h,\tilde\delta,\tilde\theta}[w_{2h}](x_{2h}) = \min_{\{\boldsymbol{w}_i\}_{i=1}^d \in \mathcal{V}_{\tilde\theta}} \left[\prod_{i=1}^d \left(\tilde\nabla^2_{\tilde\delta} \boldsymbol{w}_i \; w_{2h}(x_{2h}) \right)^+ - \sum_{i=1}^d \left(\tilde\nabla^2_{\tilde\delta} \boldsymbol{w}_i \; w_{2h}(x_{2h}) \right)^- \right], \tag{44}$$

where Ω^i_{2h} denotes the set of internal degrees of freedom of X^2_{2h}, which includes now the vertices and edge midpoints of $\mathcal{T}^2_{2h}$, and $\tilde\nabla^2_{\tilde\delta \boldsymbol{w}}$ is a more accurate, say using five points, discretization of the second derivative in direction $\boldsymbol{w}$ at scale $\tilde\delta$.

Following the ideas presented in Theorem 7 we can show that operator (44) is consistent with order $\mathcal{O}(\tilde\delta^{k+\alpha} + \frac{h^3}{\tilde\delta^2} + \tilde\theta^2)$; see (Nochetto and Ntogkas, 2018, Lemma 5.8). However, this scheme is *not* monotone. It will, instead serve as the two scale analogue of the accurate scheme (22).

By refining in a conforming way once $\mathcal{T}_{2h}^2$ we obtain $\mathcal{T}_h$, over which we can apply the two scale scheme of Section 2.7. Notice that there is a bijection between Ω_{2h}^i and Ω_h^i so that the elements of X_{2h}^2 and X_h can be compared by looking at their nodal values. In light of this observation we alleviate the notation and carry out the rest of the discussion using the scale h.

We combine (44) and (38) into a *filtered* two scale operator: for $w_h \in X_h$ and $x_h \in \Omega_h^i$

$$\mathrm{MA}^{\mathrm{F}}_{h,\delta,\theta,\tilde{\delta},\tilde{\theta}}[w_h](x_h) = \mathrm{MA}^{\mathrm{2S}}_{h,\delta,\theta}[w_h](x_h) + \tau\tilde{S}\left(\frac{\mathrm{MA}^{\mathrm{2Sq}}_{2h,\tilde{\delta},\tilde{\theta}}[w_h](x_h) - \mathrm{MA}^{\mathrm{2S}}_{h,\delta,\theta}[w_h](x_h)}{\tau}\right),$$

where $\tilde{S}(t) = \min\{S(t), 0\}$ and the function S is defined in (24). As explained in Nochetto and Ntogkas (2018, Section 2) the choice of filter function ensures discrete convexity in the case that the right-hand side degenerates, that is if $f(x_h) = 0$, for some $x_h \in \Omega_h^i$.

With these ingredients the filtered two scale scheme seeks for $u_h^F \in X_h$ such that

$$\mathrm{MA}^{\mathrm{F}}_{h,\delta,\theta,\tilde{\delta},\tilde{\theta}}[u_h^F](x_h) = f(x_h), \quad \forall x_h \in \Omega_h^i, \tag{45a}$$

$$u_h^F(x_h) = g(x_h), \quad \forall x_h \in \Omega_h^b. \tag{45b}$$

The theory of almost monotone schemes of Corollary 1 was combined with the convergence results of Section 2.7.4 in Nochetto and Ntogkas (2018, Section 6) to assert the convergence of any solution to (45).

2.8.3 Approximation of convex envelopes

Let us describe the results obtained in Li and Nochetto (2018b) regarding the approximation of the convex envelope of a function, which was introduced in Definition 10. Let $f \in C(\bar{\Omega})$. As shown in Oberman and Ruan (2017) the convex envelope $u = \Gamma f$ of f can be characterized as the viscosity solution of the problem

$$\mathrm{CE}[u](x) = 0, \quad x \in \Omega, \tag{46a}$$

$$u(x) = f(x), \quad x \in \partial\Omega, \tag{46b}$$

where the operator CE[·] is given by

$$\mathrm{CE}[w](x) = \min\left\{f(x) - u(x), \min\sigma(D^2 w(x))\right\}. \tag{47}$$

The intuition behind (46) is clear. First, we have that $u(x) \leq f(x)$ for every $x \in \bar{\Omega}$. In addition, if we define the *contact set*

$$\mathcal{C}(f) = \{x \in \bar{\Omega} : u(x) = f(x)\},$$

we obtain, upon denoting $\lambda_1(w) = \min\sigma(D^2 w(x))$, that for $x \in \mathcal{C}(f)$ we must have $\lambda_1(u) \geq 0$. On the other hand, if $x \notin \mathcal{C}(f)$, then we must have $\lambda_1(u) = 0$. In conclusion, u must be convex.

We remark, however, that problem (46) is very degenerate. Indeed, it can be shown, see for instance Li and Nochetto (2018b, Lemma 3.1), that if $\operatorname{dist}(x, \mathcal{C}(f)) \geq d\delta$ and $\boldsymbol{p} \in \partial u(x)$ there is $\boldsymbol{v} \in \mathbb{R}^d$ with $|\boldsymbol{v}| = 1$ such that

$$x_\pm = x \pm \delta \boldsymbol{v}, \quad u(x_\pm) = u(x) + \delta \boldsymbol{p} \cdot \boldsymbol{v}, \quad \nabla^2_{\delta \boldsymbol{v}} u(x) = 0, \quad \boldsymbol{p} \in \partial u(x_\pm).$$

In other words, if we are sufficiently far away from the contact set $\mathcal{C}(f)$, then the graph of u is flat in at least one direction. As a consequence, in general, the convex envelope cannot be arbitrarily smooth, regardless of the smoothness of the domain Ω and data f. Indeed, De Philippis and Figalli (2015) shows that if Ω is strictly convex with $\partial\Omega \in C^{3,1}$, and $f \in C^{3,1}(\bar{\Omega})$, then $u \in C^{1,1}(\bar{\Omega})$, and that this is optimal. This very low regularity is one of the main obstacles in the analysis of numerical schemes for (46).

Formulation (46) was already used for numerical purposes in Oberman (2008a) via wide stencil schemes like those presented in Section 2.3. Let us present here, instead, the two scale methods of Li and Nochetto (2018b). We will follow the notation of Section 2.7. In addition, if $\mathcal{S}$ denotes the unit ball in $\mathbb{R}^d$ we introduce, in full analogy to (37), a discretization $\mathcal{S}_\theta$ of $\mathcal{S}$ such that, for every $\boldsymbol{w} \in \mathcal{S}$ there is $\boldsymbol{w}^\theta \in \mathcal{S}_\theta$ that satisfies

$$|\boldsymbol{w} - \boldsymbol{w}^\theta| \leq \theta.$$

Over the space of piecewise linear functions X_h subordinated to the triangulation $\mathcal{T}_h$ we define

$$\mathrm{CE}_{h,\delta,\theta}[w_h](x_h) = \min\left\{ f(x_h) - w_h(x_h), \min_{\boldsymbol{w} \in \mathcal{S}_\theta} \nabla^2_{\delta \boldsymbol{w}} w_h(x_h) \right\} \tag{48}$$

where $w_h \in X_h$ and $x_h \in \Omega_h^i$. With the aid of this operator we define the discrete convex envelope of a function f as the function $u_h^\varepsilon \in X_h$ that solves

$$\mathrm{CE}_{h,\delta,\theta}[u_h^\varepsilon](x_h) = 0, \quad x_h \in \Omega_h^i, \tag{49a}$$

$$u_h^\varepsilon(x_h) = f(x_h), \quad x_h \in \Omega_h^b. \tag{49b}$$

The analysis of scheme (49) to a large extent follows that of two scale methods presented in Section 2.7. Namely, owing to discrete convexity we can show that the scheme has a comparison principle, from which uniqueness of solutions follows. The existence of solutions is obtained via a discrete Perron method, and the stability by noticing that $u_h^- = \mathcal{I}_h u$ and $u_h^+ = \mathcal{I}_h f$ are discrete sub- and supersolutions, respectively.

The considerations given above show that scheme (49) is monotone and stable. In addition, assuming smoothness of the arguments, one can show its consistency with similar arguments to those of Section 2.7.3. Upon realizing that the operator (47) has a comparison principle in the sense of Definition 5, this is enough to appeal to the theory of Section 2.1 and conclude that the scheme is convergent as $\varepsilon = (h, \delta, \theta) \to 0$, provided $\frac{h}{\delta} \to 0$.

The derivation of rates of convergence, however, requires special attention. This is due to the fact that, as stated above, the best regularity we can expect is $u \in C^{1,1}(\bar{\Omega})$, and this is not enough to exploit the consistency estimates that were used for convergence (which are applied to smooth test functions). To overcome this, one must take advantage of the flatness of the solution outside the contact set. To describe these results we must introduce some notation. Set, for $x_h \in \Omega_h^i$,

$$\delta_{x_h} = \min\{\delta, \operatorname{dist}(x_h, \partial\Omega)\}, \quad B_{x_h} = \bigcup_{T \in \mathcal{T}_h : \operatorname{dist}(x_h, T) < \delta_{x_h}} T,$$

and

$$W_{x_h} = \{x \in \bar{\Omega} : |x - x_h| \le d\delta\}.$$

The following is Li and Nochetto (2018b, Proposition 3.3).

Proposition 15 (consistency)**.**
Let Ω be strictly convex and u, the solution of (46) satisfy $u \in C^{k,\alpha}(\bar{\Omega})$ with $k = 0, 1$ and $\alpha \in (0, 1]$. For $x_h \in \Omega_h^i$ we have:

1. *If $dist(x_h, \mathcal{C}(f)) \ge d\delta$, then*

$$\min_{\boldsymbol{w} \in \mathcal{S}_\theta} \nabla^2_{\delta \boldsymbol{w}} \mathcal{I}_h u(x_h) \le C \left(\frac{(\delta\theta)^{k+\alpha} + h^{k+\alpha}}{\delta^2} \right) |u|_{C^{k,\alpha}(B_{x_h})}.$$

2. *If $dist(x_h, \mathcal{C}(f)) < d\delta$ but $dist(x_h, \partial\Omega) \ge d\delta$, then we have*

$$f(x_h) - u(x_h) \le C_k \delta^{k+\alpha},$$

where C_k depends on $|u|_{C^{0,\alpha}(W_{x_h})} + |f|_{C^{0,\alpha}(W_{x_h})}$ for $k = 0$ and on $|f|_{C^{1,\alpha}(W_{x_h})}$ for $k = 1$.

3. *If $0 < dist(x_h, \partial\Omega) < d\delta$, then for all $\boldsymbol{w} \in \mathcal{S}$ we have*

$$\nabla^2_{\delta \boldsymbol{w}} \mathcal{I}_h u(x_h) \le C \delta_i^{k+\alpha-2} |u|_{C^{k,\alpha}(B_{x_h})},$$

and the previous item also holds provided k $= 0$.

To take advantage of this result two new discrete barriers were constructed. One handles the first case, i.e., points sufficiently far away from the contact set. The other barrier handles points near the boundary, that is the third case in the previous result. Without going into details, we present here the main error estimate, and refer the reader to Li and Nochetto (2018b, Theorem 3.7).

Theorem 11 (convergence rate)**.**
Let Ω *be strictly convex,* u *be the viscosity solution of* (46)*, and* u_h^ε *the solution of* (49)*. If* $u \in C^{k,\alpha}(\bar{\Omega})$*, with* $k = 0, 1$ *and* $\alpha \in (0, 1]$*, then*

$$\|u - u_h^\varepsilon\|_{L^\infty(\Omega)} = \mathcal{O}\left(h^{\frac{(k+\alpha)^2}{k+\alpha+2}}\right),$$

provided, $\delta = \mathcal{O}\left(h^{\frac{k+\alpha}{k+\alpha+2}}\right)$ *and* $\theta = \mathcal{O}\left(h^{\frac{2}{k+\alpha+2}}\right)$*. In particular, if* $k = \alpha = 1$*, i.e.,* $u \in C^{1,1}(\bar{\Omega})$*, we obtain*

$$\delta = \mathcal{O}(h^{1/2}), \quad \theta = \mathcal{O}(h^{1/2}) \quad \Longrightarrow \quad \|u - u_h^\varepsilon\|_{L^\infty(\Omega)} = \mathcal{O}(h).$$

Similarly, if $k = 0$ *and* $\alpha = 1$*, i.e.,* $u \in C^{0,1}(\bar{\Omega})$*, we get*

$$\delta = \mathcal{O}(h^{1/3}), \quad \theta = \mathcal{O}(h^{2/3}) \quad \Longrightarrow \quad \|u - u_h^\varepsilon\|_{L^\infty(\Omega)} = \mathcal{O}(h^{1/3}).$$

2.8.4 The Gauss curvature problem

As an application of the wide stencil finite difference schemes that were presented in Section 2.3 let us here, following Hamfeldt (2018), describe a discretization of the prescribed Gaussian curvature problem (2). To do so, we must begin by defining what is a solution of this problem. In a similar manner to the notion of Alexandrov solutions to Monge–Ampère problem, introduced in Definition 9, we have

Definition 17 (generalized solution)**.**
A convex function $u : \bar{\Omega} \to \mathbb{R}$ is a generalized solution to (2) if the following two conditions hold:

1. It is a generalized solution of (2a). This means that, for all Borel sets $D \subset \Omega$, we have

$$\int_{\partial u(D)} \frac{1}{\left(1 + |\boldsymbol{p}|^2\right)^{(d+2)/2}} \mathrm{d}\boldsymbol{p} = \int_D \mathcal{K}(x)\mathrm{d}x.$$

2. It satisfies

$$\limsup_{y \to x} u(y) \leq g(x), \quad \forall x \in \partial\Omega$$

and, if v is any other generalized solution of (2a), then $v \leq u$ in Ω.

Under the assumptions of uniform convexity of Ω; continuity of g; continuity, boundedness, and nonnegativity of $\mathcal{K}$; and the compatibility condition

$$\int_{\mathbb{R}^d} \frac{1}{\left(1 + |\boldsymbol{p}|^2\right)^{(d+2)/2}} \mathrm{d}\boldsymbol{p} > \int_\Omega \mathcal{K}(x)\mathrm{d}x;$$

it can be shown that problem (2) has a unique generalized solution; see Bakelman (1994).

It is also possible to extend the notion of viscosity solution presented in Definition 4, by allowing the operators in Definition 2 to also depend on a variable $\boldsymbol{p} \in \mathbb{R}^d$. In doing that, we note that the operator $F_{G\mathcal{K},c} : \bar{\Omega} \times \mathbb{R} \times \mathbb{R}^d \times \mathbb{S}^d \to \mathbb{R}$ defined by

$$F_{G\mathcal{K},c}(x,r,\boldsymbol{p},M) = \begin{cases} \det M - \mathcal{K}(x)\left(1+|\boldsymbol{p}|^2\right)^{(d+2)/2}, & x \in \Omega, \\ g(x) - r, & x \in \partial\Omega, \end{cases}$$

is, as the Monge–Ampère operator F_{MA} defined in (8), only elliptic when $M \in \mathbb{S}^d_+$, which implies that to have a reasonable notion of viscosity solution, we must require sub- and supersolutions to be convex, and restrict the test functions to be convex, as in Definition 6. As we have seen throughout our discussion, the convexity constraint is rather difficult to impose explicitly during discretization.

Hamfeldt (2018) proposed to consider the following formulation of (2). If for $M \in \mathbb{S}^d$ we set $\sigma(M) = \{\lambda_1(M), \ldots, \lambda_d(M)\}$, where the eigenvalues are counted with multiplicity and arranged in nondecreasing order, then the operator

$$F_{G\mathcal{K}}(x,r,\boldsymbol{p},M) = \begin{cases} F^{\text{in}}_{G\mathcal{K}}(x,\boldsymbol{p},M), & x \in \Omega, \\ g(x) - r, & x \in \partial\Omega, \end{cases}$$

with

$$F^{\text{in}}_{G\mathcal{K}}(x,\boldsymbol{p},M) = \min\left\{\lambda_1(M),\ \prod_{i=1}^{d}\lambda_i(M)^+ - \mathcal{K}(x)\left(1+|\boldsymbol{p}|^2\right)^{(d+2)/2}\right\}$$

is elliptic in the sense of Definition 2 and, at least formally, it is clear that if

$$F_{G\mathcal{K}}(x,u(x),\nabla u(x),D^2u(x)) = 0, \quad x \in \Omega,$$

then we must have, that either, $\lambda_1(D^2u(x)) > 0$, so that u is convex, and

$$\det D^2u(x) = \mathcal{K}(x)\left(1+|\nabla u(x)|^2\right)^{(d+2)/2},$$

or $\lambda_1(D^2u(x)) = 0$ and, thus

$$0 = \det D^2u(x) \geq \mathcal{K}(x)\left(1+|\nabla u(x)|^2\right)^{(d+2)/2} \geq 0.$$

In either case, the convexity of the solution is recovered.

With these constructions we have two options to define viscosity solutions to (2). The first one is, like in Definition 6, to require that it is a viscosity solution, in the set of convex functions, of the problem

$$F_{G\mathcal{K},c}(x,u(x),\nabla u(x),D^2u(x))=0,\quad \forall x\in\bar{\Omega}. \tag{50}$$

The second, as in Definition 4, to require that it is a viscosity solution of

$$F_{G\mathcal{K}}(x,u(x),\nabla u(x),D^2u(x))=0,\quad \forall x\in\bar{\Omega}. \tag{51}$$

In full analogy to Proposition 3 it is shown in Hamfeldt (2018, Section 3) that viscosity subsolutions to problem (51) are convex and that a function is a viscosity solution of (50) over the set of convex function if and only if it is a viscosity solution of (51). In addition it is shown that, under certain assumptions on $\mathcal{K}$, this notion of solution, at least in the interior of the domain Ω, coincides with that of Definition 17.

It is important to note that incorporating the boundary conditions into the definition of the operator is *essential* in this problem, as they may not be realized in a classical sense. The following is Hamfeldt (2018, Example 1).

Example 2 (nonclassical boundary conditions)**.**
Let $d=1$, $\Omega=(0,1)$, and $\mathcal{K}\equiv 1$. We set the boundary conditions $u(0)=-1$ and $u(1)=1$. Then it is possible to show that

$$u(x)=-\sqrt{1-x^2}$$

is a viscosity solution of (51). It is a classical solution over $[0,1)$ so it remains to understand what happens at $x=1$.

Note that $u'(x)$ grows unboundedly as $x\uparrow 1$ so that it is not possible to find a smooth φ such that $u_\star-\varphi$ has a local minimum at x_0, in other words, the graph of u cannot be touched from below at $x=1$. This makes u automatically a supersolution.

To show that u is also a subsolution we note that $u(1)=0<1$ so that, if φ touches the graph of u from above at $x=1$, we must have $\varphi(1)=u(1)=0$, and

$$(F_{G\mathcal{K}})_\star(1,u(1),\varphi'(1),\varphi''(1))\geq 1-u(1)=1>0.$$

The behaviour of Example 2 was characterized in Hamfeldt (2018, Corollary 24). Namely, if u is a viscosity solution of (50) then at every $x\in\partial\Omega$ we either have that $u_\star(x)=u^\star(x)=g(x)$, or $u_\star(x)\leq u^\star(x)\leq g(x)$ and $\partial u_\star(x)=\varnothing$. The second option here corresponds to the right endpoint in Example 2.

Existence of solutions to (51) was shown using a variant of Perron's method. The usual argument to show uniqueness is obtained via a comparison principle of Definition 5. This problem, however, does not have a comparison principle, as Hamfeldt (2018, Example 3) shows.

Example 3 (lack of comparison)**.**
In the setting of Example 2 we have that $u(x)=-\sqrt{1-x^2}$ is a viscosity solution, so that necessarily it is a supersolution. Let

$$v(x) = \begin{cases} u(x), & x \in [0,1), \\ 1, & x = 1, \end{cases}$$

we see that $v \in USC([0, 1])$ and, as in Example 2, if φ touches from above the graph of u at $x = 1$, then $\varphi(1) = v(1) = 1$ and

$$(F_{GK})_\star(1, v(1), \varphi'(1), \varphi''(1)) \geq 1 - v(1) = 0,$$

showing that v is a subsolution. Note, however, that $u(1) \leq v(1)$ and this problem does not have a comparison principle.

The previous result, combined with the behaviour of solutions at the boundary shows that, in fact, a comparison principle takes place, but only in the interior of the domain; see Hamfeldt (2018, Theorem 7).

Theorem 12 (interior comparison)**.**
Let $\underline{u} \in USC(\bar{\Omega})$ be a subsolution of (51) and $\overline{u} \in LSC(\bar{\Omega})$ a supersolution. Then we have $\underline{u} \leq \overline{u}$ in Ω.

This weakened comparison principle is sufficient to guarantee uniqueness.

Having shown the existence and uniqueness of solutions to (51), it is possible now to construct numerical schemes. This is carried using variants of the wide stencil finite difference schemes of Section 2.3. With the notation introduced there we define, for $w_h \in X_h$,

$$\begin{aligned} \mathrm{GK}_{h,\theta}[w_h](x_h) = \min \Bigg\{ & \min_{\{\boldsymbol{\nu}_i\}_{i=1}^d \in \mathcal{G}_\theta} \Delta_{\boldsymbol{\nu}_i} w_h(x_h), \quad \mathrm{MA}^{\mathrm{WS}}_{h,\theta}[w_h](x_h) \\ & - \mathcal{K}(x_h) \left(1 + |\nabla_h w_h(x_h)|^2\right)^{(d+2)/2} \Bigg\}, \end{aligned}$$

where $\mathrm{MA}^{\mathrm{WS}}_{h,\theta}[\,\cdot\,]$ was defined in (19) and the vector $\nabla_h w_h(x_h)$ is such that

$$\begin{aligned} \nabla_h w_h(x_h) \cdot \boldsymbol{e}_i = \max \Bigg\{ & \frac{w_h(x_h) - w_h(x_h - h\boldsymbol{e}_i)}{h}, \\ & \frac{w_h(x_h) - w_h(x_h + h\boldsymbol{e}_i)}{h}, 0 \Bigg\}, \end{aligned} \tag{52}$$

and $\{\boldsymbol{e}_i\}_{i=1}^d$ is the canonical basis of $\mathbb{R}^d$. With this operator, the finite difference approximation of (51) is to find $u_h \in X_h$ such that

$$\mathrm{GK}_{h,\theta}[u_h](x_h) = 0, \quad x_h \in \Omega_h^i, \tag{53a}$$

$$u(x_h) = g(x_h), \quad x_h \in \Omega_h^b. \tag{53b}$$

In Hamfeldt (2018, Section 6) it is shown that scheme (53) is monotone, in the sense of (12), stable, in the sense of (13), and consistent, in the sense of (14).

Notice, however, that as Example 3 shows, problem (51) does not have a comparison principle. As a consequence, Theorem 4 cannot be applied. For this reason, the framework of Section 2.1 was extended in Hamfeldt (2018, Theorem 9) to cases where problem (7) only has an interior comparison principle like that of Theorem 12 and there exist classical sub- and supersolutions. The conclusion is the locally uniform convergence of u_h to u.

2.8.5 Transport boundary conditions

Let us conclude the discussion of wide stencil finite difference schemes by describing how these methods can be used to tackle the optimal transportation problem. Since this will be one of the main topics of chapter "Optimal transport" by Merigot in this volume, we shall be brief.

We recall that, given $\Omega, \mathcal{O} \subset \mathbb{R}^d$, which we assume bounded, with $\mathcal{O}$ convex, and measures $\rho_\Omega : \Omega \to \mathbb{R}$ and $\rho_\mathcal{O} : \mathcal{O} \to \mathbb{R}$, the optimal transportation problem (with quadratic cost) seeks for a map $T : \Omega \to \mathcal{O}$ with $T_\sharp \rho_\Omega = \rho_\mathcal{O}$ that minimizes

$$\frac{1}{2}\int_\Omega |x - T(x)|^2 \mathrm{d}\rho_\Omega(x).$$

We recall that $T_\sharp \mu$ denotes the pushforward of the measure μ under the mapping T. Assuming that the measures are absolutely continuous with respect to Lebesgue measure, with densities $f_\Omega, f_\mathcal{O}$, this condition can be written as

$$\int_E f_\mathcal{O}(x)\mathrm{d}x = \int_{T^{-1}(E)} f_\Omega(x)\mathrm{d}x,$$

and so by a change of variables, $\det(\nabla T(x)) f_\mathcal{O}(T(x)) = f_\Omega(x)$. Finally, we recall that since the cost is quadratic, it can be shown that T is given by the gradient map of a convex potential $u : \Omega \to \mathbb{R}$. This allows us to, at least at the formal level, rewrite the optimal transportation problem as a Monge–Ampère problem: find $u : \bar{\Omega} \to \mathbb{R}$ convex, such that

$$\det D^2 u(x) = F(x, \nabla u(x)), \quad x \in \Omega. \tag{54}$$

where we set $F(x, \boldsymbol{p}) = \rho_\Omega(x)/\rho_\mathcal{O}(\boldsymbol{p})$. This problem is supplemented by the so-called *transport* or *second* boundary condition

$$\nabla u(\bar{\Omega}) = \bar{\mathcal{O}}.$$

Notice that this, more than a boundary condition, is a set of constraints. It can be shown also that this condition can be replaced by

$$\nabla u(\partial\Omega) = \partial\mathcal{O}. \tag{55}$$

Thus, we want to construct numerical schemes to approximate the solution of (54) and (55).

It is clear that the main issue is the discretization of the boundary condition (55). If the boundary of the domain $\mathcal{O}$ is given as the zero level set of some function $\Phi : \mathbb{R}^d \to \mathbb{R}$, then it is clear that (55) can be equivalently written as

$$\Phi(\nabla u(x)) = 0, \quad \forall x \in \partial\Omega.$$

While we would be tempted to discretize this condition directly, the function Φ can be highly nonlinear and nonsmooth, which will make the design of monotone and consistent numerical schemes a daunting task. However, this can be achieved very easily if the domains are rectangles, say $\Omega = (0,1)^2 = \mathcal{O}$. In this case, it is shown in Froese (2012, Section 3.2) that each side must be mapped to itself. If we consider the left side of the square, that is,

$$\{(x_1, x_2) \in \mathbb{R}^2 : x_1 = 0, x_2 \in [0,1]\},$$

then the function that describes this is given by $\Phi(y_1, y_2) = y_1$. Thus, on this side we can write

$$\frac{\partial u(0, x_2)}{\partial x_1} = 0.$$

Similarly, in the right, bottom and top sides, respectively, we can write

$$\frac{\partial u(1, x_2)}{\partial x_1} = 1, \quad \frac{\partial u(x_1, 0)}{\partial x_2} = 0, \quad \frac{\partial u(x_1, 1)}{\partial x_2} = 1.$$

It is remarkable that on all sides the derivative that appears is actually the normal derivative. This motivated Froese (2012) to replace the boundary condition (55) by a Neumann-type boundary condition

$$\frac{\partial u(x)}{\partial \boldsymbol{n}} = \phi(x)$$

for some unknown function ϕ.

Obviously, the correct choice of function ϕ is $\phi(x) = \nabla u(x) \cdot \boldsymbol{n}(x)$, which motivates the introduction of the following iterative scheme: Given u_0, an initial guess, then

- For $k \geq 0$
  - Define, for $x \in \partial\Omega$,

$$\boldsymbol{p}_k(x) = \mathrm{Proj}_{\partial\mathcal{O}}(\nabla u_k(x)), \tag{56}$$

  where by $\mathrm{Proj}_S(\boldsymbol{w})$ we denoted a projection of the vector $\boldsymbol{w}$ onto the set S.

– Find $u_{k+1}:\bar{\Omega}\to\mathbb{R}$ convex, and $c_{k+1}\in\mathbb{R}$ that satisfy

$$\int_\Omega u_{k+1}(x)\mathrm{d}x=0, \tag{57a}$$

$$\det D^2 u_{k+1}(x)=c_{k+1}F(x,\nabla u_{k+1}(x)),\quad x\in\Omega, \tag{57b}$$

$$\frac{\partial u_{k+1}(x)}{\partial \boldsymbol{n}}=\boldsymbol{p}_k(x),\quad x\in\partial\Omega. \tag{57c}$$

– Set $k\leftarrow k+1$

- EndFor

Remark 14 (iterative scheme).
The iterative scheme (56) and (57) deserves several observations.

1. The introduction of the projection $\boldsymbol{p}_k$ in step (56) is due to the fact that there is no reason to expect that $\nabla u_k(\partial\Omega)\subset\partial\mathcal{O}$. Thus, we settle for the closest point on the target boundary.
2. Problem (57) is a Neumann problem for an elliptic equation so that the solution, if it exists, is unique only up to a constant. Condition (57a) forces uniqueness, while the introduction of the number c_{k+1} in (57b) relaxes the equation so that the necessary conditions for existence are fulfilled.
3. The initialization of this scheme can done by choosing $\boldsymbol{p}_0=Mx\cdot\boldsymbol{n}$, where $\boldsymbol{n}$ is the unit outer normal to $\partial\Omega$ and $M>0$ is so large that the image of the mapping $\bar{\Omega}\ni x\mapsto Mx\in\mathbb{R}^d$ contains $\bar{\mathcal{O}}$. ■

We are then going to discretize (56) and (57). Notice that now the boundary conditions (57c) are rather standard and can be approximated by, for instance, introducing a layer of ghost nodes near the boundary and computing centred differences.

It remains to discretize (57b). Setting $v_h=u_{h,\,k+1}$, the first alternative, proposed in Froese (2012), is to use

$$\mathrm{MA}^{\mathrm{WS}}_{h,\theta}[v_h](x_h)=F(x_h,\nabla_h v_h),\quad x_h\in\Omega^i_h, \tag{58}$$

where $\mathrm{MA}^{\mathrm{WS}}_{h,\theta}[\,\cdot\,](x_h)$ is the operator defined in (19) or Remark 2, and $\nabla_h v_h$ is defined as in (52). Another option, also from Froese (2012), is to take advantage of the directional difference that are already being computed to approximate the Monge–Ampère operator. Notice that if $\{\boldsymbol{\nu}_i\}_{i=1}^d\in\mathcal{V}$, we have that

$$\nabla w=\left(\frac{\partial w}{\partial x_1},\ldots,\frac{\partial w}{\partial x_d}\right)^{\mathsf{T}}=\left(\sum_{i=1}^d\frac{\partial w}{\partial \boldsymbol{\nu}_i}\boldsymbol{\nu}_i\cdot\boldsymbol{e}_1,\ldots,\sum_{i=1}^d\frac{\partial w}{\partial \boldsymbol{\nu}_i}\boldsymbol{\nu}_i\cdot\boldsymbol{e}_d\right)^{\mathsf{T}},$$

where, as usual, $\{\boldsymbol{e}_i\}_{i=1}^d$ is the canonical basis of $\mathbb{R}^d$. This allows us to write that

$$
\begin{aligned}
\det D^2 w(x) - F(x, \nabla w(x)) &= \mathrm{MA}[w](x) - F(x, \nabla w(x)) \\
&= \min_{\{\boldsymbol{w}_i\}_{i=1}^d \in \mathcal{V}} \left[\prod_{i=1}^d \left(\frac{\partial^2 w(x)}{\partial \boldsymbol{w}_i^2} \right)^+ - \sum_{i=1}^d \left(\frac{\partial^2 w(x)}{\partial \boldsymbol{w}_i^2} \right)^- \right] \\
&= \min_{\{\boldsymbol{w}_i\}_{i=1}^d \in \mathcal{V}} \left[\prod_{i=1}^d \left(\frac{\partial^2 w(x)}{\partial \boldsymbol{w}_i^2} \right)^+ - \sum_{i=1}^d \left(\frac{\partial^2 w(x)}{\partial \boldsymbol{w}_i^2} \right)^- - F(x, \nabla w(x)) \right] \\
&= \min_{\{\boldsymbol{w}_i\}_{i=1}^d \in \mathcal{V}} \left[\prod_{i=1}^d \left(\frac{\partial^2 w(x)}{\partial \boldsymbol{w}_i^2} \right)^+ - \sum_{i=1}^d \left(\frac{\partial^2 w(x)}{\partial \boldsymbol{w}_i^2} \right)^- \right. \\
&\quad \left. - F\left(x, \left(\sum_{i=1}^d \frac{\partial w(x)}{\partial \boldsymbol{\nu}_i} \frac{\boldsymbol{\nu}_i \cdot \boldsymbol{e}_1}{|\boldsymbol{\nu}_i|}, \ldots, \sum_{i=1}^d \frac{\partial w(x)}{\partial \boldsymbol{\nu}_i} \frac{\boldsymbol{\nu}_i \cdot \boldsymbol{e}_d}{|\boldsymbol{\nu}_i|} \right)^\mathsf{T} \right) \right] \\
&= \min_{\{\boldsymbol{w}_i\}_{i=1}^d \in \mathcal{V}} \mathrm{OT}_{\{\boldsymbol{w}_i\}_{i=1}^d}[w](x).
\end{aligned}
$$

In conclusion, an approximation of (57b) is obtained by setting

$$
\mathrm{OT}_{h,\theta}[v_h](x_h) = 0, \quad \forall x_h \in \Omega_h^i,
$$

where

$$
\mathrm{OT}_{h,\theta}[w_h](x_h) = \min_{\{\boldsymbol{\nu}_i\}_{i=1}^d \in \mathcal{G}\theta} \mathrm{OT}_{\{\boldsymbol{\nu}_i\}_{i=1}^d}[w_h](x_h).
$$

Benamou et al. (2014) considered a different treatment of the boundary condition (55). Since for all $x \in \partial\Omega$ we must have that $\nabla u(x) \in \partial\mathcal{O}$, then we must have

$$
H(\nabla u(x)) = 0, \quad H(y) = \begin{cases} \mathrm{dist}(y, \partial\mathcal{O}), & y \in \mathcal{O}, \\ -\mathrm{dist}(y, \partial\mathcal{O}), & y \notin \mathcal{O}, \end{cases} \tag{59}
$$

where H is nothing but the signed distance function to $\partial\mathcal{O}$. Notice that (59) is a sort of Hamilton Jacobi equation posed on $\partial\Omega$. Exploiting the convexity of $\mathcal{O}$, the authors of Benamou et al. (2014) were able to rewrite the function H as the supremum over linear expressions on y (the supporting hyperplanes of $\mathcal{O}$ at y)

$$
H(y) = \sup_{\boldsymbol{n} \in \mathbb{R}^d : |\boldsymbol{n}| = 1} \{ y \cdot \boldsymbol{n} - H^\star(\boldsymbol{n}) : \boldsymbol{n} \cdot \boldsymbol{n}_x > 0 \},
$$

where $\boldsymbol{n}_x$ is the normal to $\partial\Omega$ at x and $H^\star$ is the support function of $\mathcal{O}$, that is,

$$
H^\star(\boldsymbol{n}) = \sup_{z \in \partial\mathcal{O}} z \cdot \boldsymbol{n}.
$$

This function can be precomputed or evaluated rather cheaply in the discrete setting. The reformulation of the function H can be approximated by replacing the supremum by one over a finite set of directions, and, finally, the gradient appearing in (59) can be discretized as in (52). This gives a discretization of (55). Finally, the discretization of (54) is proposed to be carried similarly to (58).

3 Discretizations based on geometric considerations

In fact, geometrical representations, graphs and diagrams of all sorts, are used in all sciences, not only in physics, chemistry, and the natural sciences, but also in economics, and even in psychology. Using some suitable geometrical representation, we try to express everything in the language of figures, to reduce all sorts of problems to problems of geometry.

Pólya (2014)

In this section we will describe the so-called Oliker–Prussner method, which is a discrete analogue of the notion of solution in the Alexandrov sense. We recall that Alexandrov solutions to the Monge–Ampère equation were introduced in Definition 9. They make a connection between the Monge–Ampère equation and the measure of the subdifferential of its solution. This, very geometric, notion enables us to define solutions that are not smooth, say not $C^2(\Omega)$. The Oliker–Prussner method, in turn, will allow us to approximate these solutions.

3.1 Description of the scheme

To be able to present the Oliker–Prussner method, we must begin by introducing some useful notions.

3.1.1 Nodal set and domain partition

To discretize the domain Ω and its boundary $\partial\Omega$, we introduce a translation invariant nodal set and an open, disjoint partition of the domain. For a parameter $h > 0$, we define the interior nodal set as

$$\Omega_h = \left\{ x_h = h\sum_{j=1}^{d} z^j \tilde{\boldsymbol{e}}_j : z^j \in \mathbb{Z} \right\} \cap \Omega, \tag{60}$$

where $\{\tilde{\boldsymbol{e}}_j\}_{j=1}^d$ is a basis of $\mathbb{R}^d$ with $|e_j| \leq 1$ for all $1 \leq j \leq d$. To discretize the boundary $\partial\Omega$, we set the boundary nodal set $\partial\Omega_h$ as a collection of points on the boundary and require that their spacing is of order h, namely, $\partial\Omega \subset \cup_{x_h \in \partial\Omega_h} B_{h/2}(x_h)$. We set the nodal set $\bar{\Omega}_h = \Omega_h \cup \partial\Omega_h$. We remark that this is a generalization of the finite difference discretizations introduced in Section 2.3. Indeed, in that case the vectors $\{\tilde{\boldsymbol{e}}_j\}_{j=1}^d$ were the canonical basis of $\mathbb{R}^d$.

We define an open, disjoint partition $\{\omega_{x_h}\}_{x_h \in \bar{\Omega}_h}$ of the domain where, for $x_h \in \bar{\Omega}_h$,

$$\omega_{x_h} = \left\{ x_h + \sum_{j=1}^{d} h^j \tilde{\boldsymbol{e}}_j : \quad h^j \in \mathbb{R}, \quad |h^j| \leq \frac{h}{2} \right\} \cap \Omega. \tag{61}$$

Note that, by construction, the partition is translation invariant, that is, $\omega_{y_h} = y_h - x_h + \omega_{x_h}$ for all x_h, $y_h \in \Omega_h$ with $\omega_{y_h}, \omega_{x_h} \subset \Omega$.

3.1.2 Nodal functions, their subdifferentials, and convex envelopes

On the nodal set Ω_h constructed above, we define a nodal function u_h to approximate the solution of the Monge–Ampère PDE. First, to mimic the convexity constraint for the PDE, we require the notion of convexity for nodal functions (compare to Definition 16).

Definition 18 (nodal convexity)**.**
Let w_h be a (nodal) function that maps the set of nodes $\bar{\Omega}_h$ to $\mathbb{R}$. We say that w_h is *convex* if, for any node $x_h \in \bar{\Omega}_h$, there exist an affine function L, that is, $L(x) = \boldsymbol{p} \cdot (x - x_h) + c$ for some $\boldsymbol{p} \in \mathbb{R}^d$ and $c \in \mathbb{R}$, such that

$$L(y_h) \leq w_h(y_h) \quad \forall y_h \in \bar{\Omega}_h \quad \text{and} \quad L(x_h) = w_h(x_h). \tag{62}$$

We define the subdifferential of a convex nodal function w_h at a fixed node $x_h \in \bar{\Omega}_h$ as the set

$$\partial w_h(x_h) := \{\boldsymbol{p} \in \mathbb{R}^d : \ \boldsymbol{p} \cdot (y_h - x_h) + w_h(x_h) \leq w_h(y_h) \quad \forall y_h \in \bar{\Omega}_h\}. \tag{63}$$

In other words, this is the collection of slopes of affine functions that satisfy the condition that defines convexity for a nodal function. Note that nodal functions are only defined on $\bar{\Omega}_h$. To extend a nodal function to the domain Ω, we introduce its convex envelope.

Definition 19 (convex envelope of a nodal function)**.**
Let w_h be a nodal function defined on $\bar{\Omega}_h$. The *convex envelope* of w_h is the piecewise linear function

$$\Gamma(w_h)(x) = \sup_L \{L(x) : \ L \text{affine function and} L(x_h) \leq w_h(x_h) \quad \forall x_h \in \bar{\Omega}_h\}$$

for any $x \in \Omega$.

We note that, by definition, $\Gamma(w_h)(x_h) \leq w_h(x_h)$ for any node $x_h \in \bar{\Omega}_h$, and equality holds for all interior nodes if w_h is convex. Indeed, if w_h is convex, by (62), for any node $x_h \in \bar{\Omega}_h$, there exists an affine function $L(x)$ satisfying

$$L(y_h) \leq w_h(y_h) \quad \forall y_h \in \bar{\Omega}_h \quad \text{and} \quad L(x_h) = w_h(x_h).$$

Since $L(x) \leq \Gamma(w_h)(x)$ for any $x \in \Omega$ by Definition 19, we deduce that $w_h(x_h) = L(x_h) \leq \Gamma(w_h)(x_h)$. Combining this inequality with the inequality in the other direction, we have $w_h(x_h) = \Gamma(w_h)(x_h)$ for all interior nodes. Thus, $\Gamma(w_h)$ is a natural extension to $\bar{\Omega}$ of the convex nodal function w_h. With an abuse of notation, we still use w_h to denote the convex envelope of this nodal function.

The convex envelope of a nodal function w_h induces a triangulation of the domain Ω and a piecewise linear function over this triangulation. However, this triangulation is not known a priori. Here we give an example to illustrate this property.

Example 4 (convex envelope and triangulation)**.**
Define the nodal set $\bar{\Omega}_h = \{z_1, \ldots, z_5\}$ with $z_1 = (1, 0)$, $z_2 = (0, 1)$, $z_3 = (-1, 0)$, $z_4 = (0, -1)$, and $z_5 = (0, 0)$. Consider the nodal functions satisfying

$$w_1(z_1) = w_1(z_3) = 1, \quad w_2(z_2) = w_2(z_4) = 1,$$
$$w_3(z_1) = w_3(z_2) = w_3(z_3) = w_3(z_4) = 1,$$

and $w_j(z_i) = 0$ otherwise. The convex envelopes are $\Gamma(w_1) = |x_1|$, $\Gamma(w_2) = |x_2|$, and $\Gamma(w_3) = |x_1| + |x_2|$. The convex envelopes are subordinate to the meshes depicted in Fig. 4. ■

The above example shows that $\Gamma(w_h)$ is a piecewise linear function that induces a mesh $\mathcal{T}_h$ that depends on the values of w_h. The example depicted in Fig. 5 shows that, if w_h is the nodal interpolant of a function w, and if the Hessian D^2w is degenerate (or nearly degenerate), the induced mesh may be anisotropic.

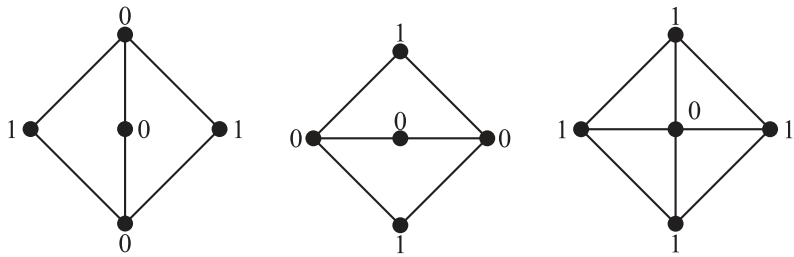

FIG. 4 Meshes corresponding to convex envelopes $\Gamma(w_1) = |x_1|$ (*left*), $\Gamma(w_2) = |x_2|$ (*middle*), and $\Gamma(w_3) = |x_1| + |x_2|$ (*right*).

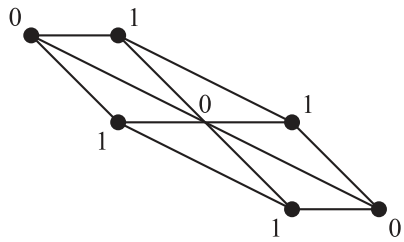

FIG. 5 Mesh induced by the nodal interpolant of $w(x) = (x \cdot e)^2$ where $e = (1,2)^{\mathsf{T}}$. Its convex envelope equals $|x \cdot e|$ in the star of (0, 0).

3.1.3 The Oliker–Prussner method

Now we are ready to introduce the Oliker–Prussner method (Nochetto and Zhang, 2019; Oliker and Prussner, 1988). We seek a convex nodal function u_h satisfying the boundary condition $u_h(x_h) = g(x_h)$ for all $x_h \in \partial\Omega_h$ and

$$|\partial u_h(x_h)| = \int_{\omega_{x_h}} f(x)\mathrm{d}x, \quad \forall x_h \in \Omega_h, \tag{64}$$

Note that, since the partition $\{\omega_{x_h}\}_{x_h\in\Omega_h}$ is nonoverlapping, for all Borel sets $D \subset \Omega$, we have

$$|\partial u_h(D)| = \sum_{x_h\in D} f_{x_h}, \quad \text{where} f_{x_h} = \int_{\omega_{x_h}} f(x)\mathrm{d}x.$$

Thus, the scheme is obtained by replacing f in (9) by a family of Dirac measures supported at the nodes in Ω_h, and by replacing g by its nodal interpolant on the boundary. To implement the method, we need to derive a formula to compute the subdifferential of a nodal function u_h. This is a nontrivial task because it is non local. In fact, it involves computing the convex envelope of u_h. The following observation is useful in the characterization of the subdifferential. For a proof, see Nochetto and Zhang (2018).

Lemma 8 (characterization of subdifferential)**.**
Let w_h be a convex nodal function and $\mathcal{T}_h$ be the mesh induced by its convex envelope $\Gamma(w_h)$. Then the subdifferential of w_h at $x_h \in \Omega_h$ is the convex hull of the constant gradients $\nabla\Gamma(w_h)|_T$ for all $T \in \mathcal{T}_h$ which contain x_h.

Fig. 6 depicts the subdifferential $\partial w_h(x_h)$ of a convex nodal function w_h at node x_h for $d = 2$.

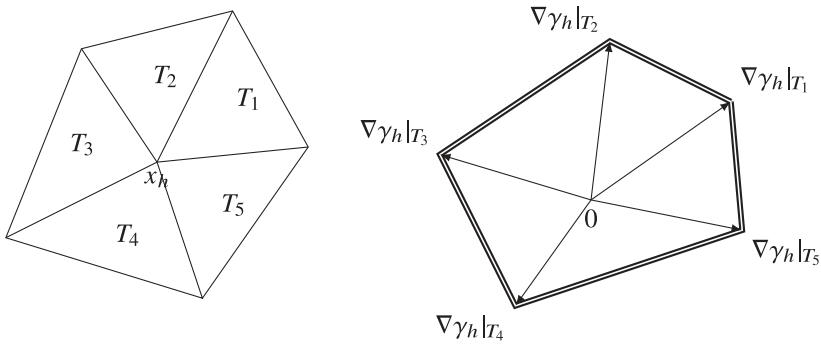

FIG. 6 Star centred at node x_h corresponding to the mesh $\mathcal{T}_h$ induced by the convex envelope $\gamma_h = \Gamma(w_h)$ and subdifferential $\partial w_h(x_h)$ of the convex nodal function w_h at node x_h. The latter is the convex hull of the constant element gradients $\nabla\gamma_h|_{T_j}$ for $1 \le j \le 5$.

3.2 Stability, continuous dependence on data, and discrete maximum principle

The Alexandrov estimate, which establishes the stability and continuous dependence of the Monge–Ampère equation, is a cornerstone in the nonlinear PDE theory. In this subsection, we introduce a discrete version of the Alexandrov estimate suitable for nodal functions. We refer the reader to Nochetto and Zhang (2018) for a complete proof.

Lemma 9 (discrete Alexandrov estimate)**.**
Let w_h be a nodal function with $w_h(x_h) \geq 0$ at all $x_h \in \partial\Omega_h$. Then

$$\sup_{\Omega_h} w_h^- \leq C \left(\sum_{x_h \in \mathcal{C}_h^-(w_h)} |\partial w_h(x_h)| \right)^{1/d}, \tag{65}$$

where $C = C(d, \Omega)$ is proportional to the diameter of Ω and $\mathcal{C}_h^-(w_h)$ is the (lower) contact set:

$$\mathcal{C}_h^-(w_h) := \{x_h \in \Omega_h, \ \ \Gamma(w_h)(x_h) = w_h(x_h)\}. \tag{66}$$

The Alexandrov estimate establishes a lower bound for a nodal function in terms of the measure of the subdifferential at the (lower) contact set. Similarly, one can obtain an upper bound for a nodal function by the measure of the superdifferential at the (upper) contact set.

Applying the discrete Alexandrov estimate, we are ready to compare two arbitrary nodal functions in terms of their subdifferentials. This is instrumental for the error analysis.

Proposition 16 (stability of numerical solution)**.**
Let v_h and w_h be two nodal functions with $v_h \geq w_h$ on $\partial\Omega_h$. Then

$$\sup_{\Omega_h} (v_h - w_h)^- \leq C \left(\sum_{x_h \in \mathcal{C}_h^-(v_h - w_h)} \left(|\partial v_h(x_h)|^{1/d} - |\partial w_h(x_h)|^{1/d} \right)^d \right)^{1/d},$$

where $C = C(d, \Omega)$ is proportional to the diameter of Ω.

Proof. Let v_h, w_h be two nodal functions. We introduce the convex envelope $\Gamma(v_h - w_h)$ as in Definition 19, and the nodal contact set $\mathcal{C}_h^-(v_h - w_h)$ defined in (66). The discrete Alexandrov estimate of Lemma 9 yields

$$\sup_{\Omega_h} (v_h - w_h)^- \leq C \left(\sum_{x_h \in \mathcal{C}_h^-(v_h - w_h)} |\partial\Gamma(v_h - w_h)(x_h)| \right)^{1/d}, \tag{67}$$

whence we only need to estimate $|\partial\Gamma(v_h - w_h)(x_h)|$ for all $x_h \in \mathcal{C}_h^-(v_h - w_h)$. For these nodes, we easily see that

$$\partial\Gamma(v_h - w_h)(x_h) \subset \partial(v_h - w_h)(x_h).$$

We claim that

$$\partial w_h(x_h) + \partial\Gamma(v_h - w_h)(x_h) \subset \partial v_h(x_h) \quad \forall\, x_h \in \mathcal{C}_h^-(v_h - w_h). \tag{68}$$

Fix $x_h \in \mathcal{C}_h^-(v_h - w_h)$, and let $\boldsymbol{p} \in \partial w_h(x_h)$ and $\boldsymbol{q} \in \partial\Gamma(v_h - w_h)(x_h)$, respectively, that is, by definition of the subdifferential (63),

$$\boldsymbol{p} \cdot (y_h - x_h) \le w_h(y_h) - w_h(x_h)$$

and

$$\boldsymbol{q} \cdot (y_h - x_h) \le \Gamma(v_h - w_h)(y_h) - \Gamma(v_h - w_h)(x_h)$$

for all nodes $y_h \in \Omega_h$. Adding both inequalities, we get

$$(\boldsymbol{p} + \boldsymbol{q}) \cdot (y_h - x_h) \le w_h(y_h) + \Gamma(v_h - w_h)(y_h) - (w_h(x_h) + \Gamma(v_h - w_h)(x_h))$$

Since x_h is in the contact set $\mathcal{C}_h^-(v_h - w_h)$, we have $\Gamma(v_h - w_h)(x_h) = (v_h - w_h)(x_h)$. For all other nodes $y_h \in \Omega_h$, we have $\Gamma(v_h - w_h)(y_h) \le (v_h - w_h)(y_h)$. Hence, we deduce

$$\begin{aligned}(\boldsymbol{p} + \boldsymbol{q}) \cdot (y_h - x_h) &\le w_h(y_h) + (v_h - w_h)(y_h) - (w_h(x_h) + (v_h - w_h)(x_h)) \\ &= v_h(y_h) - v_h(x_h).\end{aligned}$$

This inequality implies $(\boldsymbol{p} + \boldsymbol{q}) \in \partial v_h(x_h)$ and proves the claim.

The Brunn–Minkowski inequality of Lemma 2 applied to (68) yields

$$\begin{aligned}&|\partial w_h(x_h)|^{1/d} + |\partial\Gamma(v_h - w_h)(x_h)|^{1/d} \\ &\qquad \le |\partial w_h(x_h) + \partial\Gamma(v_h - w_h)(x_h)|^{1/d} \le |\partial v_h(x_h)|^{1/d},\end{aligned}$$

whence

$$|\partial\Gamma(v_h - w_h)(x_h)| \le \left(|\partial v_h(x_h)|^{1/d} - |\partial w_h(x_h)|^{1/d}\right)^d \quad \forall x_h \in \mathcal{C}_h^-(v_h - w_h).$$

This inequality gives us the desired estimate for $|\partial\Gamma(v_h - w_h)(x_h)|$. In view of (67), adding over all $x_h \in \mathcal{C}_h^-(v_h - w_h)$ concludes the proof. □

A direct consequence of this stability result is the maximum principle for nodal functions.

Corollary 6 (discrete maximum principle)**.**
Let v_h and w_h be two nodal functions over the nodal set $\bar{\Omega}_h$. If $v_h(x_h) \ge w_h(x_h)$ at all $x_h \in \partial\Omega_h$ and $|\partial v_h(x_h)| \le |\partial w_h(x_h)|$ at all $x_h \in \Omega_h$, then

$$w_h(x_h) \le v_h(x_h) \quad \forall x_h \in \Omega_h.$$

Proof. Since $v_h(y_h) \ge w_h(y_h)$ for all $y_h \in \partial\Omega_h$, then for any node $x_h \in \mathcal{C}_h^-(v_h - w_h)$, we have

$$\partial w_h(x_h) \subset \partial v_h(x_h).$$

Combining this with the assumption that $|\partial v_h(x_h)| \leq |\partial w_h(x_h)|$ for all $x_h \in \Omega_h$, we deduce $|\partial v_h(x_h)| = |\partial w_h(x_h)|$ for all $x_h \in \mathcal{C}_h^-(v_h - w_h)$. Consequently, the stability of Proposition 16 implies

$$\sup_{\Omega_h} (v_h - w_h)^- = 0,$$

whence $v_h - w_h \geq 0$. This completes the proof. □

Proposition 16 yields a lower bound on the difference between two nodal functions in terms of the difference of the measure of their subdifferentials. Similarly, to derive an upper bound, one may consider the functions $-w_h$ and $-v_h$ and derive

$$\sup_{\Omega_h} (w_h - v_h)^- \leq C \left(\sum_{x_h \in \mathcal{C}_h^-(w_h - v_h)} \left(|\partial w_h(x_h)|^{1/d} - |\partial v_h(x_h)|^{1/d} \right)^d \right)^{1/d}.$$

Combining both bounds, we can derive a bound on $\| v_h - w_h \|_{L^\infty(\Omega_h)}$ in terms of $|\partial v_h(x_i)|$ and $|\partial w_h(x_i)|$. In particular, the uniqueness of the solution of the Oliker–Prussner method follows immediately from Proposition 16.

Finally, we notice that Proposition 16 is instrumental to derive error estimates. Define the nodal interpolation of a function w as the nodal function $N_h w$ such that

$$N_h w(x_h) = w(x_h) \quad \forall x_h \in \bar{\Omega}_h. \tag{69}$$

Setting $w_h = u_h$ and $v_h = N_h u$ In Proposition 16, where u_h and u solve (64) and (1), respectively, we can derive an estimate for $\| u_h - N_h u \|_{L^\infty(\Omega)}$. It remains to estimate the discrepancy of the subdifferentials of the two nodal functions. While $|\partial u_h(x_h)| = f_{x_h}$ is known by definition of the scheme (64), the measure of the subdifferential $|\partial N_h u(x_h)|$ remains unknown. Therefore, the goal of our next step is to estimate the quantity $|\partial N_h u(x_h)|^{1/d} - f_{x_h}^{1/d}$ which will then be applied in Proposition 16 to derive a pointwise estimate.

3.3 Consistency

In general, this method (64) is consistent in the sense that the right-hand side of the (64) can be written equivalently as $\sum_{x_h \in \Omega_h} f_{x_h} \delta_{x_h}$ and this converges to f in measure. However, such a concept of convergence is too weak to derive rates of convergence. Fortunately, we realize that if internal nodes are translation invariant, then a reasonable notion of operator consistency holds for any convex quadratic polynomial; see Lemma 12. Such property is shown in Benamou et al. (2016), Mirebeau (2015), and Nochetto and Zhang (2019) for Cartesian nodes, see also Section 2.5. In contrast, we give here an

alternative proof of consistency based on the geometric interpretation of subdifferentials of convex quadratic polynomials in the interior of the domain, extend the results to $C^{2,\alpha}$ functions, and further investigate the consistency error in the region close to the boundary. To achieve this we, First, we require a definition.

Definition 20 (adjacent set)**.**
Given a convex nodal function w_h and a node $x_h \in \Omega_h$, the *adjacent set* of x_h, denoted by $A_{x_h}(w_h)$, is the collection of nodes $y_h \in \bar{\Omega}_h$ closest to x_h such that there exists a supporting hyperplane L of w_h and $L(y_h) = w_h(y_h)$. Thus, the set $A_{x_h}(w_h)$ is the collection of nodes in the star associated with x_h in the mesh $\mathcal{T}_h$ induced by $\Gamma(w_h)$.

Lemma 10 (size of adjacent sets)**.**
Let the nodal set Ω_h be translation invariant, and let p be a C^2 convex function defined in $\bar{\Omega}$. If $\lambda I \leq D^2 p \leq \Lambda I$ in Ω for some constants λ, $\Lambda > 0$ and $p_h := N_h p$ is the nodal function associated with p defined in (69), then the adjacent set of nodes $A_{x_h}(p_h)$ satisfies

$$A_{x_h}(p_h) \subset B_{Rh}(x_h)$$

where $R = \frac{\Lambda}{\lambda} d$, and $B_{Rh}(x_h)$ is the ball centred at x_h with radius Rh.

Proof. Let $z_h \in A_{x_h}(p_h)$ be such that

$$|z_h - x_h| = \max\{|y_h - x_h| : y_h \in A_{x_h}(p_h)\}.$$

Without loss of generality, we may assume that $p(x_h) = 0$ and $\nabla p(x_h) = 0$. Let ω be the convex hull of the nodal set $\{x_{\pm j} := x_h \pm h\tilde{\boldsymbol{e}}_j,\ j = 1, \ldots, d\}$ where $\{\tilde{\boldsymbol{e}}_j\}_{j=1}^d$ is the basis defined in (60). If $z_h \in \omega$, then the assertion is trivial because $R \geq 1$.

If $z_h \notin \omega$, then there is a constant $\tilde{R} \geq 1$ such that $\tilde{R}^{-1} z_h \in \omega$, which implies that $|z_h| \leq \tilde{R} h$ and $|z_h| \geq \tilde{R} d^{-1/2} h$. Because ω is convex, we may write

$$\tilde{R}^{-1} z_h = \sum_{\substack{j=1 \\ \sigma \in \{+,-\}}}^{d} \alpha_{\sigma j} x_{\sigma j}, \quad \alpha_{\sigma j} \geq 0, \qquad \sum_{\substack{=1 \\ \sigma \in \{+,-\}}}^{d} \alpha_{\sigma j} = 1.$$

We next note that $p(x_{\pm j}) \leq \frac{1}{2}\Lambda h^2$ for all $j = 1, \ldots, d$ because $D^2 p \leq \Lambda I$, $|x_{\pm j} - x_h| \leq h$, $p(x_h) = 0$ and $\nabla p(x_h) = 0$. Since $z_h \in A_{x_h}(p_h)$, there exists a supporting hyperplane L at x_h such that

$$L(z_h) = p_h(z_h), \quad L(x_{\pm j}) \leq p_h(x_{\pm j}) \leq \frac{1}{2}\Lambda h^2.$$

Exploiting that L is linear and $p_h(x_h) = 0$ yields

$$p_h(z_h) = L(z_h) = L\left(\tilde{R} \sum_{\substack{j=1 \\ \sigma \in \{+,-\}}}^{d} \alpha_{\sigma j} x_{\sigma j}\right) = \tilde{R} \sum_{\substack{j=1 \\ \sigma \in \{+,-\}}}^{d} \alpha_{\sigma j} L(x_{\sigma j}) \leq \frac{1}{2}\Lambda h^2 \tilde{R}.$$

On the other hand, since $D^2 p \geq \lambda I$ and $|z_h| \geq \tilde{R} h d^{-1/2}$, we have

$$p_h(z_h) = p(z_h) \geq \frac{\lambda}{2}|z_h|^2 \geq \frac{\lambda}{2}\tilde{R}^2 d^{-1} h^2.$$

Combining the last two inequalities implies

$$\tilde{R} \leq R = \frac{\Lambda}{\lambda} d.$$

This completes the proof. □

The previous result shows that for any node x_h with $\operatorname{dist}(x_h, \partial\Omega) > Rh$, all nodes in its adjacent set are contained in Ω_h. We apply this observation to establish the following consistency result.

Lemma 11 (properties of convex interpolation)**.**
Let p be a convex quadratic polynomial such that $\lambda I \leq D^2 p \leq \Lambda I$, and let $p_h = N_h p$ be the nodal function defined by (69). Then the following properties hold:

1. *For all $x_h \in \Omega_h$ we have $\partial p_h(x_h) \neq \emptyset$.*
2. *If the nodal set Ω_h is translation invariant and $\operatorname{dist}(x_h, \partial\Omega_h) \geq Rh$, with $R = \frac{\Lambda}{\lambda} d$, under a uniform refinement from Ω_h to $\Omega_{h/2}$, we have*

$$|\partial p_h(x_h)| = 2^d |\partial p_{h/2}(x_h)|.$$

3. *If the nodal set Ω_h is translation invariant, $\operatorname{dist}(x_h, \partial\Omega_h h) \geq Rh$, and $\operatorname{dist}(y_h, \partial\Omega_h) \geq Rh$, then $|\partial p_h(x_h)| = |\partial p_h(y_h)|$.*

Proof. To prove the first claim, we only need to note that if ℓ is the tangent plane of p at x_h, then ℓ is a supporting plane of p_h at x_h. Thus $\nabla \ell \in \partial p_h(x_h)$.

To prove the second claim, we may assume that $p(x_h) = 0$, and $\nabla p(x_h) = 0$. Note that for homogeneous quadratic polynomials, we have

$$p(x) = 4p\left(\frac{x}{2}\right).$$

A simple calculation yields

$$\partial p_h(x_h) = 2\partial p_{h/2}(x_h)$$

and therefore $|\partial p_h(x_h)| = 2^d |\partial p_{h/2}(x_h)|$.

To prove the third claim. We consider the function

$$p^{x_h}(x) = p(x) - \nabla p(x_h)\cdot(x - x_h) - p(x_h),$$

obtained by subtracting the tangent plane of p at x_h. Since adding an affine function does not change the measure of the subdifferential, we have $|\partial p_h(x_h)| = |\partial p_h^{x_h}(x_h)|$. Further note that by subtracting the tangent plane at a node y_h, we obtain the same function up to a parallel translation, that is,

$$p^{x_h}(x - x_h) = p^{y_h}(x - y_h).$$

Since the mesh is translation invariant, we have that if L is a supporting plane of $p_h^{x_h}$ at x_h, then by a parallel translation it is also a supporting plane of $p_h^{y_h}$ at y_h. Hence, we have $|\partial p_h^{x_h}(x_h)| = |\partial p_h^{y_h}(y_h)|$. Since $|\partial p_h(x_h)| = |\partial p_h^{x_h}(x_h)|$ for all nodes x_h, we conclude that $|\partial p_h(x_h)| = |\partial p_h(y_h)|$. □

Now we are ready to prove the consistency, for a proof see Nochetto and Zhang (2019, Lemma 5.3).

Lemma 12 (consistency I)**.**
Let p be a convex quadratic polynomial such that $\lambda I \le D^2 p \le \Lambda I$, and let $p_h := N_h p$ be the corresponding convex nodal function defined in (69). Let Ω_h be translation invariant. Then

$$|\partial p_h(x_h)| = \int_{\omega_{x_h}} \det D^2 p(x)\,\mathrm{d}x$$

for any node $x_h \in \Omega_h$ such that $dist(x_h, \partial\Omega) \ge Rh$ with $R = \frac{\Lambda}{\lambda} d$.

Proof. Let ϕ be any continuous function with compact support in Ω. We consider a sequence of nested refinements Ω_{h_n} with $h_n = 2^{-n}H$, for a fixed $H > 0$.

By Lemma 1 we immediately obtain, as $n \to \infty$, that

$$\sum_{y_{h_n}\in\Omega_{h_n}} \phi(y_{h_n})|\partial p_{h_n}(y_{h_n})| \to \int_\Omega \phi \det D^2 p(x)\,\mathrm{d}x = \det D^2 p(x) \int_\Omega \phi(x)\,\mathrm{d}x.$$

Thus, we only need to prove that as $n \to \infty$

$$\sum_{y_{h_n}\in\Omega} \phi(y_{h_n})|\partial p_{h_n}(y_{h_n})| \to \frac{|\partial p_H(x_H)|}{|\omega_{x_H}|} \int_\Omega \phi(x)\,\mathrm{d}x.$$

In view of second and third result in Lemma 11, we have

$$|\partial p_{h_n}(y_{h_n})| = |\partial p_{h_n}(x_H)| = 2^{-nd}|\partial p_H(x_H)|.$$

The refinement strategy implies that $|\omega_{y_{h_n}}| = 2^{-nd}|\omega_{x_H}|$. Thus, we infer that

$$\begin{aligned}\sum_{y_{h_n}\in\Omega_{h_n}} \phi(y_{h_n})|\partial p_{h_n}(y_{h_n})| &= \frac{|\partial p_H(x_H)|}{|\omega_{x_H}|} \sum_{y_{h_n}\in\Omega_{h_n}} \phi(y_{h_n})|\omega_{y_{h_n}}| \\ &\to \frac{|\partial p_H(x_H)|}{|\omega_{x_H}|} \int_\Omega \phi(x)\,\mathrm{d}x.\end{aligned}$$

This completes the proof. □

Moreover, for convex cubic polynomials, we have the following consistency error estimate. This result, to our knowledge, has not appeared elsewhere.

Lemma 13 (consistency II)**.**
Let $x_h \in \Omega_h$ and q be a convex cubic polynomial such that $\lambda I \leq D^2 q \leq \Lambda I$ in the ball $B_{Rh} := \overline{B_{Rh}(x_h)} \subset \Omega$, with $R = \frac{\Lambda}{\lambda} d$. Then

$$\left| |\partial N_h q(x_h)| - \int_{\omega_{x_h}} \det D^2 q(x)\, dx \right| \leq C h^{d+2} |q|_{C^3(B_{Rh})}..$$

Proof. Without loss of generality, we may assume that $x_h = 0$ and $q(0) = 0$ and $\nabla q(0) = 0$. We decompose the cubic polynomial $q(x)$ as

$$q(x) = p(x) + hr(x),$$

where $p(x)$ is a quadratic polynomial such that $D^2 p = D^2 q(0)$ and $r(x)$ is a homogeneous cubic polynomial. Since, by Lemma 10, the adjacent set $A_{x_h}(q)$ of the node $x_h = 0$ is contained in a ball of radius Rh we deduce that

$$|p(z_h)| \leq C_q R^2 h^2, \quad |r(z_h)| \leq C_r R^2 h^2 \quad \forall z_h \in A_{x_h}(q),$$

where C_q and C_r depends on $D^2 p$ and $D^3 r$, respectively. We set

$$q_t(x) = p(x) + tr(x) \quad t \in [-h, h],$$

and note that $\lambda I \leq D^2 q_t(0) \leq \Lambda I$ for all t. Therefore, the adjacent set of q_t at 0 remains in the ball B_{Rh}.

We set the measure of its subdifferential of q_t at x_h as a function of t

$$m(t) = |\partial N_h q_t(x_h)| = |\partial N_h q_t(0)|,$$

and note that we aim to show that

$$\left| m(h) - \int_{\omega_{x_h}} \det D^2 q(x) \mathrm{d}x \right| \leq C h^{d+2} |q|_{C^3(B_{Rh})}.$$

Now we proceed to prove the lemma in the following steps.

1. We aim to show that $m(t)$ is a polynomial of degree d

$$m(t) = \sum_{k=0}^{d} C_k t^k. \tag{70}$$

and the coefficients C_k satisfy $|C_k| \leq C h^d$ where C depends on $|D^2 p|$, $|D^3 r|$, and the dimension d. By the characterization of the subdifferential, given in Lemma 8, the subdifferential of $N_h q_t$ at 0 is the convex hull of the piecewise gradient of its convex envelope $\nabla \Gamma(N_h q_t)|_T$ for all $T \in \mathcal{T}_h$ that have x_h as a vertex; see Fig. 6. We label these simplices as $T_1, \cdots, T_N$ and, to simplify notation, we set the piecewise gradient of $\Gamma(N_h p)$ and $\Gamma(N_h r)$ at T_i as

$$\boldsymbol{v}_i = \nabla\Gamma(N_h p)|_{T_i}, \quad \boldsymbol{w}_i = \nabla\Gamma(N_h r)|_{T_i}, \quad i = 1, \ldots, N.$$

Hence, we have

$$\boldsymbol{u}_i := \nabla\Gamma(N_h q_t)|_{K_i} = \boldsymbol{v}_i + t\boldsymbol{w}_i.$$

To compute the measure of the convex hull of $\{\boldsymbol{u}_i\}$, we may divide the convex hull into a set of disjoint simplices $\{S_i,\ i = 1, \cdots, N\}$ and label the vertices of S_i as $\{\mathbf{0}, \boldsymbol{u}_{i_1}, \cdots, \boldsymbol{u}_{i_d}\}$. Thus, we obtain

$$m(t) = \sum_{i=1}^{N} |S_i| \quad \text{where } |S_i| \text{ denotes the signed volume of } S_i.$$

and so, by the volume formula of simplices, we get

$$m(t) = \frac{1}{d!}\sum_{i=1}^{N} \det \begin{pmatrix} 1 & \mathbf{0}^{\mathsf{T}} \\ 1 & \boldsymbol{u}_{i_1}^{\mathsf{T}} \\ \vdots & \vdots \\ 1 & \boldsymbol{u}_{i_d}^{\mathsf{T}} \end{pmatrix}. \qquad \boldsymbol{u}_{i_j} = \boldsymbol{v}_{i_j} + t\boldsymbol{w}_{i_j}. \tag{71}$$

Now, it is clear that

$$|S_i| = \sum_{k=0}^{d} C_k^i t^k$$

is a polynomial of t with degree at most d. Thus, $m(t)$ must be a polynomial with degree at most d as well. Furthermore, by the volume formula of simplices (71), the coefficients $|C_k^i| \le Ch^d$ because both $|\boldsymbol{v}_{i_j}| \le Ch$ and $|\boldsymbol{w}_{i_j}| \le Ch$. Finally, the number N of simplices S_i is finite and bounded by the number of vertices in the adjacent set A.

2. We show that $m'(0) = 0$. To do so, it suffices to show that the function m is even, that is $m(t) = m(-t)$ for all $-h \le t \le h$. Note that if $\boldsymbol{v} \in \partial N_h(p + tr)(0)$, then $-\boldsymbol{v} \in \partial N_h(p - tr)(0)$ for any $t \in (0, h]$. Indeed, since the subdifferential set is determined by the function values on the adjacent set which is contained in the ball $B_{Rh}(0)$, if $\boldsymbol{v} \cdot y_h \le (p + tr)(y_h)$ for all $y_h \in B_{Rh}(0)$, then

$$\boldsymbol{v} \cdot (-y_h) \le (p + tr)(-y_h) \quad \forall y_h \in B_{Rh}(0).$$

Hence, $-\boldsymbol{v} \in \partial N_h(p - tr)(0)$ because $p(y_h) = p(-y_h)$. Thanks to this symmetry property, we deduce that $|\partial N_h(p - tr)(0)| = |\partial N_h(p + tr)(0)|$, i.e., $m(t) = m(-t)$.

3. We show that

$$|m(h) - m(0)| \le Ch^{d+2}.$$

Combining the previous two steps we get that

$$m(t) = m(0) + C_2 t^2 + \cdots + C_d t^d$$

because $C_1 = m'(0) = 0$. Since $|C_j| \leq Ch^d$ for $j = 2, \ldots, d$, we deduce that

$$|m(t) - m(0)| \leq Ch^{d+2} \quad \forall t \in [0, h].$$

4. It remains to show that

$$\left| \int_{\omega_{x_h}} \det D^2 q(x) \mathrm{d}x - m(0) \right| \leq Ch^{d+2}.$$

By the consistency for quadratics given in Lemma 12, we have

$$m(0) = \int_{\omega_{x_h}} \det D^2 p(x) \mathrm{d}x.$$

Therefore, it is sufficient to show that

$$\left| \int_{\omega_{x_h}} (\det D^2 q(x) - \det D^2 p(x)) \mathrm{d}x \right| \leq Ch^{d+2}.$$

A Taylor expansion of $\det D^2 q = \det D^2 (p + hr)$ reveals that

$$\left| \det D^2 q(x) - \det D^2 p(x) - h \operatorname{cof} D^2 p(x) : D^2 r(x) \right| \leq Ch^2.$$

where the constant C depends on $D^2 p$ and $D^3 r$. This implies that

$$\left| \int_{\omega_{x_h}} (\det D^2 q(x) - \det D^2 p(x)) \mathrm{d}x \right| \leq h \left| \int_{\omega_{x_h}} \operatorname{cof} D^2 p(x) : D^2 r(x) \mathrm{d}x \right| + Ch^2 |\omega_{x_h}|.$$

Noting that $\operatorname{cof} D^2 p : D^2 r$ is an odd function and ω_{x_h} is symmetric respect to the origin, we obtain

$$\int_{\omega_{x_h}} \operatorname{cof} D^2 p(x) : D^2 r(x) \mathrm{d}x = 0$$

and

$$\left| \int_{\omega_{x_h}} (\det D^2 q(x) - \det D^2 p(x)) \mathrm{d}x \right| \leq Ch^{d+2}.$$

This completes the proof. □

Now for any function w that can be approximated locally by a quadratic polynomial such that $w(x) = p(x) + \mathcal{O}(h^{2+\alpha})$ in $B_{Rh}(x_h)$ or by a cubic polynomial such that $w(x) = q(x) + O(h^{3+\alpha})$ in $B_{Rh}(x_h)$, we show that the consistency error of the Oliker–Prussner method is of order $\mathcal{O}(h^{\alpha})$ and $\mathcal{O}(h^{1+\alpha})$, respectively.

Proposition 17 (interior consistency)**.**
Let Ω_h be a translation invariant set of nodes, and $x_h \in \Omega_h$ be such that dist $(x_h, \partial\Omega_h) \geq Rh$ with $R = \frac{\Lambda}{\lambda} d$. If $w \in C^{2+k,\alpha}(\overline{B_{Rh}})$, with $k \in \{0, 1\}$, and $\alpha \in (0, 1]$) is a convex function with $\lambda I \leq D^2 w \leq \Lambda$I, then we have

$$\left| |\partial N_h w(x_h)| - \int_{\omega_{x_h}} \det D^2 w(x) \mathrm{d}x \right| \leq C h^{k+\alpha} |w|_{C^{2+k,\alpha}(\overline{B_{Rh}})} |\omega_{x_h}|,$$

where $C = C(d, \lambda, \Lambda)$.

Proof. We divide the proof into two cases $k = 0$ and $k = 1$.

1. Case $k = 0$: We only need to show the inequality

$$|\partial N_h w(x_h)| \leq \int_{\omega_{x_h}} \det D^2 w(x) \mathrm{d}x + C h^{\alpha} |w|_{C^{2,\alpha}(\overline{B_{Rh}})} |\omega_{x_h}|,$$

because the reverse inequality can be derived similarly. Since $w \in C^{2,\alpha}(\overline{B_{Rh}})$, we estimate w by a quadratic polynomial p so that

$$w(x) \leq p(x) \quad \forall x \in B_{Rh}(x_h),$$

where $p(x_h) = w(x_h)$, $\nabla p(x_h) = \nabla w(x_h)$ and

$$D^2 p = D^2 w(x_h) + C h^{\alpha} |w|_{C^{2,\alpha}(\overline{B_{Rh}})} I$$

for a fixed, and sufficiently large, constant C. Let $p_h = N_h p$, and note that

$$|\partial N_h w(x_h)| \leq |\partial p_h(x_h)|.$$

It remains to show that

$$|\partial p_h(x_h)| \leq \int_{\omega_{x_h}} \det D^2 w(x) \mathrm{d}x + C h^{\alpha} |w|_{C^{2,\alpha}(\overline{B_{Rh}})} |\omega_{x_h}|.$$

Since $(\lambda + Ch^{\alpha})I \leq D^2 p \leq (\Lambda + Ch^{\alpha})I$ and

$$\frac{\Lambda + Ch^{\alpha}}{\lambda + Ch^{\alpha}} \leq \frac{\Lambda}{\lambda} \quad \text{because } \Lambda \geq \lambda,$$

invoking the consistency of Lemma 12 we obtain

$$|\partial p_h(x_h)| = \int_{\omega_{x_h}} \det D^2 p(x)\mathrm{d}x$$

provided that $\mathrm{dist}(x_h, \partial\Omega_h) \geq Rh$. Recalling that $w \in C^{2,\alpha}(\overline{B_{Rh}})$, we can write $D^2 p = D^2 w(x) + E(x)$ for all $x \in \overline{B_{Rh}}$, where $|E(x)| \leq Ch^\alpha |w|_{C^{2,\alpha}(\overline{B_{Rh}})}$. A Taylor expansion yields

$$|\partial p_h(x_h)| \leq \int_{\omega_{x_h}} \det D^2 w(x)\mathrm{d}x + Ch^\alpha |w|_{C^{2,\alpha}(\overline{B_{Rh}})} |\omega_{x_h}|.$$

2. **Case** $k = 1$: If $w \in C^{3,\alpha}(\overline{B}_{Rh})$, we approximate w by a cubic polynomial q so that

$$w(x) \leq q(x) \quad \forall x \in B_{Rh}(x_h),$$

where $q(x_h) = w(x_h)$, $\nabla q(x_h) = \nabla w(x_h)$,

$$D^2 q(x_h) = D^2 w(x_h) + Ch^{1+\alpha} |w|_{C^{3,\alpha}(\overline{B_{Rh}})},$$

and $D^3 q = D^3 w(x_h)$ with universal constant C. The rest of the proof is similar to the previous case.

Combing both cases, we conclude the proof of the estimate. □

3.4 Pointwise error estimate

We are now ready to show a pointwise error estimate for the method (64) under suitable regularity assumptions on the solution u. We aim to apply the stability of the numerical scheme shown in Proposition 16 to derive a lower bound of the difference $v_h - u_h$, for a suitable convex piecewise linear function v_h.

Assume that the convex solution u of the Monge–Ampère equation (1) is $C^{k,\alpha}$ near the boundary of the domain Ω where $k \in \{2, 3\}$ and $\alpha \in (0, 1]$. We first extend the solution to a larger convex domain

$$\Omega_{4Rh} = \{x \in \mathbb{R}^d, \mathrm{dist}(x, \Omega) \leq 4Rh\}$$

such that, for sufficiently small h, the extended function, which we still denote as u, remains $C^{k,\alpha}$-continuous in the extended region and satisfies

$$\frac{\lambda}{2} I \leq D^2 u(x) \leq 2\Lambda I \quad \text{for any } x \in \Omega_{4Rh}. \tag{72}$$

Next, we extend the translation invariant interior nodal set Ω_h to the extended domain Ω_{4Rh} and, by an abuse of notation, we still denote the set as Ω_h, that is,

$$\Omega_h = \left\{ x_h = \sum_{j=1}^{d} z^j \tilde{\boldsymbol{e}}_j : \quad z^j \in \mathbb{Z} \right\} \cap \Omega_{4Rh}.$$

We construct the piecewise linear function $v_h = \Gamma(N_h u)$ by taking the convex envelope of the nodal interpolation of the solution u on Ω_h in the extended domain and then restrict the piecewise linear function v_h to the domain Ω. Thus, this procedure yields a piecewise linear function v_h defined on the domain Ω.

We claim that the piecewise linear function v_h satisfies the following two conditions which are useful in the error estimate. First, the adjacent set size estimate of Lemma 10 and the bound of D^2u given in (72) imply that for any interior node $x_h \in \Omega_h \cap \Omega$, its adjacent set $A_{x_h}(v_h)$ is contained in the extended domain Ω_{4Rh}. Second, we notice that $|v_h(x) - u(x)| \leq Ch^2$ on the boundary $\partial\Omega$ where the constant C depends on $\|u\|_{C^2(\Omega)}$. This is simply due to the fact that the diameter of any patch of a node $z \in \Omega_h \cap \Omega$ is bounded by $4Rh$ and interpolation theory of piecewise linear function.

Now we are ready to derive the main error estimate.

Theorem 13 (error estimate)**.**
Let u be the solution of the Monge–Ampère equation (1), $0 \leq \lambda I \leq D^2u \leq \Lambda I$ and $u \in C^{2+k,\alpha}(\bar{\Omega})$ with $k \in \{0, 1\}$ and $\alpha \in (0, 1]$. Let Ω_h be a translation invariant nodal set satisfying (60), and let u_h be the solution of discrete Monge–Ampère equation (64) defined on Ω_h. Then we have

$$\|u - u_h\|_{L^\infty(\Omega)} \leq Ch^{k+\alpha},$$

where the constant C depends only on $\|u\|_{C^{2+k,\alpha}(\Omega)}$, λ, Λ, diam(Ω), and space dimension d.

Proof. Let v_h be the interpolation of the extension of the solution u defined above. Since $|v_h - u_h| \leq Ch^2$ on the boundary $\partial\Omega$, we have $v_h + Ch^2 \geq u_h$. By the stability of the numerical solution, Proposition 16, we obtain

$$\sup_{\Omega_h} (v_h + Ch^2 - u_h)^- \leq C \left(\sum_{x_i \in \mathcal{C}_h^-(v_h - u_h)} \left(|\partial v_h(x_i)|^{1/d} - |\partial u_h(x_i)|^{1/d} \right)^d \right)^{1/d}.$$

Invoking the consistency error estimate, Proposition 17, we immediately obtain

$$\sup_{\Omega_h} (v_h + Ch^2 - u_h)^- \leq Ch^{k+\alpha}.$$

By a simple algebraic manipulation, the estimate yields a lower bound for the error $v_h - u_h \geq -Ch^2 - Ch^{k+\alpha}$. Similarly, an estimate for the upper bound follows by considering the function $u_h + Ch^2 - v_h$. Combining both estimates, we get the desired result. □

3.5 $W^{2,p}$ error estimate

The results and arguments of the previous section have recently been extended to the derivation of $W^{2,p}$ error estimates of the Oliker–Prussner scheme (Neilan and Zhang, 2018). Here, the discrete $W^{2,p}$ norm is taken to be the sum of weighted second-order differences:

$$\|v\|_{W_f^{2,p}} = \left(\sum_{x_h \in \Omega_h} f_{x_h} |\Delta_{\boldsymbol{e}} v(x_h)|^p \right)^{1/p}.$$

The starting point is a simple observation that the contact set of a nodal function contains information of its second-order difference. In particular, if u_h is the solution to (64) and v_h is some approximation to u, then we can define the perturbed error

$$w_h^\epsilon = v_h - (1-\epsilon) u_h \tag{73}$$

with parameter $\epsilon \in (0, 1)$. Now, by using the identity $\Delta_{\boldsymbol{e}} w_h^\epsilon(x_h) \geq \Delta_{\boldsymbol{e}} \Gamma w_h^\epsilon(x_h) \geq 0$ for $x_h \in \mathcal{C}_h^-(w_h^\epsilon)$, we have, after some algebraic manipulations,

$$\Delta_e (u_h - v_h)(x_h) \leq \frac{\epsilon}{1-\epsilon} \Delta_e v_h(x_h) \quad \forall x_h \in \mathcal{C}_h^-(w_h^\epsilon).$$

The right-hand side of this expression is uniformly bounded for appropriate v_h if u is sufficiently smooth, and therefore we find that the error $\Delta_e(u_h - v_h)(x_h)$ is controlled on the contact set $\mathcal{C}_h^-(w_h^\epsilon)$. However, noting that w_h^ϵ is not necessary convex, we must estimate $\Delta_e(u_h - v_h)(x_h)$ on the complement set

$$E^\epsilon := \Omega_h \backslash \mathcal{C}_h^-(w_h^\epsilon). \tag{74}$$

This is done by estimating its cardinality in terms of the consistency of the method.

Lemma 14 (size of complement set)**.**
Let u_h and v_h be convex nodal functions with $u_h = v_h$ on $\partial\Omega_h$ and $u_h \leq v_h$ on Ω_h. Set

$$|\partial u_h(x_h)| = f_{x_h} \quad \text{and} \quad |\partial v_h(x_h)| = g_{x_h} \quad x_h \in \Omega_h.$$

Then there exists a constant $C > 0$ depending only on f such that

$$\sum_{x_h \in E^\epsilon} f_{x_h} \leq C \frac{(1-\epsilon)}{\epsilon} \|f^{1/d} - g^{1/d}\|_{\ell^d(\mathcal{C}_h^-(w_h^\epsilon))},$$

where w_h^ϵ and E^ϵ are defined by (73) and (74), respectively.

The last ingredient to develop $W^{2,p}$ estimates is a simple result of the discrete L^1 norm of a nodal function in terms of its level sets. Roughly speaking this result gives a relation between Riemann and Lebesgue sums; see Neilan and Zhang (2018, Lemma 5.1)

Lemma 15. *Let s_h be a nodal function with $|s_h(x_h)| \leq M$ for some $M > 0$. Then, for any $\sigma > 0$,*

$$\sum_{x_h \in \Omega_h} f_{x_h} |s_h(x_h)| \leq \sigma \sum_{k=0}^{M} \sum_{x_h \in A_k} f_{x_h},$$

where

$$A_k = \{x_h \in \Omega_h : \ |s_h(x_h)| \geq k\sigma\}.$$

Theorem 14 ($W^{2,p}$ error estimate)**.**
Suppose that the conditions of Theorem 14 are satisfied with $k + \alpha = 2$. Then there holds

$$\|u - u_h\|_{W^{2,p}_f} \leq \begin{cases} Ch^{1/p} & p \in (d, \infty) \\ C|\log h|^{1/d} h^{1/d} & p \in (1, d]. \end{cases}$$

We now give a sketch of the main ideas to prove Theorem 14 and refer the reader to Neilan and Zhang (2018) for details. To communicate the main ideas, we make the simplifying assumption that the consistency estimate in Proposition 17 holds up to the boundary. We also assume homogeneous boundary conditions, i.e., $g = 0$ in (1b). These assumptions, which do not hold in general, allow us to derive better rates of convergence than those stated in Theorem 14.

As a first step we set $v_h = (1 - Ch^2)^{1/d} N_h u$, where $C > 0$ is sufficiently large such that (cf. Proposition 17)

$$g_{x_h} = |\partial v_h(x_h)| = (1 - Ch^2)|\partial N_h u(x_h)| \leq f_{x_h}.$$

Therefore by the comparison principle in Corollary 6, we have $v_h \geq u_h$ on Ω_h. We also have $|f_{x_h} - g_{x_h}| \leq Ch^{2+d}$.

To deduce the estimate, it suffices bound

$$\sum_{x_h \in \Omega_h} f_{x_h} (\Delta_{\boldsymbol{e}} (u_h - v_h)(x_h))^+.$$

Bounding the negative part of the error can be obtained by similar arguments.

For parameter ϵ_k with $\epsilon_k/(1 - \epsilon_k) = Ck^{1/p} h^2$, we define

$$A_k = \left\{ x_h \in \Omega_h : \ \Delta_{\boldsymbol{e}} (u_h - v_h)(x_h) \geq \frac{\epsilon_k}{1 - \epsilon_k} \Delta_{\boldsymbol{e}} v_h(x_h) \right\},$$

and note that $A_k \subset E^{\epsilon_k}$. Let $s_h(x_h) = \left|(\Delta_{\boldsymbol{e}}(u_h - v_h))^+\right|^p$, and note that $|s_h(x_h)| \leq Ch^{-2p}$ because u_h and v_h are bounded. Applying Lemma 15, with $\sigma = Ch^{2p}$, we have

$$\sum_{x_h \in \Omega_h} f_{x_h} \left|(\Delta_{\boldsymbol{e}}(u_h - v_h)(x_h))^+\right|^p \leq Ch^{2p}\left(1 + \sum_{k=1}^{Ch^{-2p}} \sum_{x_h \in A_k} f_{x_h}\right).$$

On the other had, using Lemma 14 and the consistency of the scheme yields, for h sufficiently small,

$$\begin{aligned}\sum_{x_h \in A_k} f_{x_h} \leq \sum_{x_h \in E^{\epsilon_k}} f_{x_h} &\leq C\frac{1-\epsilon_k}{\epsilon_k}\|f^{1/d} - g^{1/d}\|_{\ell^d(\mathcal{C}_h^-(w_h^{\epsilon_k}))} \\ &\leq Ch^2\frac{1-\epsilon_k}{\epsilon_k} = Ck^{-1/p}.\end{aligned}$$

Thus, we find that

$$\begin{aligned}\sum_{x_h \in \Omega_h} f_{x_h} \left|(\Delta_{\boldsymbol{e}}(u_h - v_h)(x_h))^+\right|^p &\leq Ch^{2p}\left(1 + \sum_{k=1}^{Ch^{-2p}} \frac{1}{k^{1/p}}\right) \\ &\leq C\begin{cases} h^2|\log h| & \text{if } p = 1, \\ h^2 & \text{if } p > 1. \end{cases}\end{aligned}$$

In certain settings, Theorems 13 and 14 immediately give us $W^{1,p}$ error estimates as well. To make this precise, we assume that the basis $\{\tilde{\boldsymbol{e}}_j\}_{j=1}^d = \{\boldsymbol{e}_j\}_{j=1}^d$ defined in (60) is the canonical one. We then define the backward difference operator

$$D_{\boldsymbol{e}}^- v(x_h) = \frac{v(x_h) - v(x_h - \boldsymbol{e}h)}{h},$$

and the discrete norms/semi-norms, for $p \in (1, \infty)$,

$$\|v\|_{L_h^p(\Omega_h)} = \left(h^d \sum_{x_h \in \Omega_h} |v(x_h)|^p\right)^{1/p},$$

$$\|v\|_{W_h^{1,p}(\Omega_h)} = \left(\|v\|_{L_h^p(\Omega_h)}^p + h^d \sum_{j=1}^d \|D_{\boldsymbol{e}_j}^- v\|_{L_h^p(\Omega_h)}^p\right)^{1/p},$$

$$\|v\|_{W_h^{2,p}(\Omega_h)} = \left(\|v\|_{W_h^{1,p}(\Omega_h)}^p + h^d \sum_{j=1}^d \left(\|\Delta_{\boldsymbol{e}_j} v\|_{L_h^p(\Omega_h)}^p + \sum_{\substack{i=1 \\ j \neq i}}^d \|D_{\boldsymbol{e}_i}^- D_{\boldsymbol{e}_j}^- v\|_{L_h^p(\Omega_h)}^p\right)\right)^{1/p}.$$

We then have (Jovanović and Süli, 2014, Lemmas 2.60–2.61)

$$\|v\|_{W_h^{1,p}(\Omega_h)} \le C\,\|v\|_{L_h^p(\Omega_h)}^{1/2}\|v\|_{W_h^{2,p}(\Omega_h)}^{1/2}.$$

Therefore noting that $\|v\|_{L_h^p(\Omega_h)} \le C\,\|v\|_{L^\infty(\Omega)}$, and,

$$D_{\boldsymbol{e}_i}^- D_{\boldsymbol{e}_j}^- v(x) = \frac{1}{2}\left(\Delta_{\boldsymbol{e}_i} v(x-h\boldsymbol{e}_i) + \Delta_{\boldsymbol{e}_j} v(x-h\boldsymbol{e}_j) - \Delta_{\tilde{\boldsymbol{e}}_{i,j}} v(x-h(\boldsymbol{e}_i+\boldsymbol{e}_j))\right)$$

with $\tilde{\boldsymbol{e}}_{i,j} = \boldsymbol{e}_i - \boldsymbol{e}_j$, we have the following, by Theorems 13 and 14.

Corollary 7. *Suppose that the conditions in Theorem 13 are satisfied with $k + \alpha = 2$, and assume that $f \ge f_0 > 0$ in Ω. Then there holds*

$$\|u - u_h\|_{W_h^{1,p}(\Omega_h)} \le \begin{cases} Ch^{1+\frac{1}{2p}} & p \in (d,\infty), \\ C|\log h|^{\frac{1}{2d}} h^{1+\frac{1}{2d}} & p \in (1,d]. \end{cases}$$

Remark 15 (extensions).
In this section we showed that the stability estimate given in Proposition 16 provides a powerful tool to develop error estimates for the Monge–Ampère equation, as it allows us to derive L^∞ and W_p^2 error estimates when the solution enjoys regularity $u \in C^{2+k,\alpha}(\bar{\Omega})$. Thanks to this stability estimate, it also possible to extend these estimates if the solution is of lower regularity and/or degenerate. The key observation is that the stability estimate measures the consistency error in the ℓ^d-norm. If the solution is rough in a region of small measure and smooth elsewhere, so that the consistency error is small in ℓ^d-norm, then by the stability estimate, we may still derive a rate of convergence for the low regularity case. This is explored in Nochetto and Zhang (2019, Theorem 6.3) to prove a rate of convergence for solutions in $C^{1,1}(\Omega)$, but not in $C^2(\Omega)$. ■

4 Finite Element Methods

> *It will be found that most classical mathematical approximation procedures as well as the various direct approximations used in engineering fall into this category. It is thus difficult to determine the origins of the finite element method and the precise moment of its invention.*
>
> Zienkiewicz and Taylor (2000)

In this section, we summarize recent developments of finite element methods for the Monge–Ampère problem with Dirichlet boundary conditions (1). For simplicity, throughout this section, we assume that boundary conditions in (1) are homogeneous, i.e., $g = 0$. The extension to nonhomogenous boundary conditions is straightforward.

The main difficulty to construct (and analyze) finite element schemes for fully nonlinear problems is that the PDEs are nonvariational. Recall that a finite element method is typically derived by

(i) multiplying the PDE by a test function;
(ii) integrating the resulting product over the domain;
(iii) performing integration by parts to arrive at a variational formulation;
(iv) posing the variational formulation on a finite dimensional space, usually consisting of piecewise polynomials.

Note that the third step usually requires some structure conditions of the PDE, e.g., that the PDE is in divergence-form, which is not present for fully nonlinear problems. Another obvious difficult to construct convergent finite element schemes is that the notion of viscosity solutions, given in Definition 4, and Alexandrov solutions, as in Definition 9 for the Monge–Ampère equation are nonvariational, and it is unclear how this solution concept can be adopted within a finite element framework.

We must remark, however, that the Monge–Ampère operator (1a) does possess a divergence-form. Using well-known algebraic identities and the divergence-free property of cofactor matrices, there holds $\det D^2 u = \frac{1}{d}\nabla \cdot (\operatorname{cof} D^2 u \nabla u)$. Note however that variational formulations based on this identity would still involve second-order derivatives, and therefore, at this time, it is unclear whether numerical methods based on this approach are advantageous.

Nonetheless, assuming some regularity of the solution, well-defined finite element methods can be formulated and analyzed for fully nonlinear PDEs. One approach is to omit the third step of the four-step process described above. For example, multiplying the Monge–Ampère equation (1a) by a function v and integrating over Ω yields the identity

$$\int_\Omega (f - \det D^2 u) v \, dx = 0. \tag{75}$$

A simple calculation involving Hölder's inequality and Sobolev embeddings show that expression (75) is well-defined provided $u, v \in W^{2,d}(\Omega)$. Finite element methods can then be constructed based on the identity (75). Namely, an obvious finite element method based on the identity (75) seeks $u_h \in V_h$ satisfying

$$\int_\Omega (f - \det D^2 u_h) v_h \, dx = 0 \quad \forall v_h \in X_h, \tag{76}$$

where X_h is a finite dimensional space consisting of piecewise polynomials with respect to a partition of Ω that vanish on the boundary. While this

method may be convergent (cf. Awanou, 2014, 2015c, 2017b; Böhmer, 2008; Davydov and Saeed, 2013; Neilan, 2014b), the appearance of global second-order derivatives in the method necessitates the use of C^1 finite element spaces which can be arduous to implement and are not found in most finite element software packages. In addition, C^1 finite element generally require high-degree polynomial bases, resulting in a relatively large algebraic system.

Because of the many disadvantages of the finite element method (76) several finite element methods with simpler spaces have been developed. These include C^0 penalty methods, discontinuous Galerkin (DG) methods, mixed finite element methods, and methods based on high order regularizations. We now discuss these methods in the subsequent sections.

4.1 Continuous finite element methods

Here we summarize finite element methods presented in Brenner et al. (2011), Brenner and Neilan (2012), and Neilan (2013) for the Monge–Ampère equation which employ spaces consisting of continuous, piecewise polynomials, i.e., the Lagrange finite element space. These are arguably the simplest finite element spaces and are available on virtually all finite element software programs and libraries. In addition, we provide a slightly new and improved convergence analysis based on recent results for finite element methods for linear nondivergence form PDEs (Feng et al., 2017). To describe these methods and their accompanying analysis, we require some notation.

As before, we assume that $\Omega \subset \mathbb{R}^d$ $(d = 2,3)$ is a bounded, convex domain. Let $\mathcal{T}_h$ denote a shape-regular and simplicial triangulation of Ω. We denote the sets of interior and boundary $(d-1)$–dimensional faces of $\mathcal{T}_h$ by $\mathcal{F}_h^I$ and $\mathcal{F}_h^B$, restrictively. The jump of a vector valued function $\boldsymbol{v}$ across an interior face $F = \partial T_+ \cap \partial T_- \in \mathcal{F}_h^I$ is given by

$$[\![\boldsymbol{v}]\!] = \frac{1}{2}(\boldsymbol{v}_+ \otimes \boldsymbol{n}_+ + \boldsymbol{n}_+ \otimes \boldsymbol{v}_+ + \boldsymbol{v}_- \otimes \boldsymbol{n}_- + \boldsymbol{n}_- \otimes \boldsymbol{v}_-), \tag{77}$$

where $\boldsymbol{n}_\pm$ is the outward unit normal of $\partial T_\pm$, and $\boldsymbol{v}_\pm = \boldsymbol{v}|_{T_\pm}$. We also define the average of B (a scalar, vector, or matrix-valued function) across F as

$$\{\!\{B\}\!\} = \frac{1}{2}(B_+ + B_-). \tag{78}$$

If $F = \partial T_+ \cap \partial\Omega \in \mathcal{F}_h^B$, then we define

$$[\![\boldsymbol{v}]\!] = \frac{1}{2}(\boldsymbol{v}_+ \otimes \boldsymbol{n}_+ + \boldsymbol{n}_+ \otimes \boldsymbol{v}_+), \quad \{\!\{B\}\!\} = B_+. \tag{79}$$

For an integer $r \geq 2$, the Lagrange finite element space with homogeneous boundary conditions is given by

$$V_h = \left\{ v_h \in W_0^{1,\infty}(\Omega) : \quad v_h|_T \in \mathbb{P}_r(T) \quad \forall T \in \mathcal{T}_h \right\},$$

where $\mathbb{P}_r(T)$ is the space of polynomials with degree less than or equal to r with domain T. In addition, for a number $p \in (1, \infty)$ and integer m, we define

$$W^{m,p}(\mathcal{T}_h) = \prod_{T \in \mathcal{T}_h} W^{m,p}(T), \qquad V_p = W_0^{1,p}(\Omega) \cap W^{2,p}(\mathcal{T}_h),$$

and note that $V_h \subset V_p$ for all $p \in (1, \infty)$. We also set $H^m(\mathcal{T}_h) = W^{m,2}(\mathcal{T}_h)$.

Because of the noninclusion $V_h \not\subset W^{2,d}(\Omega)$, the finite element formulation (76) is not well defined if X_h is taken to be the Lagrange finite element space. A naïve approach to bypass this issue is to redefine this formulation so that integration is done piecewise over the mesh, i.e., to consider

$$\sum_{T \in \mathcal{T}_h} \int_T \left(f - \det D^2 u_h \right) v_h \mathrm{d}x \qquad \forall v_h \in V_h. \tag{80}$$

While this method is well defined (i.e., all quantities are defined and bounded), it is easy to see that the scheme is ill-posed. For example, if $w_h \in V_h$ is strictly piecewise linear, then $\det D^2 w_h = 0$ on each $T \in \mathcal{T}_h$, and consequently, uniqueness (and stability) is dramatically lost.

The arguments given in Brenner et al. (2011) offer an alternative explanation on why the formulation (80) leads to an ill-posed problem. Namely, the main point in Brenner et al. (2011) is that the linearization of the discrete problem (80) is not consistent with respect to the linearization of the continuous problem (1a). Instead, to ensure consistency and stability, finite element methods for the Monge–Ampère problem should be designed such that the discrete linearization at the solution u is a coercive operator over the finite element space. We now explain how to construct methods with stable linearizations. To do so, we first assume that the exact solution to the Monge–Ampère equation satisfies $u \in C^{k,\alpha}(\bar{\Omega})$ with $k + \alpha > 2$ and is strictly convex.

Define

$$\mathbb{F}[u] = f - \det D^2 u$$

to be the Monge–Ampère operator, and let L be the linearization of F at the solution u, i.e.,

$$Lw = \lim_{t \to 0} \frac{\mathbb{F}[u + tw] - \mathbb{F}[u]}{t} = -\operatorname{cof} D^2 u : D^2 w, \tag{81}$$

where $\operatorname{cof} D^2 u$ denotes the cofactor matrix of $D^2 u$, and ":" denotes the Frobenius inner product. The assumptions on u imply that matrix $\operatorname{cof} D^2 u$ is positive definite on $\bar{\Omega}$ and uniformly continuous.

A consistent discretization of linear operators in nondivergence form (such as L) was introduced in Feng et al. (2017). In the case that the linear problem is given by (81), the discretization is given by $L_h : V_p \to V_h'$ with

$$\begin{aligned}\langle L_h v, w_h\rangle = &-\sum_{T\in\mathcal{T}_h}\int_T \big(\operatorname{cof} D^2 u : D^2 v\big) w_h \mathrm{d}x \\ &+ \sum_{F\in\mathcal{F}_h^I}\int_F \big\{\!\!\big\{\operatorname{cof} D^2 u\big\}\!\!\big\} : [\![\nabla v]\!] w_h \mathrm{d}s,\end{aligned} \tag{82}$$

where $\langle\cdot,\cdot\rangle$ denotes the dual pairing between some Banach space and its dual. The operator L_h is clearly consistent with L: If $v \in W^{2,p}(\Omega)\cap W_0^{1,p}(\Omega)$, then $\langle L_h v, w_h\rangle = \langle Lv, w_h\rangle$ for all $w_h \in V_h$. In addition, the discrete operator is stable as the next lemma shows. We refer the reader to Feng et al. (2017) for a proof.

Lemma 16 (stability)**.**
Define the discrete $W^{2,p}$-norm

$$\begin{aligned}\|v\|^p_{W_h^{2,p}(\Omega)} &:= \|D_h^2 v\|^p_{L^p(\Omega)} + \sum_{F\in\mathcal{F}_h^I} h_F^{1-p}\|[\![\nabla v]\!]\|^p_{L^p(F)} \qquad 1<p<\infty,\\ \|v\|_{W_h^{2,\infty}(\Omega)} &:= \|D_h^2 v\|_{L^\infty(\Omega)} + \max_{F\in\mathcal{F}_h^I} h_F^{-1}\|[\![\nabla v]\!]\|_{L^\infty(F)},\end{aligned}$$

where $D_h^2 v$ is the piecewise Hessian of v. Assume that $u\in C^2(\bar\Omega)$ and is strictly convex over $\bar\Omega$. Then there exists $h_0 > 0$ depending on the modulus of continuity of $D^2 u$, such that for $h \in (0, h_0]$, there holds the following inf-sup condition $(2\le p<\infty)$

$$\|w_h\|_{W_h^{2,p}(\Omega)} \le C\,\|L_h w_h\|_{L_h^p(\Omega)} := \sup_{v_h\in V_h\setminus\{0\}} \frac{\langle L_h w_h, v_h\rangle}{\|v_h\|_{L^{p'}(\Omega)}} \qquad \forall w_h \in V_h,$$

where $1/p + 1/p' = 1$

Based on the definition of L_h and the stability results stated in Lemma 16 we can develop a consistent discretization for the Monge–Ampère problem as well as a convergence theory. Essentially, its construction is based on the observations that the expressions $\int_T(\operatorname{cof} D^2 u : D^2 v) w_h \mathrm{d}x$ and $\int_F \{\!\!\{\operatorname{cof} D^2 u\}\!\!\} : [\![\nabla v]\!] w_h\, ds$ are the linearizations of $\int_T (f - \det D^2 v) w_h$ and $\int_F \{\!\!\{\operatorname{cof} D^2 v\}\!\!\} : [\![\nabla v]\!] w_h \mathrm{d}s$, respectively, about the solution u. With this in mind, we define the discrete operator $\mathbb{F}_h : V \to V_h'$ via

$$\langle \mathbb{F}_h[v], w\rangle = \sum_{T\in\mathcal{T}_h}\int_T \big(f - \det D^2 v\big) w_h \mathrm{d}x + \sum_{F\in\mathcal{F}_h^I}\int_F \big\{\!\!\big\{\operatorname{cof} D^2 v\big\}\!\!\big\} : [\![\nabla v]\!] w_h \mathrm{d}s,$$

and consider the finite element method: Find $u_h \in V_h$ such that

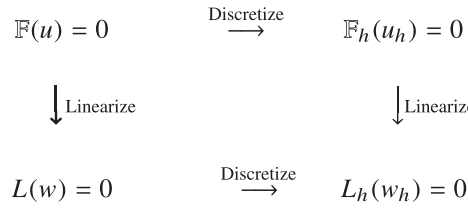

FIG. 7 A commuting diagram connecting the nonlinear problems and their discretizations.

$$\langle \mathbb{F}_h[u_h], v_h \rangle = 0 \quad \forall v_h \in V_h. \tag{83}$$

We immediately see that method (83) is consistent: There holds $[\![\nabla u]\!]|_F = 0$ over all interior faces F, and therefore $\langle \mathbb{F}_h[u], v_h \rangle = 0$ for all $v_h \in V_h$. Furthermore, the proceeding discussion implies that L_h is the linearization of $\mathbb{F}_h$:

$$L_h w = \lim_{t \to 0} \frac{\mathbb{F}_h[u + tw] - \mathbb{F}_h[u]}{t} \quad \text{in } V_h'.$$

In summary the diagram given in Fig. 7 commutes. We now show that this property (along with the regularity and convexity assumptions of u) implies that there exists a locally unique solution to (83) with optimal rates of convergence.

As a first step, we first point out that Lemma 16 implies that $L_h|_{V_h}$ is bijective. Therefore, the mapping $M_h : V_p \to V_h$ given by

$$M_h = \left(L_h|_{V_h}\right)^{-1}(L_h - \mathbb{F}_h) \tag{84}$$

is well defined. The existence of a solution to the finite element method (83) is proven by showing that M_h has a fixed point in a ball centred at $u_{c,h}$, where $u_{c,h}$ is the elliptic projection of u given by

$$u_{c,h} := \left(L_h|_{V_h}\right)^{-1} L_h u. \tag{85}$$

The basis of this argument is provided in the next lemma.

Lemma 17 (M_h is Lipschitz)**.**
Assume that the convex solution of the Monge–Ampère equation satisfies $u \in C^{k,\alpha}(\bar{\Omega})$ with $k + \alpha > 2$. Then there holds, for all $p \in [2, \infty)$ and all $v_1, v_2 \in V_p$,

$$\|M_h v_1 - M_h v_2\|_{W_h^{2,p}(\Omega)} \leq C_1 \left\| u - \frac{1}{2}(v_1 + v_2) \right\|_{W_h^{2,\infty}(\Omega)} \|v_1 - v_2\|_{W_h^{2,p}(\Omega)},$$

where $C_1 > 0$ depends on p and u, but is independent of h.

Proof. We give the proof of the two-dimensional case $d = 2$; the arguments in three dimensions are similar and can be found in Brenner and Neilan (2012).

We first use Taylor's Theorem and the fact that F_h is quadratic in two dimensions, to get

$$\mathbb{F}_h[v] = \mathbb{F}_h[u] + L_h(v-u) + R_h[v-u] = L_h(v-u) + R_h[v-u],$$

where $R_h : V \to V_h'$ is quadratic in its arguments and independent of u. Using this expansion into the mapping M_h yields

$$\begin{aligned} M_h[v_1] - M_h[v_2] &= \left(L_h|_{V_h}\right)^{-1}(L_h v_1 - L_h v_2 - (\mathbb{F}_h[v_1] - \mathbb{F}_h[v_2])) \\ &= \left(L_h|_{V_h}\right)^{-1}(R_h[v_2 - u] - R_h[v_1 - u]). \end{aligned} \tag{86}$$

Since R_h is quadratic there holds

$$\begin{aligned} R_h[v_2-u] - R_h[v_1-u] &= \int_0^1 DR_h[t(v_2-u) + (1-t)(v_1-u)](v_2-v_1)\mathrm{d}t \\ &= DR_h(\frac{1}{2}(v_2+v_1) - u)(v_2-v_1), \end{aligned}$$

where by DR_h we denoted the derivative of R_h. Therefore, by (86) and Lemma 16 we have

$$\|M_h v_1 - M_h v_2\|_{W_h^{2,p}(\Omega)} \leq C \left\| DR_h\left(\frac{1}{2}(v_2+v_1) - u\right)(v_2-v_1) \right\|_{L_h^p(\Omega)}.$$

Several applications of Hölder's inequality yields (cf. Neilan, 2013, Lemma 4.2)

$$\|DR_h(w)(q)\|_{L_h^p(\Omega)} \leq C \|w\|_{W_h^{2,\infty}(\Omega)} \|q\|_{W_h^{2,p}(\Omega)},$$

and therefore

$$\|M_h v_1 - M_h v_2\|_{W_h^{2,p}(\Omega)} \leq C \left\| \frac{1}{2}(v_1+v_2) - u \right\|_{W_h^{2,\infty}(\Omega)} \|v_1 - v_2\|_{W_h^{2,p}(\Omega)}.$$

□

Lemma 18 (contraction)**.**
Assume that the hypotheses of Lemma 17 are satisfied. For fixed $\rho > 0$ and $p \in [2, \infty)$, define the closed ball

$$B_{\rho,p} = \left\{ v_h \in V_h : \quad \|u_{c,h} - v_h\|_{W_h^{2,p}(\Omega)} \leq \rho \right\},$$

where $u_{c,h} \in V_h$ is defined by (85). Then, for all v_1, $v_2 \in B_{\rho,p}$, there holds

$$\|M_h v_1 - M_h v_2\|_{W_h^{2,p}(\Omega)} \leq C_2 h^{-d/p}\left(h^{\ell+\alpha} + \rho\right) \|v_1 - v_2\|_{W_h^{2,p}(\Omega)},$$

where $\ell = \min\{r-2, k-2\}$.

Proof. First, the smoothness assumptions on u allows us to conclude that the elliptic projection $u_{c,h}$ satisfies (Feng et al., 2017, Theorem 3.2)

$$\|u-u_{c,h}\|_{W_h^{2,p}(\Omega)} \le C_3 h^{\ell+\alpha} \quad p\in[2,\infty), \tag{87}$$

where $C_3 > 0$ depends on p and $\|u\|_{C^{k,\alpha}(\bar{\Omega})}$. Consequently, there holds by an inverse estimate, for any $w_h \in V_h$,

$$\begin{aligned}\|u-u_{c,h}\|_{W_h^{2,\infty}(\Omega)} &\le \|u-w_h\|_{W_h^{2,\infty}(\Omega)} + Ch^{-d/p}\,\|u_{c,h}-w_h\|_{W_h^{2,p}(\Omega)}\\ &\le \|u-w_h\|_{W_h^{2,\infty}(\Omega)} + Ch^{-d/p}\Big(\|u-u_{c,h}\|_{W_h^{2,p}(\Omega)} + \|u-w_h\|_{W_h^{2,p}(\Omega)}\Big).\end{aligned}$$

Taking w_h to be the nodal interpolant of u yields

$$\|u-u_{c,h}\|_{W_h^{2,\infty}(\Omega)} \le C_4 h^{\ell+\alpha-d/p}. \tag{88}$$

Applying this result to Lemma 17 and using an inverse estimate, we obtain

$$\begin{aligned}&\|M_h v_1 - M_h v_2\|_{W_h^{2,p}(\Omega)}\\ &\quad\le C\left(\|u-u_{c,h}\|_{W_h^{2,\infty}(\Omega)} + h^{-d/p}\,\|u_{c,h}-\frac{1}{2}(v_1+v_2)\|_{W_h^{2,p}(\Omega)}\right)\|v_1-v_2\|_{W_h^{2,p}(\Omega)}\\ &\quad\le Ch^{-d/p}\left(h^{\ell+\alpha}+\rho\right)\|v_1-v_2\|_{W_h^{2,p}(\Omega)}\end{aligned}$$

for all v_1, $v_2 \in B_{\rho,p}$. □

Theorem 15 (error estimate)**.**
Assume that $u\in C^{k,\alpha}(\bar{\Omega})$ with $k+\alpha > 2$ and is strictly convex. Set $\ell = \min\{r-2, k-2\}$. There exists $h_1 > 0$ such that for $h \le h_1$, there exists a solution to (83) satisfying

$$\|u-u_h\|_{W_h^{2,p}(\Omega)} \le Ch^{\ell+\alpha}. \tag{89}$$

Moreover, if $\tilde{u}_h$ is another solution to (83) then there holds $\|u-\tilde{u}_h\|_{W_h^{2,\infty}(\Omega)} \ge C$, with the constant $C > 0$ independent of h.

Proof. Fix $p\in[2,\infty)$ such that $\ell+\alpha-d/p > 0$, and let

$$h_1 = \min\{1/(4C_2), 1/(2C_1C_2C_3C_4)\}^{1/(\alpha+\ell-d/p)}.$$

Then, for $h \le \min\{h_0, h_1\}$, where h_0 was defined in Lemma 16, set $\rho_1 = h^{\ell+\alpha}/(4C_2)$. Lemma 18 then shows that, for $v_1, v_2 \in B_{\rho_1,p}$,

$$\begin{aligned}\|M_h v_1 - M_h v_2\|_{W_h^{2,p}(\Omega)} &\le C_2\left(h^{\ell+\alpha-d/p} + h^{-d/p}\rho_1\right)\|v_1-v_2\|_{W_h^{2,p}(\Omega)}\\ &\le 2C_2 h_1^{\alpha+\ell-d/p}\,\|v_1-v_2\|_{W_h^{2,p}(\Omega)} \le \frac{1}{2}\|v_1-v_2\|_{W_h^{2,p}(\Omega)},\end{aligned}$$

and therefore $M_h|_{V_h}$ is a contraction mapping on $B_{\rho_1,p}$. Likewise, we can use Lemma 17 and the fact that $u_{c,h} = M_h u$ to get (cf. (87) and (88))

$$\begin{aligned}\|u_{c,h} - M_h v\|_{W_h^{2,p}(\Omega)} &= \|M_h u - M_h v\|_{W_h^{2,p}(\Omega)} \\ &\leq \frac{C_1}{2}\|u - u_{c,h}\|_{W_h^{2,\infty}(\Omega)}\,\|u - u_{c,h}\|_{W_h^{2,p}(\Omega)} \\ &\leq \frac{C_1 C_3 C_4 h^{2\alpha+2\ell-d/p}}{2} \leq \frac{h^{\ell+\alpha}}{4C_2} = \rho_1.\end{aligned}$$

Therefore M_h maps $B_{\rho_1,p}$ to itself. By Banach's fixed point theorem, we conclude that M_h has a fixed point in $B_{\rho_1,p}$, and this fixed point is a solution to (83). The error estimate for $\alpha - d/p > 0$ (89) follows from the inclusion $u_h \in B_{\rho_1,p}$ and the definition of ρ_1. The other cases $\ell + \alpha - d/p \leq 0$ then follow from Hölder's inequality.

Finally, if $\tilde{u}_h \in V_h$ is another solution to (83), then there holds $M_h\tilde{u}_h = \tilde{u}_h$. Therefore, by Lemma 17 we conclude that

$$\begin{aligned}\|\tilde{u}_h - u_h\|_{W_h^{2,p}(\Omega)} &= \|M_h\tilde{u}_h - M_h u_h\|_{W_h^{2,p}(\Omega)} \\ &\leq \frac{C_1}{2}\left(\|u - u_h\|_{W_h^{2,\infty}(\Omega)} + \|u - \tilde{u}_h\|_{W_h^{2,\infty}(\Omega)}\right)\|u_h - \tilde{u}_h\|_{W_h^{2,p}(\Omega)}.\end{aligned}$$

Now applying similar arguments as those found in Lemma 17, we conclude that $\|u - u_h\|_{W_h^{2,\infty}(\Omega)} \leq Ch^{\alpha-d/p} \to 0$. Therefore, by dividing by $\|u_h - \tilde{u}_h\|_{W_h^{2,p}(\Omega)}$, we get $C \leq \|u - \tilde{u}_h\|_{W_h^{2,\infty}(\Omega)}$ for h sufficiently small. □

Remark 16 (extensions).
The proposed method and the conclusion of Theorem 15 deserve the following comments:

- As mentioned earlier, the analysis given here slightly improves the results given in Brenner et al. (2011) and Neilan (2013). Namely, the paper (Brenner et al., 2011) requires $d = 2$, $r \geq 3$, and $u \in H^s(\Omega)$ for $s > 3$ (implying that $u \in C^{2,\alpha}(\bar{\Omega})$ by a Sobolev embedding). The paper (Neilan, 2013) requires $r \geq 2$ and regularity $u \in W^{3,\infty}(\Omega)$ to carry out the analysis.
- Discontinuous Galerkin methods have also been developed under this methodology in Neilan (2013). The analysis carried out in this section can be applied to these methods using the recent results for nondivergence PDEs given in Feng et al. (2018).
- A two–grid method to solve the nonlinear method has recently been proposed in Awanou et al. (2018). ■

4.2 Mixed formulations

In this section we describe mixed finite element formulations for the Monge–Ampère equation proposed in Lakkis and Pryer (2011), Neilan (2014a), Awanou (2015a); Awanou and Li (2014), Awanou (2017a), and Kawecki et al. (2018). Essentially, the main idea in these approaches is to introduce the Hessian matrix of u as an additional auxiliary unknown in the formulation of the Monge–Ampère problem, that is, we write the PDE (1a) as

$$\sigma = D^2 u, \quad \det \sigma = f \quad \text{in } \Omega. \tag{90}$$

As before, assuming regularity $u \in W^{2,d}(\Omega)$ so that $\sigma \in L^d(\Omega)$, we can multiply the second equation by a smooth test function and integrate over the domain:

$$\int_\Omega (f - \det \sigma) v \mathrm{d}x = 0 \tag{91}$$

for all $v \in L^\infty(\Omega)$.

The direct analogue of this formulation in the discrete setting requires C^1 finite element spaces by the same reasons that the method described in Section 4.1 does. In other words, to ensure that the discrete version of (91) is well–defined, we require that the Hessian of the discrete approximation u_h has (global) second-order derivatives in $L^d(\Omega)$; if u_h is a piecewise polynomial, then this restriction implies that $u \in C^1(\Omega)$. To relax this restriction on the finite element spaces, one can instead develop finite element methods that only employ continuous (or discontinuous) bases based on this formulation by introducing the notion of a discrete Hessian (also known as a finite element Hessian (Lakkis and Pryer, 2011)). The discrete Hessian is defined globally via an integration by parts procedure rather than a piecewise fashion. This idea has been carried out for (linear) Kirchhoff plates in Huang et al. (2010), and its formulation is reminiscent of the construction of local discontinuous Galerkin methods for second-order problems (Arnold et al., 2002; Cockburn and Shu, 1998).

To motivate the definition of the discrete Hessian, we introduce the auxiliary space

$$\Sigma_h = \{\tau_h \in L^\infty(\Omega; \mathbb{R}^{d\times d}) : \quad \tau_h|_T \in \mathbb{P}_r(T; \mathbb{R}^{d\times d}) \quad \forall T \in \mathcal{T}_h\},$$

and note the following integration by parts identity

$$\begin{aligned} \sum_{T\in\mathcal{T}_h} \int_T D^2 w : \tau_h \mathrm{d}x = &- \sum_{T\in\mathcal{T}_h} \int_T (\nabla \cdot \tau_h) \cdot \nabla w \mathrm{d}x \\ &+ \sum_{T\in\mathcal{T}_h} \int_{\partial T} (\tau_h \boldsymbol{n}_T) \cdot \nabla w \mathrm{d}s, \end{aligned} \tag{92}$$

for all $w \in H^2(\Omega)$ and $\tau_h \in \Sigma_h$. Here, $\boldsymbol{n}_T$ is the outward unit normal of ∂T, and the divergence acting on a matrix is performed row-wise. We may then write

the integral boundary terms in (92) using the jump and average operators. In addition to (77)–(79), we define the jump of a matrix-valued function τ across $F = \partial T_+ \cap T_- \in \mathcal{F}_h^I$ as

$$[\![\tau]\!] = \tau_+ \boldsymbol{n}_+ + \tau_- \boldsymbol{n}_-,$$

and define $[\![\tau]\!] = \tau_+ \boldsymbol{n}_+$ if $F = \partial T_+ \cap \partial\Omega \in \mathcal{F}_h^B$. We then have

$$\begin{aligned}\sum_{T\in\mathcal{T}_h}\int_{\partial T}(\tau_h \boldsymbol{n}_T)\cdot\nabla w \mathrm{d}s &= \sum_{F\in\mathcal{F}_h^I}\int_F \{\!\{\tau_h\}\!\} : [\![\nabla w]\!]\mathrm{d}s + \sum_{F\in\mathcal{F}_h}\int_F [\![\tau_h]\!]\cdot\{\!\{\nabla w\}\!\}\mathrm{d}s \\ &= \sum_{F\in\mathcal{F}_h}\int_F [\![\tau_h]\!]\cdot\{\!\{\nabla w\}\!\}\mathrm{d}s,\end{aligned}$$

where we used that $[\![\nabla w]\!]|_F = 0$ for all $F \in \mathcal{F}_h^I$ due to the regularity $w \in H^2(\Omega)$. Combining this identity with (92), we arrive at

$$\sum_{T\in\mathcal{T}_h}\int_T D^2 w : \tau_h \mathrm{d}x = -\sum_{T\in\mathcal{T}_h}\int_T (\nabla\cdot\tau_h)\cdot\nabla w \mathrm{d}x + \sum_{F\in\mathcal{F}_h}\int_F [\![\tau_h]\!]\cdot\{\!\{\nabla w\}\!\}\mathrm{d}s.$$

This identity leads to the following definitions of the discrete Hessian.

Definition 21 (discontinuous discrete Hessian)**.**
The *discontinuous discrete Hessian* is the operator $\mathbb{H}_h : H^1(\Omega)\cap H^2(\mathcal{T}_h) \to \Sigma_h$ uniquely defined by the conditions

$$\int_\Omega \mathbb{H}_h(w) : \tau_h \mathrm{d}x = -\sum_{T\in\mathcal{T}_h}\int_T (\nabla\cdot\tau_h)\cdot\nabla w \mathrm{d}x + \sum_{F\in\mathcal{F}_h}\int_F [\![\tau_h]\!]\cdot\{\!\{\nabla w\}\!\}\mathrm{d}s$$

for all $\tau_h \in \Sigma_h$.

Remark 17 (characterization through liftings).
Define the lifting operator

$$\Theta : L^2(\mathcal{F}_h^I;\mathbb{R}^d) \to \Sigma_h$$

via

$$\int_\Omega \Theta(\boldsymbol{v}) : \tau_h \mathrm{d}x = -\sum_{F\in\mathcal{F}_h^I}\int_F \{\!\{\tau_h\}\!\} : [\![\boldsymbol{v}]\!]\mathrm{d}s \quad \forall \tau_h \in \Sigma_h.$$

Integrating by parts we obtain

$$\begin{aligned}\sum_{T\in\mathcal{T}_h}\int_T \mathbb{H}_h(w) : \tau_h \mathrm{d}x &= \sum_{T\in\mathcal{T}_h}\int_T D^2 w : \tau_h \mathrm{d}x - \sum_{F\in\mathcal{F}_h^I}\int_F \{\!\{\tau_h\}\!\} : [\![\nabla w]\!]\mathrm{d}s \\ &= \sum_{T\in\mathcal{T}_h}\int_T \big(D^2 w + \Theta(\nabla w)\big) : \tau_h \mathrm{d}x.\end{aligned}$$

Recalling that $D_h^2 w$ denotes the piecewise Hessian of w, and that V_h is the (scalar) Lagrange space of degree r, we then have $D_h^2 V_h \subset \Sigma_h$, and therefore

$$\mathbb{H}_h(w_h) = D_h^2 w_h + \Theta(\nabla w_h) \quad \forall w_h \in V_h.$$

■

The notion of the discrete Hessian and the formal identities (90) and (91) lead to the following scheme introduced in Neilan (2014a): Find $u_h \in V_h$ such that

$$\int_\Omega (f - \det \mathbb{H}_h(u_h)) v_h \mathrm{d}x \quad \forall v_h \in V_h. \tag{93}$$

Remark 18 (mixed formulation).
While (93) is written in primal form, the problem is in fact a mixed finite element method. Introducing $\sigma_h = \mathbb{H}_h(u_h) \in \Sigma_h$, we see from the definition of the discrete Hessian that (93) is equivalent to the system

$$\int_\Omega \sigma_h : \tau_h \mathrm{d}x + \int_\Omega (\nabla \cdot \tau_h) \cdot u_h \mathrm{d}x - \sum_{F \in \mathcal{F}_h} \int_F [\![\tau_h]\!] \cdot \{\!\{\nabla u_h\}\!\} \mathrm{d}s \ = 0, \tag{94a}$$

$$\int_\Omega (f - \det \sigma_h) v_h \mathrm{d}x \ = 0, \tag{94b}$$

for all $(\tau_h, v_h) \in \Sigma_h \times V_h$. Note that the matrix representation of the form $(\sigma_h, \tau_h) \to \int_\Omega \sigma_h : \tau_h \mathrm{d}x$ is symmetric positive definite, and more importantly, block-diagonal because Σ_h does not have any continuity constraints. As a result, the Schur complement (i.e., the primal method (93)) represents a sparse algebraic system of equations. ■

Theorem 16 (error estimate)**.**
Assume that $d = 2$, and that (1) has a unique strictly convex solution $u \in C^{r+3,\,\alpha}(\Omega)$ with $r \geq 3$ and $\alpha > 0$. Then for h sufficiently small, there exists a locally unique solution to the finite element method (93). Moreover, there holds

$$\|u - u_h\|_{H^1(\Omega)} + h \,\|\, \sigma - \sigma_h\|_{L^2(\Omega)} \leq C h^r. \tag{95}$$

Proof. See Neilan (2014a, Theorem 4.2). □

Remark 19 (regularity).
The regularity assumptions on u in Theorem 16 can be relaxed using the stability analysis for linear nondivergence form PDES found in Neilan (2017). There it is shown that, assuming $u \in C^2(\bar{\Omega})$,

$$\|w_h\|_{W_h^{2,2}(\Omega)} \leq C \,\|\, L_h w_h\|_{L_h^2(\Omega)} \quad \forall w_h \in V_h,$$

with

$$\langle L_h w_h, v_h \rangle = -\int_\Omega \operatorname{cof} D^2 u : \mathbb{H}_h(w_h) v_h \mathrm{d}x.$$

By applying the same techniques found in the previous section, it is simple to show that the solution to (93) satisfies $\|u - u_h\|_{W_h^{2,2}(\Omega)} \leq Ch^{\ell+\alpha}$ with $\ell = \min\{r-2, k-2\}$ provided that $u \in C^{k,\alpha}(\bar{\Omega})$ with $k + \alpha > 3$, $r \geq 3$, and h is sufficiently small.

To reduce the number of unknowns in the mixed system (94), continuity constraints can be added in the matrix–valued space Σ_h. This is the idea of the method proposed in Lakkis and Pryer (2013). There, the auxiliary space is defined as the matrix-valued Lagrange space, i.e.,

$$\Sigma_h^c := \Sigma_h \cap H^1(\Omega; \mathbb{R}^{d\times d}) = \{\tau_h \in H^1(\Omega) : \quad \tau_h \in \mathbb{P}_r(T; \mathbb{R}^{d\times d}) \quad \forall T \in \mathcal{T}_h\}.$$

Restricting Definition 21 to Σ_h^c leads to the following notation of the discrete Hessian.

Definition 22 (continuous discrete Hessian)**.**
The *continuous discrete Hessian* is the operator $\mathbb{H}_h^c : H^1(\Omega) \cap H^2(\mathcal{T}_h) \to \Sigma_h^c$ uniquely defined by the conditions

$$\int_\Omega \mathbb{H}_h^c(w) : \tau_h \mathrm{d}x \;\; = -\int_\Omega (\nabla \cdot \tau_h) \cdot \nabla w \mathrm{d}x + \int_{\partial\Omega} (\tau_h \boldsymbol{n}) \cdot \nabla w \mathrm{d}s$$

for all $\tau_h \in \Sigma_h^c$.

This definition leads to a finite element method proposed in Lakkis and Pryer (2013) which similar to (93), but with the continuous version of the discrete Hessian.

$$\int_\Omega \big(f - \det \mathbb{H}_h^c(u_h)\big) v_h \mathrm{d}x = 0 \quad \forall v_h \in V_h. \tag{96}$$

As before, we may set $\sigma_h = \mathbb{H}_h^c(u_h)$ as an auxiliary variable, and deduce from Definition 22 that (96) is equivalent to the mixed method

$$\int_\Omega \sigma_h : \tau_h \mathrm{d}x + \int_\Omega (\nabla \cdot \tau_h) \cdot \nabla u_h \mathrm{d}x - \int_{\partial\Omega} (\tau_h \boldsymbol{n}) \cdot \nabla u_h \mathrm{d}s \;\; = 0, \tag{97a}$$

$$\int_\Omega (f - \det \sigma_h) v_h \mathrm{d}x \;\; = 0, \tag{97b}$$

for all $(\tau_h, v_h) \in \Sigma_h^c \times V_h$ Compared with the formulation using the discontinuous discrete Hessian, the mixed problem (97) has significantly less unknowns than (94) due to the continuity restrictions of Σ_h^c. On the other hand, the

(mass) matrix associated with the form $(\sigma_h, \tau_h) \to \int_\Omega \sigma_h : \tau_h$ is not block-diagonal, and therefore the Schur complement of (97) (i.e., the algebraic system representing the primal problem (96)) is dense.

Existence, (local) uniqueness, and error estimates for method (97) are similar to the statements given in Theorem 16.

Theorem 17 (error estimates)**.**
Assume that $d \in \{2, 3\}$*, and that* (1) *has a unique strictly convex solution* $u \in H^{r+3}(\Omega)$ *with* $r \geq d$*. Then for* h *sufficiently small, there exists a locally unique solution to the finite element method* (97)*. Moreover, there holds*

$$\|u - u_h\|_{H^1(\Omega)} + h \| \sigma - \sigma_h\|_{L^2(\Omega)} \leq Ch^r. \tag{98}$$

Proof. See Awanou and Li (2014, Theorem 3.13) and Awanou (2015a, 2017a, Theorem 1). □

Remark 20 (extension to optimal transport).
The mixed finite element method (97) has recently been extended to the optimal transport problem in Kawecki et al. (2018). ■

Remark 21 (historical remark).
Our presentation follows a reverse chronological order. The first Galerkin-type method based on the concept of discrete Hessians was that of Lakkis and Pryer (2013), where they used the continuous Hessian of Definition 22. The DG version was introduced later. ■

4.3 Galerkin methods for singular solutions

The analysis of the Galerkin methods discussed thus far require relatively stringent regularity conditions to carry out the analysis (e.g., $u \in C^{2,\alpha}(\Omega)$). While numerical experiments indicate that regularity assumptions can be relaxed somewhat, they also indicate that some regularity of the solution is required for the methods to converge. For example, the numerical experiments in Brenner et al. (2011) indicate that the C^0 penalty method (83) does not converge if $u \notin H^2(\Omega)$ in two dimensions. In this section, we discuss various ways to modify the Galerkin methods and the analysis such that the resulting numerical scheme is robust with respect to the solution's regularity.

The first approach, introduced in Feng and Neilan (2009), regularizes the problem at the PDE level by adding a higher order perturbation, resulting in a fourth-order, quasi-linear problem. The motivation of this approach is that solutions of the regularized problem are defined via variational principles, so that weak formulations can be obtained via integration by parts, and therefore the resulting PDE framework is amenable to Galerkin methods. Applying this methodology to the Monge–Ampère problem results in

$$-\epsilon\Delta^2 u^\epsilon + \det D^2 u^\epsilon = f \qquad \text{in } \Omega, \tag{99a}$$

$$u = 0 \qquad \text{on } \partial\Omega, \tag{99b}$$

where $\epsilon > 0$ and $\Delta^2 = \Delta\Delta$ denotes the biharmonic operator. Note that, due to the higher order of the PDE, the Dirichlet boundary condition is no longer sufficient to close the system. In Feng and Neilan (2009), the following additional boundary conditions are proposed:

$$\Delta u^\epsilon = 0, \quad \text{or} \quad \frac{\partial \Delta u^\epsilon}{\partial \boldsymbol{n}} = 0 \qquad \text{on } \partial\Omega. \tag{99c}$$

These conditions are chosen so that the resulting boundary layer is minimized; see Feng and Neilan (2009) for details. For the sake of illustration, we take the first boundary condition in (99c) in the discussion below.

Since the problem (99) is quasi–linear and in divergence–form, the notion of weak solutions is easily defined.

Definition 23 (weak solution)**.**
A function $u \in W^{2,d}(\Omega) \cap W_0^{1,d}(\Omega)$ is a weak solution to (99) provided that

$$-\epsilon \int_\Omega \Delta u^\epsilon \Delta v \mathrm{d}x + \int_\Omega v \det D^2 u^\epsilon \mathrm{d}x = \int_\Omega f v \mathrm{d}x \quad \forall v \in W^{2,d}(\Omega) \cap W_0^{1,d}(\Omega). \tag{100}$$

The function $u = \lim_{\epsilon \downarrow 0} u^\epsilon$, if it exists, is called a weak (resp., strong) moment solution to the Monge–Ampère problem if convergence holds in a $W^{1,d}$-weak (resp., $W^{2,d}$-weak) topology.

Remark 22 (relation to other solution concepts).
Except in very simple settings (e.g., radially symmetric solutions (Feng and Neilan, 2014)), the existence of moment solutions and their relation with viscosity and Alexandrov solutions is an open problem. Nonetheless, numerical experiments indicate that this methodology leads to robust numerical methods with respect to regularity of the solution of the Monge–Ampère equation. For example, numerical methods applied to problem (99) are able to capture viscosity/Alexandrov solutions that are merely Lipschitz continuous. ■

Constructing methods for the regularized problem (100) can be done by applying any of the above Galerkin methods described above; one only needs to tack on a consistent and stable discretization of the biharmonic operator to the discrete formulation. For example, the simplest method, at least in theory, is to restrict the variational formulation (100) onto a finite dimensional subspace of $W^{2,d}(\Omega) \cap W_0^{1,d}(\Omega)$. This results in the method to find $u_h^\epsilon \in X_h$ satisfying

$$\epsilon \int_\Omega \Delta u_h^\epsilon \Delta v_h \mathrm{d}x + \int_\Omega \left(f - \det D^2 u_h^\epsilon\right) v_h \, dx = 0 \quad \forall v_h \in X_h, \tag{101}$$

with $X_h \subset C^1(\Omega) \cap W_0^{1,d}(\Omega)$. A convergence analysis of this discrete problem has been done in Feng and Neilan (2011). There it is shown that, if there exists a moment solution with sufficient regularity, then there exists a locally unique solution to the discrete problem (101).

Analogously, combining the C^0 finite element method (83) with the symmetric C^0 interior penalty method for the biharmonic problem introduced in Engel et al. (2002) and Brenner and Sung (2005) results in the method: Find $u_h^\epsilon \in V_h$ satisfying

$$\begin{aligned}
&\epsilon \sum_{T\in\mathcal{T}_h} \int_T \Delta u_h^\epsilon \Delta v_h \mathrm{d}x \\
&\quad - \epsilon \sum_{F\in\mathcal{F}_h^I} \int_F \left(\{\!\{ \Delta u_h^\epsilon \}\!\} (I : [\![\nabla v_h]\!]) + \{\!\{ \Delta v_h \}\!\} (I : [\![\nabla u_h^\epsilon]\!]) \right. \\
&\quad \left. - \frac{\sigma}{h_F} [\![\nabla u_h^\epsilon]\!] : [\![\nabla v_h]\!] \right) \mathrm{d}s + \sum_{T\in\mathcal{T}_h} \int_T \left(f - \det D^2 u_h^\epsilon\right) v_h \mathrm{d}x \\
&\quad + \sum_{F\in\mathcal{F}_h^I} \int_F \{\!\{ \operatorname{cof} D^2 u_h^\epsilon \}\!\} : [\![\nabla u_h^\epsilon]\!] v_h \mathrm{d}s = 0
\end{aligned} \tag{102}$$

for all $v_h \in V_h$. Here, $\sigma > 0$ is a penalty parameter, and we recall that I denotes the $d \times d$ identity matrix and V_h is the Lagrange finite element space of degree $r \geq 2$ with homogeneous Dirichlet boundary conditions. The method (102) can be written succinctly as

$$\epsilon \langle A_h u_h^\epsilon, v_h \rangle + \langle \mathbb{F}_h[u_h^\epsilon], v_h \rangle = 0 \quad \forall v_h \in V_h,$$

where the operator $\mathbb{F}_h$ is defined by (83), and A_h is a consistent discretization of the biharmonic operator given by

$$\begin{aligned}
\langle A_h w, v_h \rangle = &\sum_{T\in\mathcal{T}_h} \int_T \Delta w \Delta v_h \mathrm{d}x - \sum_{F\in\mathcal{F}_h^I} \int_F \left(\{\!\{ \Delta w \}\!\} (I_d : [\![\nabla v_h]\!]) \right. \\
&\left. + \{\!\{ \Delta v_h \}\!\} (I_d : [\![\nabla w_h]\!]) - \frac{\sigma}{h_F} [\![\nabla w]\!] : [\![\nabla v_h]\!] \right) \mathrm{d}s.
\end{aligned}$$

Arguments given in Brenner and Sung (2005); Engel et al. (2002) show that there exists $\sigma_0 > 0$, independent of h, such that $\langle A_h v_h, v_h \rangle \geq C \| v_h \|^2_{W_h^{2,2}(\Omega)}$ for all $v_h \in V_h$ provided that $\sigma \geq \sigma_0$. Moreover, there holds $\epsilon \langle A_h u^\epsilon, v_h \rangle + \langle \mathbb{F}_h[u^\epsilon], v_h \rangle = 0$ for all $v_h \in V_h$ provided that $u^\epsilon \in H^s(\Omega)$ for some $s > 5/2$. Thus, the method (102) is consistent.

While a convergence analysis of the regularized PDE (99) and the discretization (102) is an open problem, we show, via numerical experiments in the next section, that the method is able to capture nonsmooth solutions for the

Monge–Ampère problem in a variety of settings. In addition, as shown in Brenner et al. (2011), Newton's method is robust for the regularized solution, which allows a natural way to construct initial guesses for the (unregularized) problem (83).

4.3.1 Convergence of interior discretizations

Recent results given in Awanou (2015b, 2016, 2017b); Awanou and Awi (2016) argue that, in certain settings, standard discretizations (both finite element and finite difference) for the Monge–Ampère equation converge to the Alexandrov solution as the discretization parameter tends to zero. Here, in this section, we summarize these results and the techniques to obtain them.

As always, we assume that Ω is convex. More importantly, we assume also that the Dirichlet boundary conditions can be extended to a function $\tilde{g}$ that is convex on Ω. Note that the existence of $\tilde{g}$ is guaranteed if the domain is strictly convex. However, due to our assumption that $u|_{\partial\Omega} = 0$, we may simply take $\tilde{g} \equiv 0$ in our setting. We further assume that $f \in C(\bar{\Omega})$ with $f \geq C > 0$ on Ω. Let $\{f_m\}_{m=0}^{\infty} \subset C^{\infty}(\bar{\Omega})$ be a sequence of approximations of f with $f_m \to f$ uniformly on $\bar{\Omega}$ and $f_m \geq C > 0$ for all m. We then consider the PDE problem

$$\det D^2 u_m = f_m \quad \text{in } \Omega, \tag{103a}$$

$$u_m = 0 \quad \text{on } \partial\Omega. \tag{103b}$$

Even though the source data of this problem is smooth, in general there does not exist smooth solutions to (103) because Ω is not necessarily strictly convex nor smooth, see Theorem 1. Nonetheless, there exists a unique (convex) Alexandrov solution $u_m \in C(\bar{\Omega})$.

Let $\tilde{\Omega} \subset \Omega$ be a strict subdomain of Ω that is polyhedral and convex, and let $\tilde{\mathcal{T}}_h$ be a simplicial triangulation of $\tilde{\Omega}$. Finally, we denote by $\tilde{X}_h$ a $C^1(\bar{\tilde{\Omega}})$-conforming finite element space consisting of piecewise polynomials with respect to $\tilde{\mathcal{T}}_h$. We then consider the finite element method: Find $\tilde{u}_h \in \tilde{X}_h$ satisfying $\tilde{u}_{m,h} = u_m$ on $\partial\tilde{\Omega}$ and

$$\int_{\tilde{\Omega}} (f_m - \det D^2 \tilde{u}_{m,h}) v_h \mathrm{d}x = 0 \quad \forall v_h \in \tilde{X}_h \cap W_0^{1,d}(\tilde{\Omega}). \tag{104}$$

This method is similar to (76), the differences being

(i) the problem is posed on $\tilde{\Omega}$ instead of Ω;
(ii) the source function has been regularized;
(iii) the homogeneous Dirichlet boundary conditions have been replaced by

$$\tilde{u}_{m,h}|_{\partial\tilde{\Omega}} = u_m|_{\partial\tilde{\Omega}}.$$

It is clear that method (104) is a discretization of the PDE problem

$$\det D^2 \tilde{u}_m = f_m \qquad \text{in } \tilde{\Omega}, \tag{105a}$$

$$\tilde{u}_m = u_m \qquad \text{on } \partial\tilde{\Omega}, \tag{105b}$$

which, similar to (103), has a unique Alexandrov solution and is generally nonsmooth. In fact, it is simple to see that, due to the inclusion $\tilde{\Omega} \subset \Omega$ and the uniqueness of Alexandrov solutions, that $\tilde{u}_m = u_m$ on $\tilde{\Omega}$.

Theorem 18 (interior convergence)**.**
There exists $h_0 > 0$, which depends on $\operatorname{dist}\{\partial\Omega, \partial\tilde{\Omega}\}$*, such that for $h \leq h_0$, there exists a locally unique solution to* (104)*. In addition, as $h \to 0$, $\tilde{u}_{m,h}$ converges uniformly to $\tilde{u}_m$ (the solution to* (105)*) on compact subsets of $\tilde{\Omega}$.*

Proof. The proof relies on a series of smooth approximations to problem (105). Let $\{\Omega_s\}_{s=0}^{\infty}$ be a sequence of strictly convex and smooth domains such that $\Omega_s \subset \Omega_{s+1} \subset \Omega$ for all s, and $\Omega_s \to \Omega$ as $s \to \infty$; see Fig. 8. Consider the problem

$$\begin{aligned} \det D^2 u_{ms} &= f_m \quad \text{in } \Omega_s, \\ u_{ms} &= 0 \quad \text{on } \partial\Omega_s. \end{aligned}$$

Note that, because the data is regular, and since Ω_s is uniformly convex with smooth boundary, the solution to this problem is smooth. In particular,

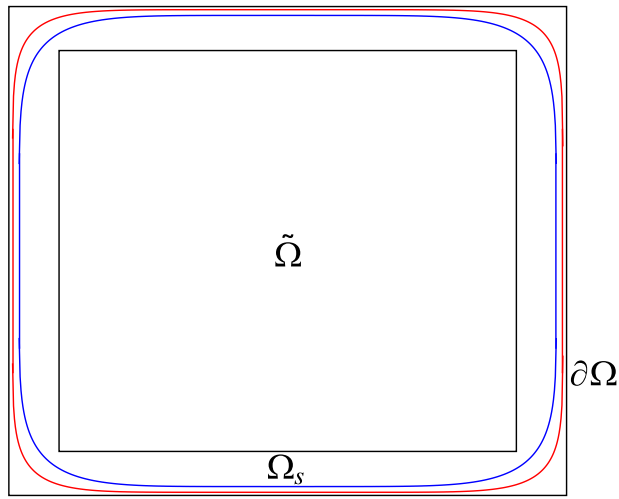

FIG. 8 Pictorial description of the proof of Theorem 18. Here, $\tilde{\Omega} \subset \Omega_s \subset \Omega_{s+1} \subset \Omega$, where Ω is the physical domain, $\tilde{\Omega}$ is the computational domain, and $\{\Omega_s\}$ are smooth and uniformly convex approximations to Ω.

interior Schauder estimates (Gilbarg and Trudinger, 2001, Section 6.1) show that, for any $D \subset\subset \Omega_s$,

$$\|u_{ms}\|_{C^{r+1}(D)} \leq C_m,$$

where $C_m > 0$ depends on m, f_m, D, and $\text{dist}\{D, \partial\Omega_s\} \leq \text{dist}\{D, \partial\Omega\}$. Moreover, results in Savin (2013) show that u_{ms} (up to subsequence) converges uniformly on compact subsets of Ω as $s \to \infty$. Now, because u_{ms} is smooth, and because the derivatives of u_{ms} are uniformly bounded on $\tilde{\Omega}$ (with respect to s), arguments similar those given in the previous section (see Awanou, 2015d; Böhmer, 2008) show that, for $h \leq h_0$ with h_0 sufficiently small, there exists a locally unique and convex solution to the following discrete problem: Find $\tilde{u}_{ms,h} \in \tilde{X}_h$ satisfying $\tilde{u}_{ms,h}|_{\partial\tilde{\Omega}} = u_{ms}|_{\partial\tilde{\Omega}}$ and

$$\int_{\tilde{\Omega}} (f_m - \det D^2 \tilde{u}_{ms,h}) v_h \mathrm{d}x = 0 \quad \forall v_h \in \tilde{X}_h \cap W_0^{1,d}(\tilde{\Omega}).$$

Furthermore, there holds $\|u_{ms} - \tilde{u}_{ms,h}\|_{W^{2,2}(\tilde{\Omega})} \leq Ch^{r-1}$ where $C > 0$ depends on $\|u_{ms}\|_{C^{r+1}(\tilde{\Omega})}$ but is independent of s. Because $\|u_{ms}\|_{C^{r+1}(\tilde{\Omega})}$ is uniformly bounded with respect to s, it follows from a Sobolev embedding theorem that $\tilde{u}_{ms,h}$ is uniformly bounded. Thus, since $u_{ms,h}$ is convex and uniformly bounded, the sequence $\{\tilde{u}_{ms,h}\}_s$ is locally uniformly equicontinuous, and thus has a pointwise convergent subsequence. Standard arguments, along with $u_{ms} \to u_m$ on $\partial\tilde{\Omega}$, then show that this limit is a solution to the discrete problem (104). □

Remark 23 (interior convergence).
Regarding Theorem 18 note that:

1. The ideas and techniques given in this section has been applied to standard finite difference discretizations of the Monge–Ampère problem in Awanou (2016).
2. While the results and techniques of Theorem 18 are interesting, it is not immediately clear how to obtain the Dirichlet boundary condition $\tilde{u}_{m,h} = u_m$ on $\partial\tilde{\Omega}$ since u_m is not given data. One can alternatively use $\tilde{u}_{m,h}|_{\partial\tilde{\Omega}} = 0$, but this condition is not consistent with problem (103). We also point out that h_0 depends on $\text{dist}\{\partial\tilde{\Omega}, \partial\Omega\}$, and therefore Theorem 18 suggests we cannot take $\tilde{\Omega}$ to be arbitrarily close to Ω. ■

5 Numerical examples

The high point of this classical algorithmic age was perhaps reached in the work of Leonhard Euler […] Innumerable numerical examples are dispersed in the (so far) seventy volumes of his collected works, showing that Euler always kept foremost in his mind the immediate numerical use of his formulas and algorithms.

Henrici (1964)

In this section we perform some simple numerical examples to show the efficiency and accuracy of some of the numerical schemes discussed in the previous sections. We consider three different test problems, each reflecting different scenarios of regularity. These are computed using the wide stencil finite difference scheme (20), the analogous filtered scheme (26), Oliker–Prussner method (64), the C^0 finite element method (83), and its regularized version using the vanishing moment methodology (102). We emphasize that these tests are not meant to form comparisons, but rather to highlight their advantages in different situations.

5.1 Example 1: Smooth solution

In the first set of experiments, we take the data such that the Monge–Ampère equation has a $C^\infty(\Omega)$ solution: $\Omega = (-1, 1)^2$,

$$f(x_1,x_2) = (1+x_1^2+x_2^2)e^{x_1^2+x_2^2}, \quad u(x_1,x_2) = e^{\frac{x_1^2+x_2^2}{2}}. \tag{106}$$

In this setting, the Galerkin methods discussed in Sections 4.1 and 4.2 are advantageous due to their relative high order. We implement the C^0 finite element method (83) and the Oliker–Prussner method (64) on a sequence of mesh refinements and report the resulting errors in Fig. 9. In agreement with Theorem 89 (with $\ell = r - 2$ and $\alpha = 1$), the plots show optimal order convergence in $W_h^{2,p}$-norm with respect to the discretization parameter h for the Galerkin methods. In terms of the degrees of freedom (DOFs), the errors scale like

$$\|u-u_h\|_{W_h^{2,p}(\Omega)} = \mathcal{O}(DOFs^{(1-r)/2}).$$

The errors in L^∞ converge with optimal order provided that the polynomial degree is sufficiently high. Fig. 9 shows that

$$\|u-u_h\|_{L^\infty(\Omega)} = \mathcal{O}(DOFs^{(-1-r)/2}) \quad r=3,4,$$
$$\|u-u_h\|_{L^\infty(\Omega)} = \mathcal{O}(DOFs^{-1}) \quad r=2.$$

These rates are proven in Neilan (2013). For the Oliker–Prussner method and finite difference methods defined on translation invariant meshes, we define its $W^{2,p}$ error on the nodal set as

$$\|u-u_h\|_{W_h^{2,p}(\Omega_h)} = \left(h^d \sum_{x_h \in \Omega_h^i,\, \boldsymbol{e}_j \in S} |\Delta_{\boldsymbol{e}_j}(u-u_h)(x_h)|^p \right)^{1/p}$$

where S is the 9-points stencil in two space dimensions and $\Delta_{\boldsymbol{e}_j} v(x_h)$ denotes the centred second difference, defined in (17), of the function v at node x_h in the direction $\boldsymbol{e}_j$. We observe in Fig. 9 that, for the Oliker–Prussner method,

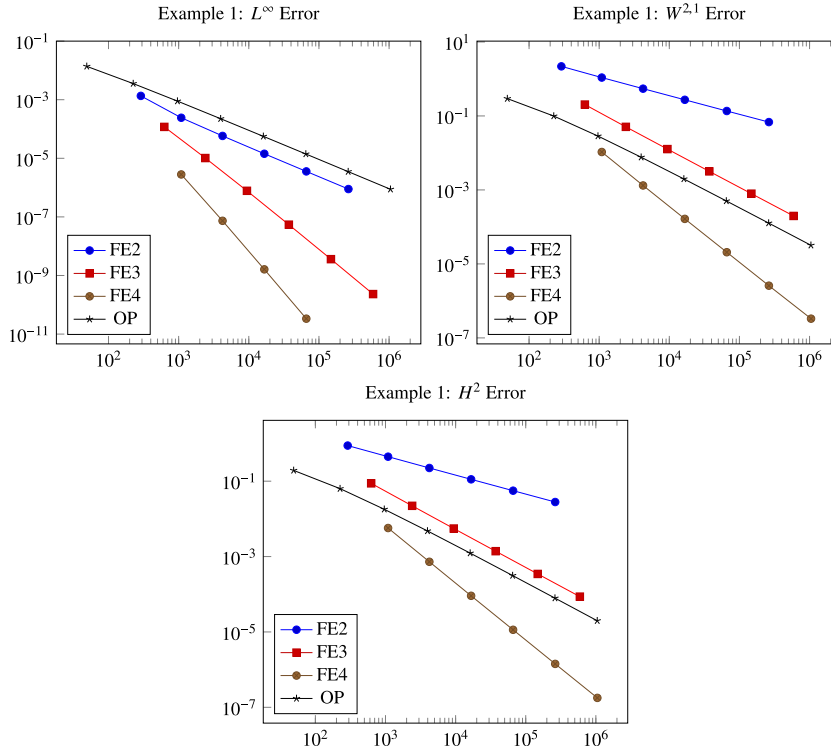

FIG. 9 Example 1: Errors versus degrees of freedom for the C^0 finite element method (83) with polynomial degrees $r = 2, 3, 4$, and the Oliker–Prussner method (64) applied to the smooth test problem (106).

$$\|u - u_h\|_{W_h^{2,p}(\Omega_h)} = \mathcal{O}(DOFs^{-1}) \quad \text{and} \quad \|u - u_h\|_{L^\infty(\Omega)} = \mathcal{O}(DOFs^{-1})$$

These results on $W^{2,\,p}$ error are consistent with the theorems proven in Neilan and Zhang (2018) and Theorem 13.

5.2 Example 2: Nonclassical solution

In this set of experiments, we again take $\Omega = (-1,\,1)^2$, but choose the data such that the resulting solution is not a classical one:

$$f(x_1,x_2) = \begin{cases} 16, & |x| \le 1/2, \\ 64 - 16|x|^{-1}, & |x| > 1/2. \end{cases}$$

$$u(x_1,x_2) = \begin{cases} 2|x|^2, & |x| \le 1/2, \\ 2(|x| - 1/2)^2 + 2|x|^2, & |x| > 1/2. \end{cases}$$

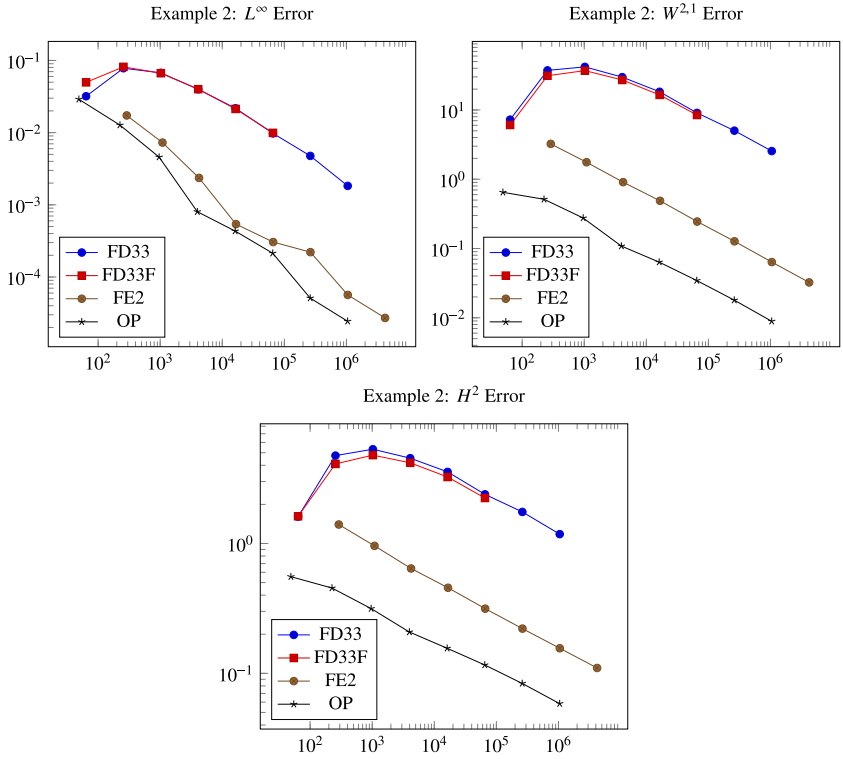

FIG. 10 Example 2: Errors versus degrees of freedom for the 33-point wide stencil scheme, 33-point wide stencil filtered scheme, the quadratic C^0 finite element method and Oliker–Prussner method.

One easily finds that $u \notin C^{1,1}(\bar{\Omega}) \backslash C^2(\Omega)$. We implement the C^0 finite element method (83), the wide stencil finite difference scheme (20) with a stencil size that consists of 33 grid points, and the Oliker–Prussner method (64). We also compare the results with the filtered scheme (26) The errors, depicted in Fig. 10, show that all methods converge with similar rates, although the finite element scheme and Oliker–Prussner method have smaller errors with similar DOFs. While the rate of convergence in the L^∞ norm is not obvious from the tests, Fig. 10 clearly shows that all three methods converge in the $W^{2,p}$-norms with rates

$$\|u - u_h\|_{H_h^2(\Omega)} = \mathcal{O}(DOFs^{-1/4}), \quad \|u - u_h\|_{W_h^{2,1}(\Omega)} = \mathcal{O}(DOFs^{-1/2}). \tag{107}$$

We note that, for the finite element, these rates seem to be the same rates of interpolation errors. Indeed, let $\mathcal{T}_h^\Gamma$ denote the set of triangles in $\mathcal{T}_h$ intersect the circle $|x| = 1/2$. Likewise, we let $\mathcal{F}_h^\Gamma$ denote the set of edges in $\mathcal{F}_h^I$ that intersect Γ. Finally, we denote by $\mathcal{I}_h u$ the nodal interpolant of u.

Because u is smooth on both $\Omega \cap \{x \in \Omega : |x| < 1/2\}$ and $\Omega \cap \{x \in \Omega : x > 1/2\}$, we have by standard interpolation estimates,

$$\begin{aligned}\|u - \mathcal{I}_h u\|^p_{W_h^{2,p}(\Omega)} &\leq Ch^{p(r-1)} + \sum_{T \in \mathcal{T}_h^\Gamma} \|D^2(u - \mathcal{I}_h u)\|^p_{L^p(T)} \\ &\quad + \sum_{F \in \mathcal{F}_h^\Gamma} h_F^{1-p} \|[\![\nabla(u - \mathcal{I}_h u)]\!]\|^p_{L^p(F)} \\ &\leq C\left(h^{p(r-1)} + \sum_{T \in \mathcal{T}_h^\Gamma} \left(\|D^2(u - \mathcal{I}_h u)\|^p_{L^p(T)} + h_T^{-p} \|\nabla(u - \mathcal{I}_h u)\|^p_{L^p(T)}\right)\right),\end{aligned}$$

where we used a standard trace inequality. Applying interpolation estimates and Hölder's inequality, noting that $u \in W^{2,\infty}(\Omega)$, yields

$$\begin{aligned}\|u - \mathcal{I}_h u\|^p_{W_h^{2,p}(\Omega)} &\leq C\left(h^{p(r-1)} + \sum_{T \in \mathcal{T}_h^\Gamma} \|D^2 u\|^p_{L^p(T)}\right) \\ &\leq C\left(h^{p(r-1)} + \sum_{T \in \mathcal{T}_h^\Gamma} h_T^2 \|D^2 u\|^p_{L^\infty(T)}\right) \\ &\leq Ch^{p(r-1)} + Ch,\end{aligned}$$

where we used that the cardinality of $\mathcal{T}_h^\Gamma$ is $\mathcal{O}(h^{-1})$. We then take the pth root of this inequality to deduce that $\|u - \mathcal{I}_h u_h\|_{W_h^{2,p}(\Omega)} = \mathcal{O}(h^{1/p}) = \mathcal{O}(DOFs^{-1/(2p)})$, which is the same rates as (107).

5.3 Example 3: Lipschitz and degenerate solution

In our last set of experiments, we take the domain to be $\Omega = (-1, 1)^2$ with data

$$f(x_1, x_2) = \begin{cases} 36 - 9x_2^2 x_1^{-6}, & |x_2| \leq |x_1|^3, \\ \frac{8}{9} - \frac{5}{9} x_1^2 x_2^{-\frac{2}{3}}, & |x_2| > |x_1|^3, \end{cases}$$

$$u(x_1, x_2) = \begin{cases} |x_1|^4 + \frac{3x_2^2}{2x_1^2}, & |x_2| \leq |x_1|^3, \\ \frac{1}{2} x_1^2 x_2^{\frac{2}{3}} + 2x_2^{\frac{4}{3}}, & |x_2| > |x_1|^3. \end{cases}$$

Similar to the previous example, u is not a classical solution to (1) as it only has regularity $u \in C^{0,\,1}(\Omega)$ and $u \notin W^{2,\,p}(\Omega)$ for any $p > 2$ (Wang, 1995). Moreover, a simple calculation shows that $|D^2 u(x)| \to \infty$ as $x \to 0$. Since the determinant in two dimensions is the product of two eigenvalues of the Hessian and $\det D^2 u(x) = f(x)$ is bounded in the domain, the largest eigenvalue blows up while the other eigenvalue of $D^2 u(x)$ approaches zero as $x \to 0$. Hence, the Hessian of the solution degenerates as $x \to 0$.

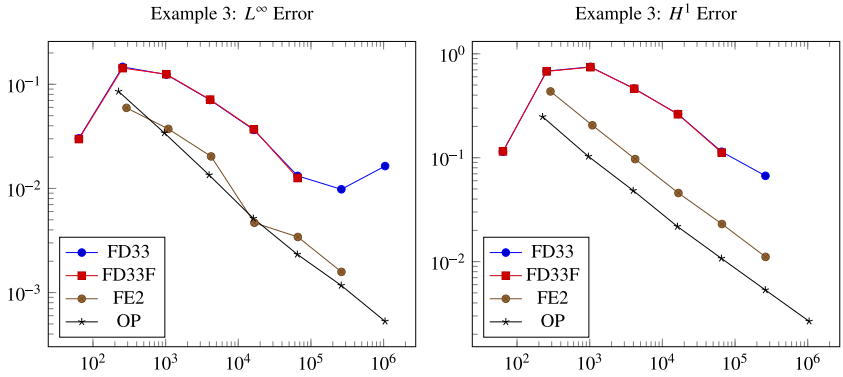

FIG. 11 Example 3: Errors versus degrees of freedom for the 33-point wide stencil scheme, the 33-point wide stencil filtered scheme, and the quadratic C^0 finite element method with regularization and Oliker–Prussner method.

While the monotone finite difference schemes presented in Section 2 are robust for problems with low regularity, Galerkin methods generally fail to capture solutions whose second derivatives are not square integrable; our numerical tests show that Newton's method applied to (83) does not converge for this example even when using very generous initial guesses. In fact, even for the monotone finite difference schemes and the Oliker–Prussner method, Newton's method is very sensitive with respect to the initial guess and the convexity of the iterates for this problem. In our implementation, we found that at each iteration, we require the solution to remain convex. As Newton's method may not give a convex solution in general, we applied, if necessary, the algorithm proposed in Oberman (2008a) to preserve convexity.

In addition to the 33-point finite difference scheme and Oliker–Prussner method, we implement the fourth-order regularization of the C^0 finite element method (83) with parameters $\sigma = 100$ and $\epsilon = 0.1h^2$. The resulting errors measured in the L^∞ and H^1 norms are plotted in Fig. 11. Similar to the previous series of experiments, the plots show that both methods have similar behaviour rates. While the rate in the L^∞ is not clear, the second plot in Fig. 11 shows that

$$\|u - u_h\|_{H^1(\Omega)} = \mathcal{O}(DOFs^{-1/2}).$$

6 Concluding remarks

"And if anyone knows anything about anything" said Bear to himself, "it's Owl who knows something about something," he said, "or my name is not Winnie-the-Pooh," he said. "Which it is," he added. "So there you are."

Hoff (1982)

In this work we have reviewed the progress that has been made concerning the approximation and numerical analysis of the Monge–Ampère problem. In doing so we highlighted how to develop a convergence analysis of wide stencil finite difference schemes as well as their generalizations, schemes based on geometric considerations, and finite element methods. A focus that we have taken, and one of recent development, is the derivation of rates of convergence for these discretizations.

Despite fundamental advances in only the past decade, there still remain several open problems in the analysis of computational methods for Monge–Ampère problems. One of these is the derivation of rates of convergence for the Oliker–Prussner scheme on unstructured grids. Another basic problem is rates of convergence of any of the schemes presented in this work assuming that the solution is not a classical one, i.e., without the assumption $u \in C^{2,\alpha}(\bar{\Omega})$. In most of the error analyses we have presented, it is assumed that $0 < \lambda I \leq D^2 u(x) \leq \Lambda I$ for all $x \in \Omega$. However, if the function $f(x)$ is discontinuous, the Hessian of the solution may be degenerate as the third example in the numerics section illustrates. The design and analysis of robust and high order numerical schemes to capture degenerate solutions remains a challenging problem. A posteriori error estimation, and adaptive methods based on the existing schemes are nonexistent. Finally let us mention that, as far as we are aware, except for the recent work (Berman, 2018), rates of convergence are restricted to the Dirichlet problem (1); extensions to, e.g., the applications discussed in Section 1.1 is still unchartered territory.

In conclusion, we know something about the numerical analysis of the Monge–Ampère problem, but there is much more that needs to be developed. It is our hope that this overview will encourage the numerical analysis community to work on the interesting, and challenging, problems found in geometry in general, and those that the Monge–Ampère equation in particular present to us.

Acknowledgements

The work of M.J.N. was supported by NSF Grant DMS–1719829. The work of A.J.S. was supported by NSF Grant DMS–1720213. The work of W.Z. was supported by NSF Grant DMS–1818861.

References

Arnol'd, V., 1998. On the teaching of mathematics. Uspekhi Mat. Nauk 53 (1), 229–234. ISSN 0042-1316. https://doi.org/10.1070/rm1998v053n01ABEH000005.

Arnold, D., Brezzi, F., Cockburn, B., Marini, L., 2002. Unified analysis of discontinuous Galerkin methods for elliptic problems. SIAM J. Numer. Anal. 39 (5), 1749–1779. ISSN 0036-1429. https://doi.org/10.1137/S0036142901384162.

Awanou, G., 2014. Isogeometric method for the elliptic Monge-Ampère equation. In: Approximation Theory XIV: San Antonio 2013, Springer Proc. Math. Stat, vol. 83. Springer, Cham, pp. 1–13. https://doi.org/10.1007/978-3-319-06404-8_1.

Awanou, G., 2015. Quadratic mixed finite element approximations of the Monge-Ampère equation in 2D. Calcolo 52 (4), 503–518. ISSN 0008-0624. https://doi.org/10.1007/s10092-014-0127-7.

Awanou, G., 2015. Smooth approximations of the Aleksandrov solution of the Monge-Ampère equation. Commun. Math. Sci. 13 (2), 427–441. ISSN 1539-6746. https://doi.org/10.4310/CMS.2015.v13.n2.a8.

Awanou, G., 2015. Spline element method for Monge-Ampère equations. BIT 55 (3), 625–646. ISSN 0006-3835. https://doi.org/10.1007/s10543-014-0524-y.

Awanou, G., 2015. Standard finite elements for the numerical resolution of the elliptic Monge-Ampère equations: classical solutions. IMA J. Numer. Anal. 35 (3), 1150–1166. ISSN 0272-4979. https://doi.org/10.1093/imanum/dru028.

Awanou, G., 2016. On standard finite difference discretizations of the elliptic Monge-Ampère equation. J. Sci. Comput. 69 (2), 892–904. ISSN 0885-7474. https://doi.org/10.1007/s10915-016-0220-y.

Awanou, G., 2017. Erratum to: Quadratic mixed finite element approximations of the Monge-Ampère equation in 2D [MR3421667]. Calcolo 54 (1), 281–297. ISSN 0008-0624. https://doi.org/10.1007/s10092-016-0187-y.

Awanou, G., 2017. Standard finite elements for the numerical resolution of the elliptic Monge-Ampère equation: Aleksandrov solutions. ESAIM Math. Model. Numer. Anal. 51 (2), 707–725. ISSN 0764-583X. https://doi.org/10.1051/m2an/2016037.

Awanou, G., Awi, R., 2016. Convergence of finite difference schemes to the Aleksandrov solution of the Monge-Ampère equation. Acta Appl. Math. 144, 87–98. ISSN 0167-8019. https://doi.org/10.1007/s10440-016-0041-x.

Awanou, G., Li, H., 2014. Error analysis of a mixed finite element method for the Monge-Ampère equation. Int. J. Numer. Anal. Model. 11 (4), 745–761. ISSN 1705-5105.

Awanou, G., Li, H., Malitz, E., 2018. A two-grid method for the C0 interior penalty discretization of the Monge-Ampère equation. Preprint.

Bakelman, I., 1994. Convex Analysis and Nonlinear Geometric Elliptic Equations. Springer-Verlag, Berlin, ISBN: 3-540-13620-7, pp. xxii–510. https://doi.org/10.1007/978-3-642-69881-1.

Barles, G., Souganidis, P.E., 1991. Convergence of approximation schemes for fully nonlinear second order equations. Asymptotic Anal. 4 (3), 271–283. ISSN 0921-7134.

Benamou, J., Duval, V., 2018. Minimal convex extensions and finite difference discretisation of the quadratic Monge–Kantorovich problem. ArXiv:1710.05594 [math.NA].

Benamou, J.-D., Froese, B., Oberman, A., 2014. Numerical solution of the optimal transportation problem using the Monge-Ampère equation. J. Comput. Phys. 260, 107–126. ISSN 0021-9991. https://doi.org/10.1016/j.jcp.2013.12.015.

Benamou, J.-D., Collino, F., Mirebeau, J.-M., 2016. Monotone and consistent discretization of the Monge-Ampère operator. Math. Comp. 85 (302), 2743–2775. ISSN 0025-5718. https://doi.org/10.1090/mcom/3080.

Berman, R., 2018. Convergence rates for discretized Monge-Ampère equations and quantitative stability of optimal transport. ArXiv:1803.00785 [math.NA].

Böhmer, K., 2008. On finite element methods for fully nonlinear elliptic equations of second order. SIAM J. Numer. Anal. 46 (3), 1212–1249. ISSN 0036-1429. https://doi.org/10.1137/040621740.

Bonito, A., Guermond, J.-L., Popov, B., 2014. Stability analysis of explicit entropy viscosity methods for non-linear scalar conservation equations. Math. Comp. 83 (287), 1039–1062. ISSN 0025-5718. https://doi.org/10.1090/S0025-5718-2013-02771-8.

Brenier, Y., 1991. Polar factorization and monotone rearrangement of vector-valued functions. Comm. Pure Appl. Math. 44 (4), 375–417. ISSN 0010-3640. https://doi.org/10.1002/cpa.3160440402.

Brenner, S., Neilan, M., 2012. Finite element approximations of the three dimensional Monge-Ampère equation. ESAIM Math. Model. Numer. Anal. 46 (5), 979–1001. ISSN 0764-583X. https://doi.org/10.1051/m2an/2011067.

Brenner, S., Scott, L., 2008. The Mathematical Theory of Finite Element Methods. In: Texts in Applied Mathematics, third ed 15. Springer, New York. ISBN: 978-0-387-75933-3, pp. xviii–397. https://doi.org/10.1007/978-0-387-75934-0.

Brenner, S., Sung, L.-Y., 2005. C^0 interior penalty methods for fourth order elliptic boundary value problems on polygonal domains. J. Sci. Comput. 22/23, 83–118. ISSN 0885-7474. https://doi.org/10.1007/s10915-004-4135-7.

Brenner, S., Gudi, T., Neilan, M., Sung, L.-Y., 2011. C^0 penalty methods for the fully nonlinear Monge-Ampère equation. Math. Comp. 80 (276), 1979–1995. ISSN 0025-5718. https://doi.org/10.1090/S0025-5718-2011-02487-7.

Caffarelli, L., Cabré, X., 1995. Fully Nonlinear Elliptic Equations. American Mathematical Society Colloquium Publications, vol. 43American Mathematical Society, Providence, RI. ISBN: 0-8218-0437-5, p. vi+104.

Calabi, E., 1990. Affine differential geometry and holomorphic curves. In: Villani, V. (Ed.), Complex Geometry and Analysis (Pisa, 1988), Lecture Notes in Math., vol. 1422. Springer, Berlin, pp. 15–21. 10.1007/BFb0089401.

Ciarlet, P., 2002. The Finite Element Method for Elliptic Problems. Classics in Applied Mathematics, vol. 40. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA. ISBN: 0-89871-514-8, p. xxviii+530. https://doi.org/10.1137/1.9780898719208.

Cockburn, B., Shu, C.-W., 1998. The local discontinuous Galerkin method for time-dependent convection-diffusion systems. SIAM J. Numer. Anal. 35 (6), 2440–2463. ISSN 0036-1429. https://doi.org/10.1137/S0036142997316712.

Courant, R., Friedrichs, K., Lewy, H., 1967. On the partial difference equations of mathematical physics. IBM J. Res. Develop. 11, 215–234. ISSN 0018-8646. https://doi.org/10.1147/rd.112.0215.

Crandall, M., Ishii, H., Lions, P.-L., 1992. User's guide to viscosity solutions of second order partial differential equations. Bull. Amer. Math. Soc. (N.S.) 27 (1), 1–67. ISSN 0273-0979. https://doi.org/10.1090/S0273-0979-1992-00266-5.

Cuesta, J., Matrán, C., 1989. Notes on the Wasserstein metric in Hilbert spaces. Ann. Probab. 17 (3), 1264–1276. ISSN 0091-1798. http:/links.jstor.org/sici?sici=0091-1798(198907)17:3%3C1264:NOTWMI%3E2.0.CO;2-J&origin=MSN.

Davydov, O., Saeed, A., 2013. Numerical solution of fully nonlinear elliptic equations by Böhmer's method. J. Comput. Appl. Math. 254, 43–54. ISSN 0377-0427. https://doi.org/10.1016/j.cam.2013.03.009.

De Philippis, G., Figalli, A., 2015. Optimal regularity of the convex envelope. Trans. Amer. Math. Soc. 367 (6), 4407–4422. ISSN 0002-9947. https://doi.org/10.1090/S0002-9947-2014-06306-X.

Dean, E., Glowinski, R., 2003. Numerical solution of the two-dimensional elliptic Monge-Ampère equation with Dirichlet boundary conditions: an augmented Lagrangian approach. C. R. Math. Acad. Sci. Paris 336 (9), 779–784. ISSN 1631-073X. https://doi.org/10.1016/S1631-073X(03)00149-3.

Dean, E., Glowinski, R., 2004. Numerical solution of the two-dimensional elliptic Monge-Ampère equation with Dirichlet boundary conditions: a least-squares approach. C. R. Math. Acad. Sci. Paris 339 (12), 887–892. ISSN 1631-073X. https://doi.org/10.1016/j.crma.2004.09.018.

Dean, E., Glowinski, R., 2005. On the numerical solution of a two-dimensional Pucci's equation with Dirichlet boundary conditions: a least-squares approach. C. R. Math. Acad. Sci. Paris 341 (6), 375–380. ISSN 1631-073X. https://doi.org/10.1016/j.crma.2005.08.002.

Dean, E., Glowinski, R., 2006a. An augmented Lagrangian approach to the numerical solution of the Dirichlet problem for the elliptic Monge-Amp'ere equation in two dimensions. Electron. Trans. Numer. Anal. 22, 71–96. ISSN 1068-9613 (electronic).

Dean, E., Glowinski, R., 2006. Numerical methods for fully nonlinear elliptic equations of the Monge-Ampère type. Comput. Methods Appl. Mech. Engrg. 195 (13–16), 1344–1386. ISSN 0045-7825. https://doi.org/10.1016/j.cma.2005.05.023.

Debrabant, K., Jakobsen, E., 2013. Semi-Lagrangian schemes for linear and fully non-linear diffusion equations. Math. Comp. 82 (283), 1433–1462. ISSN 0025-5718. https://doi.org/10.1090/S0025-5718-2012-02632-9.

Engel, G., Garikipati, K., Hughes, T., Larson, M., Mazzei, L., Taylor, R., 2002. Continuous/discontinuous finite element approximations of fourth-order elliptic problems in structural and continuum mechanics with applications to thin beams and plates, and strain gradient elasticity. Comput. Methods Appl. Mech. Engrg. 191 (34), 3669–3750. ISSN 0045-7825. https://doi.org/10.1016/S0045-7825(02)00286-4.

Feng, X., Jensen, M., 2017. Convergent semi-Lagrangian methods for the Monge-Ampère equation on unstructured grids. SIAM J. Numer. Anal. 55 (2), 691–712. ISSN 0036-1429. https://doi.org/10.1137/16M1061709.

Feng, X., Neilan, M., 2009. Vanishing moment method and moment solutions for fully nonlinear second order partial differential equations. J. Sci. Comput. 38 (1), 74–98. ISSN 0885-7474. https://doi.org/10.1007/s10915-008-9221-9.

Feng, X., Neilan, M., 2011. Analysis of Galerkin methods for the fully nonlinear Monge-Ampère equation. J. Sci. Comput. 47 (3), 303–327. ISSN 0885-7474. https://doi.org/10.1007/s10915-010-9439-1.

Feng, X., Neilan, M., 2014. Convergence of a fourth-order singular perturbation of the n-dimensional radially symmetric Monge-Ampère equation. Appl. Anal. 93 (8), 1626–1646. ISSN 0003-6811. https://doi.org/10.1080/00036811.2013.842228.

Feng, X., Glowinski, R., Neilan, M., 2013. Recent developments in numerical methods for fully nonlinear second order partial differential equations. SIAM Rev. 55 (2), 205–267. ISSN 0036-1445. https://doi.org/10.1137/110825960.

Feng, X., Hennings, L., Neilan, M., 2017. Finite element methods for second order linear elliptic partial differential equations in non-divergence form. Math. Comp. 86 (307), 2025–2051. ISSN 0025-5718. https://doi.org/10.1090/mcom/3168.

Feng, X., Neilan, M., Schnake, S., 2018. Interior penalty discontinuous Galerkin methods for second order linear non-divergence form elliptic PDEs. J. Sci. Comput. 74 (3), 1651–1676. ISSN 0885-7474. https://doi.org/10.1007/s10915-017-0519-3.

Figalli, A., 2017. The Monge-Ampère Equation and Its Applications. Zurich Lectures in Advanced MathematicsEuropean Mathematical Society (EMS), Zürich. ISBN: 978-3-03719-170-5, p. x+200. https://doi.org/10.4171/170.

Froese, B., 2012. A numerical method for the elliptic Monge-Ampère equation with transport boundary conditions. SIAM J. Sci. Comput. 34 (3), A1432–A1459. ISSN 1064-8275. https://doi.org/10.1137/110822372.

Froese, B., 2018. Meshfree finite difference approximations for functions of the eigenvalues of the Hessian. Numer. Math. 138 (1), 75–99. ISSN 0029-599X. https://doi.org/10.1007/s00211-017-0898-2.

Froese, B., Oberman, A., 2011. Convergent finite difference solvers for viscosity solutions of the elliptic Monge-Ampère equation in dimensions two and higher. SIAM J. Numer. Anal. 49 (4), 1692–1714. ISSN 0036-1429. https://doi.org/10.1137/100803092.

Froese, B., Oberman, A., 2011. Fast finite difference solvers for singular solutions of the elliptic Monge-Ampère equation. J. Comput. Phys. 230 (3), 818–834. ISSN 0021-9991. https://doi.org/10.1016/j.jcp.2010.10.020.

Froese, B., Oberman, A., 2013. Convergent filtered schemes for the Monge-Ampère partial differential equation. SIAM J. Numer. Anal. 51 (1), 423–444. ISSN 0036-1429. https://doi.org/10.1137/120875065.

Gilbarg, D., Trudinger, N., 2001. Elliptic Partial Differential Equations of Second Order. Classics in MathematicsSpringer-Verlag, Berlin. ISBN: 3-540-41160-7, p. xiv+517.

Guan, B., Spruck, J., 1993. Boundary-value problems on S^n for surfaces of constant Gauss curvature. Ann. of Math. (2) 138 (3), 601–624. ISSN 0003-486X. https://doi.org/10.2307/2946558.

Guermond, J.-L., Pasquetti, R., 2011. Entropy viscosity method for high-order approximations of conservation laws. In: Bittencourt, M., Dumont, N., Hesthaven, J.S. (Eds.), Spectral and High Order Methods for Partial Differential Equations, Lect. Notes Comput. Sci. Eng., vol. 76. Springer, Heidelberg, pp. 411–418.

Guermond, J.-L., Pasquetti, R., Popov, B., 2011. Entropy viscosity method for nonlinear conservation laws. J. Comput. Phys. 230 (11), 4248–4267. ISSN 0021-9991. https://doi.org/10.1016/j.jcp.2010.11.043.

Guermond, J.-L., Nazarov, M., Popov, B., Tomas, I., 2018. Second-order invariant domain preserving approximation of the Euler equations using convex limiting. SIAM J. Sci. Comput. 40 (5), A3211–A3239. ISSN 1064-8275. https://doi.org/10.1137/17M1149961.

Gutiérrez, C., 2001. The Monge-Ampère Equation. Progress in Nonlinear Differential Equations and their Applications, 44Birkhäuser Boston, Inc., Boston, MA, . ISBN: 0-8176-4177-7xii+127. https://doi.org/10.1007/978-1-4612-0195-3.

Hamfeldt, B., 2018. Convergent approximation of non-continuous surfaces of prescribed Gaussian curvature. Commun. Pure Appl. Anal. 17 (2), 671–707. ISSN 1534-0392. https://doi.org/10.3934/cpaa.2018036.

Henrici, P., 1964. Elements of Numerical Analysis. John Wiley & Sons, Inc., New York, London, Sydney, p. xv+328.

Hintermüller, M., Ito, K., Kunisch, K., 2002. The primal-dual active set strategy as a semismooth Newton method. SIAM J. Optim 13 (3), 865–888. ISSN 1052-6234. (2003)10.1137/S1052623401383558.

Hoff, B., 1982. The Tao of Pooh. Penguin Books, p. xii+158 0 14 0 0.6747 7.

Huang, J., Huang, X., Han, W., 2010. A new C^0 discontinuous Galerkin method for Kirchhoff plates. Comput. Methods Appl. Mech. Engrg. 199 (23-24), 1446–1454. ISSN 0045-7825. https://doi.org/10.1016/j.cma.2009.12.012.

Jensen, M., 2018. Numerical solution of the simple Monge-Ampère equation with nonconvex Dirichlet data on nonconvex domains. In: Kalise, D., Kunisch, K., Rao, Z. (Eds.), Hamilton-Jacobi-Bellman Equations: Numerical Methods and Applications in Optimal Control, Radon Series on Computational and Applied Mathematics, vol. 21. De Gryuter, Boston, Berlin, pp. 129–142.

Jensen, M., Smears, I., 2018. On the notion of boundary conditions in comparison principles. In: Kalise, D., Kunisch, K., Rao, Z. (Eds.), Hamilton-Jacobi-Bellman Equations: Numerical

Methods and Applications in Optimal Control, Radon Series on Computational and Applied Mathematics, vol. 21. De Gryuter, Boston, Berlin, pp. 143–154.

Jovanović, B.S., Süli, E., 2014. Analysis of Finite Difference Schemes, Springer Series in Computational Mathematics. 46, Springer, London, pp. xiv–408. 978-1-4471-5459-4; 978-1-4471-5460-010.1007/978-1-4471-5460-0.

Kantorovich, L., 2004. On a problem of Monge. Zap. Nauchn. Sem. S.-Peterburg. Otdel. Mat. Inst. Steklov. (POMI) 312, 15–16. ISSN 0373-2703. https://doi.org/10.1007/s10958-006-0050-9 (Teor. Predst. Din. Sist. Komb. i Algoritm. Metody. 11).

Kawecki, E., Lakkis, O., Pryer, T., 2018. A finite element method for the Monge-Ampère equation with transport boundary conditions. ArXiv:1807.03535.

Kossaczký, I., Ehrhardt, M., Günther, M., 2016. On the non-existence of higher order monotone approximation schemes for HJB equations. Appl. Math. Lett. 52, 53–57. ISSN 0893-9659. https://doi.org/10.1016/j.aml.2015.08.005.

Krasnosel'skiĭ, M., Rutickiĭ, J., 1961. Convex Functions and Orlicz Spaces. Translated from the First Russian edition by Leo F. BoronP. Noordhoff Ltd., Groningen, p. xi+249

Krylov, N., 1987. Nonlinear Elliptic and Parabolic Equations of the Second Order. Mathematics and its Applications (Soviet Series), vol. 7D. Reidel Publishing Co., Dordrecht. ISBN: 90-277-2289-7, p. xiv+462. https://doi.org/10.1007/978-94-010-9557-0

Lakkis, O., Pryer, T., 2011. A finite element method for second order nonvariational elliptic problems. SIAM J. Sci. Comput. 33 (2), 786–801. ISSN 1064-8275. https://doi.org/10.1137/100787672.

Lakkis, O., Pryer, T., 2013. A finite element method for nonlinear elliptic problems. SIAM J. Sci. Comput. 35 (4), A2025–A2045. ISSN 1064-8275. https://doi.org/10.1137/120887655.

Li, W., Nochetto, R., 2018. Optimal pointwise error estimates for two-scale methods for the Monge-Ampère equation. SIAM J. Numer. Anal. 56 (3), 1915–1941. ISSN 0036-1429. https://doi.org/10.1137/18M1165670.

Li, W., Nochetto, R., 2018b. Two-scale methods for convex envelopes. ArXiv:1812.11519 [math.NA].

Lions, P.-L., Souganidis, P., 1995. Convergence of MUSCL and filtered schemes for scalar conservation laws and Hamilton-Jacobi equations. Numer. Math. 69 (4), 441–470. ISSN 0029-599X. https://doi.org/10.1007/s002110050102.

Mirebeau, J.-M., 2015. Discretization of the 3D Monge-Ampere operator, between wide stencils and power diagrams. ESAIM Math. Model. Numer. Anal. 49 (5), 1511–1523. ISSN 0764-583X.

Mirebeau, J.-M., 2016. Minimal stencils for discretizations of anisotropic PDEs preserving causality or the maximum principle. SIAM J. Numer. Anal. 54 (3), 1582–1611. ISSN 0036-1429. https://doi.org/10.1137/16M1064854.

Motzkin, T., Wasow, W., 1953. On the approximation of linear elliptic differential equations by difference equations with positive coefficients. J. Math. Physics 31, 253–259.

Neilan, M., 2013. Quadratic finite element approximations of the Monge-Ampère equation. J. Sci. Comput. 54 (1), 200–226. ISSN 0885-7474. https://doi.org/10.1007/s10915-012-9617-4.

Neilan, M., 2014. Finite element methods for fully nonlinear second order PDEs based on a discrete Hessian with applications to the Monge-Ampère equation. J. Comput. Appl. Math. 263, 351–369. ISSN 0377-0427. https://doi.org/10.1016/j.cam.2013.12.027.

Neilan, M., 2014. A unified analysis of three finite element methods for the Monge-Ampère equation. Electron. Trans. Numer. Anal. 41, 262–288. ISSN 1068-9613.

Neilan, M., 2017. Convergence analysis of a finite element method for second order non-variational elliptic problems. J. Numer. Math. 25 (3), 169–184. ISSN 1570-2820. https://doi.org/10.1515/jnma-2016-1017.

Neilan, M., Zhang, W., 2018. Rates of convergence in W_p^2-norm for the Monge-Ampère equation. SIAM J. Numer. Anal. 56 (5), 3099–3120. ISSN 0036-1429. https://doi.org/10.1137/17M1160409.

Neilan, M., Salgado, A., Zhang, W., 2017. Numerical analysis of strongly nonlinear PDEs. Acta Numer. 26, 137–303. ISSN 0962-4929. https://doi.org/10.1017/S0962492917000071.

Nochetto, R., Ntogkas, D., 2018. Convergent two-scale filtered scheme for the Monge-Ampère equation. arXiv:1807.04866.

Nochetto, R., Zhang, W., 2018. Discrete ABP estimate and convergence rates for linear elliptic equations in non-divergence form. Found. Comput. Math. 18 (3), 537–593. ISSN 1615-3375. https://doi.org/10.1007/s10208-017-9347-y.

Nochetto, R., Zhang, W., 2019. Pointwise rates of convergence for the Oliker-Prussner method for the Monge-Ampère equation. Numer. Math. https://doi.org/10.1007/s0021. To appear.

Nochetto, R., Ntogkas, D., Zhang, W., 2019. Two-scale method for the Monge-Ampère equation: convergence to the viscosity solution. Math. Comp. 88 (316), 637–664. ISSN 0025-5718. https://doi.org/10.1090/mcom/3353.

Nochetto, R., Ntogkas, D., Zhang, W., 2019. Two-scale method for the Monge-Ampère equation: pointwise error estimates. IMA J. Numer. Anal. https://doi.org/10.1093/imanum/dry026. To appear.

Norris, A., Westcott, B., 1976. Computation of reflector surfaces for bivariate beamshaping in the elliptic case. J. Phys. A: Math. Gen. 9 (12), 2159. http://stacks.iop.org/0305-4470/9/i=12/a=020.

Oberman, A., 2006. Convergent difference schemes for degenerate elliptic and parabolic equations: Hamilton-Jacobi equations and free boundary problems. SIAM J. Numer. Anal. 44 (2), 879–895. ISSN 0036-1429. (electronic). 10.1137/S0036142903435235.

Oberman, A., 2008. Computing the convex envelope using a nonlinear partial differential equation. Math. Models Methods Appl. Sci. 18 (5), 759–780. ISSN 0218-2025. https://doi.org/10.1142/S0218202508002851.

Oberman, A., 2008. Wide stencil finite difference schemes for the elliptic Monge-Ampère equation and functions of the eigenvalues of the Hessian. Discrete Contin. Dyn. Syst. Ser. B 10 (1), 221–238. ISSN 1531-3492. https://doi.org/10.3934/dcdsb.2008.10.221.

Oberman, A., Ruan, Y., 2017. A partial differential equation for the rank one convex envelope. Arch. Ration. Mech. Anal. 224 (3), 955–984. ISSN 0003-9527. https://doi.org/10.1007/s00205-017-1092-5.

Oliker, V., 1984. Hypersurfaces in $\mathbf{R}^{n+1}$ with prescribed Gaussian curvature and related equations of Monge-Ampère type. Comm. Partial Differential Equations 9 (8), 807–838. ISSN 0360-5302. https://doi.org/10.1080/03605308408820348.

Oliker, V., 1987. Near radially symmetric solutions of an inverse problem in geometric optics. Inverse Problems 3 (4), 743–756. ISSN 0266-5611. http:/stacks.iop.org/0266-5611/3/743.

Oliker, V., Newman, E., 1993. The energy conservation equation in the reflector mapping problem. Appl. Math. Lett. 6 (1), 91–95. ISSN 0893-9659. https://doi.org/10.1016/0893-9659(93)90156-H.

Oliker, V., Prussner, L., 1988. On the numerical solution of the equation $(\partial^2 z/\partial x^2)(\partial^2 z/\partial y^2) - ((\partial^2 z/\partial x \partial y))^2 = f$ and its discretizations. I. Numer. Math. 54 (3), 271–293. ISSN 0029-599X. https://doi.org/10.1007/BF01396762.

Oliker, V., Waltman, P., 1987. Radially symmetric solutions of a Monge-Ampère equation arising in a reflector mapping problem. In: Knowles, I.W., Saitō, Y. (Eds.), Differential Equations and Mathematical Physics, Lecture Notes in Math, 1285, Springer, Berlin, pp. 361–374. https:/doi.org/10.1007/BFb0080616.

Pólya, G., 2014. How to Solve It. Princeton Science Library. Princeton University Press, Princeton, NJ, ISBN: 978-0-691-16407-6, p. xxviii+253.

Rüschendorf, L., Rachev, S., 1990. A characterization of random variables with minimum L^2-distance. J. Multivariate Anal. 32 (1), 48–54. ISSN 0047-259X. https://doi.org/10.1016/0047-259X(90)90070-X.

Rüschendorf, L., Rachev, S., 1990. Corrigendum: "A characterization of random variables with minimum L^2-distance" J. Multivariate Anal. 34 (1), 156. ISSN 0047-259X. https://doi.org/10.1016/0047-259X(90)90066-Q.

Russell, B., 1931. The Scientific Outlook. Routledge.

Savin, O., 2013. Pointwise $C^{2,\alpha}$ estimates at the boundary for the Monge-Ampère equation. J. Amer. Math. Soc. 26 (1), 63–99. ISSN 0894-0347. https://doi.org/10.1090/S0894-0347-2012-00747-4.

Schmutz, E., 2008. Rational points on the unit sphere. Cent. Eur. J. Math. 6 (3), 482–487. ISSN 1895-1074. https://doi.org/10.2478/s11533-008-0038-4.

Trudinger, N., Wang, X.-J., 2005. The affine Plateau problem. J. Am. Math. Soc. 18 (2), 253–289. ISSN 0894-0347. https://doi.org/10.1090/S0894-0347-05-00475-3.

Trudinger, N., Wang, X.-J., 2008. Boundary regularity for the Monge-Ampère and affine maximal surface equations. Ann. Math. 167 (3), 993–1028. ISSN 0003-486X. https://doi.org/10.4007/annals.2008.167.993.

Wang, X.-J., 1995. Some counterexamples to the regularity of Monge-Ampére equations. Proc. Amer. Math. Soc. 123 (3), 841–845.

Wang, X.-J., 1996. On the design of a reflector antenna. Inverse Problems 12 (3), 351–375. ISSN 0266-5611. https://doi.org/10.1088/0266-5611/12/3/013.

Zienkiewicz, O., Taylor, R., 2000. The Finite Element Method, fifth ed. Vol. 1. Butterworth-Heinemann, Oxford. ISBN: 0-7506-5049-4, p. xvi+689.