SIAM J. SCI. COMPUT.
Vol. 45, No. 3, pp. A1097–A1124

A CONVERGENT QUADRATURE-BASED METHOD FOR THE MONGE–AMPÈRE EQUATION*

JAKE BRUSCA† AND BRITTANY FROESE HAMFELDT†

Abstract. We introduce an integral representation of the Monge–Ampère equation, which leads to a new finite difference method based upon numerical quadrature. The resulting scheme is monotone and fits immediately into existing convergence proofs for the Monge–Ampère equation with either Dirichlet or optimal transport boundary conditions. The use of higher-order quadrature schemes allows for substantial reduction in the component of the error that depends on the angular resolution of the finite difference stencil. This, in turn, allows for significant improvements in both stencil width and formal truncation error. The resulting schemes can achieve a formal accuracy that is arbitrarily close to $\mathcal{O}(h^2)$, which is the optimal consistency order for monotone approximations of second-order operators. We present three different implementations of this method. The first two exploit the spectral accuracy of the trapezoid rule on uniform angular discretizations to allow for computation on a nearest-neighbors finite difference stencil over a large range of grid refinements. The third uses higher-order quadrature to produce superlinear convergence while simultaneously utilizing narrower stencils than other monotone methods. Computational results are presented in two dimensions for problems of various regularity.

Key words. Monge–Ampère equation, finite difference methods, monotone, superlinear convergence

MSC codes. 35J15, 35J60, 35J70, 35J96, 65N06, 65N12

DOI. 10.1137/22M1494658

1. Introduction.

In this article we introduce an integral representation of the Monge–Ampère equation

$$\begin{cases} -\det(D^2u(x)) + f(x) = 0, \quad x \in \Omega, \\ u \text{ is convex}, \end{cases} \tag{1.1}$$

where $\Omega \subset \mathbb{R}^n$ is convex and the right-hand side f is nonnegative. This allows us to produce new monotone approximation schemes via quadrature. Because these schemes are monotone, they fit within several recently developed numerical convergence frameworks [3, 5, 16, 19, 20, 29]. Moreover, these new schemes offer significant advantages over existing monotone methods in terms of both accuracy and efficiency.

Recent years have seen a growing interest in Monge–Ampère type equations in the context of a diverse range of problems including design of optical systems [33], geophysics [12], mesh generation [7], medical image processing [18], meteorology [10], and data science [31]. This has encouraged the design of many new methods for the Monge–Ampère equation including [4, 6, 11, 13, 32].

The development of numerical methods that are guaranteed to converge to the correct solution, particularly in the absence of classical solutions, has proven to be more challenging. An early method [30] used a geometric interpretation of weak solutions to design a convergent, but computationally expensive, method for the

* Submitted to the journal's Methods and Algorithms for Scientific Computing section May 5, 2022; accepted for publication (in revised form) November 29, 2022; published electronically May 16, 2023.
https://doi.org/10.1137/22M1494658
Funding: The authors were partially supported by NSF DMS-1751996.
† Department of Mathematical Sciences, New Jersey Institute of Technology, University Heights, Newark, NJ 07102 USA (jb327@njit.edu, bdfroese@njit.edu).

two-dimensional Monge–Ampère equation. Recently, convergence frameworks have been established for the Monge–Ampère equation with either Dirichlet boundary conditions [16, 19, 27, 29],

$$u(x) = g(x), \quad x \in \partial\Omega \tag{1.2}$$

or the second type boundary condition arising in optimal transport [3, 5, 20],

$$\nabla u(\Omega) \subset \tilde{\Omega}. \tag{1.3}$$

These convergence proofs can be viewed as extensions of the powerful Barles and Souganidis convergence framework [1], which is valid for weak (viscosity) solutions of fully nonlinear partial differential equations. Critically, they are valid only for approximation schemes that are monotone. Construction of monotone schemes for degenerate elliptic PDE operators is not trivial: in fact, given any fixed finite difference stencil, it is possible to find linear elliptic operators for which no consistent, monotone approximation is possible on the given stencil [24, 26]. Circumventing this challenge requires the use of finite difference stencils that are allowed to grow wider as the grid is refined. Monotone schemes are inherently limited in their accuracy: a monotone approximation of a second-order operator can achieve at most second-order ($\mathcal{O}(h^2)$) truncation error [28, Theorem 4].

Several monotone finite difference schemes are now available for the Monge–Ampère equation [2, 5, 15, 16, 25, 29]. Because of the wide-stencil nature of these methods, the methods are computationally expensive and typically have low (sublinear) accuracy. There are limited techniques available that are capable of achieving the optimal $\mathcal{O}(h^2)$ truncation error [2, 5], but these schemes are valid for the Monge–Ampère equation only in two dimensions and with problem data that guarantees uniform ellipticity of the PDE. These challenges are magnified in three dimensions, where even evaluating the finite difference approximations (without attempting to solve the resulting nonlinear system) can be prohibitively expensive [21].

In this article, we propose to express the Monge–Ampère operator in terms of a Gaussian integral. This allows us to utilize higher-order quadrature schemes in order to simultaneously achieve improved consistency error (of $\mathcal{O}(h^{2-\epsilon})$ for any $\epsilon > 0$) and more compact wide finite difference stencils. The schemes are nevertheless monotone and fit neatly within the existing proofs of convergence of numerical methods to the weak (viscosity) solution of the Monge–Ampère equation. We describe three different implementations of this approach in two dimensions and validate the performance using a range of standard benchmark problems for the Dirichlet problem. This new formulation of the Monge–Ampère equation holds particular promise for the development of computationally practical methods in three dimensions, as it provides a dimension-reduction as compared with a typical variational formulation of the three-dimensional Monge–Ampère equation. It also extends naturally to more general Monge–Ampère type equations in optimal transport, including equations that are posed on the sphere [23].

2. Background.

2.1. Elliptic equations.

The Monge–Ampère equation is an example of a degenerate elliptic partial differential equation, which takes the general form

$$F(x, u(x), \nabla u(x), D^2 u(x)) = 0, \quad x \in \bar{\Omega}. \tag{2.1}$$

DEFINITION 2.1 (degenerate elliptic). *Let $\Omega \subset \mathbb{R}^n$ and denote by $\mathcal{S}^n$ the set of symmetric $n \times n$ matrices. The operator $F : \overline{\Omega} \times \mathbb{R} \times \mathbb{R}^n \times \mathcal{S}^n \to \mathbb{R}$ is said to be degenerate elliptic if*

$$F(x, u, p, X) \leq F(x, v, p, Y)$$

whenever $u \leq v$ and $X \succeq Y$.

We note that the operator is defined on the closure of Ω and takes on the value of the relevant boundary conditions at $\partial\Omega$. For the Dirichlet problem, which is the setting implemented in this article, the PDE operator at the boundary is defined as

$$F(x, u(x), \nabla u(x), D^2u(x)) = u(x) - g(x), \quad x \in \partial\Omega. \tag{2.2}$$

The Monge–Ampère equation (1.1) does not immediately satisfy this definition of an elliptic equation; in fact, it holds only on the restricted class of convex functions. Going hand in hand with this difficulty is the fact that the solution of the Monge–Ampère equation is not expected to be unique; the additional constraint that u is convex is needed in order to select a unique solution. A common remedy to these challenges is to define a globally elliptic extension of the Monge–Ampère equation that automatically enforces solution convexity [19]. This can be accomplished by considering the convexified Monge–Ampère operator

$$F(x, u(x), \nabla u(x), D^2u(x)) = -\det{}^+(D^2u(x)) + f(x), \quad x \in \Omega. \tag{2.3}$$

Here the modified determinant $\det^+$ should agree with the usual determinant when operating on the Hessian of a convex function and should return a negative value otherwise. The particular choice utilized in this article is

$$\det{}^+(M) = \begin{cases} \det(M), & M \succeq 0, \\ \lambda_1(M) & \text{otherwise}, \end{cases} \tag{2.4}$$

where $\lambda_1(M) \leq \cdots \leq \lambda_n(M)$ are the eigenvalues of the symmetric matrix M.

In general, degenerate elliptic equations need not have classical solutions, and some notion of weak solution is required. The Aleksandrov solution provides a geometric interpretation in terms of the subgradient measure, which allows for very general right-hand sides, including measures that do not have an associated density [17]. Though slightly less general, the viscosity solution has proved to be particularly useful for this class of equations [9] and forms the foundation for most of the recently developed numerical convergence proofs for the Monge–Ampère equation. The idea of the viscosity solution is to use a maximum principle argument to pass derivatives onto smooth test functions that lie above or below the semicontinuous envelopes of the candidate weak solution.

DEFINITION 2.2 (semicontinuous envelopes). *Let $u : \Omega \to \mathbb{R}$ be a bounded function. Then for $x \in \overline{\Omega}$, the upper and lower semicontinuous envelopes are defined, respectively, as*

$$u^*(x) = \limsup_{y \to x} u(y), \quad u_*(x) = \liminf_{y \to x} u(y).$$

DEFINITION 2.3 (viscosity subsolutions (supersolutions)). *A bounded upper (lower) semicontinuous function u is a viscosity subsolution (supersolution) of* (2.1) *if for every $\phi \in C^2(\overline{\Omega})$, whenever $u - \phi$ has a local maximum (minimum) at $x \in \overline{\Omega}$, then*

$$F_*^{(*)}(x, u(x), \nabla\phi(x), D^2\phi(x)) \leq (\geq) 0.$$

DEFINITION 2.4 (viscosity solution). *A bounded function $u : \overline{\Omega} \to \mathbb{R}$ is a viscosity solution of* (2.1) *if $u^*(x)$ is a viscosity subsolution and $u_*(x)$ is a viscosity supersolution.*

An important characteristic of many elliptic operators, which immediately yields solution uniqueness, is the comparison principle.

DEFINITION 2.5 (comparison principle). *The operator* (2.1) *satisfies a strong comparison principle if whenever u is an upper semicontinuous subsolution and v a lower semicontinuous supersolution, then $u \leq v$ on $\overline{\Omega}$.*

We remark that the Dirichlet problem for the Monge–Ampère equation (1.1), (1.2) does satisfy a comparison principle under reasonable assumptions on the data. However, this is no longer true when the right-hand side depends on the solution gradient or when the second type boundary condition (1.3) is considered [19, 20].

2.2. Convergence framework. A fruitful technique for numerically solving fully nonlinear elliptic equations involves finite difference schemes of the form

$$F^h(x, u(x), u(x) - u(\cdot)) = 0 \tag{2.5}$$

defined on a finite set of discretization points $\mathcal{G} \subset \Omega$. Many key results on the convergence of finite difference methods to the viscosity solution of a degenerate elliptic PDE are based upon a set of criteria developed by Barles and Souganidis [1].

DEFINITION 2.6 (consistency). *The scheme* (2.5) *is consistent with* (2.1) *if, for any test function $\phi \in C^{2,1}(\overline{\Omega})$ and $x \in \overline{\Omega}$, we have*

$$\limsup_{h \to 0^+, y \to x, \xi \to 0} F^h(y, \phi(y) + \xi, \phi(y) - \phi(\cdot)) \leq F^*(x, \phi(x), \nabla\phi(x), D^2\phi(x)), \tag{2.6}$$

$$\liminf_{h \to 0^+, y \to x, \xi \to 0} F^h(y, \phi(y) + \xi, \phi(y) - \phi(\cdot)) \geq F_*(x, \phi(x), \nabla\phi(x), D^2\phi(x)). \tag{2.7}$$

To a consistent scheme we can also assign a local truncation error.

DEFINITION 2.7 (truncation error). *The truncation error of a scheme* (2.5) *on a set of admissible functions Φ is a function $\tau(h)$ such that for any $\phi \in \Phi$ there exists a constant $C \geq 0$ such that*

$$\left| F^h(x, \phi(x), \phi(x) - \phi(\cdot)) - F(x, \phi(x), \nabla\phi(x), D^2\phi(x)) \right| \leq C\tau(h)$$

for every $x \in \mathcal{G}$ and sufficiently small $h > 0$.

DEFINITION 2.8 (monotonicity). *The scheme* (2.5) *is monotone if F^h is a nondecreasing function of its last two arguments.*

DEFINITION 2.9 (stability). *The scheme* (2.5) *is stable if there exists some $M > 0$, independent of h, such that every solution u^h satisfies $||u^h||_\infty < M$.*

These simple concepts lead immediately to convergence of finite difference methods, provided the underlying PDE satisfies a strong comparison principle.

THEOREM 2.10 (convergence [1]). *Let u be the unique viscosity solution of the PDE* (2.1)*, where F is a degenerate elliptic operator with a strong comparison principle. Let u^h be any solution of* (2.5) *where F^h is a consistent, monotone, stable approximation scheme. Then u^h converges uniformly to u as $h \to 0$.*

This result does apply to the Monge–Ampère equation (1.1) with Dirichlet boundary conditions (1.2) under mild assumptions on the data. However, many other Monge–Ampère equations of interest do not possess the strong comparison principle required by the theorem. In recent years, the convergence proof has been adapted to include discontinuous solutions of the nonclassical Dirichlet problem [19] and the second boundary value problem [3, 5, 20].

2.3. Wide stencil methods. Several monotone finite difference approximations have been proposed for the Monge–Ampère operator [2, 5, 8, 16, 27]. These hinge upon different reformulations of the Monge–Ampère operator, which typically take a variational form

$$\det(D^2 u) = \min_{(\nu_1,\dots,\nu_k)\in\mathcal{A}} G\left(u_{\nu_1\nu_1},\dots,u_{\nu_k\nu_k}\right). \tag{2.8}$$

Here $u_{\nu\nu}$ denotes the second directional derivative in the direction $\nu \in \mathbb{R}^n$, $\mathcal{A}$ is some admissible set, and G is a nondecreasing function. Generating a monotone approximation then requires (1) an appropriate discretization of the relevant second directional derivatives and (2) an appropriate discretization of the admissible set.

On a structured grid, where aligned points x, $x + h^+\nu$, and $x - h^-\nu$ are available for some $h^-, h^+ > 0$, a simple (negative) monotone approximation is

$$\begin{aligned}\mathcal{D}_{\nu\nu}u(x) &\equiv 2\frac{h^- u(x+h^+\nu) + h^+ u(x-h^-\nu) - (h^+ + h^-)u(x)}{h^+h^-(h^+ + h^-)} \\ &= u_{\nu\nu}(x) + \mathcal{O}(h^+ - h^-) + \mathcal{O}\big((h^+)^2 + (h^-)^2\big).\end{aligned} \tag{2.9}$$

These approximations are typically allowed to have a wide-stencil flavor, with the spacing h^+, h^- being potentially larger than the characteristic spacing h of grid points. See Figure 1. We also remark that in the special case of equispaced neighboring points ($h^+ = h^-$), such as on a uniform Cartesian grid, this reduces to the usual centered difference approximation with second-order truncation error. Monotone approximations are also possible on unstructured grids, though they are typically less accurate [15].

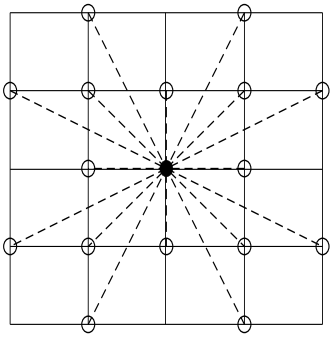

FIG. 1. *Wide finite difference stencils.*

The width of stencils required by the approximations of (2.8) is determined by the discretization of the admissible set $\mathcal{A}$. Optimal discretization of this set is itself a nontrivial problem and evaluating (2.8) may involve minimization over a prohibitively large set of candidate directions, particularly in three dimensions [21]. A typical scaling for the maximal stencil width that optimizes truncation error is at least $\mathcal{O}(\sqrt{h})$ [15], though in some cases it is not clear what the optimal choice is.

3. Integral formulation. In this section, we present an integral representation of $\det(D^2u(x))$ which can be used to create a monotone discretizaton through the use of quadrature.

To motivate this, we recall a well-known result about the integrals of multivariate Gaussians:

$$\det(M) = \pi^n \left(\int_{\mathbb{R}^n} e^{-v^T M v} dV \right)^{-2}, \tag{3.1}$$

where M is a symmetric positive-definite $n \times n$ matrix.

This provides an alternate characterization of the Monge–Ampère operator if we let $M = D^2u(x)$ be the Hessian of the potential function u. Provided u is strictly convex, its Hessian is positive definite and we can write

$$\det(D^2u(x)) = \pi^n \left(\int_{\mathbb{R}^n} e^{-v^T D^2u(x) v} dV \right)^{-2}. \tag{3.2}$$

To express this in a form that is easily discretized, we convert to spherical coordinates. Let $r = |v|$ and $\hat{v} = v/r$ and denote by $u_{\hat{v}\hat{v}} = \hat{v}^T D^2u(x)\hat{v}$ the second directional derivative of u in the direction of $\hat{v}$. Then the Monge–Ampère operator in $\mathbb{R}^n$ can be expressed as

$$\det(D^2u(x)) = \pi^n \left(\int_{\hat{v}\in\mathbb{S}^{n-1}} \int_0^\infty r^{n-1} e^{-r^2 u_{\hat{v}\hat{v}}} dr d\hat{V} \right)^{-2}. \tag{3.3}$$

Integrating out the radius r, we obtain

$$\det(D^2u(x)) = \frac{4\pi^n}{\Gamma(n/2)^2} \left(\int_{\mathbb{S}^{n-1}} (u_{\hat{v}\hat{v}})^{-n/2} d\hat{V} \right)^{-2}. \tag{3.4}$$

For simplicity, the results in the remainder of this paper are presented in two dimensions. However, they are certainly generalizable to higher dimensions. We introduce the notation

$$u_{\theta\theta} = \frac{\partial^2 u}{\partial \nu^2}, \quad \nu = (\cos\theta, \sin\theta)$$

and note that

$$u_{\theta\theta} = u_{\theta+\pi,\theta+\pi}.$$

Then we easily obtain the two-dimensional version of (3.4) in terms of polar coordinates.

THEOREM 3.1 (integral representation). *Let $\Omega \subset \mathbb{R}^2$ be convex and $u \in C^2(\Omega)$ be strictly convex. Then for every $x \in \Omega$,*

$$\det(D^2u(x)) = \left(\frac{1}{\pi} \int_0^\pi \frac{d\theta}{u_{\theta\theta}(x)} \right)^{-2}. \tag{3.5}$$

The characterization in Theorem 3.1 holds only when $D^2u(x) \succ 0$. However, we are also interested in degenerate cases where $\det(D^2u(x)) = 0$. In these cases, we know that $D^2u(x)$ has at least eigenvalue equal to zero, and the integrand in (3.5) becomes singular.

We introduce the following relaxation to approximate the integral in these cases:

$$\det{}_{\varepsilon_1}(D^2u(x)) = \left(\frac{1}{\pi}\int_0^\pi \frac{d\theta}{\max(u_{\theta\theta}(x), \varepsilon_1)}\right)^{-2}. \tag{3.6}$$

This, in turn, is used to construct a relaxed version of the convexified Monge–Ampère operator:

$$\det{}^+_{\varepsilon_1,\epsilon_2}(D^2u(x)) = \det{}_{\varepsilon_1}(D^2u(x)) + \min_{\theta\in[0,\pi)}\{\min(u_{\theta\theta}, \varepsilon_2)\}. \tag{3.7}$$

Here we have represented $\lambda_1(D^2u)$ as $\min\{u_{\theta\theta}\}$, which is equivalent via the minimax principle.

4. Quadrature scheme. In this section, we describe a very general framework for utilizing the integral formulation (3.3) to produce a consistent, monotone approximation of the two-dimensional Monge–Ampère equation. In section 5, we will describe two particular implementations.

4.1. Approximation scheme. We introduce the following notation.

Definition 4.1 (notation).

(N1) $\Omega \subset \mathbb{R}^2$ *is a bounded, open, convex domain with Lipschitz boundary* $\partial\Omega$.
(N2) $\mathcal{G} \subset \bar{\Omega}$ *is a finite set of discretization points* x_i, $i = 1, \dots, N$.
(N3) $h = \sup_{x\in\Omega} \min_{y\in\mathcal{G}} |x - y|$ *is the spatial resolution of the grid. In particular, every ball of radius* h *contained in* $\bar{\Omega}$ *contains at least one discretization point* x_i.
(N4) $r \geq h$ *is a stencil width associated to the grid.*
(N5) $0 \leq \theta_0 < \cdots < \theta_M < \pi$ *is a finite set of angles discretizing* $[0, \pi)$.
(N6) $d\theta_i = \theta_{i+1} - \theta_i$ *is the local angular resolution of the discretization, where we define* $d\theta_M = \theta_0 + \pi - \theta_M$.
(N7) $d\theta = \max_{i=0,\dots,M}\{d\theta_i\}$ *is the angular resolution of the discretization.*
(N8) $Q = \dfrac{d\theta}{\min_{i=0,\dots,M} d\theta_i}$ *is the quasi-uniformity constant of the angular discretization.*
(N9) $w_0, \dots, w_M$ *is a collection of nonnegative quadrature weights summing to* π *and satisfying*

$$w_k \geq c\,d\theta$$

for some constant $c > 0$ *that depends only on the quasi-uniformity constant.*
(N10) $\epsilon_1 > 0$ *and* $\epsilon_2 \geq 0$ *are regularization parameters associated with the grid.*
(N11) $\mathcal{N}(x) \subset \{1, \dots, N\}$ *is the set of neighboring indices for* $x \in \mathcal{G} \cap \Omega$ *such that for every* $j \in \mathcal{N}(x)$ *we have* $0 < |x_j - x| \leq r$.
(N12) $\mathcal{D}_{\theta\theta}u(x)$ *described in* (2.9) *has the form*

$$\mathcal{D}_{\theta\theta}u(x) = \sum_{j\in\mathcal{N}(x)} a_j(\theta)\,(u(x_j) - u(x))$$

for every $x \in \mathcal{G} \cap \Omega$, *where all* $a_i \geq 0$.

(N13) $\tau_\theta(r)$ *is the truncation error of the finite difference scheme* $\mathcal{D}_{\theta\theta}u$ *for approximating the second directional derivative* $u_{\theta\theta}$ *on the admissible set* $\Phi = C^{2,1}(\Omega)$.

(N14) $\tau_{FD}(r) = \max_{i=1,\dots,M} \tau_{\theta_i}(r)$ *is the maximal truncation error of the finite difference approximations.*

(N15) $\tau_Q(d\theta)$ *is the truncation error of the quadrature scheme*

$$\sum_{i=0}^{M} w_i f(\theta_i)$$

for approximating the integral $\int_0^\pi f(\theta)\,d\theta$ *on the admissible set* $\Phi = \{f \in C^\infty([0,\pi]) \mid f \text{ is periodic}\}$.

Then we propose the following scheme for approximating the convexified Monge–Ampère operator at interior points $x \in \mathcal{G} \cap \Omega$:

(4.1)
$$G^h(x, u(x), u(x) - u(\cdot)) = -\left(\frac{1}{\pi}\sum_{i=0}^{M} \frac{w_i}{\max\{\mathcal{D}_{\theta_i\theta_i}u(x), \epsilon_1\}}\right)^{-2} - \min_{i=0,\dots,M}\{\mathcal{D}_{\theta_i\theta_i}u(x), \epsilon_2\}.$$

4.2. Convergence. We now provide conditions under which the scheme (4.1) is consistent and monotone. As an immediate consequence, it fits directly into the convergence proofs developed in [3, 5, 16, 19, 20, 29].

THEOREM 4.2 (monotonicity). *The approximation scheme* (4.1) *is monotone.*

Proof. We note that the operator that appears in (4.1) can be written in the form

$$G^h(x,u,z) = -\left(\frac{1}{\pi}\sum_{i=0}^{M} \frac{w_i}{\max\{-\sum_{j\in\mathcal{N}(x)} a_{ij}z_j,\, \epsilon_1\}}\right)^{-2} - \min_{i=0,\dots,M}\left\{-\sum_{j\in\mathcal{N}(x)} a_{ij}z_j,\, \epsilon_2\right\},$$

where $z_j = u(x) - u(x_j)$ and the a_{ij} are nonnegative by the (negative) monotonicity of the approximations $\mathcal{D}_{\theta_i\theta_i}u(x)$.

Let $\delta \in \mathbb{R}^N$ have nonnegative components. We notice that

$$-\sum_{j\in\mathcal{N}(x)} a_{ij}(z_j + \delta_j) \le -\sum_{j\in\mathcal{N}(x)} a_{ij}z_j.$$

Since the max and min operators preserve monotonicity and the weights w_i are nonnegative, we can immediately conclude that

$$G^h(x,u,z+\delta) \ge G^h(x,u,z).$$

Since G^h has no dependence on its second argument, this completes the proof of monotonicity. ∎

THEOREM 4.3 (consistency). *Consider discretizations* $\mathcal{G}^h$ *of* $\bar{\Omega}$ *such that the corresponding parameters*

$$r, d\theta, \frac{\epsilon_1}{d\theta}, \frac{\tau_{FD}(r)}{d\theta}, \tau_Q(d\theta), \epsilon_2 \to 0$$

as $h \to 0$ *and the corresponding quasi-uniformity constants* Q *are bounded uniformly. Then the approximation scheme* (4.1) *is consistent with the convexified Monge–Ampère operator* (2.3).

We will break this result into three separate cases (Lemmas 4.4–4.6), depending on the sign of $\lambda_1(D^2u)$, the smallest eigenvalue of the Hessian. We note that the scheme $G^h(x,u,z)$ appearing in (4.1) has no dependence on the first two arguments, which allows us to simplify slightly the verification of consistency.

LEMMA 4.4 (consistency with positive eigenvalues). *Under the assumptions of Theorem* 4.3*, let* $u \in C^{2,1}(\Omega)$ *and consider* $x \in \Omega$ *such that* $\lambda_1(D^2u(x)) > 0$*. Then the scheme* (4.1) *satisfies*

$$\lim_{y\in\mathcal{G}\to x, h\to 0} G^h(y,u(y),u(y)-u(\cdot)) = -\det{}^+(D^2u(x)).$$

Proof. Since the smallest eigenvalue $\lambda_1(D^2u(x))$ is strictly positive and $u \in C^{2,1}$, we are assured that

$$\lambda_1(D^2u(y)) > \frac{1}{2}\lambda_1(D^2u(x)) > \epsilon_k + \mathcal{O}(\tau_{FD}(r)), \quad k \in \{1,2\},$$

for all y sufficiently close to x and sufficiently small $\epsilon_1, \epsilon_2, r$. Then using the consistency error in the components of this scheme, we can compute

$$\begin{aligned}
&G^h(y,u(y),u(y)-u(\cdot))\\
&\quad= -\left(\frac{1}{\pi}\sum_{i=0}^{M}\frac{w_i}{\max\{\mathcal{D}_{\theta_i\theta_i}u(y),\epsilon_1\}}\right)^{-2} - \min_{i=0,\dots,M}\{\mathcal{D}_{\theta_i\theta_i}u(y),\epsilon_2\}\\
&\quad= -\left(\frac{1}{\pi}\sum_{i=0}^{M}\frac{w_i}{\max\{u_{\theta_i\theta_i}(y)+\mathcal{O}(\tau_{FD}(r)),\epsilon_1\}}\right)^{-2} - \min_{i=0,\dots,M}\{u_{\theta_i\theta_i}(y)+\mathcal{O}(\tau_{FD}(r)),\epsilon_2\}\\
&\quad= -\left(\frac{1}{\pi}\sum_{i=0}^{M}\frac{w_i}{u_{\theta_i\theta_i}(y)+\mathcal{O}(\tau_{FD}(r))}\right)^{-2} - \epsilon_2\\
&\quad= -\left(\frac{1}{\pi}\sum_{i=0}^{M}\frac{w_i}{u_{\theta_i\theta_i}(y)} + \mathcal{O}(\tau_{FD}(r))\sum_{i=0}^{M}\frac{w_i}{u_{\theta_i\theta_i}(y)^2}\right)^{-2} - \epsilon_2.
\end{aligned}$$

Since $u_{\theta_i\theta_i}(y)$ is continuous in θ and bounded away from zero, the two sums in the last line can both be interpreted as consistent quadrature schemes. Thus we can further estimate

$$G^h(y,u(y),u(y)-u(\cdot)) = -\left(\frac{1}{\pi}\int_0^\pi \frac{1}{u_{\theta\theta}(y)}\,d\theta + \mathcal{O}(\tau_Q(d\theta)+\tau_{FD}(r))\right)^{-2} - \epsilon_2.$$

Recalling now the integral formulation of the Monge–Ampère operator (3.5), we conclude that

$$G^h(y,u(y),u(y)-u(\cdot)) = -\det(D^2u(y)) + \mathcal{O}(\tau_Q(d\theta)+\tau_{FD}(r)+\epsilon_2).$$

Since $u \in C^{2,1}$ and all eigenvalues of the Hessian are strictly positive, we conclude that

$$\lim_{y\in\mathcal{G}\to x, h\to 0} G^h(y,u(y),u(y)-u(\cdot)) = -\det{}^+(D^2u(x)).$$

∎

LEMMA 4.5 (consistency with a negative eigenvalue). *Under the assumptions of Theorem* 4.3*, let* $u \in C^{2,1}(\Omega)$ *and consider* $x \in \Omega$ *such that* $\lambda_1(D^2u(x)) < 0$*. Then the scheme* (4.1) *satisfies*

$$\lim_{y \in \mathcal{G} \to x, h \to 0} G^h(y, u(y), u(y) - u(\cdot)) = -\det{}^+(D^2u(x)).$$

Proof. Suppose without loss of generality that the coordinates are chosen so that the eigenvector corresponding to the smallest eigenvalue $\lambda_1(D^2u(y))$ is $v_1 = (1, 0)$. Then the second directional derivative of u in the direction θ can be expressed as

$$u_{\theta\theta}(y) = \lambda_1(y)\cos^2\theta + \lambda_2(y)\sin^2\theta.$$

Let us consider in particular the first angle $\theta_0 \le d\theta$ in the angular discretization. Since $u \in C^{2,1}$, the second directional derivative of u in this direction satisfies

$$\begin{aligned} u_{\theta_0\theta_0}(y) &= \lambda_1(D^2u(y)) + \mathcal{O}(d\theta^2) \\ &= \lambda_1(D^2u(x)) + \mathcal{O}(d\theta^2 + |x - y|). \end{aligned}$$

Since $\lambda_1(D^2u(x))$ is strictly negative, it is certainly the case that for y sufficiently close to x and small enough $r, d\theta$,

$$\begin{aligned} \mathcal{D}_{\theta_0\theta_0}u(y) &= u_{\theta_0\theta_0}(y) + \mathcal{O}(\tau_{FD}(r)) \\ &= \lambda_1(D^2u(x)) + \mathcal{O}(\tau_{FD}(r) + d\theta^2 + |x - y|) \\ &< 0 \\ &< \epsilon_k, \qquad\qquad k \in \{1, 2\}. \end{aligned}$$

Now we perform a crude estimate on the sum in (4.1) by considering only a single term:

$$\begin{aligned} 0 &\le \left(\frac{1}{\pi}\sum_{i=0}^{M}\frac{w_i}{\max\{\mathcal{D}_{\theta_i\theta_i}u(y), \epsilon_1\}}\right)^{-2} \\ &\le \left(\frac{1}{\pi}\frac{w_0}{\max\{\mathcal{D}_{\theta_0\theta_0}u(y), \epsilon_1\}}\right)^{-2} \\ &= \frac{\pi^2{\epsilon_1}^2}{w_0^2} \\ &\le \frac{\pi^2{\epsilon_1}^2}{c^2 d\theta^2}. \end{aligned}$$

Using the same estimates on the discrete second directional derivatives, we can also estimate the term

$$\begin{aligned} \min_{i=0,\dots,M}\{\mathcal{D}_{\theta_i\theta_i}, \epsilon_2\} &= \min\{\lambda_1(D^2u(x)) + \mathcal{O}(\tau_{FD}(r) + d\theta^2 + |x - y|), \epsilon_2\} \\ &= \lambda_1(D^2u(x)) + \mathcal{O}(\tau_{FD}(r) + d\theta^2 + |x - y|). \end{aligned}$$

By combining these estimates and recalling that $\epsilon_1/d\theta \to 0$, we conclude that

$$\begin{aligned} \lim_{y \in \mathcal{G}, h \to 0} G^h(y, u(y), u(y) - u(\cdot)) &= -\lambda_1(D^2u(x)) \\ &= -\det{}^+(D^2u(x)). \end{aligned}$$

∎

LEMMA 4.6 (consistency with a vanishing eigenvalue). *Under the assumptions of Theorem* 4.3*, let* $u \in C^{2,1}(\Omega)$ *and consider* $x \in \Omega$ *such that* $\lambda_1(D^2u(x)) = 0$*. Then the scheme* (4.1) *satisfies*

$$\lim_{y \in \mathcal{G} \to x, h \to 0} G^h(y, u(y), u(y) - u(\cdot)) = -\det{}^+(D^2u(x)).$$

Proof. By the regularity of u, we know that

$$\lambda_1(D^2u(y)) = \mathcal{O}(|x - y|).$$

Suppose without loss of generality that the coordinates are chosen so that the eigenvector corresponding to the smallest eigenvalue $\lambda_1(D^2u(y))$ is $v_1 = (1, 0)$. Then the second directional derivative of u in the direction θ can be expressed as

$$u_{\theta\theta}(y) = \lambda_1(y)\cos^2\theta + \lambda_2(y)\sin^2\theta.$$

Now we are going to estimate the sum in (4.1) by considering only the angles θ that are close to zero, which corresponds to the direction of the eigenvector v_1. To this end, we define

$$s = \max\{d\theta, \sqrt{|x - y|}\}$$

and let $K = \mathcal{O}(s/d\theta) \geq 1$ be the number of nodes $\theta_0, \ldots, \theta_{K-1}$ in the interval $[0, s]$.

We notice that for any $i = 0, \ldots, K - 1$ we have

$$\begin{aligned}
\max\{\mathcal{D}_{\theta_i\theta_i}u(y), \epsilon_1\} &\leq u_{\theta_i\theta_i}(y) + \tau_{FD}(r) + \epsilon_1 \\
&= \lambda_1(y)\cos^2\theta_i + \lambda_2(y)\sin^2\theta_i + \tau_{FD}(r) + \epsilon_1 \\
&= \mathcal{O}(|x - y| + s^2 + \tau_{FD}(r) + \epsilon_1).
\end{aligned}$$

Then using the lower bound $w_i > c\,d\theta$ allows us to obtain the following bounds on the sum:

$$\begin{aligned}
\sum_{i=0}^{M} \frac{w_i}{\max\{\mathcal{D}_{\theta_i\theta_i}u(y), \epsilon_1\}} &\geq \sum_{i=0}^{K-1} \frac{w_i}{\max\{\mathcal{D}_{\theta_i\theta_i}u(y), \epsilon_1\}} \\
&\geq K \frac{c\,d\theta}{\mathcal{O}(|x - y| + s^2 + \tau_{FD}(r) + \epsilon_1)} \\
&= \frac{s}{d\theta} \frac{c\,d\theta}{\mathcal{O}(|x - y| + s^2 + \tau_{FD}(r) + \epsilon_1)} \\
&= \frac{cs}{\mathcal{O}(|x - y| + s^2 + \tau_{FD}(r) + \epsilon_1)}.
\end{aligned}$$

This allows us to obtain bounds on the following value appearing in the scheme:

$$\left(\frac{1}{\pi}\sum_{i=0}^{M} \frac{w_i}{\max\{\mathcal{D}_{\theta_i\theta_i}u(y), \epsilon_1\}}\right)^{-2} \leq \mathcal{O}\left(\frac{|x - y|^2}{s^2} + \frac{s^4}{s^2} + \frac{\tau_{FD}(r)^2}{s^2} + \frac{{\epsilon_1}^2}{s^2}\right).$$

Recalling from the definition of s that $s \geq d\theta$ and $s \geq \sqrt{|x-y|}$ allows us to simplify this as follows:

$$\begin{aligned} 0 &\leq \left(\frac{1}{\pi} \sum_{i=0}^{M} \frac{w_i}{\max\{\mathcal{D}_{\theta_i\theta_i} u(y), \epsilon_1\}} \right)^{-2} \\ &\leq \mathcal{O}\left(\frac{|x-y|^2}{|x-y|} + \max\{d\theta^2, |x-y|\} + \frac{\tau_{FD}(r)^2}{d\theta^2} + \frac{{\epsilon_1}^2}{d\theta^2} \right). \end{aligned}$$

Thus under the conditions of Theorem 4.3, we find that

$$\lim_{y \in \mathcal{G} \to x, h \to 0} \left(\frac{1}{\pi} \sum_{i=0}^{M} \frac{w_i}{\max\{\mathcal{D}_{\theta_i\theta_i} u(y), \epsilon_1\}} \right)^{-2} = 0.$$

We also observe that

$$\begin{aligned} \epsilon_2 &\geq \min_{i=0,\dots,M} \{\mathcal{D}_{\theta_i\theta_i} u(y), \epsilon_2\} \\ &\geq \min_{i=0,\dots,M} \{\lambda_1(D^2u(y)) \cos^2\theta_i + \lambda_2(D^2u(y)) \sin^2\theta_i - \mathcal{O}(\tau_{FD}(r)), \epsilon_2\} \\ &\geq \min_{i=0,\dots,M} \{-\mathcal{O}(|x-y|) \cos^2\theta_i + \lambda_2(D^2u(y)) \sin^2\theta_i - \mathcal{O}(\tau_{FD}(r)), \epsilon_2\} \\ &\geq -\mathcal{O}(|x-y| + \tau_{FD}(r)), \end{aligned}$$

which implies that

$$\lim_{y \in \mathcal{G} \to x, h \to 0} \min_{i=0,\dots,M} \{\mathcal{D}_{\theta_i\theta_i} u(y), \epsilon_2\} = 0.$$

We conclude that

$$\lim_{y \in \mathcal{G} \to x, h \to 0} G^h(y, u(y), u(y) - u(\cdot)) = 0,$$

which coincides with the value of

$$\det^+(D^2u(x)) = \lambda_1(D^2u(x))\lambda_2(D^2u(x)) = 0. \qquad \square$$

4.3. Quadrature rules. In designing a scheme of the form (4.1), the choice of quadrature rule

$$\sum_{i=0}^{M} w_i f(\theta_i) \approx \int_0^{\pi} f(\theta)\, d\theta$$

is a key factor that will influence the overall cost and accuracy.

A simple choice is the trapezoid rule, which utilizes the weights

$$w_i = \begin{cases} \dfrac{\theta_0 + \pi - \theta_M}{2}, & i = 0, \\ \dfrac{\theta_{i+1} - \theta_{i-1}}{2}, & i = 1, \dots, M-1, \\ \dfrac{\pi - \theta_{M-1} + \theta_0}{2}, & i = M. \end{cases} \tag{4.2}$$

As required, the weights are all positive. As required by (N8) (Definition 4.1), they can also be bounded from below in terms of the quasi-uniformity constant via

$$w_i = \frac{d\theta_{i-1} + d\theta_i}{2} \geq \frac{d\theta}{Q}.$$

In general, the truncation error of the trapezoid rule is $\tau_Q(d\theta) = d\theta^2$. However, in the special case of a uniform angular discretization ($d\theta_i = d\theta$ for all $i = 0, \ldots, M$), the trapezoid rule is spectrally accurate. In this case, the truncation error satisfies $\tau_Q(d\theta) \leq d\theta^p$ for every $p > 0$ given a C^∞ integrand.

Higher-order quadrature is also possible on nonuniform angular discretizations. In fact, as we will demonstrate in section 5, this can be exploited in order to design approximation schemes that simultaneously improve the formal consistency error and reduce the required stencil width.

As an example, we consider Simpson's rule. Suppose that $M+1$, the number of angles in the angular discretization, is even. Then Simpson's rule takes the form

$$\begin{aligned}\int_0^\pi f(\theta)\,d\theta \approx \sum_{i=0}^{(M-1)/2} \frac{d\theta_{2i} + d\theta_{2i+1}}{6}\Bigg[&\left(2 - \frac{d\theta_{2i+1}}{d\theta_{2i}}\right) f(\theta_{2i}) \\ &+ \frac{(d\theta_{2i} + d\theta_{2i+1})^2}{d\theta_{2i} d\theta_{2i+1}} f(\theta_{2i+1}) + \left(2 - \frac{d\theta_{2i}}{d\theta_{2i+1}}\right) f(\theta_{2i+2})\Bigg],\end{aligned} \tag{4.3}$$

where we identify $\theta_{j+M+1} = \theta_j + \pi$ and $d\theta_j = d\theta_{j+M+1}$ because of the periodicity of f.

Rearranging, we find that the corresponding quadrature weights are

$$w_j = \begin{cases} \dfrac{(d\theta_{j-1} + d\theta_j)^3}{6 d\theta_{j-1} d\theta_j}, & j \text{ odd}, \\ \dfrac{d\theta_j + d\theta_{j+1}}{6}\left(2 - \dfrac{d\theta_{j+1}}{d\theta_j}\right) + \dfrac{d\theta_{j-2} + d\theta_{j-1}}{6}\left(2 - \dfrac{d\theta_{j-2}}{d\theta_{j-1}}\right), & j \text{ even}. \end{cases} \tag{4.4}$$

The truncation error associated with Simpson's rule is $\tau_Q(d\theta) = d\theta^4$.

However, unlike with the trapezoid rule, these quadrature weights are not automatically positive. Instead, positivity is guaranteed only if the quasi-uniformity constant of the angular discretization is not too large. In particular, we note that $Q < 2$ is sufficient to guarantee that

$$2 - \frac{d\theta_{j\pm1}}{d\theta_j} \geq 2 - \frac{d\theta}{d\theta_j} \geq 2 - Q > 0.$$

Under the same assumption on quasi-uniformity, we use the fact that

$$\frac{d\theta}{Q} \leq d\theta_j \leq d\theta$$

to verify that

$$w_j \geq \min\left\{\frac{(2d\theta/Q)^3}{6d\theta^2}, 2\frac{2d\theta/Q}{6}\left(2 - \frac{d\theta}{d\theta/Q}\right)\right\} = \min\left\{\frac{4}{3Q^3}, \frac{2}{3Q}(2-Q)\right\} d\theta,$$

as required by (N8) (Definition 4.1).

Similar results can be obtained using other higher-order quadrature schemes, which will place differing requirements on the quasi-uniformity constant Q in order to ensure positivity of the weights.

4.4. Truncation error. As an immediate consequence of the consistency proofs (in particular, Lemma 4.4), we obtain the formal truncation error of the scheme as points where the function is "locally" strictly convex. This will be used to inform and optimize the particular implementations of this method in section 5.

COROLLARY 4.7 (truncation error). *Under the assumptions of Theorem* 4.3*, let* $u \in C^{2,1}(\Omega)$ *and consider* $x \in \Omega$ *such that* $\lambda_1(D^2u(x)) > 0$*. Then there exists a constant* $C > 0$ *such that for all sufficiently small* $h > 0$*,*

$$\left|G^h(x, u(x), u(x) - u(\cdot)) + \det{}^+(D^2u(x))\right| \leq C\left(\tau_Q(d\theta) + \tau_{FD}(r) + \epsilon_2\right).$$

We are also interested in approximating functions that are convex, but not necessarily strictly convex. In this case, the integrand in (3.5) is singular and we cannot directly use the formal truncation error $\tau_Q(d\theta)$ of the quadrature rule. However, we can easily bound the resulting sums directly in the case where at least one eigenvalue $\lambda_1(D^2u(x))$ is known to vanish. We consider two separate cases: (1) the fully degenerate case ($\lambda_1(D^2u(x)) = \lambda_2(D^2u(x)) = 0$) and (2) the semidegenerate case ($\lambda_1(D^2u(x)) = 0 < \lambda_2(D^2u(x))$).

LEMMA 4.8 (truncation error (fully degenerate)). *Under the assumptions of Theorem* 4.3*, let* $u \in C^{2,1}(\Omega)$ *and consider* $x \in \Omega$ *such that* $\lambda_1(D^2u(x)) = \lambda_2(D^2u(x)) = 0$*. Then there exists a constant* $C > 0$ *such that for all sufficiently small* $h > 0$*,*

$$\left|G^h(x, u(x), u(x) - u(\cdot)) + \det{}^+(D^2u(x))\right| \leq C\left(\tau_{FD}(r)^2 + \epsilon_1\right).$$

LEMMA 4.9 (truncation error (semidegenerate)). *Under the assumptions of Theorem* 4.3*, let* $u \in C^{2,1}(\Omega)$ *and consider* $x \in \Omega$ *such that* $\lambda_1(D^2u(x)) = 0 < \lambda_2(D^2u(x))$*. Then there exists a constant* $C > 0$ *such that for all sufficiently small* $h > 0$*,*

$$\left|G^h(x, u(x), u(x) - u(\cdot)) + \det{}^+(D^2u(x))\right| \leq C\left(d\theta^2 + \frac{\tau_{FD}(r)^2}{d\theta^2} + \frac{{\epsilon_1}^2}{d\theta^2} + \epsilon_2\right).$$

Finally, we observe that with appropriate symmetry in the discretization, our quadrature-based schemes can sometimes result in even better formal consistency error than that predicted by Corollary 4.7. This observation motivates one of the implementations (on hexagonal grids) that will be introduced in section 5.

We consider the special case of applying the trapezoid rule using equally spaced angles ($d\theta_i = d\theta$ for all $i = 0, \ldots, M$), which is spectrally accurate as discussed previously. We suppose that we use grid-aligned differences, which may be centered or uncentered, to discretize the finite difference operators. That is, the error in (2.9) takes the form

$$\mathcal{D}_{\theta_i\theta_i}u = u_{\theta_i\theta_i} + \frac{1}{3}u_{\theta_i\theta_i\theta_i}(r(\theta_i) - r(\theta_{i+\pi})) + \mathcal{O}(r^2),$$

where $|r(\theta_i)| \leq r$ for all $i = 0, \ldots, M$. Notice that by symmetry and periodicity, we have that

$$u_{\theta+\pi,\theta+\pi} = u_{\theta\theta}, \quad u_{\theta+\pi,\theta+\pi,\theta+\pi} = -u_{\theta\theta\theta}, \quad r(\theta + 2\pi) = r(\theta).$$

For ease of notation, we extend the indexing such that $\theta_{i+M+1} = \theta_i + \pi$ for $i = 0, \ldots, M$.

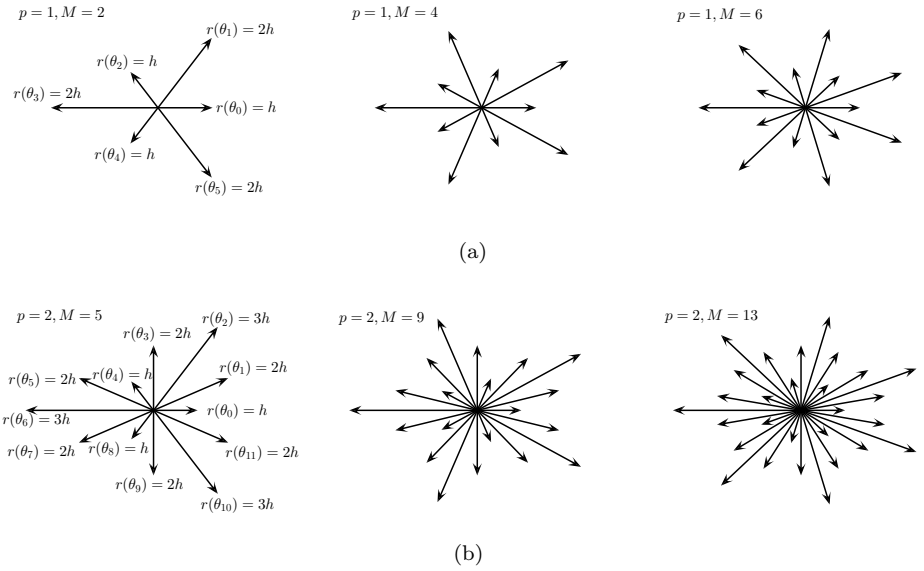

FIG. 2. *Examples of sequences of stencils corresponding to* (a) $p=1$ *and* (b) $p=2$.

Now we assume there is sufficient symmetry in the grids $\mathcal{G}^h$ so that, for some fixed $p \in \mathbb{N}$ and every $h>0$, we have $r(\theta_{i+p}) - r(\theta_{i+p}+\pi) = -(r(\theta_i) - r(\theta_i+\pi))$. Note that this condition requires that $M+1$, the number of terms in the angular discretization of $[0,\pi)$, is an odd multiple of p for each grid $\mathcal{G}^h$ and corresponding stencil. See Figure 2. We argue that if $\lambda_1(D^2u) > 0$, we can expect the formal truncation error of (4.1) to be $\mathcal{O}(\tau_Q(d\theta) + r^2 + \epsilon_2)$ despite the fact that the underlying finite difference approximations have only $\mathcal{O}(r)$ accuracy.

We first note that under this symmetry condition, there are at most p possible values that $|r(\theta_i) - r(\theta_i+\pi)|$ can take. To access these, we rewrite the indices $i = 0,\dots,2M+1$ in the form $i = kp+j$ where $j = 0,\dots,p-1$ and $k = 0,\dots,(2M+2)/p-1$. Then by p-periodicity, we find that

$$|r(\theta_{kp+j}) - r(\theta_{kp+j}+\pi)| = |r(\theta_j) - r(\theta_j+\pi)|.$$

Because the sign alternates every p steps, we can further characterize

$$r(\theta_{kp+j}) - r(\theta_{kp+j}+\pi) = (-1)^k (r(\theta_j) - r(\theta_j+\pi)),$$

which can take at most $2p$ distinct values.

In the setting $\lambda_1(D^2u) > 0$ (so that all $u_{\theta\theta} \geq \lambda_1(D^2u) > 0$), we have that for sufficiently small $r, \epsilon > 0$, the summation appearing in (4.1) can be expressed as

$$\begin{aligned}\sum_{i=0}^{M} \frac{w_i}{\max\{\mathcal{D}_{\theta_i\theta_i} u, \epsilon_1\}} &= \frac{\pi}{M+1} \sum_{i=0}^{M} \frac{1}{u_{\theta_i\theta_i} + \frac{1}{3} u_{\theta_i\theta_i\theta_i}(r(\theta_i) - r(\theta_{i+\pi})) + \mathcal{O}(r^2)} \\ &= \frac{\pi}{M+1} \sum_{i=0}^{M} \frac{1}{u_{\theta_i\theta_i}} - \frac{\pi}{3(M+1)} \sum_{i=0}^{M} \frac{u_{\theta_i\theta_i\theta_i}(r(\theta_i) - r(\theta_i+\pi))}{u_{\theta_i\theta_i}} + \mathcal{O}(r^2),\end{aligned}$$

where the last line here follows from a binomial expansion.

Now we exploit symmetry and the fact that $1/u_{\theta\theta}$ is smooth (since $\lambda_1(D^2u) > 0$) to re-express this as

$$\begin{aligned}
&\sum_{i=0}^{M} \frac{w_i}{\max\{\mathcal{D}_{\theta_i\theta_i} u, \epsilon_1\}} \\
&\quad = \int_0^\pi \frac{1}{u_{\theta\theta}} d\theta + \mathcal{O}(\tau_Q(d\theta) + r^2) - \frac{\pi}{6(M+1)} \sum_{i=0}^{M} \Bigg(\frac{u_{\theta_i\theta_i\theta_i}(r(\theta_i) - r(\theta_i+\pi))}{u_{\theta_i\theta_i}} \\
&\qquad\qquad + \frac{-u_{\theta_i+\pi,\theta_i+\pi,\theta_i+\pi}(r(\theta_i+2\pi) - r(\theta_i+\pi))}{u_{\theta_i+\pi,\theta_i+\pi}} \Bigg) \\
&\quad = \int_0^\pi \frac{1}{u_{\theta\theta}} d\theta + \mathcal{O}(\tau_Q(d\theta) + r^2) - \frac{\pi}{6(M+1)} \sum_{i=0}^{2M+1} \frac{u_{\theta_i\theta_i\theta_i}(r(\theta_i) - r(\theta_i+\pi))}{u_{\theta_i\theta_i}}.
\end{aligned}$$

Next, we utilize the periodicity of the terms $r(\theta_i) - r(\theta_i+\pi)$ with respect to shifts of $2p$ in the index. This allows us to rewrite the sum as

$$\begin{aligned}
&\sum_{i=0}^{M} \frac{w_i}{\max\{\mathcal{D}_{\theta_i\theta_i} u, \epsilon_1\}} \\
&\quad = \int_0^\pi \frac{1}{u_{\theta\theta}} d\theta + \mathcal{O}(\tau_Q(d\theta) + r^2) \\
&\qquad - \frac{\pi}{6(M+1)} \sum_{j=0}^{p-1} \sum_{k=0}^{(2M+2)/p-1} \frac{u_{\theta_{pk+j}\theta_{pk+j}\theta_{pk+j}}(-1)^k (r(\theta_j) - r(\theta_j+\pi))}{u_{\theta_{pk+j}\theta_{pk+j}}} \\
&\quad = \int_0^\pi \frac{1}{u_{\theta\theta}} d\theta + \mathcal{O}(\tau_Q(d\theta) + r^2) \\
&\qquad - \frac{1}{6p} \sum_{j=0}^{p-1} (r(\theta_j) - r(\theta_j+\pi)) \Bigg(\frac{\pi p}{M+1} \sum_{k \text{ even}} \frac{u_{\theta_{pk+j}\theta_{pk+j}\theta_{pk+j}}}{u_{\theta_{pk+j}\theta_{pk+j}}} \\
&\qquad\qquad - \frac{\pi p}{M+1} \sum_{k \text{ odd}} \frac{u_{\theta_{pk+j}\theta_{pk+j}\theta_{pk+j}}}{u_{\theta_{pk+j}\theta_{pk+j}}} \Bigg).
\end{aligned}$$

Now we notice that each of the sums in the last line can be interpreted as the trapezoid rule applied to integral

$$\int_0^{2\pi} \frac{u_{\theta\theta\theta}}{u_{\theta\theta}} d\theta = 0$$

using equally spaced angles with $d\tilde{\theta} = 2p\, d\theta$. Thus formally, we expect that

$$\sum_{i=0}^{M} \frac{w_i}{\max\{\mathcal{D}_{\theta_i\theta_i} u, \epsilon_1\}} = \int_0^\pi \frac{1}{u_{\theta\theta}} d\theta + \mathcal{O}(\tau_Q(d\theta) + r^2 + r\,\tau_Q(d\theta)).$$

Substituting this into the quadrature scheme (4.1) for the Monge–Ampère equation, we obtain an expected consistency error of $\mathcal{O}(\tau_Q(d\theta) + r^2 + \epsilon_2)$, which is better than the truncation error predicted by Corollary 4.7.

5. Implementation. We now present three specific implementations of a quadrature scheme based upon the formulation of (4.1).

The first two implementations rely on a hexagonal and a triangular tiling of the domain, respectively. The underlying structure of the grid allows us to design an angular discretization of $[0, 2\pi]$ involving 12 equispaced angles. The use of the trapezoid rule then leads to a compact finite difference stencil that in practice achieves spectral accuracy in the angular parameter $d\theta$ (which is held fixed).

The third implementation relies on a simple Cartesian grid combined with Simpson's rule for quadrature. As required by the convergence analysis (Theorems 4.24.3), the stencil does grow wider as the grid is refined. However, the optimal stencil width is asymptotically narrower than that required by existing monotone schemes, while simultaneously improving the order of the formal consistency error [15].

5.1. Discretization of domain. The implementations we describe rely on a discretization of the domain that consists of two components: (1) a structured mesh restricted to the interior of the domain and (2) a list of boundary points chosen to preserve the desired angular resolution. Hand in hand with this grid, we include the list of angles θ_j used to discretize the integral in (3.5).

We begin by presenting an algorithm for discretizing the domain, which applies to all types of mesh structure (hexagonal, triangular, and Cartesian) considered in this work. In the subsequent subsections, we will fill in the remaining details about each specific implementation of the quadrature scheme.

As a starting point, suppose we are given a structured mesh $\mathcal{M}$ that tiles $\mathbb{R}^2$ and a set of angles $0 \le \theta_0 < \cdots < \theta_M < \pi$. Moreover, the angles are chosen such that for every $x \in \mathcal{M}$ and $j = 0, \dots, M$, we have

$$x \pm r_j^{\pm}(x)(\cos\theta_j, \sin\theta_j) \in \mathcal{M}$$

for some $r_j^{\pm}(x) > 0$. That is, we are able to identify neighboring grid points aligned with all the directions in our angular discretization. To this underlying grid, we associate a stencil width defined by

$$r = \max\{r_j^{\pm}(x) \mid x \in \mathcal{M}; j = 0, \dots, M\}.$$

From this tiling of $\mathbb{R}^2$, we generate a set of discretization points $\mathcal{G}$ by (1) including all mesh points lying in the interior of the domain Ω and (2) supplementing with points in $\partial\Omega$ in order to preserve the existence of grid points perfectly aligned with the given set of angles. That is, given any interior node $x \in \mathcal{G} \cap \Omega$ and $j = 0, \dots, M$, we have

$$x \pm r_j^{\pm}(x)(\cos\theta_j, \sin\theta_j) \in \mathcal{G}$$

for some $r_j^{\pm}(x) > 0$.

As an example, consider the case where the domain Ω is a square, $\mathcal{M}$ is either a hexagonal or a triangular tiling of $\mathbb{R}^2$, and the desired angles are $\theta_j = \frac{j\pi}{6}$ for $j = 0, \dots, 5$. The resulting meshes are pictured in Figure 3. An example involving an underlying Cartesian grid is shown in Figure 4.

The meshing of the domain can be easily accomplished if the domain Ω is represented through the signed distance function $d_{\partial\Omega}(x)$ to its boundary $\partial\Omega$. See Algorithm 5.1.

5.2. Implementation on hexagonal or triangular grids. The first implementation we suggest is motivated by a desire to exploit the spectral accuracy of the trapezoid rule when applied to a uniform discretization of the angles ($d\theta_j = d\theta$ for all $j = 0, \dots, M$).

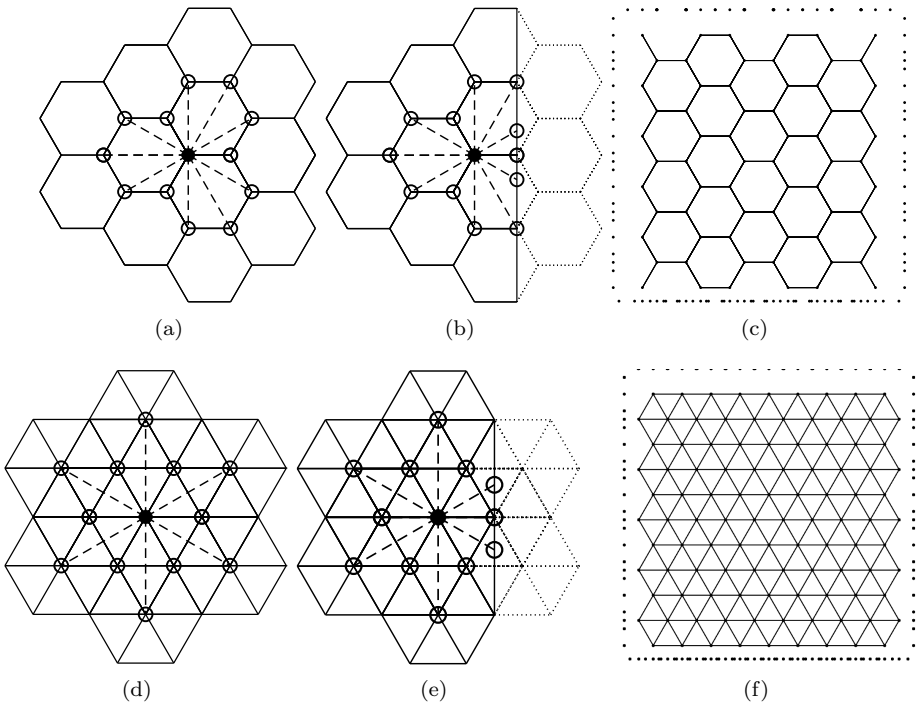

FIG. 3. (a) *A hexagonal mesh, discrete set of angles, and neighboring mesh points aligned with those angles.* (b) *An example of boundary points that are added to the grid.* (c) *Meshing of a square domain using a hexagonal mesh augmented with boundary points.* (d) *A triangular mesh, discrete set of angles, and neighboring mesh points aligned with those angles.* (e) *An example of boundary points that are added to the grid.* (f) *Meshing of a square domain using a triangular mesh augmented with boundary points.*

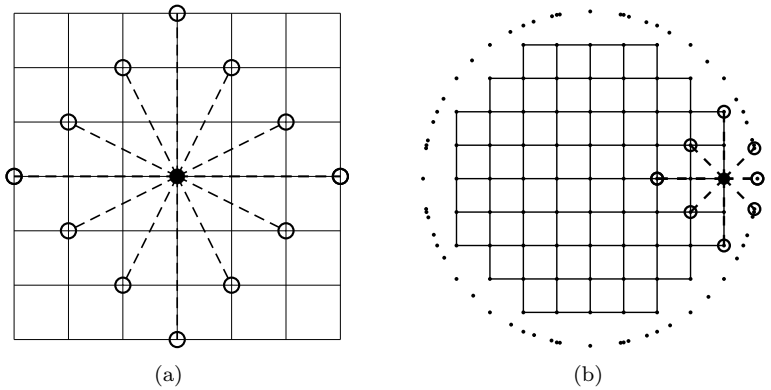

FIG. 4. (a) *A stencil with L^1 width $r = 3h$ on a Cartesian grid and* (b) *an example set of grid points used to discretize a disc.*

In order to obtain as many equispaced angles as possible, we propose to let the underlying mesh $\mathcal{M}$ be a tiling of $\mathbb{R}^2$ with either regular hexagons or equilateral triangles. Then we can achieve a grid-aligned scheme using the uniform angular discretization $\theta_j = \frac{j\pi}{6}$ for $j = 0, \ldots, 5$. See Figure 3.

Algorithm 5.1 Discretization of the domain Ω.

$\mathcal{G} \leftarrow \{x \in \mathcal{M} \mid d_{\partial\Omega}(x) < 0\}$
for $x \in \mathcal{G}$ such that $0 < |d_{\partial\Omega}(x)| < r$ **do**
  for $j = 0, \dots, M$ **do**
    $r^{\pm} \leftarrow \min\{h > 0 \mid x \pm h(\cos\theta_j, \sin\theta_j) \in \mathcal{M}\}$
    if $r^{\pm}(\cos\theta_j, \sin\theta_j) \notin \mathcal{G}$ **then**
      $t^{\pm} \leftarrow$ Positive solution of

$$d_{\partial\Omega}(x \pm t(\cos\theta_j, \sin\theta_j)) = 0$$

      $\mathcal{G} \leftarrow \mathcal{G} \cup \{x \pm t^{\pm}(\cos\theta_j, \sin\theta_j)\}$
    end if
  end for
end for

The resulting angular resolution is $d\theta = \frac{\pi}{6}$. Applying the trapezoid rule (4.2), we obtain equal quadrature weights $w_i = \frac{\pi}{6}$.

We notice that fixing $d\theta$ also has the effect of fixing the stencil width $r = \mathcal{O}(h)$. We recall that second directional derivatives in the directions $\nu = (\cos\theta, \sin\theta)$ are discretized by (2.9) as follows:

$$\mathcal{D}_{\nu\nu}u(x) = 2\frac{h^- u(x + h^+\nu) + h^+ u(x - h^-\nu) - (h^+ + h^-)u(x)}{h^+h^-(h^+ + h^-)}.$$

In general, the use of narrow stencils on a hexagonal mesh leads to stencils that are aligned but not necessarily centered ($h^+ \neq h^-$); see Figure 3(a). Thus the truncation error of these finite differences satisfies $\tau_{FD}(r) = h$. An improvement to centered stencils with $\tau_{FD}(r) = h^2$ is possible by allowing each stencil to extend across the width of two hexagons. However, we also note that the compact stencils illustrated in Figure 3(a) satisfy the symmetry condition discussed in section 4.4 with $p = 2$. Thus we choose to limit our implementation to the uncentered compact stencils with the expectation (which is confirmed by numerical experiments) that the resulting scheme for approximating Monge–Ampère will nevertheless display second-order accuracy in the spatial resolution parameter. The use of narrow stencils on the triangular mesh leads to stencils that are both aligned and centered so that $\tau_{FD}(r) = h^2$ automatically; see Figure 3(d).

Finally, we choose the regularization parameters $\epsilon_1 > 0, \epsilon_2 \geq 0$. From the perspective of consistency error, choosing these to be as small as possible ($\epsilon_1 \ll h$, $\epsilon_2 = 0$) might seem ideal. However, allowing $\epsilon_2 < \epsilon_1$ is undesirable for many solution methods such as Newton's method. In the regime where $\epsilon_2 I < D^2u(x) < \epsilon_1 I$, the resulting scheme F^h would be insensitive to perturbations in u and the corresponding Jacobian ∇F^h would be singular. Moreover, larger values of ϵ_1, ϵ_2 are preferable in this regime since increasing these parameters tends to improve the conditioning of the scheme and its Jacobian. With these factors in mind, we suggest a choice of $\epsilon_1 = \epsilon_2 = h^2$, which is smaller than the other terms appearing in the truncation error and will not impact the overall order of scheme.

From Corollary 4.7, the overall formal truncation error of the quadrature scheme (4.1) is $\mathcal{O}(h + d\theta^p)$ for every $p > 0$, though in our implementation $d\theta$ is held fixed and the scaling constant depends on p.

We notice that with a fixed stencil, the truncation error $\tau_Q(d\theta)$ of the quadrature scheme does not converge to zero. Nevertheless, at points $x \in \Omega$ where u is smooth, we expect the overall truncation error of the scheme to be dominated by the remaining terms $\tau_{FD}(r)$ and ϵ unless the grid is very highly resolved. Thus in principle the scheme (4.1) is not consistent. It is certainly possible to create wider-stencil extensions of this as $h \to 0$, though at the expense of a uniform angular discretization. However, in practice we expect that these wider stencils will not need to be engaged until the grid spacing h is very small. Thus we do expect to see this scheme outperform lower-order schemes ($\mathcal{O}(h^p)$, $p < 2$) for most practical refinements of the grid when solutions are smooth enough. Indeed, this is what we observe for all but the most singular and/or degenerate of our computational examples; see section 6.

5.3. Implementation on Cartesian grids. The second implementation we propose is based upon a uniform Cartesian grid. In order to achieve true consistency and convergence, we will allow for stencils that grow wider as the grid is refined. To maintain grid-alignment, we are then forced to utilize a nonuniform angular discretization. This prevents the use of a spectrally accurate trapezoid rule. However, by exploiting higher-order quadrature schemes, we can still produce monotone schemes with improved consistency error on more compact stencils.

Our particular implementation will perform quadrature using Simpson's rule, as outlined in (4.3)–(4.4). The truncation error of this quadrature rule is $\tau_Q(d\theta) = d\theta^4$.

Given a desired stencil width $r = Kh$ for some $K \in \mathbb{N}$, we select an angular discretization by considering neighboring grid points that are a distance r from the reference point x as measured by the L^1 norm.

Specifically, we consider a set of angles $\theta_0 < \cdots < \theta_M$, where $M = 2K - 1$ is odd. Letting h be the standard grid point spacing in the Cartesian grid, we let r_j and θ_j be the polar coordinates of the grid points

$$r_j(\cos\theta_j, \sin\theta_j) = h\left(K - j, K - |K - j|\right), \quad j = 0, \ldots, 2K - 1. \tag{5.1}$$

See Figure 4.

An important consequence of this choice of angles is that it has a uniformly bounded quasi-uniformity constant, as required for consistency (Theorem 4.3). Moreover, as $d\theta \to 0$ the ratios

$$\frac{d\theta_{j+1}}{d\theta_j} \to 1.$$

This ensures that the quadrature weights (4.4) are strictly positive, as required for monotonicity (Theorem 4.2).

LEMMA 5.1 (quasi-uniformity). *The angular discretization defined in* (5.1) *has a uniformly bounded quasi-uniformity constant.*

Proof. We bound the ratio $d\theta/d\theta_j$ for $j = 0, \ldots, K$. The remaining cases are identical by symmetry.

Notice that the local stencil width is given by

$$r_j^2 = h^2\left((K - j)^2 + j^2\right).$$

This is bounded by

$$\frac{hK}{\sqrt{2}} \le r_j \le hK.$$

We can also compute the local angular resolution via

$$\begin{aligned}\sin d\theta_j &= \sin(\theta_{j+1} - \theta_j)\\ &= \sin\theta_{j+1}\cos\theta_j - \cos\theta_{j+1}\sin\theta_j\\ &= \frac{h^2}{r_j r_{j+1}}\left((j+1)(K-j) - (K-j-1)j\right)\\ &= \frac{Kh^2}{r_j r_{j+1}}.\end{aligned}$$

Bounds on r_j imply that for every j,

$$\frac{\sin d\theta}{\sin d\theta_j} \le \frac{2/K}{1/K} = 2.$$

Thus the upper bound on the quasi-uniformity constant Q converges to 2 as $K \to \infty$. ∎

LEMMA 5.2 (ratios of angles). *The angular discretization defined in* (5.1) *satisfies*

$$\frac{d\theta_j}{d\theta_{j+1}} \to 1$$

as $K \to \infty$.

Proof. As in the proof of the previous lemma, we can use symmetry to limit ourselves to considering $j = 0, \dots, K$ and compute

$$\frac{\sin d\theta_j}{\sin d\theta_{j+1}} = \frac{r_{j+2}}{r_j} = \frac{(K-j-2)^2 + (j+2)^2}{(K-j)^2 + j^2} = 1 + \frac{8 - 4K + 8j}{(K-j)^2 + j^2}.$$

We notice that

$$\frac{|8 - 4K + 8j|}{(K-j)^2 + j^2} \le \frac{20K}{K^2/2},$$

which converges to zero as $N \to \infty$.

Therefore

$$\frac{\sin d\theta_j}{\sin d\theta_{j+1}} \to 1$$

as $K \to \infty$. ∎

From the proofs of the previous lemmas, we notice that $d\theta = \mathcal{O}(1/K)$. Since we initially chose the search radius $r = Kh$, we find the following relationship between the grid parameters described in Definition 4.1:

$$d\theta = \mathcal{O}\left(\frac{h}{r}\right).$$

The uniform Cartesian grid allows us to use centered differences to discretize the second derivatives (2.9), so that $\tau_{FD}(r) = r^2$. We recall also that Simpson's rule satisfies $\tau_Q(d\theta) = d\theta^4$.

Combining these terms, we find that the formal truncation error of the quadrature scheme (Corollary 4.7) is given by

$$\mathcal{O}(\tau_Q(d\theta) + \tau_{FD}(r) + \epsilon) = \mathcal{O}(h^4/r^4 + r^2 + \epsilon).$$

An optimal choice is obtained by the stencil width

$$r = \mathcal{O}(h^{2/3}),$$

which leads to an angular resolution of $d\theta = \mathcal{O}(h^{1/3})$. Choosing $\epsilon \leq r^2 = \mathcal{O}(h^{4/3})$, we find that the formal consistency error of the scheme is given by

$$\mathcal{O}(h^{4/3}).$$

Moreover, these choices satisfy all the requirements of consistency (Theorem 4.3).

We should remark that in a small band of radius r near the boundary of Ω, it may not be possible to use centered finite differences. Instead, we must fall back on uncentered differences in (2.9) ($h^+ \neq h^-$) so that $\tau_{FD}(r) = r = \mathcal{O}(h^{2/3})$. This reduces the overall truncation error of the scheme to $\mathcal{O}(h^{2/3})$. However, we emphasize that this occurs only in a narrow band, which vanishes as $h \to 0$. In our computational experiments (section 6), we found that the reduced accuracy at a small number of points had *no* impact on the global accuracy of the method.

We notice that with a careful selection of grid parameters, the quadrature-based scheme can produce substantial improvements over monotone schemes such as the work of [15], which requires a much larger stencil width $r = \mathcal{O}(\sqrt{h})$ and produces a significantly worse truncation error $\mathcal{O}(\sqrt{h})$. This is possible because higher-order quadrature rules allow for substantial improvements in the component of the error coming from the angular resolution $d\theta$, which can be made arbitrarily small with the use of higher-order quadrature schemes.

The formal convergence of our quadrature scheme is superlinear in h, which is of great value when the goal is to approximate solution gradients (which is common in problems related to optimal transport). Moreover, higher-order quadrature rules could be substituted in place of Simpson's rule to provide even greater improvements in both stencil width and truncation error. In general, a quadrature rule satisfying $\tau_Q(d\theta) = h^p$ can be combined with a stencil width $r = h^{1-2/p}$ to produce a scheme with a formal truncation error of $\mathcal{O}(h^{2-4/p})$.

6. Computational results. In this section, we present numerical results for the Monge–Ampère equation ($\det^+(D^2u(x)) = f(x)$) with Dirichlet boundary conditions ($u(x) = g(x)$). To accomplish this, we solve a system of the form

(6.1)
$$F^h(x, u^h(x), u^h(x) - u^h(\cdot)) = \begin{cases} G^h(x, u^h(x), u^h(x) - u^h(\cdot)) + f(x), & x \in \mathcal{G} \cap \Omega, \\ u^h(x) - g(x), & x \in \mathcal{G} \cap \partial\Omega, \end{cases}$$

where G^h is a consistent, monotone approximation of the convexified Monge–Ampère operator.

We will compare the results of the following four schemes:

- the quadrature scheme on a hexagonal grid described in section 5.2,
- the quadrature scheme on a triangular grid described in section 5.2,
- the quadrature scheme on a Cartesian grid described in section 5.3,
- the method of [22], which relies on a variational formulation of the Monge–Ampère operator,

$$\det^+(D^2u) = \min_{\nu_1 \cdot \nu_2 = 0} \prod_{j=1}^{2} \max\{u_{\nu_j \nu_j}, 0\},$$

discretized using centered differences on the Cartesian grid described in section 5.3.

6.1. Numerical implementation. To discretize the domain using a Cartesian grid (for either the quadrature or variational schemes), we begin with an underlying $N \times N$ grid that contains the domain. To discretize the domain using a hexagonal tiling, we begin with a tiling covering the domain that contains N points along the vertical dimension, with (approximately) $\frac{1}{2}N$ points along the horizontal dimension at each fixed nodal value of y. To discretize the domain using a triangular tiling, we begin with a tiling covering the domain that contains N points along the vertical dimension, with (approximately) N points along the horizontal dimension at each fixed nodal value of y. The grids are then restricted to the interior and augmented with boundary points using Algorithm 5.1. In every case, we have $N \approx \frac{k}{h}$, where the constant k depends only on the size of the domain.

Each of the three discretizations we consider results in a nonlinear algebraic system of equations. We solve these using a damped Newton's method

$$\nabla F^h[u_n]y_n = -(f + F^h[u_n]),$$
$$u_{n+1} = u_n + \alpha_n y_n,$$

where the value of α_n is chosen at each step to ensure that the residual $r_n = \|F^h[u_n] + f\|_\infty$ is always decreasing. We run Newton's method until the residual falls below the threshold $r_n < h^2$ since, for a more degenerate/singular example, quadratic convergence is not always observed until the residual is very small.

To obtain an initial guess u_0 for the Newton solver, we first solve the following Poisson equation, which is obtained through linearization of the Monge–Ampère equation [4, 16]:

$$\begin{cases} \Delta u(x) = \sqrt{2f(x)}, & x \in \Omega, \\ u(x) = g(x), & x \in \partial\Omega. \end{cases}$$

The solution process can be accelerated slightly by first solving the linearized problem (6.2) on a coarse $N \times N$ grid, solving the nonlinear problem via Newton's method on the same coarse grid, then interpolating onto the desired refined grid to initialize the final Newton solver.

6.2. Representative examples. We test our methods using four representative benchmark examples. For simplicity of comparison, each example is posed on a square domain. However, we should note that this is *not* a simplifying assumption for the quadrature method, which performs equally well on general convex domains.

The first example is defined on the domain $\Omega = (-1, 1)^2$ and has a smooth, radially symmetric solution $u \in C^\infty(\Omega)$:

$$u(\mathbf{x}) = \exp\left(\frac{|\mathbf{x}|^2}{2}\right), \quad f(\mathbf{x}) = \left(1 + |\mathbf{x}|^2\right)\exp\big(|\mathbf{x}|^2\big). \tag{6.2}$$

The second example is defined on the domain $\Omega = (0, 1)^2$ and includes a "fully degenerate" region where both eigenvalues of D^2u are 0. The solution $u \in C^1(\Omega)$ is only continuously differentiable. We introduce the constant $\mathbf{x}_0 = (0.5, 0.5)$ and let

$$u(\mathbf{x}) = \frac{1}{2}\big((|\mathbf{x} - \mathbf{x}_0| - 0.2)^+\big)^2, \quad f(\mathbf{x}) = \left(1 - \frac{0.2}{|\mathbf{x} - \mathbf{x}_0|}\right)^+. \tag{6.3}$$

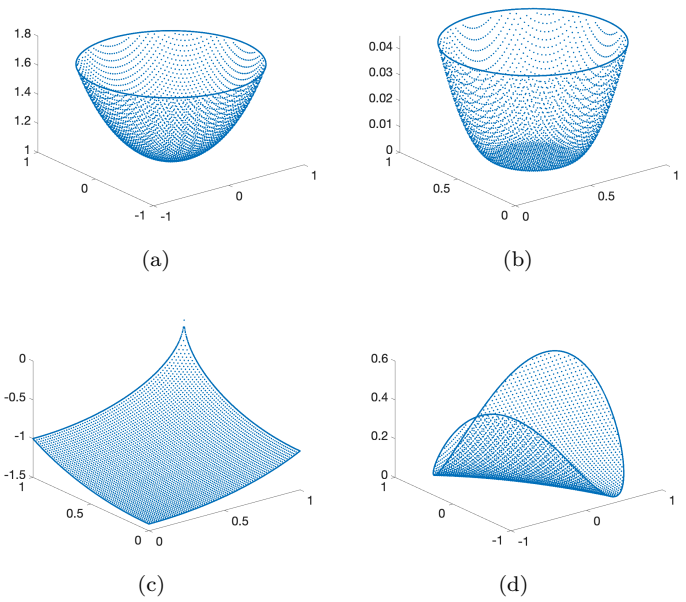

FIG. 5. *Numerical solutions for* (a) C^2 *example,* (b) C^1 *example,* (c) *example with gradient blow-up, and* (d) *semidegenerate example.*

The third example has domain $\Omega = (0,1)^2$, and the solution is twice differentiable in the interior of the domain. However, the solution gradient becomes unbounded near the boundary point $(1,1)$:

$$u(\mathbf{x}) = -\sqrt{2 - |\mathbf{x}|^2}, \quad f(\mathbf{x}) = 2\big(2 - |\mathbf{x}|^2\big)^{-2}. \tag{6.4}$$

The final example is defined on the domain $\Omega = (-1,1)^2$ and the solution $u \in C^2(\Omega)$ is in fact a polynomial. We introduce the vector $\vec{\gamma} = (\frac{1}{\sqrt{2}}, 1 - \frac{1}{\sqrt{2}})$ and let

$$u(\mathbf{x}) = (\vec{\gamma} \cdot \mathbf{x})^2, \quad f(\mathbf{x}) = 0. \tag{6.5}$$

The solution is "semidegenerate" on the entire domain, with $D^2u(x)$ having one positive and one vanishing eigenvalue at each point in the domain. This fully semidegenerate example, while somewhat artificial, should be viewed as an "edge case" for the quadrature scheme since the truncation error degrades in this setting (Lemma 4.9).

See Figure 5 for graphs of the solutions u, which were obtained using the Cartesian quadrature scheme.

6.3. Numerical results. The maximum error for each test is displayed in Figure 6. We find that the hexagonal and triangular implementations display effectively quadratic convergence for smooth enough tests and small enough values of N. As expected, the error eventually levels off for less regular examples and larger values of N, though these implementations continue to outperform the others over a large range of refinements. The Cartesian implementation of the quadrature scheme displays the expected superlinear $\mathcal{O}(N^{-4/3})$ convergence; surprisingly, this continues to be true even for the less regular examples. On the semidegenerate example, we observe nonmonotonic convergence of the Cartesian scheme as the grid is refined. This is likely

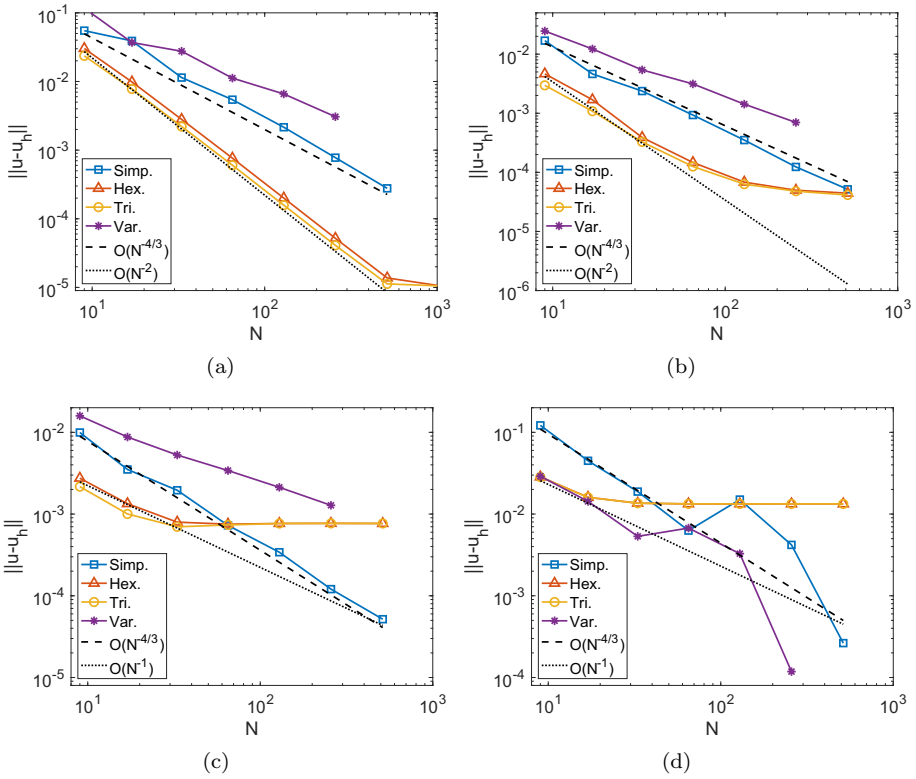

FIG. 6. *Convergence tests for the* (a) C^2 *example,* (b) C^1 *example,* (c) *example with gradient blow-up, and* (d) *semidegenerate example.*

due to the fact that the parameter $\vec{\gamma}$ is non-grid-aligned; a coarser resolution can, by chance, include an angle that aligns closely to $\vec{\gamma}$, leading to improved accuracy.

Comparison with the variational scheme demonstrates the clear superiority of the quadrature-based method that is made possible by reducing the angular component of the error. Because the variational implementation has limited accuracy in the angular component (truncation error is $\mathcal{O}(d\theta^2 + r^2)$), while $d\theta$ itself goes to zero very slowly in the wide-stencil schemes ($d\theta = \mathcal{O}(h/r) \gg h$), solution error is at best $\mathcal{O}(h)$.

The improvement achieved by the quadrature schemes becomes even more pronounced when the improvement in computational cost is factored in. See Figure 7 for plots of solution error as a function of computation time. It is clear that for smooth, and even moderately nonsmooth, examples, the hexagonal and triangular implementations provide the best results despite the fact that they are not technically consistent in the limit $N \to \infty$. On the most singular examples (e.g., blow-up in the gradient), the Cartesian implementation takes over as the most efficient. All quadrature schemes dramatically outperform the variational scheme, which requires a much wider stencil ($r = \mathcal{O}(\sqrt{h})$) to optimize truncation error. The only exception to this trend is the semidegenerate example. As noted before, this can be viewed as an "edge case" where the variational scheme will sometimes perform unusually well because (1) the centered finite difference approximations are exact on quadratics and (2) chance near-alignment between the eigenvectors of the Hessian and the underlying Cartesian grid can drastically reduce the truncation error. Indeed, the

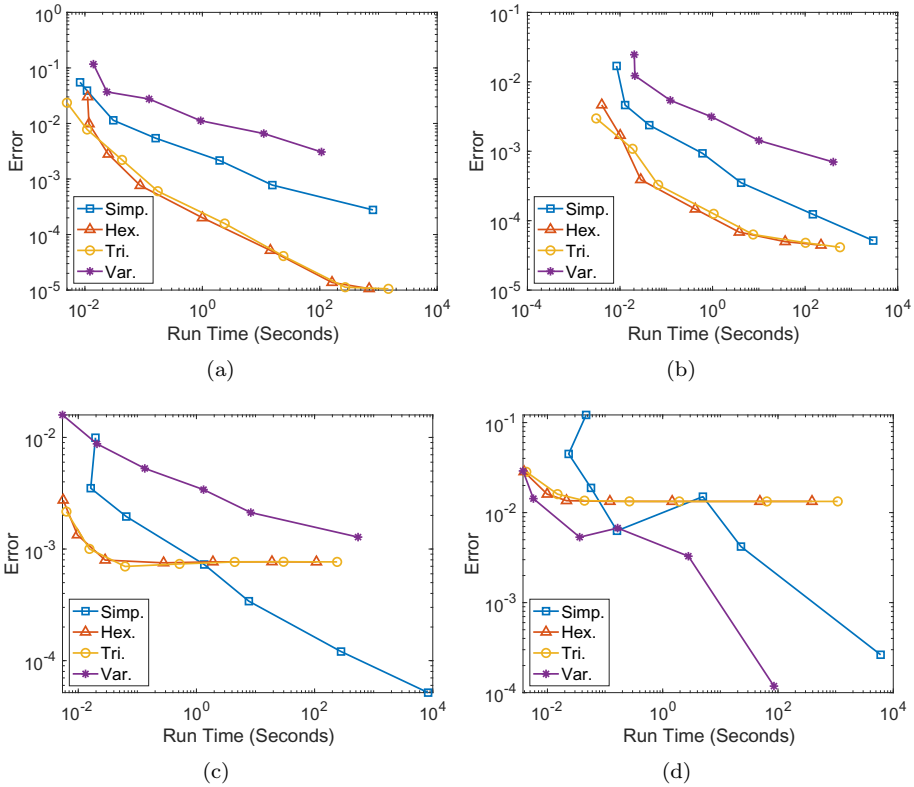

FIG. 7. *Efficiency results for the* (a) C^2 *example,* (b) C^1 *example,* (c) *example with gradient blow-up, and* (d) *semidegenerate example.*

performance of the Cartesian quadrature scheme is still good even on this challenging test problem.

7. Conclusion. In this paper we presented a new integral representation of the Monge–Ampère operator. We showed that this can be combined with different quadrature rules to produce a family of monotone finite difference methods. Importantly, these methods fit directly into existing convergence proofs for the Dirichlet [16, 19, 27, 29] or optimal transport problems [3, 5, 20].

Existing monotone methods for the Monge–Ampère equation rely on wide finite difference stencils. The resulting truncation error depends upon several factors: the typical spacing of grid points h, the width of the stencil r, and the angular resolution of the stencil $d\theta$. The use of higher-order quadrature schemes allows us to substantially reduce the component of the error coming from the angular resolution. This, in turn, allows for significant reductions in both the stencil width r and the overall truncation error of the scheme. The end result is a monotone (convergent) method that achieves significant gains in both accuracy and efficiency.

We provided three implementations of this method. The first two combined the spectrally accurate trapezoid rule with an underlying hexagonal or triangular mesh. The resulting methods involved a simple nearest-neighbors scheme which is highly efficient and achieves second-order convergence in practice for smooth enough solutions and reasonable grid refinements. The third method utilized a nonuniform Simpson's

rule on a Cartesian mesh. The method is provably convergent, relies on relatively narrow stencils of width $r = \mathcal{O}(h^{2/3})$, is highly robust with respect to solution regularity, and provides superlinear convergence of order $\mathcal{O}(h^{4/3})$. Moreover, this implementation could easily be adapted to accommodate other higher-order quadrature rules, making possible a formal convergence rate of $\mathcal{O}(h^{2-4/p})$ for any $p > 0$.

This approach holds particularly great promise for the three-dimensional Monge–Ampère equation, for which existing discretizations can be prohibitively expensive [21]. Typical schemes rely on some variational form of the Monge–Ampère equation, which requires performing optimization over a three-dimensional set at each point in the three-dimensional domain. A method based upon the integral reformulation would reduce this to the cost of integrating over the sphere, which is two-dimensional. We also expect this approach to adapt well to generalized Monge–Ampère equations arising in optimal transport problems in the plane [14] or on the sphere [23].

REFERENCES

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

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

[3] J.-D. Benamou and V. Duval, *Minimal convex extensions and finite difference discretisation of the quadratic Monge-Kantorovich problem*, European J. Appl. Math., 30 (2019), pp. 1041–1078.

[4] J.-D. Benamou, B. D. Froese, and A. M. Oberman, *Two numerical methods for the elliptic Monge-Ampère equation*, Math. Model. Numer. Anal., 44 (2010), pp. 737–758.

[5] G. Bonnet and J.-M. Mirebeau, *Monotone discretization of the Monge-Ampère equation of optimal transport*, ESAIM Math. Model. Numer. Anal., 56 (2022), pp. 815–865.

[6] S. C. Brenner, T. Gudi, M. Neilan, and L.-Y. Sung, *C^0 penalty methods for the fully nonlinear Monge-Ampère equation*, Math. Comp., 80 (2011), pp. 1979–1995.

[7] C. Budd and J. Williams, *Moving mesh generation using the parabolic Monge-Ampère equation*, SIAM J. Sci. Comput., 31 (2009), pp. 3438–3465.

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

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

[10] M. Cullen, J. Norbury, and R. Purser, *Generalised Lagrangian solutions for atmospheric and oceanic flows*, SIAM J. Appl. Math., 51 (1991), pp. 20–31.

[11] E. J. Dean and R. Glowinski, *Numerical methods for fully nonlinear elliptic equations of the Monge-Ampère type*, Comput. Methods Appl. Mech. Engrg., 195 (2006), pp. 1344–1386.

[12] B. Engquist and B. D. Froese, *Application of the Wasserstein metric to seismic signals*, Commun. Math. Sci., 12 (2014), pp. 979–988.

[13] X. Feng and M. Neilan, *Vanishing moment method and moment solutions for fully nonlinear second order partial differential equations*, J. Sci. Comput., 38 (2009), pp. 74–98.

[14] B. D. Froese, *Generalised finite difference methods for Monge-Ampère equations*, in Mathematisches Forschungsinstitut Oberwolfach Report 7/2017, 2017, pp. 383–386.

[15] B. D. Froese, *Meshfree finite difference approximations for functions of the eigenvalues of the Hessian*, Numer. Math., 138 (2018), pp. 75–99.

[16] B. D. Froese and A. M. Oberman, *Convergent finite difference solvers for viscosity solutions of the elliptic Monge-Ampère equation in dimensions two and higher*, SIAM J. Numer. Anal., 49 (2011), pp. 1692–1714.

[17] C. E. Gutiérrez, *The Monge-Ampère Equation*, Progr. Nonlinear Differential Equations Appl. 44, Springer, New York, 2001.

[18] S. Haker, L. Zhu, A. Tannenbaum, and S. Angenent, *Optimal mass transport for registration and warping*, Int. J. Comput. Vis., 60 (2004), pp. 225–240.

[19] B. Hamfeldt, *Convergent approximation of non-continuous surfaces of prescribed Gaussian curvature*, Commun. Pure Appl. Anal., 17 (2018), pp. 671–707.

[20] B. Hamfeldt, *Convergence framework for the second boundary value problem for the Monge-Ampère equation*, SIAM J. Numer. Anal., 57 (2019), pp. 945–971.
[21] B. F. Hamfeldt and J. Lesniewski, *Convergent finite difference methods for fully nonlinear elliptic equations in three dimensions*, J. Sci. Comput., 90 (2022).
[22] B. F. Hamfeldt and T. Salvador, *Higher-order adaptive finite difference methods for fully nonlinear elliptic equations*, J. Sci. Comput., 75 (2018), pp. 1282–1306.
[23] B. F. Hamfeldt and A. G. R. Turnquist, *A convergent finite difference method for optimal transport on the sphere*, J. Comput. Phys., 445 (2021).
[24] M. Kocan, *Approximation of viscosity solutions of elliptic partial differential equations on minimal grids*, Numer. Math., 72 (1995), pp. 73–92.
[25] J.-M. Mirebeau, *Discretization of the* 3D *Monge-Ampere operator, between wide stencils and power diagrams*, ESAIM Math. Model. Numer. Anal., 49 (2015), pp. 1511–1523.
[26] T. S. Motzkin and W. Wasow, *On the approximation of linear elliptic differential equations by difference equations with positive coefficients*, J. Math. Phys., 31 (1952), pp. 253–259.
[27] R. Nochetto, D. Ntogkas, and W. Zhang, *Two-scale method for the Monge-Ampère equation: Convergence to the viscosity solution*, Math. Comp., 88 (2019), pp. 637–664.
[28] A. M. Oberman, *Convergent difference schemes for degenerate elliptic and parabolic equations: Hamilton–Jacobi equations and free boundary problems*, SIAM J. Numer. Anal., 44 (2006), pp. 879–895.
[29] A. M. Oberman, *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 (2008), pp. 221–238.
[30] V. I. Oliker and L. D. Prussner, *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 (1988), pp. 271–293.
[31] G. Peyré and M. Cuturi, *Computational optimal transport: With applications to data science*, Found. Trends Mach. Learn., 11 (2019), pp. 355–607.
[32] C. R. Prins, R. Beltman, J. H. M. ten Thije Boonkkamp, W. L. IJzerman, and T. W. Tukker, *A least-squares method for optimal transport using the Monge-Ampère equation*, SIAM J. Sci. Comput., 37 (2015), pp. B937–B961.
[33] L. B. Romijn, M. J. H. Anthonissen, J. H. M. ten Thije Boonkkamp, and W. L. IJzerman, *Generating-function approach for double freeform lens design*, J. Opt. Soc. Am. A, 38 (2021), pp. 356–368.