SIAM J. SCI. COMPUT.
Vol. 41, No. 2, pp. B295–B319

CONVERGENT TWO-SCALE FILTERED SCHEME FOR THE MONGE–AMPÈRE EQUATION*

RICARDO H. NOCHETTO† AND DIMITRIOS NTOGKAS†

Abstract. We propose an extension to our monotone and convergent method for the Monge–Ampère equation in dimension $d \geq 2$ that incorporates the idea of filtered schemes. The method combines our original monotone operator with a more accurate nonmonotone modification, using an appropriately chosen filter. This results in a remarkable improvement of accuracy, but without sacrificing the convergence to the unique viscosity solution.

Key words. Monge–Ampère equation, filtered scheme, monotone, viscosity solution, convergence, accurate

AMS subject classifications. 65N30, 65N12, 35J96

DOI. 10.1137/18M1191634

1. Introduction.

We consider the Monge–Ampère equation with Dirichlet boundary condition:

$$\begin{cases} \det D^2 u = f & \text{in } \Omega \subset \mathbb{R}^d, \\ u = g & \text{on } \partial\Omega, \end{cases} \tag{1.1}$$

where Ω is a uniformly convex domain and $f \geq 0$ and g are uniformly continuous functions. We seek a *convex* solution u of (1.1), which is critical for (1.1) to be elliptic and have a unique viscosity solution [20].

The Monge–Ampère equation has a wide spectrum of applications in optimal mass transport problems, geometry, nonlinear elasticity, optics, and meteorology. These applications lead to an increasing interest in the investigation of efficient numerical methods. Existing methods for the Monge–Ampère equation include the early work by Oliker and Prussner [31, 28] for space dimension $d = 2$, the vanishing moment methods by Feng and Neilan [15, 16], the penalty method of Brenner et al. [8], least squares and augmented Lagrangian methods by Dean and Glowinski [11, 12, 19], and the finite difference methods proposed by Benamou, Froese, and Oberman [4], Froese and Oberman [17, 18], and Benamou, Collino, and Mirebeau [3, 24]. Feng and Jensen [13] have also recently proposed a semi-Lagrangian method that relies on an equivalent Hamilton–Jacobi–Bellman formulation of the Monge–Ampère equation. Schemes in [3, 13, 17, 18, 24] are closely related to ours and hinge on a wide stencil approach. We refer the reader to the survey [14].

In this work we extend our two-scale method from [25, 26], where we use continuous piecewise linear polynomials on a quasi-uniform mesh of size h and an approximation of the determinant that hinges on a second coarser scale δ, in order to solve the

*Submitted to the journal's Computational Methods in Science and Engineering section June 1, 2018; accepted for publication (in revised form) January 14, 2019; published electronically April 2, 2019.

http://www.siam.org/journals/sisc/41-2/M119163.html

Funding: This work was partially supported by the NSF under grant DMS-1411808. The work of the first author was partially supported by the Institute Henri Poincaré (Paris) and the Hausdorff Institute (Bonn). The work of the second author was partially supported by the 2016–2017 Patrick and Marguerite Sung Fellowship of the University of Maryland.

†Department of Mathematics, University of Maryland, College Park, MD 20742 (rhn@math.umd.edu, dimitnt@gmail.com).

Monge–Ampère equation numerically. In [25] we introduce the two-scale method and prove uniform convergence to the viscosity solution of (1.1), whereas in [26] we derive rates of convergence in L^∞ for classical and viscosity solutions that belong to certain Hölder and Sobolev spaces. The idea of a filtered scheme that we employ here is motivated by the work of Froese and Oberman in [18], but follows a different approach; we refer the reader to [30] and [6] for stationary and time dependent Hamilton–Jacobi equations. Instead of combining two different methods, we modify our monotone two-scale method into a more accurate, two-scale nonmonotone version that still relies on the same variational formulation for the determinant and combine it with the original monotone operator through a filter function. In order to computationally examine the performance of the scheme, we compare the L^∞ error of the monotone, the accurate, and the filtered schemes, using the two main examples from [25]. We observe that the filtered operator inherits the improved errors from the accurate operator, but allows the monotone operator to dominate the calculations whenever there is a discontinuity of the Hessian. We investigate this behavior and conclude with some computational observations about the scheme. We prove convergence to the viscosity solution of (1.1).

1.1. Our contribution. As in [17, 25] our method hinges on the following formula for the determinant of the positive semidefinite Hessian D^2w of a smooth convex function w:

$$\det D^2 w(x) = \min_{\boldsymbol{v}\in\mathbb{S}^\perp} \prod_{j=1}^{d} v_j^T D^2 w(x)\ v_j, \tag{1.2}$$

where $\mathbb{S}^\perp$ is the set of all d-orthonormal bases $\boldsymbol{v} = (v_j)_{j=1}^d,\ v_j \in \mathbb{R}^d$. The minimum in (1.2) is achieved by the eigenvectors of $D^2w(x)$ and is equal to the product of the respective eigenvalues. We can discretize the above formula in various ways, employing different polynomial spaces and approximations for the directional derivatives given by $v_j^T D^2 w v_j$. These choices lead to schemes with different theoretical properties and levels of accuracy. We first briefly recall the discretization used in [25, 26] and then introduce a more accurate approach. Combining the two leads to the main contribution of this work, which we call the filtered scheme, due to the use of a filter function that allows us to appropriately combine the two discretizations.

Monotone operator [25, 26]. We discretize the domain Ω by a shape regular and quasi-uniform mesh $\mathcal{T}_h^1$ with spacing h, the fine scale, and construct a space $\mathbb{V}_h^1$ of continuous piecewise linear functions over $\mathcal{T}_h^1$. The superscript 1 of $\mathbb{V}_h^1$ indicates the use of linear polynomials whereas that of $\mathcal{T}_h^1$ entails the use of straight (affine equivalent) simplices. We denote by Ω_h the computational domain, namely the union of the elements. We also denote by $\mathcal{N}_h$ the nodes of $\mathcal{T}_h$, and by

$$\mathcal{N}_h^b := \{x_i \in \mathcal{N}_h : x_i \in \partial\Omega_h\}, \quad \mathcal{N}_h^0 := \mathcal{N}_h \setminus \mathcal{N}_h^b$$

the boundary and interior nodes, respectively. We require that $\mathcal{N}_h^b \subset \partial\Omega$, which in view of the convexity of Ω implies that Ω_h is also convex and $\Omega_h \subset \Omega$. The second and coarser scale, which from now on we call δ_m whenever we refer to the monotone operator, is the length of directions we use to approximate second directional derivatives by central second order differences:

$$\nabla^2_{\delta_m} w(x;v) := \frac{w(x+\delta_m v) - 2w(x) + w(x-\delta_m v)}{\delta_m^2} \quad \text{and} \quad |v| = 1 \tag{1.3}$$

for any $w \in C^0(\overline{\Omega})$. Let $\varepsilon = (h, \delta_m, \theta_m)$ represent the two scales and a third parameter θ_m that is utilized to discretize $\mathbb{S}^\perp$ with precision θ_m. We ask that for any v in the unit sphere $\mathbb{S}$, there exists v^{θ_m} that belongs in our discrete approximate set $\mathbb{S}_{\theta_m}$ such that

$$|v - v^{\theta_m}| \leq \theta_m.$$

Likewise, we define the finite set $\mathbb{S}^\perp_{\theta_m}$: for any $\mathbf{v}^{\theta_m} = (v_j^{\theta_m})_{j=1}^d \in \mathbb{S}^\perp_{\theta_m}$, $v_j^{\theta_m} \in \mathbb{S}_{\theta_m}$ and there exists $\mathbf{v} = (v_j)_{j=1}^d \in \mathbb{S}^\perp$ such that $|v_j - v_j^{\theta_m}| \leq \theta_m$ for all $1 \leq j \leq d$ and conversely. We can now define the discrete monotone operator to be

$$T_{\varepsilon,m}[w](x_i) := \min_{\mathbf{v} \in \mathbb{S}^\perp_{\theta_m}} \left(\prod_{j=1}^d \nabla^{2,+}_{\delta_m} w(x_i; v_j) - \sum_{j=1}^d \nabla^{2,-}_{\delta_m} w(x_i; v_j) \right), \tag{1.4}$$

where $\nabla^{2,+}_{\delta_m}$ and $\nabla^{2,-}_{\delta_m}$ denote the positive and negative parts of $\nabla^2_{\delta_m}$, respectively, and $x_i \in \mathcal{N}_h^0$. The discrete solution $u_\varepsilon \in \mathbb{V}_h^1$ satisfies

$$T_{\varepsilon,m}[u_\varepsilon](x_i) = f(x_i) \quad \forall x_i \in \mathcal{N}_h^0, \quad u_\varepsilon(x_i) = g(x_i) \quad \forall x_i \in \mathcal{N}_h^b.$$

In [25] we prove that this discretization of (1.2) is monotone and consistent and that u_ε converges uniformly in Ω to the unique viscosity solution of (1.1). In [26] we derive rates of convergence in $L^\infty(\Omega_h)$ for classical solutions in Hölder and Sobolev spaces and $f > 0$ as well as for some special cases of viscosity solutions and $f \geq 0$. Our numerical experiments of [25] indicate linear convergence rates, which is rigorously proven in [23] for classical solutions. Therefore, a linear rate is an accuracy barrier for two-scale monotone schemes with piecewise linear elements. A viable way to reduce this error is to increase the polynomial degree, which, given the two-scale nature of our scheme, would require higher order approximation of second directional derivatives. This pair of corrections leads us to introduce what we call the accurate operator.

Accurate operator. This time we use quadratic polynomials in order to achieve a better interpolation error and a more accurate discretization of second directional derivatives in order to decrease the truncation error of the operator. To this end, we introduce again two scales h and δ_a, where $\delta_a \neq \delta_m$ is the coarse scale corresponding to the length of directions used for accurate discretization of second derivatives. We define the space $\mathbb{V}_h^2$ of continuous, piecewise quadratic functions and, in order to maximize the effect of polynomial degree, we employ isoparametric finite elements [7, 9, 33]. We assume that our domain Ω is piecewise uniformly convex and piecewise $C^{1,1}$, so that we can guarantee the existence of invertible and quadratic maps that transform the master element into elements with curved sides connecting boundary nodes $\mathcal{N}_h^b$ [9]. We call the resulting mesh $\mathcal{T}_h^2$, the superscript indicating quadratic isoparametric mappings for boundary elements. We also employ a more accurate approximation of the second directional derivatives that relies on five, rather than three, point stencils. Consequently, second differences for $u_\varepsilon \in \mathbb{V}_h^2$ are now given by

$$\nabla^2_{\delta_a} u_\varepsilon(x_i; v) := \frac{-u_\varepsilon(x_i + \delta_a v) + 16u_\varepsilon(x_i + \frac{\delta_a}{2} v) - 30u_\varepsilon(x_i) + 16u_\varepsilon(x_i - \frac{\delta_a}{2} v) - u_\varepsilon(x_i - \delta_a v)}{3\delta_a^2}, \tag{1.5}$$

where $x_i \in \mathcal{N}_h^0$ and $v \in \mathbb{S}_{\theta_a}$. The symbol $\mathbb{S}_{\theta_a}$ indicates that we use a different angle discretization parameter θ_a for the accurate operator. The accurate scheme then becomes the following: We seek $u_\varepsilon \in \mathbb{V}_h^2$ such that $u_\varepsilon(x_i) = g(x_i)$ for $x_i \in \mathcal{N}_h^b$ and for

$x_i \in \mathcal{N}_h^0$

$$T_{\varepsilon,a}[u_\varepsilon](x_i) := \min_{\mathbf{v}\in\mathbb{S}^{\perp}_{\theta_a}} \left(\prod_{j=1}^{d} \nabla^{2,+}_{\delta_a} u_\varepsilon(x_i; v_j) - \sum_{j=1}^{d} \nabla^{2,-}_{\delta_a} u_\varepsilon(x_i; v_j) \right) = f(x_i). \tag{1.6}$$

We observe that this discretization is no longer monotone, since a change in u_ε at a certain x_j could affect the second difference at another node x_i in two possible ways. It can either decrease or increase it, depending on whether it affects the behavior of $-u_\varepsilon(x_i \pm \delta_a v)$ or $u_\varepsilon(x_i \pm \frac{\delta_a}{2} v)$, respectively. We also note that $T_{\varepsilon,a}$ is defined on a space of piecewise quadratic functions, which means that the behavior at nodes does not translate monotonically to the behavior inside simplices. Since the only convergence technique in the literature that we are aware of for such nonlinear schemes is that of Barles and Souganidis [2], which heavily relies on monotonicity, we believe that the lack of this property would make convergence to viscosity solutions of (1.1) either wrong or very hard to prove. This is the motivation behind the use of a filtered scheme, along the lines of [18].

Filtered scheme. The idea is to make use of a filter function that combines the accurate and the monotone operators and guarantees that the monotone operator will be active if the accurate operator fails, due to lack of monotonicity. This allows for a notion of an "almost monotone" operator that is flexible enough to deliver better accuracy for each fixed mesh size. We introduce the scheme here briefly and expand on its theoretical properties later.

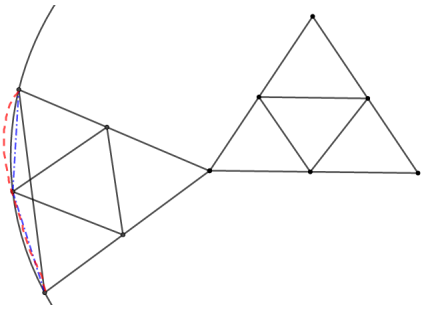

FIG. 1. *Illustration of the refinement close to and away from the boundary for $d = 2$. In the interior we create four new triangles upon each refinement, by connecting the midpoints of the coarser cell. The space $\mathbb{V}_h^1$ corresponding to $\mathcal{T}_h^1$ is defined over these four triangles and shares the same nodal values as the space $\mathbb{V}_{2h}^2$ corresponding to $\mathcal{T}_{2h}^2$ and defined over the original triangle. On boundary elements, the boundary point that is used in the construction and as a nodal value of the isoparametric element in $\mathcal{T}_{2h}^2$ for $\mathbb{V}_{2h}^2$ becomes the new boundary node for $\Omega_h^1 \supset \Omega_{2h}^1$ that is used for the four new triangles of $\mathcal{T}_h^1$. The blue "– · – · –" dashed line corresponds to the new edges introduced after the refinement. The red "– – –" dashed line illustrates the curved edge of the curved element that is isoparametric to the original triangle. (Color available online.)*

We start with the two meshes and function spaces used. Let $\mathcal{T}_h^1$ be a shape regular and quasi-uniform mesh of size h and $\mathbb{V}_h^1$ be the corresponding space of continuous piecewise linear elements. Let $\mathcal{T}_{2h}^2$ be an isoparametric mesh of size $2h$ with the same nodes as $\mathcal{T}_h^1$ and $\mathbb{V}_{2h}^2$ be the corresponding space of continuous isoparametric piecewise quadratic elements; see Figure 1 for $d = 2$. An important consequence of this two-grid approach is that functions in $\mathbb{V}_h^1$ and $\mathbb{V}_h^2$ have degrees of freedom at the same nodes, including midpoints on the curvilinear boundary of Ω_{2h}^2.

We now exploit this structure as follows. Let $\mathbf{U}_\varepsilon = \{U_\varepsilon^i\}_i \in \mathbb{R}^N$ be a grid function

where N is the number of nodes of either $\mathbb{V}_h^1$ or $\mathbb{V}_{2h}^2$. We define two functions $u_\varepsilon^1 \in \mathbb{V}_h^1$ and $u_\varepsilon^2 \in \mathbb{V}_{2h}^2$ with nodal values dictated by $\mathbf{U}_\varepsilon$,

$$u_\varepsilon^1(x_i) = u_\varepsilon^2(x_i) = U_\varepsilon^i \quad \forall\, x_i \in \mathcal{N}_h,$$

and compare them via a filter function F; see section 2 for an explicit definition. We thus seek $\mathbf{U}_\varepsilon \in \mathbb{R}^N$ such that $U_\varepsilon^i = g(x_i)$ for all $x_i \in \mathcal{N}_h^b$ and for all $x_i \in \mathcal{N}_h^0$

$$T_{\varepsilon,f}[\mathbf{U}_\varepsilon](x_i) := T_{\varepsilon,m}[u_\varepsilon^1](x_i) + \tau F\left(\frac{T_{\varepsilon,a}[u_\varepsilon^2](x_i) - T_{\varepsilon,m}[u_\varepsilon^1](x_i)}{\tau}\right) = f(x_i). \tag{1.7}$$

The filter function F is required to be compactly supported and continuous (hence uniformly bounded), as well as equal to the identity close to the origin. Therefore, the difference of operators $T_{\varepsilon,a}[u_\varepsilon^2](x_i) - T_{\varepsilon,m}[u_\varepsilon^1](x_i)$ relative to the filter scale $\tau = \tau(\varepsilon)$ is a decisive factor for the performance of the scheme. We observe that $\tau(\varepsilon)$ depends on the scales $\varepsilon = (h, \delta_a, \delta_m, \theta_a, \theta_m)$ of the accurate and monotone operators and, since they in turn depend ultimately on h, it is important to realize that $\tau \to 0$ as $h \to 0$. We later provide some insight, based on heuristics and experimental evidence, on how to choose τ. We now emphasize here the two main properties of F that we wish to exploit:

- *Minimize the risk:* The accurate operator (1.6) exhibits a smaller consistency error than the monotone operator (1.4) (see Lemma 5.5 (consistency of $T_{\varepsilon,m}[\mathcal{I}_h^1 u]$) and Lemma 5.8 (consistency of $T_{\varepsilon,a}[\mathcal{I}_{2h}^2 u]$)), but the improved accuracy comes at the cost of lack of monotonicity. Therefore, given the importance of monotonicity for convergence to viscosity solutions, the solution of (1.6) cannot be guaranteed to converge. This is especially relevant when the right-hand side f degenerates and a classical solution of (1.1) might not exist. Since lack of monotonicity could in principle lead to the failure of the accurate operator, the filter F examines the quantity $T_{\varepsilon,a}[u_\varepsilon^2](x_i) - T_{\varepsilon,m}[u_\varepsilon^1](x_i)$ relative to the filter scale τ. If this diference is smaller than τ, thus signaling that $T_{\varepsilon,a}[u_\varepsilon^2](x_i)$ is well behaved, then F is the identity and $T_{\varepsilon,f}[\mathbf{U}_\varepsilon](x_i) = T_{\varepsilon,a}[u_\varepsilon^2](x_i)$ as desired. If instead this difference is larger than τ, thereby indicating erratic behavior of $T_{\varepsilon,a}[u_\varepsilon^2](x_i)$, then F vanishes and $T_{\varepsilon,f}[\mathbf{U}_\varepsilon](x_i) = T_{\varepsilon,m}[u_\varepsilon^1](x_i)$ signifies that the monotone operator dominates and yields convergence. We note, however, that this is a rather simplistic approach that could lead to using the monotone operator even in cases where the accurate operator is much better. We explore and discuss this further in section 4.
- *Almost monotonicity:* A key feature of F is uniform boundedness, namely $|F(s)| \leq 1$ for all $s \in \mathbb{R}$. This leads to Lemma 6.1 (almost monotonicity), which, in turn, is critical to prove existence of a solution of (1.7) and its convergence to the viscosity solution of (1.1). Proving such results is an essential component of this work, which entails suitable definitions of the filter F and of the filter operator $T_{\varepsilon,f}$ in (1.7). We discuss this in section 2, including the possible degeneracy of the right-hand side f. We provide explicit definitions of F.

In order to explore the performance of the accurate operator and the filtered scheme, a substantial part of our presentation is devoted to numerical experiments. We first verify computationally the increased accuracy of the higher order operator and provide some computational remarks. We then repeat our experiments for the filtered scheme and obtain the anticipated results: the scheme has a smaller error compared to the monotone operator and appears to detect singularities when the solution is not smooth. In fact, we observe that in the case of singularities the filtered scheme performs even better than the accurate operator. We explore this behavior

and examine the interplay between the monotone and the accurate operator in the nonsmooth case in order to elucidate the role of filter F. Since this is a significant component of this work, we present the numerical examples first in sections 3 and 4. We conclude with a discussion of consistency of the monotone and accurate operators in section 5, as well as proofs of existence and convergence of solutions to the filtered scheme in section 6.

2. Definition of filter function. We start with explicit definitions for the filter function F, including potential degeneracy of the right-hand side f. We recall that $\delta_m \neq \delta_a$ are the coarse scales in the definition of the operators $T_{\varepsilon,m}$ and $T_{\varepsilon,a}$. We introduce the simplifying notation

$$A[\mathbf{U}_\varepsilon](x_i) := T_{\varepsilon,a}[u_\varepsilon^2](x_i) - T_{\varepsilon,m}[u_\varepsilon^1](x_i) \tag{2.1}$$

for the argument of F in (1.7). We choose the following continuous and uniformly bounded filter function:

$$F_\sigma(s) := \begin{cases} s, & |s| \le 1, \\ 0, & |s| \ge 1+\sigma, \\ -\frac{1}{\sigma}\, s + \frac{1+\sigma}{\sigma}, & 1 < s < 1+\sigma, \\ -\frac{1}{\sigma}\, s - \frac{1+\sigma}{\sigma}, & -1-\sigma < s < -1. \end{cases} \tag{2.2}$$

The parameter σ encodes a smooth transition of F_σ to zero; its choice and use are further discussed in section 4. Function F_σ satisfies the desirable properties mentioned in section 1; i.e., it is uniformly bounded by one and coincides with the identity in the interval $[-1, 1]$.

However, one important feature of $T_{\varepsilon,m}$ reported in [25] is that it can guarantee the discrete convexity of the discrete solution to $T_{\varepsilon,m}[u_\varepsilon^1](x_i) = f(x_i)$, i.e.,

$$\nabla^2_{\delta_m} u_\varepsilon^1(x_i; v_j) \ge 0 \quad \forall\, x_i \in \mathcal{N}_h^0,\ v_j \in \mathbb{S}_{\theta_m},$$

provided $f \ge 0$. This may not be the case with (2.2) when the right-hand side f touches zero. Although it is possible to deal with this issue asymptotically, it is also desirable to mimic the properties of the continuous problem, which justifies preserving the discrete convexity of discrete solutions.

We explain now why (2.2) may not guarantee discrete convexity and suggest a way to enforce it. We observe that for every grid function $\mathbf{U}_h$ with corresponding functions $u_h^1 \in \mathbb{V}_h^1$ and $u_h^2 \in \mathbb{V}_{2h}^2$ with nodal values dictated by $\mathbf{U}_h$, and for all $x_i \in \mathcal{N}_h^0$, there exists $|k_i| \le 1$ depending on $T_{\varepsilon,a}[u_h^2](x_i)$ and $T_{\varepsilon,m}[u_h^1](x_i)$ such that

$$T_{\varepsilon,f}[\mathbf{U}_\varepsilon](x_i) = \begin{cases} T_{\varepsilon,m}[u_h^1](x_i), & |A[\mathbf{U}_h](x_i)| \ge (1+\sigma)\tau, \\ T_{\varepsilon,m}[u_h^1](x_i) - |k_i|\tau, & -(1+\sigma)\tau < A[\mathbf{U}_h](x_i) \le 0, \\ T_{\varepsilon,m}[u_h^1](x_i) + |k_i|\tau, & 0 < A[\mathbf{U}_h](x_i) < (1+\sigma)\tau. \end{cases} \tag{2.3}$$

Suppose that $f(x_i) = 0$ for some $x_i \in \mathcal{N}_h^0$. If we want u_ε^1 to be discretely convex at x_i, we need to make sure that $T_{\varepsilon,m}[u_\varepsilon^1](x_i) \ge 0$. This property is also instrumental in Lemma 6.2 (existence and stability) to prove existence of a discrete solution using results from [25]. The above calculation shows that

$$T_{\varepsilon,m}[u_\varepsilon^1](x_i) = \begin{cases} f(x_i) = 0, & |A[\mathbf{U}_\varepsilon](x_i)| \ge (1+\sigma)\tau, \\ f(x_i) + |k_i|\tau \ge 0, & -(1+\sigma)\tau < A[\mathbf{U}_\varepsilon](x_i) \le 0, \\ f(x_i) - |k_i|\tau \le 0, & 0 < A[\mathbf{U}_\varepsilon](x_i)(x_i) < (1+\sigma)\tau. \end{cases} \tag{2.4}$$

This reveals that in order to preserve $T_{\varepsilon,m}[u_\varepsilon^1](x_i) \geq 0$ for all $x_i \in \mathcal{N}_h^0$, we must exclude the third case. We thus introduce a nonsymmetric modification of the filter function:

$$\widetilde{F}_\sigma(s) := \begin{cases} s, & -1 \leq s \leq 0, \\ -\frac{1}{\sigma}\, s - \frac{1+\sigma}{\sigma}, & -1-\sigma \leq s \leq -1, \\ 0 & \text{otherwise.} \end{cases} \tag{2.5}$$

From now on we make the convention that (2.2) is used in (1.7) whenever the right-hand side $f(x) \geq f_0 > 0$ for all $x \in \Omega$, whereas (2.5) is our choice provided f touches zero. We emphasize that this decision depends on f but not on the space location x. In both cases, we have

$$T_{\varepsilon,m}[u_\varepsilon^1](x_i) \geq f(x_i) - |k_i|\tau \geq 0, \tag{2.6}$$

provided $\tau \leq f_0$ in the first case and by construction in the degenerate case. Consequently, u_ε^1 is discretely convex.

In Lemma 6.2 (existence and stability) we show the existence of a discretely convex solution of (1.7). The restriction $\tau \leq f_0$ is not stringent in practice because τ is some positive power of h and thus tends to zero. On the other hand, choosing the nonsymmetric filter $\widetilde{F}_\sigma$ destroys the symmetry of the resulting system and excludes parts of the domain where $T_{\varepsilon,a}$ could have still been used, i.e., whenever $f \geq \tau$. Consequently, we only employ (2.5) if necessary. In section 4 we test both (2.2) and (2.5) on a smooth example with strictly positive f and a $C^{1,1}$ example with vanishing f, respectively. We also explore briefly a space-dependent choice of filter for the degenerate case.

3. Numerical experiments: Accurate scheme. In this section we illustrate the improved performance of the accurate operator. Its implementation follows closely that of the monotone operator $T_{\varepsilon,m}$ in [25]. We refer the reader to [25], as well as [5, 10, 21], for details about the implementation of the method.

3.1. Comparison between $T_{\varepsilon,m}$ and $T_{\varepsilon,a}$. We present in Table 1 the L^∞ error of the solution of (1.4) versus that of (1.6) for the following two examples taken from [25] and defined on $\Omega = [0,1]^2$.

Smooth Hessian. Let the exact solution u and forcing f be

$$u(x) = e^{|x|^2/2}, \quad f(x) = (1+|x|^2)e^{|x|^2} \quad \forall x \in \Omega.$$

Discontinuous Hessian. Let $x_0 = (0.5, 0.5)$ and u and f be

$$u(x) = \frac{1}{2}\left(\max(|x-x_0| - 0.2, 0)\right)^2, \quad f(x) = \max\left(1 - \frac{0.2}{|x-x_0|}, 0\right) \quad \forall x \in \Omega.$$

Since Ω is polygonal, the computational domain $\Omega_h = \Omega$ and the isoparametric maps of $\mathcal{T}_{2h}^2$ for boundary elements are simply affine. This choice simplifies the numerics and allows us to compare with earlier experiments from [25]. We indeed compare $\|u - u_\varepsilon^1\|_{L^\infty(\Omega)}$ and $\|u - u_\varepsilon^2\|_{L^\infty(\Omega)}$ where $u_\varepsilon^1 \in \mathbb{V}_h^1$ solves (1.4) and $u_\varepsilon^2 \in \mathbb{V}_{2h}^2$ solves (1.6), whence the number of degrees of freedom is the same in both examples.

For the smooth case, we observe that the accurate operator exhibits a significant improvement beyond one order of magnitude. Although there is no theoretical result to support this fact, it can be formally explained by the regularity of the solution and

the higher order operator consistency error in Lemma 5.8 (consistency of $T_{\varepsilon,a}[\mathcal{I}^2_{2h}u]$). It is worth noting that the accuracy improvement for the monotone operator for the smooth example exhibits saturation in the last refinement. Upon examining where the error is largest, we realize that it appears on the boundary layer that arises from the definition of $T_{\varepsilon,m}$. As shown in [26, Theorem 5.3], this error does not obey operator consistency, but is instead bounded by $C\ \|u\|_{W^2_\infty(\Omega)}\ \delta_m$ through a barrier argument. This effect manifests in our experiments, where for $h = 2^{-9}$ the above upper bound is already of order 10^{-2}.

TABLE 1

L^∞ error for monotone and accurate operators, $T_{\varepsilon,m}$ and $T_{\varepsilon,a}$, for the smooth Hessian (top) and the discontinuous Hessian (bottom). P: Number of points $x_i \pm \delta_a v_j, x_i \pm \frac{\delta_a}{2} v_j$ used for $T_{\varepsilon,a}$ at each $x_i \in \mathcal{N}_h^0$; $P = 8(D-1)$, where D is the number of directions v_j in a quarter circle, as determined by the value of θ_a. The operator $T_{\varepsilon,a}$ has an accuracy of one to two orders higher than $T_{\varepsilon,m}$ [25]. The number of Newton iterations corresponds to $T_{\varepsilon,a}$.

DoFs	P: # of points	$T_{\varepsilon,m}$	$T_{\varepsilon,a}$	Newton steps
N = 4225, $h = 2^{-6}$	56	$2.8\ 10^{-3}$	$1.91\ 10^{-4}$	6
N = 16641, $h = 2^{-7}$	88	$1.5\ 10^{-3}$	$8.60\ 10^{-5}$	5
N = 66049, $h = 2^{-8}$	144	$7.8\ 10^{-4}$	$4.17\ 10^{-5}$	5
N = 263169, $h = 2^{-9}$	224	$6.4\ 10^{-4}$	$2.42\ 10^{-5}$	7
DoFs	P: # of points	$T_{\varepsilon,m}$	$T_{\varepsilon,a}$	Newton steps
N = 4225, $h = 2^{-6}$	40	$1.9\ 10^{-3}$	$2.48\ 10^{-4}$	9
N = 16641, $h = 2^{-7}$	56	$9.0\ 10^{-4}$	$1.51\ 10^{-4}$	12
N = 66049, $h = 2^{-8}$	72	$5.7\ 10^{-4}$	$8.34\ 10^{-5}$	14
N = 263169, $h = 2^{-9}$	96	$3.83\ 10^{-4}$	$5.31\ 10^{-5}$	20

For the example with discontinuous Hessian we observe again an accuracy improvement from $T_{\varepsilon,m}$ to $T_{\varepsilon,a}$, despite the fact that the predicted error in [26, Theorem 5.7] for $T_{\varepsilon,m}$ and a degenerate f is determined by the dimension of the problem rather than the regularity of the solution. We also notice, in contrast to [25], an increase in the number of Newton iterations with each refinement for $T_{\varepsilon,a}$. This may be attributed to the lack of monotonicity of $T_{\varepsilon,a}$.

3.2. Computational remarks. We now explain implementation issues for the accurate method, which are in turn relevant for the filtered scheme.

Sparsity. The evaluation of $\nabla^2_{\delta_a} u^2_\varepsilon(x_i; v_j)$ using (1.5) requires about twice the number of points as $\nabla^2_{\delta_m} u^1_\varepsilon(x_i; v_j)$ using (1.3), for each point $x_i \in \mathcal{N}_h^0$ and direction $v_j \in \mathbb{S}_\theta$. This results in a sparsity pattern with a wider bandwidth and more nonzero elements, since for each extra point $x_i \pm \frac{\delta}{2} v_j$, we need to use the degrees of freedom of the simplex where $x_i \pm \frac{\delta}{2} v_j$ belongs.

Solver. The monotone operator $T_{\varepsilon,m}$ in [25] was implemented using a direct solver for the linear system, i.e., the MATLAB backslash operator. However, for very fine meshes which yield many directions an iterative method like conjugate gradient may be more appropriate. The situation is more critical for the accurate operator $T_{\varepsilon,a}$ due to its worse sparsity pattern discussed above. This makes a direct solver a less favorable option for very fine meshes. Choosing a small tolerance for the conjugate gradient method seems to provide computational results similar to the direct solver, thus without sacrificing accuracy but gaining efficiency. The design of suitable preconditioners is essential, but remains an open issue.

4. Numerical experiments: Filtered scheme. We now explore computationally the accuracy of the filtered scheme in (1.7), which combines the monotone

and accurate operators. In fact, we determine the *active set* of the filtered scheme, which is the region where the monotone operator dominates. We start with a brief discussion about the choice of σ in F_σ and fix a value for our implementation.

4.1. Choice of $\boldsymbol{\sigma}$ in $\boldsymbol{F_\sigma}$. We observe that the function F_σ in (2.2) is Lipschitz for any $\sigma > 0$ whereas for $\sigma = 0$ it is the discontinuous function

$$F_0(s) := \begin{cases} s, & |s| \le 1, \\ 0, & |s| > 1. \end{cases}$$

Since continuity of F_σ is only used in Lemma 6.2 (existence and stability) in order to apply [25, Lemma 3.1], we can choose σ as small as we want and then perform computations with decreasing values of h. In practice, we take $\sigma = 10^{-4}$, which leads to such a tiny window for the last two cases in (2.2) that they do not occur in practice. This is desirable because our goal is to give full control to $T_{\varepsilon,a}$ in regions of smoothness and to employ the monotone operator $T_{\varepsilon,m}$ otherwise.

The use of semismooth Newton for the range $1 \le |s| \le 1+\sigma$ is however questionable; we refer the reader to [10, 21]. For example, if $\sigma = 1$ and $-2 \le \tau^{-1}\big(T_{\varepsilon,a}[u_\varepsilon^1](x_i) - T_{\varepsilon,m}[u_\varepsilon^2](x_i)\big) \le -1$ in (2.2), we obtain

$$T_{\varepsilon,f}[\mathbf{U}_\varepsilon](x_i) = 2\ T_{\varepsilon,m}[u_\varepsilon^1](x_i) - T_{\varepsilon,a}[u_\varepsilon^2](x_i) - 2\tau.$$

This would result in the ith row $\nabla T_{\varepsilon,f}[\mathbf{U}_\varepsilon](x_i)$ of the Jacobian matrix being

$$\nabla T_{\varepsilon,f}[\mathbf{U}_\varepsilon](x_i) = 2\ \nabla T_{\varepsilon,m}[u_\varepsilon^1](x_i) - \nabla T_{\varepsilon,a}[u_\varepsilon^2](x_i),$$

where $\nabla T_{\varepsilon,m}[u_\varepsilon^1]$ and $\nabla T_{\varepsilon,a}[u_\varepsilon^2]$ are the Jacobian matrices associated with the monotone and accurate operators. Both Jacobians behave well computationally, but there is no reason to expect that a matrix resulting from subtracting them will be nonsingular. In fact, we observe computationally that the semismooth Newton becomes very slow and for very fine meshes it does not even converge. Concerns about the solvability of the Newton system are also raised in [18], where the following approximation is advocated:

$$\nabla T_{\varepsilon,f}[\mathbf{U}_\varepsilon](x_i) \approx 2\ \nabla T_{\varepsilon,m}[u_\varepsilon^1](x_i).$$

We prefer, instead, to avoid these issues altogether by choosing a rather small value of σ. For any fixed value of σ, we still employ a semismooth Newton iteration and treat all the corners of the filter F_σ similarly to the min and max functions in [25]. Moreover, using $\sigma = 10^{-4}$ the likelihood of $|T_{\varepsilon,a}[u_\varepsilon^2](x_i) - T_{\varepsilon,m}[u_\varepsilon^1](x_i)| \in [1, 1+\sigma]$ is rather small and indeed it rarely occurs in practice. We see in Tables 2 and 3 in subsection 4.2 that our choice leads computationally to a number of Newton iterations similar to that of the accurate scheme in Table 1.

4.2. Numerical experiments. Before presenting our results in detail, using the same examples as in section 3, we make two general observations.

- **Choice of filter scale $\boldsymbol{\tau}$.** We stress that the behavior of the scheme depends strongly on the choice of the filtered scale τ, which must obey $\tau \to 0$ as $h \to 0$. A bigger τ allows the accurate operator to take control, while the monotone operator guarantees convergence when the accurate operator has a very large consistency error. On the other hand, smaller values of τ lead to the presence of the active set of nodes, where the monotone operator dominates. Since we measure the error in the L^∞-norm, a large active set could prevent us from achieving better accuracy

than in [25]. Although there is no obvious recipe for choosing τ, the definition (1.7) of $T_{\varepsilon,f}$ indicates that, in order for the accurate operator to be active at a point x_i, we need $|T_{\varepsilon,a}[u_\varepsilon^2](x_i) - T_{\varepsilon,m}[u_\varepsilon^1](x_i)| \le \tau$. Consequently, τ has to be greater than the truncation error of the monotone operator, because the active set may otherwise include nodes where the solution is smooth. We follow this approach to generate Tables 2 and 3 and Figure 5.

- **Boundary layer.** Our experiments reveal that the active set may contain nodes near $\partial\Omega$. This is due to the different boundary layer effect of each operator. In fact, the consistency error for the monotone operator is of order one because $\frac{\delta_m^2}{h^2} \approx 1$ (see Lemma 5.5 (consistency of $T_{\varepsilon,m}[\mathcal{I}_h^1 u]$)), while for the accurate operator the order becomes $\frac{\delta_a^2}{h^2} \approx \frac{h^3}{h^2} \approx h$ (see Lemma 5.8 (consistency of $T_{\varepsilon,a}[\mathcal{I}_{2h}^2 u]$)).

We now document the performance of the filtered scheme for the two examples of section 3, upon comparing the ℓ^∞-norm of the difference between $\mathbf{U}_\varepsilon$ and u at the nodes, and investigate the effect of τ in the size and location of the active set. We also compare the performance of the scheme with that of the monotone and accurate operators and recall that $\Omega_h = \Omega$ for all $h > 0$.

TABLE 2

Smooth Hessian, $\tau = 6e^2 h$. The L^∞ error for $T_{\varepsilon,f}$ is up to an order of magnitude smaller than $T_{\varepsilon,m}$. Active set is the number of nodes where the low-accuracy operator $T_{\varepsilon,m}$ dominates. The number of Newton iterations for the filtered scheme is also displayed.

h	$T_{\varepsilon,m}$	$T_{\varepsilon,a}$	$T_{\varepsilon,f}$	$T_{\varepsilon,f}$ Newton	Active set
$h = 2^{-5}$	$5.4\ 10^{-3}$	$5.16\ 10^{-4}$	$1.01\ 10^{-3}$	6	17
$h = 2^{-6}$	$2.8\ 10^{-3}$	$1.91\ 10^{-4}$	$3.16\ 10^{-4}$	6	57
$h = 2^{-7}$	$1.5\ 10^{-3}$	$8.60\ 10^{-5}$	$1.20\ 10^{-4}$	6	214
$h = 2^{-8}$	$7.8\ 10^{-4}$	$4.17\ 10^{-5}$	$5.00\ 10^{-5}$	7	918
$h = 2^{-9}$	$6.4\ 10^{-4}$	$2.42\ 10^{-5}$	$1.99\ 10^{-5}$	8	5035

Experiment 1: Smooth Hessian. We start by illustrating the error estimates for the smooth example for $\tau = 6e^2 h$ in Table 2. This choice is motivated by the theoretical truncation error of $T_{\varepsilon,m}$ for the corresponding choice of δ_m. For this example we use the original, symmetric, filter F_σ with $\sigma = 10^{-4}$. This falls under the existence and convergence results of section 6, since $f(x) = (1 + |x|^2)e^{|x|^2} \ge e > \tau$ in Ω for all values of h that are used in Table 2. We observe a small active set, with relative size around 1%–2% for all refinements.

The active set for $\tau = 6e^2 h$ and $h = 2^{-8}$ is displayed in Figure 2. We observe the aforementioned boundary layer effect, especially close to the upper-right corner, where u and its derivatives are larger.

Experiment 2: Discontinuous Hessian. For the $C^{1,1}$ example of subsection 3.1 we choose $\tau = 0.62\ h^{2/5}$, which is motivated by the theoretical consistency error of the monotone operator for $\delta_m = 2.44\ h$. We use the nonsymmetric filter $\widetilde{F}_\sigma$ with $\sigma = 10^{-4}$ in (2.5). We observe that the active set is located on the circle of discontinuity of the Hessian and near the boundary. We also see that the error in the L^∞-norm of $T_{\varepsilon,f}$ is slightly better than in Table 1. We observe computationally that coarser choices of τ, as for example $\tau = O(h^{1/5})$, lead to an empty active set. In contrast to the smooth example, we now notice a gradual increase of the relative size of the active set. This is more prominent in the last three refinements, where this relative size increases from around 1.5% to 2.9% and then to 4.1%.

In order to explain the smaller errors of Table 3 with respect to Table 1, we present in Figure 3 a contour plot with $|u_\varepsilon - u|$ for the accurate scheme and meshsize

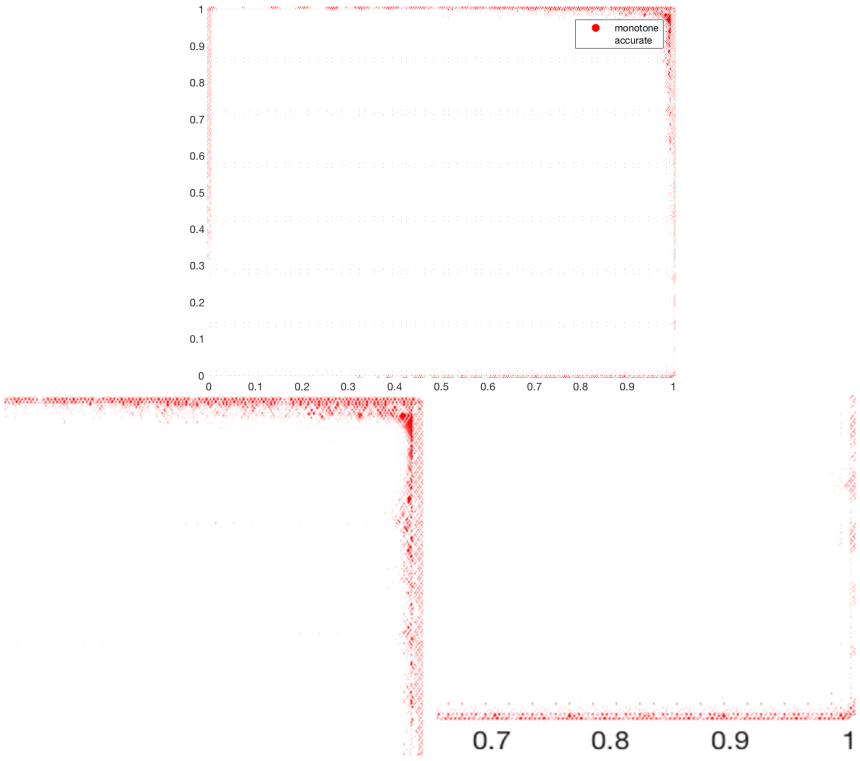

FIG. 2. *Active set for smooth example:* $\tau = 6e^2h$. *The active set is concentrated on the upper corner and, less, on the lower and left side. To enhance visualization of this boundary layer effect, the bottom pictures display zooms of the upper-right and lower-right corners within a square of size about* 0.35.

TABLE 3

Discontinuous Hessian, $\tau = 0.62\ h^{2/5}$. *The* L^∞ *error for* $T_{\varepsilon,f}$ *is about one order of magnitude better than that of* $T_{\varepsilon,m}$ *and slightly better than of* $T_{\varepsilon,a}$. *The latter is explained in Figure* 3.

h	$T_{\varepsilon,m}$	$T_{\varepsilon,a}$	$T_{\varepsilon,f}$	$T_{\varepsilon,f}$ Newton	Active set
$h = 2^{-5}$	$4.0\ 10^{-3}$	$5.67\ 10^{-4}$	$5.50\ 10^{-4}$	5	22
$h = 2^{-6}$	$1.9\ 10^{-3}$	$2.48\ 10^{-4}$	$2.48\ 10^{-4}$	8	8
$h = 2^{-7}$	$9.0\ 10^{-4}$	$1.51\ 10^{-4}$	$1.40\ 10^{-4}$	12	248
$h = 2^{-8}$	$5.7\ 10^{-4}$	$8.34\ 10^{-5}$	$7.58\ 10^{-5}$	14	1904
$h = 2^{-9}$	$3.83\ 10^{-4}$	$5.31\ 10^{-5}$	$4.89\ 10^{-5}$	16	10825

$2h = 2^{-6}$. In the same figure we depict the active set associated with the filtered scheme. Both illustrations correspond to the last Newton iteration. We observe that the circle of discontinuity of the Hessian dominates the active set and is precisely the set of nodes where the accurate operator exhibits the biggest error. This explains why using the monotone operator at these points increases the accuracy and provides experimental justification of the filtered scheme.

We now explore further the behavior of the active set. We illustrate it in Figure 4 (left) for the final iteration that corresponds to $h = 2^{-7}$. We observe that it includes the layer δ_m, away from the circle of discontinuity, the center of the domain, where $f = 0$ and the problem degenerates, and a small boundary layer at the corners of the

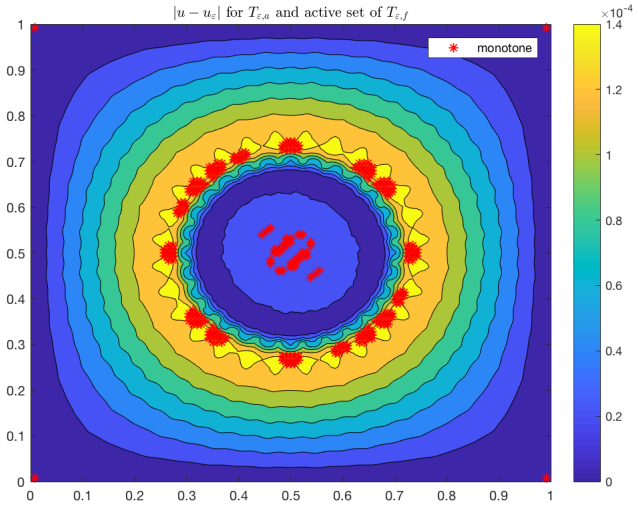

FIG. 3. *Error $|u_\varepsilon - u|$ of $T_{\varepsilon,a}$ for meshsize $2h = 2^{-6}$ and active set for filtered scheme (in red). Most nodes on the circle of discontinuity of the Hessian, as well as the corners of Ω, belong to the active set and are the nodes x_i with the highest absolute error for $T_{\varepsilon,a}[u_\varepsilon^2](x_i)$. (Color available online.)*

domain.

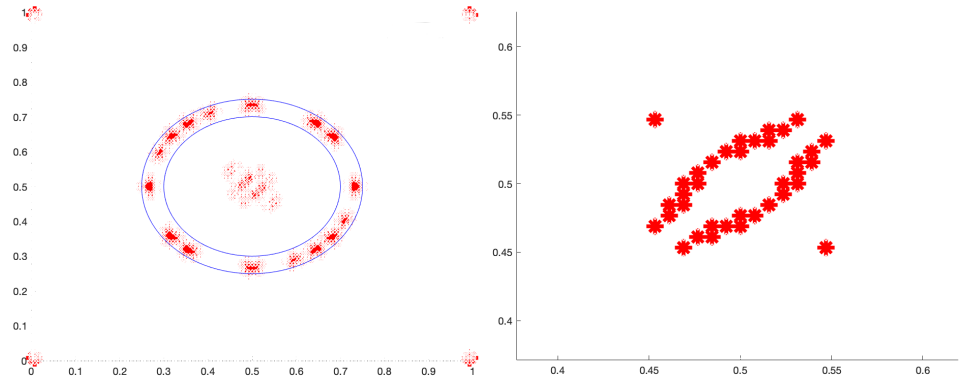

FIG. 4. (a) *Active set for discontinuous Hessian: $h = 2^{-7}$, $\tau = 0.62\ h^{2/5}$, and $\widetilde{F}_\sigma$ with $\sigma = 10^{-4}$ (left). The two circles correspond to $\{|x - x_0| = 0.2\}$ and $\{|x - x_0| = 0.2 + \delta_m\}$.* (b) *Active set for a space-dependent filter function (right). The active set is depicted in red in both plots. (Color available online.)*

That the active set reduces to a layer around the circle of discontinuity of the Hessian is a desirable property of the filtered scheme, whereas active nodes near the corners are caused by the disparate consistency errors of the operators near the boundary. In order to investigate the presence of active nodes within $\{|x - x_0| \le 0.2\}$, we experiment with a space-dependent filter function: we restrict the use of the accurate operator only for those nodes $x_i \in \mathcal{N}_h^0$ such that $f(x_i) \le \tau$. This function helps to shed some light on the behavior of the two operators. We present in Figure 4(b) a zoomed-in illustration of the active set that corresponds to the same parameters as

in Figure 4(a), but using this space-dependent filter. We observe that the active set is now only restricted to the center of the circle. By examining the values of the two operators in the active set, we realize that the monotone operator is of order 10^{-18}, while the accurate operator is of order 10^{-8}, which explains why this active set remains present. Since $f = 0$ in $\{|x - x_0| \leq 0.2\}$, we want to choose the operator whose value is closer to zero. This is why both (2.5) and its space-dependent version enforce the monotone operator whenever the accurate operator is "positive enough" in this region. For (2.5) this includes parts of the Hessian discontinuity at the center of the circle, while for the space-dependent version of the filter, this is only enforced in the center of the domain. We see that, although at first sight the definition of (2.5) may raise concerns about restricting the accurate operator too much, it actually allows us to employ the monotone operator precisely at the critical nodes. This is further portrayed in Figure 3, which illustrates that the active set due to (2.5) includes nodes of lowest accuracy of $T_{\varepsilon,a}$. This is why Figure 5 displays smaller errors for the filtered operator than for the accurate one.

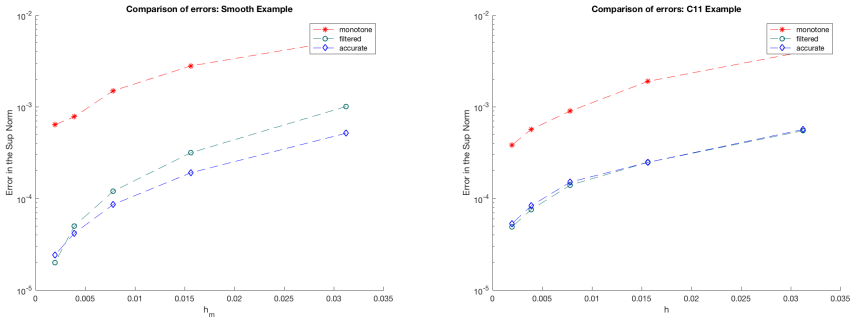

FIG. 5. *Error for monotone, accurate, and filtered operator. Left: Smooth example, for filtered scheme:* $\tau = 4e^2 h$. *Right:* $C^{1,1}$ *example, for filtered scheme:* $\tau = 0.41\ h^{2/5}$.

Conclusions. We see that in both examples the choice of the filtered scale τ is not a trivial task and it can have a dramatic effect on the outcome. Since the accurate operator performs better close to the boundary, it is expected that for many choices of τ there will be a boundary layer, where the monotone operator will be active. Unfortunately, this is due to the bad behavior of the monotone operator near the boundary and not to the filter capturing a singularity. This can be explained by the simplicity of the filter functions being used, because they compare only the values of the two operators and do not take into account the respective value of the right-hand side.

A crucial observation is that the convergence result allows for a great deal of flexibility in the choice of τ. We can always choose τ to be relatively big with respect to ε but still satisfy $\tau \to 0$ as $\varepsilon \to 0$. This results in the accurate operator being the one that is always active, hence allowing us to fully exploit the higher accuracy it offers, without sacrificing the convergence result. We do not present a table for this case, because it corresponds exactly to the results of section 3. Instead we present in Figure 5 two comparative plots for the errors due to the monotone, the accurate, and the filtered schemes, which correspond to the results from Table 2 and Table 3. We see that the filtered scheme is much more efficient than the monotone one even in the presence of boundary layers, and outperforms the accurate scheme in the nonsmooth case.

4.3. Comparison of $T_{\varepsilon,a}$ and $T_{\varepsilon,f}$ based on a singular example. We now turn our attention to an example beyond the scope of our theory. It is the third experiment illustrated in [25]; namely the exact solution u and the forcing f are given by

$$u(x) = -\sqrt{2 - |x|^2} \quad \text{and} \quad f(x) = 2(2 - |x|^2)^{-2} \quad \forall x \in \Omega = (0,1)^2.$$

We see that the right-hand side f is unbounded near the corner $(1,1)$ of Ω; hence the convergence theory for $T_{\varepsilon,m}$ in [25], as well as for $T_{\varepsilon,f}$ in subsection 6.2, does not apply due to the requirement that f must be uniformly bounded. As a result, we choose to examine this problem separately, aiming to exhibit a computational advantage of the filtered scheme over the accurate one, which is not backed by theory, yet is stronger than what we observe in subsection 4.2. Before we present our results, we note that similarly to [25] and section 3, we have to make some choices about the various parameters present in our methods. To simplify the presentation, we follow the same approach as in [25], i.e., we choose the same parameters as for the smooth example, due to the high regularity of u away from the boundary. This is not justified by the theory in [26] and the choice of parameters is in general a delicate issue that is open for further investigation and varies for every example. We also follow [25] in terms of the initial guess; i.e., we use the solution of $\Delta u_0 = (d!f)^{1/d}$, as proposed in [17]. Lastly, we choose $\tau \approx 13e^2h^{2/5}$. This is a large parameter which, we observe numerically, allows for the accurate operator to take over away from the singular corner. Note that the choice of τ is motivated in subsection 4.2 by the truncation error of $T_{\varepsilon,m}$. As a result, a large parameter for this singular example is not surprising. We are now ready to present our experiments in Table 4.

TABLE 4

Unbounded f, $\tau = 13e^2\ h^{2/5}$. The L^∞ error for $T_{\varepsilon,f}$ is slightly better than that of $T_{\varepsilon,m}$, but one order of magnitude better than $T_{\varepsilon,a}$.

h	$T_{\varepsilon,m}$	$T_{\varepsilon,a}$	$T_{\varepsilon,f}$	Newton $T_{\varepsilon,a}, T_{\varepsilon,f}$	Active set
$h = 2^{-5}$	$8.3\ 10^{-3}$	$6.4\ 10^{-2}$	$4.5\ 10^{-3}$	20 13	6
$h = 2^{-6}$	$5.0\ 10^{-3}$	$4.6\ 10^{-2}$	$2.9\ 10^{-3}$	13 36	19
$h = 2^{-7}$	$3.3\ 10^{-3}$	$3.3\ 10^{-2}$	$3.1\ 10^{-3}$	33 49	89
$h = 2^{-8}$	$2.0\ 10^{-3}$	$2.3\ 10^{-2}$	$1.9\ 10^{-3}$	75 51	336

A first interesting observation from Table 4 is that $T_{\varepsilon,a}$ has one order of magnitude larger error than both $T_{\varepsilon,m}$ and $T_{\varepsilon,f}$, showcasing the decreased performance of the accurate scheme when the regularity assumptions are severely violated. We can see that the filtered scheme manages to converge, but has only a minor improvement over the monotone operator, especially for finer meshes. This is, however, to be expected for this example, since the accurate operator performs worse than the monotone and the filtered scheme has the highest L^∞ error where the regularity is decreased, which is exactly where the active set appears. We illustrate this in Figure 6, where we present a contour plot of the absolute error between the discrete and the exact solution for the filtered scheme, along with the active set. We zoom where the error is larger to indicate this more clearly. A last observation is that the error for $T_{\varepsilon,f}$ performs slightly worse for $h = 2^{-7}$ than for $h = 2^{-6}$. This anomalous behavior might be caused by a rather large filtered scale $\tau \approx 13e^2h^{2/5}$, which is too permissive with the accurate scheme for coarse meshes; this illustrates once more how delicate the choice of τ is. Due to the large truncation error close to the corner, it is hard to predict the exact active set (which in this case is what determines the error differences); however,

we can see that the error is never bigger than the error of the monotone operator. These observations allow us to showcase a better performance of the filtered scheme over the accurate operator also for a singular example outside our theory.

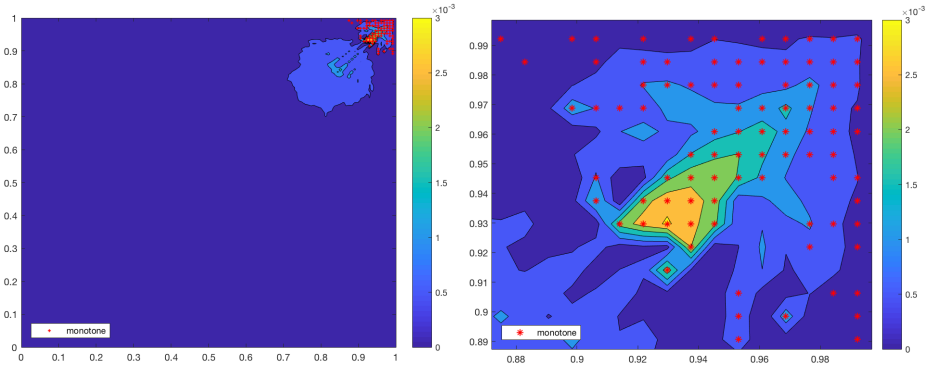

FIG. 6. *Contour plot of absolute error for filtered scheme and active set for* $h = 2^{-6}$ *(left). Zoom-in of the active set, where the error for the filtered scheme is higher (right).*

4.4. Choice of solver. The discussion in subsection 3.2 indicates that we need to take into account the sparsity of the resulting Jacobian matrix and possibly choose between a direct and an iterative solver. This choice depends on the problem at hand. For instance, for a strictly positive right-hand side $f \geq f_0 > 0$, we may choose δ_m and θ_m as well as δ_a and θ_a (with some modifications) on the basis of [26, Theorem 5.3 (rates of convergence for classical solutions)]. On the other hand, a right-hand side f that touches zero may yield a choice of scales as described in [26, Theorem 5.6 (degenerate forcing $f \geq 0$)]. Note that the ensuing constants are not accessible in either case. The first choice corresponds to a smaller angular parameter and leads to more directions, and hence to a larger bandwidth for the Jacobian matrix. If in addition f is sufficiently smooth so that the solution u is $C^2(\Omega)$ and strictly convex, then the Jacobian matrix is likely to be positive definite, which, combined with the reduced sparsity, makes it preferable to use an iterative solver such as the conjugate gradient. In contrast, a degenerate $f \geq 0$ that touches zero does not guarantee strict convexity and global regularity of u. In this case, [26, Theorem 5.6 (degenerate forcing $f \geq 0$)] suggests a coarser choice of θ_m which leads to a sparser Jacobian matrix that makes a direct solver competitive.

4.5. Implementation challenges. The above exposition shows that the filtered scheme provides the desirable combination of provable convergence and increased accuracy and adds to the family of similar approaches, such as [18]. However, this improvement is not without challenges. In particular, implementing the filtered scheme requires special care in the following three aspects, which are tightly related to our choice of accurate scheme.

- *Second differences.* Similarly to [25], for each node $x_i \in \mathcal{N}_h^0$ we need to locate the appropriate simplex to which $x_i \pm \frac{\delta}{2} v_j$ and $x_i \pm \delta v_j$ belong, in order to calculate the second differences for the monotone and accurate operator. This is a process that now needs to take place for both operators. We employ the efficient searching techniques of FELICITY to achieve this in minimal time, as in [25].
- *Initialization.* To construct the initial guess, we solve a Laplace problem on the coarsest mesh and, for each subsequent refinement, we interpolate the solution of

(1.7) on the previous mesh. This is an efficient choice from [25], suggested earlier in [18], that we preserve for both the accurate and filtered schemes. To achieve this, we need to interpolate a quadratic function on a finer mesh. We thus create the new mesh $\mathcal{T}_h^2$ at the end of each iteration and use the searching and interpolation capabilities of FELICITY to associate each node of $\mathcal{T}_h^2$ with a simplex of $\mathcal{T}_{2h}^2$ and interpolate u_ε^2 in $\mathcal{T}_h^2$.

- *Two-mesh approach.* The definition (1.7) of a filter operator utilizes a piecewise linear function $u_\varepsilon^1 \in \mathbb{V}_h^1$ and a piecewise quadratic function $u_\varepsilon^2 \in \mathbb{V}_{2h}^2$, each defined on a different mesh $\mathcal{T}_h^1$ and $\mathcal{T}_{2h}^2$, and each involving the same number, N, of degrees of freedom. Even though $\mathcal{T}_h^1$ and $\mathcal{T}_{2h}^2$ are compatible, the global numbering of nodes, and thus of degrees of freedom, is in practice often different. Consequently, in order to compare $T_{\varepsilon,m}[u_\varepsilon^1](x_i)$ and $T_{\varepsilon,a}[u_\varepsilon^2](x_i)$ at each node x_i, we need to communicate between $\mathcal{T}_h^1$ and $\mathcal{T}_{2h}^2$ and their degrees of freedom. However, exploiting the efficiency of MATLAB for vectorized quantities, creating a map between the degrees of freedom for the two meshes takes minimal time.

5. Properties of monotone and accurate operators. We now embark on the theoretical analysis of our method. To this end, we first briefly recall key properties of the monotone operator of [25] and next present the notion of consistency of both the monotone and accurate operators and compare them. These properties are important to prove the existence and convergence results for the filtered scheme. One of the critical properties of the Monge–Ampère equation is the convexity of its solution u; see [1, 32] for a discussion of convexity for piecewise polynomial functions. We mimic this property at the discrete level, using the notion of discrete convexity [25, Definition 2.1].

DEFINITION 5.1 (discrete convexity). *We say that $w_h \in \mathbb{V}_h^1$ is discretely convex if*

$$\nabla^2_{\delta_m} w_h(x_i; v_j) \geq 0 \qquad \forall x_i \in \mathcal{N}_h^0, \quad \forall v_j \in \mathbb{S}_\theta.$$

We have the following lemma for the monotone operator $T_{\varepsilon,m}$ [25, Lemma 2.2].

LEMMA 5.2 (discrete convexity of $T_{\varepsilon,m}$). *If $w_h \in \mathbb{V}_h^1$ satisfies*

$$T_{\varepsilon,m}[w_h](x_i) \geq 0 \quad \forall x_i \in \mathcal{N}_h^0, \tag{5.1}$$

then w_h is discretely convex and as a consequence

$$T_{\varepsilon,m}[w_h](x_i) = \min_{\mathbf{v} \in \mathbb{S}^\perp_{\theta_m}} \prod_{j=1}^{d} \nabla^2_{\delta_m} w_h(x_i; v_j), \tag{5.2}$$

namely

$$\nabla^{2,+}_{\delta_m} w_h(x_i; v_j) = \nabla^2_{\delta_m} w_h(x_i; v_j), \quad \nabla^{2,-}_{\delta_m} w_h(x_i; v_j) = 0 \quad \forall x_i \in \mathcal{N}_h^0, \quad \forall v_j \in \mathbb{S}_{\theta_m}.$$

Conversely, if w_h is discretely convex, then (5.1) *is valid.*

A critical feature for the convergence of $T_{\varepsilon,m}[u_\varepsilon^1]$ is monotonicity [25, Lemma 2.3].

LEMMA 5.3 (monotonicity of $T_{\varepsilon,m}$). *Let $u_h, w_h \in \mathbb{V}_h^1$ be discretely convex. If $u_h - w_h$ attains a maximum at an interior node $z \in \mathcal{N}_h^0$, then*

$$T_{\varepsilon,m}[w_h](z) \geq T_{\varepsilon,m}[u_h](z).$$

Another important property that relies on monotonicity is the following discrete comparison principle [25, Lemma 2.4].

LEMMA 5.4 (discrete comparison principle for $T_{\varepsilon,m}$). *Let $u_h, w_h \in \mathbb{V}_h^1$ with $u_h \leq w_h$ on the boundary $\partial\Omega_h$ be such that*

$$T_{\varepsilon,m}[u_h](x_i) \geq T_{\varepsilon,m}[w_h](x_i) \geq 0 \quad \forall x_i \in \mathcal{N}_h^0. \tag{5.3}$$

Then, $u_h \leq w_h$ everywhere.

We now provide a consistency estimate for $T_{\varepsilon,m}$ and later compare it with $T_{\varepsilon,a}$. To this end, given a node $x_i \in \mathcal{N}_h^0$ we denote

$$B_i := \cup\{\overline{T} : T \in \mathcal{T}_h, \text{ dist } (x_i, T) \leq \hat{\delta}\}, \tag{5.4}$$

where $\hat{\delta} := \rho\delta$ with $0 < \rho \leq 1$ is so that $x_i \pm \hat{\delta} v_j \in \overline{\Omega}_h$ for all $v_j \in \mathbb{S}_\theta$. We also introduce the δ-interior region

$$\Omega_{h,\delta} = \{T \in \mathcal{T}_h \ : \ \text{dist}(x, \partial\Omega_h) \geq \delta \ \forall x \in T\}.$$

The above notation is introduced for general δ and θ and is adjusted accordingly for each of the following consistency lemmas. The following result is proven in [25, Lemma 4.2].

LEMMA 5.5 (consistency of $T_{\varepsilon,m}[\mathcal{I}_h^1 u]$). *Let $x_i \in \mathcal{N}_h^0 \cap \Omega_{h,\delta_m}$ and B_i be defined as in (5.4). Let also $u \in C^{2+k,\alpha}(B_i)$ with $0 < \alpha \leq 1$ and $k = 0, 1$ convex, and let $\mathcal{I}_h^1 u$ be its piecewise linear interpolant. Then*

$$\left|\det D^2 u(x_i) - T_{\varepsilon,m}[\mathcal{I}_h^1 u](x_i)\right| \leq C_1(d,\Omega,u)\delta_m^{k+\alpha} + C_2(d,\Omega,u)\left(\frac{h^2}{\delta_m^2} + \theta_m^2\right), \tag{5.5}$$

where

$$C_1(d,\Omega,u) = C|u|_{C^{2+k,\alpha}(B_i)}|u|^{d-1}_{W^2_\infty(B_i)}, \quad C_2 = C|u|^d_{W^2_\infty(B_i)}. \tag{5.6}$$

If $x_i \in \mathcal{N}_h^0$ and $u \in W^2_\infty(B_i)$, then (5.5) remains valid with $\alpha = k = 0$ and $C^{2+k,\alpha}(B_i)$ replaced by $W^2_\infty(B_i)$.

Remark 5.6 (optimal consistency error). We observe that, for a smooth function u, Lemma 5.5 gives a consistency error of order h upon equating δ_m^2, θ_m^2, and $\frac{h^2}{\delta_m^2}$ and taking $\delta_m = h^{1/2}$; this explains the accuracy barrier alluded to in section 1.

We now present a consistency lemma for $T_{\varepsilon,a}$ to showcase the improved formal accuracy of $T_{\varepsilon,a}$ relative to $T_{\varepsilon,m}$. We only sketch the proof of the result, since it mostly follows along the lines of [25, Lemma 4.1 (consistency of $\nabla^2_{\delta_m}\mathcal{I}_h u$)]. Note that the use of isoparametric finite elements in $\mathcal{T}_{2h}^2$ does not affect the following results, because in the interior of the domain the elements of $\mathcal{T}_{2h}^2$ are straight and the δ_a-interior domain Ω_{h,δ_a} does not contain curved boundary cells.

LEMMA 5.7 (consistency of $\nabla^2_{\delta_a}\mathcal{I}^2_{2h}u$). *Let $u \in W^3_\infty(B_i)$, $\mathcal{I}^2_{2h}u$ be its Lagrange interpolant in $\mathbb{V}^2_{2h}$, and B_i be as defined in (5.4). The following two estimates are then valid:*

(i) *For all $x_i \in \mathcal{N}_h^0$ and all $v_j \in \mathbb{S}_{\theta_a}$, we have*

$$\left|\nabla^2_{\delta_a}\mathcal{I}^2_{2h}u(x_i; v_j)\right| \leq C\ |u|_{W^2_\infty(B_i)}.$$

(ii) *If in addition $u \in C^{2+k,\alpha}(B_i)$ for $k = 0, \ldots, 3$ and $\alpha \in (0, 1]$, then for all $x_i \in \mathcal{N}_h^0 \cap \Omega_{h,\delta_a}$ and all $v_j \in \mathbb{S}_{\theta_a}$, we have*

$$\left|\nabla^2_{\delta_a}\mathcal{I}^2_{2h}u(x_i; v_j) - \frac{\partial^2 u}{\partial v_j^2}(x_i)\right| \leq C\left(|u|_{C^{2+k,\alpha}(B_i)}\delta_a^{k+\alpha} + |u|_{W^3_\infty(B_i)}\frac{h^3}{\delta_a^2}\right).$$

In both cases C stands for a constant independent of the two scales h and δ_a, the parameter θ_a, and the function u.

Proof. We rewrite (1.5) as

$$\nabla^2_{\delta_a} u(x;v) = \frac{4}{3}\frac{u(x+\frac{\delta_a}{2}) - 2u(x) + u(x-\frac{\delta_a}{2})}{\left(\frac{\delta_a}{2}\right)^2} - \frac{1}{3}\frac{u(x+\delta_a v) - 2u(x) + u(x-\delta_a v)}{\delta_a^2}$$

and then proceed as in [25, Lemma 4.1 (consistency of $\nabla^2_{\delta_m}\mathcal{I}_h u$)] to obtain

$$\nabla^2_{\delta_a} u(x_i;v) \le C|u|_{W^2_\infty(B_i)}.$$

We then incorporate the L^∞-interpolation error estimate for quadratics [7]

$$\|u - \mathcal{I}^2_{2h}u\|_{L^\infty(\Omega_h)} \le C\ |u|_{W^m_\infty(B_i)} h^m, \quad m = 2, 3,$$

with $m = 2$, along with the relation $h \le \delta_a$, to complete the proof of (i). To prove (ii) we make use of $m = 3$ and exploit cancellation of higher order derivatives built in the definition of $\nabla^2_{\delta_a} u(x_i;v)$. We refer the reader to [25, Lemma 4.1(ii)] for similar estimates for $T_{\varepsilon,m}$. ∎

Lemma 5.7 implies the following result, which is similar to [25, Lemma 4.2].

LEMMA 5.8 (consistency of $T_{\varepsilon,a}[\mathcal{I}^2_{2h}u]$). *Let $x_i \in \mathcal{N}^0_h \cap \Omega_{h,\delta_a}$ and B_i be defined as in* (5.4). *If $u \in C^{2+k,\alpha}(B_i)$ with $0 < \alpha \le 1$ and $k = 0, \dots, 3$ is convex, and $\mathcal{I}^2_{2h}u$ is its piecewise quadratic interpolant, then*

$$\begin{aligned}\left|\det D^2 u(x_i) - T_{\varepsilon,a}[\mathcal{I}^2_{2h}u](x_i)\right| &\le C_1(d,\Omega,u)\delta_a^{k+\alpha} \\ &\quad + C_2(d,\Omega,u)\frac{h^3}{\delta_a^2} + C_3(d,\Omega,u)\theta_a^2,\end{aligned} \tag{5.7}$$

where

$$\begin{aligned} C_1(d,\Omega,u) &= C|u|_{C^{2+k,\alpha}(B_i)}|u|^{d-1}_{W^2_\infty(B_i)}, \\ C_2 &= C|u|_{W^3_\infty(B_i)}|u|^{d-1}_{W^2_\infty(B_i)}, \\ C_3 &= C|u|^d_{W^2_\infty(B_i)}. \end{aligned} \tag{5.8}$$

If $x_i \in \mathcal{N}^0_h$ and $u \in W^2_\infty(B_i)$, then (5.7) *remains valid with $\alpha = k = 0$ and $C^{2+k,\alpha}(B_i)$ replaced by $W^2_\infty(B_i)$.*

Proof. We argue as in [25, Lemma 4.2]; namely we use Lemma 5.7 (consistency of $\nabla^2_{\delta_a}\mathcal{I}^2_{2h}u$) along with the bound $|\lambda_j| \le |u|_{W^2_\infty(B_i)}$ on the eigenvalues λ_j of D^2u. ∎

Remark 5.9 (improved consistency error). We observe that for a smooth function $u \in C^{5,1}(\overline{\Omega})$ Lemma 5.8 gives a consistency error of order h^2, instead of only order h, which corresponds again to the optimal choice $\delta_a = h^{1/2}$.

6. Filtered scheme: Analysis. In this section we prove the existence of discrete solutions of (1.7) and their convergence to the unique viscosity solution of (1.1), the two main theoretical results of this paper. An important component for this analysis is the following property of the filtered operator $T_{\varepsilon,f}$, which mimics Lemma 5.3 (monotonicity of $T_{\varepsilon,m}$) and hinges on the uniform bound of the filter F.

LEMMA 6.1 (almost monotonicity of $T_{\varepsilon,f}$). *For two grid functions $\mathbf{V}_h, \mathbf{W}_h$ such that $\mathbf{V}_h - \mathbf{W}_h$ attains a maximum at an interior node $x_i \in \mathcal{N}^0_h$, we have that*

$$T_{\varepsilon,f}[\mathbf{W}_h](x_i) + 2\tau \ge T_{\varepsilon,f}[\mathbf{V}_h](x_i).$$

Proof. Let $v_h^1, w_h^1 \in \mathbb{V}_h^1$ and $v_h^2, w_h^2 \in \mathbb{V}_{2h}^2$ be the corresponding functions with the same nodal values as $\mathbf{V}_h$ and $\mathbf{W}_h$, respectively. We use the definition (1.7) of $T_{\varepsilon,f}$ and property $|F(s)| \leq 1$ for all s to write

$$T_{\varepsilon,f}[\mathbf{W}_h](x_i) = T_{\varepsilon,m}[w_h^1](x_i) + k_i\ \tau, \quad T_{\varepsilon,f}[\mathbf{V}_h](x_i) = T_{\varepsilon,m}[v_h^1](x_i) + \tilde{k}_i\ \tau,$$

where $|k_i|, |\tilde{k}_i| \leq 1$. We now apply [25, Lemma 2.3 (monotonicity)] to $T_{\varepsilon,m}$ and v_h^1, w_h^1, whose difference attains a maximum at x_i, to deduce that

$$T_{\varepsilon,m}[w_h^1](x_i) \geq T_{\varepsilon,m}[v_h^1](x_i).$$

Combining these expressions, we obtain

$$T_{\varepsilon,f}[\mathbf{W}_h](x_i) + (\tilde{k}_i - k_i)\tau \geq T_{\varepsilon,f}[\mathbf{V}_h](x_i),$$

whence the assertion follows immediately. ∎

6.1. Existence of discrete solution. In this section we prove that (1.7) has a discrete solution $\mathbf{U}_\varepsilon$. This hinges on Lemma 6.1 (almost monotonicity of $T_{\varepsilon,f}$) and the existence results in [25] for $T_{\varepsilon,m}$. In fact, we combine the latter with a fixed point argument as in [18]. Moreover, we show that, although we cannot guarantee uniqueness, we can control the $l^\infty(\mathcal{N}_h)$ difference between distinct discrete solutions.

LEMMA 6.2 (existence and stability). *Let the filter F_σ be defined by either* (2.2) *if $f \geq f_0 > 0$ or* (2.5) *if $f \geq 0$ and $0 < \tau = \tau(h) < f_0$ or $\tau = \tau(h) > 0$, respectively. Then, there exists a grid function $\mathbf{U}_\varepsilon$ that solves* (1.7) *and so that the corresponding function u_ε^1 is discretely convex. Moreover, $\mathbf{U}_\varepsilon$ is stable in the sense that $\|\mathbf{U}_\varepsilon\|_{l^\infty(\mathcal{N}_h)}$ does not depend on the parameters $\varepsilon = (h, \delta_a, \delta_m, \theta_a, \theta_m)$ and τ.*

Proof. We first show that for any grid function $\mathbf{U}_h$, with corresponding functions $u_h^1 \in \mathbb{V}_h^1$ discretely convex and $u_h^2 \in \mathbb{V}_{2h}^2$, we can find a grid function $\mathbf{Y}_h(\mathbf{U}_h)$ with corresponding function $y_h^1 = y_h^1(\mathbf{U}_h^1) \in \mathbb{V}_h^1$ such that for all $x_i \in \mathcal{N}_h^0$

$$T_{\varepsilon,m}[y_h^1(\mathbf{U}_h)](x_i) = f(x_i) - \tau F_\sigma\Big(\frac{A[\mathbf{U}_h](x_i)}{\tau}\Big). \tag{6.1}$$

We proceed in two steps.

Step 1: *Existence of y_h^1.* Let $\mathbf{U}_h$ be a grid function with corresponding function $u_h^1 \in \mathbb{V}_h^1$ discretely convex, or equivalently $T_{\varepsilon,m}[u_h^1](x_i) \geq 0$ for all $x_i \in \mathcal{N}_h^0$. If $f \geq f_0 > 0$, then for $0 < \tau \leq f_0$ we infer that

$$f(x_i) - \tau F_\sigma\Big(\frac{A[\mathbf{U}_h](x_i)}{\tau}\Big) \geq 0 \quad \forall\, x_i \in \mathcal{N}_h^0.$$

On the other hand, if $f \geq 0$, then (2.5) implies $F_\sigma(s) \leq 0$ for all s and the above inequality holds again. We next extend $F_\sigma\big(\tau^{-1}A[\mathbf{U}_h](x_i)\big)$ as a continuous piecewise linear function to Ω_h and apply the existence result for $T_{\varepsilon,m}$ from [25, Lemma 3.1] to conclude that there exists a unique solution $y_h^1(\mathbf{U}_h) \in \mathbb{V}_h^1$ to (6.1) with $\|y_h^1(\mathbf{U}_h)\|_{L^\infty} \leq \Lambda$. Since τ and F_σ are uniformly bounded, the latter by 1 for both (2.2) and (2.5), Λ only depends on $\|g\|_{L^\infty(\partial\Omega)}$ and $\|f\|_{L^\infty(\Omega)}$. We have thus constructed a grid function $\mathbf{Y}_h(\mathbf{U}_h)$ with nodal values given by $y_h^1(\mathbf{U}_h)$ and such that $\|\mathbf{Y}_h(\mathbf{U}_h)\|_{\ell^\infty(\mathcal{N}_h)} \leq \Lambda$.

Step 2: *Fixed point argument.* Since F_σ is continuous for any $\sigma > 0$, and the solution of (6.1) depends continuously on data in $L^\infty(\Omega_h)$, according to [26, Proposition 4.6], we deduce that the map $\mathbf{U}_h \mapsto \mathbf{Y}_h(\mathbf{U}_h)$ is continuous. In addition, the set

of grid functions $\mathbf{U}_h$ with corresponding discretely convex $u_h^1 \in \mathbb{V}_h^1$ that satisfy both the boundary condition $\mathbf{U}_h(x_i) = g(x_i)$ for all $x_i \in \mathcal{N}_h^b$ as well as the uniform bound $\|\mathbf{U}_h\|_{l^\infty(\mathcal{N}_h)} \le \Lambda$ is compact and convex. Since $\mathbf{U}_h \mapsto \mathbf{Y}_h(\mathbf{U}_h)$ maps this set into itself, we can apply Brouwer's fixed point theorem to find $\mathbf{U}_h$ such that $\mathbf{Y}_h(\mathbf{U}_h) = \mathbf{U}_h$, which is thus a solution to (1.7). This concludes the proof. ∎

Remark 6.3 (nonuniqueness). We emphasize that the above proof does not guarantee the existence of a unique solution to (1.7), since in principle we can have more than one fixed point for (6.1). However, the next lemma shows that two different solutions of (1.7) are very close to each other. Their distance in the l^∞-norm is dictated by the filter scale τ.

LEMMA 6.4 (control of the lack of uniqueness). *Let $\mathbf{U}_\varepsilon, \mathbf{V}_\varepsilon$ be two discrete solutions of* (1.7) *with filter F satisfying* (2.6)*, and let the corresponding discretely convex functions $u_\varepsilon^1, v_\varepsilon^1 \in \mathbb{V}_h^1$. Then,*

$$\|\mathbf{U}_\varepsilon - \mathbf{V}_\varepsilon\|_{l^\infty(\Omega)} \le C\ \tau^{1/d},$$

where C depends only on the dimension d and Ω.

Proof. The result is an immediate consequence of [25, Lemma 2.4 (discrete comparison principle)] for $T_{\varepsilon,m}$ and the fact that (2.6) implies

$$T_{\varepsilon,m}[u_\varepsilon^1](x_i) \ge 0,\ T_{\varepsilon,m}[v_\varepsilon^1](x_i) \ge 0 \quad \forall\, x_i \in \mathcal{N}_h^0.$$

In fact, we use the discrete barrier $q_h = \mathcal{I}_h^1(|x - x_0|^2 - R^2) \in \mathbb{V}_h^1$, introduced in [25, Lemma 5.2], where x_0 and $R > 0$ are such that $\Omega \subset B_R(x_0)$. Since $T_{\varepsilon,f}[\mathbf{U}_\varepsilon](x_i) = T_{\varepsilon,f}[\mathbf{V}_\varepsilon](x_i) = f(x_i)$ for all $x_i \in \mathcal{N}_h^0$, we have that

$$T_{\varepsilon,m}[u_\varepsilon^1](x_i) \le T_{\varepsilon,m}[v_\varepsilon^1](x_i) + 2\tau \le T_{\varepsilon,m}\Big[v_\varepsilon^1 + \frac{(2\tau)^{1/d}}{2} q_h\Big](x_i)$$

and

$$T_{\varepsilon,m}[v_\varepsilon^1](x_i) \le T_{\varepsilon,m}[u_\varepsilon^1](x_i) + 2\tau \le T_{\varepsilon,m}\Big[u_\varepsilon^1 + \frac{(2\tau)^{1/d}}{2} q_h\Big](x_i)$$

for all $x_i \in \mathcal{N}_h^0$. Applying [25, Lemma 2.4], we obtain

$$\|u_\varepsilon^1 - v_\varepsilon^1\|_{L^\infty(\Omega_h)} \le \frac{(2\tau)^{1/d}}{2} \|q_h\|_{L^\infty(\Omega_h)} \le C\tau^{1/d},$$

which concludes the proof. ∎

6.2. Convergence. We now prove convergence of the function u_ε^1 associated with the solution $\mathbf{U}_\varepsilon$ of (1.7) to the unique viscosity u solution of (1.1). To this end, we follow the proof of convergence of [25, Theorem 5.7], which in turn modifies that of [2] to account for the Dirichlet boundary conditions and the lack of operator consistency near $\partial\Omega_h$. In addition, we exploit the almost monotone nature of the scheme, as in [18], to further adjust the proof of [2] and derive uniform convergence of u_ε^1 to u in Ω. As in [25], we also resort to a discrete barrier argument to control the behavior of the discrete solution close to the boundary. We start with the discrete barrier function from [25, Lemma 5.1].

LEMMA 6.5 (discrete boundary barrier). *Let Ω be uniformly convex and $E > 0$ be arbitrary. For each node $z \in \mathcal{N}_h^0$ with $\operatorname{dist}(z, \partial\Omega_h) \le \delta$, there exists a function $p_h \in \mathbb{V}_h^1$ such that $T_{\varepsilon,m}[p_h](x_i) \ge E$ for all $x_i \in \mathcal{N}_h^0$, $p_h \le 0$ on $\partial\Omega_h$ and*

$$|p_h(z)| \le CE^{1/d}\delta,$$

with C depending on the curvature of the boundary.

Before proceeding further, we recall the following continuous version of the Monge–Ampère operator:

$$T[u] := \min_{\mathbf{v}=(v_j)_{j=1}^d \in \mathbb{S}^\perp} \left(\prod_{j=1}^d \partial^{2,+}_{v_j v_j} u - \sum_{j=1}^d \partial^{2,-}_{v_j v_j} u \right),$$

where $\partial^{2,+}_{v_j v_j} u := \max\left(\partial^2_{v_j v_j} u, 0\right)$ and $\partial^{2,-}_{v_j v_j} u := -\min\left(\partial^2_{v_j v_j} u, 0\right)$. The following equivalence between convex viscosity solutions of (1.1) and viscosity solutions of

$$T[u] = f \quad \text{in } \Omega, \qquad u = g \quad \text{on } \partial\Omega, \tag{6.2}$$

is proven in [25, Lemma 5.6].

LEMMA 6.6 (equivalence of viscosity solutions). *If $f \in C(\Omega)$ satisfies $f \geq 0$, and $u \in C(\overline{\Omega})$, then u is a viscosity solution of* (6.2) *if and only if u is a convex viscosity solution of* (1.1).

We are now in a position to prove the uniform convergence of $u^1_\varepsilon \in \mathbb{V}^1_h$ in Ω. Since u^1_ε is defined in the computational domain Ω_h, and $\Omega_h \subset \Omega$, we extend u^1_ε to Ω as follows. Given $x \in \Omega \setminus \Omega_h$ let $z \in \partial\Omega_h$ be the closest point to x, which is unique because Ω_h is convex, and let

$$u^1_\varepsilon(x) := u^1_\varepsilon(z) = \mathcal{I}^1_h g(z) \quad \forall\, x \in \Omega \setminus \Omega_h. \tag{6.3}$$

This will allow control of the behavior of u^1_ε close to $\partial\Omega$ using techniques from [25].

We also introduce the limit supremum and the limit infimum of u^1_ε, namely

$$u^*(x) = \limsup_{\varepsilon \to 0, z \to x} u^1_\varepsilon(z), \qquad u_*(x) = \liminf_{\varepsilon \to 0, z \to x} u^1_\varepsilon(z) \quad \forall x \in \Omega,$$

where we require without explicit statement that $\frac{h}{\delta_m}, \frac{h}{\delta_a} \to 0$ as $h \to 0$. We observe that u^* is upper semicontinuous and u_* is lower semicontinuous. Since the proof closely follows the one in [25], we emphasize only the parts of it that are different for the filtered operator. In the following calculations we do not rely on the precise definition of the filter function F_σ but use Lemma 6.1 (almost monotonicity of $T_{\varepsilon,f}$).

THEOREM 6.7 (uniform convergence). *Let Ω be uniformly convex, $f \in C(\overline{\Omega}) \cap L^\infty(\Omega)$ satisfy $f \geq 0$, and $g \in C(\partial\Omega)$. Let the filter function F_σ satisfy $|F_\sigma(\cdot)| \leq 1$ and $u^1_\varepsilon \in \mathbb{V}^1_h$ be so that the vector of its nodal values $\mathbf{U}_\varepsilon \in \mathbb{R}^N$ solves* (1.7). *Then u_ε converges uniformly to the unique viscosity solution $u \in C(\overline{\Omega})$ of* (1.1) *as $\varepsilon = \varepsilon(h) \to 0$.*

Proof. In view of Lemma 6.6 (equivalence of viscosity solutions), we prove instead that u^1_ε converges to the viscosity solution u of (6.2) uniformly. To this end, we have to deal with a test function $\phi \in C^2(\Omega)$ and the respective grid function $\mathbf{\Phi}_h$ with corresponding piecewise polynomial functions $\phi^1_h = \mathcal{I}^1_h \phi \in \mathbb{V}^1_h$ and $\phi^2_h = \mathcal{I}^2_{2h} \phi \in \mathbb{V}^2_{2h}$. Without loss of generality we may assume $\phi \in C^{2,\alpha}(\Omega)$. We split the proof into five steps.

Step 1: *Consistency.* Let $x_0 \in \Omega$ and $x_i \in \mathcal{N}^0_h \cap \Omega_{h,\delta_m}$. We have the following consistency estimate for the operator T in (6.2), which is an immediate consequence of Lemma 5.5 (consistency of $T_{\varepsilon,m}[\mathcal{I}^1_h u]$), the Lipschitz continuity of the min and max

functions, and the fact that $|T_{\varepsilon,f}[\mathbf{\Phi}_h](x_i) - T_{\varepsilon,m}[\phi_h^1](x_i)| \le \tau$:

$$\big|T[\phi](x_0) - T_{\varepsilon,f}[\mathbf{\Phi}_h](x_i)\big| \le C_1(\phi)\Big(\delta_m^\alpha + |x_0 - x_i|^\alpha\Big) + C_2(\phi)\Big(\frac{h^2}{\delta_m^2} + \theta_m^2\Big) + \tau.$$

Here the constants C_1, C_2 are defined as in Lemma 5.5 and depend on $|\phi|_{C^{2,\alpha}(B_i)}$ and $|\phi|_{W^2_\infty(B_i)}$ with B_i as defined in (5.4).

Step 2: *Subsolutions.* We show that u^* is a viscosity subsolution of (6.2); likewise u_* is a viscosity supersolution. This hinges on monotonicity and consistency [2]. In our case, we employ Lemma 5.3 (monotonicity of $T_{\varepsilon,m}$) and Lemma 6.1 (almost monotonicity of $T_{\varepsilon,f}$). We must show that if $u^* - \phi$ attains a local maximum at $x_0 \in \Omega$, we have

$$T[\phi](x_0) \ge f(x_0);$$

note that $u^* - \phi$ is upper semicontinuous and the local maximum is well defined. Without loss of generality, we may assume that $u^* - \phi$ attains a strict global maximum at x_0 [22, Remark in p.31] and $x_0 \in \Omega_h$ for h sufficiently small. Let $x_h \in \mathcal{N}_h$ be a sequence of nodes so that $\mathbf{U}_\varepsilon - \mathbf{\Phi}_h$ attains a maximum at x_h. We claim that, as in [25], $x_h \to x_0$ as $h \to 0$. Exploiting the fact that $\mathbf{U}_\varepsilon - \mathbf{\Phi}_h$ attains a maximum at x_h, Lemma 6.1 (almost monotonicity of $T_{\varepsilon,f}$) yields

$$T_{\varepsilon,f}[\mathbf{\Phi}_h](x_h) + 2\tau \ge T_{\varepsilon,f}[u_\varepsilon](x_h) = f(x_h),$$

where $\tau \to 0$ as $\varepsilon \to 0$. Since $f \in C(\overline{\Omega})$, to prove $T[\phi](x_0) \ge f(x_0)$ we only need to show that as $\varepsilon \to 0$

$$T_{\varepsilon,f}[\mathbf{\Phi}_h](x_h) \to T[\phi](x_0).$$

This is a consequence of Step 1 and the fact that $x_h \in \Omega_{h,\delta_m}$ for δ_m sufficiently small, because $x_0 \in \Omega$, $x_h \to x_0$, and the sequence of $\Omega_h \uparrow \Omega$ is nondecreasing.

Step 3: *Boundary behavior.* We now prove that $u^* = u_* = g$ on $\partial\Omega$ via a barrier argument similar to those in [13, 25, 27]; we proceed as in [13]. This is essential in order to apply the comparison principle for operator T to relate u_*, u^* and u in Step 4.

Let p_k be the quadratic function in the proof of Lemma 6.5 (discrete boundary barrier) associated with an arbitrary boundary point $x \in \partial\Omega$ (the origin in the construction of p_k) and with constant $E = k$. We recall that $p_k(x) = 0$ and $p_k(z) \le 0$ for all $z \in \partial\Omega$ can be made arbitrarily large for $k \to \infty$ by virtue of the uniform convexity of Ω. A simple consequence is that the sequence of points $x_k \in \partial\Omega$ where $g + p_k$ (resp., $g - p_k$) attains a maximum (resp., a minimum) over $\partial\Omega$ converges to x as $k \to \infty$.

Lemma 5.3 (monotonicity of $T_{\varepsilon,m}$) implies the following maximum principle for the monotone operator: if $T_{\varepsilon,m}[u_h^1](x_i) > 0$ for all $x_i \in \mathcal{N}_h^0$ is valid for a discretely convex function $u_h^1 \in \mathbb{V}_h^1$, then u_h^1 attains a maximum over Ω_h on $\mathcal{N}_h^b \subset \partial\Omega$. We now see that, since $T_{\varepsilon,m}[u_\varepsilon^1] \ge 0$, we have that for k big enough and all $x_i \in \mathcal{N}_h^0$

$$\begin{aligned} T_{\varepsilon,m}[u_\varepsilon^1 + \mathcal{I}_h^1 p_k](x_i) &\ge T_{\varepsilon,m}[u_\varepsilon^1](x_i) + T_{\varepsilon,m}[\mathcal{I}_h^1 p_k](x_i) \\ &\ge T_{\varepsilon,f}[\mathbf{U}_\varepsilon](x_i) - \tau + E \\ &= f(x_i) - \tau + E > 0, \end{aligned}$$

whence $u_\varepsilon^1 + \mathcal{I}_h^1 p_k$ attains its maximum on $\mathcal{N}_h^b$. In view of (6.3), we may assume

$z \in \Omega_h$ in the limit $u^*(x) = \limsup_{\varepsilon, \frac{h}{\delta} \to 0, z \to x} u_\varepsilon^1(z)$. Consequently,

$$\begin{aligned} u^*(x) &\leq \limsup_{\varepsilon \to 0, z \to x} \big(u_\varepsilon^1(z) + \mathcal{I}_h^1 p_k(z)\big) - \liminf_{\varepsilon \to 0, z \to x} \mathcal{I}_h^1 p_k(z) \\ &\leq \limsup_{\varepsilon \to 0} \max_{z \in \mathcal{N}_h^b} \big(g + p_k\big)(z) - p_k(x) \leq g(x_k) + p_k(x_k) \leq g(x_k), \end{aligned}$$

because $\max_{\mathcal{N}_h^b} g + p_k \leq \max_{\partial\Omega} g + p_k$. Hence taking $k \to \infty$ yields $u^*(x) \leq g(x)$.

On the other hand, since $T_{\varepsilon,m}[\mathcal{I}_h^1 p_k](x_i) > T_{\varepsilon,m}[u_\varepsilon^1](x_i)$ for all $x_i \in \mathcal{N}_h^0$ and k big enough, Lemma 5.3 (monotonicity of $T_{\varepsilon,m}$) implies that $u_\varepsilon^1 - \mathcal{I}_h^1 p_k$ attains a minimum on $\mathcal{N}_h^b$. Therefore, arguing as before,

$$u_*(x) \geq \liminf_{\varepsilon \to 0} \min_{z \in \partial\Omega} \big(g - p_k\big)(z) + p_k(x) \geq g(x_k) - p_k(x_k) \geq g(x_k),$$

whence $u_*(x) \geq g(x)$. This in turn gives $u^* \leq g \leq u_* \leq u^*$ on $\partial\Omega$ as asserted.

Step 4: *Comparison.* To prove that $u^* = u_*$ in $\overline{\Omega}$ we make use of the comparison principle in [25] for (6.2). Since u^* and u_* are a subsolution and supersolution, respectively, of (6.2) and they agree on the boundary, we can deduce that $u^* \leq u_*$ in $\overline{\Omega}$. Combining with $u^* \geq u_*$, by definition, this results in $u^* = u_*$ in $\overline{\Omega}$.

Step 5: *Uniform convergence.* This is identical to [25] and is thus omitted. The proof is complete. ∎

7. Conclusions. In this paper we introduce two methods to solve the Monge–Ampère equation (1.1). The first one is an accurate scheme that hinges on quadratic interpolation and a higher order approximation of directional derivatives in (1.2). It exhibits errors in the L^∞-norm of one to two orders of magnitude lower than the monotone operator introduced in [25]. However, formal higher order accuracy comes at the cost of monotonicity, which prevents us from proving convergence in L^∞ for this operator. The second method circumvents this issue by combining the monotone and the accurate operators into a filtered scheme. This yields convergence to the viscosity solution relying on stability and monotonicity properties of the monotone operator and the fact that the filter scale $\tau \to 0$ as $h \to 0$. We employ two filter functions according to whether the forcing f is strictly positive or degenerate. In both cases, the discrete piecewise linear solution $u_\varepsilon^1 \in \mathbb{V}_h^1$ is discretely convex. The filter detects parts of the domain where the accurate operator could underperform due to lack of regularity of the solution, as happens in our degenerate example, and switches to the monotone operator. We explore the two methods computationally and illustrate the enhanced performance of both schemes by comparing them with the numerical experiments from [25]. Finally, we investigate the effect of filter function and filter scale and discuss some computational challenges of the method.

Acknowledgments. We are indebted to S. W. Walker for providing assistance and guidance with the software FELICITY and to H. Antil for numerous discussions about the implementation of the method. We also thank W. Zhang for early discussions about the filter methodology.

REFERENCES

[1] N. E. Aguilera and P. Morin, *On convex functions and the finite element method*, SIAM J. Numer. Anal., 47 (2009), pp. 3139–3157, https://doi.org/10.1137/080720917.

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

[3] J.-D. Benamou, F. Collino, and J.-M. Mirebeau, *Monotone and consistent discretization of the Monge-Ampère operator*, Math. Comp., 85 (2016), pp. 2743–2775.
[4] J.-D. Benamou, B. Froese, and A. Oberman, *Two numerical methods for the elliptic Monge-Ampère equation*, M2AN Math. Model. Numer. Anal., 44 (2010), pp. 737–758.
[5] D. P. Bertsekas, A. Nedic, and A. E. Ozdaglar, *Convex Analysis and Optimization*, Athena Scientific, Belmont, MA, 2003.
[6] O. Bokanowski, M. Falcone, and S. Sahu, *An efficient filtered scheme for some first order time-dependent Hamilton–Jacobi equations*, SIAM J. Sci. Comput., 38 (2016), pp. A171–A195, https://doi.org/10.1137/140998482.
[7] S. C. Brenner and R. Scott, *The Mathematical Theory of Finite Element Methods*, Springer, New York, 2008.
[8] 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.
[9] P. G. Ciarlet, *The Finite Element Method for Elliptic Problems*, SIAM, Philadelphia, 2002, https://doi.org/10.1137/1.9780898719208.
[10] X. Chen, Z. Nashed, and L. Qi, *Smoothing methods and semismooth methods for nondifferentiable operator equations*, SIAM J. Numer. Anal., 38 (2000), pp. 1200–1216, https://doi.org/10.1137/S0036142999356719.
[11] E. J. Dean and R. Glowinski, *An augmented Lagrangian approach to the numerical solution of the Dirichlet problem for the elliptic Monge-Ampère equation in two dimensions*, Electron. Trans. Numer. Anal., 22 (2006), pp. 71–96.
[12] E. J. Dean and R. Glowinski, *On the numerical solution of the elliptic Monge-Ampère equation in dimension two: A least-squares approach*, in Partial Differential Equations, Comput. Methods Appl. Sci. 16, Springer, Dordrecht, The Netherlands, 2008, pp. 43–63.
[13] X. Feng and M. Jensen, *Convergent semi-Lagrangian methods for the Monge–Ampère equation on unstructured grids*, SIAM J. Numer. Anal., 55 (2017), pp. 691–712, https://doi.org/10.1137/16M1061709.
[14] X. Feng, R. Glowinski, and M. Neilan, *Recent developments in numerical methods for fully nonlinear second order partial differential equations*, SIAM Rev., 55 (2013), pp. 205–267, https://doi.org/10.1137/110825960.
[15] X. Feng and M. Neilan, *Mixed finite element methods for the fully nonlinear Monge–Ampère equation based on the vanishing moment method*, SIAM J. Numer. Anal., 47 (2009), pp. 1226–1250, https://doi.org/10.1137/070710378.
[16] 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.
[17] B. Froese and A. 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, https://doi.org/10.1137/100803092.
[18] B. Froese and A. Oberman, *Convergent filtered schemes for the Monge–Ampère partial differential equation*, SIAM J. Numer. Anal., 51 (2013), pp. 423–444, https://doi.org/10.1137/120875065.
[19] R. Glowinski, *Numerical methods for fully nonlinear elliptic equations*, in Proceedings of the 6th International Congress on Industrial and Applied Mathematics (ICIAM 07), R. Jeltsch and G. Wanner, eds., European Math. Soc., Zürich, 2009, pp. 155–192.
[20] C. Gutiérrez, *The Monge-Ampère Equation*, Birkhäuser, Basel, 2001.
[21] M. Hintermüller, K. Ito, and K. Kunisch, *The primal-dual active set strategy as a semismooth Newton method*, SIAM J. Optim., 13 (2003), pp. 865–888, https://doi.org/10.1137/S1052623401383558.
[22] H. Ishii and P. L. Lions, *Viscosity solutions of fully nonlinear second-order elliptic partial differential equations*, J. Differential Equations, 83 (1990), pp. 26–78.
[23] W. Li and R. H. Nochetto, *Optimal pointwise error estimates for two-scale methods for the Monge–Ampère equation*, SIAM J. Numer. Anal., 56 (2018), pp. 1915–1941, https://doi.org/10.1137/18M1165670.
[24] J.-M. Mirebeau, *Discretization of the 3D Monge–Ampère Operator, between Wide Stencils and Power Diagrams*, preprint, https://arxiv.org/abs/1503.00947, 2015.
[25] R. H. 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.
[26] R. H. Nochetto, D. Ntogkas, and W. Zhang *Two-scale method for the Monge–Ampère equation: Pointwise error estimates*, IMA J. Numer. Anal., (2018), https://doi.org/10.1093/imanum/dry026.
[27] R. H. Nochetto and W. Zhang, *Discrete ABP estimate and convergence rates for linear elliptic equations in non-divergence form*, Found. Comput. Math., 18 (2017), pp. 537–593.

[28] R. H. Nochetto and W. Zhang, *Pointwise rates of convergence for the Oliker–Prussner method for the Monge–Ampère equation*, Numer. Math., 141 (2019), pp. 253–288.

[29] A. 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, https://doi.org/10.1137/S0036142903435235.

[30] A. Oberman and T. Salvador, *Filtered schemes for Hamilton-Jacobi equations: A simple construction of convergent accurate difference schemes*, J. Comput. Phys., 284 (2015), pp. 367–388.

[31] 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.

[32] G. Wachsmuth, *Conforming approximation of convex functions with the finite element method*, Numer. Math., 137 (2017), pp. 741–772.

[33] S. W. Walker, *FELICITY: Finite ELement Implementation and Computational Interface Tool for You*, https://www.mathworks.com/matlabcentral/fileexchange/31141-felicity.