J Sci Comput (2018) 76:1839–1867
https://doi.org/10.1007/s10915-018-0685-y

Monotone Mixed Finite Difference Scheme for Monge–Ampère Equation

Yangang Chen[1] · **Justin W. L. Wan**[2] · **Jessey Lin**[3]

Received: 11 May 2017 / Revised: 2 November 2017 / Accepted: 24 February 2018 /
Published online: 3 March 2018

Abstract In this paper, we propose a monotone mixed finite difference scheme for solving the two-dimensional Monge–Ampère equation. In order to accomplish this, we convert the Monge–Ampère equation to an equivalent Hamilton–Jacobi–Bellman (HJB) equation. Based on the HJB formulation, we apply the standard 7-point stencil discretization, which is second order accurate, to the grid points wherever monotonicity holds, and apply semi-Lagrangian wide stencil discretization elsewhere to ensure monotonicity on the entire computational domain. By dividing the admissible control set into six regions and optimizing the sub-problem in each region, the computational cost of the optimization problem at each grid point is reduced from $O(M^2)$ to $O(1)$ when the standard 7-point stencil discretization is applied and to $O(M)$ otherwise, where the discretized control set is $M \times M$. We prove that our numerical scheme satisfies consistency, stability, monotonicity and strong comparison principle, and hence is convergent to the viscosity solution of the Monge–Ampère equation. In the numerical results, second order convergence rate is achieved when the standard 7-point stencil discretization is applied monotonically on the entire computation domain, and up to order one convergence is achieved otherwise. The proposed mixed scheme yields a smaller discretization error and a faster convergence rate compared to the pure semi-Lagrangian wide stencil scheme.

Keywords Nonlinear elliptic partial differential equations · Monge–Ampère equations · Hamilton–Jacobi–Bellman equations · Viscosity solutions · Finite difference methods · Monotone schemes · Mixed schemes

✉ Yangang Chen
y493chen@uwaterloo.ca

[1] Department of Applied Mathematics, University of Waterloo, 200 University Avenue West, Waterloo, ON N2L 3G1, Canada

[2] David R. Cheriton School of Computer Science, University of Waterloo, 200 University Avenue West, Waterloo, ON N2L 3G1, Canada

[3] Centre for Computational Mathematics in Industry and Commerce, University of Waterloo, 200 University Avenue West, Waterloo, ON N2L 3G1, Canada

1 Introduction

The goal of this paper is to compute the numerical solution of the two-dimensional Monge–Ampère equation with Dirichlet boundary condition:

$$\begin{aligned} u_{xx}u_{yy} - u_{xy}^2 &= f, && \text{in } \Omega, \\ u &= g, && \text{on } \partial\Omega, \\ u &\text{ is convex}, \end{aligned} \tag{1}$$

where Ω is a bounded convex open set in $\mathbb{R}^2$, $\partial\Omega$ is its boundary, $\overline{\Omega} = \Omega \cup \partial\Omega$, $u : \overline{\Omega} \to \mathbb{R}$ is the unknown function, and $f : \Omega \to \mathbb{R}$ and $g : \partial\Omega \to \mathbb{R}$ are given functions.

The Monge–Ampère equation is of great interest due to a wide range of applications, including differential geometry, optimal mass transport (or Monge–Kantorovich) problem, image registration, mesh generation, etc. We direct the interested readers to [8] for an extensive review of applications.

The Monge–Ampère equation is a fully nonlinear partial differential equation (PDE), since the left hand side consists of products of the second derivatives. As a result, it may have multiple weak solutions. Among all these weak solutions, we are interested in computing the viscosity solution [10,11], since it is often considered the correct one in many practical applications [17]. The viscosity solution of the Monge–Ampère equation is globally convex, while the other solutions may not be convex [17]. We note that a convexity constraint is imposed in the Dirichlet problem (1) in order to select the viscosity solution and circumvent the issue of multiple weak solutions.

Due to the nonlinearity of the Monge–Ampère Eq. (1) with the additional convexity constraint, it is challenging to design a numerical scheme that converges to the viscosity solution. Some numerical schemes have been proposed in recent years. One approach is using finite difference methods. Some finite difference schemes, such as [4], use the standard central differencing to discretize u_{xy}, and are thus not monotone. The significance of monotonicity is that together with consistency, stability and strong comparison principle, they provide sufficient conditions for a numerical scheme to converge to the viscosity solution [2].

Very few finite difference schemes that are monotone and thus convergent in the viscosity sense have been proposed. One of the schemes, proposed in [28], is to exploit the geometrical interpretation of the Monge–Ampère equation. The grid structure, constrained by the geometry of the equation, is usually not rectangular or triangular. Another scheme, proposed in [17,27], uses wide stencils to achieve monotonicity. However, in order for the scheme to converge, the number of the stencil points must increase towards infinity when the mesh size h decreases towards 0, thus resulting in high computational costs for solving problems on fine grids. Some improvements on this wide stencil scheme have been proposed. For instance, in [18,19], the same authors use hybrid and filtered schemes, both integrating the wide stencil scheme with the more accurate non-monotone central difference scheme in order to improve the accuracy. That being said, the issue of infinite stencil points still exists. Recently, Ref. [3] improves on the previous wide stencil approach so that it is the least nonlocal among all wide stencils of the same family. The number of stencil points does not need to grow to infinity as $h \to 0$, but it still grows and can reach as high as 48.

Galerkin-type methods have also been developed for solving the Monge–Ampère equation. An immediate challenge is that it is not obvious how to write down the variational formulation of (1) using the common integration-by-parts approach. The L^2 projection methods, proposed in [5,7], build up the Galerkin-type schemes based on the linearized Monge–Ampère equation. Similar idea can be found in the nonvariational finite element

method in [23]. In [12], the authors reformulate the Monge–Ampère equation into an augmented Lagrangian problem or a least-squares problem, which allows the use of mixed finite element methods. The authors in [15] add an artificial fourth order elliptic differential operator $\epsilon\Delta^2 u$. They show that with this additional term, a variational formulation, and thus a finite element scheme, becomes possible. However, a common issue for these Galerkin-type methods is that convergence to the viscosity solutions for non-regular solutions remains unclear.

Our approach, which is distinct from many of the existing methods, is to first convert (1) into an equivalent Hamilton–Jacobi–Bellman (HJB) equation [22,25], and then numerically solve the equivalent HJB equation. The application of the HJB formulation in the numerical computation of the Monge–Ampère equation is first investigated by the coauthors of this paper; see the essay [24]. Another recent investigation on this approach, [14], is made public at the completion of our paper. There are some important benefits using the HJB formulation. One is that the differential operator of the HJB equation under fixed control parameters is linear. Another benefit is that the convexity constraint in (1) is already implicitly incorporated into the HJB differential operator. In other words, there is no need to impose the convexity constraint in the HJB formulation. In addition, many convergent numerical schemes for HJB equations or HJB differential operators have been developed, such as [1,6,13,16,21,26,33]. As a result, it is more tractable to design a numerical scheme that converges in the viscosity sense for the equivalent HJB equation than for the Monge–Ampère Eq. (1) with the convexity constraint.

Our primary goal is to design a monotone finite difference scheme for the equivalent HJB equation. We note that the cross derivative u_{xy} is still present in the HJB equation, and the standard central differencing or the standard 7-point stencil discretization for u_{xy} may be non-monotone. In order to achieve monotonicity, Ref. [14] follows the idea in [13,26] and applies "semi-Lagrangian scheme" on the entire computational domain, where a local coordinate rotation is performed to remove the cross derivative from the HJB equation, and then central differencing is applied with a stencil length greater than the mesh size h, resulting in *at most 17 stencil points for any* h. In some literature, such semi-Lagrangian scheme is also called wide stencil scheme, which should not to be confused with the wide stencil scheme in [17–19,27] that requires infinity stencil size as $h \to 0$. However, monotonicity is achieved at the expense of large truncation error and slow convergence. In particular, the convergence rate is no better than $O(h)$.

In order to improve the accuracy and meanwhile *strictly* maintain monotonicity, our approach is to apply a mixed standard 7-point stencil and semi-Lagrangian wide stencil discretization on the equivalent HJB equation. More specifically, the standard 7-point stencil discretization, which is second order accurate, is applied to discretize u_{xy} at a grid point if monotonicity is fulfilled. Otherwise, the semi-Lagrangian wide stencil scheme, which is less accurate but guaranteed to be monotone, is implemented. We emphasize that our discretization scheme is designed such that consistency, stability, monotonicity and strong comparison principle are fulfilled on the entire computational domain. As a result, our numerical scheme is guaranteed to converge to the viscosity solution of the Monge–Ampère equation [2]. Meanwhile, by maximal use of the standard 7-point stencil discretization, the discretization error of the numerical solution is significantly reduced, compared to the pure semi-Lagrangian wide stencil scheme in [14]. Moreover, our numerical scheme yields a convergence rate of $O(h^2)$ whenever the standard 7-point stencil discretization can be applied monotonically on the entire computation domain, and up to $O(h)$ otherwise. The second order convergence rate in the optimal cases is another significant improvement over the numerical scheme in [14].

To solve the resulting nonlinear discretized system, one of the most expensive steps is to optimize two control parameters at every grid point. Reference [14] does not discuss the computational cost of the optimization problem. Typically a bilinear search is implemented on an $M \times M$ discretized control set, resulting in $O(M^2)$ computational complexity. We propose an approach that reduces the computational cost for the optimization problem to $O(1)$ whenever the standard 7-point stencil discretization is applied, and at most $O(M)$ otherwise.

Finally, we want to emphasize that our method is the only method that fulfills all the following properties: monotone and thus convergent to the viscosity solution, second order accurate in the optimal cases, and having at most 17 stencil points independent of the mesh size h. None of the references in our paper have the same properties.

To illustrate our numerical scheme, we will briefly review the notion of viscosity solution in Sect. 2. In Sect. 3, we will establish the equivalent HJB formulation for the Monge–Ampère Eq. (1). In Sect. 4, we will describe our mixed standard 7-point stencil and semi-Lagrangian wide stencil finite difference discretization for the HJB formulation. Section 5 solves the nonlinear discretized system using policy iteration, with a detailed discussion on speeding up computation for the optimization of control parameters. Section 6 proves that our numerical scheme satisfies consistency, stability, monotonicity and strong comparison principle, and thus converges to the viscosity solution of (1). Section 7 shows numerical results. We also demonstrate the discretization error and the rate of convergence for each case. Section 8 is the conclusion.

2 Viscosity Solution of the Monge–Ampère Equation

The objective of this paper is to compute the viscosity solution of the Monge–Ampère Eq. (1). An overview on the topic of viscosity solution can be found in [10,11].

Before defining the viscosity solution of (1), we rewrite (1) as

$$\begin{aligned} \mathcal{F}\left(\mathbf{x}, u(\mathbf{x}), D^2u(\mathbf{x})\right) \equiv \begin{cases} -\det\left[D^2u(\mathbf{x})\right] + f(\mathbf{x}), & \mathbf{x} \in \Omega, \\ u(\mathbf{x}) - g(\mathbf{x}), & \mathbf{x} \in \partial\Omega, \end{cases} = 0, \\ u \text{ is convex } \Rightarrow D^2u(\mathbf{x}) \text{ is positive semi-definite,} \end{aligned} \tag{2}$$

where $\mathbf{x} = (x, y) \in \overline{\Omega}$, and D^2u is the Hessian matrix of u.

To introduce the notion of viscosity solution, we define the upper (respectively lower) semi-continuous envelope of a function $z : C \to \mathbb{R}$ on a closed set C as

$$z^*(x) \equiv \limsup_{y \to x,\, y \in C} z(y) \quad \left(\text{respectively } z_*(x) \equiv \liminf_{y \to x,\, y \in C} z(y) \right). \tag{3}$$

Definition 1 (*Viscosity solution*) A convex upper (respectively lower) semi-continuous function $u : \overline{\Omega} \to \mathbb{R}$ is a viscosity subsolution (respectively supersolution) of the Monge–Ampère equation $\mathcal{F}\left(\mathbf{x}, u(\mathbf{x}), D^2u(\mathbf{x})\right) = 0$, if for all the test functions $\varphi(\mathbf{x}) \in C^2(\overline{\Omega})$ and all $\mathbf{x} \in \overline{\Omega}$, such that $u^* - \varphi$ (respectively $u_* - \varphi$) has a local maximum (respectively minimum) at $\mathbf{x}$, we have

$$\mathcal{F}_*(\mathbf{x}, u^*(\mathbf{x}), D^2\varphi(\mathbf{x})) \le 0 \quad \left(\text{respectively } \mathcal{F}^*(\mathbf{x}, u_*(\mathbf{x}), D^2\varphi(\mathbf{x})) \ge 0 \right). \tag{4}$$

Furthermore, the function u is a viscosity solution if it is both a viscosity sub-solution and super-solution.

We note that the convexity of u (or equivalently, D^2u being positive semi-definite, $\det(D^2u) = f \geq 0$) already implies that the differential operator of (2) is degenerate elliptic. Furthermore, degenerate ellipticity, plus $\overline{\Omega}$ being bounded and convex, ensures the existence and uniqueness of the viscosity solution of (2). See [10,20] for details.

3 HJB Formulation of the Monge–Ampère Equation

Since the Monge–Ampère Eq. (2) is nonlinear, it is challenging to design a finite difference scheme that converges to the viscosity solution. Our approach is to convert the Monge–Ampère equation into an equivalent HJB equation. The equivalence of the two PDEs is first established in [22,25] for classical solutions. Recently, Ref. [14] extends the equivalence to the setting of viscosity solutions. Here we state the equivalence of the two PDEs as the following theorem:

Theorem 1 *Let Ω be a convex open set in $\mathbb{R}^2$. Let $f \in C(\Omega)$ be a non-negative function. Let a convex function u be the viscosity solution of the following HJB equation,*

$$\max_{A(\mathbf{x})\in S_1^+} \left\{ -\operatorname{tr}\left[A(\mathbf{x})D^2u(\mathbf{x})\right] + 2\sqrt{\det(A(\mathbf{x}))\, f(\mathbf{x})} \right\} = 0, \tag{5}$$

where $S_1^+ \equiv \{A \in \mathbb{R}^{2\times 2} : A$ is positive semi-definite, $A^T = A$, $\operatorname{tr}(A) = 1\}$ and $A(\mathbf{x}) \in S_1^+$ is the control at point $\mathbf{x}$. Then u is the viscosity solution of the Monge–Ampère Eq. (2).

Proof We refer interested readers to the proof in [32] when u is a classical solution, and the proof in [14] for the extension to the viscosity solution. □

We notice that due to the positive semi-definite property of the matrix $A(\mathbf{x})$, it can be diagonalized by an order-two orthogonal matrix. More specifically, $A(\mathbf{x}) \in S_1^+$ can be parametrized as follows:

$$\begin{aligned} A(\mathbf{x}) &= \begin{pmatrix} \cos\theta(\mathbf{x}) & \sin\theta(\mathbf{x}) \\ -\sin\theta(\mathbf{x}) & \cos\theta(\mathbf{x}) \end{pmatrix} \begin{pmatrix} a(\mathbf{x}) & 0 \\ 0 & 1-a(\mathbf{x}) \end{pmatrix} \begin{pmatrix} \cos\theta(\mathbf{x}) & -\sin\theta(\mathbf{x}) \\ \sin\theta(\mathbf{x}) & \cos\theta(\mathbf{x}) \end{pmatrix}, \\ a(\mathbf{x}) &\in [0,1],\ \theta(\mathbf{x}) \in [-\pi, \pi). \end{aligned} \tag{6}$$

This parametrization gives rise to the following HJB equation, which we aim at solving.

Corollary 1 *Under the parametrization (6), the HJB Eq. (5) becomes*

$$\begin{aligned} \max_{(a(\mathbf{x}),\theta(\mathbf{x}))\in\Gamma} \Big\{ &-\alpha_{11}\,(a(\mathbf{x}),\theta(\mathbf{x}))u_{xx}(\mathbf{x}) - 2\alpha_{12}(a(\mathbf{x}),\theta(\mathbf{x}))u_{xy}(\mathbf{x}) \\ &-\alpha_{22}(a(\mathbf{x}),\theta(\mathbf{x}))u_{yy}(\mathbf{x}) + 2\sqrt{a(\mathbf{x})(1-a(\mathbf{x}))f(\mathbf{x})} \Big\} = 0, \end{aligned} \tag{7}$$

where $(a(\mathbf{x}),\theta(\mathbf{x}))$ is the pair of controls at point $\mathbf{x}$, $\Gamma = [0,1]\times\left[-\frac{\pi}{4},\frac{\pi}{4}\right)$ is the set of admissible controls,[1] *and the coefficients are*

[1] Although (6) defines the admissible control set to be in the range of $[0,1]\times[-\pi,\pi)$, the optimal control pair (a^*,θ^*) that maximizes (7) may not be unique in $[0,1]\times[-\pi,\pi)$. We notice that since $\mathcal{L}_{a,\theta}\,u = \mathcal{L}_{a,\theta+\pi}\,u$, and $\mathcal{L}_{a,\theta}\,u = \mathcal{L}_{1-a,\theta+\frac{\pi}{2}}\,u$, the admissible control set Γ can be reduced to $[0,1]\times[-\frac{\pi}{4},\frac{\pi}{4})$. Such removal of the redundancy of Γ ensures that the optimal control pair (a^*,θ^*) is unique in Γ, except when $a^* = \frac{1}{2}$ or when $f = 0$.

$$\alpha_{11}(a(\mathbf{x}),\theta(\mathbf{x})) = \frac{1}{2}\left[1-(1-2a(\mathbf{x}))\cos 2\theta(\mathbf{x})\right],$$
$$\alpha_{22}(a(\mathbf{x}),\theta(\mathbf{x})) = \frac{1}{2}\left[1+(1-2a(\mathbf{x}))\cos 2\theta(\mathbf{x})\right],$$
$$\alpha_{12}(a(\mathbf{x}),\theta(\mathbf{x})) = \frac{1}{2}(1-2a(\mathbf{x}))\sin 2\theta(\mathbf{x}). \tag{8}$$

For convenience, we rewrite the HJB Eq. (7) as

$$\mathcal{F}\left(\mathbf{x}, u(\mathbf{x}), D^2u(\mathbf{x})\right) \equiv \max_{(a(\mathbf{x}),\theta(\mathbf{x}))\in\Gamma} \mathcal{L}_{a(\mathbf{x}),\theta(\mathbf{x})}\, u(\mathbf{x}) = 0, \tag{9}$$

where the differential operator of the HJB equation is given by

$$\mathcal{L}_{a,\theta}\, u \equiv -\alpha_{11}(a,\theta)u_{xx} - 2\alpha_{12}(a,\theta)u_{xy} - \alpha_{22}(a,\theta)u_{yy} + 2\sqrt{a(1-a)f}. \tag{10}$$

We note that since the HJB Eqs. (9)–(10) and the Monge–Ampère Eq. (2) are mathematically equivalent, we still use the notation $\mathcal{F}\left(\mathbf{x}, u(\mathbf{x}), D^2u(\mathbf{x})\right)$ to denote the HJB equation.

The HJB formulation introduces some favorable properties over the Monge–Ampère Eq. (2). We first notice that in the equivalent HJB Eqs. (5) or (7), the convexity constraint of the Monge–Ampère equation disappears. Indeed, the convexity constraint is implicitly enforced in the HJB formulation. The reason is that the proof of Theorem 1, where the Monge–Ampère equation is converted to the HJB equation, has already taken into account that u is a convex function. We remark that the convexity constraint poses a major difficulty in designing a convergent numerical scheme for Monge–Ampère equation; see [17] for a discussion. However, in the HJB formulation, there is no need to explicitly impose the convexity constraint any more, which makes the numerical computation more manageable.

Another useful property of the HJB Eqs. (9)–(10) is that for a fixed given control pair (a,θ), the differential operator $\mathcal{L}_{a,\theta}\, u$ is linear. We note, however, that the HJB equation itself is still nonlinear, since the maximization depends on u. Unlike (2), the linear differential operator $\mathcal{L}_{a,\theta}\, u$ does not contain products of the second derivatives. The linearity of $\mathcal{L}_{a,\theta}\, u$ allows us to develop finite difference schemes based on numerical methods for linear PDEs.

Considering these advantages of the HJB formulation, our approach is to solve the HJB Eq. (7) instead of the Monge–Ampère Eq. (2).

4 Mixed Finite Difference Discretization

In this section, we will construct a monotone finite difference discretization for the HJB Eq. (7). Monotonicity is a desirable property, since [2] has proved that monotonicity is one of the sufficient conditions for a numerical scheme to converge to the viscosity solution.

To set up notation, let us consider an $N\times N$ square grid $\left\{\mathbf{x}_{i,j} = (x_i, y_j)\right\}$, where $\mathbf{x}_{i,j}\in\Omega$ when $i, j = 1,\cdots,N$, and $\mathbf{x}_{i,j}\in\partial\Omega$ when $i, j = 0$ or $N+1$. Also, let h be the mesh size and let $u_{i,j}$, $a_{i,j}$, $\theta_{i,j}$ and $f_{i,j}$ be the grid functions of $u(\mathbf{x}_{i,j})$, $a(\mathbf{x}_{i,j})$, $\theta(\mathbf{x}_{i,j})$ and $f(\mathbf{x}_{i,j})$, respectively. Our goal is to solve the set of the unknowns $\{u_{i,j} \mid 1\le i\le N, 1\le j\le N\}$.

4.1 Standard 7-Point Stencil Discretization

Consider discretizing the HJB Eq. (7) at a grid point $\mathbf{x}_{i,j}$. We can use the standard central differencing to approximate $u_{xx}(\mathbf{x}_{i,j})$ and $u_{yy}(\mathbf{x}_{i,j})$ as follows:

$$(\delta_{xx}u)_{i,j} \equiv \frac{u_{i+1,j}-2u_{i,j}+u_{i-1,j}}{h^2}, \quad (\delta_{yy}u)_{i,j} \equiv \frac{u_{i,j+1}-2u_{i,j}+u_{i,j-1}}{h^2}. \tag{11}$$

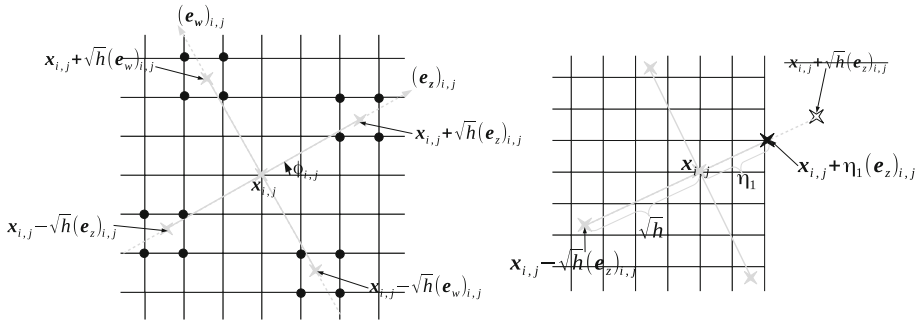


Fig. 1 (left) Local coordinate rotation at the grid point $\mathbf{x}_{i,j}$, and semi-Lagrangian wide stencil discretization of $u_{zz}(\mathbf{x}_{i,j})$ and $u_{ww}(\mathbf{x}_{i,j})$ under the rotation. The rotation angle is $\phi_{i,j}$, counter-clockwise. The grey dashed lines are the orthogonal axis $\{(\mathbf{e}_z)_{i,j}, (\mathbf{e}_w)_{i,j}\}$. The stencil length is $\sqrt{h}$ ($\sqrt{h} > h$). The grey stars are the stencil points $\mathbf{x}_{i,j} \pm \sqrt{h}(\mathbf{e}_z)_{i,j}$ and $\mathbf{x}_{i,j} \pm \sqrt{h}(\mathbf{e}_w)_{i,j}$. The unknowns at these stencil points are approximated by the bilinear interpolation from the neighboring points (black dots). Standard central differencing associated with this wide stencil is applied to approximate $u_{zz}(\mathbf{x}_{i,j})$ and $u_{ww}(\mathbf{x}_{i,j})$. (right) Semi-Lagrangian wide stencil discretization near the boundary. One of the wide stencil points $\mathbf{x}_{i,j} + \sqrt{h}(\mathbf{e}_z)_{i,j}$ falls outside $\overline{\Omega}$ (hollow star). The wide stencil is truncated and the stencil point is relocated to the point $\mathbf{x}_{i,j} + \eta_1(\mathbf{e}_z)_{i,j} \in \partial\Omega$ (black star). The corresponding stencil length has shrunk from $\sqrt{h}$ to η_1

It can be shown that the standard 7-point stencil discretization for $u_{xy}(\mathbf{x}_{i,j})$ can lead to a monotone scheme in the following two cases:

Case 1 When the coefficients α_{11}, α_{22} and α_{12} in (8) satisfy

$$\begin{aligned} &\alpha_{11}(a_{i,j}, \theta_{i,j}) \geq |\alpha_{12}(a_{i,j}, \theta_{i,j})|,\ \ \alpha_{22}(a_{i,j}, \theta_{i,j}) \geq |\alpha_{12}(a_{i,j}, \theta_{i,j})|, \\ &\qquad \text{and } \alpha_{12}(a_{i,j}, \theta_{i,j}) \geq 0 \text{ at the grid point } \mathbf{x}_{i,j}, \end{aligned} \tag{12}$$

we approximate $u_{xy}(\mathbf{x}_{i,j})$ using

$$(\delta_{xy}^{[1]} u)_{i,j} \equiv \frac{2u_{i,j} + u_{i+1,j+1} + u_{i-1,j-1} - u_{i+1,j} - u_{i-1,j} - u_{i,j+1} - u_{i,j-1}}{2h^2}. \tag{13}$$

Case 2 When the coefficients α_{11}, α_{22} and α_{12} in (8) satisfy

$$\begin{aligned} &\alpha_{11}(a_{i,j}, \theta_{i,j}) \geq |\alpha_{12}(a_{i,j}, \theta_{i,j})|,\ \ \alpha_{22}(a_{i,j}, \theta_{i,j}) \geq |\alpha_{12}(a_{i,j}, \theta_{i,j})|, \\ &\qquad \text{and } \alpha_{12}(a_{i,j}, \theta_{i,j}) \leq 0 \text{ at the grid point } \mathbf{x}_{i,j}, \end{aligned} \tag{14}$$

we approximate $u_{xy}(\mathbf{x}_{i,j})$ using

$$(\delta_{xy}^{[2]} u)_{i,j} \equiv \frac{-2u_{i,j} - u_{i+1,j-1} - u_{i-1,j+1} + u_{i+1,j} + u_{i-1,j} + u_{i,j+1} + u_{i,j-1}}{2h^2}. \tag{15}$$

4.2 Semi-Lagrangian Wide Stencil Discretization

However, if neither (12) nor (14) is fulfilled at the grid point $\mathbf{x}_{i,j}$, then it is unclear how to directly discretize the cross derivative $u_{xy}(\mathbf{x}_{i,j})$ in (7) monotonically. Our approach, following [13,26], is to eliminate the cross derivative $u_{xy}(\mathbf{x}_{i,j})$ by a local coordinate transformation. Let $\{(\mathbf{e}_z)_{i,j}, (\mathbf{e}_w)_{i,j}\}$ be a local orthogonal basis which is obtained by a rotation of the standard axes $\{(\mathbf{e}_x)_{i,j}, (\mathbf{e}_y)_{i,j}\}$ at an angle $\phi_{i,j}$; see Fig. 1 (left). If the rotation angle is chosen as

$\phi_{i,j} = \frac{1}{2}\arctan\frac{2\alpha_{12}(a_{i,j},\theta_{i,j})}{\alpha_{11}(a_{i,j},\theta_{i,j})-\alpha_{22}(a_{i,j},\theta_{i,j})} = -\theta_{i,j}$, then the cross derivative vanishes under the basis $\{(\mathbf{e}_z)_{i,j}, (\mathbf{e}_w)_{i,j}\}$. By straightforward algebra, one can show that (7) becomes

$$\max_{(a_{i,j},\theta_{i,j})\in\Gamma}\left\{-a_{i,j}\,u_{zz}(\mathbf{x}_{i,j}) - \left(1-a_{i,j}\right)\,u_{ww}(\mathbf{x}_{i,j}) + 2\sqrt{a_{i,j}\left(1-a_{i,j}\right)f_{i,j}}\right\} = 0. \quad (16)$$

Here $u_{zz}(\mathbf{x}_{i,j})$ and $u_{ww}(\mathbf{x}_{i,j})$ are the directional derivatives along the basis $(\mathbf{e}_z)_{i,j}$ and $(\mathbf{e}_w)_{i,j}$, which depend on the rotation $\theta_{i,j}$.

We may consider the finite difference discretization of (16) by applying the standard central differencing to $u_{zz}(\mathbf{x}_{i,j})$ and $u_{ww}(\mathbf{x}_{i,j})$. For instance, we approximate $u_{zz}(\mathbf{x}_{i,j})$ by $\frac{1}{h^2}\left[u(\mathbf{x}_{i,j}+h(\mathbf{e}_z)_{i,j}) - 2u_{i,j} + u(\mathbf{x}_{i,j}-h(\mathbf{e}_z)_{i,j})\right]$. However, since the stencil is rotated, the stencil points $\mathbf{x}_{i,j}\pm h(\mathbf{e}_z)_{i,j}$ may no longer coincide with any grid points. In such cases, bilinear interpolation from the neighboring grid points can be used to approximate $u(\mathbf{x}_{i,j}\pm h(\mathbf{e}_z)_{i,j})$. However, a consequence of the bilinear interpolation is that the truncation error of this central difference approximation becomes $O(1)$ if the stencil length is h. In order to maintain consistency, we choose the stencil length $\sqrt{h}$, which yields $O(h)$ truncation error. Note that when h is small, $\sqrt{h} > h$, which means the stencil length appears to be wide. The details of the discretization is explained in Fig. 1 (left). As a result, the finite difference discretization for $u_{zz}(\mathbf{x}_{i,j})$ and $u_{ww}(\mathbf{x}_{i,j})$ is given by

$$(\delta_{zz}u)_{i,j} \equiv \frac{\mathcal{I}_h u|_{\mathbf{x}_{i,j}+\sqrt{h}(\mathbf{e}_z)_{i,j}} - 2u_{i,j} + \mathcal{I}_h u|_{\mathbf{x}_{i,j}-\sqrt{h}(\mathbf{e}_z)_{i,j}}}{h}, \quad (17)$$

$$(\delta_{ww}u)_{i,j} \equiv \frac{\mathcal{I}_h u|_{\mathbf{x}_{i,j}+\sqrt{h}(\mathbf{e}_w)_{i,j}} - 2u_{i,j} + \mathcal{I}_h u|_{\mathbf{x}_{i,j}-\sqrt{h}(\mathbf{e}_w)_{i,j}}}{h}, \quad (18)$$

where we have used the stencil length $\sqrt{h}$, and used bilinear interpolation to approximate the unknown values at the stencil points $\mathbf{x}_{i,j}\pm\sqrt{h}(\mathbf{e}_z)_{i,j}$ and $\mathbf{x}_{i,j}\pm\sqrt{h}(\mathbf{e}_w)_{i,j}$, denoted as $\mathcal{I}_h u|_{\mathbf{x}_{i,j}\pm\sqrt{h}(\mathbf{e}_z)_{i,j}}$ and $\mathcal{I}_h u|_{\mathbf{x}_{i,j}\pm\sqrt{h}(\mathbf{e}_w)_{i,j}}$. Such discretization scheme is called semi-Lagrangian wide stencil discretization [13,26].

If we apply the semi-Lagrangian wide stencil discretization at a grid point $\mathbf{x}_{i,j}$ that is close to the boundary, some of its associated stencil points may fall outside the computational domain $\overline{\Omega}$. In such case, our solution is to shrink the corresponding stencil length(s) such that the stencil point(s) are relocated onto the boundary $\partial\Omega$. Without loss of generality, we analyze one scenario; see Fig. 1 (right). Let us assume that $\mathbf{x}_{i,j}+\sqrt{h}(\mathbf{e}_z)_{i,j}$ falls outside $\overline{\Omega}$. We truncate the corresponding stencil length from $\sqrt{h}$ to η_1 along the $\mathbf{e}_z$ axis, such that the stencil point is relocated to $\mathbf{x}_{i,j}+\eta_1(\mathbf{e}_z)_{i,j}\in\partial\Omega$. Since $\eta_1\neq\sqrt{h}$, the finite difference approximation for $u_{zz}(\mathbf{x}_{i,j})$ in (17) is replaced by

$$(\delta_{zz}u)_{i,j} \equiv \frac{\frac{g(\mathbf{x}_{i,j}+\eta_1(\mathbf{e}_z)_{i,j})-u_{i,j}}{\eta_1} - \frac{u_{i,j}-\mathcal{I}_h u|_{\mathbf{x}_{i,j}-\sqrt{h}(\mathbf{e}_z)_{i,j}}}{\sqrt{h}}}{\frac{\eta_1+\sqrt{h}}{2}}, \quad (19)$$

where we have used the Dirichlet boundary condition of (1): $u(\mathbf{x}_{i,j}+\eta_1(\mathbf{e}_z)_{i,j}) = g(\mathbf{x}_{i,j}+\eta_1(\mathbf{e}_z)_{i,j})$. We note that such procedure can be used whenever $\mathbf{x}_{i,j}$ is close to the boundary and a truncation of stencil is needed.

4.3 Mixed Discretization

Sections 4.1 and 4.2 describe the standard 7-point stencil and semi-Lagrangian wide stencil finite difference discretization for the HJB Eq. (7). The advantage of the semi-Lagrangian

wide stencil discretization is that it is unconditionally monotone. Reference [14] applies the semi-Lagrangian wide stencil discretization at every grid point. However, it is only first order accurate, while the standard 7-point stencil discretization is second order accurate, as will be proved in Sect. 6. In order to combine the advantages of both discretization schemes, we will only apply the semi-Lagrangian wide stencil discretization at the grid points where neither (12) nor (14) is satisfied. For the other grid points where either (12) or (14) is fulfilled, we will apply the standard 7-point stencil discretization. The purpose is to strictly maintain monotonicity at every grid point and meanwhile to make the numerical scheme as accurate as possible. As a result, the discrete equation at each grid point $\mathbf{x}_{i,j}$ is given by the following mixed scheme:

Standard 7-point stencil discretization. When the control pair $(a_{i,j}, \theta_{i,j})$ satisfies Condition (12) or (14), the discrete equation is given by

$$\max_{(a_{i,j},\theta_{i,j})\in\Gamma} \Big\{ -\alpha_{11}(a_{i,j},\theta_{i,j})(\delta_{xx}u)_{i,j} - 2\alpha_{12}(a_{i,j},\theta_{i,j}) \left(\delta_{xy}^{[disc]}u\right)_{i,j} - \alpha_{22}(a_{i,j},\theta_{i,j})(\delta_{yy}u)_{i,j} + 2\sqrt{a_{i,j}(1-a_{i,j})f_{i,j}} \Big\} = 0, \tag{20}$$

where $disc = 1$ or 2 if (12) or (14) is satisfied respectively.

Semi-Lagrangian wide stencil discretization. Otherwise, the discrete equation is given by

$$\max_{(a_{i,j},\theta_{i,j})\in\Gamma} \Big\{ -a_{i,j}\,(\delta_{zz}u)_{i,j} - \left(1-a_{i,j}\right)\,(\delta_{ww}u)_{i,j} + 2\sqrt{a_{i,j}\left(1-a_{i,j}\right)f_{i,j}} \Big\} = 0, \tag{21}$$

where $(\delta_{zz}u)_{i,j}$ and $(\delta_{ww}u)_{i,j}$ are defined by (17) and (18) when $\mathbf{x}_{i,j}$ is inside the computational domain, and by (19) or similar expressions when $\mathbf{x}_{i,j}$ is near the boundary.

4.4 The Nonlinear Discrete System

The mixed discretization scheme, defined by (20) and (21), gives rise to a nonlinear discrete system that contains N^2 discrete equations. If we define a vector of the unknowns $u_h \equiv (u_{1,1}, u_{1,2}, \ldots, u_{1,N}, u_{2,1}, \ldots, u_{N,N})^T \in \mathbb{R}^{N^2\times 1}$, and similarly, vectors of controls $a_h \in \mathbb{R}^{N^2\times 1}$, $\theta_h \in \mathbb{R}^{N^2\times 1}$, then the entire nonlinear discrete system can be written into the following matrix form:

$$\max_{(a_h,\theta_h)\in\Gamma} \{\mathbf{A}(a_h,\theta_h)\,u_h - F_h(a_h,\theta_h)\} = 0, \tag{22}$$

where $\mathbf{A}(a_h,\theta_h) \in \mathbb{R}^{N^2\times N^2}$ is a matrix that consists of the coefficients of u^h, and $F_h(a_h,\theta_h) \in \mathbb{R}^{N^2\times 1}$ is a vector that does not explicitly contain u^h. We note that this nonlinear system can be treated as a combination of an optimization problem and a linear system as follows:

$$\mathcal{F}_h(u_h) \equiv \max_{(a_h,\theta_h)\in\Gamma} \mathcal{L}_h(a_h,\theta_h;u_h) = 0, \tag{23}$$

where the to-be-maximized linear system is

$$\mathcal{L}_h(a_h,\theta_h;u_h) \equiv \mathbf{A}(a_h,\theta_h)\,u_h - F_h(a_h,\theta_h). \tag{24}$$

Here the symbols $\mathcal{F}_h$ and $\mathcal{L}_h$ in (23)–(24) represent the discretization of $\mathcal{F}$ and $\mathcal{L}$ in (9)–(10), respectively.

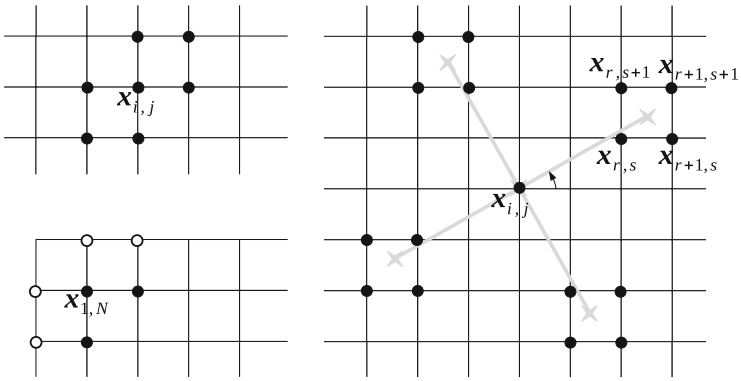


Fig. 2 (left-top) Case 1: Suppose Condition (12) is satisfied at $\mathbf{x}_{i,j}$ and the standard 7-point stencil discretization (20) is used. The discrete equation contains 7 unknown values of u_h, labelled by the black dots. (left-bottom) Case 2: Consider $\mathbf{x}_{1,N}$, which is close to the boundary. The hollow dots sit on the boundary and the values of u on these points are determined by the Dirichlet boundary condition. As a result, the discrete equation contains 3 unknown values of u_h, labeled by the black dots. (right) Case 3: Suppose neither (12) nor (14) is satisfied at $\mathbf{x}_{i,j}$ and thus semi-Lagrangian wide stencil discretization (21) is used. Since bilinear interpolation of each stencil point contains 4 unknown values, the resulting discrete equation has 17 unknown values in total (black dots)

To show how the standard 7-point stencil discretization (20) and the semi-Lagrangian wide stencil discretization (21) can be written into the general form (22), we analyze four cases.

Standard 7-point stencil discretization, grid point $\mathbf{x}_{i,j}$ *inside* Ω. Suppose Condition (12) is satisfied at $\mathbf{x}_{i,j}$. Then we use the standard 7-point stencil discretization (20) with $disc = 1$. This is illustrated in Fig. 2 (left-top). Some simple algebra shows that (20) can be transformed into (22) where

$$\begin{aligned}(\mathbf{A}u_h)_{i,j} &= \frac{2}{h^2}(\alpha_{11}+\alpha_{22}-\alpha_{12})u_{i,j} - \frac{1}{h^2}(\alpha_{11}-\alpha_{12})u_{i+1,j} \\ &\quad - \frac{1}{h^2}(\alpha_{11}-\alpha_{12})u_{i-1,j} - \frac{1}{h^2}(\alpha_{22}-\alpha_{12})u_{i,j+1} \\ &\quad - \frac{1}{h^2}(\alpha_{22}-\alpha_{12})u_{i,j-1} - \frac{1}{h^2}\alpha_{12}\,u_{i+1,j+1} - \frac{1}{h^2}\alpha_{12}\,u_{i-1,j-1}, \\ F_{i,j} &= -2\sqrt{a_{i,j}(1-a_{i,j})f_{i,j}}, \end{aligned} \tag{25}$$

where $(\mathbf{A}u_h)_{i,j}$ and $F_{i,j}$ are the values of $\mathbf{A}(a_h,\theta_h)u_h$ and $F_h(a_h,\theta_h)$ at the grid point $\mathbf{x}_{i,j}$. For simplicity, we have suppressed the dependency of $\mathbf{A}$, $F_{i,j}$, α_{11}, α_{22} and α_{12} on $(a_{i,j},\theta_{i,j})$. This equation contains 7 unknown values of u_h. Similarly, interested readers can also write down the expressions when Condition (14) is satisfied at $\mathbf{x}_{i,j}$ and the standard 7-point stencil discretization (20) with $disc = 2$ is applied.

Standard 7-point stencil discretization, grid point $\mathbf{x}_{i,j}$ *near* $\partial\Omega$. Without loss of generality, we assume that $\mathbf{x}_{i,j} = \mathbf{x}_{1,N}$, as shown in Fig. 2 (left-bottom). Now $u_{i-1,j}$, $u_{i,j+1}$, $u_{i+1,j+1}$ and $u_{i-1,j-1}$ can be determined by the Dirichlet boundary condition $u = g$. These terms become part of $F_{i,j}$. As a result, $(\mathbf{A}u_h)_{i,j}$ contains only 3 unknown values.

Semi-Lagrangian wide stencil discretization, grid point $\mathbf{x}_{i,j}$ *inside* Ω. Suppose neither (12) nor (14) is fulfilled at $\mathbf{x}_{i,j}$, so semi-Lagrangian wide stencil discretization (21) is applied;

see Fig. 2 (right). Then (21) can be written into (22) where

$$\begin{aligned}(\mathbf{A}u_h)_{i,j} &= \frac{2}{h}u_{i,j} - \frac{a_{i,j}}{h}\,\mathcal{I}_h u|_{\mathbf{x}_{i,j}+\sqrt{h}(\mathbf{e}_z)_{i,j}} - \frac{a_{i,j}}{h}\,\mathcal{I}_h u|_{\mathbf{x}_{i,j}-\sqrt{h}(\mathbf{e}_z)_{i,j}} \\ &\quad - \frac{1-a_{i,j}}{h}\,\mathcal{I}_h u|_{\mathbf{x}_{i,j}+\sqrt{h}(\mathbf{e}_w)_{i,j}} - \frac{1-a_{i,j}}{h}\,\mathcal{I}_h u|_{\mathbf{x}_{i,j}-\sqrt{h}(\mathbf{e}_w)_{i,j}}, \\ F_{i,j} &= -2\sqrt{a_{i,j}(1-a_{i,j})f_{i,j}}. \end{aligned} \tag{26}$$

We note that each bilinear interpolation term contains 4 unknowns. For instance, $\mathcal{I}_h u|_{\mathbf{x}_{i,j}+\sqrt{h}(\mathbf{e}_z)_{i,j}}$ can be written as the linear combination of the unknowns at the four neighboring points $u_{r,s}$, $u_{r+1,s}$, $u_{r,s+1}$ and $u_{r+1,s+1}$, which are labeled in Fig. 2 (right). As a result, (26) has 17 unknown values.

Semi-Lagrangian wide stencil discretization, grid point $\mathbf{x}_{i,j}$ *near* $\partial\Omega$. The analysis is similar to the previous cases. The number of the unknowns is less than 17.

5 Solving the Nonlinear Discrete System

5.1 Policy Iteration

After setting up the complete nonlinear discrete system (23)–(24), the next objective is to solve it. We apply a well-known fixed point iteration algorithm, called policy iteration (or Howard's algorithm) [16,21] as follows:

1. Start with an initial guess of the solution $u_h^{(0)}$.
2. For $k = 0, 1, \ldots$ until convergence:

   (a) Solve for the optimal control pair $(a_h^{(k)}, \theta_h^{(k)})$ under the current solution $u_h^{(k)}$:

   $$(a_{i,j}^{(k)}, \theta_{i,j}^{(k)}) = \arg\max_{(a_{i,j},\theta_{i,j})\in\Gamma_{i,j}} \mathcal{L}_{i,j}(a_{i,j}, \theta_{i,j}; u_h^{(k)}), \quad \text{for all } \mathbf{x}_{i,j} \in \Omega, \tag{27}$$

   where $\mathcal{L}_{i,j}$ is the pointwise component of $\mathcal{L}_h \in \mathbb{R}^{N^2\times 1}$ defined in (24) and $\Gamma_{i,j} = [0,1]\times[-\frac{\pi}{4}, \frac{\pi}{4}]$ is the control set at $\mathbf{x}_{i,j}$.

   Meanwhile, obtain the residual $R_h^{(k)} \in \mathbb{R}^{N^2\times 1}$, where each pointwise component reads $R_{i,j}^{(k)} \equiv \mathcal{L}_{i,j}(a_{i,j}^{(k)}, \theta_{i,j}^{(k)}; u_h^{(k)})$.

   (b) If $\|R_h^{(k)}\| \le$ tolerance: break

   Else, solve the following linear system for the solution $u_h^{(k+1)}$ under the current optimal control pair $(a_h^{(k)}, \theta_h^{(k)})$:

   $$\mathbf{A}(a_h^{(k)}, \theta_h^{(k)})\, u_h^{(k+1)} = F_h(a_h^{(k)}, \theta_h^{(k)}) \quad \Rightarrow \quad u_h^{(k+1)}. \tag{28}$$

It is proved that policy iteration is guaranteed to converge for any initial guess $u_h^{(0)}$, if by applying a monotone discretization to an HJB equation, the resulting matrix $\mathbf{A}(a_h, \theta_h)$ is an M-matrix under all admissible controls [1,6]. We will show in Sect. 6.2 that the resulting matrix $\mathbf{A}(a_h, \theta_h)$ in (22) is indeed an M-matrix.

Policy iteration consists of two sub-steps. One sub-step is to solve the linear system under a given control pair; see (28). We use Krylov subspace methods, such as the GMRES with the incomplete LU preconditioner. The other sub-step of the policy iteration is to solve the optimization problem at each grid point $\mathbf{x}_{i,j}$; see (27). We will discuss speeding up computation of the optimization problem in detail in the next section.

5.2 Speeding Up Computation of Optimal Controls

Since the semi-Lagrangian wide stencil discretization of $(\delta_{zz}u)_{i,j}$ and $(\delta_{ww}u)_{i,j}$ in (21) depends on the control $\theta_{i,j}$, there is no simple closed-form formula to evaluate the optimal $(a_{i,j}^{(k)}, \theta_{i,j}^{(k)})$ directly. In this case, one typical approach is to use bilinear search algorithm for the optimization problem. More specifically, consider the optimization problem at a grid point $\mathbf{x}_{i,j}$. We discretize the continuous admissible control set $\Gamma_{i,j} = [0,1] \times [-\frac{\pi}{4}, \frac{\pi}{4})$ into an $M \times M$ discrete set, denoted as $\Gamma_{i,j}^h$. We note that the discretization of the control set introduces additional truncation error. In order to maintain consistency, we must let $M \to \infty$ as $h \to 0$. A typical choice of M is $M = N$. Then we compute the $M \times M$ values of the objective function $\mathcal{L}_{i,j}(a_{i,j}, \theta_{i,j}; u_h^{(k)})$ with $(a_{i,j}, \theta_{i,j}) \in \Gamma_{i,j}^h$ and then find the global maximal value, which gives the optimal $(a_{i,j}^{(k)}, \theta_{i,j}^{(k)})$. However, the computational cost of the bilinear search per grid point $\mathbf{x}_{i,j}$ is $O(M^2)$. Furthermore, if we denote the total number of grid points as $\#\Omega = N^2$, then the computational cost on the entire computational domain Ω is as high as $O(M^2\#\Omega)$, or $O(\#\Omega^2)$ if we choose $M = N$.

In order to speed up computation for the optimal controls, we divide the continuous admissible control set $\Gamma_{i,j} = [0,1] \times [-\frac{\pi}{4}, \frac{\pi}{4})$ into six regions, as shown in Fig. 3.[2] The six regions are identified by whether a control pair $(a_{i,j}, \theta_{i,j})$ satisfies (12), or (14), or neither. Our approach is to find the optimal control pair within each region, and then find the global optimal control pair among the six regional optimal control pairs. This approach enables us to make full use of the analytical property of each region, and to improve the optimization algorithm within each region and eventually on the entire admissible control set $\Gamma_{i,j}$.

Using our approach, the computational cost of solving the optimization problem on $\Gamma_{i,j}$ can be significantly reduced. More specifically, if the standard 7-point stencil discretization can be applied monotonically on all or most of the grid points, then the computational cost is $O(1)$ per grid point and $O(\#\Omega)$ on the entire computational domain. In general, the computational cost is at most $O(M)$ per grid point and at most $O(M\#\Omega)$ on the entire computational domain. For the typical choice $M = N$, the total computational cost of solving the optimization problem is $O(\#\Omega^{3/2})$.

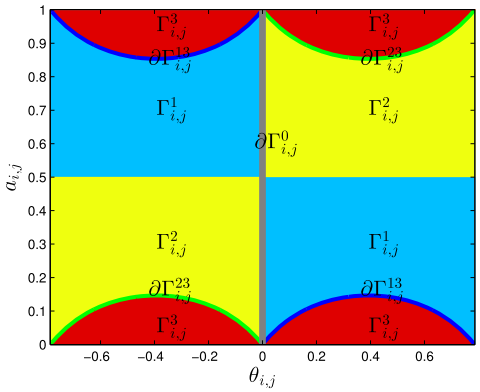
To explain the details of the regional optimization, consider again a given grid point $\mathbf{x}_{i,j}$ and its associated control set $\Gamma_{i,j}$. In Region $\Gamma_{i,j}^1$, $\Gamma_{i,j}^2$, $\partial\Gamma_{i,j}^0$, $\partial\Gamma_{i,j}^{13}$ and $\partial\Gamma_{i,j}^{23}$ (see Fig. 3), where the standard 7-point stencil discretization (20) is applied, the discretization of $(\delta_{xx}u)_{i,j}$, $(\delta_{yy}u)_{i,j}$, $(\delta_{xy}^{[1]}u)_{i,j}$ and $(\delta_{xy}^{[2]}u)_{i,j}$ does not depend on the controls $(a_{i,j}, \theta_{i,j})$. This enables us to derive a closed-form formula for the optimal controls in these regions using first derivative test, which can be evaluated by $O(1)$ operation and introduces no additional truncation error. More specifically:

Region $\Gamma_{i,j}^1$. The region is defined where Condition (12) is satisfied. Equation (20) gives the objective function in $\Gamma_{i,j}^1$:

$$\begin{aligned}\mathcal{L}_{i,j}(a_{i,j}, \theta_{i,j}) = &-\alpha_{11}(a_{i,j}, \theta_{i,j})(\delta_{xx}u)_{i,j} - 2\alpha_{12}(a_{i,j}, \theta_{i,j})(\delta_{xy}^{[1]}u)_{i,j} \\ &-\alpha_{22}(a_{i,j}, \theta_{i,j})(\delta_{yy}u)_{i,j} + 2\sqrt{a_{i,j}(1-a_{i,j})f_{i,j}},\end{aligned} \tag{29}$$

where we only manifest the dependency of $\mathcal{L}_{i,j}$ on the control pair $(a_{i,j}, \theta_{i,j})$. One can verify that this function is smooth in $(a_{i,j}, \theta_{i,j}) \in \Gamma_{i,j}^1$, concave in $a_{i,j} \in [0,1]$, and its stationary

[2] It is unnecessary to consider the line $a_{i,j} = \frac{1}{2}$, since the objective function is a constant on this line. Also it is unnecessary to consider the line $\theta_{i,j} = \pm\frac{\pi}{4}$, since $\mathcal{L}_{a,\theta}\, u = \mathcal{L}_{1-a,\theta+\frac{\pi}{2}}\, u$ indicates that $\theta_{i,j} = \pm\frac{\pi}{4}$ is indeed an interior part of $\Gamma_{i,j}^1$ and $\Gamma_{i,j}^2$.



Region	Definition	Discretization	Optimization algorithm in each region	Cost	Extra truncation error introduced?
$\Gamma^1_{i,j}$	The region where Condition (12) is satisfied	Standard 7-point stencil with $disc = 1$	Closed-form formula from first derivative test	$O(1)$	No
$\Gamma^2_{i,j}$	The region where Condition (14) is satisfied	Standard 7-point stencil with $disc = 2$			
$\Gamma^3_{i,j}$	The region where neither (12) nor (14) is satisfied	Semi-Lagrangian wide stencil	Linear search over a single control $\theta_{i,j} \in [-\frac{\pi}{4}, \frac{\pi}{4})$	$O(M)$	Yes
$\partial\Gamma^0_{i,j}$	The line $\theta_{i,j} = 0$	Standard 7-point stencil with $disc = 1$ or 2	Closed-form formula from first derivative test	$O(1)$	No
$\partial\Gamma^{13}_{i,j}$	The boundary between $\Gamma^1_{i,j}$ and $\Gamma^3_{i,j}$	Standard 7-point stencil with $disc = 1$			
$\partial\Gamma^{23}_{i,j}$	The boundary between $\Gamma^2_{i,j}$ and $\Gamma^3_{i,j}$	Standard 7-point stencil with $disc = 2$			

Fig. 3 Division of the admissible control set $\Gamma_{i,j} = [0, 1] \times [-\frac{\pi}{4}, \frac{\pi}{4})$ into regions. For each region, the characterization, discretization, optimization algorithm and the corresponding cost/truncation error of the optimization algorithm are listed

point in $\Gamma^1_{i,j}$ is unique, if it exists. This allows us to use first derivative test to find the optimal control pair in $\Gamma^1_{i,j}$:

$$\theta^*_{i,j} = \frac{1}{2}\arctan\frac{2(\delta^{[1]}_{xy}u)_{i,j}}{(\delta_{yy}u)_{i,j} - (\delta_{xx}u)_{i,j}}, \quad a^*_{i,j} = \frac{1}{2}\left(1 - \frac{\lambda_{i,j}}{\sqrt{4f_{i,j} + \lambda^2_{i,j}}}\right), \tag{30}$$

where $\lambda_{i,j} \equiv [(\delta_{xx}u)_{i,j} - (\delta_{yy}u)_{i,j}]\cos 2\theta^*_{i,j} - 2(\delta^{[1]}_{xy}u)_{i,j}\sin 2\theta^*_{i,j}$. With a slight abuse of notations, here and for the rest of Sect. 5.2, we use $(a^*_{i,j}, \theta^*_{i,j})$ to denote the the regional (rather than global) optimal control pair at $\mathbf{x}_{i,j}$. We note that $(a^*_{i,j}, \theta^*_{i,j})$ given by (30) may not necessarily be inside $\Gamma^1_{i,j}$. If $(a^*_{i,j}, \theta^*_{i,j}) \in \Gamma^1_{i,j}$, then the maximum in $\Gamma^1_{i,j}$ must occur at

$(a^*_{i,j}, \theta^*_{i,j})$. Otherwise, the maximum must occur on the boundary of $\Gamma^1_{i,j}$, or more specifically, either $\partial\Gamma^0_{i,j}$ or $\partial\Gamma^{13}_{i,j}$, which will be investigated separately.

Region $\Gamma^2_{i,j}$. The region is defined where Condition (14) is satisfied. The analysis for solving the optimization problem in $\Gamma^2_{i,j}$ is the same as $\Gamma^1_{i,j}$, except that $(\delta^{[1]}_{xy}u)_{i,j}$ in (29), (30) is replaced by $(\delta^{[2]}_{xy}u)_{i,j}$.

Region $\partial\Gamma^0_{i,j}$. This is the line $\theta_{i,j} = 0$ which separates Region $\Gamma^1_{i,j}$ and $\Gamma^2_{i,j}$. The objective function in $\partial\Gamma^0_{i,j}$ can be found in (29), where $\alpha_{12} = 0$ and thus the cross derivative term disappears. The optimal control pair in $\partial\Gamma^0_{i,j}$ is simply

$$\theta^*_{i,j} = 0, \quad a^*_{i,j} = \frac{1}{2}\left[1 - \frac{(\delta_{xx}u)_{i,j} - (\delta_{yy}u)_{i,j}}{\sqrt{4f_{i,j} + ((\delta_{xx}u)_{i,j} - (\delta_{yy}u)_{i,j})^2}}\right]. \tag{31}$$

Region $\partial\Gamma^{13}_{i,j}$. This is the boundary between Region $\Gamma^1_{i,j}$ and $\Gamma^3_{i,j}$. If we define the signs of $a_{i,j} - \frac{1}{2}$ and $\theta_{i,j}$ as

$$s_{a-1/2} \equiv \begin{cases} -1, & a_{i,j} - \frac{1}{2} < 0, \\ 1, & a_{i,j} - \frac{1}{2} > 0, \end{cases} \quad s_\theta \equiv \begin{cases} -1, & \theta_{i,j} < 0, \\ 1, & \theta_{i,j} > 0, \end{cases} \tag{32}$$

then $\partial\Gamma^{13}_{i,j}$ contains two sections: (i) $(s_{a-1/2}, s_\theta) = (1, -1)$, (ii) $(s_{a-1/2}, s_\theta) = (-1, 1)$.

The objective function on $\partial\Gamma^{13}_{i,j}$ is the same as (29). First derivative test shows that for each of the two sections of $\partial\Gamma^{13}_{i,j}$, the maximum of the objective function occurs at

$$\theta^*_{i,j} = \frac{s_\theta}{2}\arctan\left(1 + \gamma^2_{i,j} - \gamma_{i,j}\sqrt{2 + \gamma^2_{i,j}}\right), \tag{33}$$

where $\gamma_{i,j} \equiv \frac{s_{a-1/2}}{2\sqrt{f_{i,j}}}\left((\delta_{yy}u)_{i,j} - (\delta_{xx}u)_{i,j} - 2s_\theta(\delta^{[1]}_{xy}u)_{i,j}\right)$. The corresponding $a^*_{i,j} \in \partial\Gamma^{13}_{i,j}$, derived from Condition (12), is

$$a^*_{i,j} = \frac{1}{2}\left(1 + \frac{s_{a-1/2}}{\sqrt{2}\sin(2|\theta^*_{i,j}| + \frac{\pi}{4})}\right). \tag{34}$$

Region $\partial\Gamma^{23}_{i,j}$. This is the boundary between Region $\Gamma^2_{i,j}$ and $\Gamma^3_{i,j}$. The analysis on $\partial\Gamma^{23}_{i,j}$ is then the same as $\partial\Gamma^{13}_{i,j}$, except that the two sections of $\partial\Gamma^{23}_{i,j}$ become (i) $(s_{a-1/2}, s_\theta) = (1, 1)$, (ii) $(s_{a-1/2}, s_\theta) = (-1, -1)$, and $(\delta^{[1]}_{xy}u)_{i,j}$ is replaced by $(\delta^{[2]}_{xy}u)_{i,j}$.

Region $\Gamma^3_{i,j}$. The region is defined where neither (12) nor (14) is satisfied. The semi-Lagrangian wide stencil discretization (21) is applied. Accordingly, the objective function reads

$$\mathcal{L}_{i,j}(a_{i,j}, \theta_{i,j}) = -a_{i,j}\,(\delta_{zz}u)_{i,j} - (1 - a_{i,j})\,(\delta_{ww}u)_{i,j} + 2\sqrt{a_{i,j}(1 - a_{i,j})\,f_{i,j}}. \tag{35}$$

The dependency of the discretization of $(\delta_{zz}u)_{i,j}$ and $(\delta_{ww}u)_{i,j}$ on the control $\theta_{i,j}$ prevents us from deriving a closed-form formula for $\theta^*_{i,j} \in \Gamma^3_{i,j}$. However, we note that the discretization of $(\delta_{zz}u)_{i,j}$ and $(\delta_{ww}u)_{i,j}$ is independent of the control $a_{i,j}$, which implies that a two dimensional bilinear search on the controls $(a_{i,j}, \theta_{i,j}) \in \Gamma_{i,j}$ can be reduced to a one-dimensional linear search on the single control $\theta_{i,j} \in [-\frac{\pi}{4}, \frac{\pi}{4})$.

One can prove that the regional optimal control pair $(a^*_{i,j}, \theta^*_{i,j}) \in \Gamma^3_{i,j}$ must sit on the following parametrized curve

$$a_{i,j}(\theta_{i,j}) = \begin{cases} \mathcal{C}^\lambda(\theta_{i,j}), & \text{if} \quad \mathcal{C}^\lambda(\theta_{i,j}) \le \mathcal{C}^-(\theta_{i,j}) \text{ or } \mathcal{C}^\lambda(\theta_{i,j}) \ge \mathcal{C}^+(\theta_{i,j}), \\ \mathcal{C}^-(\theta_{i,j}), & \text{if} \quad \mathcal{C}^-(\theta_{i,j}) \le \mathcal{C}^\lambda(\theta_{i,j}) \le \frac{1}{2}, \\ \mathcal{C}^+(\theta_{i,j}), & \text{if} \quad \frac{1}{2} \le \mathcal{C}^\lambda(\theta_{i,j}) \le \mathcal{C}^+(\theta_{i,j}). \end{cases} \tag{36}$$

Here the curves

$$\mathcal{C}^\pm(\theta_{i,j}) \equiv \frac{1}{2}\left(1 \pm \frac{1}{\sqrt{2}\sin(2|\theta_{i,j}| + \frac{\pi}{4})}\right), \quad \theta_{i,j} \in \left[-\frac{\pi}{4}, \frac{\pi}{4}\right) \tag{37}$$

are given by Condition (12) and (14). The other curve

$$\mathcal{C}^\lambda(\theta_{i,j}) \equiv \frac{1}{2}\left[1 - \frac{(\delta_{zz}u)_{i,j} - (\delta_{ww}u)_{i,j}}{\sqrt{4f_{i,j} + ((\delta_{zz}u)_{i,j} - (\delta_{ww}u)_{i,j})^2}}\right], \quad \theta_{i,j} \in \left[-\frac{\pi}{4}, \frac{\pi}{4}\right), \tag{38}$$

where the directions of z and w depend on $\theta_{i,j}$, is given by the first derivative test of (35) with respect to $a_{i,j}$. Taking the parametrization (36) into account, the objective function (35) becomes $\mathcal{L}_{i,j}(a_{i,j}(\theta_{i,j}), \theta_{i,j})$, which is a function of the single control variable $\theta_{i,j} \in [-\frac{\pi}{4}, \frac{\pi}{4})$. This motivates us to discretize the set $[-\frac{\pi}{4}, \frac{\pi}{4})$ into an M-element control set, and perform a linear search for the maximum of the parametrized objective function $\mathcal{L}_{i,j}(a_{i,j}(\theta_{i,j}), \theta_{i,j})$ over the single control variable $\theta_{i,j} \in [-\frac{\pi}{4}, \frac{\pi}{4})$. The computational cost is thus reduced to $O(M)$.

Once we obtain the six regional optimal control pairs and their corresponding objective function values, we search within them for the global optimal control pair on $\Gamma_{i,j}$. This step is cheap and straightforward.

As a side remark, in Section 8 of [14], the authors discretize θ with 64 different angles, regardless of the mesh size N. Indeed, if θ is discretized with fixed number of angles, then the numerical scheme in [14] is no longer consistent in theory. This is different from our scheme, where θ is discretized with M angles, and we choose $M = N$ such that consistency is still maintained.

6 Convergence Analysis

As proved by Barles and Souganidis [2], there are four sufficient conditions for the numerical scheme of a nonlinear PDE to converge in the viscosity sense. In this section, we will prove that our numerical scheme does fulfill all the four requirements and is therefore guaranteed to converge to the viscosity solution of (2).

6.1 Consistency

One sufficient condition for convergence is consistency. Intuitively, consistency claims that the discretized equation of a PDE should be close to the continuous PDE. In particular, when $h \to 0$, the discretized equation should converge to the PDE. The main result of this subsection is to prove that our numerical scheme is consistent in the viscosity sense:

Lemma 1 (Consistency) *For the Monge–Ampère equation* $\mathcal{F}\left(\mathbf{x}, u(\mathbf{x}), D^2u(\mathbf{x})\right) = 0$, *the numerical scheme* $\mathcal{F}_h\left(\mathbf{x}_{i,j}, u_h\right) = 0$, *given in* (23)–(24), *is consistent in the viscosity sense. More specifically, for any function* $\varphi(\mathbf{x}) \in C^\infty(\overline{\Omega})$ *with* $\varphi_{i,j} \equiv \varphi(\mathbf{x}_{i,j})$ *and*

$\varphi_h \equiv (\varphi_{1,1}, \varphi_{1,2}, \ldots, \varphi_{N,N})^T \in \mathbb{R}^{N^2\times 1}$, for any $\hat{\mathbf{x}} \in \overline{\Omega}$, and for h and ξ that are arbitrary small constants independent of $\mathbf{x}$, we have

$$\limsup_{\substack{h\to 0,\,\xi\to 0\\ \mathbf{x}_{i,j}\to\hat{\mathbf{x}}}} \mathcal{F}_h(\mathbf{x}_{i,j}, \varphi_h + \xi) \le \mathcal{F}^*\left(\hat{\mathbf{x}}, \varphi(\hat{\mathbf{x}}), D^2\varphi(\hat{\mathbf{x}})\right), \tag{39}$$

$$\liminf_{\substack{h\to 0,\,\xi\to 0\\ \mathbf{x}_{i,j}\to\hat{\mathbf{x}}}} \mathcal{F}_h(\mathbf{x}_{i,j}, \varphi_h + \xi) \ge \mathcal{F}_*\left(\hat{\mathbf{x}}, \varphi(\hat{\mathbf{x}}), D^2\varphi(\hat{\mathbf{x}})\right). \tag{40}$$

In practise, we prove a sufficient condition for consistency, called local consistency, as follows:

Lemma 2 (Local consistency) *Under the assumptions in Lemma 1, we have*

$$\mathcal{F}(\mathbf{x}_{i,j}, \varphi(\mathbf{x}_{i,j}), D^2\varphi(\mathbf{x}_{i,j})) - \mathcal{F}_h(\mathbf{x}_{i,j}, \varphi_h + \xi) = \begin{cases} O(h^2) + O(\xi), & \text{standard 7-point stencil,}\\ O(h) + O(\xi), & \text{semi-Lagrangian wide stencil, with all the 4}\\ & \text{wide stencil points } \in \Omega,\\ O(\sqrt{h}) + O(\xi), & \text{semi-Lagrangian wide stencil, otherwise.}\end{cases} \tag{41}$$

Proof We note that the proof with $\xi = 0$ is equivalent to the proof with a general ξ. Such equivalence can be easily verified if we substitute φ by $\varphi + \xi$ in the following proof. Hence, we will only prove the case where $\xi = 0$.

Truncation error of the standard 7-point stencil discretization. Suppose the standard 7-point stencil discretization is applied at $\mathbf{x}_{i,j}$. It is easy to show that the truncation errors for $(\delta_{xx}\varphi)_{i,j}$, $(\delta_{yy}\varphi)_{i,j}$, $(\delta^{[1]}_{xy}\varphi)_{i,j}$ and $(\delta^{[2]}_{xy}\varphi)_{i,j}$ are all $O(h^2)$. Hence, the local truncation error of the discrete linear Eq. (24) is then $\mathcal{L}_{a(\mathbf{x}_{i,j}),\theta(\mathbf{x}_{i,j})}\varphi(\mathbf{x}_{i,j}) - \mathcal{L}_h(\mathbf{x}_{i,j}; a_{i,j}, \theta_{i,j}; \varphi_h) = O(h^2)$. Furthermore, the local truncation error of the finite difference scheme at $\mathbf{x}_{i,j}$ is

$$\begin{aligned} &\left|\, \mathcal{F}(\mathbf{x}_{i,j}, \varphi(\mathbf{x}_{i,j}), D^2\varphi(\mathbf{x}_{i,j})) - \mathcal{F}_h(\mathbf{x}_{i,j}, \varphi_h) \,\right| \\ &= \left| \max_{(a(\mathbf{x}_{i,j}),\theta(\mathbf{x}_{i,j}))\in\Gamma} \mathcal{L}_{a(\mathbf{x}_{i,j}),\theta(\mathbf{x}_{i,j})}\varphi(\mathbf{x}_{i,j}) - \max_{(a_{i,j},\theta_{i,j})\in\Gamma} \mathcal{L}_h(\mathbf{x}_{i,j}; a_{i,j}, \theta_{i,j}; \varphi_h) \right| \\ &\le \max_{(a_{i,j},\theta_{i,j})\in\Gamma} \left|\, \mathcal{L}_{a_{i,j},\theta_{i,j}}\varphi(\mathbf{x}_{i,j}) - \mathcal{L}_h(\mathbf{x}_{i,j}; a_{i,j}, \theta_{i,j}; \varphi_h) \,\right| = O(h^2). \end{aligned} \tag{42}$$

The inequality comes from $\left|\max_x f(x) - \max_x g(x)\right| \le \max_x |f(x) - g(x)|$.

Truncation error of semi-Lagrangian wide stencil discretization. Suppose semi-Lagrangian wide stencil discretization is applied at $\mathbf{x}_{i,j}$. We focus on the truncation error for $(\delta_{zz}\varphi)_{i,j}$ only and analyze three cases. The first case is that both stencil points of $(\delta_{zz}\varphi)_{i,j}$ are in the computational domain. The expression for $(\delta_{zz}\varphi)_{i,j}$ is given by (17). The truncation error for $(\delta_{zz}\varphi)_{i,j}$ is then

$$\begin{aligned} &\varphi_{zz}(\mathbf{x}_{i,j}) - (\delta_{zz}\varphi)_{i,j} \\ &= \varphi_{zz}(\mathbf{x}_{i,j}) - \frac{\mathcal{I}_h\varphi|_{\mathbf{x}_{i,j}+\sqrt{h}(\mathbf{e}_z)_{i,j}} - 2\varphi_{i,j} + \mathcal{I}_h\varphi|_{\mathbf{x}_{i,j}-\sqrt{h}(\mathbf{e}_z)_{i,j}}}{h} \\ &= \varphi_{zz}(\mathbf{x}_{i,j}) - \frac{\varphi(\mathbf{x}_{i,j} + \sqrt{h}(\mathbf{e}_z)_{i,j}) - 2\varphi(\mathbf{x}_{i,j}) + \varphi(\mathbf{x}_{i,j} - \sqrt{h}(\mathbf{e}_z)_{i,j}) + O(h^2)}{h} \\ &= O(h) + O(h) = O(h). \end{aligned}$$

From the first to the second line we have used the fact that the truncation error of the bilinear interpolation is $O(h^2)$.

Now we consider another case, where one of the stencil points of $(\delta_{zz}\varphi)_{i,j}$ falls outside the computational domain and is thus relocated. Without loss of generality, let us assume again that $\mathbf{x}_{i,j} + \eta_1(\mathbf{e}_z)_{i,j} \in \partial\Omega$ is the relocated point. The expression for $(\delta_{zz}\varphi)_{i,j}$ is given by (19). The truncation error for $(\delta_{zz}\varphi)_{i,j}$ is then

$$\begin{aligned}
&\varphi_{zz}(\mathbf{x}_{i,j}) - (\delta_{zz}\varphi)_{i,j} \\
&= \varphi_{zz}(\mathbf{x}_{i,j}) - \frac{\frac{\varphi(\mathbf{x}_{i,j}+\eta_1(\mathbf{e}_z)_{i,j})-\varphi_{i,j}}{\eta_1} - \frac{\varphi_{i,j}-\mathcal{I}_h\varphi|_{\mathbf{x}_{i,j}-\sqrt{h}(\mathbf{e}_z)_{i,j}}}{\sqrt{h}}}{\frac{\eta_1+\sqrt{h}}{2}} \\
&= \varphi_{zz}(\mathbf{x}_{i,j}) - \frac{\frac{\varphi(\mathbf{x}_{i,j}+\eta_1(\mathbf{e}_z)_{i,j})-\varphi(\mathbf{x}_{i,j})}{\eta_1} - \frac{\varphi(\mathbf{x}_{i,j})-\varphi(\mathbf{x}_{i,j}-\sqrt{h}(\mathbf{e}_z)_{i,j})}{\sqrt{h}} + O(h^2)}{\frac{\eta_1+\sqrt{h}}{2}} \\
&= O(\sqrt{h}-\eta_1) + O\left(\frac{h^2}{\sqrt{h}\frac{\eta_1+\sqrt{h}}{2}}\right) = O(\sqrt{h}).
\end{aligned}$$

There is one more case, where $\mathbf{x}_{i,j} + \eta_1(\mathbf{e}_z)_{i,j} \in \partial\Omega$ and $\mathbf{x}_{i,j} - \eta_2(\mathbf{e}_z)_{i,j} \in \partial\Omega$ are both relocated points. Using the similar argument, one can show that the truncation error for $(\delta_{zz}\varphi)_{i,j}$ is again $O(\sqrt{h})$.

Then, similar to (42), one can show that the local truncation error of the finite difference scheme at $\mathbf{x}_{i,j}$, where the semi-Lagrangian wide stencil discretization is applied, is given by

$$\begin{aligned}
&\left| \mathcal{F}(\mathbf{x}_{i,j}, \varphi(\mathbf{x}_{i,j}), D^2\varphi(\mathbf{x}_{i,j})) - \mathcal{F}_h(\mathbf{x}_{i,j}, \varphi_h) \right| \\
&= \begin{cases} O(h), & \text{semi-Lagrangian wide stencil, with all the 4} \\ & \text{wide stencil points } \in \Omega, \\ O(\sqrt{h}), & \text{semi-Lagrangian wide stencil, otherwise.} \end{cases}
\end{aligned} \tag{43}$$

Finally, we note that the previous proof has assumed that the optimal control pair is solved exactly, or does not introduce additional truncation error. In Sect. 5, we have mentioned that using linear search for the optimal control pair under the semi-Lagrangian wide stencil discretization introduces truncation error. In particular, if we choose $M = O(N)$, then $O(h)$ truncation error is introduced [33]. As a result, (43) holds. □

6.2 Stability

Another condition for convergence is stability, which means that the discrete system has a bounded solution u_h. Stability condition is very closely related to the matrix $\mathbf{A}(a_h, \theta_h)$ in (22) being an M-matrix [29], which will be proved in this section. For convenience, given vectors u_h and v_h, we use $u_h \geq 0$ and $u_h \geq v_h$ to denote $(u_h)_i \geq 0$ and $(u_h)_i \geq (v_h)_i$ for all i. Similarly, given a matrix $\mathbf{A}$, we use $\mathbf{A} \geq 0$ to denote $\mathbf{A}_{ij} \geq 0$ for all i, j. In other words, the inequalities for vectors and matrices hold for all the elements.

Lemma 3 (M-matrix) *Suppose an $n \times n$ matrix $\mathbf{A}$ satisfies the following:*

1. $\mathbf{A}$ *is an L-matrix: $\mathbf{A}_{ii} > 0$ for all i, and $\mathbf{A}_{ij} \leq 0$ for all $i \neq j$;*
2. $\mathbf{A}$ *is weakly diagonally dominant: $|\mathbf{A}_{ii}| \geq \sum_{j\neq i} |\mathbf{A}_{ij}|$; and*
3. $\mathbf{A}$ *has the following connectivity property: Let $\mathcal{G}(\mathbf{A}) = \left\{ i \,\middle|\, |\mathbf{A}_{ii}| > \sum_{j\neq i} |\mathbf{A}_{ij}| \right\} \neq \emptyset$ be the set of rows where strict inequality is achieved. For any $i \notin \mathcal{G}(\mathbf{A})$, there exists a sequence $i_1, i_2, \cdots, i_k$ with $\mathbf{A}_{i_r, i_{r+1}} \neq 0$, $0 \leq r \leq k-1$, such that $i_0 = i$ and $i_k \in \mathcal{G}(\mathbf{A})$.*

Then $\mathbf{A}$ *is an M-matrix. In particular,*

1. $\mathbf{A}$ *is non-singular; and*
2. $\mathbf{A}^{-1} \geq 0$, *namely,* $(\mathbf{A}^{-1})_{ij} \geq 0$ *for all* i, j.

Proof We refer the readers to [1,29,31]. □

Lemma 4 *The matrix* $\mathbf{A}(a_h, \theta_h)$, *defined in* (22), *is an M-matrix under the set of admissible controls* $(a_h, \theta_h) \in \Gamma$.

Proof For the matrix $\mathbf{A}(a_h, \theta_h)$, the L-matrix condition and the weakly diagonal dominance condition can be easily verified by checking the four cases in Sect. 4.4. We remark that the strictly diagonally dominant rows correspond to the grid points near the boundary $\partial\Omega$, while the weakly diagonally dominant rows correspond to those inside the computation domain Ω.

The connectivity property of $\mathbf{A}(a_h, \theta_h)$ is yet to be verified. For the grid points $\mathbf{x}_{i,j}$ that are near the boundary, the lexicographical index satisfies $N(i-1)+j \in \mathcal{G}(\mathbf{A})$. For those points that are inside the computational domain, or $N(i-1)+j \notin \mathcal{G}(\mathbf{A})$, there must exist non-zero entries $\mathbf{A}_{N(i-1)+j, N(i'-1)+j'} \neq 0$, where $i' \geq i$, $j' \geq j$, with at lease one strict inequality satisfied. Hence, given any $\mathbf{x}_{i_0,j_0}$, where $N(i_0-1)+j_0 \notin \mathcal{G}(\mathbf{A})$, there exist monotonically increasing sequences $i_0 \leq i_1 \leq \cdots \leq i_k \leq N$ and $j_0 \leq j_1 \leq \cdots \leq j_k \leq N$, such that $N(i_k-1)+j_k \in \mathcal{G}(\mathbf{A})$. □

Before investigating the stability for the nonlinear problem (22), we first prove the stability for the corresponding linear problem.

Lemma 5 *Define a circle* $B_R(0) : \{(x,y) | x^2+y^2 \leq R^2\}$, *where the radius* $R = \max_{(x,y)\in\overline{\Omega}} \sqrt{x^2+y^2}$, *such that* $B_R(0)$ *covers the entire computational domain* $\overline{\Omega}$. *Let* $\varphi(\mathbf{x}) \equiv -\frac{1}{2}\|\sqrt{f}\|_\infty (R^2 - x^2 - y^2)$ *be a lower-bound estimate function that is smooth and non-positive in* $\overline{\Omega}$. *Denote its corresponding grid function as* $\varphi_h \in \mathbb{R}^{N^2\times 1}$. *Then the vector* $\mathbf{A}\varphi_h \in \mathbb{R}^{N^2\times 1}$ *satisfies*

$$\mathbf{A}\varphi_h \leq -\|\sqrt{f}\|_\infty, \text{ for all } h. \tag{44}$$

Proof Without loss of generality, let us consider a grid point $\mathbf{x}_{i,j}$ where semi-Lagrangian wide stencil discretization is applied and boundary terms occur with $\mathbf{x}_{i,j} + \sqrt{h}(\mathbf{e}_z)_{i,j}$ relocated to $\mathbf{x}_{i,j} + \eta_1(\mathbf{e}_z)_{i,j}$. Then

$$\begin{aligned}
(\mathbf{A}\varphi_h)_{i,j} =& \, 2\left(\frac{a_{i,j}}{\eta_1\sqrt{h}} + \frac{1-a_{i,j}}{h}\right)\varphi(\mathbf{x}_{i,j}) - \frac{a_{i,j}}{\sqrt{h}\frac{\eta_1+\sqrt{h}}{2}}\,\mathcal{I}_h\varphi|_{\mathbf{x}_{i,j}-\sqrt{h}(\mathbf{e}_z)_{i,j}} \\
&- \frac{1-a_{i,j}}{h}\,\mathcal{I}_h\varphi|_{\mathbf{x}_{i,j}+\sqrt{h}(\mathbf{e}_w)_{i,j}} - \frac{1-a_{i,j}}{h}\,\mathcal{I}_h\varphi|_{\mathbf{x}_{i,j}-\sqrt{h}(\mathbf{e}_w)_{i,j}} \\
\leq& \, 2\left(\frac{a_{i,j}}{\eta_1\sqrt{h}} + \frac{1-a_{i,j}}{h}\right)\varphi(\mathbf{x}_{i,j}) - \frac{a_{i,j}}{\eta_1\frac{\eta_1+\sqrt{h}}{2}}\varphi(\mathbf{x}_{i,j}+\eta_1(\mathbf{e}_z)_{i,j}) \\
&- \frac{a_{i,j}}{\sqrt{h}\frac{\eta_1+\sqrt{h}}{2}}\varphi(\mathbf{x}_{i,j}-\sqrt{h}(\mathbf{e}_z)_{i,j}) - \frac{1-a_{i,j}}{h}\varphi(\mathbf{x}_{i,j}+\sqrt{h}(\mathbf{e}_w)_{i,j}) \\
&- \frac{1-a_{i,j}}{h}\varphi(\mathbf{x}_{i,j}-\sqrt{h}(\mathbf{e}_w)_{i,j}) \\
=& \, -\|\sqrt{f}\|_\infty,
\end{aligned}$$

where we have used $\varphi(\mathbf{x}_{i,j} + \eta_1(\mathbf{e}_z)_{i,j}) \leq 0$, and $\mathcal{I}_h\varphi|_{\mathbf{x}_{i,j}-\sqrt{h}(\mathbf{e}_z)_{i,j}} \geq \varphi(\mathbf{x}_{i,j} - \sqrt{h}(\mathbf{e}_z)_{i,j})$ and similarly for the other stencil points. Interested readers can prove the other cases in the same fashion. □

Lemma 6 (Stability for linear problem) *Assume that a control pair* (a, θ) *is given, such that the HJB Eq.* (7) *becomes linear:*

$$-\alpha_{11}(a,\theta)u_{xx} - 2\alpha_{12}(a,\theta)u_{xy} - \alpha_{22}(a,\theta)u_{yy} = -2\sqrt{a(1-a)f}, \text{ in } \Omega,$$
$$u = g, \text{ on } \partial\Omega.$$

Suppose the mixed discretization gives the linear system $\mathbf{A}(a_h, \theta_h)\, u_h = F_h(a_h, \theta_h)$, *which is the linear version of* (22). *Then the solution* u_h *is bounded as follows:*

1. *If* $g = 0$ *(homogeneous boundary condition) and* $f \geq 0$ *is a bounded function,*

$$-\frac{1}{2}\|\sqrt{f}\|_\infty R^2 \leq u_h \leq 0, \text{ independent of } h. \tag{45}$$

2. *If* $f = 0$ *(homogeneous PDE) and* g *is a bounded function,*

$$\|u_h\|_\infty \leq \|g\|_\infty, \text{ independent of } h. \tag{46}$$

3. *In general, if* $f \geq 0$ *and* g *are bounded functions,*

$$\|u_h\|_\infty \leq \frac{1}{2}\|\sqrt{f}\|_\infty R^2 + \|g\|_\infty, \text{ independent of } h. \tag{47}$$

Proof 1. The proof follows the idea in [30]. In this case, the N^2-vector F_h is simply given by $F_{i,j} = -2\sqrt{a_{i,j}(1-a_{i,j})f_{i,j}}$. Since $a_{i,j} \in [0,1]$, we have $-\|\sqrt{f}\|_\infty \leq F_h \leq 0$. Lemma 4 has proved that $\mathbf{A}$ is an M-matrix, and thus $\mathbf{A}^{-1} \geq 0$. Also, we note that $F_h \leq 0$. Hence, the upper bound of u_h is given by $u_h = \mathbf{A}^{-1}F_h \leq 0$.
Lemma 5 has proved that $\mathbf{A}\varphi_h \leq -\|\sqrt{f}\|_\infty$. Since $-\|\sqrt{f}\|_\infty \leq F_h = \mathbf{A}u_h$, we have $\mathbf{A}\varphi_h \leq \mathbf{A}u_h$. Since $\mathbf{A}^{-1} \geq 0$, we have $\varphi_h \leq u_h$. Hence, the lower bound of u_h is given by $u_h \geq \varphi_h \geq -\|\varphi\|_\infty = -\frac{1}{2}\|\sqrt{f}\|_\infty R^2$.
2. By Lemma 4, $\mathbf{A}$ is an M-matrix. Then following the proof in [9], the solution u_h under the M-matrix discretization satisfies the discrete comparison principle, and furthermore, (46).
3. This can be obtained by applying the superposition principle of the linear PDEs on 1 and 2. □

Eventually, we come back to our original nonlinear problem (22).

Lemma 7 (Stability for nonlinear problem) *Assume that* f *and* g *are bounded in* L_∞ *norm. Given that Lemma* 4 *is satisfied, the solution of the discrete system* (22), u_h, *is bounded by*

$$\|u_h\|_\infty \leq \frac{1}{2}\|\sqrt{f}\|_\infty R^2 + \|g\|_\infty, \tag{48}$$

where the bound is independent of the mesh size h *and the controls* (a_h, θ_h).

Proof Since the solution for the linear PDE under the mixed discretization is bounded by (47) under all admissible controls $(a_h, \theta_h) \in \Gamma$, and the bound is independent of the controls (a_h, θ_h) and the mesh size h, we conclude that the same bound applies to the solution for the nonlinear PDE under the mixed discretization. □

6.3 Monotonicity

For nonlinear PDEs, monotonicity is another sufficient condition for convergence in the viscosity sense. Monotonicity means that the discretization scheme at a grid point $\mathbf{x}_{i,j}$ must be a non-decreasing function of the unknown $u_{i,j}$ and a non-increasing function of the unknowns at the other points $\{u_{p,q}|_{(p,q)\neq(i,j)}\}$. Monotonicity of our numerical scheme (23)–(24) is inherited from the M-matrix property in Lemma 3.

Lemma 8 (Monotonicity) *The finite difference discretization* $\mathcal{F}_h(\mathbf{x}_{i,j}, u_h) = \mathcal{F}_h(\mathbf{x}_{i,j}, u_{i,j}, \{u_{p,q}|_{(p,q)\neq(i,j)}\}) = 0$, *given in* (23)–(24), *is monotone. More specifically, for all* $u_h \leq v_h$, *we have*

$$\begin{aligned}\mathcal{F}_h(\mathbf{x}_{i,j}, u_{i,j}, \{u_{p,q}|_{(p,q)\neq(i,j)}\}) &\leq \mathcal{F}_h(\mathbf{x}_{i,j}, v_{i,j}, \{u_{p,q}|_{(p,q)\neq(i,j)}\}),\\ \mathcal{F}_h(\mathbf{x}_{i,j}, u_{i,j}, \{u_{p,q}|_{(p,q)\neq(i,j)}\}) &\geq \mathcal{F}_h(\mathbf{x}_{i,j}, u_{i,j}, \{v_{p,q}|_{(p,q)\neq(i,j)}\}).\end{aligned} \tag{49}$$

Proof The proof follows [16]. Our goal is to verify the monotonicity condition (49). Without loss of generality, let us analyze one example: $u_h \leq v_h$ with $u_{i,j} = v_{i,j}$. Then

$$\begin{aligned}&\mathcal{F}_h(\mathbf{x}_{i,j}, u_{i,j}, \{u_{p,q}|_{(p,q)\neq(i,j)}\}) - \mathcal{F}_h(\mathbf{x}_{i,j}, u_{i,j}, \{v_{p,q}|_{(p,q)\neq(i,j)}\})\\ &\quad= \max_{(a_{i,j},\theta_{i,j})\in\Gamma}\left\{(\mathbf{A}(a_{i,j},\theta_{i,j})\,u_h)_{i,j} - F_{i,j}(a_{i,j},\theta_{i,j})\right\}\\ &\qquad- \max_{(a_{i,j},\theta_{i,j})\in\Gamma}\left\{(\mathbf{A}(a_{i,j},\theta_{i,j})\,v_h)_{i,j} - F_{i,j}(a_{i,j},\theta_{i,j})\right\}\\ &\quad\geq \min_{(a_{i,j},\theta_{i,j})\in\Gamma}[(\mathbf{A}(a_{i,j},\theta_{i,j})(u_h - v_h)]_{i,j} \;\geq\; 0,\end{aligned}$$

where the first inequality uses $\max_x f(x) - \max_x g(x) \geq \min_x [f(x) - g(x)]$, and the last inequality considers that $u_h - v_h \leq 0$ and that all the off-diagonal entries of $\mathbf{A}$ are non-positive under all admissible controls. □

6.4 Strong Comparison Principle

There is one more sufficient condition for convergence, called strong comparison principle [2]. Strong comparison principle holds if the boundary condition is satisfied in the viscosity sense. Unfortunately, there is no proof in the literature that this necessarily holds for the Dirichlet problem (2). Hence, we provide a proof in the setting of our proposed numerical scheme.

Lemma 9 *Let* $\zeta(\mathbf{x};\mathbf{p}) \equiv \frac{1}{2}\|\sqrt{f}\|_\infty\|\mathbf{x}-\mathbf{p}\|_2^2$, *where* $\mathbf{p}\in\mathbb{R}^2$ *is a random vector. Let* $\hat{u}(\mathbf{x}) : \{\mathbf{x}_{i,j} \in \Omega\} \cup \partial\Omega \to \mathbb{R}$, *where* $\hat{u}(\mathbf{x}) \equiv \begin{cases} u_h(\mathbf{x}_{i,j}), & \text{if} \quad \mathbf{x}\in\{\mathbf{x}_{i,j}\in\Omega\},\\ g(\mathbf{x}), & \text{if} \quad \mathbf{x}\in\partial\Omega.\end{cases}$ *Then* $\mathcal{I}_h\zeta \pm \hat{u}$ *achieves its maximum on* $\partial\Omega$.

Proof Without loss of generality, let us consider again a grid point $\mathbf{x}_{i,j}\notin\partial\Omega$ where semi-Lagrangian wide stencil discretization is applied and boundary terms occur with $\mathbf{x}_{i,j} + \sqrt{h}(\mathbf{e}_z)_{i,j}$ relocated to $\mathbf{x}_{i,j} + \eta_1(\mathbf{e}_z)_{i,j}$. Assume that the control pair is fixed. Define a linear stencil operator on an arbitrary function u at $\mathbf{x}_{i,j}$ as

$$\begin{aligned}\mathcal{S}[u](\mathbf{x}_{i,j}) \equiv\; & 2\left(\frac{a_{i,j}}{\eta_1\sqrt{h}} + \frac{1-a_{i,j}}{h}\right)u|_{\mathbf{x}_{i,j}} - \frac{a_{i,j}}{\sqrt{h}\frac{\eta_1+\sqrt{h}}{2}}u|_{\mathbf{x}_{i,j}-\sqrt{h}(\mathbf{e}_z)_{i,j}}\\ &- \frac{a_{i,j}}{\eta_1\frac{\eta_1+\sqrt{h}}{2}}u|_{\mathbf{x}_{i,j}+\eta_1(\mathbf{e}_z)_{i,j}} - \frac{1-a_{i,j}}{h}u|_{\mathbf{x}_{i,j}+\sqrt{h}(\mathbf{e}_w)_{i,j}} - \frac{1-a_{i,j}}{h}u|_{\mathbf{x}_{i,j}-\sqrt{h}(\mathbf{e}_w)_{i,j}}.\end{aligned}$$

We note that the relocated stencil point is also included in the operator. Then we have $\mathcal{S}[\mathcal{I}_h\zeta](\mathbf{x}_{i,j}) \le \mathcal{S}[\zeta](\mathbf{x}_{i,j}) = -\|\sqrt{f}\|_\infty$, and $\mathcal{S}[\hat{u}](\mathbf{x}_{i,j}) = -2\sqrt{a_{i,j}(1-a_{i,j})f_{i,j}}$. As a result, we have $\mathcal{S}[\mathcal{I}_h\zeta \pm \hat{u}](\mathbf{x}_{i,j}) = -\|\sqrt{f}\|_\infty \pm 2\sqrt{a_{i,j}(1-a_{i,j})f_{i,j}} \le 0$.

Now assume that $\mathcal{I}_h\zeta \pm \hat{u}$ achieves its maximum at this grid point $\mathbf{x}_{i,j}$. Next we prove that $(\mathcal{I}_h\zeta \pm \hat{u})|_{\mathbf{y}} = (\mathcal{I}_h\zeta \pm \hat{u})|_{\mathbf{x}_{i,j}}$ for any stencil point $\mathbf{y}$ connected to $\mathbf{x}_{i,j}$, namely, for any $\mathbf{y} \in \{\mathbf{x}_{i,j}+\eta_1(\mathbf{e}_z)_{i,j}, \mathbf{x}_{i,j}-\sqrt{h}(\mathbf{e}_z)_{i,j}, \mathbf{x}_{i,j}\pm\sqrt{h}(\mathbf{e}_w)_{i,j}\}$. This can be proved by contradiction. Assume that there exists at least one stencil point where the strict inequality holds, namely, $(\mathcal{I}_h\zeta \pm \hat{u})|_{\mathbf{y}} < (\mathcal{I}_h\zeta \pm \hat{u})|_{\mathbf{x}_{i,j}}$. Then

$$\mathcal{S}[\mathcal{I}_h\zeta \pm \hat{u}](\mathbf{x}_{i,j}) > \left[2\left(\frac{a_{i,j}}{\eta_1\sqrt{h}} + \frac{1-a_{i,j}}{h}\right) - \frac{a_{i,j}}{\sqrt{h}\frac{\eta_1+\sqrt{h}}{2}} - \frac{a_{i,j}}{\eta_1\frac{\eta_1+\sqrt{h}}{2}} - \frac{1-a_{i,j}}{h} - \frac{1-a_{i,j}}{h}\right](\mathcal{I}_h\zeta \pm \hat{u})|_{\mathbf{x}_{i,j}} = 0,$$

which contradicts with $\mathcal{S}[\mathcal{I}_h\zeta \pm \hat{u}](\mathbf{x}_{i,j}) \le 0$. The key point of this result is that $(\mathcal{I}_h\zeta \pm \hat{u})|_{\mathbf{x}_{i,j}+\eta_1(\mathbf{e}_z)_{i,j}} = (\mathcal{I}_h\zeta \pm \hat{u})|_{\mathbf{x}_{i,j}}$. That is, $\mathcal{I}_h\zeta \pm \hat{u}$ achieves its maximum at the boundary point $\mathbf{x}_{i,j} + \eta_1(\mathbf{e}_z)_{i,j} \in \partial\Omega$.

In general, consider any grid point $\mathbf{x}_{i,j} \notin \partial\Omega$. Assume that $\mathcal{I}_h\zeta \pm \hat{u}$ achieves its maximum at $\mathbf{x}_{i,j}$. One can prove in the same fashion that $(\mathcal{I}_h\zeta \pm \hat{u})|_{\mathbf{y}} = (\mathcal{I}_h\zeta \pm \hat{u})|_{\mathbf{x}_{i,j}}$ for any stencil point $\mathbf{y}$ connected to $\mathbf{x}_{i,j}$. Then by the connectivity property (see the proof of Lemma 4), there exists a boundary point $\mathbf{z} \in \partial\Omega$, such that $(\mathcal{I}_h\zeta \pm \hat{u})|_{\mathbf{z}} = (\mathcal{I}_h\zeta \pm \hat{u})|_{\mathbf{x}_{i,j}}$. Hence, $\mathcal{I}_h\zeta \pm \hat{u}$ achieves its maximum at the boundary point $\mathbf{z} \in \partial\Omega$. □

Lemma 10 *Let Ω be a strictly convex domain. Assume that Lemma 9 holds. Define*

$$\overline{u}(\mathbf{x}) \equiv \limsup_{h\to 0,\, \mathbf{y}\to\mathbf{x}} u_h(\mathbf{y}), \quad \underline{u}(\mathbf{x}) \equiv \liminf_{h\to 0,\, \mathbf{y}\to\mathbf{x}} u_h(\mathbf{y}).$$

Then $\overline{u}(\mathbf{x}) = \underline{u}(\mathbf{x}) = g(\mathbf{x})$ for all $\mathbf{x} \in \partial\Omega$.

Proof Once Lemma 9 holds, the proof follows Lemma 6.4 in [14]. □

Lemma 10 is essentially the comparison result on the boundary $\partial\Omega$. Now we are ready to extend the comparison result to the entire computational domain $\overline{\Omega}$.

Lemma 11 *Given that the finite difference discretization (23)–(24) satisfies consistency, stability and monotonicity, $\overline{u}(\mathbf{x})$ and $\underline{u}(\mathbf{x})$ are respectively the viscosity subsolution and supersolution of the Dirichlet problem (2).*

Proof See the proof of Theorem 2.1 in [2]. □

Lemma 12 (Strong comparison principle) *Let Ω be a strictly convex domain. Then the finite difference discretization (23)–(24) satisfies $\overline{u} \le \underline{u}$ in $\overline{\Omega}$.*

Proof Since $\overline{u}$ and $\underline{u}$ are respectively the viscosity subsolution and supersolution (Lemma 11), and $\overline{u} \le \underline{u}$ on $\partial\Omega$ (Lemma 10), by Theorem 3.3 in [10], we conclude that $\overline{u} \le \underline{u}$ in $\overline{\Omega}$. □

6.5 Convergence of the Numerical Solution to the Viscosity Solution

Once consistency, stability, monotonicity and strong comparison principle are proved, Barles–Souganidis theorem [2] guarantees the convergence of the numerical solution to the viscosity solution.

Theorem 2 (Barles–Souganidis theorem) *Let Ω be a strictly convex domain. Given that the finite difference discretization (23)–(24) satisfies consistency, stability, monotonicity and strong comparison principle, the numerical solution converges to the viscosity solution of the Dirichlet problem (2).*

Proof See Barles and Souganidis's proof of Theorem 2.1 in [2]. □

7 Numerical Results

In this section, we will present numerical results for the Monge–Ampère equation using our proposed mixed standard 7-point stencil and semi-Lagrangian wide stencil scheme. These numerical results show that the mixed scheme can achieve second order convergence rate whenever the standard 7-point stencils can be applied monotonically on the entire computational domain, and up to order one convergence rate otherwise. Compared to the pure semi-Lagrangian wide stencil scheme in [14], our proposed mixed scheme yields a smaller discretization error $\|u - u_h\|$ and a faster convergence rate. The examples we consider in this section come from [4,17]. We choose the tolerance of residual for the policy iteration to be 10^{-6}. We let the initial guess of the numerical solution be the solution of

$$\begin{aligned} u_{xx} + u_{yy} &= 2\sqrt{f}, \qquad \text{in } \Omega, \\ u &= g, \qquad \text{on } \partial\Omega, \end{aligned} \tag{50}$$

which corresponds to the solution of (7) with $a = \frac{1}{2}$ and arbitrary θ. We choose the grid size $N^2 = 32^2, 64^2, \ldots, 512^2$, and define the numerical convergence rate as $\log_2 \frac{\|u-u_h(\frac{N}{2})\|}{\|u-u_h(N)\|}$, where $u_h(N)$ is the numerical solution on an $N \times N$ grid.

Example 1 Start with

$$f(x, y) = \left(1 + x^2 + y^2\right) e^{x^2+y^2}, \qquad g(x, y) = e^{\frac{1}{2}(x^2+y^2)}, \qquad \overline{\Omega} = [-1, 1] \times [-1, 1],$$

where the exact solution $u(x, y) = e^{\frac{1}{2}(x^2+y^2)}$ is smooth. For this example, it turns out that the standard 7-point stencil discretization can be applied on the entire computational domain and still results in a monotone scheme, since the optimal control pair (a^*, θ^*) at every grid point is inside the 7-point-stencil regions $\Gamma^1 \cup \Gamma^2 \cup \partial\Gamma^0$. Consequentially, the numerical solution converges at the optimal theoretical convergence rate $O(h^2)$; see Fig. 4(2, red-solid) and Table 1(1). We observe that the computation is efficient, in the sense that the number of policy iterations remains a small constant 4 as N increases.

We compare the proposed mixed scheme with the pure semi-Lagrangian wide stencil scheme in [14], where the wide stencils are applied on the entire computation domain. Figure 4(2, blue-dashed) and Table 1(2) show that the convergence rate of the pure wide stencil scheme is approximately first order. We note that order one is the optimal theoretical convergence rate for the pure wide stencil scheme; see Lemma 2. The convergence rate using the proposed mixed scheme is significantly faster than the rate using the pure semi-Lagrangian wide stencil scheme.

Example 2 Consider

$$f(x, y) = \frac{2}{(2 - x^2 - y^2)^2}, \qquad g(x, y) = -\sqrt{2 - x^2 - y^2}, \qquad \overline{\Omega} = [0, 1] \times [0, 1],$$

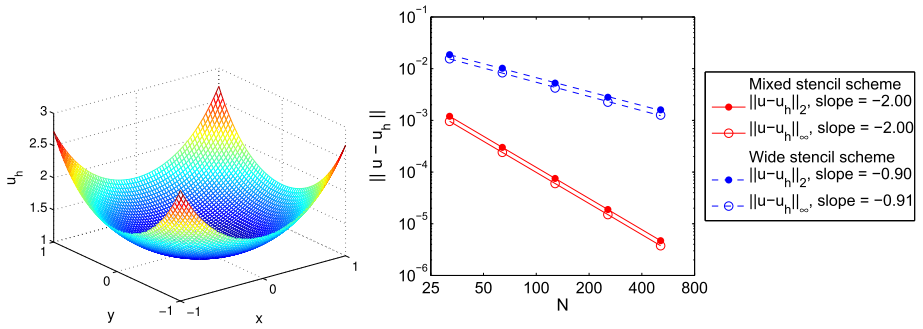


Fig. 4 Numerical results of Example 1, where the exact solution is $u(x, y) = e^{\frac{1}{2}(x^2+y^2)}$. **1** Numerical solution. **2** Norms of the errors $\|u - u_h\|$. For the proposed mixed stencil scheme (red-solid), the convergence rates, indicated by the slopes, are $O(h^2)$ in both L_2 and L_∞ norms. For the pure semi-Lagrangian wide stencil scheme (blue-dashed), the convergence rates are approximately $O(h)$ in both L_2 and L_∞ norms (Color figure online)

Table 1 Numerical results of Example 1, where the exact solution is $u(x, y) = e^{\frac{1}{2}(x^2+y^2)}$

N	$\|u - u_h\|_2$	Numerical convergence rate	$\|u - u_h\|_\infty$	Numerical convergence rate	Number of policy iterations
(1) Proposed mixed stencil scheme					
32	1.201×10^{-3}		9.598×10^{-4}		4
64	3.009×10^{-4}	2.00	2.404×10^{-4}	2.00	4
128	7.526×10^{-5}	2.00	6.013×10^{-5}	2.00	4
256	1.882×10^{-5}	2.00	1.504×10^{-5}	2.00	4
512	4.705×10^{-6}	2.00	3.759×10^{-6}	2.00	4
(2) Pure semi-Lagrangian wide stencil scheme					
32	1.868×10^{-2}		1.557×10^{-2}		5
64	1.020×10^{-2}	0.87	8.364×10^{-3}	0.90	5
128	5.263×10^{-3}	0.95	4.240×10^{-3}	0.98	6
256	2.801×10^{-3}	0.91	2.259×10^{-3}	0.91	5
512	1.600×10^{-3}	0.81	1.268×10^{-3}	0.83	5

(1) Proposed mixed stencil scheme. The convergence rates in both L_2 and L_∞ norms are $O(h^2)$. (2) Pure semi-Lagrangian wide stencil scheme. The convergence rates in both L_2 and L_∞ norms are approximately $O(h)$

where f is singular at (1, 1), and the exact solution is $u(x, y) = -\sqrt{2 - x^2 - y^2}$. Similar to Example 1, we can apply the standard 7-point stencil discretization monotonically on the entire Ω. The convergence rates are $O(h^2)$ and $O(h^{1.5})$ in L_2 and L_∞ norms respectively; see Fig. 5(2, red-solid) and Table 2(1). As a comparison, if we applied the pure semi-Lagrangian wide stencil scheme, then the convergence rate is worse than $O(h)$; see Fig. 5(2, blue-dashed) and Table 2(2).

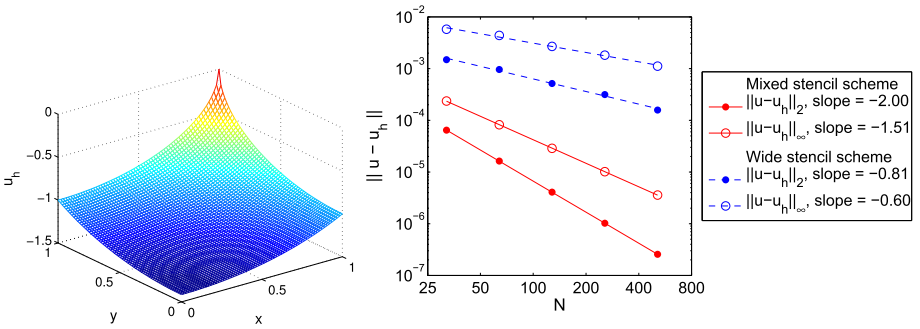


Fig. 5 Numerical results of Example 2, where the exact solution is $u(x, y) = -\sqrt{2 - x^2 - y^2}$. **1** Numerical solution. **2** Norms of the errors $\|u - u_h\|$. For the proposed mixed stencil scheme (red-solid), the convergence rates, indicated by the slopes, are $O(h^2)$ in L_2 norm and $O(h^{1.5})$ in L_∞ norm, respectively. For the pure semi-Lagrangian wide stencil scheme (blue-dashed), the convergence rates are worse than $O(h)$ in both L_2 and L_∞ norms (Color figure online)

Table 2 Numerical results of Example 2, where the exact solution is $u(x, y) = -\sqrt{2 - x^2 - y^2}$

N	$\|u - u_h\|_2$	Numerical convergence rate	$\|u - u_h\|_\infty$	Numerical convergence rate	Number of policy iterations
(1) Proposed mixed stencil scheme					
32	6.450×10^{-5}		2.359×10^{-4}		4
64	1.628×10^{-5}	1.99	8.211×10^{-5}	1.52	5
128	4.084×10^{-6}	2.00	2.882×10^{-5}	1.51	5
256	1.022×10^{-6}	2.00	1.015×10^{-5}	1.51	5
512	2.557×10^{-7}	2.00	3.583×10^{-6}	1.50	5
(2) Pure semi-Lagrangian wide stencil scheme					
32	1.493×10^{-3}		5.799×10^{-3}		5
64	9.634×10^{-4}	0.63	4.394×10^{-3}	0.40	4
128	5.166×10^{-4}	0.90	2.697×10^{-3}	0.70	5
256	3.153×10^{-4}	0.71	1.824×10^{-3}	0.56	5
512	1.583×10^{-4}	0.99	1.120×10^{-3}	0.70	5

(1) Proposed mixed stencil scheme. The convergence rates in L_2 and L_∞ norms are $O(h^2)$ and $O(h^{1.5})$, respectively. (2) Pure semi-Lagrangian wide stencil scheme. The convergence rates in both L_2 and L_∞ norms are worse than $O(h)$

Example 3 Consider

$$f(x, y) = \max\left(1 - \frac{0.1}{\sqrt{x^2 + y^2}}, 0\right), \qquad g(x, y) = \frac{1}{2}(\sqrt{x^2 + y^2} - 0.1)^2,$$
$$\overline{\Omega} = [-0.5, 0.5] \times [-0.5, 0.5].$$

The exact solution is given by $u(x, y) = \frac{1}{2}\max\left(\sqrt{x^2 + y^2} - 0.1, 0\right)^2$. This is a C^1 function where the singularity occurs at the ring $x^2 + y^2 = 0.1^2$. First we consider the proposed mixed scheme. Semi-Lagrangian wide stencils need to be applied near the ring $x^2 + y^2 = 0.1^2$.

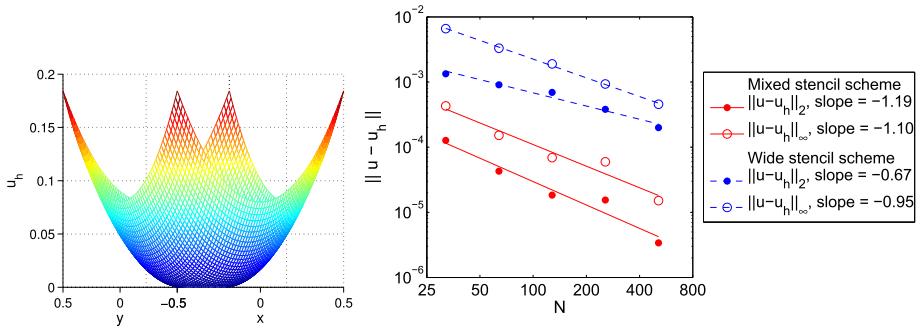


Fig. 6 Numerical results of Example 3, where the exact solution is $\frac{1}{2}\max\left(\sqrt{x^2+y^2}-0.1,\ \ 0\right)^2$. **1** Numerical solution. **2** Norms of the error $\|u-u_h\|$. For the proposed mixed stencil scheme (red-solid), the convergence rates, indicated by the slopes, are approximately $O(h)$ in both L_2 and L_∞ norms. For the pure semi-Lagrangian wide stencil scheme (blue-dashed), the errors are larger than the mixed scheme, and the convergence rates are worse than $O(h)$ in both L_2 and L_∞ norms (Color figure online)

Table 3 Numerical results for Example 3, where the exact solution is $\frac{1}{2}\max\left(\sqrt{x^2+y^2}-0.1, 0\right)^2$

N	$\|u-u_h\|_2$	Numerical convergence rate	$\|u-u_h\|_\infty$	Numerical convergence rate	Number of policy iterations
(1) Proposed mixed stencil scheme					
32	1.270×10^{-4}		4.298×10^{-4}		4
64	4.273×10^{-5}	1.57	1.520×10^{-4}	1.50	6
128	1.835×10^{-5}	1.22	6.907×10^{-5}	1.14	7
256	1.544×10^{-5}	0.25	5.959×10^{-5}	0.21	9
512	3.396×10^{-6}	2.18	1.513×10^{-5}	1.98	20
(2) Pure semi-Lagrangian wide stencil scheme					
32	1.337×10^{-3}		6.604×10^{-3}		5
64	9.084×10^{-4}	0.56	3.304×10^{-3}	1.00	6
128	6.940×10^{-4}	0.39	1.901×10^{-3}	0.80	7
256	3.815×10^{-4}	0.86	9.335×10^{-4}	1.03	7
512	1.998×10^{-4}	0.93	4.563×10^{-4}	1.03	9

(1) Proposed mixed stencil scheme. (2) Pure semi-Lagrangian wide stencil scheme. The errors $\|u-u_h\|$ by the proposed mixed stencil scheme are smaller than those by the pure wide stencil scheme

Figure 6(2, red-solid) and Table 3(1) show the numerical results. We note that the error reduction rates for the sequence of $N = 32, 64, \ldots, 512$ do not look as regular as the previous examples. The reason is that wide stencil introduces interpolation error, which fluctuates as N increases, despite converging towards 0. However, a clear error reduction, and thus convergence, can be observed. For comparison, we also test the pure semi-Lagrangian wide stencil scheme, as shown in Fig. 6(2, blue-dashed) and Table 3(2). Our proposed mixed scheme performs better than the pure wide stencil scheme, in the sense that the error $\|u-u_h\|$ is significantly smaller, and the convergence rate is faster.

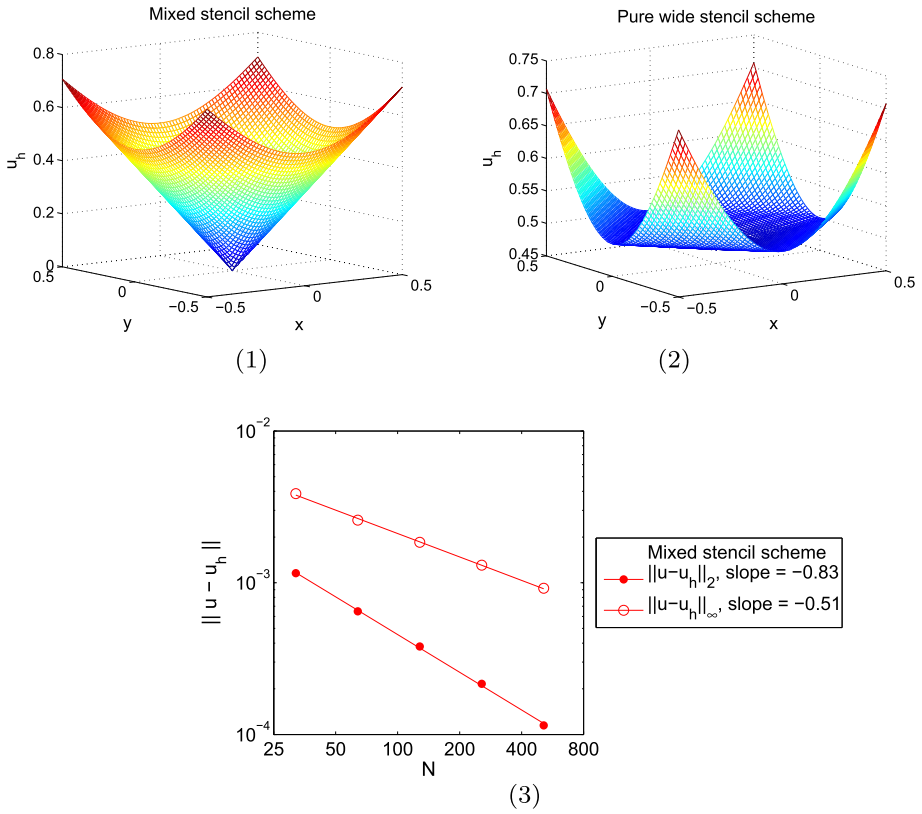


Fig. 7 Numerical results of Example 4, where the exact solution is $u(x, y) = \sqrt{x^2 + y^2}$. **1** Numerical solution by the proposed mixed stencil scheme, which converges to the exact solution. **2** Numerical solution by the pure semi-Lagrangian wide stencil scheme, which does not converge to the exact solution. **3** Norms of the error $\|u - u_h\|$. The proposed mixed stencil scheme is used. The convergence rates, indicated by the slopes, are $O(h^{0.8})$ in L_2 norm and $O(h^{0.5})$ in L_∞ norm, respectively

Example 4 In practice, our numerical scheme can converge to not only viscosity solutions, but also a type of more general weak solutions, called Aleksandrov solutions [20]. In this example, the corresponding f is a delta function at the origin and is zero elsewhere:

$$f(x, y) = \pi\delta(0, 0), \quad g(x, y) = \sqrt{x^2 + y^2}, \quad \overline{\Omega} = [-0.5, 0.5] \times [-0.5, 0.5].$$

The exact solution $u(x, y) = \sqrt{x^2 + y^2}$ is an Aleksandrov solution. It is a C^0 function and is singular at the origin. Figure 7(1) shows that our proposed mixed scheme converges to the cone-shaped Aleksandrov solution. Conversely, Fig. 7(2) shows that the pure semi-Lagrangian wide stencil scheme in [14] does not give the cone-shaped Aleksandrov solution. Indeed, there is no theoretical proof that the pure wide stencil scheme can converge to Aleksandrov solutions. Figure 7(3) and Table 4 report the convergence results by the proposed mixed scheme. The orders of convergence are close to 0.8 and 0.5 in L_2 and L_∞ norms respectively.

Table 4 Numerical results of Example 4. The exact solution is $u(x, y) = \sqrt{x^2 + y^2}$. The proposed mixed stencil scheme is used

N	$\|u - u_h\|_2$	Numerical convergence rate	$\|u - u_h\|_\infty$	Numerical convergence rate	Number of policy iterations
Proposed mixed stencil scheme					
32	1.156×10^{-3}		3.868×10^{-3}		9
64	6.484×10^{-4}	0.83	2.583×10^{-3}	0.58	15
128	3.803×10^{-4}	0.77	1.848×10^{-3}	0.48	17
256	2.159×10^{-4}	0.82	1.305×10^{-3}	0.50	23
512	1.148×10^{-4}	0.91	9.203×10^{-4}	0.50	27

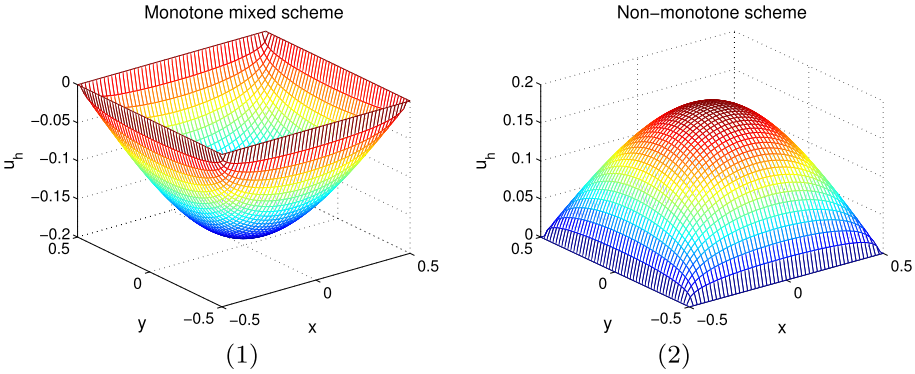


Fig. 8 Example 5: **1** The solution given by the monotone mixed scheme, which is convex and is convergent in the viscosity sense. **2** One possible solution given by a non-monotone scheme, which is concave and is not a viscosity solution

Example 5 In order to make a case for designing a monotone numerical scheme that converges to the viscosity solution (which is convex), we show explicitly that non-monotone numerical scheme may converge to a non-viscosity solution (which may be non-convex). More analysis on this issue can be found in [4,17]. We consider

$$f(x, y) = 1, \qquad g(x, y) = 0, \qquad \overline{\Omega} = [-0.5, 0.5] \times [-0.5, 0.5].$$

For this example, the exact solution u is not smooth near $\partial\Omega$ [4]. Since a closed-form expression for u is not available, we follow [4] and study the convergence behavior of u_h towards u by checking the values of $u_h(0, 0)$ as $h \to 0$. The numerical solution using our monotone mixed scheme converges to the convex viscosity solution as $h \to 0$; see Fig. 8 and Table 5. Alternatively, we consider a possible non-monotone discretization for $u_{xx}u_{yy} - u_{xy}^2 = f$, which is the direct application of the standard central differencing on u_{xx}, u_{yy} and the standard 4-point central differencing on u_{xy}. In our numerical experiment, the numerical solution under the non-monotone discretization converges to a concave function as $h \to 0$. We note that [4] has considered the same example using non-monotone discretization, and obtained another non-viscosity solution that is non-convex near $\partial\Omega$.

Table 5 Example 5: (1) The minimum values of the numerical solutions u_{min} given by the monotone mixed scheme, which provides an evidence that the numerical solution converges to a convex solution. (2) The maximum values of the numerical solutions u_{max} given by a non-monotone scheme, which provides an evidence that the numerical solution converges to a non-convex solution

N	$u_h(0,0)$ by monotone scheme	$u_h(0,0)$ by non-monotone scheme
32	−0.18380	0.18063
64	−0.18444	0.18312
128	−0.18461	0.18436
256	−0.18485	0.18499
512	−0.18507	0.18530

8 Conclusion

In this paper, we convert the Monge–Ampère equation into the equivalent HJB equation, and propose a mixed finite difference discretization for solving the equivalent HJB equation. The discretization satisfies consistency, stability, monotonicity and strong comparison principle, and thus convergent to the viscosity solution of the Monge–Ampère equation. Our proposed mixed scheme significantly improves the accuracy over the pure semi-Lagrangian scheme in [14]. More specifically, the proposed mixed scheme yields a smaller discretization error $\|u - u_h\|$. Furthermore, if the standard 7-point stencils can be applied on the entire computational domain monotonically, then our proposed mixed stencil scheme can improve the convergence rate to $O(h^2)$.

Our mixed scheme can be potentially extended to higher dimensional cases. Assuming that the dimension is d, the idea is to parametrize the control of the HJB Eq. (5), namely to parametrize $A(\mathbf{x}) = Q(\mathbf{x})\Lambda(\mathbf{x})Q(\mathbf{x})^T$, where $Q(\mathbf{x}) \in SO(d)$ and $\Lambda(\mathbf{x})$ is a trace-1 non-negative diagonal matrix. Then the standard 7-point stencil discretization can be applied if $A(\mathbf{x})$ is weakly diagonal dominant, and the semi-Lagrangian wide stencil discretization is applied otherwise. We leave this topic as a future work.

References

1. Azimzadeh, P., Forsyth, P.A.: Weakly chained matrices, policy iteration, and impulse control. SIAM J. Numer. Anal. **54**(3), 1341–1364 (2016). https://doi.org/10.1137/15M1043431
2. Barles, G., Souganidis, P.E.: Convergence of approximation schemes for fully nonlinear second order equations. Asymptot. Anal. **4**(3), 271–283 (1991)
3. Benamou, J.D., Collino, F., Mirebeau, J.M.: Monotone and consistent discretization of the Monge–Ampère operator. Math. Comput. **85**(302), 2743–2775 (2016). https://doi.org/10.1090/mcom/3080
4. Benamou, J.D., Froese, B.D., Oberman, A.M.: Two numerical methods for the elliptic Monge–Ampère equation. M2AN. Math. Model. Numer. Anal. **44**(4), 737–758 (2010). https://doi.org/10.1051/m2an/2010017
5. Böhmer, K.: On finite element methods for fully nonlinear elliptic equations of second order. SIAM J. Numer. Anal. **46**(3), 1212–1249 (2008). https://doi.org/10.1137/040621740
6. Bokanowski, O., Maroso, S., Zidani, H.: Some convergence results for Howard's algorithm. SIAM J. Numer. Anal. **47**(4), 3001–3026 (2009). https://doi.org/10.1137/08073041X
7. Brenner, S.C., Gudi, T., Neilan, M., Sung, Ly: C^0 penalty methods for the fully nonlinear Monge–Ampère equation. Math. Comput. **80**(276), 1979–1995 (2011). https://doi.org/10.1090/S0025-5718-2011-02487-7
8. Caffarelli, L.A., Milman, M. (eds.): Monge–Ampère equation: applications to geometry and optimization, Contemporary Mathematics, vol. 226. American Mathematical Society, Providence (1999). https://doi.org/10.1090/conm/226

9. Ciarlet, P.G.: Discrete maximum principle for finite-difference operators. Aequ. Math. **4**, 338–352 (1970)
10. Crandall, M.G., Ishii, H., Lions, P.L.: User's guide to viscosity solutions of second order partial differential equations. Bull. Am. Math. Soc. (N.S.) **27**(1), 1–67 (1992). https://doi.org/10.1090/S0273-0979-1992-00266-5
11. Crandall, M.G., Lions, P.L.: Viscosity solutions of Hamilton–Jacobi equations. Trans. Am. Math. Soc. **277**(1), 1–42 (1983). https://doi.org/10.2307/1999343
12. Dean, E.J., Glowinski, R.: Numerical methods for fully nonlinear elliptic equations of the Monge–Ampère type. Comput. Methods Appl. Mech. Eng. **195**(13–16), 1344–1386 (2006). https://doi.org/10.1016/j.cma.2005.05.023
13. Debrabant, K., Jakobsen, E.R.: Semi-Lagrangian schemes for linear and fully non-linear diffusion equations. Math. Comput. **82**(283), 1433–1462 (2013). https://doi.org/10.1090/S0025-5718-2012-02632-9
14. Feng, X., Jensen, M.: Convergent semi-Lagrangian methods for the Monge–Ampère equation on unstructured grids. SIAM J. Numer. Anal. **55**(2), 691–712 (2017)
15. Feng, X., Neilan, M.: Vanishing moment method and moment solutions for fully nonlinear second order partial differential equations. J. Sci. Comput. **38**(1), 74–98 (2009). https://doi.org/10.1007/s10915-008-9221-9
16. Forsyth, P.A., Labahn, G.: Numerical methods for controlled Hamilton–Jacobi–Bellman PDEs in finance. J. Comput. Finance **11**(2), 1 (2007)
17. Froese, B.D., Oberman, A.M.: Convergent finite difference solvers for viscosity solutions of the elliptic Monge–Ampère equation in dimensions two and higher. SIAM J. Numer. Anal. **49**(4), 1692–1714 (2011). https://doi.org/10.1137/100803092
18. Froese, B.D., Oberman, A.M.: Fast finite difference solvers for singular solutions of the elliptic Monge–Ampère equation. J. Comput. Phys. **230**(3), 818–834 (2011). https://doi.org/10.1016/j.jcp.2010.020
19. Froese, B.D., Oberman, A.M.: Convergent filtered schemes for the Monge–Ampère partial differential equation. SIAM J. Numer. Anal. **51**(1), 423–444 (2013). https://doi.org/10.1137/120875065
20. Gutiérrez, C.E.: The Monge–Ampère Equation, vol. 42. Springer, Berlin (2012)
21. Howard, R.A.: Dynamic Programming and Markov Processes. The Technology Press of M.I.T, Cambridge (1960)
22. Krylov, N.V.: The control of the solution of a stochastic integral equation. Teor. Verojatnost. i Primenen. **17**, 111–128 (1972)
23. Lakkis, O., Pryer, T.: A finite element method for nonlinear elliptic problems. SIAM J. Sci. Comput. **35**(4), A2025–A2045 (2013). https://doi.org/10.1137/120887655
24. Lin, J.: Wide stencil for the Monge–Ampère equation. Technical report, University of Waterloo master essay, supervised by Justin WL Wan, https://uwaterloo.ca/computational-mathematics/sites/ca.computational-mathematics/files/uploads/files/cmmain1.pdf (2014)
25. Lions, P.L.: Hamilton–Jacobi–Bellman equations and the optimal control of stochastic systems. In: Proceedings of the International Congress of Mathematicians, vol. 1, 2 (Warsaw, 1983), pp. 1403–1417. PWN, Warsaw (1984)
26. Ma, K., Forsyth, P.: An unconditionally monotone numerical scheme for the two factor uncertain volatility model. Preprint (2014)
27. Oberman, A.M.: Wide stencil finite difference schemes for the elliptic Monge–Ampère equation and functions of the eigenvalues of the Hessian. Discrete Contin. Dyn. Syst. Ser. B **10**(1), 221–238 (2008). https://doi.org/10.3934/dcdsb.2008.10.221
28. Oliker, V.I., Prussner, L.D.: On the numerical solution of the equation $(\partial^2 z/\partial x^2)(\partial^2 z/\partial y^2) - ((\partial^2 z/\partial x \partial y))^2 = f$ and its discretizations. I. Numer. Math. **54**(3), 271–293 (1988). https://doi.org/10.1007/BF01396762
29. Saad, Y.: Iterative Methods for Sparse Linear Systems, 2nd edn. Society for Industrial and Applied Mathematics, Philadelphia (2003). https://doi.org/10.1137/1.9780898718003
30. Samarskii, A.A.: The Theory of Difference Schemes, Monographs and Textbooks in Pure and Applied Mathematics, vol. 240. Marcel Dekker, Inc., New York (2001). https://doi.org/10.1201/9780203908518
31. Shivakumar, P.N., Williams, J.J., Ye, Q., Marinov, C.A.: On two-sided bounds related to weakly diagonally dominant M-matrices with application to digital circuit dynamics. SIAM J. Matrix Anal. Appl. **17**(2), 298–312 (1996). https://doi.org/10.1137/S0895479894276370
32. Smears, I.: Hamilton–Jacobi–Bellman equations analysis and numerical analysis. Technical report, research report available on www.math.dur.ac.uk/Ug/projects/highlights/PR4/Smears_HJB_report.pdf
33. Wang, J., Forsyth, P.A.: Maximal use of central differencing for Hamilton–Jacobi–Bellman PDEs in finance. SIAM J. Numer. Anal. **46**(3), 1580–1601 (2008). https://doi.org/10.1137/060675186