High-order filtered schemes for time-dependent second order HJB equations

Olivier Bokanowski*, Athena Picarelli†, and Christoph Reisinger†

August 23, 2017

Abstract

In this paper, we present and analyse a class of "filtered" numerical schemes for second order Hamilton-Jacobi-Bellman (HJB) equations. Our approach follows the ideas recently introduced in B.D. Froese and A.M. Oberman, Convergent filtered schemes for the Monge-Ampère partial differential equation, *SIAM J. Numer. Anal.*, 51(1):423–444, 2013, and more recently applied by other authors to stationary or time-dependent first order Hamilton-Jacobi equations. For high order approximation schemes (where "high" stands for greater than one), the inevitable loss of monotonicity prevents the use of the classical theoretical results for convergence to viscosity solutions. The work introduces a suitable local modification of these schemes by "filtering" them with a monotone scheme, such that they can be proven convergent and still show an overall high order behaviour for smooth enough solutions. We give theoretical proofs of these claims and illustrate the behaviour with numerical tests from mathematical finance, focussing also on the use of backward differencing formulae for constructing the high order schemes.

Keywords. monotone schemes, high-order schemes, backward difference formulae, viscosity solutions, second order Hamilton-Jacobi-Bellman equations.

1 Introduction

We consider second order Hamilton-Jacobi-Bellman (HJB) equations in $\mathbb{R}^d$:

$$\begin{cases} v_t + \sup_{a\in A} \mathcal{L}^a(t,x,v,D_x v, D_x^2 v) = 0 & x\in\mathbb{R}^d, t\in(0,T) \\ v(0,x) = v_0(x) & x\in\mathbb{R}^d, \end{cases} \tag{1.1}$$

where $\mathcal{L}^a : [0,T]\times\mathbb{R}^d\times\mathbb{R}\times\mathbb{R}^d\times\mathcal{S}^{d\times d}\to\mathbb{R}$ (where $\mathcal{S}$ is the set of symmetric matrices) takes the form

$$\mathcal{L}^a(t,x,r,p,Q) = \Big\{ -\frac{1}{2}Tr(\sigma\sigma^T(t,x,a)Q) + b(t,x,a)\cdot p + f(t,x,a)r + \ell(t,x,a)\Big\} \tag{1.2}$$

and $A\subset\mathbb{R}^m$ is a nonempty and compact set. We also define the Hamiltonian of the system:

$$H(t,x,r,p,Q) := \sup_{a\in A} \mathcal{L}^a(t,x,r,p,Q). \tag{1.3}$$

*Laboratoire Jacques-Louis Lions, Université Paris-Diderot, 5 Rue Thomas Mann, 75205 Paris, Cedex13, and Laboratoire UMA, Ensta ParisTech, Palaiseau, boka@math.univ-paris-diderot.fr

†Mathematical Institute, University of Oxford, Andrew Wiles Building, Woodstock Rd, Oxford OX2 6GG, {athena.picarelli,christoph.reisinger}@maths.ox.ac.uk

Existence and uniqueness of viscosity solutions to (1.1) are obtained for instance under the classical assumptions that the initial datum v_0 is bounded and Lipschitz continuous, and the coefficients b, σ, f, ℓ are bounded, Lipschitz and Hölder continuous respectively in space and time, i.e. there is a constant K such that, for any $\varphi \in \{b, \sigma, f, \ell\}$,

$$\sup_{a \in A} \left\{ \sup_{(t,x) \neq (s,y)} \frac{|\varphi(t,x,a) - \varphi(s,y,a)|}{|t-s|^{1/2} + |x-y|} + \sup_{(t,x)} |\varphi(t,x,a)| \right\} \leq K. \tag{1.4}$$

Moreover, the solution is then also Lipschitz continuous in space and locally 1/2-Hölder continuous in time [16]. The boundedness assumption can be removed, see also [16]. Moreover, in the paper we consider the equation on bounded numerical domains so that the boundedness assumption is automatically satisfied.

In this article, we propose approximation schemes for (1.1) for which convergence is guaranteed in a general setting, and which exhibit high order convergence under sufficient regularity of the solution. As we will explain in the following, these are in a sense two conflicting goals, and we will meet them by application of a so-called "filter", an idea introduced in [21].

The seminal work by Barles and Souganidis [6] establishes that a consistent and stable scheme converges to the viscosity solution of (1.1) if it is also monotone. That this is not simply a requirement of the proof, but can be crucial in practice, is demonstrated, e.g., by [37] (and also in Section 4.2 here). It is shown there empirically that for the uncertain volatility model from [30], perhaps the simplest non-trivial second order HJB equation there is, the (consistent and stable but) non-monotone Crank-Nicolson scheme fails to converge to the correct viscosity solution in the presence of Lipschitz initial data without higher regularity. This is in contrast to classical solutions where there is no such monotonicity requirement. As the setting of solutions above (i.e., Lipschitz in space and 1/2-Hölder in time) is the natural setting for HJB equations and often no higher global regularity is observed, a wide literature on monotone schemes has developed.

It is easy to see that for linear schemes, monotonicity is equivalent to a sign condition on the coefficients [20]. For nonlinear schemes, monotonicity requires that the discretized Hamiltonian is nondecreasing in the "central" variable and nonincreasing in the remaining variables [6, 19, 34].

By Godunov's theorem [22], in the case of explicit linear schemes for the approximation of the linear advection equation, the monotonicity property restricts the scheme to be of order at most one. Also, in [41], a similar result is given for the approximation of a diffusion equation with order of consistency limited to two. To the best of our knowledge, for general diffusions in more than one dimension, no monotone schemes of order higher than one are available in the literature. However, the provable order of convergence for second order HJB equations under the weak assumptions above is significantly less than one. By a technique pioneered by Krylov based on "shaking the coefficients" and mollification to construct smooth sub- and/or super-solutions, [25, 26, 3, 4, 5] prove certain fractional convergence orders. More encouragingly, it was remarked (Augoula and Abgrall [2]) that a weaker "ε-monotonicity" property was sufficient for proving convergence towards the viscosity solution.

To make matters worse, in more than one dimension, in the presence of general cross-derivative terms even first order consistent monotone schemes are necessarily "non-local", by which we mean that the length of the finite difference stencil grows relative to the mesh size as it is refined.

The minimal stencil size can be characterised by number theoretical tools using Diophantine equations [24], discrete geometrical methods [33] and Stern-Brocot trees [8], and duality techniques [38].

While standard finite difference schemes (see, e.g., [27]) are generally non-monotone unless the diffusion matrix is diagonally dominant (see [17]), schemes which are monotone by construction include semi-Lagrangian schemes [32, 15, 18] and generalized finite difference schemes [14, 13]. In order to utilize the second-order accuracy of standard finite differences for smooth solutions, the authors of [31] use those schemes in regions where the coefficients and controls are such that the

scheme is monotone, and switches to wide stencils only if monotonicity of the standard scheme, easily checkable by the signs of the discretization matrix, fails.

The strong uniform convergence of monotone P1 finite elements to the viscosity solution of possibly degenerate HJB equations is established in [23]. Again, contrasting this with the case of more regularity, [40] proves high order convergence of discontinuous finite element approximations under a Cordes condition on the coefficients which guarantees high order Sobolev regularity for smooth enough data.

The simple idea of "filtered" schemes is to use a combination of a high order scheme and a low order monotone scheme, where the latter is known to converge *a priori* by standard results [6]. The filter ensures that the low order scheme is used locally where and when the discrepancy to the high order scheme is too large, thus ensuring at least the same convergence order as the low order scheme (see, e.g., [3, 4, 5]), but otherwise uses the high order scheme and benefits from its accuracy for smooth solutions. Such schemes have been proposed, analysed and used in [21] for the Monge-Ampere equation, and in [35, 11] for first order Hamilton-Jacobi equations (see also [10] for convergence results on non-monotone value iteration schemes for first order stationary equations). We continue this program by studying (time-dependent) second order Hamilton-Jacobi-Bellmann equations as they arise from stochastic control problems.

Our results parallel the ones in [11] to prove, in the second order setting, that suitable filtered schemes converge at least of the same order as the underlying monotone scheme if there are solutions (only) in the viscosity sense and exhibit higher order truncation error for sufficiently smooth solutions.

On one hand, the presence of the diffusion term implies more regularity of the solution and therefore makes it possible to recover the high order behavior of the scheme (see Sections 4.1 and 4.3). On the other hand, if compared with the results in [11], the diffusion has a detrimental effect when the filter is activated to correct some pathological behavior of the high order scheme. In fact, in this case, the loss of accuracy does not remain localized and it diffuses in a neighborhood of the region of interest, as clearly shown by our example in Section 4.2.

Given the much more restrictive CFL condition on the time step in the second order case, we include implicit time stepping schemes in our analysis. This requires an extension of the arguments in [11] to work with monotone implicit operators. The monotone schemes covered by the analysis include the most commonly used one-step finite difference and semi-Lagrangian schemes. The theoretical results of the paper do not make use of any particular assumption on the high order scheme, beyond a higher order truncation error. However, in the numerical examples we focus on backward differentiation formulae of second order (BDF2). Although non-monotone, these schemes show good stability properties and have been recently used for solving obstacle problems for parabolic differential equations for American-style options in [36] and [9].

The rest of this article is organized as follows. In Section 2, we present the general framework for the monotone scheme, the higher order scheme, and the filtered scheme, and give examples of such schemes. Section 3 is devoted to convergence results, as well as useful existence results for some implicit schemes. Numerical examples mostly motivated by problems from mathematical finance are given in Section 4, and we conclude with some remarks in Section 5.

Acknowledgments. This work was partially supported by the program GNCS-INdAM. We are grateful to Peter Forsyth for the provision of his code for Example 2 (Section 4.2).

2 Main assumptions and definition of the scheme

In order to simplify the presentation we will focus our analysis on the one-dimensional case:

$$v_t + \sup_{a\in A} \left(-\tfrac{1}{2}\sigma^2(t,x,a)v_{xx} + b(t,x,a)v_x + f(t,x,a)v + \ell(t,x,a)\right) = 0, \tag{2.1}$$

but the main analysis can be extended to higher dimensions. Let $N \geq 1$ and let us introduce a time step

$$\tau := T/N.$$

We denote by Δx the space step. A uniform mesh in time and space is defined in one dimension by:

$$t_n = n\tau, \quad n \in \{0, \dots, N\} \quad \text{and} \quad x_i \equiv i\Delta x \quad i \in \mathbb{I} \subseteq \mathbb{Z}.$$

We also denote by $\mathcal{G}_{\Delta x} := \{x_i : i \in \mathbb{I}\}$ the space grid. The analysis can be adapted to nonuniform grids (see Section 4.2) and higher dimensions (see Section 4.3) by interpreting x_i as a general mesh point in a potentially non-uniform or higher-dimensional mesh, and Δx as the maximum mesh size.

We will denote by $u = (u_i^n)$ the numerical approximation of the solution v, so that

$$u_i^n \approx v(t_n, x_i)$$

and furthermore u^n will denote the vector $(u_i^n)_{i \in \mathbb{I}}$.

We aim to define a high order convergent scheme ("high" stands for greater than one) for the approximation of (1.1). However, in order to be in the convergence framework of the theorem of Barles and Souganidis [6], monotonicity of the scheme is fundamental, restricting the attainable order, as described in the introduction. Hence, in order to devise our convergent high order scheme, we consider the framework of Froese and Oberman [21] using filtered schemes, which is a special form of ε-monotone schemes. Three main ingredients are needed: a monotone scheme, a higher order scheme and a filter function.

Let us consider the numerical approximation given by a monotone convergent (one-step) scheme written, in abstract form, for $n = 0, \dots, N-1$:

$$u_i^{n+1} := S_M(u^n)_i, \quad \forall i \in \mathbb{I}, \tag{2.2}$$

with initialization

$$u_i^0 := v_0(x_i), \quad \forall i \in \mathbb{I}. \tag{2.3}$$

Although here (2.2) is written in explicit form, the scheme may be implicitly defined.

Analogously we consider a two-step high order scheme (high order consistent, but possibly neither monotone nor stable), for $n \geq 1$:

$$u_i^{n+1} = S_H(u^n, u^{n-1})_i, \quad \forall i \in \mathbb{I} \tag{2.4}$$

(some particular definition of the scheme might be needed for u^1). As above, the scheme written here in explicit form may also correspond to an implicit scheme.

A more precise characterization of S_M and S_H will be given below.

We consider the following filter function as introduced in [11, 35]:

$$F(x) := \begin{cases} x & \text{if } |x| \leq 1 \\ 0 & \text{otherwise.} \end{cases} \tag{2.5}$$

Analogous theoretical results may be obtained using different filter functions such that $\|F\|_\infty \leq 1$ and $F(x) = x$ in a neighborhood of $x = 0$, as in [21]. The filtered scheme is then defined in the following form, for $n \geq 1$:

$$u_i^{n+1} = S_F(u^n, u^{n-1})_i := S_M(u^n)_i + \varepsilon\tau F\left(\frac{S_H(u^n, u^{n-1})_i - S_M(u^n)_i}{\varepsilon\tau}\right), \quad i \in \mathbb{I}, \tag{2.6}$$

where $\varepsilon = \varepsilon_{\tau,\Delta x} > 0$ and such that

$$\lim_{(\tau,\Delta x)\to 0} \varepsilon_{\tau,\Delta x} = 0.$$

Specific choices of $\varepsilon_{\tau,\Delta x}$ will be made precise later on.

Although the form of the filtered scheme (2.6) is explicit, we emphasize again that the computation of $S_M(u^n)$ and of $S_H(u^n, u^{n-1})$ may require the solution of implicit schemes.

Notice that the construction principle of (2.6) with the filter function (2.5) bears a resemblance to flux limiter schemes for conservation laws; see Section 16.2 in [28]. In such schemes, a sensor function interpolates between a high order and a low order numerical flux, such that in regions where the solution is smooth the accuracy of the high order scheme can be exploited, while near discontinuities the monotone behaviour of the low order scheme is utilised. A closely related class are slope limiter methods (Section 16.3 in [28]), which choose the weights of the schemes based on the ratio of the numerical gradients of the solution at neighbouring mesh points. The requirement for second order consistency is that the high order scheme is used if the ratio is close to one. If the limiter lies between certain bounds (the minmod and superbee limiters), the scheme preserves the property of the continuous equation that the total variation diminishes over time (TVD).

Despite some similarities, there are marked differences between those schemes and the filtered schemes studied in this paper. First, the flux and gradient limiter schemes use information about the gradient and its relative change, while the filtered schemes use information about the truncation error, through local comparison of the low and high order solutions. In the case of a central second spatial difference, for instance, the leading order term in the truncation error of a smooth function is the fourth derivative. Second, where the flux and slope limiter schemes are designed to be stable, specifically in the TVD sense, but convergence is not guaranteed, the filtered schemes are explicitly constructed to be convergent by tying the solution to one which is known to converge. Lastly, there is a relatively large family of limiter functions studied in the literature which share the aforementioned conditions of second order accuracy and TVD stability but are found to have numerically different properties for different applications. We did not observe any significant change in behaviour for different filter functions as long as the conditions discussed after (2.5) were satisfied. There may, however, be scope to construct more sophisticated filter functions by including more information other than the point values of the high and low order solutions.

2.1 The monotone scheme

For convenience, the monotone scheme (2.2) shall also be denoted in the following abstract form, for $n = 0, \ldots, N-1$:

$$\mathscr{S}_M(t_{n+1}, x_i, u_i^{n+1}, u)_i = 0, \quad i \in \mathbb{I}, \tag{2.7}$$

where u denotes all the components (u_ℓ^k). This formulation may include both explicit and implicit schemes.

For computational purposes it will be necessary to define our scheme on a bounded domain. Therefore, from now on we consider

$$\mathbb{I} = \{1, \ldots, J\},$$

which also means that the scheme (2.7) may take into account some boundary conditions.

We consider a particular family of schemes $\mathscr{S}_M$ with the following form:

$$\mathscr{S}_M(t_{n+1}, x_i, u_i^{n+1}, u)_i \equiv \frac{1}{\tau} \sup_{a \in A} \left\{ M^{a,n+1} u^{n+1} - G^{a,n}(u^n) \right\}_i \tag{2.8}$$

where $M^{a,n+1} \in \mathbb{R}^{J \times J}$ and $G^{a,n}(u^n) \in \mathbb{R}^J$. More explicitly, for any $\varphi : [0,T] \times \mathbb{R} \to \mathbb{R}^J$, $\mathscr{S}_M$ can be written in the form

$$\mathscr{S}_M(t_{n+1}, x_i, r, \varphi)_i = \frac{1}{\tau} \sup_{a \in A} \left\{ M_{ii}^{a,n+1} r + \sum_{\mathbb{I} \ni j \neq i} M_{ij}^{a,n+1} \varphi(t_{n+1}, x_j) - G^{a,n}(\varphi(t_n, .))_i \right\}.$$

If the scheme is defined in explicit form, i.e., $u_i^{n+1} = S_M(u^n)_i$, then it suffices to take $M^{a,n+1} = I_J$ (the identity matrix in $\mathbb{R}^{J\times J}$) and $S_M(u^n)_i := \inf_{a\in A} G^{n,a}(u^n)_i$.

For any $x \in \mathbb{R}^J$ and $A \in \mathbb{R}^{J\times J}$ we denote the usual vector and matrix supremum norm by

$$\|x\|_\infty := \sup_{i\in\mathbb{I}} |x_i| \qquad \text{and} \qquad \|A\|_\infty := \sup_{x\neq 0} \frac{\|Ax\|_\infty}{\|x\|_\infty} = \sup_{i\in\mathbb{I}} \sum_{j\in\mathbb{I}} |A_{i,j}|.$$

One can observe that in expression (2.8) only the contribution of the i-th line of $M^{a,n+1}$ and $G^{a,n}$ appears. Therefore, denoting for any $a \equiv (a_1, \dots, a_J) \in A^J$

$$(M^{a,n})_{i,j} := (M^{a_i,n})_{i,j} \qquad \text{and} \qquad (G^{a,n}(u))_i := (G^{a_i,n}(u))_i, \tag{2.9}$$

the scheme can also be written in the equivalent vector form:

$$\sup_{a\in A^J} \left\{ M^{a,n+1}u^{n+1} - G^{a,n}(u^n) \right\} = 0, \quad \text{in } \mathbb{R}^J. \tag{2.10}$$

Remark 2.1. *The form of the scheme* (2.8) *is natural and will be satisfied by all the schemes considered in this paper. They are of the "discretize, then optimise" type (see [20]), where we discretize the linear operator in* (1.2) *by a monotone linear scheme for a fixed control a and then carry out the optimisation in* (1.1)*.*

The following assumptions are considered on $(M^{a,n+1})$ and $(G^{a,n})$:

Assumption (A1):

(*i*) For all $a \in A^J$ and $n \geq 1$,

$$M^{a,n} \text{ is an } M\text{-matrix} \tag{2.11}$$

(A is said to be an M-matrix if $A_{ij} \leq 0$, $\forall i \neq j$, and $A^{-1} \geq 0$ componentwise);

(*ii*) For all $n \geq 1$ (and for all $\tau, \Delta x$), there exists a constant $C_n = C_n(\tau, \Delta x) \geq 0$ such that

$$\sup_{a\in A^J} \|M^{a,n}\|_\infty \leq C_n; \tag{2.12}$$

(*iii*) There exists $C > 0$ (independent of $n, \tau, \Delta x$), such that for all $n \geq 1$,

$$\sup_{a\in A^J} \|(M^{a,n})^{-1}\|_\infty \leq 1 + C\tau; \tag{2.13}$$

(*iv*) For all $a \in A^J$ and $n \geq 0$, $G^{n,a}$ is monotone increasing, i.e., $\forall \varphi, \psi \in \mathbb{R}^J$:

$$\varphi \leq \psi \qquad \Rightarrow \qquad G^{a,n}(\varphi) \leq G^{a,n}(\psi) \tag{2.14}$$

(where here "$\leq$" denotes the componentwise inequality between vectors);

(*v*) There exists $C > 0$ (independent of $n, \tau, \Delta x$) such that for all $a \in A^J$, $n \geq 0$,

$$\forall \varphi, \psi \in \mathbb{R}^J, \quad \|G^{a,n}(\varphi) - G^{a,n}(\psi)\|_\infty \leq (1 + C\tau)\|\varphi - \psi\|_\infty, \tag{2.15}$$

and

$$\|G^{a,n}(0)\|_\infty \leq C\tau. \tag{2.16}$$

Existence of solutions of the scheme (2.10) will be shown under assumption (A1) (see Section 3, Proposition 3.1).

Remark 2.2. *Assumption (A1) is more specific than the basic requirements of monotonicity and stability in [6]. However, to show monotonicity for a given scheme for HJB equations, typically (A1)(i) and (iv) are used, and similarly (A1)(ii), (iii), and (v) for stability. Hence, we consider these as natural assumptions for schemes of the form (2.8).*

Hereafter, we will denote by $C^{p,q}$ the set of functions that are continuously p-differentiable with repect to the time variable t and q-differentiable with respect to the space variable x, and will also denote φ_{pt} or φ_{qx} the corresponding partial derivatives. For any function $\varphi \in C^{1,2}$, the consistency error of the scheme $\mathscr{S}_M$ is defined, for a given (t,x), by:

$$\mathcal{E}^{\varphi}_{\mathscr{S}_M}(\tau,\Delta x) := \Big|\mathscr{S}_M(t+\tau,x,\varphi(t+\tau,x),\varphi) - \Big(\varphi_t(t,x) + H(t,x,\varphi,\varphi_x,\varphi_{xx})\Big)\Big| \tag{2.17}$$

where H is defined by (1.3). The following natural consistency property of the scheme will be needed.

Assumption (A2) [consistency]:

For any (t,x), for any function φ sufficiently regular in a neighbourhood of (t,x):

$$\lim_{(\tau,\Delta x,\xi)\to 0} \mathcal{E}^{\varphi+\xi}_{\mathscr{S}_M}(\tau,\Delta x) = 0. \tag{2.18}$$

Notice that the consistency property (A2), introduced for simplicity, implies the weaker consistency condition of Barles and Souganidis [6] in the interior of the domain.

As recalled before, a monotone scheme has a limited order of convergence. More precisely, in our setting, we will typically assume to have first order consistency:

Assumption (A2') [first-order consistency]:

There exists $q \in \{1,2\}$ such that, for any function $\varphi \in C^{2,2+q}$:

$$\big|\mathcal{E}^{\varphi}_{\mathscr{S}_M}(\tau,\Delta x)\big| \le C_M \left(\|\varphi_{tt}\|_\infty + \sum_{2\le k\le 2+q} \|\varphi_{kx}\|_\infty\right) \max(\tau,\Delta x) \tag{2.19}$$

as $(\tau,\Delta x)\to 0$, where $C_M \ge 0$ is a constant independent of φ, τ and Δx.

We will eventually assume also that the scheme has the same order of convergence in τ and Δx, in order to obtain the bound (2.19) (see the case of the semi-Lagrangian scheme below).

The last inequality reads $|\mathcal{E}^{\varphi}_{\mathscr{S}_M}(\tau,\Delta x)| \le C^{\varphi}_M \max(\tau,\Delta x)$ for some constant C^{φ}_M that can be bounded explicitly.

Furthermore, it is easily verified that (A2') $\Rightarrow$ (A2), so that we will focus on (A2').

2.1.1 Two examples of monotone schemes

We now consider two types of monotone schemes. We focus again on the one-dimensional case (2.1). The first type is the Implicit Euler (IE) finite difference scheme, defined as follows:

Implicit Euler (IE) Scheme. *For $n \ge 0$ the scheme is defined by:*

$$\begin{aligned}\frac{u_i^{n+1}-u_i^n}{\tau} + \sup_{a\in A}\Big\{ &-\frac{1}{2}\sigma^2(t_{n+1},x_i,a)D^2u_i^{n+1} + b^+(t_{n+1},x_i,a)D^{1,-}u_i^{n+1} \\ &- b^-(t_{n+1},x_i,a)D^{1,+}u_i^{n+1} + f(t_{n+1},x_i,a)u_i^{n+1} + \ell(t_{n+1},x_i,a)\Big\} \;=\; 0,\end{aligned} \tag{2.20}$$

where we have denoted

$$D^2 v_i := \frac{v_{i-1} - 2v_i + v_{i+1}}{\Delta x^2}, \tag{2.21}$$

$$D^{1,-} v_i := \frac{v_i - v_{i-1}}{\Delta x} \quad \textit{and} \quad D^{1,+} v_i := \frac{v_{i+1} - v_i}{\Delta x}, \tag{2.22}$$

and where we have used the decomposition $b = b^+ - b^-$ *with*

$$b^{\pm}(t, x, a) := \max(\pm b(t, x, a), 0). \tag{2.23}$$

The scheme (2.20) can also be written in the equivalent form (2.10) with $M^{a,n+1}$ the tridiagonal matrix such that

$$M^{a,n+1}_{i,i} := 1 + \frac{\tau}{\Delta x^2}\sigma^2(t_{n+1}, x_i, a) + \frac{\tau}{\Delta x}|b(t_{n+1}, x_i, a)| + \tau f(t_{n+1}, x_i, a) \tag{2.24}$$

$$M^{a,n+1}_{i,i\pm1} := -\frac{1}{2}\frac{\tau}{\Delta x^2}\sigma^2(t_{n+1}, x_i, a) - \frac{\tau}{\Delta x} b_{\pm}(t_{n+1}, x_i, a), \tag{2.25}$$

and with $G^{a,n}$ defined by

$$G^{a,n}(u^n)_i := u^n_i - \tau\ell(t_{n+1}, x_i, a). \tag{2.26}$$

We summarize here some basic results concerning the IE scheme:

Proposition 2.1. *Assume that* σ, b, f, ℓ *are bounded functions. The (IE) scheme satisfies assumptions (A1) and (A2'), with* $q = 1$.

Proof. We note that for diagonally dominant matrices, i.e. such that $\forall i$, $|M_{ii}| \geq \delta + \sum_{j\neq i} |M_{ij}|$ for some $\delta > 0$, it holds $\|M^{-1}\|_\infty \leq 1/\delta$, see [42]. From this follows the estimate (A1)(*iii*). Other properties are immediate or classical. □

Remark 2.3. *Existence and uniqueness results for such implicit schemes will be stated in Section 3, using monotonicity properties of the matrix* $M^{a,n}$. *We can also solve* (2.20) *efficiently by the policy iteration algorithm (see [12]).*

We recall that, in multiple dimensions, standard finite difference schemes are in general non-monotone. In [14, 13] it is shown how to get monotone schemes for second order equations with general diffusion matrices, but also at the cost of a wider stencil as well as limited order of consistency.

As an alternative to finite difference schemes, simple and explicit monotone schemes known as semi-Lagrangian (SL) schemes [32, 15, 18] can be considered. They are based on a discrete time approximation of the Dynamic Programming Principle satisfied by the exact solution, combined with a spatial grid interpolation.

In the one-dimensional case of (2.1), this leads to the following approximation, for $n \geq 0$:

$$u^{n+1}_i = \inf_{a\in A}\Big\{\frac{1}{2}\sum_{\epsilon=\pm1}\big[u^n\big]\big(x_i - \tau b(t_n, x_i, a) + \epsilon\sqrt{\tau}\sigma(t_n, x_i, a)\big) \\ -\tau f(t_n, x_i, a)u^n_i - \tau\ell(t_n, x_i, a)\Big\}, \tag{2.27}$$

where $[\,\cdot\,]$ stands for a monotone linear interpolation operator on the spatial grid.

A straightforward equivalent of (2.27), which shows the similarity with the finite difference scheme (2.20) and is convenient for computing the truncation error, is then:

Semi-Lagrangian (SL) Scheme. *For $n \geq 0$ the scheme is defined by:*

$$\frac{u_i^{n+1} - u_i^n}{\tau} + \sup_{a \in A} \Big\{ - \frac{1}{2\tau} \Big(\sum_{\epsilon = \pm 1} \big[u^n\big]\big(x_i - \tau b(t_n, x_i, a) + \epsilon \sqrt{\tau} \sigma(t_n, x_i, a)\big) - 2u_i^n \Big) + f(t_n, x_i, a) u_i^n + \ell(t_n, x_i, a) \Big\} \; = \; 0. \tag{2.28}$$

As this is an explicit scheme, we can choose $M^{a,n+1} := I_J$ (the identity matrix in $\mathbb{R}^{J \times J}$) and $G^{a,n}(u^n)_i$ defined as the right-hand-side of (2.27). A two-dimensional version will be presented on a numerical example in Section 4.3.

Proposition 2.2. *Assume that σ, b, f, ℓ are bounded functions. The (SL) scheme satisfies assumption (A1). Assumption (A2') is satisfied with $q = 2$ and for τ and Δx of the same order.*

Proof. For data $\varphi \in C^2$, the interpolation error satisfies $\|\varphi - [\varphi]\|_\infty \leq \frac{1}{8} \|\varphi_{xx}\|_\infty \Delta x^2$, and therefore it is easy to see that the consistency error satisfies the following bound:

$$\big|\mathcal{E}^{\varphi}_{\mathscr{S}_M}(\tau, k, \Delta x)\big| \leq C \left(\tau + \frac{\Delta x^2}{\tau} \right).$$

Then for $\Delta x \equiv \tau$, the desired consistency estimate is obtained. □

Notice that for the exact solution, in general, only Lipschitz spatial regularity holds and precise error estimates are of the order of $O(\tau^{1/4}) + O(\frac{\Delta x}{\tau})$, where $O(\Delta x)$ is the interpolation error for Lipschitz regular data, see [18] (see also [1] for the case of unbounded data).

We refer to [32, 15] for the introduction of SL schemes in the context of second order equations and to [18] for an exhaustive discussion and main results.

2.2 The high order scheme

The high order scheme (2.4) will be written in the following form, for $n = 0, \ldots, N-1$:

$$\mathscr{S}_H(t_{n+1}, x, u_i^{n+1}, u)_i = 0, \qquad i \in \mathbb{I} \tag{2.29}$$

with an initialization of u^0 as in (2.3) and, possibly, a particular definition of u^1 to handle the case of two-step schemes (general multi-step schemes can be handled similarly).

We want to make minimal assumptions on the high order scheme to allow flexibility for obtaining the high order (in particular, we do not assume monotonicity). As a minimum, we require that the scheme is well-defined, i.e., that (2.29) uniquely determines u^{n+1}, such that we can write a general two-step scheme in explicit form

$$u^{n+1} = S_H(u^n, u^{n-1}).$$

We consider the following assumption:

Assumption (A3) [high order consistency]:

> There exist $k \geq 2$ and $q \in \mathbb{N}$ such that, for any function φ with regularity $C^{1+k,2+q}$ in the neighborhood of some point (t, x) and such that
>
> $$\varphi_t(.,.) + H(.,.,\varphi, \varphi_x, \varphi_{xx}) = 0$$

in a neighborhood of (t,x), one has

$$\left|\frac{1}{\tau}\Big(\varphi(t+\tau,x) - S_H(\varphi(t,\cdot),\varphi(t-\tau,\cdot))(x)\Big)\right| \\ \leq C_{H,k}\left(\|\varphi_{(1+k)t}\|_\infty + \sum_{2\leq p\leq 2+q}\|\varphi_{px}\|_\infty\right)\ \max(\tau^k,\Delta x^k) \tag{2.30}$$

for some constant $C_{H,k}\geq 0$ that is independent of φ,τ,Δx.

Remark 2.4. *We point out that for the high order schemes we present here, assumption (A3) is satisfied if*

(i) there exist $k\geq 2$, $\alpha\in[0,1]$ and $q\in\mathbb{N}$ such that,

$$\begin{aligned}\mathcal{E}^{\varphi}_{\mathscr{S}_H}(\tau,\Delta x) &:= \left|\mathscr{S}_H(t+\tau,x,\varphi(t+\tau,x),\varphi) - \Big(\varphi_t(t+\alpha\tau,x) + H(t+\alpha\tau,x,\varphi,\varphi_x,\varphi_{xx})\Big)\right| \\ &\leq C^{\varphi}_{H,k}\ \max(\tau^k,\Delta x^k)\end{aligned} \tag{2.31}$$

for some constant $C^{\varphi}_{H,k}\geq 0$ independent of τ, Δx,

(ii) the high order scheme is also of the form (2.8), i.e.,

$$\mathscr{S}_H(t_{n+1},x_i,u_i^{n+1},u)_i \equiv \frac{1}{\tau}\sup_{a\in A}\Big\{\widetilde{M}^{a,n+1}u^{n+1} - \widetilde{G}^{a,n}(u^n,u^{n-1})\Big\}_i$$

(where $\widetilde{M}^{a,n+1}\in\mathbb{R}^{J\times J}$ and $\widetilde{G}^{a,n}(u^n,u^{n-1})\in\mathbb{R}^J$)

(iii) $\sup_{a\in A^J}\|(\widetilde{M}^{a,n+1})^{-1}\|_\infty$ is bounded by a constant independent of τ and Δx.

We could have asked for a consistency error of orders $O(\tau^k+\Delta x^{k'})$ which are different in time and space. However, in this work we will use the same order of consistency in τ and Δx.

2.2.1 Examples of high order schemes

Different choices for the high order scheme are possible. In this paper, we consider mainly second order schemes ($k=2$) based on finite differences (see Lemma 2.9).

We first focus on a particular second order Backward Difference Formula (BDF2) both in time and space as follows (here for the one-dimensional case), and then discuss the Crank-Nicolson (CN) scheme, also of second order.

Remark 2.5. *Note that (A3) will hold for the BDF2 scheme using $\alpha=1$ in Remark 2.4, and for the CN scheme by using $\alpha=\frac{1}{2}$, which justifies the introduction of this parameter α.*

<u>BDF2 Scheme</u>. *For $n\geq 1$, the scheme is defined by:*

$$\begin{aligned}\frac{3u_i^{n+1}-4u_i^n+u_i^{n-1}}{2\tau} + \sup_{a\in A}\Big\{ &-\frac{1}{2}\sigma^2(t_{n+1},x_i,a)D^2u_i^{n+1} + b^+(t_{n+1},x_i,a)D^{1,-}u_i^{n+1} \\ &- b^-(t_{n+1},x_i,a)D^{1,+}u_i^{n+1} + f(t_{n+1},x_i,a)u_i^{n+1} + \ell(t_{n+1},x_i,a)\Big\} \ =\ 0,\end{aligned} \tag{2.32}$$

where D^2u_i corresponds to the usual second order approximation (2.21), $b^{\pm}$ denote the positive (resp. negative) part of b as in (2.23), and a BDF2 approximation (with stencil shifted left or right) is used for the first derivative in space:

$$D^{1,-}v_i := \frac{3v_i-4v_{i-1}+v_{i-2}}{2\Delta x} \quad \text{and} \quad D^{1,+}v_i := -\left(\frac{3v_i-4v_{i+1}+v_{i+2}}{2\Delta x}\right). \tag{2.33}$$

The first time step ($n=0$) needs some special treatment and in this case we consider the implicit Euler scheme with the same spatial BDF2 discretization as in (2.32), i.e., the time approximation term $(3u_i^{n+1}-4u_i^n+u_i^{n-1})/(2\tau)$ is replaced by $(u_i^{n+1}-u_i^n)/\tau$.

Remark 2.6. *Of course, considering the equation in a bounded domain, some modification of the scheme might also be necessary at the boundary.*

Remark 2.7. *The particular treatment of the first order drift term $b(t,x,a)v_x$ is in order to take into account possibly vanishing diffusion terms. Indeed, this approximation leads to a second order consistent scheme in both time and space, and appears to be stable even when the diffusion term vanishes, $\sigma(t,x_i,a)\equiv 0$, as in part of the domain in Example 1, Section 4.1.*

To the best of our knowledge, the use of the second order approximations (2.33) in (2.32) is new. It avoids to switch to a first order backward or forward approximation as in [31].

Remark 2.8. *We can attempt to solve* (2.32) *by policy iteration. This is done in the numerical section, with no problem encountered. However, in presence of a non-vanishing drift term b, no theoretical results are available at the moment for justifying the existence of a solution for this BDF2 scheme. In particular, one can easily observe that if the finite discretization matrix in front of u^{n+1} is not an M-matrix, results such as in [12] do not apply.*

For comparison purposes, and because of its popularity for applications in financial mathematics and engineering, the classical CN scheme will also be tested on Example 2, Section 4.2. The precise definition of the scheme used is given in [37] (where centered finite differences are involved to obtain a second order approximation of the first order derivatives, and upwinding when needed for monotonicity).

Proposition 2.3. *The CN and BDF2 schemes satisfy the high-order consistency condition (A3) with $k=2$ and $q=2$ under the following CFL condition:*

$$\|b\|_\infty \frac{\tau}{\Delta x} \le C_1 - \eta \tag{2.34}$$

for some $\eta>0$ independent of τ and Δx, and $C_1=1$ (resp. $C_1=\frac{3}{2}$) for the CN (resp. BDF2) scheme.

Proof. The result follows by Remark 2.4. In particular, properties (i) and (ii) hold by the very definition of the schemes. Moreover, under the CFL condition (2.34), $\widetilde{M}^{a,n+1}$ is a strictly diagonally dominant M-matrix with, for instance in the case when $b(t_{n+1},x_i,a)\ge 0$,

$$|\widetilde{M}^{a,n+1}_{ii}| - \sum_{i\neq j} |\widetilde{M}^{a,n+1}_{ij}| = C_1 + 3\tau\frac{b^+}{2\Delta x} - (4+1)\tau\frac{b^+}{2\Delta x} = C_1 - \frac{\tau}{\Delta x}b^+ \ge \eta \tag{2.35}$$

(where $b^+ := b(t_{n+1},x_i,a)$ here and the diffusion terms vanish). Consequently, using classical estimates for diagonaly dominant matrices,

$$\|(\widetilde{M}^{a,n+1})^{-1}\|_\infty \le \max_i \frac{1}{|\widetilde{M}^{a,n+1}_{ii}| - \sum_{i\neq j}|\widetilde{M}^{a,n+1}_{ij}|} \le \frac{1}{\eta}$$

and the single step stability property, Remark 2.4 (iii), follows. □

Remark 2.9. *Condition* (2.34) *above is not a very restrictive requirement as the scheme is second order consistent in both τ and Δx, such that it is reasonable to keep the ratio fixed. It can be easily verified that, in case the diffusion is not degenerate and a centered finite difference approximation for the drift term is used, condition (2.35) holds in absence of a CFL condition.*

Remark 2.10. *We conjecture that a stronger global stability property in the maximum norm holds for the BDF2 timestepping scheme. Firstly, this is confirmed by the numerical experiments. Secondly, stability results for the maximum norm, in the linear setting with constant coefficients, can be found in [7].*

In the one-dimensional case, the previous schemes can be extended to non-uniform grids (x_i). For instance for D^2u_i one can use the following expression, with $h_i := x_{i+1} - x_i$:

$$D^2u_i = \frac{2}{h_{i-1}+h_i}\left(\frac{1}{h_{i-1}}u_{i-1} - \left(\frac{1}{h_{i-1}}+\frac{1}{h_i}\right)u_i + \frac{1}{h_i}u_{i+1}\right).$$

This finite difference is generally of first order consistent, and of second order if $x_i = q(y_i)$ with a uniform grid y_i and a piecewise smooth Lipschitz function $q(\cdot)$, as we will have in Example 2, Section 4.2.

3 Main results

We first state the main result on the convergence of the filtered schemes introduced in the previous section.

Theorem 3.1. *Let assumptions (A1), (A2') be satisfied. Let u (resp. u_M) denote the solution of the filtered (resp. monotone) scheme. Let v be the viscosity solution of* (2.1).

(i) (Convergence of filtered scheme) If the monotone scheme satisfies the error estimate, for some $\beta > 0$,

$$\max_{0\le n\le N} \left\|u_M^n - v^n\right\|_\infty \le C_1 \max(\tau, \Delta x)^\beta, \tag{3.1}$$

and if in the filtered scheme ε is chosen such that, for some constant $C \ge 0$,

$$0 < \varepsilon \le C \max(\tau, \Delta x)^\beta, \tag{3.2}$$

then the filtered scheme u^n will satisfy the same estimate as for u_M^n, i.e.

$$\max_{0\le n\le N} \left\|u^n - v^n\right\|_\infty \le C \max(\tau, \Delta x)^\beta$$

for some constant $C \ge 0$.

(ii) (First order convergence in regular cases) Assume that the viscosity solution has regularity $v \in C^{2,2+q}([0,T],\mathbb{R})$ (with q as in (A2')). Assume furthermore that ε is chosen such that

$$0 < \varepsilon \le c_0 \max(\tau, \Delta x)$$

for some constant $c_0 \ge 0$. Then a first-order estimate holds for the filtered scheme:

$$\max_{0\le n\le N} \|u^n - v^n\|_\infty \le C \max(\tau, \Delta x)$$

for some constant $C \ge 0$.

(iii) (Local high order consistency) Assume (A3). Let (t,x) be given and assume that v is sufficiently regular in a neighborhood B of (t,x). Assume that

$$\varepsilon := c_0 \max(\tau, \Delta x),$$

with a constant c_0 such that

$$C_M^v := C_M\left(\|v_{tt}\|_{B,\infty} + \sum_{2\le p\le 2+q} \|v_{px}\|_{B,\infty}\right) < c_0 \tag{3.3}$$

where the constants C_M and q are as in (A2'). Then, for sufficiently small τ, Δx, the filtered scheme satisfies the same high order consistency estimate at (t,x) as the high-order scheme.

Remark 3.1. *Motivated by this result, we shall choose ε of the form*

$$\varepsilon = c_0 \max(\tau, \Delta x).$$

If the solution v is sufficiently smooth, we choose c_0 such that

$$c_0 > C_M^v$$

(where the constant C_M^v is as in the left-hand side of (3.3)) in order to get the high order consistency (iii). Moreover, in general the monotone scheme will satisfy a bound of the form (3.1) for some $\beta \in]0, 1]$, and then this choice of ε ensures that (3.2) is satisfied and this gives the convergence of the filtered scheme for the case of nonsmooth solutions.

We give a proof of Theorem 3.1 at the end of this section, and start by proving some preliminary results on the monotone scheme presented in Section 2.1.

First of all we give an elementary result for solving implicit schemes. Let Φ be such that

$$\Phi(x) = \sup_{a \in A}(M^a x - G^a) \quad \text{in } \mathbb{R}^J. \tag{3.4}$$

In every timestep of the monotone scheme (2.7) with (2.8), an equation of the type $\Phi(x) = 0$ has to be solved for $x = u^{n+1} \in \mathbb{R}^J$.

Notice that the supremum in (3.4) may not be attained in general if the maps $a \to M^a$ and $a \to G^a$ are not continuous.[1]

However, considering a maximizing sequence (M^{a_j}, G^{a_j}) such that $\lim_{j\to\infty} M^{a_j} x - G^{a_j} = \Phi(x)$, since (M^{a_j}, G^{a_j}) is bounded in a finite dimensional space, it is possible to extract a convergent subsequence so that $(M^{a'_j}, G^{a'_j}) \to (M^*, G^*)$ and $M^* x - G^* = \Phi(x)$. Hence defining

$$Q := \{(M^a, G^a),\ a \in A^J\},$$

we can write, for any $x \in \mathbb{R}^J$:

$$\Phi(x) = \max_{(M,G)\in\bar{Q}}(Mx - G), \tag{3.5}$$

where now the supremum is attained in $\bar{Q}$.

The policy iteration algorithm is then defined as follows:

1. Start from some $x_0 \in \mathbb{R}^J$.
2. Then for $k \geq 0$, define
$$(M^k, G^k) \in \arg\max_{(M,G)\in\bar{Q}}(Mx_k - G),$$
i.e., an element (M^k, G^k) in $\bar{Q}$ such that $M^k x_k - G^k = \Phi(x_k)$.
3. Then take x_{k+1} the solution of
$$M^k x_{k+1} - G^k = 0.$$
4. Iterate from 2. until convergence.

We have the following result, the proof of which is given in Appendix A.

[1] We avoided making a continuity assumptions not only because the PDE coefficients may be discontinuous functions of the control, but also because the discretisation may introduce discontinuities if switches between different schemes are utilised to ensure monotonicity (see, e.g., [31]).

Proposition 3.1. *Let A be a non-empty compact set, and let matrices $M^a \in \mathbb{R}^{J\times J}$, vectors $G^a \in \mathbb{R}^J$ be defined for $a \in A^J$ as in (2.9). Assume that*

$$\forall a \in A^J, \quad (M^a)^{-1} \geq 0 \tag{3.6}$$

and

$$\sup_{a\in A} \|M^a\|_\infty \leq C, \quad \sup_{a\in A^J} \|(M^a)^{-1}\|_\infty \leq C, \quad \sup_{a\in A} \|G^a\|_\infty \leq C \tag{3.7}$$

for some constant $C \geq 0$.

(i) There exists a unique $x \in \mathbb{R}^J$ such that $\Phi(x) = 0$, with Φ from (3.5).

(ii) For any $x_0 \in \mathbb{R}^J$, the policy iteration algorithm converges to x.

(iii) The convergence is superlinear.

Remark 3.2. *The result of Proposition 3.1 is given in [12] under a continuity condition of the maps $a \to M^a$ and $a \to G^a$. It was more recently shown in [31] that the continuity condition is not necessary for (i) and (ii).*

Proposition 3.2 (Stability and monotonicity)**.** *Let (A1) be satisfied.*
(i) For any $(\tau, \Delta x)$ there exists a unique solution of (2.7) (denoted $(u_M^n)_{n\geq 0}$).
(ii) The scheme $\mathscr{S}_M$ is stable in the following sense: there exists a constant $C \geq 0$ (independent of τ and Δx) such that for any $0 \leq n \leq N$

$$\|u_M^n\|_\infty \leq e^{CT}(\|v_0\|_\infty + CT). \tag{3.8}$$

(iii) The scheme is monotone in the sense of Barles and Souganidis [6], i.e.

$$\phi \leq \psi \quad \Rightarrow \quad \mathscr{S}_M(t,x,r,\phi) \geq \mathscr{S}_M(t,x,r,\psi). \tag{3.9}$$

Remark 3.3. *For the monotonicity property (3.9), only assumptions (A1)(i) and (A1)(iv) are needed.*

Remark 3.4. *As a consequence, under assumptions (A1) and (A2) the scheme $\mathscr{S}_M$ (with solution u_M) satisfies the stability, monotonicity and consistency conditions of the convergence theorem of Barles and Souganidis [6]. As long as a comparison principle holds for the HJB equation (1.1), this proves that the monotone scheme u_M converges to the (unique) viscosity solution. For numerical purposes we choose to deal with a bounded domain, and in principle this would require a precise statement for the viscosity solution of the related HJB equation with specific boundary conditions. However it is not the focus of the paper to go into such a study. We will rather assume that the monotone scheme deals correctly with the boundary conditions and focus on obtaining a high-order scheme in the interior of the domain.*

Proof of Proposition 3.2. (i) The existence and uniqueness of a solution for (2.7) follows from Proposition 3.1.

(ii) Using assumptions (A1)(iii) and (A1)(v) one has for $0 < \tau \leq 1$

$$\begin{aligned} \|u_M^{n+1}\|_\infty &\leq (1 + C_1\tau)(C_2\tau + \|u_M^n\|_\infty) \\ &\leq (1 + C\tau)\|u_M^n\|_\infty + C\tau \end{aligned} \tag{3.10}$$

(for some constants $C_1, C_2, C \geq 0$). By recursion, the bound (3.8) is obtained.

(iii) Using that

$$\mathscr{S}_M(t_{n+1}, x_i, r, \varphi)_i = \frac{1}{\tau} \sup_{a\in A} \Big\{ M_{ii}^{a,n+1} r + \sum_{j\neq i} M_{ij}^{a,n+1} \varphi(t_{n+1}, x_j) - G^{a,n}(\varphi(t_n, .))_i \Big\},$$

the non-positivity of the off-diagonal elements of $M^{a,n+1}$ (assumption (A1)(i)) and the monotonicity of G (assumption (A1)(iv)), the monotonicity of $\mathscr{S}_M$ follows. □

Proposition 3.3. *Let assumptions (A1)(i) and (A1)(iv) be satisfied.*

(i) For any $\phi, \psi \in \mathbb{R}^J$,

$$\phi \leq \psi \quad \Rightarrow \quad S_M(\phi) \leq S_M(\psi).$$

(ii) Let also assumptions (A1)(iii) and (A1)(v) be satisfied. Then, there exists $C \geq 0$ such that, for any $\psi, \phi \in \mathbb{R}^J$:

$$\|S_M(\phi) - S_M(\psi)\|_\infty \leq (1 + C\tau)\|\phi - \psi\|_\infty. \tag{3.11}$$

Proof. (i) For explicit schemes the result is straightforward. Let us now show that this remains true for any implicit finite difference scheme written in the general form (2.8).

Let the optimizing matrix and vector $(M^{*,n+1}, G^{*,n})$ be such that

$$M^{*,n+1}S_M(\psi) - G^{*,n} = \sup_{a \in A^J} \left(M^{a,n+1}S_M(\psi) - G^{a,n}(\psi) \right) = 0, \tag{3.12}$$

where $(M^{*,n+1}, G^{*,n})$ is obtained as a limit of a convergent subsequence $(M^{a_j,n+1}, G^{a_j,n})_j$ that realizes the supremum in (3.12), and notice that $(M^{*,n+1})^{-1} \geq 0$ from (A1)(i). One also has for any ϕ and $a \in A^J$:

$$M^{a,n+1}S_M(\phi) - G^{a,n}(\phi) \leq 0,$$

so that, with $a = a_j$ and passing to the limit as $j \to \infty$:

$$M^{*,n+1}S_M(\phi) - G^{*,n}(\phi) \leq 0.$$

Hence

$$M^{*,n+1}(S_M(\phi) - S_M(\psi)) \leq G^{*,n}(\phi) - G^{*,n}(\psi). \tag{3.13}$$

If $\phi \leq \psi$, this implies by A1(iv)

$$M^{*,n+1}(S_M(\phi) - S_M(\psi)) \leq 0,$$

and (i) follows from the monotonicity property of $M^{*,n+1}$.

Let us now prove (ii). Again from (3.13) together with assumption (A1)(v) one has

$$M^{*,n+1}(S_M(\phi) - S_M(\psi)) \leq (1 + C_1\tau)\|\phi - \psi\|_\infty,$$

where the quantity on the right-hand side has to be considered as a constant vector. Hence, from assumptions (A1)(i) and (A1)(iii) we conclude to

$$(S_M(\phi) - S_M(\psi)) \leq (1 + C_2\tau)(1 + C_1\tau)\|\phi - \psi\|_\infty \leq (1 + C\tau)\|\phi - \psi\|_\infty.$$

Switching the roles of ϕ and ψ, we arrive at the desired estimate (3.11).

□

Theorem 3.2. *Let assumption (A1) be satisfied. Let u_M and u respectively denote the solution of the monotone and filtered scheme, i.e.:*

$$u_M^{n+1} = S_M(u_M^n), \qquad u^{n+1} = S_F(u^n, u^{n-1}).$$

(i) *There exists a constant $C \geq 0$ such that the following estimate holds:*

$$\max_{0 \leq n \leq N} \|u^n - u_M^n\|_\infty \leq Te^{CT}\varepsilon.$$

(ii) *In particular, if the solution of the monotone scheme u_M converges to the viscosity solution of (1.1) as $(\tau, \Delta x) \to 0$, then the solution of the filtered scheme u also.*

Proof. (ii) is a straighforward consequence of (i) recalling that $\varepsilon \to 0$ as $(\tau, \Delta x) \to 0$. To prove (i), by the very definition of the filtered scheme (2.6) one has

$$u^{n+1} - u_M^{n+1} = S_M(u^n) - S_M(u_M^n) + \varepsilon\tau F(\cdot).$$

Hence, thanks to Proposition 3.3 and the fact that $\|F\|_\infty \leq 1$, it holds:

$$\left\|u^{n+1} - u_M^{n+1}\right\|_\infty \leq (1 + C\tau)\left\|u^n - u_M^n\right\|_\infty + \varepsilon\tau.$$

By recursion, using that $1 + C\tau \leq e^{C\tau}$ we obtain

$$\|u_i^n - u_M^n\|_\infty \leq e^{Ct_n}\big(\|u^0 - u_M^0\|_\infty + t_n\varepsilon\big).$$

Then using the fact $u_M^0 = u^0$, the desired estimate follows. □

Before going on, we state the following preliminary result establishing the first order of convergence of the monotone scheme in the case of smooth solutions.

Lemma 3.3. *Let assumptions (A1), (A2') be satisfied and let u_M denote the solution of the monotone scheme. Assume that the viscosity solution v of (2.1) has regularity $C^{2,2+q}([0,T],\mathbb{R})$ (with q as in (A2')), and let $v_j^n = v(t_n, x_j)$ and $v^n = (v_j^n)_{1\leq j\leq J}$. The following estimate holds:*

$$\max_{0\leq n\leq N} \|u_M^n - v^n\|_\infty \leq C\max(\tau, \Delta x),$$

for some constant $C \geq 0$ independent of τ and Δx.

Proof. For a given n,

$$\sup_{a\in A^J} \frac{1}{\tau}(M^{a,n+1}u_M^{n+1} - G^{a,n}(u_M^n)) \;=\; \mathscr{S}_M(t_{n+1}, x_i, u_M^{n+1}, u_M) \;=\; 0.$$

We can also write

$$\sup_{a\in A^J} \frac{1}{\tau}(M^{a,n+1}v^{n+1} - G^{a,n}(v^n)) = \mathscr{S}_M(t_{n+1}, x_i, v^{n+1}, v).$$

Therefore, by the same argument as in Proposition 3.3(ii), one obtains for any $n \geq 0$ and for some constants $C \geq 0$:

$$\|u_M^{n+1} - v^{n+1}\|_\infty \leq (1 + C\tau)\|u_M^n - v^n\|_\infty + C\tau\|(\mathscr{S}_M(t_{n+1}, x_i, v^{n+1}, v_M))_i\|_\infty.$$

By the consistency assumption (A2') and the fact that v is a classical solution of (2.1), one has $|\mathcal{E}^v_{\mathscr{S}_M}| = |\mathscr{S}_M(t_{n+1}, x_i, v^{n+1}, v)| \leq c_1\max(\tau, \Delta x)$ for some constant c_1. Hence, with $c_2 := C\,c_1$:

$$\|u_M^{n+1} - v^{n+1}\|_\infty \leq (1 + C\tau)\|u_M^n - v^n\|_\infty + c_2\max(\tau, \Delta x).$$

By recursion, recalling that $u_M^0 = v^0$ and using that $1 + C\tau \leq e^{C\tau}$, we obtain

$$\begin{aligned} \|u_M^n - v^n\|_\infty &\leq e^{Ct_n}(\|u_M^0 - v^0\|_\infty + c_2 t_n \max(\tau, \Delta x)) \\ &\leq c_2 T e^{CT}\max(\tau, \Delta x). \end{aligned}$$

□

We can now give a proof of Theorem 3.1.

Proof of Theorem 3.1. (i) follows by Theorem 3.2 and the fact that

$$\left\|u^n - v^n\right\|_\infty \le \left\|u^n - u_M^n\right\|_\infty + \left\|u_M^n - v^n\right\|_\infty \le Te^{CT}\varepsilon + C_1 \max(\tau, \Delta x)^\beta.$$

(ii) is a consequence of Lemma 3.3. In fact, as in (i), we have

$$\left\|u^n - v^n\right\|_\infty \le \left\|u^n - u_M^n\right\|_\infty + \left\|u_M^n - v^n\right\|_\infty \le Te^{CT}\varepsilon + c_2 Te^{CT} \max(\tau, \Delta x)$$

from which we deduce the desired estimate because of the bound on ε.

(iii) The filtered scheme will be equivalent to the high-order scheme ($S_F \equiv S_H$) if

$$\frac{|S_H(v^n, v^{n-1})_i - S_M(v^n)_i|}{\varepsilon\tau} \le 1.$$

Let us first remark that $\frac{1}{\tau}|v_i^{n+1} - S_M(v^n, v^{n-1})_i|$ is of the same order as $\mathcal{E}^v_{\mathscr{S}_M}$. Indeed, being v a classical solution one has

$$\left|\mathscr{S}_M(t_{n+1}, x_i, v_i^{n+1}, v)_i\right| \le \mathcal{E}^v_{\mathscr{S}_M}(\tau, \Delta x).$$

Let $(M^{*,n+1}, G^{*,n})$ be the optimal matrix and vector such that

$$\mathscr{S}_M(t_{n+1}, x_i, v^{n+1}, v) \;=\; \frac{1}{\tau}(M^{*,n+1}v^{n+1} - G^{*,n}(v^n)) \;=\; \frac{1}{\tau}M^{*,n+1}(v^{n+1} - S_M(v^n)).$$

Therefore by assumptions (A1)(i) and (A1)(iii) one has

$$\left|\frac{1}{\tau}(v^{n+1} - S_M(v^n))_i\right| \le \|(M^{*,n+1})^{-1}\|_\infty \mathcal{E}^v_{\mathscr{S}_M}(\tau, \Delta x) \le (1 + C\tau)\mathcal{E}^v_{\mathscr{S}_M}(\tau, \Delta x).$$

Using the high-order consistency assumption (A3) one obtains, for some $k \ge 2$:

$$\begin{aligned}\frac{|S_H(v^n, v^{n-1}) - S_M(v^n)|}{\epsilon\tau} &\le \frac{1}{\epsilon\tau}\left|v^{n+1} - S_M(v^n)\right| + \frac{1}{\epsilon\tau}\left|v^{n+1} - S_H(v^n, v^{n-1})\right| \\ &\le \frac{1}{\epsilon}(1 + C\tau)|\mathcal{E}^v_M(\tau, \Delta x)| + \frac{1}{\epsilon}C\max(\tau, \Delta x)^k\end{aligned}$$

with $|\mathcal{E}^v_M(\tau, \Delta x)| \le C^v_M \max(\tau, \Delta x)$ and

$$C^v_M = C_M\left(\|v_{tt}\|_\infty + \sum_{2\le p\le 2+q} \|v_{px}\|_\infty\right) < c_0.$$

Therefore $C^v_M/c_0 < 1$ and for $\max(\tau, \Delta x)$ sufficiently small,

$$\frac{|S_H(v^n, v^{n-1}) - S_M(v^n)|}{\epsilon\tau} \;\le\; \frac{C^v_M}{c_0}(1 + C\tau) + C\max(\tau, \Delta x)^{k-1} \;\le\; 1.$$

The desired result follows. □

4 Numerical tests

In this section, we test the filter on three numerical examples which were chosen to illustrate different features.

In the first example, the solution is regular enough for the high order non-monotone scheme to converge and here the filter is not active for appropriately chosen ε – it ensures convergence without diminishing the accuracy. In the second example, a non-smoothness in the initial data causes a non-monotone time-stepping scheme to converge to a wrong value, while the filtered scheme corrects the behaviour at the cost of lower order convergence. Finally, the last example is of a degenerate but smooth two-dimensional diffusion problem, where a local high-order finite difference scheme is necessarily non-monotone, and again the filter can be used to ensure convergence without sacrificing the observed high order.

4.1 Example 1: Mean-variance asset allocation problem

We study the mean-variance asset allocation problem (see [43, 29, 44]) formulated as follows on $\Omega := (x_{min}, x_{max}) \subset \mathbb{R}$:

$$v_t + \sup_{a \in A} \Big(-\tfrac{1}{2}(\sigma a x)^2 v_{xx} - (c + x(r + a\sigma\xi))v_x \Big) = 0, \quad t \in (0,T),\ x \in \Omega \tag{4.1}$$

$$v(0,x) = \Big(x - \frac{\gamma}{2}\Big)^2, \quad x \in \Omega. \tag{4.2}$$

Here, x is the wealth of an investor who controls the fraction a to invest in a risky asset in order to minimize the portfolio variance under a return target. The above equation assumes a Black-Scholes model with volatility σ and Sharpe ratio ξ, r the rate on a risk-free bond and c the contribution rate, γ a measure of the risk aversion. We use the parameters from [43], see Table 1.

r	σ	ξ	c	T	γ
0.03	0.15	0.33	0.1	20	14.47

Table 1: Parameters used in numerical experiments for mean-variance problem.

If bankruptcy ($x < 0$) is not allowed, the PDE (4.1–4.2) holds on $\Omega = (0, \infty)$.

As in [43], the control is assumed to take values in the bounded set

$$A := [0, 1.5].$$

The equation at $x = 0$ simplifies to

$$v_t(t,0) - c v_x(t,0) = 0 \tag{4.3}$$

(see [43] for a discussion), which is a pure transport equation and, since $c > 0$, there is an influx boundary so that *no boundary condition* is needed at $x = 0$.

For numerical purposes, we truncate the computational domain to

$$\Omega = (0, 5).$$

As in [39], at the boundary $x_{max} = 5$ we consider the Dirichlet boundary condition obtained with the solution of the equation associated with the asymptotic optimal control $a \equiv 0$ (obtained for large values of x), leading us to solve $\bar{v}_t - (c + rx)\bar{v}_x = 0$. By using the method of characteristics, we obtain the following boundary value:

$$\bar{v}(t, x_{max}) := v_0\Big(\frac{e^t(c + r x_{max}) - c}{r}\Big).$$

Let (x_j) be a uniform mesh on $[x_{min}, x_{max}]$: $x_j = x_{min} + j\Delta x$, $j = 0, \ldots, J$, with

$$\Delta x = \frac{x_{max} - x_{min}}{J},$$

and $t_n = n\tau$, $\tau = \frac{T}{N}$.

We consider a filtered scheme using the BDF2 scheme (2.32) as the high order scheme, and the implicit Euler scheme (2.20) as the monotone scheme. We choose here $\varepsilon := c_0 \max(\tau, \Delta x)$, with $N = J$ and therefore τ and Δx are of the same order (in particular $\tau = 4\Delta x$).

Figure 1 shows the approximate shape of the value function and of the optimal control computed with the filtered scheme. The optimal control is at the upper bound for small x, then decreases linearly to zero, the lower bound, and then stays constant at 0. A loss of regularity can be observed

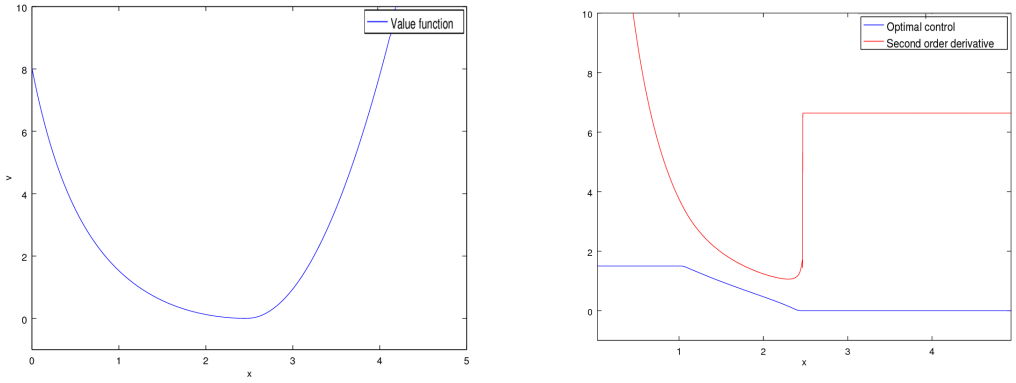


Figure 1: (Example 1) Left: Plot of the value function at time $t = T$. Right: Optimal control (blue) at time $t = T$, and plot of $v_{xx}(T, \cdot)$ (red).

for $x \sim 2.5$ corresponding to a switching point of the control, as shown in Figure 1 (right) where the second order derivative in space has a jump.

The global errors for different norms, obtained with the choice $\varepsilon = c_0\tau$ and $c_0 = 5$, are given in Table 2. Table 3 also gives the local errors computed away from the singularity located at $x \sim 2.5$ (i.e. on $[0, 5]\backslash[2.3, 2.7]$). For the computation of the errors we have used as reference an accurate numerical solution obtained with $N = J = 163840$ and with the high order scheme.

The scheme shows a clear second order of convergence for the L^1- and L^2-norms. The order in the L^∞-norm deteriorates around $x \sim 2.5$. Away from the singularity, the scheme is also second order in the L^∞-norm.

Here the use of the filter secures (via Theorem 3.1) the convergence of the overall approximation towards the viscosity solution, while the BDF2 approximation leads to practically observed second order behavior in both time and space.

We use policy iteration to solve the discretized nonlinear systems, and although convergence is not guaranteed in the case of the non-monotone scheme (see Remark 2.8), we did not encounter any problems.

N	J	Error L^1	order	Error L^2	order	Error L^∞	order	CPU (s)
40	40	2.97E-01	1.94	1.64E-01	1.98	2.24E-01	1.85	0.10
80	80	7.64E-02	1.96	4.15E-02	1.98	5.99E-02	1.90	0.19
160	160	1.95E-02	1.97	1.05E-02	1.98	1.56E-02	1.94	0.34
320	320	4.97E-03	1.97	2.68E-03	1.97	4.01E-03	1.96	0.77
640	640	1.26E-03	1.98	6.86E-04	1.97	1.02E-03	1.98	1.86
1280	1280	3.16E-04	1.99	1.77E-04	1.95	3.35E-04	1.61	5.05
2560	2560	7.93E-05	2.00	4.62E-05	1.94	1.35E-04	1.31	15.43

Table 2: (Example 1) Global errors for the filtered scheme ($c_0 = 5$).

Figure 2 shows the rate of convergence of the scheme depending on the values of c_0 (in this figure the errors are computed only at the point $x = 1$). The thick blue and red lines correspond to order one and two, respectively. For small values of c_0, after some refinements we observe convergence of order one. Applying the analysis of Remark 3.1 to this case, in order to obtain high order we should choose c_0 to be such that, roughly,

$$c_0 > C_M^v, \tag{4.4}$$

N	J	Error L^1	order	Error L^2	order	Error L^∞	order	CPU (s)
40	40	2.85E-01	1.96	1.63E-01	1.99	2.24E-01	1.85	0.10
80	80	7.13E-02	2.00	4.07E-02	2.00	5.99E-02	1.90	0.19
160	160	1.80E-02	1.99	1.02E-02	1.99	1.56E-02	1.94	0.34
320	320	4.52E-03	1.99	2.57E-03	2.00	4.01E-03	1.96	0.77
640	640	1.12E-03	2.01	6.42E-04	2.00	1.02E-03	1.98	1.86
1280	1280	2.79E-04	2.01	1.61E-04	2.00	2.58E-04	1.98	5.05
2560	2560	6.94E-05	2.01	4.03E-05	1.99	6.53E-05	1.98	15.43

Table 3: (Example 1) Local errors for the filtered scheme in the subdomain $[0, 5] \setminus [2.3, 2.7]$ ($c_0 = 5$).

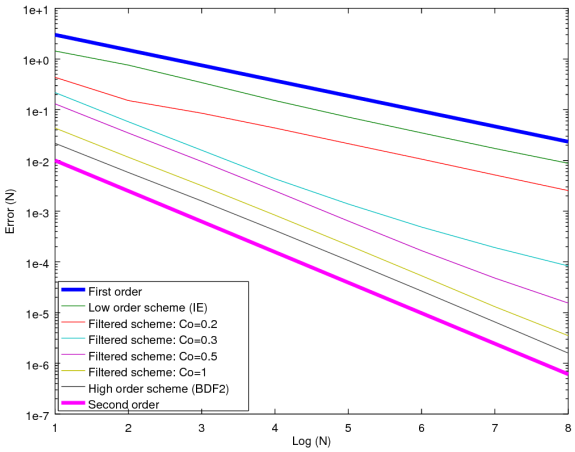


Figure 2: (Example 1) Rate of convergence for the filtered BDF2 scheme, with $\tau = 4\Delta x$, and for different values of c_0 (errors estimated at $x = 1$)

where C_M^v is the constant that appears in the truncation error (2.19) of the monotone scheme.

For the implicit Euler (the monotone) scheme, the following bound is easily obtained, assuming v is sufficiently regular:

$$|\mathcal{E}^v_{\mathscr{S}_M}(\tau, \Delta x)| \leq \tfrac{1}{2}\|v_{tt}\|_\infty \tau + \tfrac{1}{2}\|(c + x(r + a_{\max}\sigma\xi))v_{xx}\|_\infty \Delta x + \tfrac{1}{24}\|\sigma^2 a_{max}^2 x^2 v_{4x}\|_\infty \Delta x^2.$$

Hence, neglecting the Δx^2 term, we obtain a bound in the form of (2.19), i.e., $|\mathcal{E}^v_{\mathscr{S}_M}| \leq C_M^v \tau$ for $\tau = 4\Delta x$, with the constant

$$C_M^v = \tfrac{1}{2}\|v_{tt}\|_\infty + \tfrac{1}{2}\|(c + x(r + a_{\max}\sigma\xi))v_{xx}\|_\infty.$$

Approximating the derivatives on the right-hand side by using the numerical solution, one finds the rough upper bound $C_M^v \leq 40$. This means that we should choose $c_0 \geq 40$ (and $\varepsilon = c_0 \max(\tau, \Delta x)$ for the filter) in order to ensure the local error to be of high order. However for all reported results we have numerically observed that there was practically no difference for $5 \leq c_0 \leq 40$.

4.2 Example 2: Butterfly option pricing with uncertain volatility

We consider the price of a financial option with "butterfly spread" payoff, under a Black-Scholes model with uncertain volatility [30], which is given by the worst scenario with respect to all the possible values of the volatility σ in the interval $[\sigma_{\min}, \sigma_{\max}]$. This problem is associated with the following PDE in $\mathbb{R}_+ \times (0, T)$ for some expiry $T > 0$:

$$v_t + \sup_{\sigma \in \{\sigma_{\min}, \sigma_{\max}\}} \Big(-\frac{\sigma^2}{2} x^2 v_{xx} \Big) - rxv_x + rv = 0, \tag{4.5}$$

$$v(0, x) = v_0(x), \tag{4.6}$$

where

$$v_0(x) := \max(x - K_1, 0) - 2\max(x - (K_1 + K_2)/2, 0) + \max(x - K_2, 0).$$

We follow the example presented in [37], with parameters given by Table 4, on the computational domain $[0, 200]$, using the same non-uniform space grid, shown in Figure 3. The corresponding number of grid points is $J = 60 \times 2^k$, for $k = 0, 1, 2, \ldots$. The time steps are uniform, given by $N = 25 \times 2^k$ and $\tau = \frac{T}{N}$.

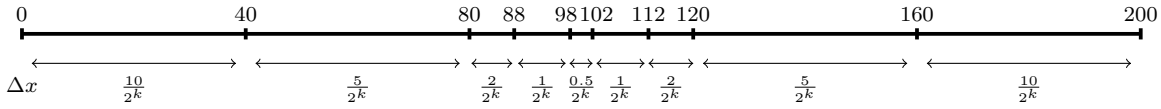


Figure 3: (Example 2) Non-uniform mesh steps for $k = 0, 1, 2, \ldots$.

r	$\sigma_{\min}$	$\sigma_{\max}$	T	K_1	K_2
0.1	0.15	0.25	0.25	90	110

Table 4: Parameters used in numerical experiments for uncertain volatility problem.

Figure 4 shows the result obtained using the CN scheme (left) as in [37], the corresponding result obtained adding a filter to the CN scheme (center), and the BDF2 scheme (right). The monotone scheme used for filtering is the standard implicit Euler finite difference scheme (2.20).

In [37], it is shown that, in absence of a CFL condition constraining τ to be of the order of Δx^2, the Crank-Nicolson (CN) scheme is not monotone, and may not converge to the unique viscosity

solution of the problem. This is also illustrated in Figure 4 (left). We refer to [37] for the details of the CN scheme used.

Convergence results are given in Table 5, for the CN scheme, the "CN-Rannacher" scheme (see below), the BDF2 scheme (2.32), the filtered CN scheme (with $\varepsilon := 50\Delta x_{\min}$) and the filtered BDF2 scheme (2.32) (with $\varepsilon := 50\Delta x_{\min}$). The table reports the L^∞-norm of the error using as reference an accurate numerical solution computed with the BDF2 scheme with $N = 12800$ and $J = 30720$. Here, the CN scheme is not convergent to the correct value. However, by considering the error estimated as differences of two successive values obtained at $x = 100$, and the corresponding "order" of convergence (see Table 5), one may observe a slow convergence – towards a wrong value, as already explained in [37]. As the solution stays bounded, the lack of convergence does not appear to be the result of an instability.

The CN-Rannacher scheme corresponds to a CN scheme with Rannacher time-stepping, i.e., the fully implicit Euler scheme is used for computing u^1 and u^2 (first two steps $n = 0, 1$), then it switches to the CN scheme (for computing u^{n+1} for steps $n \geq 2$), as in [37].

On this example, the filter is able to correct the wrong behavior of the CN scheme, but the overall order of convergence is not greater than one. This is due to the fact that the filter is applied in a quite wide area, as seen in Figure 4, where the dots on the x-axis in the middle panel indicate the mesh points where the filter is active. In particular, the measure of the area where the filter is applied does not diminish as the mesh is refined and therefore the contribution to the overall error from the low order scheme dominates.

Table 5 also compares the orders of convergence. Both the CN-Rannacher scheme and the BDF2 scheme appear to converge to the correct viscosity solution with second order. However, for these last two schemes, because of the lack of monotonicity, convergence is not theoretically established.

In order to ensure convergence to the viscosity solution, we can apply the filter to the BDF2 scheme. The convergence for different values of c_0 is shown in Figure 5. For this example, the unboundedness of the derivatives for $t \downarrow 0$ leads to the unboundedness of the truncation error and subsequently of the constant C_M^v in (3.3). This prevents us from finding a parameter ε such that the high order of convergence is achieved in the relevant region of the domain. The filter is unable to distinguish between a convergent and non-convergent scheme, as it uses only local information about the absolute size of the difference between the high and low order scheme relative to the stepsize.

N	CN		CN-Rannacher		BDF2		CN+filter		BDF2+filter	
	value	"order"	error	order	error	order	error	order	error	order
25	1.884	-	3.38E-02	-	3.19E-02	-	5.58E-01	-	3.19E-02	-
50	1.060	-	9.51E-03	1.82	9.53E-03	1.74	3.34E-01	0.74	9.50E-03	1.75
100	0.957	0.32	2.38E-03	1.99	2.58E-03	1.88	1.10E-01	1.60	2.88E-03	1.72
200	0.884	0.50	5.94E-04	2.00	6.71E-04	1.94	5.10E-02	1.11	1.07E-03	1.43
400	0.835	0.55	1.48E-04	2.00	1.71E-04	1.97	2.01E-02	1.34	3.79E-04	1.50
800	0.802	0.58	3.69E-05	2.00	4.30E-05	1.99	1.87E-02	0.10	1.97E-04	0.95
1600	0.780	0.62	9.11E-06	2.01	1.07E-05	2.00	1.48E-02	0.34	9.84E-05	1.00
3200	0.767	0.64	2.15E-06	2.08	2.55E-06	2.06	8.74E-03	0.76	5.04E-05	0.97

Table 5: (Example 2) Orders of convergence for the CN scheme (convergent towards a wrong solution), the CN scheme with Rannacher time-stepping, the BDF2 scheme, the CN scheme with filter, and the BDF2 scheme with filter ($\varepsilon = 50\Delta x_{\min}$). Here using $N = 25 \times 2^k$ and $J = 60 \times 2^k$, $k = 0, 1, 2....$

Remark 4.1. *Here, the local mesh refinement was chosen in order to reproduce the non-convergence in [37] and then to study the effect of the filter in this setting. The refinement increases the spatial resolution in the area where the solution changes most rapidly. However, the failure of convergence of CN is not restricted to this setting. We also tested the following two mesh constructions: (i) a uniform mesh of the same width as the finest mesh width of the locally refined mesh; (ii) a uniform*

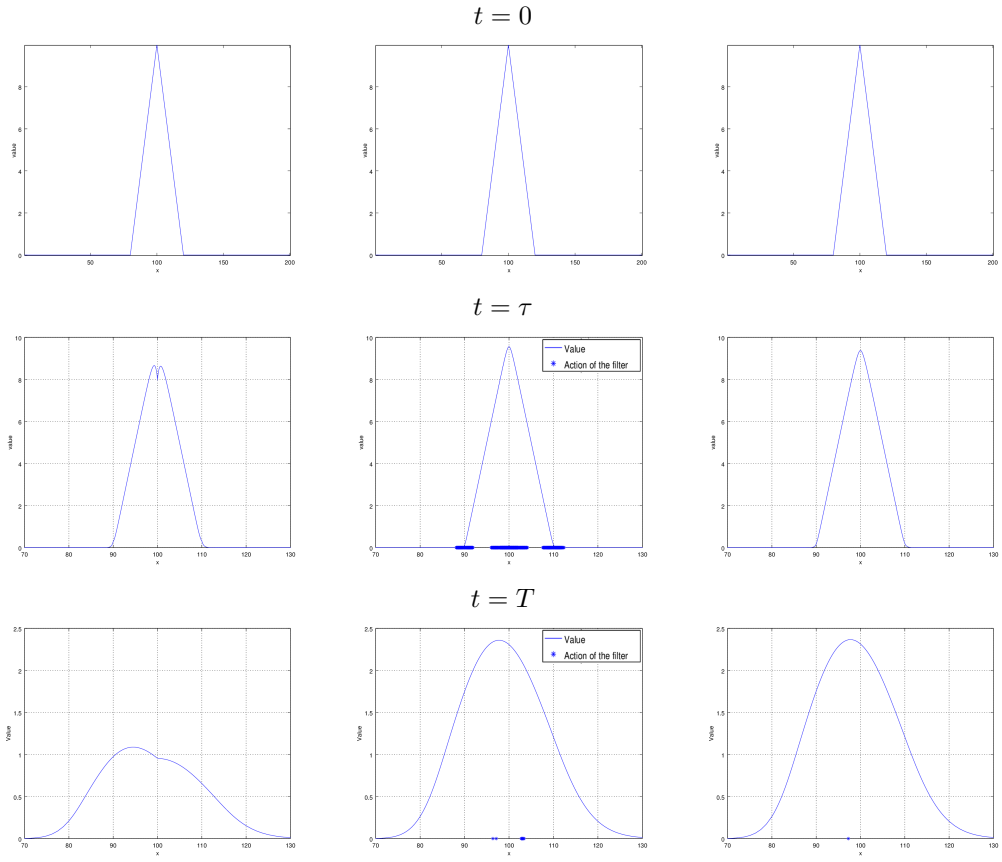


Figure 4: (Example 2) From top to bottom: value function at $t = 0$ (the payoff), after one time-step $t = \tau$, and at terminal time $t = T$. From left to right: CN scheme, CN+filter scheme, BDF2 scheme.

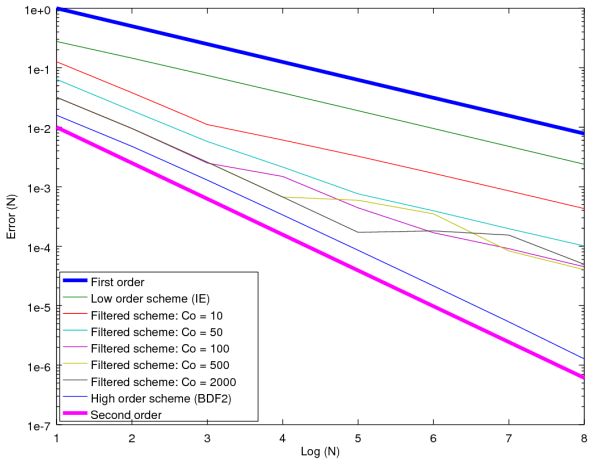


Figure 5: (Example 2) Convergence rate of the L^∞-error obtained for $\varepsilon = c_0 \Delta x_{\min}$ and different values of c_0, using the non uniform mesh defined in [37]. The error is computed by comparison with an accurate numerical solution obtained for $N = 12800$ and $J = 30720$.

mesh in logarithmic coordinates, i.e., the PDE transformed to $y = \log(S/K)$ with $K = (K_1+K_2)/2$, was solved on a uniform mesh, such that again the mesh size at K was of approximately the same size as in the first test (the latter transformation is popular in computational finance as it removes the x-dependence from the coefficients). However, in both tests, the same non-convergence behaviour was observed, as in the left panel of Figure 4.

4.3 Example 3: A two-dimensional example

We now test a filtered scheme on a two-dimensional example set on $\Omega = (-\pi, \pi)^2$:

$$v_t + \sup_{a\in A}\Big(-\frac{1}{2}Tr\big(\sigma_a\sigma_a^T D_x^2 v\big) - \ell(t,x,a)\Big) \;=\; 0, \qquad t\in(0,T), x\in\Omega \tag{4.7a}$$

$$v(0,x) \;=\; 2\sin x_1 \sin x_2, \qquad x\in\Omega \tag{4.7b}$$

with periodic boundary conditions, $T := 0.5$,

$$A := \{a = (a_1, a_2) \in \mathbb{R}^2 : a_1^2 + a_2^2 = 1\},$$

$$\sigma_a := \sqrt{2}\begin{pmatrix} a_1 \\ a_2 \end{pmatrix}, \qquad \ell(t,x,a) := (1-t)\sin x_1 \sin x_2 + (2-t)(a_1^2\cos^2 x_1 + a_2^2 \cos^2 x_2).$$

This problem has the following exact solution

$$v(t,x) = (2-t)\sin x_1 \sin x_2.$$

This example is similar – but more complex as concerns the solution of the optimization problem – to the one discussed in [18, Section 9.3 (B)]. In our case, the minimization over a is non-trivial, in contrast to [18], where the problem reduces to a linear one and the solution satisfies the PDE for any control – not only for the optimal one.

Because of the presence of cross-derivatives and the fact that the diffusion matrix $\sigma_a\sigma_a^T$ may not be diagonally dominant, standard finite difference schemes (such as the seven-point stencil) are generally not monotone. Therefore, for the monotone scheme we consider here a semi-Lagrangian scheme (as in [18]).

Let $N \geq 1$ and let $\tau = T/N$ be the time step. We use a uniform space mesh $x_{ij} = (x_{1,i}, x_{2,j})$ in the two dimensions with mesh steps $\Delta x_1, \Delta x_2$ such that

$$\Delta x_1 = \Delta x_2 = \frac{2\pi}{J}.$$

The monotone scheme gives an approximation u_{ij}^n of $v(t_n, x_{1,i}, x_{2,j})$ as in (2.28), i.e.

$$\frac{u_{ij}^{n+1} - u_{ij}^n}{\tau} = \inf_{a\in A}\left(\frac{[u^n](x_{ij} + \sqrt{\tau}\sigma_a) - 2u_{ij}^n + [u^n](x_{ij} - \sqrt{\tau}\sigma_a)}{2\tau} + \ell(t_n, x_{ij}, a)\right)$$

or, equivalently (as in (2.27))

$$u_{ij}^{n+1} = \inf_{a\in A}\left(\tfrac{1}{2}\sum_{\epsilon=\pm 1}[u^n](x_{ij} + \epsilon\sqrt{\tau}\sigma_a) + \tau\ell(t_n, x_{ij}, a)\right)$$

where $[\,\cdot\,]$ stands for the bilinear – Q_1 – interpolation, and $u_{ij}^0 = v(0, x_{ij})$.

The infimum is approximated using the following discretization of the controls, replacing A by

$$A_P = \{a_k,\ 0 \leq k \leq P-1\}, \quad a_k = \left(\cos\left(\frac{2k\pi}{P}\right), \sin\left(\frac{2k\pi}{P}\right)\right), \tag{4.8}$$

for some $P \in \mathbb{N}$, $P \geq 1$, and we denote also $\Delta a := \frac{2\pi}{P}$ a control mesh step.

The high order scheme we consider here is an implicit finite difference scheme based on the following naive approximation of the second order derivatives for $\phi^n_{ij} \equiv \phi(t_n, x_{1,i}, x_{2,j})$:

$$\partial^2_{xx}\phi_{ij} := \frac{\phi_{i+1,j} - 2\phi_{i,j} + \phi_{i-1,j}}{\Delta x^2}, \qquad \partial^2_{yy}\phi_{ij} := \frac{\phi_{i,j+1} - 2\phi_{i,j} + \phi_{i,j-1}}{\Delta y^2},$$

$$\partial^2_{xy}\phi_{ij} := \frac{\phi_{i+1,j+1} - \phi_{i+1,j-1} + \phi_{i-1,j-1} - \phi_{i-1,j+1}}{4\Delta x \Delta y}.$$

Hence, denoting $\partial^2 u := \begin{pmatrix} \partial^2_{xx} u & \partial^2_{xy} u \\ \partial^2_{xy} u & \partial^2_{yy} u \end{pmatrix}$, the scheme is

$$\frac{u^{n+1}_{ij} - u^n_{ij}}{\tau} + \sup_{a \in A_P} \Big(- \tfrac{1}{2} Tr\big(\sigma_a \sigma_a^T \, \partial^2 u^{n+1}\big)_{ij} + \ell(t_{n+1}, x_{ij}, a) \Big) = 0. \tag{4.9}$$

The problem with the monotonicity of the high-order scheme here does not come from the timestepping scheme, but only from the finite difference approximation of the spatial derivatives.

Equation (4.9) is solved by policy iteration [12]. Even though we have no proof of convergence of the policy iteration algorithm in this setting, we have numerically observed fast convergence.

In order to find a suitable value of c_0 for the filtered scheme (see Remark 3.1), the constant C_M that appears in the truncation error of the monotone scheme needs to be estimated. For the SL scheme above, a consistency estimate similar to (2.17) holds with

$$\big|\mathcal{E}^v_{\mathscr{S}_M}(\tau, \Delta x, \Delta a)\big| \leq \frac{\tau}{2}\|v_{tt}\|_\infty + \frac{2}{3}\tau\|D^4 v\|_\infty + \frac{1}{8}\frac{(\Delta x_1^2 + \Delta x_2^2)}{\tau}\|D^2 v\|_\infty + \frac{\sqrt{10}\|D^2 v\|_\infty + 2}{4}\Delta a^2,$$

where we have denoted $\|D^p v\|_\infty := \max_{k=0,\dots,p} \|\frac{\partial^4 v}{\partial x^k y^{p-k}}\|_\infty$. We take Δa of the same order as Δx and τ (i.e. $\Delta a = \frac{2\pi}{P} = \Delta x_1 = \Delta x_2 = \frac{2\pi}{J}$, and, with $T = 0.5$, $\tau = \frac{1}{2N} \equiv \frac{\Delta x_1}{4\pi}$), so that the error coming from Δa becomes negligible. Using that $v_{tt} = 0$ and $\|D^2 v\|_\infty = \|D^4 v\|_\infty \leq 2$ in this example, this gives the bound $\big|\mathcal{E}^v_{\mathscr{S}_M}(\tau, \Delta x, \Delta a)\big| \leq C^v_M \tau$ with a constant C^v_M such that:

$$C^v_M \leq \frac{4}{3} + 4\pi^2 \simeq 40.$$

In Figure 6 different convergence orders are observed, using $\varepsilon = c_0\tau$ with different values of c_0. We observe convergence of second order already for $c_0 = 0.8$, which is consistent with the upper bound 40 (see Remark 3.1).

Table 6 shows the results for $c_0 = 0.8$. As the table shows, we obtain second order convergence for all norms and refinement levels considered. The computational complexity is $O(NJ^2P)$, as is confirmed in the last column of Table 6.

Figure 7, left, shows the solution, and the right plot the points of activity of the filter for different values of c_0. As soon as $c_0 \geq 0.8$ we do not observe any use of the filter.

5 Conclusions

Filtered schemes are designed to combine the advantages of the guaranteed convergence of low order monotone schemes and the superior accuracy – in regions where the solution is smooth – of higher order non-monotone schemes. The theoretical results in this paper confirm these properties.

In our numerical tests, the schemes delivered the accuracy of the high order scheme if the solution is smooth. For an example with a locally non-smooth solution, the filter was seen to turn a

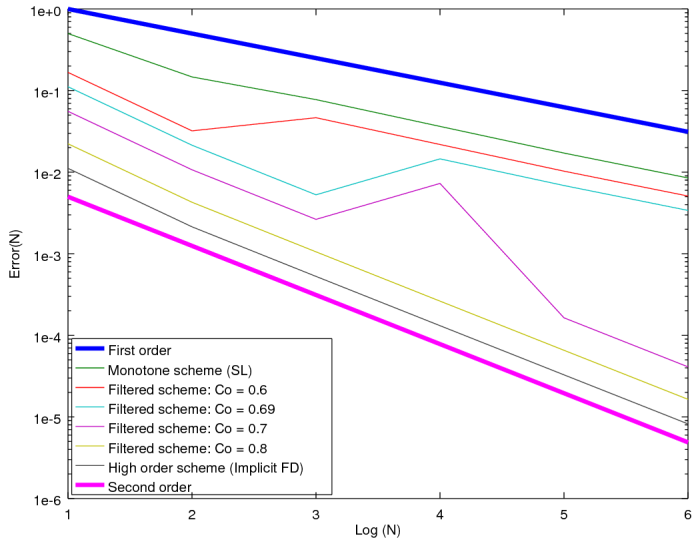


Figure 6: (Example 3) Order of convergence in the L^∞-norm for different values of c_0. Second order convergence is observed for $c_0 \simeq 0.8$ or greater values of c_0.

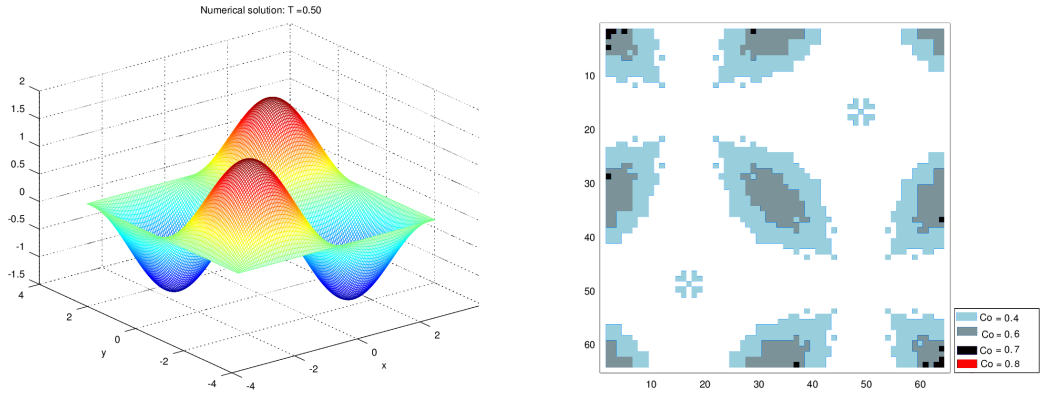


Figure 7: (Example 3) Left: value function at time $T = 0.5$. Right: points of activity of the filter corresponding to different values of c_0. For $c_0 \geq 0.4$ we do not observe any use of the filter.

N	J	P	Error L^1	order	Error L^2	order	Error L^∞	order	CPU (s)
4	4	4	1.13E+00	-	9.27E-01	-	5.57E-01	-	1.3
8	8	8	4.16E-01	1.44	2.79E-01	1.73	1.08E-01	2.37	5.8
16	16	16	1.39E-01	1.58	9.61E-02	1.54	3.13E-02	1.78	71.9
32	32	32	3.76E-02	1.89	2.70E-02	1.83	8.82E-03	1.83	1068.2
64	64	64	9.58E-03	1.97	7.06E-03	1.94	2.31E-03	1.93	19380.0
128	128	128	2.51E-03	1.93	1.83E-03	1.95	6.10E-04	1.92	307450.0

Table 6: (Example 3) Error and order of convergence for the filtered scheme with $c_0 = 0.8$.

divergent higher order time stepping scheme (the Crank-Nicolson scheme) into a convergent scheme, albeit only at the order of the low-order scheme. Although non-monotone high order schemes with better stability (such as the BDF2 scheme) empirically gave second order convergence, the filter reduced this order to one due to singularities of the higher order derivatives of the solution resulting in a wide application of the filter.

Ongoing works concern a more intrinsic choice of the ε parameter that is used in the filtered scheme.

A Proof of Proposition 3.1

Recall that $\Phi(x) := \sup_{a\in A^J}(M^a x - G^a)$, and the following identity in $\mathbb{R}^{J\times J}\times\mathbb{R}^J$:

$$\bar{Q} = \overline{\{(M^a, G^a),\ a \in A^J\}}.$$

Let us furthermore consider the following subset of $\mathbb{R}^{J\times J}\times\mathbb{R}^J$:

$$\bar{Q}_x := \{(M,G) \in \bar{Q},\ \Phi(x) = Mx - G\}.$$

It is clear that all elements (M, G) of $\bar{Q}$ still satisfy the bounds (3.6) and (3.7), that is :

$$M^{-1} \geq 0 \tag{A.1}$$

and

$$\sup_{(M,G)\in\bar{Q}} \|M\|_\infty \leq C, \qquad \sup_{(M,G)\in\bar{Q}} \|M^{-1}\|_\infty \leq C, \qquad \sup_{(M,G)\in\bar{Q}} \|G\|_\infty \leq C \tag{A.2}$$

for some constant $C \geq 0$, that $\bar{Q}$ and $\bar{Q}_x$ (for all $x \in \mathbb{R}^J$) are non empty compact subsets of $\mathbb{R}^{J\times J}\times\mathbb{R}^J$, and that $\Phi(x) = \max_{(M,G)\in\bar{Q}} Mx - G$ for all x.

Notice also that the following holds:

$$\Phi(x) \geq Mx - G \text{ for all } (M,G) \in \bar{Q}, \tag{A.3}$$

as well as the following convexity-like property:

$$\forall (M,G) \in \bar{Q}_x,\ \Phi(y) \geq \Phi(x) + M(y-x). \tag{A.4}$$

The existence and uniqueness of a solution of $\Phi(x) = 0$, as well as as the convergence of Howard's algorithm, in this context of possibly discontinuous M_a and G_a, was already stated in [31]. We give here a proof for the sake of clarity.

For the existence, we first consider, for a given $x_0 \in \mathbb{R}^J$, a sequence (x_k) as in Howard's algorithm: for all k, $(M^k, G^k) \in \bar{Q}_{x_k}$ (i.e. $\Phi(x_k) = M^k x_k - G^k$ with $(M^k, G^k) \in \bar{Q}$), and $x_{k+1} := (M^k)^{-1}G^k$. We remark that $x_k \geq x_{k+1}$ in $\mathbb{R}^J$, $\forall k \geq 1$, i.e., $(x_k)_{k\geq 1}$ is decreasing. Also $x_{k+1} = (M^k)^{-1}G^k$ is bounded (by using the bounds (A.2)). Hence x_k is convergent (pointwise decreasing and bounded from below), towards some $x \in \mathbb{R}^J$. The function Φ is Lipschitz continuous by usual arguments and using the bounds (A.2). Passing to the limit in $\Phi(x_k) = M^k x_k - G^k = M^k x_{k+1} - G^k + O(x_k - x_{k+1}) = O(x_k - x_{k+1})$ gives $\Phi(x) = 0$. Hence any sequence of Howard's algorithm converges to a solution of $\Phi(x) = 0$, which also proves the existence of a solution.

For the uniqueness, consider x, y such that $\Phi(x) = \Phi(y)$. There exists $(M^x, G^x) \in \bar{Q}$ as well as $(M^y, G^y) \in \bar{Q}$ such that $\Phi(x) = M^x x - G^x$ and $\Phi(y) = M^y y - G^y$. Hence $M^x x - G^x = M^y y - G^y \geq M^x y - G^x$ (where we have used (A.3)). So $M^x(x-y) \geq 0$. By using the fact that M_x is a monotone matrix, we obtain $x - y \geq 0$. In the same way we can show that $x - y \leq 0$, hence $x = y$ and the result is proved.

We now focus on the proof of the superlinear convergence by following the lines of [12]. Let $e_k := x_k - x$. Since $(x_k)_{k\geq 1}$ is decreasing, $e_{k+1} \geq 0$. Also we notice that

$$x_{k+1} = x_k - (M^k)^{-1}(M^k x_k - G^k) = x_k - (M^k)^{-1}\Phi(x_k),$$

hence

$$0 \leq e_{k+1} = e_k - (M^k)^{-1}\Phi(x_k).$$

On the other hand, for any $(M, G) \in \bar{Q}_x$, by using (A.4) it holds

$$\Phi(x_k) \geq \Phi(x) + M(x_k - x) = Me_k,$$

and we obtain the following inequalities in $\mathbb{R}^J$, for all $(M, G) \in \bar{Q}_x$

$$0 \leq e_{k+1} \leq (I - (M^k)^{-1}M)e_k. \tag{A.5}$$

Let us remark that $\lim_{k\to\infty} d((M^k, G^k), \bar{Q}_x) = 0$. Indeed, assume this is not the case. We could extract a subsequence of $\bar{Q}$ (still denoted (M^k, G^k)) such that $d((M^k, G^k), \bar{Q}_x) \geq \delta > 0$, $\forall k$. By using compactness arguments, we can extract a convergent subsequence $(M^k, G^k) \to (M, G)$ and the limit is still in $\bar{Q}$, with $d((M, G), \bar{Q}_x) \geq \delta$. Passing to the limit in $\Phi(x_k) = M^k x_k - G^k$ we obtain $\Phi(x) = Mx - G$ and therefore $(M, G) \in \bar{Q}_x$, which is a contradiction.

Hence we can choose a sequence of matrices $(M^{*,k}, G^{*,k})$ of $\bar{Q}_x$ such that $\lim_{k\to\infty} \|M^{*,k} - M^k\|_\infty = 0$ (as well as $\lim_{k\to\infty} \|G^{*,k} - G^k\|_\infty = 0$). Therefore $I - (M^k)^{-1}M^{*,k} \to 0$ and in combination with (A.5), we deduce that $e_{k+1} = o(e_k)$ which is the desired superlinear convergence property. □

References

[1] M. Assellaou, O. Bokanowski, and H. Zidani. Error estimates for second order Hamilton-Jacobi-Bellman equations. Approximation of probabilistic reachable sets. *Discrete Contin. Dyn. Syst.*, 35(9):3933–3964, 2015.

[2] S. Augoula and R. Abgrall. High order numerical discretization for Hamilton-Jacobi equations on triangular meshes. *J. Sci. Comput.*, 15(2):197–229, 2000.

[3] G. Barles and E.R. Jakobsen. On the convergence rate of approximation schemes for Hamilton-Jacobi-Bellman equations. *ESAIM: M2AN*, 36:33–54, 2002.

[4] G. Barles and E.R. Jakobsen. Error bounds for monotone approximation schemes for Hamilton-Jacobi-Bellman equations. *SIAM J. Numer. Anal.*, 43(2):540–558, 2005.

[5] G. Barles and E.R. Jakobsen. Error bounds for monotone approximation schemes for parabolic Hamilton-Jacobi-Bellman equations. *Mathematics of Computation*, 74(260):1861–1893, 2007.

[6] G. Barles and P.E. Souganidis. Convergence of approximation schemes for fully nonlinear second order equations. *Asymptotic Anal.*, 4:271–283, 1991.

[7] T. Beale. Smoothing properties of implicit finite difference methods for a diffusion equation in maximum norm. *SIAM J. Numer. Anal.*, 47(4):2476–2495, 2009.

[8] J.-D. Benamou, F. Collino, and J.-M. Mirebeau. Monotone and consistent discretization of the Monge-Ampère operator. *Math. Comp.*, 85(302):2743–2775, 2016.

[9] O. Bokanowski and K. Debrabant. High order finite difference schemes for some nonlinear diffusion equations with an obstacle term. Preprint, 2016.

[10] O. Bokanowski, M. Falcone, R. Ferretti, D. Kalise, L. Grüne, and H. Zidani. Value iteration convergence of ϵ-monotone schemes for stationary Hamilton-Jacobi equations. *Discrete and Continuous Dynamical Systems – Serie A*, 35(9):4041–4070, 2015.

[11] O. Bokanowski, M. Falcone, and S. Sahu. An efficient filtered scheme for some first order Hamilton-Jacobi-Bellman equations. *SIAM J. Sci. Comput.*, 38(1):A171–A195, 2015.

[12] O. Bokanowski, S. Maroso, and H. Zidani. Some convergence results for Howard's algorithm. *SIAM Journal on Numerical Analysis*, 47(4):3001–3026, 2009.

[13] J.F. Bonnans, E. Ottenwaelter, and H. Zidani. Numerical schemes for the two dimensional second-order HJB equation. *ENSAIM: M2AN*, 38:723–735, 2004.

[14] J.F. Bonnans and H. Zidani. Consistency of generalized finite difference schemes for the stochastic HJB equation. *SIAM J. Numer. Anal.*, 41(3):1008–1021, 2003.

[15] F. Camilli and M. Falcone. An approximation scheme for the optimal control of diffusion processes. *RAIRO Modél. Math. Anal. Numér.*, 29(1):97–122, 1995.

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

[17] M.G. Crandall and P.-L. Lions. Convergent difference schemes for nonlinear parabolic equations and mean curvature motion. *Numerische Mathematik*, 75(1):17–41, 1996.

[18] K. Debrabant and E. R Jakobsen. Semi-Lagrangian schemes for linear and fully non-linear diffusion equations. *Mathematics of Computation*, 82(283):1433–1462, 2013.

[19] X. Feng, C.-Y. Kao, and T. Lewis. Convergent finite difference methods for one-dimensional fully nonlinear second order partial differential equations. *Journal of Computational and Applied Mathematics*, 254:81–98, 2015.

[20] P. A. Forsyth and G. Labahn. Numerical methods for controlled Hamilton-Jacobi-Bellman PDEs in finance. *Journal of Computational Finance*, 11(2):1–44, 2007.

[21] B. D. Froese and A. M. Oberman. Convergent filtered schemes for the Monge-Ampère partial differential equation. *SIAM J. Numer. Anal.*, 51(1):423–444, 2013.

[22] S. K. Godunov. A difference method for numerical calculation of discontinuous solutions of the equations of hydrodynamics. *Matematicheskii Sbornik*, 89(3):271–306, 1959.

[23] M. Jensen and I. Smears. On the convergence of finite element methods for Hamilton–Jacobi–Bellman equations. *SIAM J. Numer. Anal.*, 51(1):137–162, 2013.

[24] M. Kocan. Approximation of viscosity solutions of elliptic partial differential equations on minimal grids. *Numerische Mathematik*, 72(1):73–92, 1995.

[25] N. V. Krylov. On the rate of convergence of finite-difference approximations for Bellman's equations. *St. Petersburg Math. J.*, 9:639–650, 1997.

[26] N. V. Krylov. On the rate of convergence of finite-difference approximations for Bellman's equations with variable coefficients. *Probab. Theory Relat. Fields*, 117:1–16, 2000.

[27] H. J. Kushner and P. G. Dupuis. *Numerical methods for stochastic control problems in continuous time*, volume 24. Springer, 2013.

[28] R. J. LeVeque. *Numerical methods for conservation laws*. Springer Science & Business Media, 1992.

[29] D. Li and W.-L. Ng. Optimal dynamic portfolio selection: multiperiod mean variance formulation. *Mathematical Finance*, 10:387–406, 2000.

[30] T. Lyons. Uncertain volatility and the risk-free synthesis of derivatives. *Applied Mathematical Finance*, 2(2):117–133, 1995.

[31] K. Ma and P. A. Forsyth. An unconditionally monotone numerical scheme for the two factor uncertain volatility model. *IMA J. Numer. Anal.*, 2016. doi: 10.1093/imanum/drw025.

[32] J.L. Menaldi. Some estimates for finite difference approximations. *SIAM J. Control Optim.*, 27:579–607, 1989.

[33] J.-M. Mirebeau. Minimal stencils for discretizations of anisotropic pdes preserving causality or the maximum principle. *SIAM J. Numer. Anal.*, 54(3):1582–1611, 2016.

[34] A. M. Oberman. Convergent difference schemes for degenerate elliptic and parabolic equations: Hamilton–jacobi equations and free boundary problems. *SIAM J. Numer. Anal.*, 44(2):879–895, 2006.

[35] A. M. Oberman and T. Salvador. Filtered schemes for Hamilton-Jacobi equations: A simple construction of convergent accurate difference schemes. *J. Comp. Phys.*, 284:367–388, 2015.

[36] C. W. Oosterlee. On multigrid for linear complementarity problems with application to American-style options. *Electronic Transactions on Numerical Analysis*, pages 165–185, 2003.

[37] D.M. Pooley, P.A. Forsyth, and K.R. Vetzal. Numerical convergence properties of option pricing PDEs with uncertain volatility. *IMA Journal on Numerical Analysis*, 23(2):241–267, 2003.

[38] C. Reisinger. The non-locality of Markov chain approximations to two-dimensional diffusions. *Mathematics and Computers in Simulation*, 2016. doi: 10.1016/j.matcom.2016.06.001.

[39] C. Reisinger and P.A. Forsyth. Piecewise constant policy approximations to Hamilton–Jacobi–Bellman equations. *Applied Numerical Mathematics*, 103:27–47, 2016.

[40] I. Smears and E. Süli. Discontinuous Galerkin finite element methods for time-dependent Hamilton–Jacobi–Bellman equations with Cordes coefficients. *Numerische Mathematik*, 133(1):141–176, 2016.

[41] S. P. van der Pijl and C. W. Oosterlee. An ENO-based method for second-order equations and application to the control of dike levels. *J. Sci. Comput.*, 49(2):462–492, 2012.

[42] J.M. Varah. A lower bound for the smallest singular value of a matrix. *Linear Algebra and its Applications*, 11:3–5, 1975.

[43] J. Wang and P.A. Forsyth. Numerical solution of the Hamilton-Jacobi-Bellman formulation for continuous time mean variance asset allocation. *Journal of Economic Dynanmics and Control*, 34:207–230, 2010.

[44] X. Zhou and D. Li. Continuous time mean variance portfolio selection: A stochastic LQ framework. *Applied Mathematics and Optimization*, 42:19–33, 2000.