# Backward differentiation formula finite difference schemes for diffusion equations with an obstacle term

Olivier Bokanowski, Kristian Debrabant

# BACKWARD DIFFERENTIATION FORMULA FINITE DIFFERENCE SCHEMES FOR DIFFUSION EQUATIONS WITH AN OBSTACLE TERM

OLIVIER BOKANOWSKI AND KRISTIAN DEBRABANT

Abstract. Finite difference schemes, using Backward Differentiation Formula (BDF), are studied for the approximation of one-dimensional diffusion equations with an obstacle term, of the form

$$\min(v_t - a(t,x)v_{xx} + b(t,x)v_x + r(t,x)v, v - \varphi(t,x)) = f(t,x).$$

For the scheme building on the second order BDF formula (BDF2), we discuss unconditional stability, prove an $L^2$-error estimate and show numerically second order convergence, in both space and time, unconditionally on the ratio of the mesh steps. In the analysis, an equivalence of the obstacle equation with a Hamilton-Jacobi-Bellman equation is mentioned, and a Crank-Nicolson scheme is tested in this context. Two academic problems for parabolic equations with an obstacle term with explicit solutions and the American option problem in mathematical finance are used for numerical tests.

**Keywords**: diffusion equation, obstacle equation, viscosity solution, numerical methods, finite difference scheme, Crank Nicolson scheme, Backward Differentiation Formula, high order schemes.

## 1. Introduction

We consider a second order partial differential equation with an obstacle term, of the following form:

$$\min(v_t + \mathcal{A}v, v - \varphi(t,x)) = f(t,x), \quad t \in (0,T), \quad x \in \Omega, \tag{1a}$$

$$v(0,x) = v_0(x), \quad x \in \Omega, \tag{1b}$$

with

$$\mathcal{A}v := -\frac{1}{2}\sigma^2(t,x)v_{xx} + b(t,x)v_x + r(t,x)v. \tag{2}$$

We will assume that $b$, $r$, $\sigma$, $f$, $\varphi$ and $v_0$ are Lipschitz continuous functions with respect to all variables, and also $v_0(x) \geq \varphi(0,x) + f(0,x)$ for compatibility reasons with (1a), and $\Omega$ is a subset of $\mathbb{R}$.

When $\Omega = \mathbb{R}$, the solution $v$ can be defined as the unique uniformly continuous viscosity solution of (1) in the viscosity sense (see [28] for a precise statement, for the case of $x$-dependent obstacle functions and $f \equiv 0$, using even less restrictive assumptions on the remaining data, and these results can easily be generalized to the case of $(t,x)$-dependent obstacle functions and $f \not\equiv 0$). The PDE (1) can also be considered on a bounded domain $\Omega = (X_{min}, X_{max})$ with Dirichlet boundary conditions, see [23, 24, 8] and Section 2. For the well-posedness of (1), a variational framework can also be used [14, 1].

In the recent years there has been a lot of interest in the approximation of such obstacle problems. Related to stochastic optimal stopping time problems, we will

consider in particular

$$\min(v_t - \frac{1}{2}\lambda^2 x^2 v_{xx} - rxv_x + rv,\ v - \varphi(x)) = 0, \quad t \in (0,T),\ x \in \Omega, \tag{3a}$$

$$v(0,x) = \varphi(x), \quad x \in \Omega, \tag{3b}$$

with $\Omega = (0,\infty)$, with constant coefficients $\lambda > 0$, $r > 0$, $f = 0$, and with initial data identical to the obstacle function. The American put option problem in mathematical finance corresponds in particular to the case of the initial data (or "payoff" function) $\varphi(x) := \max(K - x, 0)$. For this problem it is known that the solution presents a singular point $x_s(t)$ moving with time, such that $v(t,x) = \varphi(x)$ for $x \le x_s(t)$, and $v(t,x) > \varphi(x)$ for $x > x_s(t)$, and $t \to x_s(t)$ has a Hölder continuity behavior near $t = 0$ (see Remark 4.10 as well as [11], [2], and [1, Chap 6]). Some results on the structure of the interface related to (1) can also be found in [4] (and see related references).

A finite element scheme in the American option setting has been considered by Jaillet, Lamberton and Lapeyre in [19], where its convergence is also proved under conditions on the mesh steps. For a comprehensive study of finite difference schemes as well as finite element schemes in this context we refer to Achdou and Pironneau [1]. In relation with the obstacle problem, a finite volume method is also studied in Berton and Eymard [3]. In connection with the present work, Windcliff, Forsyth, and Vetzal applied in [33] a second order backward differentiation formula (BDF2) scheme to shout options, which can be understood as a sequence of interdependent American option type problems. Also Oosterlee [26] applied BDF2 in the context of the American option problem, in combination with a multigrid approach (see also Oosterlee et al. [27]). Le Floc'h [22] applied the trapezoidal rule combined with BDF2 as a one-step method (TR-BDF2) to the American option problem.

In relation with viscosity theory, a precise error analysis is given in [20] for monotone finite difference schemes and semi-Lagrangian schemes.

Now we remark that in the case of $v_0 \equiv \varphi + f$ (with $\Omega = \mathbb{R}$) in (1b), and for an operator of the form

$$\mathcal{A}v := -\frac{1}{2}\sigma^2(x)v_{xx} + b(x)v_x + r(x)v, \tag{4}$$

i. e., with coefficients and source term $f = f(x)$ which do not depend on time (and which are otherwise Lipschitz continuous) the solution $v$ of (1) is also the unique viscosity solution of the following Hamilton-Jacobi-Bellman (HJB) equation:

$$v_t + \min(0, \mathcal{A}v) = f(x), \quad t \in (0,T), \quad x \in \Omega, \tag{5a}$$

$$v(0,x) = \varphi(x) + f(x), \quad x \in \Omega. \tag{5b}$$

(For the well-posedness of (5) in the viscosity framework, see [28].) The equivalence between (1) and (5) was signaled to us by R. Eymard. It is proved in Martini [25]. A sketch of the proof of independent interest is given in Appendix A (see Remark A.1).

In this article, we first study in Section 2 two elementary Crank-Nicolson (CN) schemes: a classical CN scheme adapted to the obstacle problem (1), and an other CN scheme adapted to the PDE (5). Although these CN schemes are both second order consistent (and their results even agree in certain cases), we numerically observe that they tend to switch back to first order convergence for bad ratios of the mesh parameters (corresponding to large time steps or "high CFL numbers"). It is known that a change of variable in time, as in [29], or the use of refined time steps near singularity (i. e. near $t = 0$), as in [13], can be used as a remedy to recover second order convergence. However these remedies somehow correspond to using smaller time steps, which one may want to avoid. Stability results exist for the CN scheme, in the $L^\infty$ setting, for the approximation of the linear heat equation [31].

However, to the best of our knowledge, there is no convergence proof for the CN finite difference schemes adapted to the obstacle equation (1) without assuming a CFL condition of the form $\tau/h^2$ small enough (where $\tau$ is the time step and $h$ a mesh step).

To circumvent some of these problems, we then consider the use of the Backward Differentiation Formula (BDF) for the approximation of the time derivative $v_t$, adapted to the obstacle problem. In Section 3, we introduce implicit BDF schemes in the same way as Windcliff, Forsyth, and Vetzal [33] and Oosterlee [26]. In particular second- and third-order BDF schemes are considered. When formulated on an obstacle problem (1), these schemes are non-linear and implicit, but they can be solved by using a simple Newton-type algorithm. In Section 4, a new unconditional $L^2$ stability estimate is obtained in the case of the second order BDF obstacle scheme (BDF2). This is achieved by using estimates similar to the "Gear" scheme for parabolic PDEs (see for instance [1], see also [9]) with some new ingredients in order to deal with the non-linearity coming from the obstacle term. We then obtain also a new error estimate in an $L^2$ norm. This estimate holds under some specific assumptions on the regularity of the exact solution $v$ which allows $v_{xx}$ bounded but possibly discontinuous at some finite number of singular points $(t, y_j(t))_{1\leq j\leq p}$ that do not evolve too rapidly.

In Section 5, two academic models are introduced, with explicit solutions, one of them being very close to the American option model. These models allow us to study precisely and more easily the numerical convergence and allow for a slightly smoother behavior of the interface (compared to the American option model). This allows also to observe third order behavior for a third order BDF obstacle scheme (BDF3) on a specific model with bounded $u_{xxx}$ derivative at the free boundary.

Appendix A is devoted to a sketch of the proof for the equivalence between PDE (1) and PDE (5) in case the coefficients are not time dependent.

Our study concerns here only one-dimensional obstacle problems, but the proposed schemes based on BDF approximations can be extended to higher dimensions (see [26, 6]).

**Acknowledgments.** We are very grateful for the many helpful comments of the referees, especially concerning the analysis and the useful references related to the CN scheme, and for remarks that helped us improve the convergence result for the BDF2 scheme.

## 2. CN finite difference schemes revisited

In this section we revisit the CN schemes and related approaches for a diffusion equation in presence of an obstacle term. Although the presented schemes are all theoretically second order consistent in smooth regions (in a sense that is made precise in Lemma 2.4), we will show that the order may numerically deteriorate and switch back to first order for "high CFL" numbers, i. e., for large time steps with respect to space steps. (As mentioned in the introduction, a change of variable in time, as in [29], or the use of refined time steps near the $t = 0$ singularity, as in [13], can lead back to second order convergence). The BDF scheme presented in Section 3 will not suffer from this problem.

For the numerical approximation of (1) we will consider $\Omega = (X_{min}, X_{max})$ together with Dirichlet boundary conditions:

$$v(t, X_{min}) = v_\ell(t), \quad t \in (0, T), \tag{6a}$$

$$v(t, X_{max}) = v_r(t), \quad t \in (0, T). \tag{6b}$$

We consider a uniform mesh with $J \geq 1$ points inside:

$$x_j = X_{min} + jh, \quad j = 0, \ldots, J+1,$$

where $h := \frac{X_{max} - X_{min}}{J+1}$. Let $N \geq 1$, $\tau = \frac{T}{N}$ and $t_n = n\tau$.

We shall say that we have a "high CFL number" when $\frac{\tau}{h} \gg 1$ (or $\frac{J}{N} \gg 1$), compared to a situation where $\frac{\tau}{h} \simeq 1$ (or $\frac{J}{N} \simeq 1$).

Denoting $v_j^n := v(t_n, x_j)$, we consider the following centered finite difference approximation for the operator $\mathcal{A}v$:

$$(\mathcal{A}v)(t_n, x_j) \simeq \frac{1}{2}\sigma^2(t_n, x_j)\left(\frac{-v_{j-1}^n + 2v_j^n - v_{j+1}^n}{h^2}\right) + b(t_n, x_j)\frac{v_{j+1}^n - v_{j-1}^n}{2h} + r(t_n, x_j)v_j^n. \tag{7}$$

The diffusion part will always dominate the advection part ($\frac{1}{2}\frac{\sigma^2}{h^2} \geq \frac{|b|}{2h}$) to avoid stability issues with the centered approximation. Note that for the American put option problem this requires $h \leq x_1\lambda^2/r$.

*Remark* 2.1. 4th order finite difference approximations for $v_x$ and $v_{xx}$ can also be used instead of (7), in particular for the numerical tests, as detailed in Section 5.2.

Let us denote by $A^{(n)}u^n + q^n$ the approximation of $\mathcal{A}v(t_n, \cdot)$ on a given set of grid points, with $u^n = (u_1^n, \ldots, u_J^n)^T$, where $u_j^n$ are approximations of $v_j^n$, $A^{(n)} = (a_{i,j}^{(n)})_{1 \leq i,j \leq J}$ with

$$a_{i,i-1}^{(n)} := -\beta_i^n - \gamma_i^n, \quad i = 2, \ldots, J, \tag{8a}$$

$$a_{i,i}^{(n)} := 2\beta_i^n + r(t_n, x_i), \quad i = 1, \ldots, J, \tag{8b}$$

$$a_{i,i+1}^{(n)} := -\beta_i^n + \gamma_i^n, \quad i = 1, \ldots, J-1, \tag{8c}$$

where $\beta_i^n := \frac{1}{2h^2}\sigma^2(t_n, x_i)$ and $\gamma_i^n := \frac{b(t_n, x_i)}{2h}$, and

$$q^n := \left((-\beta_1^n - \gamma_1^n)u_0^n, 0, \ldots, 0, (-\beta_J^n + \gamma_J^n)u_{J+1}^n\right)^\top,$$

and with given Dirichlet boundary conditions, for $n = 0, \ldots, N$:

$$u_0^n = v_\ell(t_n) \quad \text{and} \quad u_{J+1}^n = v_r(t_n). \tag{9}$$

The matrix $A^{(n)}$ is in general time-dependent, but for simplicity of presentation we will write $A^{(n)} \equiv A \equiv (a_{i,j})$ without explicit time-dependency. The vector $q^n$ may depend on the time also because of the time dependency in the boundary conditions (9). We have second order consistency in space, that is, assuming $v$ sufficiently regular,

$$(Av^n + q^n)_j = (\mathcal{A}v)(t_n, x_j) + O(h^2).$$

A first simple CN scheme for the obstacle equation (1) is, for $n = 0, \ldots, N-1$ and $j = 1, \ldots, J$, given by

$$\mathcal{S}_j^{1,n}(u) := \min\left(\frac{u_j^{n+1} - u_j^n}{\tau} + \frac{1}{2}(Au^{n+1} + Au^n)_j + q_j^{n+1/2} - f_j^{n+1/2}, u_j^{n+1} - \varphi_j^{n+1} - f_j^{n+1}\right) = 0, \tag{10}$$

where we have denoted

$$\varphi_j^{n+1} := \varphi(t_{n+1}, x_j) \quad \text{and} \quad f_j^p := f(t_p, x_j), \ p \in \{n + \frac{1}{2}, n+1\}$$

and use the boundary conditions (9) and initial condition

$$u_j^0 := v_0(x_j), \quad 1 \leq j \leq J. \tag{11}$$

Looking now at equation (5), an other possible CN scheme is

$$\mathcal{S}_j^{2,n}(u) := \min\Bigg(\frac{u_j^{n+1}-u_j^n}{\tau}+\frac{1}{2}(Au^{n+1}+Au^n)_j+q_j^{n+1/2}-f_j^{n+1/2}, \\ \frac{u_j^{n+1}-u_j^n}{\tau}-f_j^{n+1/2}\Bigg)=0, \qquad 1\le j\le J, \quad (12)$$

initialized with $u_j^0 := v_0(x_j)$. Because $\tau > 0$, this scheme is also equivalent to

$$\min\Bigg(\frac{u_j^{n+1}-u_j^n}{\tau}+\frac{1}{2}(Au^{n+1}+Au^n)_j+q_j^{n+1/2}-f_j^{n+1/2}, \\ u_j^{n+1}-u_j^n-\tau f_j^{n+1/2}\Bigg)=0, \qquad 1\le j\le J. \quad (13)$$

*Remark* 2.2. For both schemes, the unknown $u^{n+1}$ is unique, well defined, and can be obtained by using fix point methods. One can use the fact that there exists a unique solution of the obstacle problem $\min(Bx-\delta, x-g)=0$ as soon as, for instance, $B$ is strictly diagonally dominant with $B_{ii}>0$ (see [1]).

*Remark* 2.3. If furthermore $B$ is a strictly diagonally dominant $M$-matrix (i. e. $B_{ij}\le 0$ for all $i\neq j$ and $B_{ii} > \sum_{j\neq i}|B_{ij}|$ for all $i$) then a Newton-like algorithm [5] can be implemented, which is particularly efficient for solving obstacle problems exactly (up to machine precision) in a few number of iterations. We refer also to [18] for convergence of semi-smooth Newton methods or related algorithms applied to solve discretized PDE obstacle problems. Note that the number of iterations before convergence may increase linearly with the number of mesh points [5]. Penalization methods for solving the obstacle problem can also be used in order to approximate the equation with a controlled penalization error and then significantly reduce the number of needed iterations [13, 34, 30].

For the linear part $v_t+\mathcal{A}v$, the CN scheme appears to be second order consistent only precisely at time $t_{n+1/2}$, while the obstacle term $v_j^{n+1}-\varphi_j^{n+1}$ is evaluated at time $t_{n+1}$ in (10), so appears to be only first order consistent with the value $v(t_{n+1/2},x_j)-\varphi(t_{n+1/2},x_j)$. Hence the consistency error seems to be in general of order $O(\tau)+O(h^2)$ but not better. The following lemma shows that a better order may hold.

**Lemma 2.4.** (*i*) *The CN scheme* (10) *is second order consistent in time and space, in the following sense: for any regular* $v$ *that is solution to* (1a) *it holds*

$$\mathcal{S}_j^{1,n}(v)=O(\tau^2+h^2).$$

(*ii*) *The CN scheme* (12) *is second order consistent in time and space, in the following sense: for any regular* $v$ *it holds*

$$\mathcal{S}_j^{2,n}(v)=\min(v_t+\mathcal{A}v-f, v_t-f)(t_n,x_j)+O(\tau^2+h^2).$$

*Proof.* (*i*) $v$ regular implies

$$\frac{v_j^{n+1}-v_j^n}{\tau}+\frac{1}{2}(Av^{n+1}+Av^n)_j+q_j^{n+1/2} \\ =(v_t+\mathcal{A}v)(t_{n+1/2},x_j)+O(\tau^2\|v_{3t}\|_\infty)+O(h^2(\|v_{3x}\|_\infty+\|v_{4x}\|_\infty)). \quad (14)$$

Inserting (14) into (10) we obtain the estimate

$$\mathcal{S}_j^{1,n}(v)=\min((v_t+\mathcal{A}v)(t_{n+1/2},x_j)-f_j^{n+1/2}, \\ v(t_{n+1},x_j)-\varphi_j^{n+1}-f_j^{n+1})+O(\tau^2+h^2). \quad (15)$$

Now we consider three possible cases. First case: $v(t_{n+1},x_j)=\varphi_j^{n+1}+f_j^{n+1}$. Since $v_t+\mathcal{A}v-f\geq 0$ by (1), it follows $\mathcal{S}_j^{1,n}(v)=O(\tau^2+h^2)$.

Second case: $v(t_{n+1},x_j)>\varphi_j^{n+1}+f_j^{n+1}$ and $v(t_{n+1/2},x_j)>\varphi_j^{n+1/2}+f_j^{n+1/2}$. In that case, since $v$ is solution to (1) at $(t_{n+1/2},x_j)$, it holds

$$(v_t+\mathcal{A}v-f)(t_{n+1/2},x_j)=0,$$

using (15) yields therefore

$$\begin{aligned}\mathcal{S}_j^{1,n}(v) &= \min(0,\ v(t_{n+1},x_j)-\varphi_j^{n+1}-f_j^{n+1})+O(\tau^2+h^2)\\ &= O(\tau^2+h^2).\end{aligned}$$

Third case: $v(t_{n+1},x_j)>\varphi_j^{n+1}+f_j^{n+1}$ and $v(t_{n+1/2},x_j)=\varphi_j^{n+1/2}+f_j^{n+1/2}$. Since $v(t,x_j)\geq\varphi(t,x_j)+f(t,x_j)$ for all $t$, $t\to v(t,x_j)-\varphi(t,x_j)-f(t,x_j)$ reaches a minimum at $t=t_{n+1/2}$, and by using the regularity of $v$, $f$ and $\varphi$ we obtain $v_t(t_{n+1/2},x_j)-\varphi_t(t_{n+1/2},x_j)-f_t(t_{n+1/2},x_j)=0$. Then $v(t_{n+1},x_j)-\varphi(t_{n+1},x_j)-f(t_{n+1},x_j)=v(t_{n+1/2},x_j)-\varphi(t_{n+1/2},x_j)-f(t_{n+1/2},x_j)+O(\tau^2)=O(\tau^2)$, from which we deduce

$$\mathcal{S}_j^{1,n}(v)=\min((v_t+\mathcal{A}v-f)(t_{n+1/2},x_j),\ 0)+O(\tau^2+h^2)\ =O(\tau^2+h^2).$$

(*ii*) Inserting (14) and

$$\frac{v_j^{n+1}-v_j^n}{\tau}=v_t(t_{n+1/2},x_j)+O(\tau^2\|v_{3t}\|_\infty)$$

into (12) we obtain the second order estimate

$$\mathcal{S}_j^{2,n}(v)=\min(v_t+\mathcal{A}v-f,v_t-f)(t_{n+1/2},x_j)+O(\tau^2+h^2).$$

□

*Remark* 2.5. In the proof of (*i*) we use the fact that $v$ is solution of the PDE, as well as the regularity of $t\to v(t,x_j)$ in a region where it switches from $v(t_n,x_j)=\varphi_j^n$ to $v(t_n,x_j)>\varphi_j^n$, and we know in general that this corresponds to a jump of $v_t$ (or $v_{xx}$). Therefore this analysis cannot be applied in general. If $v$ is regular, without assuming that $v$ is solution of (1), then by (15) we would obtain only a first order estimate in time.

On the other hand, the following assertions hold:

**Lemma 2.6.** *Assume that $f$ and $\varphi$ are independent of $t$.*

(*i*) *If for given $u^n$ value, the solution $u^{n+1}$ of the scheme* (10) *satisfies $u^{n+1}\geq u^n$ (with $u^n\geq\varphi+f$), then $u^{n+1}$ is also solution of the scheme* (12) *starting from $u^n$.*

(*ii*) *In particular, if $u^1\geq u^0$ and the following conditions hold:*

- *the vector $q$ and the matrix $A$ do not depend on time,*
- *the matrix $I-\frac{\tau}{2}A$ is positive componentwise,*
- *the matrix $I+\frac{\tau}{2}A$ is a strictly diagonally dominant $M$-matrix,*

*then the solution of the scheme* (10) *satisfies $u^{n+1}\geq u^n$ for all $n$, and thus by (i) schemes* (10) *and* (12) *give identical values.*

*Proof.* Proof of (*i*): Let $c^n:=\frac{1}{2}(Au^{n+1}+Au^n)_j+q_j^{n+1/2}-f_j$, so that the first scheme (10) reads $\min(\frac{u_j^{n+1}-u_j^n}{\tau}+c_j^n,\ u_j^{n+1}-\varphi_j-f_j)=0$. Assuming that $u^{n+1}$ is solution of scheme (10), with $u^{n+1}\geq u^n$, if $u_j^{n+1}=\varphi_j+f_j$, then it is clear that also $u_j^n=\varphi_j+f_j=u_j^{n+1}$ and therefore since $\frac{u_j^{n+1}-u_j^n}{\tau}+c_j^n\geq 0$ it can be deduced that $\mathcal{S}_j^{2,n}(u)=0$. On the other hand, if $u_j^{n+1}>\varphi_j+f_j$, then it implies $\frac{u_j^{n+1}-u_j^n}{\tau}+c_j^n=0$, from which we conclude $\mathcal{S}_j^{2,n}(u)=0$, so $u^{n+1}$ is also a solution of the scheme (12).

Proof of (*ii*): Denoting $F_n(x) := \min(Bx-\delta_n, x-\varphi-f)$ where $B = I+\frac{\tau}{2}A$ and $\delta_n = (I-\frac{\tau}{2}A)u^n - \tau q - \tau f$, the scheme (10) is equivalent to $F_n(u^{n+1}) = 0$. Because $B$ is an $M$-matrix, the function $F_n$ is monotone in the sense that $F_n(y) \le F_n(z) \Rightarrow y \le z$ (componentwise).

Assume now that $u^n \ge u^{n-1}$ for some $n \ge 1$. Due to $I - \frac{\tau}{2}A \ge 0$, it holds $\delta_n \ge \delta_{n-1}$, and therefore $F_n(y) \le F_{n-1}(y)$ for all $y$. In particular, $F_{n-1}(u^n) = 0 = F_n(u^{n+1}) \le F_{n-1}(u^{n+1})$, and by the monotonicity of $F_{n-1}$ we conclude $u_n \le u_{n+1}$. By induction it follows that $u^{n+1} \ge u^n$ for all $n$. □

*Remark* 2.7. For the American put option problem (3) with $\varphi(x) := \max(K-x, 0)$ the left boundary condition will be $u_0^n = K - X_{min}$, and the right boundary condition $u_{J+1}^n = 0$. Thus the vector $q^{n+1/2} = q$ does not depend on time. Further, for $A$ determined by (7) and $\lambda^2(X_{\min}+h) > rh$, the matrix $I - \frac{\tau}{2}A$ is positive componentwise under the CFL condition $\left(\frac{\lambda^2 X_{\max}^2}{h^2} + r\right)\frac{\tau}{2} \le 1$, and the matrix $I+\frac{\tau}{2}A$ is a strictly diagonally dominant $M$-matrix under the condition $\frac{\tau}{h} < \frac{2+\tau r}{rX_{\max}}$. Finally, with $u^0 = g$, it is easy to see that the scheme (10) satisfies $u^1 \ge u^0 = g$. Thus, by Lemma 2.6, scheme (10) and scheme (12) give identical values.

This explains why the CN scheme (10) gives the same results as scheme (12) for low CFL number.

In order to verify the expected orders, we have tested the CN schemes numerically on the American put option model with initial data

$$\varphi(x) := \max(K-x, 0) \tag{16a}$$

and parameters

$$\lambda = 0.2, \quad r = 0.1, \quad T = 1, \quad K = 100. \tag{16b}$$

In this setting, we observe that the singular point $x_s(T)$ is greater than 80, so for the numerical approximation we have considered the subdomain $\Omega = (X_{min}, X_{max}) \equiv (75, 275)$ and boundary conditions of Dirichlet type:

$$v(t, X_{\min}) = K - X_{\min}, \quad 0 < t < T,$$

and

$$v(t, X_{\max}) = 0, \quad 0 < t < T.$$

We have numerically estimated that the truncation error from the right, using $X_{max} = 275$ instead of $X_{max} = +\infty$, is less than $10^{-8}$.

The errors of the CN schemes in $L^2$, $L^1$ and $L^\infty$ norms are computed at time $t_N = T$ as follows:

$$e_{L^p} := (h\sum_{j=1}^{J} |u_j^N - v_j^N|^p)^{1/p} \text{ and } e_{L^\infty} := \max_{j=1}^{J} |u_j^N - v_j^N|. \tag{17}$$

The reference values are computed using a BDF obstacle scheme of second order and with $N = J+1 = 20480$ that will be made precise in Section 3. In the Newton-like algorithm to solve the obstacle problem $\min(Bx-\delta, x-g) = 0$, we iterated until the numerical approximation $\hat{x}$ fulfilled $\|\min(B\hat{x}-\delta, \hat{x}-g)\|_\infty < 10^{-10}$.

Results are given in Table 1 with discretization parameters $N = J+1$ and $N = (J+1)/10$ (this second case corresponds to large time steps or "high CFL numbers"). Note that the quite restrictive CFL condition in part (ii) of Lemma 2.6 (cmp. Remark 2.7) is not fulfilled.

However, for lower $N$ values (higher CFL numbers) the CN scheme is no more second order and goes back to first order behavior. This is illustrated in Table 1. In particular we observe that the pointwise inequality $u^{n+1} \ge u^n$ is no more true (due

to the fact that the amplification matrix $(I + \frac{\tau}{2}A)^{-1}(I - \frac{\tau}{2}A)$ does not have only positive coefficients anymore).

Now, taking the obstacle to be $u^n$ instead of $\varphi$, hence solving the scheme (12), enforces that $u^{n+1} \geq u^n$. Results for the scheme (12) are similar to that for scheme (10) for low CFL numbers (both schemes give identical values for the case $N = J+1$, here). For higher CFL numbers, results obtained with the scheme (12) differ (see Table 2), but again switch back to first order behavior. They are even less precise than the CN scheme (10) for the $L^1$ and the $L^2$ errors.

| Mesh | | Error $L^1$ | | Error $L^2$ | | Error $L^\infty$ | | time(s) |
|---|---|---|---|---|---|---|---|---|
| $J+1$ | $N$ | error | order | error | order | error | order | |
| 80 | 80 | 7.21e-01 | 1.88 | 1.27e-01 | 1.80 | 4.49e-02 | 1.20 | 0.01 |
| 160 | 160 | 1.42e-01 | 2.35 | 2.28e-02 | 2.48 | 5.29e-03 | 3.09 | 0.01 |
| 320 | 320 | 3.79e-02 | 1.90 | 6.04e-03 | 1.92 | 1.40e-03 | 1.92 | 0.04 |
| 640 | 640 | 1.01e-02 | 1.91 | 1.58e-03 | 1.93 | 3.57e-04 | 1.97 | 0.12 |
| 1280 | 1280 | 2.79e-03 | 1.85 | 4.29e-04 | 1.88 | 9.21e-05 | 1.95 | 0.45 |
| 2560 | 2560 | 7.98e-04 | 1.80 | 1.20e-04 | 1.84 | 2.76e-05 | 1.74 | 1.72 |
| 5120 | 5120 | 2.20e-04 | 1.86 | 3.24e-05 | 1.89 | 6.48e-06 | 2.09 | 7.20 |
| 80 | 8 | 7.93e-01 | 1.66 | 1.45e-01 | 1.62 | 4.90e-02 | 1.51 | 0.00 |
| 160 | 16 | 2.00e-01 | 1.99 | 3.62e-02 | 2.00 | 2.18e-02 | 1.17 | 0.00 |
| 320 | 32 | 6.40e-02 | 1.64 | 1.21e-02 | 1.58 | 9.91e-03 | 1.14 | 0.01 |
| 640 | 64 | 2.31e-02 | 1.47 | 4.38e-03 | 1.47 | 4.75e-03 | 1.06 | 0.03 |
| 1280 | 128 | 8.71e-03 | 1.41 | 1.61e-03 | 1.44 | 2.32e-03 | 1.04 | 0.12 |
| 2560 | 256 | 3.55e-03 | 1.29 | 6.22e-04 | 1.37 | 1.14e-03 | 1.02 | 0.62 |
| 5120 | 512 | 1.50e-03 | 1.24 | 2.49e-04 | 1.32 | 5.65e-04 | 1.01 | 3.41 |

TABLE 1. CN scheme (10) with different mesh parameters $N = J+1$ and $N = (J+1)/10$.

| Mesh | | Error $L^1$ | | Error $L^2$ | | Error $L^\infty$ | | time(s) |
|---|---|---|---|---|---|---|---|---|
| $J+1$ | $N$ | error | order | error | order | error | order | |
| 80 | 80 | 7.21e-01 | 1.88 | 1.27e-01 | 1.80 | 4.49e-02 | 1.20 | 0.01 |
| 160 | 160 | 1.42e-01 | 2.35 | 2.28e-02 | 2.48 | 5.29e-03 | 3.09 | 0.02 |
| 320 | 320 | 3.79e-02 | 1.90 | 6.04e-03 | 1.92 | 1.40e-03 | 1.92 | 0.04 |
| 640 | 640 | 1.01e-02 | 1.91 | 1.58e-03 | 1.93 | 3.57e-04 | 1.97 | 0.13 |
| 1280 | 1280 | 2.79e-03 | 1.85 | 4.29e-04 | 1.88 | 9.21e-05 | 1.95 | 0.44 |
| 2560 | 2560 | 7.98e-04 | 1.80 | 1.20e-04 | 1.84 | 2.76e-05 | 1.74 | 1.79 |
| 5120 | 5120 | 2.20e-04 | 1.86 | 3.24e-05 | 1.89 | 6.48e-06 | 2.09 | 8.30 |
| 80 | 8 | 7.04e-01 | 1.83 | 1.35e-01 | 1.73 | 4.90e-02 | 1.51 | 0.00 |
| 160 | 16 | 8.05e-02 | 3.13 | 2.52e-02 | 2.42 | 1.61e-02 | 1.60 | 0.00 |
| 320 | 32 | 5.17e-02 | 0.64 | 1.05e-02 | 1.26 | 6.51e-03 | 1.31 | 0.01 |
| 640 | 64 | 3.22e-02 | 0.68 | 5.24e-03 | 1.00 | 3.38e-03 | 0.94 | 0.03 |
| 1280 | 128 | 1.83e-02 | 0.82 | 2.72e-03 | 0.94 | 1.76e-03 | 0.95 | 0.11 |
| 2560 | 256 | 9.76e-03 | 0.91 | 1.40e-03 | 0.96 | 8.95e-04 | 0.97 | 0.55 |
| 5120 | 512 | 5.07e-03 | 0.94 | 7.17e-04 | 0.97 | 4.53e-04 | 0.98 | 2.36 |

TABLE 2. CN scheme (12) (for solving (5)) with different mesh parameters $N = J+1$ and $N = (J+1)/10$.

## 3. BDF obstacle schemes

We now consider BDF type approximations for the first derivative $u_t$, leading to implicit schemes. We propose two implicit schemes (BDF2 and BDF3) which have the same complexity as the previous CN implicit schemes but give improved numerical results with respect to precision and to stability. Furthermore a stability and error analysis will be carried out for the BDF2 scheme.

### 3.1. BDF2 obstacle scheme.

Our first scheme is therefore the following two-step implicit scheme (hereafter also referred to as BDF2 obstacle scheme), for $n \geq 1$:

$$\mathcal{H}_j^{n+1}(u) :\equiv \min\left(\frac{3u_j^{n+1} - 4u_j^n + u_j^{n-1}}{2\tau} + (Au^{n+1} + q^{n+1})_j,\ u_j^{n+1} - \varphi_j^{n+1}\right) - f_j^{n+1} = 0, \tag{18}$$

initialized with appropriate $u^0$ and $u^1$ values. Such approximations for the linear term $u_t$, known as BDF approximations, are well known and used in various contexts [10, 15]. For $u^1$, e. g. the implicit Euler method (IE) (corresponding to a first order BDF method)

$$\min\left(\frac{u_j^1 - u_j^0}{\tau} + (Au^1 + q^1)_j,\ u_j^1 - \varphi_j^1\right) - f_j^1 = 0,$$

or the CN scheme (10) with $n = 0$ could be used. In the following, for the numerical tests, the first step $u^1$ is always computed by a CN scheme (see in particular Remark 3.3).

The use of a BDF scheme for a diffusion plus obstacle problem is not new (Windcliff et al [33], Oosterlee et al [26, 27], the idea was also suggested by Seydel in [32, see pages 187 and 217]). To the best of our knowledge, a precise analysis of the scheme was missing so far.

By construction the scheme has the following consistency error, when $v$ is regular, for $v_j^n = v(t_n, x_j)$:

$$\begin{aligned}\mathcal{H}_j^{n+1}(v) &= \min(v_t + \mathcal{A}v, v - \varphi)(t_{n+1}, x_j) - f(t_{n+1}, x_j) \\ &\quad + O(\tau^2 \|v_{3t}\|_\infty) + O(h^2(\|v_{3x}\|_\infty + \|v_{4x}\|_\infty)).\end{aligned} \tag{19}$$

We summarize this in the following Lemma, to be compared to Lemma 2.4.

**Lemma 3.1.** *If $v$ is regular, the BDF scheme* (18) *is second order consistent in time and space with respect to the obstacle problem* (1).

This consistency error justifies the introduction of BDF schemes that precisely approximate $u_t + \mathcal{A}u$ at time $t_{n+1}$ without the need of other particular requirements (there is no requirement that $v_t + \mathcal{A}v = 0$ at previous times $t < t_{n+1}$, which would not hold in the presence of an obstacle term).

Let $\min(X, Y) := (\min(x_j, y_j))_j$ denote the minimum of two vectors $X = (x_j), Y = (y_j)$ of $\mathbb{R}^J$. For convenience, the scheme (18) will also be written as follows:

$$\min\left((I_J + \frac{2}{3}\tau A)\, u^{n+1} - \frac{4}{3}u^n + \frac{1}{3}u^{n-1} + \frac{2}{3}\tau q^{n+1} - \frac{2}{3}\tau f^{n+1},\ u^{n+1} - \varphi^{n+1} - f^{n+1}\right) = 0 \tag{20}$$

with $I_J$ denoting the $J$-dimensional identity matrix. (After subtracting $f^{n+1}$, a multiplication by $\frac{2\tau}{3} > 0$ of the left part of the min term does not change the equation.)

*Remark* 3.2 (Newton's method). As already mentioned before, the scheme can be solved by a semi-smooth Newton method [5]. More precisely, denoting $B := I_J + \frac{2}{3}\tau A$ (a real valued $J \times J$ matrix), $\delta := \frac{4}{3}u^n - \frac{1}{3}u^{n-1} - \frac{2}{3}\tau q^{n+1} + \frac{2}{3}\tau f^{n+1}$, and $g := \varphi^{n+1} + f^{n+1}$, the problem is to solve for $x \in \mathbb{R}^J$

$$\min(Bx - \delta, x - g) = 0 \quad \text{in } \mathbb{R}^J. \tag{21}$$

The matrix $B$ satisfies the conditions of Remark 2.3 ensuring the convergence of Newton's algorithm provided that $\frac{\tau}{h}|b| \leq \frac{3}{2} + \tau r$.

Results for the BDF2 obstacle scheme are given in Table 3 for $N = J+1$ and $N = (J+1)/10$ (larger time steps) using the same parameters as in (16). These results show robustness of the scheme even for large time steps and also an improvement of the convergence with respect to the CN schemes (the order is closer to 2 even for $N = (J+1)/10$). Note that the results indicate second order convergence although by estimate (19) we can only expect second order convergence for solutions that are three times continuously differentiable with respect to time and four times continuously differentiable with respect to space.

| Mesh | | Error $L^1$ | | Error $L^2$ | | Error $L^\infty$ | | time(s) |
|---|---|---|---|---|---|---|---|---|
| $J+1$ | $N$ | error | order | error | order | error | order | |
| 80 | 80 | 7.11e-01 | 1.87 | 1.26e-01 | 1.79 | 4.48e-02 | 1.18 | 0.01 |
| 160 | 160 | 1.35e-01 | 2.40 | 2.18e-02 | 2.53 | 5.09e-03 | 3.14 | 0.01 |
| 320 | 320 | 3.52e-02 | 1.94 | 5.66e-03 | 1.94 | 1.33e-03 | 1.93 | 0.04 |
| 640 | 640 | 9.12e-03 | 1.95 | 1.46e-03 | 1.96 | 3.41e-04 | 1.97 | 0.13 |
| 1280 | 1280 | 2.43e-03 | 1.91 | 3.84e-04 | 1.93 | 8.88e-05 | 1.94 | 0.48 |
| 2560 | 2560 | 6.60e-04 | 1.88 | 1.03e-04 | 1.90 | 2.72e-05 | 1.70 | 1.56 |
| 5120 | 5120 | 1.73e-04 | 1.93 | 2.61e-05 | 1.98 | 5.79e-06 | 2.23 | 7.25 |
| 80 | 8 | 4.56e-01 | 1.71 | 7.75e-02 | 1.49 | 3.59e-02 | 0.35 | 0.00 |
| 160 | 16 | 7.29e-02 | 2.65 | 9.73e-03 | 2.99 | 2.20e-03 | 4.02 | 0.00 |
| 320 | 32 | 2.25e-02 | 1.69 | 3.38e-03 | 1.53 | 8.82e-04 | 1.32 | 0.01 |
| 640 | 64 | 7.31e-03 | 1.62 | 1.17e-03 | 1.53 | 2.98e-04 | 1.56 | 0.02 |
| 1280 | 128 | 2.17e-03 | 1.75 | 3.52e-04 | 1.74 | 8.65e-05 | 1.79 | 0.11 |
| 2560 | 256 | 5.22e-04 | 2.06 | 8.21e-05 | 2.10 | 2.40e-05 | 1.85 | 0.38 |
| 5120 | 512 | 1.07e-04 | 2.29 | 1.39e-05 | 2.56 | 5.79e-06 | 2.05 | 1.24 |

TABLE 3. BDF2 scheme for (1).

*Remark* 3.3. We have numerically observed that if we compute the first step with an IE obstacle scheme (corresponding to BDF1) and the BDF2 scheme is otherwise unchanged for the next steps, then the results are not as clear as in Table 3 (second order convergence does not appear clearly), and with $N = (J+1)/10$ we observe rather first order convergence.

3.2. **BDF3 obstacle scheme.** In the same way, we propose the following three-step (BDF3) implicit scheme, for $n \geq 2$:

$$\mathcal{H}_j^{n+1}(u) :\equiv \min\left(\frac{\frac{11}{6}u_j^{n+1} - 3u_j^n + \frac{3}{2}u_j^{n-1} - \frac{1}{3}u_j^{n-2}}{\tau} + (Au^{n+1} + q^{n+1})_j,\right.$$
$$\left. u_j^{n+1} - \varphi_j^{n+1}\right) - f_j^{n+1} = 0.$$

The scheme may be initialized by any second order approximation for the first two steps $u^1$ and $u^2$, we have chosen the CN scheme for $u^1$ and the BDF2 scheme for $u^2$.

As we have done for the BDF2 scheme, we can multiply the left term by $6\tau$, define $B := 11\, I_J + 6\tau A$, and obtain then an equivalent scheme in the following form, for $n \geq 2$, in $\mathbb{R}^J$:

$$\min\left( Bu^{n+1} - 18u^n + 9u^{n-1} - 2u^{n-2} + 6\tau q^{n+1} - 6\tau f^{n+1},\ u^{n+1} - \varphi^{n+1} - f^{n+1} \right) = 0.$$

The unknown $u^{n+1}$ can be solved by using again a semi-smooth Newton method.

We have observed that the numerical results with the BDF3 scheme are not as good as in Table 3 with the BDF2 scheme for the American option problem (the order of convergence is two for $N = J+1$ and closer to one for $N = (J+1)/10$). Since there is a jump in the second order derivative $v_{xx}$, we do not expect better than second order convergence in this case. We do not have a convergence result for BDF3 since even in the linear case the scheme is known not to be $A$-stable (see Remark 4.8).

The performance of the BDF3 obstacle scheme (using a 4th order approximation in space) will be tested in Section 5.3 on a model problem with a bounded $v_{3x}$ derivative, showing third order in that case.

## 4. Stability and error estimate for the BDF2 scheme

Throughout this section, we will consider the following assumptions:

**Assumption (A1):**

- $a \equiv \frac{1}{2}\sigma^2$, $b$ and $r$ are bounded functions (this follows already from $\sigma, b, r$ being Lipschitz continuous on the finite domain $\Omega$),
- there exists $\eta_0 > 0$ such that $a(t,x) \geq \eta_0 > 0$ for all $t, x$,
- $a$ is Lipschitz continuous in $x$ uniformly w.r.t. $t$, that is:

$$\exists L \geq 0,\ |a(t,x) - a(t,y)| \leq L|x-y|, \quad t \in (0,T),\ (x,y) \in \Omega^2. \tag{22}$$

*Remark* 4.1. For the error analysis, no regularity assumption will be needed neither on the obstacle $\varphi$ nor the source term $f$. Indeed, these terms vanish in the consistency error analysis.

*Remark* 4.2. In the case that the diffusion coefficient may degenerate some analysis may hold without the obstacle term (see [6]).

### 4.1. Stability estimate.

Let us first start by considering an abstract obstacle problem of the form

$$\min(By - \delta, y - g) = 0 \quad \text{for} \quad y \in \mathbb{R}^J,$$

where $B$ is a square matrix of size $J$ and $\delta, g$ are given vectors of $\mathbb{R}^J$. We will use the following elementary result.

**Lemma 4.3.** *For any matrix $B$, the following equivalence holds:*

$$\min(By - \delta, y - g) = 0 \quad \Leftrightarrow \quad y \geq g \text{ and } \left( \langle By - \delta, v - y \rangle \geq 0,\ \forall v \geq g \right). \tag{23}$$

*Proof.* It is known [7] that if $B$ is a positive definite symmetric matrix, the following equivalences hold:

$$\min(By - \delta, y - g) = 0 \quad \Leftrightarrow \quad y \text{ solves } \min_{y \geq g} \frac{1}{2}\langle y, By \rangle - \langle \delta, y \rangle \tag{24}$$

$$\Leftrightarrow \quad y \geq g \text{ and } \left( \langle By - \delta, v - y \rangle \geq 0,\ \forall v \geq g \right). \tag{25}$$

When $B$ is not symmetric, the equivalence between the min equation and (25) is still true:

$\Rightarrow$: For $v \geq g$, $\langle By-\delta, v-y\rangle = \langle By-\delta, v-g\rangle + \underbrace{\langle By-\delta, g-y\rangle}_{=0}$ so is nonnegative since $By-\delta \geq 0$ and $v-g \geq 0$.
$\Leftarrow$: By taking $v = y + \lambda e_j$ with $\lambda \to +\infty$ we get $(By-\delta)_j \geq 0$, hence $By-\delta \geq 0$. Then, $\langle By-\delta, y-g\rangle \geq 0$, and also $\langle By-\delta, y-g\rangle \leq 0$ by taking $v = g$ as a test function in the inequality. Hence $\langle By-\delta, y-g\rangle = 0$. Together with $By-\delta \geq 0$, $y-g \geq 0$, this implies that $\min(By-\delta, y-g) = 0$. □

The idea now is to use the inequality of Lemma 4.3 in order to obtain a stability estimate. For parabolic problems, it is possible to obtain stability estimates in the $L^2$ norm for the Gear (or BDF2) scheme (see for instance [12]). We are going to obtain similar estimates for the scheme (20) applied to the obstacle problem (1).

Let $v(t,x)$ be a regular enough function, $v_j^n := v(t_n, x_j)$, and $\bar\epsilon^n \in \mathbb{R}^J$ be defined by

$$\bar\epsilon_j^n = \frac{1}{2\tau}(3v_j^{n+1} - 4v_j^n + v_j^{n-1}) + (Av^{n+1} + q^{n+1})_j - (v_t + \mathcal{A}v)(t_{n+1}, x_j), \qquad (26)$$

$n = 1, \dots, N-1$. The term $\bar\epsilon^n$ corresponds to a consistency error for the linear part of the PDE, here written in discrete form on the grid mesh.

*Remark* 4.4. If $v$ is continuous but $v_t, v_x, v_{xx}$ are not well defined at $(t_{n+1}, x_j)$, we can still define $\bar\epsilon_j^n$ as follows. We consider a definition of $v_t(t_{n+1}, x_j)$, $v_x(t_{n+1}, x_j)$ and $v_{xx}(t_{n+1}, x_j)$ such that $\bar\epsilon_j^n$ (defined by (26)) satisfies the bound

$$|\bar\epsilon_j^n| \leq C\Big( \|v_t(t_{n+1}, .)\|_{L^\infty(\Omega)} + \|v_x(t_{n+1}, .)\|_{L^\infty(\Omega)} + \|v_{xx}(t_{n+1}, .)\|_{L^\infty(\Omega)} \Big) \qquad (27)$$

with a constant $C \geq 0$ (independent of $n, j$). This bound assumes that the exact derivatives $v_t(t_{n+1}, .)$, $v_{xx}(t_{n+1}, .)$ exist a.e. on $\Omega$ (with possible discontinuities) and are bounded, and this will be considered later on in assumption (A2). For instance, extending the domain of definition of $v_t$, $v_x$ and $v_{xx}$ to whole $\Omega$ by $v_t = v_x = v_{xx} := 0$ at places of non-differentiability is a possible choice.

Then we have

$$\min\big(\frac{1}{2\tau}(3v^{n+1} - 4v^n + v^{n-1}) + Av^{n+1} + q^{n+1} - f^{n+1} - \bar\epsilon^n,\ v^{n+1} - g^{n+1}\big) = 0 \qquad (28)$$

with $g^{n+1} := \varphi^{n+1} + f^{n+1}$. Therefore $v^n$ satisfies a perturbed scheme, as follows:

$$\min\big((I_J + \frac{2\tau}{3}A)v^{n+1} - \frac{4}{3}v^n + \frac{1}{3}v^{n-1} + \frac{2\tau}{3}q^{n+1} - \frac{2\tau}{3}f^{n+1} - \frac{2\tau}{3}\bar\epsilon^n,\ v^{n+1} - g^{n+1}\big) = 0. \qquad (29)$$

*Remark* 4.5. Typically $\bar\epsilon^n$ is of order $O(\tau^2 + h^2)$ where $v$ is regular.

Our aim is now to show a stability estimate in order to control the error $\|u^n - v^n\|_2^2$ in terms of $\sum_{1\leq k\leq n-1} \tau\|\bar\epsilon^n\|^2$.

For a vector $x = (x_j)_{1\leq j\leq J}$, let

$$N(x) := \left( \sum_{j=1}^{J+1} |x_j - x_{j-1}|^2 \right)^{1/2} \qquad (30)$$

(with the convention $x_0 := 0$ and $x_{J+1} := 0$).

The following shows a coercivity bound for the matrix $A$.

**Lemma 4.6.** *Under assumption (A1), there exist $\eta > 0$ and $\gamma \geq 0$ such that for $A$ given by* (8) *and for all $e \in \mathbb{R}^J$:*

$$\langle e, Ae\rangle \geq \eta N(e/h)^2 - \gamma\|e\|_2^2. \qquad (31)$$

*Proof.* Considering $\mathcal{A}u = -a(t,x)u_{xx} + b(t,x)u_x + r(t,x)u$ and hereafter not explicitly mentioning the time variable, it holds

$$A = \frac{1}{h^2}\,\mathrm{tridiag}(-a_i,\ 2a_i,\ -a_i) + \frac{1}{2h}\,\mathrm{tridiag}(-b_i,\ 0,\ b_i) + \mathrm{diag}(r_i)$$

where $a_i = a(x_i)$, $b_i = b(x_i)$ and $r_i = r(x_i)$. By straightforward calculations,

$$\langle e, Ae\rangle = \frac{1}{h^2}\sum_{i=1}^{J+1}(a_ie_i - a_{i-1}e_{i-1})(e_i - e_{i-1}) + \frac{1}{2h}\sum_{i=1}^{J} b_i(e_{i+1} - e_{i-1})e_i + \sum_{i=1}^{J} r_ie_i^2. \tag{32}$$

Now we make use of $|a_i - a_{i-1}| \le Ch$ for some constant $C \ge 0$ (since $a(\cdot)$ is Lipschitz continuous), and $a_i \ge \eta_0$, to obtain:

$$\begin{aligned}\frac{1}{h^2}\sum_{i=1}^{J+1}(a_ie_i - a_{i-1}e_{i-1})(e_i - e_{i-1}) &\ge \frac{1}{h^2}\sum_{i=1}^{J+1}\eta_0(e_i - e_{i-1})^2 - \frac{1}{h}\sum_{i=1}^{J} C|e_i|\,|e_i - e_{i-1}| \\ &\ge \eta_0 N(e/h)^2 - C\|e\|_2 N(e/h). \end{aligned}\tag{33}$$

We have also, by using $e_{i+1} - e_{i-1} = (e_{i+1} - e_i) + (e_i - e_{i-1})$:

$$\sum_{i=1}^{J}|b_i(e_{i+1} - e_{i-1})e_i| \le \|b\|_\infty 2N(e)\|e\|_2.$$

Hence there exists a lower bound of the form:

$$\langle e, Ae\rangle \ge \eta_0 N(e/h)^2 - (C + \|b\|_\infty)N(e/h)\|e\|_2 - C\|e\|_2^2$$

for some constant $C$. Denoting $C' := C + \|b\|_\infty$ and applying the inequality $C'N(e/h)\|e\|_2 \le \frac{\eta_0}{2}N(e/h)^2 + \frac{C'^2}{2\eta_0}\|e\|_2^2$, we finally obtain

$$\langle e, Ae\rangle \ge \frac{\eta_0}{2}N(e/h)^2 - (C + \frac{C'^2}{2\eta_0})\|e\|_2^2$$

which gives the desired lower bound with $\eta = \eta_0/2$ and $\gamma = C + \frac{C'^2}{2\eta_0}$. □

From now on we shall denote the error by

$$e^n := v^n - u^n.$$

**Proposition 4.7.** *Consider the scheme* (20)*, and a perturbed scheme* (29)*. Let $\tau > 0$ be sufficiently small. Then there exist a constant $C_1$ independent of $n$ and a constant $\bar\gamma > 0$ such that for all $t_n \le T$*

$$\begin{aligned} e^{-\bar\gamma t_n}\|e^{n+1}\|_2^2 + \tau\eta\sum_{k=1}^{n} e^{-\bar\gamma t_k}N(e^{k+1}/h)^2 & \\ \le\ C_1\left(\|e^0\|_2^2 + \|e^1\|_2^2 + \tau\sum_{k=1,\dots,n} e^{-\bar\gamma t_k}\|\bar\epsilon^k\|_2^2\right) & \end{aligned}\tag{34}$$

*where $N(e^k/h)$ is defined by* (30)*.*

*Proof of Proposition 4.7.* Let

$$B := I_J + \frac{2\tau}{3}A,$$

and vectors $b_u$, $b_v$ be such that

$$b_u := \frac{4}{3}u^n - \frac{1}{3}u^{n-1} - \frac{2\tau}{3}q^{n+1} + \frac{2\tau}{3}f^{n+1}$$

and

$$b_v := \frac{4}{3}v^n - \frac{1}{3}v^{n-1} - \frac{2\tau}{3}q^{n+1} + \frac{2\tau}{3}f^{n+1}.$$

Then, by Lemma 4.3, the min equation (20) for the exact scheme is equivalent to $u^{n+1} \geq g^{n+1}$ and

$$\left\langle Bu^{n+1} - b_u,\ w - u^{n+1} \right\rangle \geq 0, \quad \forall w \geq g^{n+1}. \tag{35}$$

The min equation (29) for the perturbed scheme is equivalent to $v^{n+1} \geq g^{n+1}$ and

$$\left\langle Bv^{n+1} - (b_v + \frac{2}{3}\tau\bar{\epsilon}^n),\ w - v^{n+1} \right\rangle \geq 0, \quad \forall w \geq g^{n+1}. \tag{36}$$

Taking $w = v^{n+1}$ in (35) gives

$$\left\langle Bu^{n+1} - b_u,\ v^{n+1} - u^{n+1} \right\rangle \geq 0,$$

and $w = u^{n+1}$ in (36) gives

$$\left\langle Bv^{n+1} - (b_v + \frac{2}{3}\tau\bar{\epsilon}^n),\ u^{n+1} - v^{n+1} \right\rangle \geq 0.$$

Combining the last two relations gives

$$\left\langle Be^{n+1} - \frac{4}{3}e^n + \frac{1}{3}e^{n-1} - \frac{2\tau}{3}\bar{\epsilon}^n,\ e^{n+1} \right\rangle \leq 0 \tag{37}$$

and therefore

$$\left\langle 3e^{n+1} - 4e^n + e^{n-1},\ e^{n+1} \right\rangle + 2\tau\langle e^{n+1}, Ae^{n+1}\rangle \leq 2\tau\langle \bar{\epsilon}^n,\ e^{n+1}\rangle. \tag{38}$$

Now let $x_n$, $y_n$ and $z_n$ be defined by

$$x_n := \|e^n\|_2^2, \qquad y_n := \|e^{n+1} - e^n\|_2^2, \qquad z_n := 2\tau\langle Ae^{n+1},\ e^{n+1}\rangle.$$

The following estimate holds:

$$3x_{n+1} - 4x_n + x_{n-1} + 2y_n + 2z_n \leq 2y_{n-1} + 4\tau\langle \bar{\epsilon}^n, e^{n+1}\rangle. \tag{39}$$

To prove (39), we first use the properties $\langle a - b,\ a\rangle = \frac{1}{2}(\|a\|_2^2 + \|a - b\|_2^2 - \|b\|_2^2)$ as well as $\frac{1}{2}\|a + b\|_2^2 \leq \|a\|_2^2 + \|b\|_2^2$, to obtain

$$\begin{aligned}
&2\langle 3e^{n+1} - 4e^n + e^{n-1},\ e^{n+1}\rangle \\
&\quad = \quad 2(4\langle e^{n+1} - e^n,\ e^{n+1}\rangle - \langle e^{n+1} - e^{n-1},\ e^{n+1}\rangle) \\
&\quad = \quad 4(x_{n+1} + y_n - x_n) - (x_{n+1} + \|e^{n+1} - e^{n-1}\|_2^2 - x_{n-1}) \\
&\quad \geq \quad 4(x_{n+1} + y_n - x_n) - (x_{n+1} + 2(y_n + y_{n-1}) - x_{n-1}) \\
&\quad \geq \quad 3x_{n+1} - 4x_n + x_{n-1} + 2y_n - 2y_{n-1}
\end{aligned}$$

and we conclude by using (38).

Let

$$w_n := 4\tau\eta N(e^{n+1}/h)^2.$$

By using the bound $2\tau\langle \bar{\epsilon}^n, e^{n+1}\rangle \leq 2\tau\|\bar{\epsilon}^n\|_2\|e^{n+1}\|_2 \leq \tau\|\bar{\epsilon}^n\|_2^2 + \tau x_{n+1}$ and the coercivity (31), we obtain, for $n \geq 1$:

$$3x_{n+1} - 4x_n + x_{n-1} + 2y_n + w_n \leq 2y_{n-1} + 2\tau\|\bar{\epsilon}^n\|_2^2 + (2\tau + 4\tau\gamma)x_{n+1}. \tag{40}$$

Let

$$\bar{\gamma} := 2 + 4\gamma \quad \text{ and } \quad \beta := \tau\bar{\gamma}.$$

It follows

$$(3 - \beta)x_{k+1} - 4x_k + x_{k-1} + 2y_k + w_k \leq 2y_{k-1} + 2\tau\|\bar{\epsilon}^k\|_2^2, \quad k \geq 1. \tag{41}$$

We multiply (41) by $e^{-k\beta}$ and sum up the inequalities from $k=1$ to $n\geq 1$. Let $f(x)=x^2-4x+3-\beta$ and notice that for $\tau$ small enough (and therefore small $\beta>0$), $f(e^{-\beta})\sim\beta>0$. We deduce that for some constant $C_{01}$

$$\begin{aligned}
&e^{-n\beta}((3-\beta)x_{n+1}-(4-(3-\beta)e^{\beta})x_n)+\sum_{k=1}^{n}e^{-k\beta}w_k\\
&+\sum_{k=2}^{n-1}e^{-(k-1)\beta}f(e^{-\beta})x_k+2\sum_{k=1}^{n}e^{-k\beta}y_k\\
&\qquad\leq\ C_{01}(x_0+x_1)+2\sum_{k=1}^{n}e^{-k\beta}y_{k-1}+2\sum_{k=1}^{n}e^{-k\beta}\tau\|\bar{\epsilon}^k\|_2^2. \qquad (42)
\end{aligned}$$

Using that $e^{-k\beta}y_{k-1}\leq e^{-(k-1)\beta}y_{k-1}$ and $f(e^{-\beta})>0$ we deduce that

$$\begin{aligned}
&e^{-n\beta}((3-\beta)x_{n+1}-(4-(3-\beta)e^{\beta})x_n)+\sum_{k=1}^{n}e^{-k\beta}w_k\\
&\qquad\leq\ C_{01}(x_0+x_1)+2e^{-\beta}y_0+\sum_{k=1}^{n}e^{-k\beta}2\tau\|\bar{\epsilon}^k\|_2^2\\
&\qquad\leq\ C_{02}(x_0+x_1+\tau\sum_{k=1}^{n}e^{-k\beta}\|\bar{\epsilon}^k\|_2^2)=:Q \qquad (43)
\end{aligned}$$

with $C_{02}:=C_{01}+4$ (where we have used that $y_0\leq 2(x_0+x_1)$).

Let us prove that $e^{-\bar{\gamma}t_n}x_n\leq x_1+C_{02}Q$, which will give the desired bound. By using $k\beta{=}k\tau\bar{\gamma}{=}\bar{\gamma}t_k$, we deduce from (43)

$$x_{k+1}\leq\rho x_k+\frac{e^{\bar{\gamma}t_n}Q}{3-\beta},\quad 1\leq k\leq n,$$

where $\rho:=\frac{4-(3-\beta)e^{\beta}}{3-\beta}\sim\frac{1}{3}$ as $\beta=\tau\bar{\gamma}\to 0$. By recursion we get for $1\leq k\leq n$:

$$\begin{aligned}
x_k\ &\leq\ \rho^{k-1}x_1+\frac{e^{\bar{\gamma}t_n}Q}{3-\beta}(1+\rho+\cdots+\rho^{k-2})\\
&\leq\ x_1+\frac{e^{\bar{\gamma}t_n}Q}{3-\beta}\frac{1}{1-\rho}.
\end{aligned}$$

By using this bound for $x_n$ into (43), we obtain the desired result (34) with a possibly different universal constant $C_1$. This concludes the proof of Proposition 4.7. □

*Remark* 4.8 (BDF3 scheme). The previous stability estimate does not extend easily to the BDF3 obstacle scheme. Indeed, it is known that BDF3 is not $A$-stable (as well as any BFD$k$ for $k\geq 3$, see [17, 16]), which prevents the same stability analysis to apply for diffusion equations.

4.2. **Error estimate for the BDF2 scheme.** The following assumptions will be used.

**Assumption (A2).** We assume that there exist an integer $p\geq 0$ and continuous functions $t\to y_i(t)$ for $i=1,\dots,p$ with $y_i(t)\in[X_{min},X_{max}]$ such that, defining for $0\leq\tau<T$

$$\Omega_{\tau,T}:=\{(t,x)\in(\tau,T)\times\Omega,\ x\notin(y_i(t))_{1\leq i\leq p}\}$$

($\Omega_{\tau,T}$ is a subdomain of $(0,T)\times\Omega$), the following holds:

- $(i)$ $(t,x)\to v(t,x)$ is regular (i. e. $C^{2,3}$) on $\Omega_{0,T}$,

- $(ii)$ there exist constants $\alpha_i \geq 0$, $C_i \geq 0$, $i = 1, \ldots, 4$, such that for all $\epsilon > 0$:

$$\|v_t\|_{L^\infty(\Omega_{\epsilon,T})} \leq C_1 \epsilon^{-\alpha_1}, \quad \|v_{xx}\|_{L^\infty(\Omega_{\epsilon,T})} \leq C_3 \epsilon^{-\alpha_3},$$
$$\|v_{tt}\|_{L^\infty(\Omega_{\epsilon,T})} \leq C_2 \epsilon^{-\alpha_2}, \quad \|v_{xxx}\|_{L^\infty(\Omega_{\epsilon,T})} \leq C_4 \epsilon^{-\alpha_4}.$$

*Remark* 4.9. Assumption (A2) allows for $v_{xx}(t,.)$ to have "jumps" at the singular points $x = y_k(t)$.

**Assumption (A3).** There exists $\alpha_0 \in (0,1]$ such that, for $i= 1, \ldots, p$, $t \to y_i(t)$ is $\alpha_0$-Hölder continuous on $[0,T]$.

If the first step of the BDF2 scheme is initialized with the CN scheme, the following assumption will also be needed:

**Assumption (A4).** $v_0$ is Lipschitz continuous and piecewise $C^2$ regular on $\Omega$.

Explicit examples satisfying assumptions (A2)–(A4) will be given in the numerical section, see Remark 4.13.

*Remark* 4.10. For the American put option problem (3) with $\varphi(x) := (K-x)_+$ it is known that there is a unique singular point $y_1(t) \equiv x_s(t)$ such that $v(t,x) = \varphi(x)$ for $x < x_s(t)$, $v(t,x) > \varphi(x)$ for $x > x_s(t)$, and that

$$1 - x_s(t)/K \overset{t \to 0^+}{\sim} \lambda(t|\ln(t)|)^{1/2} \tag{44}$$

(see [2], and [1, Chap 6], as well as [11]). Furthermore, a function of the form of $(t|\ln(t)|)^{1/2}$ satisfies assumption (A3) for any $\alpha < 1/2$. We do not know if the American option problem satisfies (A2). Nevertheless, assumptions (A2)-(A3) allow for closely related problems where $\varphi$ is Lipschitz continuous and piecewise $C^2$ and by allowing some rapidly moving singularities $y_i(t)$ as $t \to 0$ (such as (44)).

In the following error analysis, we consider the continuous $L^2$ norm on $\Omega$, $\|f\|_{L^2(\Omega)} := (\int_\Omega |f(x)|^2 dx)^{1/2}$. We denote by $u^n$ and $\bar{v}^n$ the following piecewise constant functions of $L^2(\Omega)$:

$$\begin{aligned} u^n(x) &:= \sum_j u_j^n 1_{I_j}(x), \\ \bar{v}^n(x) &:= \sum_j v_j^n 1_{I_j}(x) = \sum_j v(t_n, x_j) 1_{I_j}(x), \end{aligned}$$

where $I_j = (x_j - h/2, x_j + h/2)$ and $1_{I_j}(x) = 1$ if $x \in I_j$ and $1_{I_j}(x) = 0$ otherwise, and we denote also the corresponding error $\bar{e}^n := u^n - \bar{v}^n$. Our aim is therefore to bound the following continuous $L^2$ error

$$\|\bar{e}^n\|_{L^2(\Omega)} = \left( \sum_i h |e_i^n|^2 \right)^{1/2}. \tag{45}$$

*Remark* 4.11. Notice that there is a uniform Lipschitz bound for $\|v_x(t_n,.)\|_{L^\infty}$ (for instance by using the representation formula (87) for the obstacle problem), therefore the error introduced by the projection on piecewise constant functions is roughly bounded by $\|\bar{v}^n - v^n\|_{L^2(\Omega)} \leq Ch$. This projection error will not be considered hereafter.

**Theorem 4.12. (error estimate).** *Assume that the exact solution $v$ of* (1) *satisfies (A1), (A2) and (A3). We consider the BDF2 scheme initialized with an*

*IE or a CN step for $u^1$. In the second case furthermore (A4) is assumed. Then the BDF2 scheme satisfies the following error bound for sufficiently small $\tau$ and $h$:*

$$\max_{1\le n\le N}\|\bar e^n\|^2_{L^2(\Omega)} \le C\big(h^2\tau^{1-2\alpha_4}+h\tau^{1-2\alpha_3}+\tau^2\tau^{1-2\alpha_2}+(\tau^{\alpha_0}+h)\tau^{1-2\alpha_1}+\tau^{\alpha_0}+h + \frac{\tau^2}{h}\big) \quad (46)$$

*for some constant $C\ge 0$ independent of $(\tau,h)$ if the powers $(\beta_i)_{1\le i\le 4} := \{(1-2\alpha_i)_{1\le i\le 4}$ are all non-zero (otherwise any $\tau^{\beta_i}$ with $\beta_i=0$ should be replaced by $\ln(\tau)$), and $\alpha_2<1$.*

*The term $\frac{\tau^2}{h}$ is not needed if $u^1$ is initialized with IE.*

In particular the scheme is convergent as soon as all factors in (46) converge to 0 as $(\tau,h)\to 0$.

*Remark* 4.13. In the case of the Model 1 presented in Section 5, for some $\alpha\in(0,1)$ we will have $\alpha_0=\alpha$, $\alpha_1=1-\alpha$, $\alpha_2=2-\alpha$, $\alpha_3=\alpha$, $\alpha_4=2\alpha$. Taking furthermore $\tau\equiv h$, the error estimate (46) is of order

$$\begin{aligned} & h^{3-2\alpha_4}+h^{2-2\alpha_3}+h^{3-2\alpha_2}+h^{1+\alpha_0-2\alpha_1}+h^{\alpha_0}+h \\ \le\ & C_1(h^{3-4\alpha}+h^{2-2\alpha}+h^{2\alpha-1}+h^{\alpha}+h) \\ \le\ & C_2(h^{\min(3-4\alpha,2-2\alpha,2\alpha-1)}) \end{aligned}$$

and does only give convergence for $\alpha$ in $(\frac12,\frac34)$, with a square error estimate of order $O(h^{\min(3-4\alpha,2\alpha-1)})$.

*Proof of Theorem 4.12.* Let us first consider the approximation in the $x$ variable. For $n=0,\dots,N-1$, let $\bar\epsilon^{n,1}$ be a consistency error in space, defined by

$$(Av^{n+1}+q^{n+1})_i = (\mathcal{A}v)(t_{n+1},x_i)+\bar\epsilon_i^{n,1}. \quad (47)$$

If $(t_{n+1},x_i)$ corresponds to a singular point of $v$, we consider for $\mathcal{A}v(t_{n+1},x_i)$ definitions of $v_t$ and $v_{xx}$ that satisfy the bounds $|v_t|\le\|v_t\|_{L^\infty}$ and $|v_{xx}|\le\|v_{xx}\|_{L^\infty}$, see Remark 4.4. In the region where $x\to v(t_{n+1},x)$ is regular, assuming that $v_{3x}(t_{n+1},.)$ is bounded on the interval $[x_{i-1},x_{i+1}]$, by using Taylor expansions up to the 3-rd order derivatives, it holds

$$|\bar\epsilon_i^{n,1}| = C_4 t_{n+1}^{-\alpha_4}\,O(h).$$

On the contrary in a region $[x_{i-1},x_{i+1}]$ that may encounter a singularity $y_j(t)$, we have no more than a bounded second order derivative ($v_{xx}\in L^\infty$). By using $|(Av^{n+1}+q^{n+1})_i|\le C(\|v_{xx}(t_{n+1},.)\|_{L^\infty}+\|v_x(t_{n+1},.)\|_{L^\infty}+\|v(t_{n+1},.)\|_{L^\infty})$, we have

$$|\bar\epsilon_i^{n,1}| = C_3 t_{n+1}^{-\alpha_3} O(1).$$

Moreover,

$$\mathrm{Card}\Big\{i,\ [x_{i-1},x_{i+1}]\cap\{y_j(t_{n+1})\}_{1\le j\le p}\neq\emptyset\Big\}\le 3p. \quad (48)$$

Using that the number of regular terms is bounded by $J\le C/h$, we obtain

$$\begin{aligned} \|\bar\epsilon^{n,1}\|^2 = \sum_i|\bar\epsilon_i^{n,1}|^2 &= \sum_{i,regular}|\bar\epsilon_i^{n,1}|^2+\sum_{i,singular}|\bar\epsilon_i^{n,1}|^2 && (49)\\ &\le C\frac1h(C_4t_{n+1}^{-\alpha_4}h)^2+C3p\,(C_3t_{n+1}^{-\alpha_3})^2 && (50)\\ &\le Cht_{n+1}^{-2\alpha_4}+Ct_{n+1}^{-2\alpha_3} && (51)\end{aligned}$$

for some constant $C$.

Notice that for any $n\tau \leq T$, and $\tau$ sufficiently small, we have

$$\tau \sum_{k=1}^{n} \frac{1}{t_k^\beta} \leq \begin{cases} C \max(\tau^{1-\beta}, 1) & \text{for } \beta > 0,\ \beta \neq 1, \\ C\,|\ln(\tau)| & \text{for } \beta = 1, \end{cases} \tag{52}$$

where $C$ may depend on $T, \beta$ but is independent of $\tau, n$.

Therefore we obtain, for $\alpha_3, \alpha_4 \neq \frac{1}{2}$:

$$h\big(\tau \sum_{k=1}^{n-1} \|\bar\epsilon^{k,1}\|^2\big) \leq C\Big(h^2 \max(\tau^{1-2\alpha_4}, 1) + h \max(\tau^{1-2\alpha_3}, 1)\Big) \tag{53}$$

$$\leq C\Big(h^2\tau^{1-2\alpha_4} + h\tau^{1-2\alpha_3} + h\Big) \tag{54}$$

(if $\alpha_3 = \frac{1}{2}$ or $\alpha_4 = \frac{1}{2}$ then the corresponding term $\tau^{1-2\alpha_i}$ should be replaced by $\ln(\tau)$).

Now we consider the approximation by BDF2 in time. Let $\bar\epsilon_i^{n,2}$ be such that, for $n \geq 1$:

$$\frac{3v_i^{n+1} - 4v_i^n + v_i^{n-1}}{2\tau} = v_t(t_{n+1}, x_i) + \bar\epsilon_i^{n,2}. \tag{55}$$

If $t \to v(t, x_i)$ is regular on $[t_{n-1}, t_{n+1}]$ with bounded $v_{tt}$ derivative, elementary Taylor expansions and (A2) give, for $n \geq 2$ (the cases $n = 0$ and $n = 1$ will be treated separately):

$$|\bar\epsilon_i^{n,2}| \leq C \max_{[t_{n-1}, t_{n+1}]} \|v_{tt}\|_{L^\infty(\Omega)}\, \tau \leq C t_{n-1}^{-\alpha_2} \tau,$$

while otherwise in a singular region we have

$$|\bar\epsilon_i^{n,2}| \leq C \max_{[t_{n-1}, t_{n+1}]} \|v_t\|_{L^\infty(\Omega)} \leq C t_{n-1}^{-\alpha_1}.$$

Let us introduce a set of singular indices as follows:

$$\mathcal{I}_s^n := \Big\{ i,\ x_i \in \bigcup_{j=1,\dots,p} y_j([t_{n-1}, t_{n+1}]) \Big\}, \qquad n = 1, \dots, N-1. \tag{56}$$

For $n \geq 1$ and for $t \in \Theta_n := [t_{n-1}, t_{n+1}]$, we get

$$|y_j(t) - y_j(t_{n+1})| \leq C(2\tau)^{\alpha_0}.$$

So if $x_i \in y_j(\Theta_n)$ then $|x_i - y_j(t_{n+1})| \leq C(2\tau)^{\alpha_0}$. Then, for any $A > 0$, the number of integers $i$ such that $x_i \in [-A + c, A + c]$ is bounded by $2A/h + 1$. Hence, for $n \geq 1$, we deduce a bound in the form

$$\mathrm{Card}(\mathcal{I}_s^n) \leq C\left(\frac{\tau^{\alpha_0}}{h} + 1\right) \tag{57}$$

for some constant $C \geq 0$. This bound holds also for

$$\mathcal{I}_s^0 := \Big\{ i,\ x_i \in \bigcup_{j=1,\dots,p} y_j([t_0, t_1]) \Big\},$$

as $\mathcal{I}_s^0 \subset \mathcal{I}_s^1$.

Now we can bound the $\bar\epsilon^{n,2}$ terms, for $n \geq 2$, as follows. We have

$$\|\bar\epsilon^{n,2}\|^2 = \sum_{i \notin \mathcal{I}_s^n} |\bar\epsilon_i^{n,2}|^2 + \sum_{i \in \mathcal{I}_s^n} |\bar\epsilon_i^{n,2}|^2 \tag{58}$$

$$\leq C \sum_{i \notin \mathcal{I}_s^n} (t_{n-1}^{-\alpha_2}\, \tau)^2 + \sum_{i \in \mathcal{I}_s^n} (t_{n-1}^{-\alpha_1})^2 \tag{59}$$

$$\leq C \frac{1}{h} \tau^2\ t_{n-1}^{-2\alpha_2} + C(\frac{\tau^{\alpha_0}}{h} + 1) t_{n-1}^{-2\alpha_1}. \tag{60}$$

Combining the previous bounds and using (52), for $\alpha_1, \alpha_2 \neq \frac{1}{2}$, we obtain

$$\begin{aligned} h\big(\tau \sum_{k=2}^{n} \|\bar\epsilon^{k,2}\|^2\big) &\leq C\big(\tau^2 \ \max(\tau^{1-2\alpha_2},1) + (\tau^{\alpha_0}+h)\max(\tau^{1-2\alpha_1},1) \\ &\leq C\big(\tau^2\tau^{1-2\alpha_2} + (\tau^{\alpha_0}+h)\tau^{1-2\alpha_1} + \tau^{\alpha_0} + h\big) \end{aligned} \tag{61}$$

(powers of $\tau$ with exponent $1-2\alpha_1 = 0$ or $1-2\alpha_2 = 0$ need to be replaced by $\ln(\tau)$).

Using the stability estimate (34) and the fact that $e^0 = 0$ we obtain

$$\|\bar e^n\|^2_{L^2(\Omega)} = h\|e^n\|^2 \tag{62}$$

$$\leq Ch\bigg(\|e^1\|^2 + \tau\sum_{k=1}^{n-1}\|\bar\epsilon^{k,1}\|^2 + \tau\sum_{k=1}^{n-1}\|\bar\epsilon^{k,2}\|^2\bigg). \tag{63}$$

By using the estimates (54) and (61), we obtain the desired error bound (46) as long as we can bound $\|\bar\epsilon^{1,2}\|^2$ (the first time-consistency error term that appears in the estimates for BDF2) as well as $\|e^1\|^2$ (the IE rsp. CN scheme error) accordingly.

By using similar techniques as for BDF2, and $e^0 = 0$, it is easy to see that

$$h\|e^1\|^2 \leq h\left(\frac{1}{1-s\tau\gamma}(\|e^0\| + \tau\|\bar\epsilon^0\|)\right)^2 = \frac{h\tau^2}{(1-\tau s\gamma)^2}\|\bar\epsilon^0\|^2, \tag{64}$$

where $\bar\epsilon^0$ is the consistency error for the IE (rsp. CN) scheme and $s = 1$ (rsp. $s = \frac{1}{2}$).

For both the IE and the CN scheme, we can write $\bar\epsilon^0 = \bar\epsilon^{0,1} + \bar\epsilon^{0,2}$ where $\bar\epsilon^{0,1}$ represents the spatial consistency error and $\bar\epsilon^{0,2}$ the time-consistency error given by

$$\frac{v_i^1 - v_i^0}{\tau} = v_t(t_1, x_i) + \bar\epsilon_i^{0,2}. \tag{65}$$

The term $|v_t(t,x)|$ is not assumed to be bounded for $t = 0^+$, but we have $|v_t(t_1,x)| \leq C\tau^{-\alpha_1}$ since $t_1 = \tau$ and using (A2). By using again (A2) we obtain

$$|\frac{1}{\tau}(v(t_1,x_i) - v(t_0,x_i)) - v_t(t_1,x_i)| = \frac{1}{\tau}|\int_{t_0}^{t_1} v_t(s,x_i)\mathrm{d}s| \leq \frac{1}{\tau}\int_{t_0}^{t_1} Cs^{-\alpha_1}\mathrm{d}s \leq C\tau^{-\alpha_1}. \tag{66}$$

This estimate holds for all $i$. Therefore

$$|\bar\epsilon_i^{0,2}| \leq C\tau^{-\alpha_1}.$$

If $t \to v(t,x_i)$ is regular on $[t_0,t_1]$ with bounded $v_{tt}$ derivative, the estimate can be improved to

$$\begin{aligned} |\bar\epsilon_i^{0,2}| \leq |\bar\epsilon_i^{0,2}||\frac{1}{\tau}(v(t_1,x_i) - v(t_0,x_i)) - v_t(t_1,x_i) - v_t(t_1,x_i)| \\ = \frac{1}{\tau}|\int_{t_0}^{t_1}\int_{s}^{t_1} v_{tt}(u,x_i)\mathrm{d}u\mathrm{d}s| \leq \frac{1}{\tau}\int_{t_0}^{t_1} C_2 s^{-\alpha_2}(t_1 - s)\mathrm{d}s \leq C\tau^{1-\alpha_2} \end{aligned}$$

as $\alpha_2 < 1$. In the end, we obtain a contribution to the error as follows (to be multiplied by $\tau$):

$$h\tau\|\bar\epsilon^{0,2}\|^2 = h\tau\sum_{i\notin\mathcal{I}_s^n}|\bar\epsilon_i^{0,2}|^2 + h\tau\sum_{i\in\mathcal{I}_s^n}|\bar\epsilon_i^{0,2}|^2 \leq C\tau^{3-2\alpha_2} + C(\tau^{\alpha_0}+h)\tau^{1-2\alpha_1}.$$

The same bound for $h\tau\|\bar\epsilon^{1,2}\|^2$ can be obtained by using similar estimates.

For the IE scheme, the consistency error in space $\bar\epsilon^{0,1}$ can be bounded by (51) with $n = 0$, and in conclusion we obtain the desired bound for the IE scheme as starting scheme.

Now, it remains to bound the spatial consistency error $\bar\epsilon^{0,1}$ in the case that $u^1$ is computed by a CN scheme. We split $\bar\epsilon^{0,1}$ into two parts, $\bar\epsilon^{0,1} = \frac{1}{2}\bar\epsilon^{0,1,1} + \frac{1}{2}\bar\epsilon^{0,1,2}$, with

$$(Av_0 + q^0)_i = (\mathcal{A}v)(t_1, x_i) + \bar\epsilon_i^{0,1,1}, \qquad (Av^1 + q^1)_i = (\mathcal{A}v)(t_1, x_i) + \bar\epsilon_i^{0,1,2}. \quad (67)$$

$\bar\epsilon^{0,1,2}$ can be estimated by (51) with $n = 0$. To bound $\bar\epsilon^{0,1,1}$, we take advantage of assumption (A4) being true. Due to $v_0$ being Lipschitz regular, $Av_0 + q^0$ is bounded by $O(\frac{1}{h})$. Since $\mathcal{A}v_0$ is also assumed to be bounded, it results a bound of the form

$$|\bar\epsilon_i^{0,1,1}| \le \frac{C}{h}.$$

If $v_0$ is $C^2$ regular on $[x_{i-1}, x_{i+1}]$, then, by standard estimates,

$$|(Av_0 + q^0)_i| \le C\,(\|v_0''\|_{L^\infty} + \|v_0'\|_{L^\infty} + \|v_0\|_{L^\infty}).$$

This, together with $v_{xx}$ being bounded, shows that $|\bar\epsilon_i^{0,1,1}|$ is bounded for such indices. Summing up the estimates in the singular region (which by (48) involves not more than $3p$ cases), and in the regular region (which involves $O(\frac{1}{h})$ cases), we obtain the bound

$$\begin{aligned} \|\bar\epsilon^{0,1,1}\|^2 &\le \sum_{i,\ \text{singular}} |\bar\epsilon_i^{0,1,1}|^2 + \sum_{i,\ \text{regular}} |\bar\epsilon_i^{0,1,1}|^2 \\ &\le C(\frac{1}{h})^2 + O(\frac{1}{h})C = O(\frac{1}{h^2}). \end{aligned}$$

Altogether the contribution of $h\|e^1\|^2$ to the overall error can be bounded by

$$h\|e^1\|^2 \le Ch\tau^2(\frac{1}{h^2} + h\tau^{-2\alpha_4} + \tau^{-2\alpha_3} + \frac{1}{h}\tau^{2-2\alpha_2} + \frac{1}{h}(\tau^{\alpha_0} + h)\tau^{-2\alpha_1}).$$

□

## 5. Numerical results on two model test problems

In this section we introduce two model test problems for diffusion with obstacle, with source terms and analytic solutions, to better analyze the performance of the proposed BDF schemes. A first problem mimics the American option problem with a jump in the $v_{xx}$ derivative at a given singular position $x_s(t)$ that can be user-defined (in the numerical simulations, we will assume a $\sqrt{t}$ behavior for small times, see (69)). The second problem allows for a bounded $v_{xxx}$ derivative with a jump at $x = x_s(t)$. These two models allow us to better check numerically the performance of the BDF2 and BDF3 schemes, respectively. Without an analytic solution, it is otherwise difficult to precisely compute a reference solution with very fine mesh.

### 5.1. Two model test problems.

We first define two model test problems. In the case of (1) we do in general not know about exact solutions. Therefore we construct simple model obstacle problems with explicit solutions (or solutions that can be easily computed with machine precision) and also with the main features of the one-dimensional American option problem.

This is obtained by choosing an explicit function $v = v(t, x)$ and adding a corresponding source term $f = f(t, x)$ to the original PDE (3), thus considering

$$\min\left(v_t - \frac{\lambda^2}{2}x^2 v_{xx} - rxv_x + rv,\ v - \varphi(x)\right) = f(t, x). \quad (68)$$

More precisely, let $K$, $X_{max}$, $c_0$, $T$ and $\alpha$ be given constants such that $0 < K < X_{max}$, $c_0 > 0$, $T > 0$, $\alpha \in (0, 1]$ and such that $K - c_0\,T^\alpha > 0$. Let $\varphi(x) := \max(K - x, 0)$ denote the payoff function and let $x_s$ be defined by

$$x_s(t) := K(1 - c_0\,t^\alpha). \quad (69)$$

In the numerical experiments we will use $\alpha = \frac{1}{2}$, to be close to the American option case, even though the error estimate in Theorem 4.12 only yields convergence for $\alpha \in (\frac{1}{2}, \frac{3}{4})$ for the below Model 1, see Remark 4.13.

We construct explicit functions $v(t,x)$ defined for $x \in [0, X_{max}]$ and such that

$(i)$ $v(t,x) = \varphi(x) = K - x$ for $x \leq x_s(t)$,
$(ii)$ $v(t,x) > \varphi(x) = \max(K-x, 0)$ for $x \in ]x_s(t), X_{max}]$,
$(iii)$ for all $t \in (0,T]$, $v(t,\cdot)$ is at least $C^1$ on $[0, X_{max}]$,
$(iv)$ $v(t, X_{max}) = 0$.

Note that requirement $(iii)$ implies $v_x(t, x_s(t)) = \varphi'(x_s(t)) = -1$ for $t > 0$.

**Model 1.** Let $v = v(t,x)$ be the function defined by:

$$v(t,x) := \begin{cases} \varphi(x) & \text{for } x < x_s(t) \\ \varphi(x_s(t)) - \dfrac{x - x_s(t)}{1 + (x - x_s(t))/C(t)} & \text{otherwise} \end{cases} \tag{70}$$

where $C(t) > 0$ is a constant such that $v(t, X_{max}) = 0$:

$$C(t) := \left( \frac{1}{\varphi(x_s(t))} - \frac{1}{X_{max} - x_s(t)} \right)^{-1}.$$

Then the requirements $(i) - (iv)$ are satisfied.

**Model 2.** Let $v = v(t,x)$ be the function defined by:

$$v(t,x) := \begin{cases} \varphi(x) & \text{for } x < x_s(t) \\ \varphi(x_s(t)) - C(t)\operatorname{atan}\left( \dfrac{x - x_s(t)}{C(t)} \right) & \text{otherwise} \end{cases} \tag{71}$$

for a given $C(t) > 0$. Notice that $v(t,x)$ is a non-increasing function of the variable $x$. This function will satisfy requirements $(i) - (iv)$ if furthermore $C(t)$ is such that

$$\frac{\varphi(x_s(t))}{C(t)} = \operatorname{atan}\left( \frac{X_{max} - x_s(t)}{C(t)} \right). \tag{72}$$

Letting $a := X_{max} - x_s(t)$ and $b = \varphi(x_s(t)) = K - x_s(t)$ it is clear that $0 < b < a$ and therefore there exists a unique $\theta > 0$ such that $b\theta = \operatorname{atan}(a\theta)$. This value can be numerically obtained by using a fixed-point method. We then define $C(t) := 1/\theta$ to obtain a solution of (72). Therefore the function $v$ is in explicit form but for the computation of the $C(t)$ function which can be computed to arbitrary precision.

*Remark* 5.1. For Model 2, in order to compute $v_t(t,x)$ the derivative $\dot{C}(t)$ is needed. Denoting $a = a(t) = X_{max} - x_s(t)$ and $b = b(t) = \varphi(x_s(t))$, and $\theta = \theta(t) = 1/C(t)$, by derivation of $b\theta = \operatorname{atan}(a\theta)$ we obtain $\dot{C}/C = -\dot{\theta}/\theta = (q\dot{b} - \dot{a})/(qb - a)$ where $q = 1 + (a/C)^2$, with $\dot{a} = -\dot{x}_s(t)$ and $\dot{b} = \varphi'(x_s(t))\dot{x}_s(t)$.

*Remark* 5.2. The main difference between the two models is the regularity of the data near the singularity $x = x_s(t)$. More precisely, for the first model there is a jump in the second derivative: $v$ is of class $C^1$ and $v_{xx}$ is discontinuous. For the second model, $v$ is of class $C^2$ and there is a jump in the third derivative $v_{3x}$.

5.2. **A 4th order approximation of the spatial operator $\mathcal{A}$.** We furthermore introduce a 4th order numerical matrix approximation of the $\mathcal{A}$ operator in order to better observe the time discretization error.

Let $D^2 u_j := \frac{-u_{j-1} + 2u_j - u_{j+1}}{h^2}$. By using Taylor expansions, if $u_j = u(x_j)$, we have

$$-u_{xx}(x_j) = \frac{-u_{j-1} + 2u_j - u_{j+1}}{h^2} + \frac{1}{12h^2}(u_{j-2} - 4u_{j-1} + 6u_j - 4u_{j+1} + u_{j+2}) + O(h^4)$$

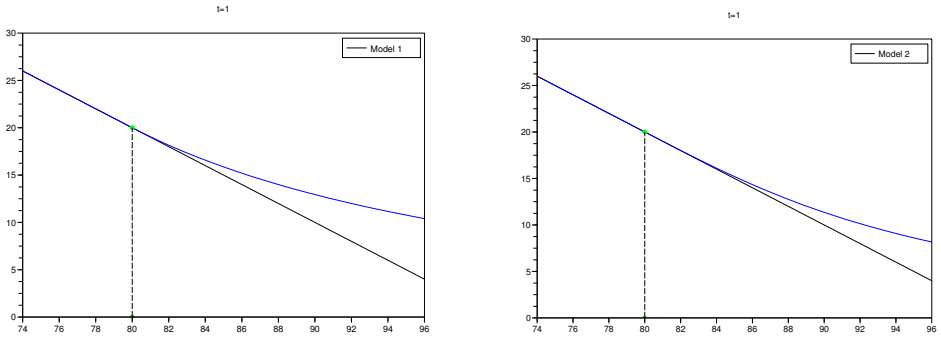


FIGURE 1. Zooming around the singular point $(x_s, \varphi(x_s))$ for model 1 (left) and 2 (right).

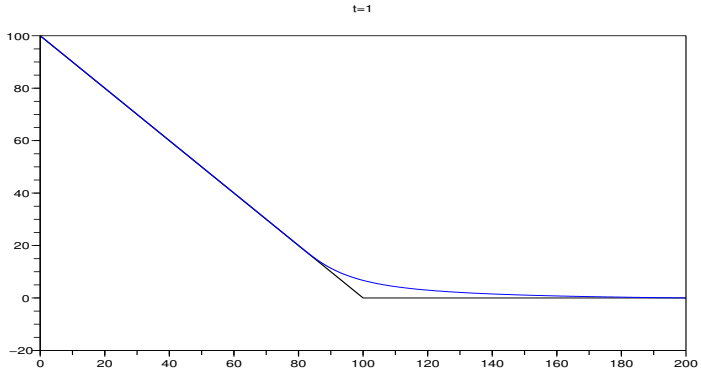


FIGURE 2. Model 2.

and

$$u_x(x_j) = \frac{u_{j+1} - u_{j-1}}{2h} + \frac{1}{12h}(u_{j-2} - 2u_{j-1} + 2u_{j+1} - u_{j+2}) + O(h^4).$$

Therefore we have a 4-$th$ order approximation of the spatial derivatives, and will denote by $\tilde{A}$ (instead of $A$) the corresponding approximation matrix. At the boundaries, for the American option problem we use $u_{-1} = K - X_{min} - h$, $u_0 = K - X_{min}$, and $u_{J+1} = u_{J+2} = 0$ (the left boundary condition is consistent with fourth order because we expect that $v(t,x) = \varphi(x) = K - x$ near the left boundary, also the right boundary condition is consistent with the fact that the exact solution $v(t,x)$ (obtained for the case $X_{\max} = \infty$) decays faster than any polynomial as $x \to \infty$). For our model problems, we will use the known exact solution values at the boundaries.

The BDF2 and BDF3 schemes are otherwise unchanged concerning the time discretization. Hence for $\tau \equiv c\ h$ we expect to mainly see the error of the time discretization. In particular we aim to observe, whenever possible, second or third order behavior in time.

The following lemma shows that the coercivity property of Lemma 4.6 extends to the $4th$ order approximation:

**Lemma 5.3.** *Under assumption (A1), there exist $\eta_2 > 0$ and $\gamma_2 \geq 0$ such that for all $x \in \mathbb{R}^J$:*

$$\langle x, \tilde{A}x\rangle \geq \eta_2 N(x/h)^2 - \gamma_2 \|x\|_2^2. \tag{73}$$

*Proof.* Let us focus on the matrix part that concerns the approximation of the diffusion $-a(x)u_{xx}(x)$, where we have removed the time dependency to simplify the notation. The other terms coming from the drift term $b(x)u_x$ and the term $ru$ can be bounded as before. The new matrix reads as follows:

$$\tilde{A} := A + B$$

where $A$ stands for the second order approximation of the diffusion term, i. e.,

$$A := \frac{1}{h^2}\,\mathrm{pdiag}(-a_j, 2a_j, -a_j) \equiv \frac{1}{h^2}\Delta A_0$$

and

$$B := \frac{1}{12h^2}\,\mathrm{pdiag}(a_j,\ -4a_j,\ 6aj, -4a_j, a_j) \equiv \frac{1}{12h^2}\Delta B_0,$$

where $a_j = a(x_j)$, and pdiag stands for p-band diagonal matrices (a tridiagonal matrix for $A$, rsp. a pentadiagonal matrix for $B$), and where we have also denoted $A_0 := \mathrm{pdiag}(-1, 2, -1)$, $B_0 := \mathrm{pdiag}(1, -4, 6, -4, 1)$, and $\Delta := \mathrm{diag}(a_j)$ (a diagonal matrix with $\Delta_{jj} = a_j$).

First notice that

$$B_0 = A_0^2 + \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 0 & \dots & 0 \\ \vdots & \vdots & \dots & \vdots \\ 0 & \dots & 0 & 1 \end{pmatrix}$$

and therefore $\langle B_0 x, x\rangle \geq \langle A_0^2 x, x\rangle$, and, since $\Delta \geq 0$ and diagonal,

$$\langle Bx, x\rangle \geq \langle \frac{1}{12h^2}\Delta A_0^2 x, x\rangle.$$

Because $x \to a(x)$ is assumed Lipschitz continuous, we have

$$\max_{i,j} |(\Delta A_0 - A_0\Delta)_{ij})|_\infty = O(h).$$

As the matrix $\Delta A_0 - A_0\Delta$ is tridiagonal, it follows

$$\|\Delta A_0 x - A_0 \Delta x\| \leq Ch\|x\|.$$

Therefore we have also

$$\begin{aligned}
\langle \Delta A_0^2 x, x\rangle &= \langle A_0\Delta A_0 x, x\rangle + \langle (\Delta A_0 - A_0\Delta)A_0 x, x\rangle && (74)\\
&= \langle \Delta A_0 x, A_0 x\rangle + \langle (\Delta A_0 - A_0\Delta)A_0 x, x\rangle && (75)\\
&\geq \eta_0 \|A_0 x\|^2 - Ch\|A_0 x\|\|x\| && (76)\\
&\geq \eta_0 \|A_0 x\|^2 - \frac{\eta_0}{2}\|A_0 x\|^2 - \frac{C^2h^2}{2\eta_0}\|x\|^2 && (77)\\
&\geq \frac{\eta_0}{2}\|A_0 x\|^2 - \frac{C^2h^2}{2\eta_0}\|x\|^2 \geq -\frac{C^2h^2}{2\eta_0}\|x\|^2 && (78)
\end{aligned}$$

and

$$\frac{1}{h^2}\langle \Delta A_0^2 x, x\rangle \quad \geq \quad -\frac{C^2}{2\eta_0}\|x\|^2. \tag{79}$$

Thus we have a lower bound for $\langle Bx, x\rangle$ of the desired type, i. e. $\langle Bx, x\rangle \geq -C'\|x\|^2$, and the lower bound for $\langle \tilde{A}x, x\rangle$ will be of the same type as for $\langle Ax, x\rangle$. □

Consequently, Theorem 4.12 extends to the 4th order approximation.

5.3. **Numerical results for models 1 and 2.** From now on the parameters used are $\lambda = 0.3$, $r = 0.1$, $K = 100$. The singularity motion is defined by $x_s(t) = K(1 - c_0\sqrt{t})$ with $c_0 = 0.2$.

The errors in $L^2$, $L^1$ and $L^\infty$ norms are computed at time $t_N = T$ using (17). For Model 1, the numerical domain is defined by $\Omega = (75, 275)$ (i. e., $X_{\min} = 75$ and $X_{\max} = 275$), and $T = 1$. Therefore the singularity at $T = 1$ is located at $x_s(T) = 80$. Numerical results for Model 1 with the CN and BDF2 schemes and second order spatial discretization (results for fourth order approximation in space are similar) are given in Table 4 and Table 5.

For Model 1, with bounded $v_{xx}$ derivative, we observe that the order of the CN scheme, when $N = J+1$, is two in the $L^1$ and $L^2$ norm (and around 1.6 in the $L^\infty$ norm) and goes down to order one in the $L^\infty$ norm when $N = (J+1)/10$ (i. e. the mesh ratio $\tau/h$ is large). On the contrary, the BDF2 scheme keeps roughly an error of order two for different mesh ratios and for all norms.

*Remark* 5.4. For models 1 and 2, we have observed that for the first step of BDF2, using the BDF1 scheme (i. e., the IE scheme) instead of CN yields nearly unchanged results (this is different to the American option problem, see Remark 3.3).

*Remark* 5.5. We have also tested the BDF3 scheme on Model 1, and the numerical results are (both for second and fourth order approximation in space) very similar to the BDF2 scheme, and in particular the numerical order is not greater than two. This comes from the fact that the solution has a bounded second order derivative, with a jump.

| Mesh | | Error $L^1$ | | Error $L^2$ | | Error $L^\infty$ | | time(s) |
|---|---|---|---|---|---|---|---|---|
| $J+1$ | $N$ | error | order | error | order | error | order | |
| 80 | 80 | 1.09e+00 | 1.79 | 1.67e-01 | 1.75 | 3.89e-02 | 1.64 | 0.04 |
| 160 | 160 | 2.85e-01 | 1.94 | 4.53e-02 | 1.88 | 1.14e-02 | 1.77 | 0.08 |
| 320 | 320 | 8.41e-02 | 1.76 | 1.36e-02 | 1.73 | 3.63e-03 | 1.65 | 0.22 |
| 640 | 640 | 2.12e-02 | 1.99 | 3.51e-03 | 1.95 | 9.94e-04 | 1.87 | 0.53 |
| 1280 | 1280 | 5.68e-03 | 1.90 | 9.46e-04 | 1.89 | 2.76e-04 | 1.85 | 1.40 |
| 2560 | 2560 | 1.41e-03 | 2.00 | 2.40e-04 | 1.98 | 7.36e-05 | 1.91 | 4.52 |
| 5120 | 5120 | 3.65e-04 | 1.96 | 6.23e-05 | 1.95 | 1.95e-05 | 1.92 | 17.04 |
| 10240 | 10240 | 9.08e-05 | 2.01 | 1.57e-05 | 1.99 | 7.00e-06 | 1.48 | 64.24 |
| 80 | 8 | 5.07e+00 | 1.61 | 1.02e+00 | 1.39 | 4.35e-01 | 0.69 | 0.00 |
| 160 | 16 | 8.93e-01 | 2.51 | 1.83e-01 | 2.48 | 9.27e-02 | 2.23 | 0.01 |
| 320 | 32 | 3.00e-01 | 1.57 | 6.83e-02 | 1.42 | 4.34e-02 | 1.09 | 0.03 |
| 640 | 64 | 6.12e-02 | 2.29 | 1.79e-02 | 1.93 | 2.10e-02 | 1.05 | 0.06 |
| 1280 | 128 | 1.79e-02 | 1.77 | 5.86e-03 | 1.61 | 1.01e-02 | 1.06 | 0.18 |
| 2560 | 256 | 4.17e-03 | 2.10 | 1.86e-03 | 1.66 | 4.94e-03 | 1.03 | 0.58 |
| 5120 | 512 | 1.11e-03 | 1.91 | 6.21e-04 | 1.58 | 2.38e-03 | 1.06 | 2.25 |
| 10240 | 1024 | 2.73e-04 | 2.02 | 2.15e-04 | 1.53 | 1.19e-03 | 1.00 | 8.23 |

TABLE 4. (Model 1) CN scheme for (1) (using 2nd order spatial approximation), with different mesh ratios $N = J+1$, $N = (J+1)/10$.

Then we focus on numerical results for Model 2. We have tested again the CN, BDF2 and BDF3 schemes. In that case we consider the problem with $\Omega = (50, 450)$ and $T = 0.5$, the other parameters being as in Model 1.

| Mesh | | Error $L^1$ | | Error $L^2$ | | Error $L^\infty$ | | time(s) |
|---|---|---|---|---|---|---|---|---|
| $J$+1 | $N$ | error | order | error | order | error | order | |
| 80 | 80 | 1.27e+00 | 1.87 | 2.03e-01 | 1.77 | 7.15e-02 | 1.39 | 0.02 |
| 160 | 160 | 3.59e-01 | 1.82 | 5.89e-02 | 1.78 | 2.50e-02 | 1.51 | 0.05 |
| 320 | 320 | 9.18e-02 | 1.97 | 1.55e-02 | 1.93 | 8.17e-03 | 1.61 | 0.12 |
| 640 | 640 | 2.39e-02 | 1.94 | 4.06e-03 | 1.93 | 2.52e-03 | 1.70 | 0.32 |
| 1280 | 1280 | 5.97e-03 | 2.00 | 1.03e-03 | 1.98 | 7.35e-04 | 1.78 | 0.96 |
| 2560 | 2560 | 1.51e-03 | 1.98 | 2.60e-04 | 1.98 | 2.06e-04 | 1.84 | 3.25 |
| 5120 | 5120 | 3.76e-04 | 2.01 | 6.49e-05 | 2.00 | 5.58e-05 | 1.88 | 11.56 |
| 10240 | 10240 | 9.43e-05 | 1.99 | 1.63e-05 | 2.00 | 1.48e-05 | 1.91 | 43.10 |
| 80 | 8 | 1.83e+00 | 2.57 | 4.24e-01 | 2.10 | 1.93e-01 | 1.76 | 0.00 |
| 160 | 16 | 3.81e-01 | 2.26 | 9.19e-02 | 2.21 | 5.32e-02 | 1.86 | 0.01 |
| 320 | 32 | 7.93e-02 | 2.27 | 2.03e-02 | 2.18 | 1.44e-02 | 1.89 | 0.01 |
| 640 | 64 | 1.96e-02 | 2.02 | 4.71e-03 | 2.11 | 3.81e-03 | 1.92 | 0.04 |
| 1280 | 128 | 4.47e-03 | 2.13 | 1.05e-03 | 2.17 | 9.90e-04 | 1.94 | 0.12 |
| 2560 | 256 | 1.13e-03 | 1.98 | 2.48e-04 | 2.08 | 2.55e-04 | 1.96 | 0.40 |
| 5120 | 512 | 2.72e-04 | 2.06 | 5.74e-05 | 2.11 | 6.50e-05 | 1.97 | 1.51 |
| 10240 | 1024 | 6.99e-05 | 1.96 | 1.41e-05 | 2.02 | 1.65e-05 | 1.98 | 5.15 |

TABLE 5. (Model 1) BDF2 scheme (using 2nd order spatial approximation), for different mesh ratios.

Results for CN and BDF3 schemes are given in Tables 6 and 7 respectively. For this model, by construction, we recall that the exact solution has bounded third order spacial derivatives. The CN scheme gives good results when $N = J+1$ (second order convergence), but goes back to first order convergence when $N = (J+1)/10$ (in the $L^\infty$ norm). The results for the BDF2 scheme, which are not shown, demonstrate second order convergence but unconditionally on the mesh parameters. On the other hand the BDF3 scheme shows at least third order convergence for the $L^\infty$ norm, as well for both ratios of the mesh parameters.

In conclusion, for the type of obstacle problems studied here, we advise using the BDF2 scheme instead of the CN scheme because it keeps its expected numerical order unconditionally on the mesh parameters.

## APPENDIX A. AN HJB EQUATION FOR OBSTACLE PROBLEMS

This appendix is devoted to a sketch of proof for the equivalence between PDE (1) and PDE (5) in case the coefficients are not time dependent.

In order to simplify the presentation we assume that $f \equiv 0$ and $\Omega \equiv \mathbb{R}$. We consider the problem (1) after a change of variable $t \to T - t$:

$$\min(-v_t + \mathcal{A}v,\ v - \varphi(x)) = 0, \quad t \in (0,T), \quad x \in \Omega, \tag{80a}$$

$$v(T,x) = \varphi(x), \quad x \in \Omega \tag{80b}$$

We aim to prove that $v$ is also a viscosity solution of (5). In the following, we will first prove that $(i)$ $-v_t + \mathcal{A}v \geq 0$, then $(ii)$ that $-v_t \geq 0$, then $(iii)$ that $\min(-v_t + \mathcal{A}v, -v_t) = 0$ and will $(iv)$ conclude by a uniqueness argument.

$(i)$ By uniqueness of the continuous solutions of (1), $v$ is also given by the expectation formula

$$v(t,x) = \sup_{\tau \in \mathcal{T}_{[t,T]}} \mathbb{E}(e^{-\int_t^\tau r\mathrm{d}s} \varphi(X_\tau^{t,x}) | \mathcal{F}_t).$$

| Mesh | | Error $L^1$ | | Error $L^2$ | | Error $L^\infty$ | | time(s) |
|---|---|---|---|---|---|---|---|---|
| $J$+1 | $N$ | error | order | error | order | error | order | |
| 80 | 80 | 8.04e-01 | 3.14 | 2.13e-01 | 2.56 | 8.32e-02 | 1.67 | 0.16 |
| 160 | 160 | 1.03e-01 | 2.96 | 2.14e-02 | 3.31 | 7.25e-03 | 3.52 | 0.26 |
| 320 | 320 | 8.46e-03 | 3.61 | 1.64e-03 | 3.71 | 4.63e-04 | 3.97 | 0.74 |
| 640 | 640 | 3.11e-04 | 4.77 | 5.80e-05 | 4.82 | 1.31e-05 | 5.14 | 1.33 |
| 1280 | 1280 | 6.26e-06 | 5.64 | 1.29e-06 | 5.50 | 5.35e-07 | 4.61 | 3.23 |
| 2560 | 2560 | 5.81e-07 | 3.43 | 1.22e-07 | 3.40 | 4.06e-08 | 3.72 | 7.99 |
| 5120 | 5120 | 1.43e-07 | 2.02 | 2.96e-08 | 2.04 | 8.70e-09 | 2.22 | 22.41 |
| 10240 | 10240 | 3.57e-08 | 2.01 | 7.40e-09 | 2.00 | 2.17e-09 | 2.00 | 64.65 |
| 80 | 8 | 8.07e-01 | 3.33 | 2.20e-01 | 2.70 | 8.83e-02 | 1.76 | 0.01 |
| 160 | 16 | 1.44e-01 | 2.49 | 3.37e-02 | 2.71 | 1.40e-02 | 2.66 | 0.02 |
| 320 | 32 | 1.48e-02 | 3.28 | 3.34e-03 | 3.34 | 1.36e-03 | 3.37 | 0.06 |
| 640 | 64 | 1.04e-03 | 3.83 | 3.62e-04 | 3.21 | 2.36e-04 | 2.52 | 0.14 |
| 1280 | 128 | 2.50e-04 | 2.06 | 8.43e-05 | 2.10 | 7.91e-05 | 1.58 | 0.28 |
| 2560 | 256 | 7.81e-05 | 1.68 | 3.30e-05 | 1.36 | 4.02e-05 | 0.98 | 0.71 |
| 5120 | 512 | 2.02e-05 | 1.95 | 1.09e-05 | 1.60 | 1.97e-05 | 1.03 | 2.05 |
| 10240 | 1024 | 5.27e-06 | 1.94 | 3.80e-06 | 1.52 | 1.01e-05 | 0.97 | 6.52 |

TABLE 6. (Model 2) CN scheme for (1) (using 4th order spatial approximation).

| Mesh | | Error $L^1$ | | Error $L^2$ | | Error $L^\infty$ | | time(s) |
|---|---|---|---|---|---|---|---|---|
| $J$+1 | $N$ | error | order | error | order | error | order | |
| 80 | 80 | 8.07e-01 | 3.13 | 2.13e-01 | 2.55 | 8.33e-02 | 1.67 | 0.09 |
| 160 | 160 | 1.01e-01 | 2.99 | 2.10e-02 | 3.34 | 7.12e-03 | 3.55 | 0.18 |
| 320 | 320 | 8.30e-03 | 3.61 | 1.60e-03 | 3.71 | 4.52e-04 | 3.98 | 0.40 |
| 640 | 640 | 3.04e-04 | 4.77 | 5.67e-05 | 4.82 | 1.29e-05 | 5.13 | 0.93 |
| 1280 | 1280 | 7.27e-06 | 5.38 | 1.40e-06 | 5.34 | 4.86e-07 | 4.73 | 2.17 |
| 2560 | 2560 | 1.34e-07 | 5.77 | 4.43e-08 | 4.98 | 2.46e-08 | 4.30 | 5.53 |
| 5120 | 5120 | 1.13e-08 | 3.57 | 2.80e-09 | 3.99 | 1.41e-09 | 4.12 | 18.26 |
| 10240 | 10240 | 1.07e-09 | 3.40 | 2.13e-10 | 3.72 | 7.88e-11 | 4.16 | 49.88 |
| 80 | 8 | 1.02E+00 | 2.74 | 2.41e-01 | 2.38 | 9.12e-02 | 1.84 | 0.01 |
| 160 | 16 | 6.54e-02 | 3.96 | 1.69e-02 | 3.84 | 6.99e-03 | 3.71 | 0.02 |
| 320 | 32 | 9.64e-03 | 2.76 | 1.86e-03 | 3.18 | 5.28e-04 | 3.73 | 0.04 |
| 640 | 64 | 3.53e-04 | 4.77 | 6.58e-05 | 4.82 | 1.54e-05 | 5.10 | 0.10 |
| 1280 | 128 | 1.46e-05 | 4.60 | 2.69e-06 | 4.61 | 6.31e-07 | 4.61 | 0.22 |
| 2560 | 256 | 8.46e-07 | 4.11 | 1.57e-07 | 4.10 | 3.88e-08 | 4.02 | 0.54 |
| 5120 | 512 | 8.62e-08 | 3.29 | 1.64e-08 | 3.26 | 4.25e-09 | 3.19 | 1.53 |
| 10240 | 1024 | 1.01e-08 | 3.10 | 1.96e-09 | 3.06 | 5.21e-10 | 3.03 | 5.00 |

TABLE 7. (Model 2) BDF3 scheme for (1) (using 4th order spatial approximation).

(see for instance [21]) where we have considered a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, a filtration $(\mathcal{F}_t)_{t\geq 0}$, $\mathcal{T}_{[t,T]}$ is the set of stopping times taking values a.s. in $[t,T]$, $X_\tau := X_\tau^{t,x}$ is the strong solution of the stochastic differential equation (SDE):

$$\mathrm{d}X_s = b(X_s)\mathrm{d}s + \sigma(X_s)\mathrm{d}W_s, \quad s \geq t,$$

with $X_t = x$, $W_s$ denotes an $\mathcal{F}_t$-adapted Brownian motion on $\mathbb{R}$, and the "sup" is an essential supremum over $\mathcal{T}_{[t,T]}$. First one can use the semi-Martingale property

$$v(t,x) \leq \mathbb{E}(e^{-rh}\, v(t+h, X_{t+h}^{t,x})|\mathcal{F}_t),$$

in order to deduce (in the viscosity sense), that

$$-v_t + \mathcal{A}v \geq 0.$$

$(ii)$ Then we aim to show that $v(t,x) \geq v(t+h,x)$, for any $h > 0$. This will imply $-v_t \geq 0$ (in the viscosity sense). By definition,

$$\begin{aligned} v(t+h,x) &= \sup_{\tau \in \mathcal{T}_{[t+h,T]}} \mathbb{E}(e^{-\int_{t+h}^{\tau} r\,\mathrm{d}s}\varphi(X_\tau^{t+h,x})|\mathcal{F}_{t+h}) && (81)\\ &= \sup_{\tau \in \mathcal{T}_{[t,T-h]}} \mathbb{E}(e^{-\int_{t+h}^{\tau+h} r\,\mathrm{d}s}\varphi(X_{\tau+h}^{t+h,x})|\mathcal{F}_{t+h}) && (82)\\ &= \sup_{\tau \in \mathcal{T}_{[t,T-h]}} \mathbb{E}(e^{-\int_{t}^{\tau} r\,\mathrm{d}s}\varphi(X_\tau^{t,x})|\mathcal{F}_t). && (83) \end{aligned}$$

We have used the fact that the process $X^{t,x}$ satisfies an SDE with no time dependency in the coefficients, and also, since $\tau \in \mathcal{T}_{[t,T-h]}$, $X_{\tau+h}$ a.s. stops before time $T$, the fact that $\mathbb{E}(X_{\tau+h}^{t+h,x}|\mathcal{F}_{t+h}) = \mathbb{E}(X_\tau^{t,x}|\mathcal{F}_t)$ - which corresponds to an averaging during a period of time $T-(t+h)$. Then, in particular,

$$v(t+h,x) \leq \sup_{\tau \in \mathcal{T}_{[t,T]}} \mathbb{E}(e^{-\int_{t}^{\tau} r\,\mathrm{d}s}\varphi(X_\tau^{t,x})|\mathcal{F}_t) = v(t,x). \quad (84)$$

At this point we therefore have shown that

$$\min(-v_t + \mathcal{A}v, -v_t) \geq 0.$$

$(iii)$ Let us assume that $-v_t(t,x) > 0$ (in the viscosity sense), and $t < T$. It implies that $v(t,x) > v(t+h,x)$ for all $h > 0$ small enough. Because $v(t,x) > v(t+h,x) \geq \varphi(x)$, we have $v(t,x) > \varphi(x)$. The following dynamic programming principle holds:

$$v(t,x) = \mathbb{E}(e^{-\int_t^{\tau_{t,x}^*} r\,\mathrm{d}s}\varphi(X_{\tau_{t,x}^*}^{t,x})|\mathcal{F}_t) = \mathbb{E}(e^{-\int_t^{\tau_{t,x}^*} r\,\mathrm{d}s} v(\tau_{t,x}^*, X_{\tau_{t,x}^*}^{t,x})|\mathcal{F}_t)$$

where $\tau_{t,x}^*$ is the optimal stopping time for the obstacle problem, defined by

$$\tau_{t,x}^* = \inf\left\{\theta \geq t,\ v(\theta, X_\theta^{t,x}) = \varphi(X_\theta^{t,x})\right\}.$$

It can be shown that $\tau_{t,x}^* > t$ a.s. (since $v(t,x) > \varphi(x)$, these functions being continuous). Let us show that $-v_t + \mathcal{A}v = 0$ at $(t,x)$ in the viscosity sense. By using Ito's formula between $t$ and $\tau_{t,x}^*$, and from the dynamic programming principle, we deduce that

$$0 = \mathbb{E}\left(\int_t^{\tau_{t,x}^*} e^{-\int_t^\theta r\,\mathrm{d}s}(v_t - \mathcal{A}v)_{(\theta, X_\theta^{t,x})}\mathrm{d}\theta \mid \mathcal{F}_t\right).$$

We already have proved that $v_t - \mathcal{A}v \leq 0$ a.s., so we deduce that $(v_t - \mathcal{A}v)(\theta,x) = 0$ a.e. for $\theta \in (t, \tau_{t,x}^*)$. For some random parameter $w$ we have $t^* := \tau_{t,x}^*(w) > t$, from which it is deduced that $(v_t - \mathcal{A}v)(t,x) = 0$. Therefore we have proved in this case that $\min(-v_t(t,x) + \mathcal{A}v(t,x),\ -v_t) = 0$.

$(iv)$ Conversely, we can use a uniqueness argument for the solutions of (5) in order to conclude the equivalence between (1) and (5).

*Remark* A.1. *In the same way, it can be proved that the following PDE with source term and $x$-dependent coefficients in the operator $\mathcal{A}$:*

$$\min(-u_t + \mathcal{A}u, u - \varphi(x)) = f(x), \quad t \in (0,T), \quad x \in \Omega, \quad (85a)$$

$$u(T,x) = \varphi(x) + f(x), \quad x \in \Omega, \quad (85b)$$

*is equivalent to the following Hamilton-Jacobi-Bellman equation*

$$-u_t + \min(\mathcal{A}u,\ 0) = f(x), \quad t \in (0,T), \quad x \in \Omega, \tag{86a}$$

$$u(T,x) = \varphi(x) + f(x), \quad x \in \Omega. \tag{86b}$$

*Problem* (85) *is associated with the following stopping time problem*

$$u(t,x) = \sup_{\tau \in \mathcal{T}_{[t,T]}} \mathbb{E}\left( e^{-\int_t^\tau r \mathrm{d}s}(\varphi(X_\tau^{t,x}) + f(X_\tau^{t,x})) + \int_t^\tau e^{-\int_t^\theta r \mathrm{d}s} f(X_\theta^{t,x}) \mathrm{d}\theta | \mathcal{F}_t \right). \tag{87}$$

The Hamilton-Jacobi equation (86) (or (5)) admits also a representation formula corresponding to a stochastic optimal control problem with controlled diffusion, drift and rate term $(\theta\sigma(t,x),\ \theta b(t,x),\ \theta r(t))$ with $\theta \in [0,1]$, see for instance [23].

## References


[1] Yves Achdou and Olivier Pironneau. *Computational methods for option pricing*, volume 30 of *Frontiers in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2005.

[2] Guy Barles, Julien Burdeau, Marc Romano, and Nicolas Samsœn. Estimation de la frontière libre des options américaines au voisinage de l'échéance. *C. R. Acad. Sci. Paris Sér. I Math.*, 316(2):171–174, 1993.

[3] Julien Berton and Robert Eymard. Finite volume methods for the valuation of American options. *M2AN Math. Model. Numer. Anal.*, 40(2):311–330, 2006.

[4] Adrien Blanchet, Jean Dolbeault, and Régis Monneau. On the continuity of the time derivative of the solution to the parabolic obstacle problem with variable coefficients. *J. Math. Pures Appl. (9)*, 85(3):371–414, 2006.

[5] Olivier Bokanowski, Stefania Maroso, and Hasnaa Zidani. Some convergence results for Howard's algorithm. *SIAM J. Numer. Anal.*, 47(4):3001–3026, 2009.

[6] Olivier Bokanowski, Athena Picarelli, and Christoph Reisinger. Stability and convergence of second order backward differentiation schemes for parabolic Hamilton-Jacobi-Bellman equations. Preprint, 2018.

[7] Philippe G. Ciarlet. *Introduction à l'analyse numérique matricielle et à l'optimisation*. Collection Mathématiques Appliquées pour la Maîtrise. [Collection of Applied Mathematics for the Master's Degree]. Masson, Paris, 1982.

[8] Michael Grain Crandall, Hitoshi Ishii, and Pierre-Louis Lions. User's guide to viscosity solutions of second order partial differential equations. *Bull. Amer. Math. Soc. (N.S.)*, 27(1):1–67, 1992.

[9] Michel Crouzeix and Alain L. Mignot. *Analyse numérique des équations différentielles*. Collection Mathématiques Appliquées pour la Maîtrise. [Collection of Applied Mathematics for the Master's Degree]. Masson, Paris, 1984.

[10] Charles Francis Curtiss and Joseph Oakland Hirschfelder. Integration of stiff equations. *Proc. Nat. Acad. Sci. U. S. A.*, 38:235–243, 1952.

[11] Jeffrey N. Dewynne, Sam D. Howison, I. Rupf, and Paul Wilmott. Some mathematical results in the pricing of American options. *European J. Appl. Math.*, 4(4):381–398, 1993.

[12] Etienne Emmrich. Stability and error of the variable two-step BDF for semilinear parabolic problems. *J. Appl. Math. & Computing*, 19(1-2):33–55, 2005.

[13] Peter A. Forsyth and Kenneth R. Vetzal. Quadratic convergence for valuing American options using a penalty method. *SIAM J. Sci. Comput.*, 23(6):2095–2122, 2002.

[14] Avner Friedman. *Variational principles and free-boundary problems*. Robert E. Krieger Publishing Co., Inc., Malabar, FL, second edition, 1988.

[15] C. William Gear. *Numerical initial value problems in ordinary differential equations*. Prentice-Hall, Inc., Englewood Cliffs, N.J., 1971.

[16] Ernst Hairer and Gerhard Wanner. *Solving ordinary differential equations. II*. Springer-Verlag, Berlin, second edition, 1996.

[17] Ernst Hairer and Gerhard Wanner. Linear multistep method. *Scholarpedia*, 5(4):4591, 2010.

[18] Michael Hintermüller, Kazufumi Ito, and Karl Kunisch. The primal-dual active set strategy as a semismooth Newton method. *SIAM J. Optim.*, 13(3):865–888 (2003), 2002.

[19] Patrick Jaillet, Damien Lamberton, and Bernard Lapeyre. Variational inequalities and the pricing of American options. *Acta Appl. Math.*, 21(3):263–289, 1990.

[20] Espen R. Jakobsen. On the rate of convergence of approximation schemes for Bellman equations associated with optimal stopping time problems. *Mathematical Models and Methods in Applied Sciences*, 13(05):613–644, 2003.

[21] Damien Lamberton and Bernard Lapeyre. *Introduction to stochastic calculus applied to finance*. Chapman & Hall/CRC Financial Mathematics Series. Chapman & Hall/CRC, Boca Raton, FL, second edition, 2008.
[22] Fabien Le Floc'h. TR-BDF2 for fast stable American option pricing. *Journal of Computational Finance*, 17(3):31–561, 2014.
[23] Pierre-Louis Lions. Optimal control of diffusion processes and Hamilton-Jacobi-Bellman equations. I. The dynamic programming principle and applications. *Comm. Partial Differential Equations*, 8(10):1101–1174, 1983.
[24] Pierre-Louis Lions. Optimal control of diffusion processes and Hamilton-Jacobi-Bellman equations. II. Viscosity solutions and uniqueness. *Comm. Partial Differential Equations*, 8(11):1229–1276, 1983.
[25] Claude Martini. American option prices as unique viscosity solutions to degenerated Hamilton-Jacobi-Bellman equations. Research Report RR-3934, INRIA, 2000.
[26] Cornelis W. Oosterlee. On multigrid for linear complementarity problems with application to American-style options. *Electron. Trans. Numer. Anal.*, 15:165–185 (electronic), 2003. Tenth Copper Mountain Conference on Multigrid Methods (Copper Mountain, CO, 2001).
[27] Cornelis W. Oosterlee, Francisco José Gaspar, and J. C. Frisch. WENO and blended BDF discretizations for option pricing problems. In *Numerical mathematics and advanced applications*, pages 419–428. Springer Italia, Milan, 2003.
[28] Huyên Pham. Optimal stopping of controlled jump diffusion processes: a viscosity solution approach. *J. Math. Systems Estim. Control*, 8(1):27 pp. 1998.
[29] Christoph Reisinger and Alan Whitley. The impact of a natural time change on the convergence of the Crank-Nicolson scheme. *IMA J. Numer. Anal.*, 34(3):1156–1192, 2014.
[30] Christoph Reisinger and Yufei Zhang. A penalty scheme for monotone systems with interconnected obstacles: convergence and error estimates. Preprint, 2018.
[31] S. I. Serdjukova. Uniform stability of a six-point scheme of higher order accuracy for the heat equation. *Ž. Vyčisl. Mat. i Mat. Fiz.*, 7(1):214–218, 1967.
[32] Rüdiger U. Seydel. *Tools for computational finance*. Universitext. Springer London, fifth edition, 2012.
[33] Heath Windcliff, Peter A. Forsyth, and Kenneth R. Vetzal. Shout options: a framework for pricing contracts which can be modified by the investor. *J. Comput. Appl. Math.*, 134(1-2):213–241, 2001.
[34] Jan Hendrik Witte and Christoph Reisinger. Penalty methods for the solution of discrete HJB equations—continuous control and obstacle problems. *SIAM J. Numer. Anal.*, 50(2):595–625, 2012.

LABORATOIRE JACQUES-LOUIS LIONS, UNIVERSITÉ DE PARIS (PARIS DIDEROT) PARIS, FRANCE AND ENSTA PARISTECH (*OLIVIER.BOKANOWSKI@MATH.UNIV-PARIS-DIDEROT.FR*)

UNIVERSITY OF SOUTHERN DENMARK, DEPARTMENT OF MATHEMATICS AND COMPUTER SCIENCE, ODENSE (*DEBRABANT@IMADA.SDU.DK*)