SAPIENZA
UNIVERSITÀ DI ROMA

# Semi-Lagrangian schemes for parabolic equations: second order accuracy and boundary conditions

Scuola di Dottorato in Matematica

Dottorato di Ricerca in Matematica – XXXIV Ciclo


Candidate

Elisa Calzola

ID number 1560756


Thesis Advisors

Prof. Elisabetta Carlini

Prof. Francisco J. Silva

2022

Thesis defended on 30th March 2022
in front of a Board of Examiners composed by:

Diogo Gomes, Espen Jakobsen, Maurizio Falcone (chairman)

---

**Semi-Lagrangian schemes for parabolic equations: second order accuracy and boundary conditions**

Ph.D. thesis. Sapienza – University of Rome



This thesis has been typeset by LaTeX and the Sapthesis class.

Author's email: calzola@mat.uniroma1.it

*Dedicated to*
*Roberto, Graziella,*
*Giada and Niccolò*

# Abstract

This thesis deals with second order parabolic differential equations and some semi-Lagrangian methods to approximate their solutions. We start with a brief survey of the main theoretical results concerning linear and nonlinear parabolic equations, recalling some existence and uniqueness to the Cauchy problem on $\mathbb{R}^d$ and to the Initial-Boundary value problem with Dirichlet and Neumann type boundary conditions. In the following three chapters, we present our approach to the numerical solution to three different problems. First, we introduce a semi-Lagrangian method for advection-diffusion-reaction systems of equations on bounded domains, with Dirichlet boundary conditions. Afterwards, we present a semi-Lagrangian technique for approximating the solution to Hamilton-Jacobi-Bellman equations on bounded domain, with Neumann-type boundary conditions. Finally, we present a Lagrange-Galerkin approximation of the Fokker-Planck equation, and we show how to apply such a method to obtain a second-order accurate solution to Mean Field Games. Every method is accompanied with numerical simulations.

# Contents

# Introduction

In this thesis we focus on numerical methods for second order parabolic Partial Differential Equations (PDEs), which are used for describing a wide variety of time-dependent phenomena.

The work is divided into four parts and treats three different problems: in the first chapter, we recall the main theoretical results for parabolic equations, starting from the linear and semilinear ones, both in $\mathbb{R}^d$ and in a bounded domain. We then briefly present the theory of viscosity solutions to nonlinear parabolic equations and we conclude the chapter with a short rewiew on the Fokker-Planck (FP) equation and its application to the theory of Mean Field Games. In the second chapter we present a second order semi-Lagrangian (SL) method for systems of advection-diffusion-reaction equations on a bounded spatial domain, introducing a novel approach in the treatment of Dirichlet boundary conditions. In the third chapter we focus on the approximation of a second order Hamilton-Jacobi-Bellman (HJB) equations on a bounded spatial domain, with generalized Neumann boundary conditions. Finally, in the fourth and last chapter, we propose a second order Lagrange-Galerkin (LG) scheme to approximate the solution to the Fokker-Planck equation, and we show how to couple it with a second order semi-Lagrangian method for HJB equations in order to obtain a second order scheme for Mean Field Games (MFGs).

### Part I - Parabolic PDEs: linear and non linear type

In Chapter 1 we recall some classical results concerning parabolic equations. Given a function $u$, we consider the linear operator $\mathcal{L}_t$

$$(\mathcal{L}_t u)(t,x) = \sum_{i,j=1}^{d} a_{ij}(t,x)\frac{\partial^2 u}{\partial x_i \partial x_j} + \sum_{i=1}^{d} \mu_i(t,x)\frac{\partial u}{\partial x_i} + c(t,x)u$$

and the differential equation

$$\partial_t u = \mathcal{L}_t u.$$

We say that the operator $\mathcal{L}_t$ is *parabolic* if the matrix $(a_{ij}(t,x))$ is symmetric and positive definite. Section 1.1 is devoted to a brief presentation of the results concerning linear parabolic equations, collecting various results regarding the Cauchy problem on unbounded domains and the Cauchy-Dirichlet problem on spatially bounded domains. An important part is the one devoted to the results concerning the link between stochastic differential equations and second-order

parabolic equations. Both kind of equations represent, in Physics, diffusion-type fenomena, and there exist some results, called *Feynman-Kac formulae*, that link explicitly the solution to a parabolic equation to the solution to an associated stochastic differential equation. These formulas are the starting point to derive semi-Lagrangian methods. In Section 1.2 we present some results for Hamilton-Jacobi-Bellman equations, a class of nonlinear parabolic equations of the form

$$\begin{cases} \partial_t u + H(t, x, u, Du, D^2u) = 0 & \text{in } (0, T] \times \mathbb{R}^d, \\ u(0, x) = u_0(x) & \text{for } x \in \mathbb{R}^d. \end{cases}$$

We recall the definition of viscosity solutions and results about existence and uniqueness for the Cauchy problem, with Dirichlet and Neumann boundary conditions. We give a brief presentation of deterministic and stochastic optimal control theories, showing that the value function of an optimal control problem is the solution to an associated HJB equation. Section 1.3 deals with the FP equation, a particular type of linear parabolic equation which has a great importance in many application, for example in Biology, Physics and, also, in Mean Field Games problems. Also the Fokker-Planck equation has a representation formula, which will be crucial in Chapter 4 to derive a scheme to approximate its solution.

### Part II - Second order fully semi-Lagrangian discretizations of advection-diffusion-reaction systems

In Chapter 2 we deal with systems of advection–diffusion–reaction (ADR) equations, which model the chemical or biochemical processes involving several species transported by a fluid. These systems are responsible for most of the computational cost of typical environmental fluid dynamics models, such as those applied in climate, water and air quality and oceanic biogeochemistry modeling for long term simulations [20, 47, 48]. Also in applications to medium range weather forecasting, which consider shorter time ranges, the number of interacting transported species can be quite large. This implies that a very large number of ADR equations have to be solved simultaneously, in order to achieve a complete description of the relevant physical processes. As a consequence, even minor efficiency gains in the solution to this very classical problem are of paramount practical importance. This explains why numerical methods that allow the use of large time steps are favoured for these applications, see e.g. the discussion in [108]. The standard ways to enhance efficiency for the solution of the advection step are either the use of implicit schemes or the application of SL techniques, [49, 102]. These are then coupled to implicit methods for the diffusion and reaction step. As discussed in [47, 48], SL methods have the advantage that all the computational work that makes them computationally more expensive per time step than standard Eulerian techniques is indeed independent of the number of tracers, which allows to achieve easily a superior efficiency level in the limit of a large number of tracers.
In the recent papers [18, 19], a fully SL approach to both the advection and diffusion steps was pursued, which combines the standard SL treatment of

advection with SL-like schemes for diffusion. In particular, it was shown in [19] that, even for a single advection–diffusion equation, the fully SL approach can be more efficient than standard implicit techniques. SL methods for parabolic, second order problems have also been studied, among others, in [26, 52, 84, 85, 86, 45, 14, 92]. A complete review of the earlier literature on this topic can be found in [49, 72]. We remark that, among the proposals in the literature, the formulation first introduced in [18] is an original contribution, since it allows to treat straightforwardly parabolic problems in divergence form, such as those usually encountered in computational fluid dynamics applications.
Since the technique under consideration stems from the Feynman–Kac stochastic representation formula, it could also be possible to mix SL schemes with a Monte-Carlo approach, as proposed, for instance, in [24]. However, while this latter strategy might be more scalable on massively parallel architectures, on more conventional platforms it suffers from a slow convergence with respect to the number of sample trajectories. By exploiting the concept of *weak convergence* of schemes for stochastic differential equations, the deterministic approach pursued here usually results in a lower computational complexity.
The outline of the chapter is the following. In Section 2.2, at least in small space dimensions, we introduce the PDE system

$$\begin{cases} \partial_t u + \langle \mu, Du \rangle - \dfrac{\sigma^2}{2}\Delta u = f(u) & (t,x) \in (0,T] \times \mathcal{O}, \\ u(t,x) = b(t,x) & (t,x) \in (0,T] \times \partial\mathcal{O}, \\ u(0,x) = u_0(x) & x \in \mathcal{O}, \end{cases}$$

where $\mathcal{O} \subset \mathbb{R}^2$ is a bounded domain. Section 2.3 describes the SL advection–diffusion solver: we approximate the stochastic characteristics using a Crank-Nicolson approach, then we reconstruct the numerical solution at the foot of such approximated characteristics using a $\mathbb{P}_2$ interpolation operator. A stability and convergence analysis of the method is outlined in Section 2.4. The possible approaches to the treatment of boundary conditions are discussed in Section 2.5. A numerical validation of the proposed approach on both structured and unstructured meshes is presented in Section 2.6, while some conclusions and perspectives for future developments are outlined in Section 2.7.

### Part III - A semi-Lagrangian scheme for Hamilton-Jacobi-Bellman equations with oblique boundary conditions

In Chapter 3 we deal with the numerical approximation of the parabolic Hamilton-Jacobi-Bellman (HJB) equation on $[0,T] \times \mathcal{O}$

$$\begin{cases} \partial_t u + H\left(t, x, Du, D^2 u\right) = 0 & \text{in } (0,T] \times \mathcal{O}, \\ L(t,x,Du) = 0 & \text{on } (0,T] \times \partial\mathcal{O}, \\ u(0,x) = \Psi(x) & \text{in } \overline{\mathcal{O}}, \end{cases} \tag{0.1}$$

where $\mathcal{O} \subset \mathbb{R}^d$ is a bounded domain and $H$ and $L$ are nonlinear functions having a specific form. The study of the numerical approximation of solutions to HJB

and, more generally, fully nonlinear second order PDEs, has made important progress over the last few decades. Most of the related literature consider the case where $\mathcal{O} = \mathbb{R}^d$, or where a Dirichlet boundary condition is imposed on the boundary $\partial\mathcal{O}$. We refer the reader to [49, 50, 90] and the references therein for the state of the art on this topic. By contrast, the numerical approximation of solutions to (0.1) has been much less explored. Indeed, to the best of our knowledge only the methods in [100, 1] can be applied to approximate (0.1) in the particular first order case ($\sigma \equiv 0$). Moreover, in [100], where a finite difference scheme is proposed, the function defining the boundary condition has the particular form $L(t, x, p, b) = \langle n(x), p\rangle$. On the other hand, both references consider Hamiltonians which are not necessarily convex with respect to $p$. Let us also mention the reference [3], where, in the context of mean curvature motion with nonlinear Neumann boundary conditions, the authors propose a discretization that combines a SL scheme in the main part of the domain with a finite difference scheme near the boundary.

The main purpose of this chapter is to provide a consistent, stable, monotone and convergent SL scheme to approximate the unique viscosity solution to (0.1). By the results in [6], the latter is well-posed in $C([0, T] \times \overline{\mathcal{O}})$ under the assumptions in Section 3.1. Semi-Lagrangian schemes to approximate the solution to (0.1) when $\mathcal{O} = \mathbb{R}^d$ (see e.g. [26, 45]) can be derived from the optimal control interpretation of (0.1) and a suitable discretization of the underlying controlled trajectories. These schemes enjoy the feature that they are explicit and stable under an inverse Courant-Friedrichs-Lewy (CFL) condition and, consequentely, they allow large time steps. A second important feature is that they permit a simple treatement of the possibly degenerate second order term in $H$. The scheme that we propose for $\mathcal{O} \neq \mathbb{R}^d$ preserves these two properties and seems to be the first convergent scheme to approximate (0.1) with the rather general asumptions in Section 3.1. In particular, our results cover the stochastic and degenerate case. Consequently, from the stochastic control point of view, our scheme allows to approximate the so-called value function of the optimal control of a controlled diffusion process with possibly oblique reflection on the boundary $\partial\mathcal{O}$ (see [22]). The main difficulty in devising such a scheme is to be able to obtain a consistency type property at points in the space grid which are near the boundary $\partial\mathcal{O}$ while maintaining the stability. This is achieved by considering a discretization of the underlying controlled diffusion which suitably emulates its reflection at the boundary in the continuous case. We refer the reader to [83] for a related construction of a semi-discrete in time approximation of a second order non-degenerate linear parabolic equation.

The remainder of this chapter is structured as follows. In Section 3.1 we state our assumptions, we recall the notion of viscosity solution to (0.1) and the well-posedness result. In Section 3.2 we provide the SL scheme as well as its probabilistic interpretation (in the spirit of [83]). The latter will play an important role in Section 3.3, which is devoted to show a consistency type property and the stability of the SL scheme. By using the half-relaxed limits technique introduced in [9], we show in Section 3.4 our main result, which is the convergence of solutions to the SL scheme to the unique viscosity solution to (0.1). The convergence is uniform in $[0, T] \times \overline{\mathcal{O}}$ and holds under the same

asymptotic condition between the space and time steps than in the case $\mathcal{O} = \mathbb{R}^d$. Next, in Section 3.5 we first illustrate the numerical convergence of the SL scheme in the case of a one-dimensional linear equation with homogeneous Neumann boundary conditions. In this case the numerical results confirm that the boundary condition in (0.1) is not satisfied at every $x \in \partial\mathcal{O}$, but it is satisfied in the viscosity sense recalled in Section 3.1 below. In a second example, we consider a two dimensional degenerate second order nonlinear equation on a circular domain with non-homogeneous Neumann and oblique boundary conditions. In the last example, we consider a two-dimensional non-degenerate nonlinear equation on a non-smooth domain. Due to the lack of regularity of $\partial\mathcal{O}$, our convergence result does not apply. However, the SL scheme can be successfully applied, which suggests that our theoretical findings could hold for more general domains. This extension as well as the corresponding study in the stationary framework remain as interesting subjects of future research. Finally, we provide in section 3.6 some theoretical results concerning oblique projections and the regularity of the distance to $\partial\mathcal{O}$, which play a key role in the definition of the scheme and in the proof of its main properties.

### Part IV - A second order Lagrange-Galerkin scheme for Fokker-Planck equations and applications to MFGs

The Fokker-Planck equations have broad areas of interest, starting with physics, biology and chemistry. We refer the reader to [97] for the theory of linear FP equations and their probabilistic interpretation. The main application we have in mind is to approximate evolutive Mean Field Games problems, recently introduced in [67, 75, 76], in order to model dynamic games with a large number of indistinguishable small players. We consider a MFG problem consisting of a backward HJB equation coupled with a forward FP equation. The two equations are linked through the cost function of the HJB equation, depending on the solution of the FP equation, and the drift of the FP, being the gradient of the value function solving the HJB. The solution of the MFG problem is the fixed point of the system.

The main purpose of Chapter 4 is to provide a Lagrange-Galerkin approximation scheme for the FP equations with constant diffusion, having the form

$$\begin{cases} \partial_t m - \dfrac{\sigma^2}{2}\Delta m + \operatorname{div}(\mu m) = 0 & \text{in } (0,T]\times\mathbb{R}^d, \\ m(0,\cdot) = m_0 & \text{in } \{0\}\times\mathbb{R}^d, \end{cases} \tag{0.2}$$

which is conservative, second-order accurate, explicit and stable with quite large time steps. Furthermore we propose a scheme which, coupled with an accurate second order semi-Lagrangian scheme for the HJ equation, approximates the solution to MFG problems with second order of accuracy.

The numerical solution of Fokker-Planck equations has been widely studied. There are several methods based on the popular finite difference scheme proposed by Chang and Cooper [38], which, in order to be stable and explicit, requires a parabolic CFL condition. Lagrange-Galerkin (LG) and SL methods have been mostly developed for advection and advection-diffusion problems, see [88, 12, 49]

and the references therein. The relation between SL and LG schemes have been analyzed in [53].

The main idea here is to couple these two techniques in order to develop a scheme for a particular class of FP equations. We begin in Section 4.1 with a brief presentation of an existing first order SL method for such equations, introduced in [34]. In Section 4.2 in order to derive our scheme, we first discretize in time the representation formula, true in a time interval $[t, s] \subset [0, T]$,

$$\int_{\mathbb{R}^d} \phi(x) m(s, x) \mathrm{d}x = \int_{\mathbb{R}^d} \mathbb{E}\left(\phi(X^{t,x}(s)) m(t, x)\right) \mathrm{d}x,$$

where $\phi$ is a continuous function with compact support and $X^{t,x}(s)$ is the characteristic starting at $x$ at time $t$. We apply a Crank-Nicolson approximation to the Stochastic Differential Equation (SDE), whose probability density is the solution of the FP. This first step is developed as in semi-Lagrangian schemes for second order parabolic equations, see [17]. Then, in Section 4.2.1 we introduce the symmetric Lagrangian basis of odd order to obtain a fully discrete and exactly integrated scheme. The choice of odd degree basis functions is inspired by the results in [53, 54], where the equivalence between semi-Lagrangian schemes, based on odd symmetric Lagrange interpolation, with Lagrange-Galerkin schemes has been analyzed. It has been shown that the symmetric odd basis have a better behavior, in terms of stability, when applied to transport problems. In all the simulations carried on during the work, the order of reconstruction has always been chosen to be 3. We prove consistency, $L^2$ stability and we provide a convergence result to the unique classical solution of (0.2).

Section 4.3 presents the MFGs problem. We introduce a second order semi-Lagrangian method for the HJB equation and we couple it with our Lagrange-Galerkin scheme for the FP to obtain a second order method for MFGs. In Section 4.4, we show a possible implementation when the spatial dimension $d = 1$, in which we consider Simpson's quadrature to approximate the integral terms. Numerical simulations endorse this choice in terms of stability and efficiency, compared to a more costly quadrature formula such as Gauss Legendre. The resulting scheme is the adjoint of a second order accurate semi-Lagrangian scheme applied to the backward equation, adjoint to the FP. We conclude our work with three numerical simulations, one for the Fokker-Planck in spatial dimension $d = 2$ and two Mean Field Games problems. For all these tests we present an error analysis, both in $L^\infty$ and in $L^2$ norms, that confirms order two of convergence of our scheme.

The results presented in Chapter 2 have been published in 2021 on *Journal of Scientific Computing*, while the results in Chapter 3 have been submitted to *Numerische Matematik* in 2021.
The results presented in Chapter 4 are still a work in progress.

# Chapter 1

# Parabolic PDEs: linear and non linear type

This first chapter has the aim to collect some results on the well-posedness for both linear and nonlinear equations of parabolic type. Let us first provide some notation and recall some useful definitions. The norm of $x = (x_1, \dots, x_d) \in \mathbb{R}^d$, i.e. the distance of $x$ from the origin, is defined as the standard Euclidean norm

$$|x| = \left( \sum_{i=1}^{d} x_i^2 \right)^{1/2}.$$

**Definition 1.** *Given an open set $S \subseteq \mathbb{R}^d$, a function $f : S \to \mathbb{R}$ is* Hölder continuous of exponent $\alpha$ $(0 < \alpha \leq 1)$ *in $S$ if there exists a constant $A > 0$ such that*

$$|f(x) - f(y)| \leq A\,|x - y|^{\alpha}, \quad \text{for all } x, y \in S. \tag{1.1}$$

*The smallest $\alpha$ for which* (1.1) *holds is called the* Hölder exponent *of $f$.*

A function $f$ is said to be *locally* Hölder continuous in $S$ if (1.1) holds in every bounded closed set $B \subset S$ with constant $A$, which may depend on $B$.

If $\alpha = 1$ in (1.1) the function $f(x)$ is said to be *Lipschitz continuous*.

## 1.1 Linear parabolic PDEs

Let $\mathcal{O} \subseteq \mathbb{R}^d$ be a bounded open domain and $T > 0$. Consider the operator $\mathcal{L}_t$, $t \in [0, T]$, defined on smooth functions $u : [0, T] \times \mathcal{O} \to \mathbb{R}$ as

$$(\mathcal{L}_t u)(t, x) = \sum_{i,j=1}^{d} a_{ij}(t, x) \frac{\partial^2 u(t, x)}{\partial x_i \partial x_j} + \sum_{i=1}^{d} \mu_i(t, x) \frac{\partial u(t, x)}{\partial x_i} + c(t, x) u(t, x) \tag{1.2}$$

and the differential equation

$$\partial_t u - \mathcal{L}_t u = 0. \tag{1.3}$$

We assume that the matrix $(a_{ij}(t, x))$ is symmetric, i.e. for every $(t, x)$ we have $a_{ij}(t, x) = a_{ji}(t, x)$. If the matrix $(a_{ij}(t, x))$ is positive definite, meaning that

for every non-null real vector $\xi = (\xi_1, \dots, \xi_d) \in \mathbb{R}^d$, $\sum a_{ij}(t,x)\xi_i\xi_j > 0$, then we say that the operator $\mathcal{L}_t$ is *parabolic*. If there exist two positive constants $\lambda_1$ and $\lambda_2$ such that

$$(\forall (t,x) \in [0,T] \times \mathcal{O},\ \forall \xi \in \mathbb{R}^d) \quad \lambda_1 |\xi|^2 \leq \sum_{i,j=1}^{d} a_{ij}(t,x)\xi_i\xi_j \leq \lambda_2 |\xi|^2 ,$$

then $\mathcal{L}_t$ is said to be *uniformly parabolic*. From now on we will assume that

(A1) $\mathcal{L}_t$ is parabolic in $[0,T] \times \mathcal{O}$;

(A2) the coefficients of $\mathcal{L}_t$ are continuous functions in $[0,T] \times \mathcal{O}$ and there exists $\alpha \in (0,1)$ such that, for all $(t,x),(s,y) \in [0,T] \times \mathcal{O}$, there exists $A > 0$ such that

$$|a_{ij}(t,x) - a_{ij}(s,y)| \leq A \left( |x-y|^\alpha + |t-s|^{\alpha/2} \right), \tag{1.4}$$

$$|\mu_i(t,x) - \mu_i(s,y)| \leq A |x-y|^\alpha , \tag{1.5}$$

$$|c(t,x) - c(s,y)| \leq A |x-y|^\alpha . \tag{1.6}$$

We can now give a definition of solution and fundamental solution of (1.3).

**Definition 2.** *A smooth function $u : [0,T] \times \mathcal{O} \to \mathbb{R}$ is a* solution *to* (1.3) *in a domain $\mathcal{O}$ if all the derivatives of $u$ occourring in* (1.3) *are continuous function in $\mathcal{O}$ and the equation* (1.3) *is satisfied at each $(t,x) \in [0,T] \times \mathcal{O}$.*

**Definition 3.** *A* fundamental solution *to* (1.3) *in $[0,T] \times \mathcal{O}$ is a function $\Gamma(t,x;\tau,\xi)$ defined for all $(t,x),(\tau,\xi) \in [0,T] \times \mathcal{O}$, $t > \tau$, such that:*

(i) *for fixed $(\tau,\xi) \in [0,T] \times \mathcal{O}$ it satisfies* (1.3) *as a function of $(t,x)$, with $x \in \mathcal{O}$ and $\tau < t \leq T$;*

(ii) *for every $f \in C(\overline{\mathcal{O}})$ and $x \in \mathcal{O}$, we have*

$$\lim_{t \searrow \tau} \int_D \Gamma(t,x;\tau,\xi) f(\xi) \mathrm{d}\xi = f(x). \tag{1.7}$$

It is possible to construct a fundamental solution to (1.3) on a bounded domain using a procedure called the *parametric method*. First, let $(a^{ij}(t,x))$ be the inverse matrix of $(a_{ij}(t,x))$; for every $(\sigma,y) \in [0,T] \times \mathcal{O}$, $\xi, x \in \mathcal{O}$, $t > \tau$ we define

$$\vartheta^{\sigma,y}(x,\xi) = \sum_{i,j=1}^{d} (a^{ij}(\sigma,y))(x_i - \xi_i)(x_j - \xi_j), \tag{1.8}$$

$$\omega^{\sigma,y}(t,x;\tau,\xi) = (t-\tau)^{-d/2} \exp\left\{ \frac{\vartheta^{\sigma,y}(x,\xi)}{4(t-\tau)} \right\}, \tag{1.9}$$

$$Z(t,x;\tau,\xi) = (2\sqrt{\pi})^{-d} \left[ \det(a^{ij}(\tau,\xi)) \right]^{1/2} \omega^{\tau,\xi}(t,x;\tau,\xi). \tag{1.10}$$

For fixed $(\tau,\xi)$ the function $Z(t,x;\tau,\xi)$ in (1.10) satisfies

$$\partial_t u(t,x) - \sum_{i,j=1}^{d} a_{ij}(\tau,\xi) \frac{\partial^2 u}{\partial x_i \partial x_j}(t,x) = 0, \quad \text{for } (t,x) \in (0,T] \times \mathcal{O}. \tag{1.11}$$

Moreover, the following holds:

**Theorem 4.** *Let $f \in C([0,T] \times \mathcal{O})$. Then,*

$$J(t,x,\tau) = \int_D Z(t,x;\tau,\xi) f(\tau,\xi) \mathrm{d}\xi$$

*is continuous in $(t,x,\tau)$, where $x \in \overline{\mathcal{O}}$ and $0 \le \tau < t \le T$. Moreover,*

$$\lim_{\tau \to t} J(t,x,\tau) = f(t,x)$$

*uniformly with respect to $(t,x)$, $\mathcal{O} \supseteq S \ni x$ closed and $0 < t \le T$.*

From Theorem 4 and the fact that $Z$ solves (1.11), it follows that $Z$ is a fundamental solution to (1.11). The idea under the parametric method is to look upon (1.11) as an approximation of (1.3) and view $Z$ as a principal part of the fundamental solution $\Gamma$ of (1.3). In the end, the method constructs a fundamental solution for (1.3) in the form

$$\Gamma(t,x;\tau,\xi) = Z(t,x;\tau,\xi) + \int_\tau^t \int_{\mathcal{O}} Z(t,x;\sigma,\eta) \Phi(\sigma,\eta,\tau,\xi) \mathrm{d}\eta \mathrm{d}\sigma, \tag{1.12}$$

where, for each $(\tau,\xi)$, $\Phi(t,x;\tau,\xi)$ is a solution of a Volterra integral equation with kernel

$$\begin{aligned} LZ(t,x;\sigma,y) = \sum_{i,j=1}^{d} [a_{ij}(t,x) - a_{ij}(\sigma,y)] \frac{\partial^2 Z(t,x;\sigma,y)}{\partial x_i \partial x_j} \\ + \sum_{i=1}^{d} \mu_i(t,x) \frac{\partial Z(t,x;\sigma,y)}{\partial x_i} + c(t,x) Z(t,x;\sigma,y), \end{aligned} \tag{1.13}$$

i.e.

$$\Phi(t,x;\tau,\xi) = LZ(t,x;\tau,\xi) + \int_\tau^t \int_{\mathcal{O}} LZ(t,x;\sigma,y) \cdot \Phi(t,x;\sigma,y) \mathrm{d}y \mathrm{d}\sigma. \tag{1.14}$$

Let us consider functions of the form

$$[0,T] \times \mathcal{O} \ni (t,x) \to W(t,x) = \int_0^t \int_{\mathcal{O}} \Gamma(t,x;\tau,\xi) f(\tau,\xi) \mathrm{d}\xi \mathrm{d}\tau \in \mathbb{R}, \tag{1.15}$$

where $f \in C([0,T] \times \overline{\mathcal{O}})$

**Theorem 5.** *If $f \in C([0,T] \times \overline{\mathcal{O}})$, then $W$ and $\partial W/\partial x_i$, $i = 1, dots, d$, are continuous. If $f$ is locally Hölder continuous in $x \in \mathcal{O}$, then also $\partial^2 W/\partial x_i \partial x_j$ and $\partial W/\partial t$ are continuous in $(0,T) \times \mathcal{O}$ and*

$$\partial W_t - \mathcal{L}_t W = f(t,x). \tag{1.16}$$

Proofs of Theorems 4 and 5 can be found in [59] (respectively, Chapter 1, Section 2, Theorem 1 and Chapter 1, Section 5, Theorem 9).

### 1.1.1 The Cauchy problem

It is important to extend the previous results to the case of $\mathcal{O}$ unbounded, since the case $\mathcal{O} = \mathbb{R}^d$ is of particular interest in order to present the theory for the Cauchy problem associated with (1.3). If $\mathcal{O}$ is unbounded, assumptions (A1) and (A2) must be modified as follows:

(A1)' $\mathcal{L}_t$ is uniformly parabolic in $[0,T] \times \overline{\mathcal{O}}$;

(A2)' the coefficients $a, \mu$ and $c$ are bounded continuous functions in $[0,T] \times \overline{\mathcal{O}}$ and (1.4), (1.5) and (1.6) hold in $[0,T] \times \overline{\mathcal{O}}$.

Definition 3 is still valid for $\mathcal{O}$ unbounded, with the additional requirement in (ii) that

$$(\exists h_1, h_2 > 0,\ \forall x \in \overline{\mathcal{O}}) \quad |f(x)| \leq h_1 \exp\left\{h_2 |x|^2\right\}. \tag{1.17}$$

The existence of a fundamental solution in an arbitrary domain $\mathcal{O} \subseteq \mathbb{R}^d$ is ensured by the following result, which is an extension to an unbounded domain of the one in Theorem 5.

**Theorem 6.** *Let $\mathcal{O}$ be any domain in $\mathbb{R}^d$ and assume that* (A1)' *and* (A2)' *hold. Then, there exists a fundamental solution $\Gamma(t,x;\tau,\xi)$ to* (1.3) *given by* (1.12) *and* (1.14). *If $f \in C([0,T] \times \mathcal{O})$ is such that* (1.17) *holds, then the function $W$ defined in* (1.15) *is uniformly continuous in $[0,T] \times \mathcal{O}$. If, for all $t \in [0,T]$, $f(t,\cdot)$ is also locally Hölder continuous, then $\partial W/\partial x_i, \partial^2 W/\partial x_i \partial x_j, \partial W/\partial t$ exist, are continuous functions and* (1.16) *holds.*

We now recall some results concerning the *Cauchy problem*, defined as follows. Given continuous functions $f : [0,T] \times \mathbb{R}^d \to \mathbb{R}$ and $u_0 : \mathbb{R}^d \to \mathbb{R}$, such that

$$|f(t,x)| \leq h_1 \exp\left\{h_2 |x|^2\right\}, \qquad \text{for } (t,x) \in [0,T] \times \mathbb{R}^d, \tag{1.18}$$

$$|u_0(x)| \leq h_1 \exp\left\{h_2 |x|^2\right\}, \qquad \text{for } x \in \mathbb{R}^d, \tag{1.19}$$

with $h_1, h_2$ positive constants, find a smooth function $u : [0,T] \times \mathbb{R}^d \to \mathbb{R}$ such that

$$\begin{cases} \partial_t u - \mathcal{L}_t u = f(t,x) & \text{in } [0,T] \times \mathbb{R}^d, \\ u(0,x) = u_0(x) & \text{in } \mathbb{R}^d. \end{cases} \tag{1.20}$$

**Theorem 7.** *Suppose that $\mathcal{L}_t$ satisfies* (A1)', (A2)', *with $\mathcal{O} = \mathbb{R}^d$, let $f$ and $u_0$ be continuous functions on $[0,T] \times \mathbb{R}^d$ and $\mathbb{R}^d$, respectively, satisfying assumptions* (1.18) *and* (1.19). *Assume also that, for all $t \in [0,T]$, $f(t,\cdot)$ is locally Hölder continuous with exponent $\alpha \in (0,1)$. Then $u : [0,T] \times \mathcal{O} \to \mathbb{R}$, defined by*

$$u(t,x) = \int_{\mathbb{R}^d} \Gamma(t,x;0,\xi) u_0(\xi) \mathrm{d}\xi - \int_0^t \int_{\mathbb{R}^d} \Gamma(t,x;\tau,\xi) f(\tau,\xi) \mathrm{d}\xi \mathrm{d}\tau, \tag{1.21}$$

*is a solution to* (1.20) *and*

$$|u(t,x)| \leq k_1 \exp\left\{k_1 |x|^2\right\}, \quad \textit{for all } (t,x) \in [0,T] \times \mathbb{R}^d \tag{1.22}$$

*for some $k_1 > 0$ and $k_2 > 0$.*

We conclude this section with a uniqueness result for the Cauchy problem. The following additional assumption on the data is required:

(A3)' the functions $a_{ij}, \partial a_{ij}/\partial x_h, \partial^2 a_{ij}/\partial x_h \partial x_k, \mu_i, \partial \mu_i/\partial x_h, c$, for $i, j, h, k = 1, \dots, d$, are bounded and continuous functions on $[0,T] \times \mathbb{R}^d$. In addition, they are uniformly Hölder continuous with exponent $\alpha \in (0,1)$ with respect to $x \in \mathbb{R}^d$, and (1.4) holds in $[0,T] \times \mathbb{R}^d$.

**Theorem 8.** *Let the operator $\mathcal{L}_t$ in* (1.2) *satisfy* (A1)' *and* (A3)'. *Then there exists at most one solution to the Cauchy problem* (1.20) *satisfying*

$$\int_0^T \int_{\mathbb{R}^d} |u(t,x)| \exp\left\{-k\,|x|^2\right\} \mathrm{d}x\mathrm{d}t < \infty,$$

*for some $k > 0$.*

We refer to [59] for the proofs of Theorem 7 (Chapter 1, Section 7, Theorem 12) and Theorem 8 (Chapter 1, Section 9, Theorem 16).

### 1.1.2 Initial-Boundary value problem

In this section, we deal with existence and uniqueness of solutions to the *Initial-Boundary value* problem

$$\begin{cases} \partial_t u - \mathcal{L}u = f(t,x) & \text{in } [0,T] \times \mathcal{O}, \\ u = g & \text{on } (0,T] \times \partial\mathcal{O}, \\ u(0,\cdot) = u_0 & \text{on } \overline{\mathcal{O}}, \end{cases} \tag{1.23}$$

where $\mathcal{O} \subseteq \mathbb{R}^d$ is a bounded domain and $f, u_0$ and $g$ are given functions. In this section, we will use the following notation for the initial and boundary conditions: for $(t,x) \in ((0,T] \times \partial\mathcal{O}) \cup \left(\{0\} \times \overline{\mathcal{O}}\right)$ we define $\Phi(t,x)$ as

$$\Phi(t,x) = \begin{cases} g(t,x) & \text{if } (t,x) \in (0,T] \times \partial\mathcal{O}, \\ u_0(x) & \text{if } (t,x) \in \{0\} \times \overline{\mathcal{O}}. \end{cases} \tag{1.24}$$

Let us recall some *maximum principle* satisfied by $\mathcal{L}$. We list below some assumptions that will be useful in the upcoming results:

(A) the coefficients of $\mathcal{L}$ in (1.2) are continuous on $\mathcal{O}$.

(B) $c(t,x) \geq 0$ in $\mathcal{O}$.

For any point in $P_0 = (t_0, x_0) \in [0,T] \times \mathcal{O}$ we denote by $\mathcal{S}(P_0)$ the set of points $P = (t,x) \in [0,T] \times \mathcal{O}$ that can be connected to $P_0$ by a continuous curve in $[0,T] \times \mathcal{O}$, along which the $t$-coordinate is nondecreasing from $P$ to $P_0$.

The strong maximum principle, which does not require $[0,T] \times \mathcal{O}$ to be bounded, asserts the following.

**Theorem 9.** *Assume* (A), (B), *and that $\mathcal{L}$ is parabolic. Then the following hold:*

(i) *If* $\partial_t u - \mathcal{L}u \leq 0$ *in* $[0,T]\times\mathcal{O}$ *and* $u$ *attains a positive maximum over* $[0,T]\times\mathcal{O}$ *at* $P_0=(t_0,x_0)\in[0,T]\times\mathcal{O}$, *then* $u$ *is constant on* $\mathcal{S}(P_0)$.

(ii) *If* $\partial_t u - \mathcal{L}u \geq 0$ *in* $[0,T]\times\mathcal{O}$ *and* $u$ *attains a negative minimum over* $[0,T]\times\mathcal{O}$ *at* $P_0=(t_0,x_0)\in[0,T]\times\mathcal{O}$, *then* $u$ *is constant on* $\mathcal{S}(P_0)$.

The following result is known as the weak maximum principle.

**Theorem 10.** *Assume* (A), (B), *that* $\mathcal{L}$ *is parabolic, that* $[0,T]\times\mathcal{O}$ *is bounded, and that* $u\in C([0,T]\times\overline{\mathcal{O}})$. *Then the following hold:*

(i) *If* $\partial_t u \leq \mathcal{L}u$ *in* $[0,T]\times\mathcal{O}$, *then for each* $P=(t,x)$ *such that* $u$ *has a positive maximum in* $\overline{\mathcal{S}(P)}$, *the maximum is obtained at some point in* $\mathcal{S}(P)^c$.

(ii) *If* $\partial_t u \geq \mathcal{L}u$ *in* $[0,T]\times\mathcal{O}$ *then for each* $P=(t,x)$ *such that* $u$ *has a negative minimum in* $\overline{\mathcal{S}(P)}$, *the minimum is obtained at some point in* $\mathcal{S}(P)^c$.

For each $P=(t,x)$, Theorem 10 does not exclude that the maximum (minimum) can be reached also at points of $\mathcal{S}(P)$. The results presented in Theorem 9 and in Theorem 10 (proven in [59, Chapter 2, Section 2]) can be used to show the uniqueness of the solution to (1.23).

**Theorem 11.** *Let* $\mathcal{L}$ *be parabolic on* $[0,T]\times\mathcal{O}$ *and assume that* (A) *holds. Then there exists at most one classical solution to the initial-boundary value problem* (1.23).

For the existence of the solution to (1.23) we need a different formulation of the concept of Hölder continuity in Definition 1. Let us define $\mathcal{O}_T=(0,T]\times\mathcal{O}$ and

$$d(P,Q):=\Big(|x-y|^2+|t-s|\Big)^{1/2}\quad\text{for } P=(t,x),Q=(s,y)\in\mathcal{O}_T.$$

We will use the following notations, for $\alpha\in(0,1)$, we set

$$\begin{aligned}|u|_0^{\mathcal{O}_T} &:= \sup_{\mathcal{O}_T}|u|\,,\\ \overline{H}_\alpha^{\mathcal{O}_T}(u) &:= \sup_{P,Q\in\mathcal{O}_T}\frac{|u(P)-u(Q)|}{d(P,Q)^\alpha},\\ \overline{|u|}_\alpha^{\mathcal{O}_T} &= |u|_0^{\mathcal{O}_T}+\overline{H}_\alpha^{\mathcal{O}_T}(u).\end{aligned}\tag{1.25}$$

Notice that $\overline{H}_\alpha^{\mathcal{O}_T}(u)<\infty$ if and only if $u$ is bounded and uniformly Hölder continuous of exponent $\alpha$. Since $\overline{|\cdot|}_\alpha^{\mathcal{O}_T}$ is a norm (see, e.g., [59]), we can denote by $\overline{C}_\alpha(\mathcal{O}_T)$ the normed space defined as

$$\overline{C}_\alpha(\mathcal{O}_T):=\{u:\mathcal{O}_T\to\mathbb{R}\mid \overline{|u|}_\alpha^{\mathcal{O}_T}<\infty\},$$

and by $D^m$ any partial derivative of order $m$ with respect to $x\in\mathbb{R}^d$: if $Du, D^2u$ and $\partial_t u$ exist, then we define

$$\overline{|u|}_{2+\alpha}^{\mathcal{O}_T}=\overline{|u|}_\alpha^{\mathcal{O}_T}+\sum\overline{|D_xu|}_\alpha^{\mathcal{O}_T}+\sum\overline{|D_x^2u|}_\alpha^{\mathcal{O}_T}+|\partial_t u|_\alpha^{\mathcal{O}_T},$$

and the normed space

$$\overline{C}_{2+\alpha}(\mathcal{O}_T) := \{u : \mathcal{O}_T \to \mathbb{R} \mid \overline{|u|}_{2+\alpha}^{\mathcal{O}_T} < \infty\}.$$

We can now state the following existence result for (1.23) (see e.g [59]).

**Theorem 12.** *Assume that $\mathcal{O}_T$ is such that, for every $(t,x) \in [0,T] \times \partial\mathcal{O}$ there exists a $(d+1)$-dimensional neighborhood $V$ such that $V \cap [0,T] \times \partial\mathcal{O}$ can be represented, for some $i = 1, \dots, d$, in the form*

$$x_i = h(t, x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_d),$$

*with $h, \partial_x h, \partial_x^2 h, \partial_t h$ continuous. Suppose that $\mathcal{L}$ is uniformly parabolic and that $a_{ij}, \mu_i, c, f$ are uniformly Hölder continuous of exponent $\alpha$ in $\mathcal{O}$. Moreover, let $K > 0$ be such that*

$$\overline{|a_{ij}|}_{\alpha}^{\mathcal{O}_T} \le K, \quad \overline{|\mu_i|}_{\alpha}^{\mathcal{O}_T} \le K, \quad \overline{|c|}_{\alpha}^{\mathcal{O}_T} \le K.$$

*Suppose that $\Phi$, defined in (1.24), belongs to $\overline{C}_{2+\alpha}(\mathcal{O}_T)$ and that $\partial_t\Phi - \mathcal{L}\Phi = f$ on $\{0\} \times \partial\mathcal{O}$. Then there exists a unique classical solution $u \in \overline{C}_{2+\alpha}$ to the Initial-Boundary value problem (1.23).*

### 1.1.3 Feynman-Kac formulae

There are intrinsic relations between stochastic differential equations and second-order parabolic equations because, from the physics point of view, both type of equations describe diffusion-type phenomena. In this line, it is possible to use the solution of some stochastic differential equations to represent the solutions of some second-order PDEs: such results are called *Feynman-Kac formulae*. Let us first consider the backward Cauchy problem:

$$\begin{cases} \partial_t u + \sum_{i,j=1}^{d} a_{ij} \dfrac{\partial^2 u}{\partial x_i \partial x_j} + \sum_{i=1}^{d} \mu_i \dfrac{\partial u}{\partial x_i} + cu + f = 0 & \text{in } [0,T) \times \mathbb{R}^d, \\ u(T,x) = u_T(x) & \text{on } \mathbb{R}^d, \end{cases} \tag{1.26}$$

with $a_{ij}, \mu_i, c, f : [0,T] \times \mathbb{R}^d \to \mathbb{R}$ and $u_T : \mathbb{R}^d \to \mathbb{R}$. In this framework, the operator $\mathcal{L}$ does not need to be uniformly elliptic, it is possible to prove the following results also for degenerate diffusion terms. We also assume that, for some $r \in \mathbb{N}$, there exists $\sigma : [0,T] \times \mathbb{R}^d \to \mathbb{R}^{d \times r}$, with $r \le d$, such that

$$a(t,x) = (a_{ij}(t,x)) = \frac{1}{2}\sigma(t,x)\sigma(t,x)^\top, \quad \text{for all } (t,x) \in [0,T] \times \mathbb{R}^d. \tag{1.27}$$

Let us also assume that:

(F) the maps $\mu, c, f : [0,T] \times \mathbb{R}^d \to \mathbb{R}$, and $\sigma : [0,T] \times \mathbb{R}^d \to \mathbb{R}^{d \times r}$ are uniformly continuous, $c$ is bounded and there exists a constant $L > 0$ such that, for $\varphi(t,x) = \mu(t,x), \sigma(t,x), f(t,x)$

$$\begin{aligned} |\varphi(t,x) - \varphi(t,y)| &\le L\,|x-y| && \text{for all } t \in [0,T], x, y \in \mathbb{R}^d, \\ |\varphi(t,0)| &\le L && \text{for all } t \in [0,T]. \end{aligned}$$

**Theorem 13.** *Assume that* (F) *hold. Then* (1.26) *admits a unique solution* $u$ *which has the following representation: for* $(t,x) \in [0,T) \times \mathbb{R}^d$

$$u(t,x) = \mathbb{E}\left[\int_t^T f(s, X(s;t,x)) \exp\left\{-\int_t^s c(r, X(r;t,x))\mathrm{d}r\right\}\mathrm{d}s \right. \\ \left. + u_T(X(T;t,x)) \exp\left\{-\int_t^T c(r, X(r;t,x))\mathrm{d}r\right\}\right], \tag{1.28}$$

*where* $X(\cdot) = X(\cdot;t,x)$ *is the unique strong solution to*

$$\begin{cases} \mathrm{d}X(s) = \mu(s, X(s))\mathrm{d}s + \sigma(s, X(s))\mathrm{d}W(s), & \text{for } s \in [t,T], \\ X(t) = x, \end{cases} \tag{1.29}$$

*where* $W(\cdot)$ *is an* $r$*-dimensional Brownian motion starting at time* $t$ $(W(t) = 0)$.

There exists an analogue result for the terminal-boundary value problem for a parabolic equation:

$$\begin{cases} \partial_t u + \sum_{i,j=1}^d a_{ij} \frac{\partial^2 u}{\partial x_i \partial x_j} + \sum_{i=1}^d \mu_i \frac{\partial u}{\partial x_i} + cu + f = 0 & \text{in } [0,T) \times \mathcal{O}, \\ u(t,x) = g(t,x) & \text{in } [0,T) \times \partial\mathcal{O}, \\ u(T,x) = u_T(x) & \text{on } \overline{\mathcal{O}}, \end{cases} \tag{1.30}$$

where $\mathcal{O} \subseteq \mathbb{R}^d$ is a bounded domain with smooth ($C^1$) boundary $\partial\mathcal{O}$.

**Theorem 14.** *Assume that* (F) *holds with all the functions defined on* $[0,T] \times \overline{\mathcal{O}}$ *and let* $\Psi$*, defined as*

$$\Psi(t,x) = \begin{cases} g(t,x) & (t,x) \in [0,T) \times \partial\mathcal{O}, \\ u_T(x) & (t,x) \in \{T\} \times \overline{\mathcal{O}}, \end{cases}$$

*be continuous on* $([0,T) \times \partial\mathcal{O}) \cup \left(\{T\} \times \overline{\mathcal{O}}\right)$*. Then* (1.26) *admits a unique solution* $u$ *such that, for every* $(t,x) \in [0,T) \times \mathbb{R}^d$*,*

$$u(t,x) = \mathbb{E}\left[\int_t^\tau f(s, X(s;t,x)) \exp\left\{-\int_t^s c(r, X(r;t,x))\mathrm{d}r\right\}\mathrm{d}s \right. \\ \left. + \Psi(X(\tau;t,x)) \exp\left\{-\int_t^\tau c(r, X(r;t,x))\mathrm{d}r\right\}\right], \tag{1.31}$$

*where* $X(\cdot) = X(\cdot;t,x)$ *is the unique strong solution of* (1.29) *and*

$$\tau = \tau(t,x) = \inf\{s \in [t,T] \,|\, X(s;t,x) \notin \mathcal{O}\}.$$

A proof for Theorems 13 and 14 can be found in [109]. The solutions of the stochastic differential equation in (1.29) are called the *characteristic curves* of equation (1.26).

We conclude this section with a representation formula for the initial-boundary value problem with mixed boundary conditions of Neumann and

Dirichlet types. Consider the problem

$$\begin{cases} \partial_t u + \sum_{i,j=1}^{d} a_{ij} \dfrac{\partial^2 u}{\partial x_i \partial x_j} + \sum_{i=1}^{d} \mu_i \dfrac{\partial u}{\partial x_i} + cu + f = 0 & \text{in } [0,T) \times \mathcal{O}, \\ u(t,x) = g_1(t,x) & \text{in } [0,T) \times \partial_1 \mathcal{O}, \\ \dfrac{\partial}{\partial n} u(t,x) = g_2(t,x) & \text{in } [0,T) \times \partial_2 \mathcal{O}, \\ u(T,x) = u_T(x) & \text{on } \overline{\mathcal{O}}, \end{cases} \tag{1.32}$$

where $\mathcal{O} \subseteq \mathbb{R}^d$ is a convex bounded domain with a $C^2$ boundary $\partial\mathcal{O}$, $\partial_1\mathcal{O} \subset \partial\mathcal{O}$ and $\partial_2\mathcal{O} = \partial\mathcal{O} \setminus \partial_1\mathcal{O}$. For this problem, if $\mu$ and $\sigma$ are uniformly Lipschitz continuous in the space variable, the diffusion process with coefficients $\mu$ and $\sigma$ and normal reflection in $\overline{\mathcal{O}}$ starting at $x_0 \in \overline{\mathcal{O}}$ is well defined. This means that there exists a unique increasing stochastic process $\{\xi(s)\}_{t \le s \le T}$ called *local time* and a unique stochastic proces $\{n(s)\}_{t \le s \le T}$ such that $n(s)$ is a normalized inward vector at $X(s) \in \partial\mathcal{O}$ and $(X, \xi, n)$ satisfies

$$\begin{cases} \mathrm{d}X(s) = \mu\left(s, X(s)\right) \mathrm{d}s + \sigma(s, X(s)) \mathrm{d}W(s) + n(s) \mathrm{d}\xi(s), & s \in [t,T], \\ \xi(s) = \displaystyle\int_t^s \mathbb{I}_{\partial\mathcal{O}}\left(X(r)\right) \mathrm{d}\xi(r), & s \in [t,T] \\ X(t) = x. \end{cases} \tag{1.33}$$

The following result holds (see e.g. [40] for a proof).

**Theorem 15.** *Let $u$ be a classical solution to* (1.32) *and suppose that $\mathcal{O}$ is convex. Then, for every $(t,x) \in [0,T] \times \mathcal{O}$ it holds that*

$$\begin{aligned} u(t,x) = \mathbb{E}\Bigg[ \int_t^\tau f(s, X(s;t,x)) \exp\left\{ -\int_t^s c(r, X(r;t,x)) \mathrm{d}r \right\} \mathrm{d}s \\ + \Psi(X(\tau;t,x)) \exp\left\{ -\int_t^\tau c(r, X(r;t,x)) \mathrm{d}r \right\} \\ - \int_t^{\min\{\tau,T\}} g_2(s, X(s;t,x)) \exp\left\{ -\int_t^s c(r, X(r;t,x)) \mathrm{d}r \right\} \mathrm{d}\xi(s) \Bigg], \end{aligned} \tag{1.34}$$

*where $(X, \xi)$ solves* (1.33) *and $\tau$ is defined as*

$$\tau = \begin{cases} \inf\{s \,|\, t \le s \le T, X(s) \in \partial_1\mathcal{O}\} & \text{if } \{s \,|\, t \le s \le T, X(s) \in \partial_1\mathcal{O}\} \neq \emptyset, \\ +\infty & \text{otherwise.} \end{cases}$$

### 1.1.4 The case of nonlinear source term: semilinear parabolic equations

Let us now briefly recall some important results on the theory of semilinear parabolic equations of the form

$$\begin{cases} \partial_t u(t,x) - \mathcal{L}_t u(t,x) = f(t,x,u) & \text{for } (t,x) \in (0,T] \times \mathcal{O}, \\ u(t,x) = g(t,x) & \text{for } (t,x) \in (0,T] \times \partial\mathcal{O}, \\ u(0,x) = u_0(x) & \text{for } x \in \overline{\mathcal{O}}, \end{cases} \tag{1.35}$$

where

$$(\mathcal{L}_t u)(t,x) = \sum_{i,j=1}^{d} a_{ij}(t,x)\frac{\partial^2 u}{\partial x_i \partial x_j} + \sum_{i=1}^{d} \mu_i(t,x)\frac{\partial u}{\partial x_i}.$$

A procedure often used for to solve (1.35) is the following. Set $w = Tv$ if $w$ solves

$$\partial_t w - \mathcal{L}_t w = f(t,x,v)$$

with same initial condition as $u$. The idea is to prove that, if we restrict $v$ to an appropriate functional space, then the operator $Tv$ is well-defined and has a fixed point, which is the solution $u$ to (1.35). We can now state some existence and uniqueness results for the solution to (1.35), whose proofs can be found, for instance, in [59, Chapter 7]. Let us start with the uniqueness results, for various types of functions $f$.

**Theorem 16.** *Let $\partial_t - \mathcal{L}_t$ be a parabolic operator, with $a$ and $\mu$ continuous and bounded functions. Let $f$ be such that, for all $(t,x) \in [0,T] \times \mathcal{O}$, $f(t,x,\cdot)$ is nondecreasing. Then there exists al most one solution of problem* (1.35).

**Theorem 17.** *Let $\partial_t - \mathcal{L}_t$ be a parabolic operator, with $a$ and $\mu$ continuous bounded functions, and let $f$ be such that, for all $(t,x) \in [0,T] \times \mathcal{O}$, $f(t,x,\cdot)$ is locally Lipschitz, uniformly with respect to $(t,x)$. Then, there exists at most one solution of problem* (1.35).

If the function $f$ satisfies more restrictive conditions, then it is also possible to provide some esimates for the solution $u$ to (1.35).

**Theorem 18.** *Let $\partial_t - \mathcal{L}_t$ be a parabolic operator, with $a$ and $\mu$ continuous, and let $f$ be a continuous function satisfying that*

$$vf(t,x,v) \le C_1 v^2 + C_2, \quad \text{with } C_1, C_2 \ge 0,$$

*for all $(t,x) \in (0,T) \times \mathcal{O}$ and $v \in \mathbb{R}$. Then, setting*

$$\Phi(t,x) = \begin{cases} g(t,x) & \text{if } (t,x) \in (0,T] \times \partial\mathcal{O}, \\ u_0(x) & \text{if } (t,x) \in \{0\} \times \overline{\mathcal{O}}, \end{cases}$$

*for any solution $u$ to* (1.35)*, the following estimate holds*

$$|u(t,x)| \le \left[ \left( \frac{C_2}{k - C_1} \right)^{1/2} + \sup_{\partial\mathcal{O}_T} \{|\Phi\} \right] e^{kt},$$

*where $\partial\mathcal{O}_T = \{0\} \times \overline{\mathcal{O}} \cup (0,T] \times \partial\mathcal{O}$, for any $(t,x) \in (0,T] \times \mathcal{O}$ and $k > C_1$.*

We can now recall the existence results: first, we need to define some notations that extend the ones defined in (1.25). Let $0 < \delta < 1$, and set

$$L^D[v] = \sup_{(t,x),(t',x') \in (0,T) \times \mathcal{O}} \frac{|v(t,x) - v(t',x')|}{|x - x'| + |t - t'|},$$

and

$$\begin{aligned} \overline{|v|}^{\mathcal{O}}_{1+\delta} &:= \overline{|v|}^{\mathcal{O}}_{\delta} + \textstyle\sum_i \overline{\left|\frac{\partial}{\partial x_i} u\right|}^{\mathcal{O}}_{\delta}, \\ \overline{|v|}^{\mathcal{O}}_{1-0} &:= \overline{|v|}^{\mathcal{O}}_{0} + L^D[v], \\ \overline{|v|}^{\mathcal{O}}_{2-0} &:= \overline{|v|}^{\mathcal{O}}_{1-0} + \textstyle\sum_i \overline{\left|\frac{\partial}{\partial x_i} u\right|}^{\mathcal{O}}_{\delta}. \end{aligned} \tag{1.36}$$

**Theorem 19.** *Suppose that $\partial\mathcal{O}$ can be parametrized with a function in $\overline{C}_{2-0} \cap \overline{C}_{2+\alpha}$, that $\mathcal{L}_t$ is parabolic, that $a_{ij}$ and $\mu_i$ are Hölder continuous of exponent $\alpha$ and that*

$$\sum_{i,j} \overline{|a_{ij}|}^{\mathcal{O}}_{\alpha} + \sum_i \overline{|\mu_i|}^{\mathcal{O}}_{\alpha} \leq K$$

*for some $K \geq 0$. Moreover, assume that $f$ is locally Hölder continuous and that there exist two positive constants $K, M_0$ such that, for any $M \geq M_0$, $2K|f(t,x,u)| \leq M$ in $(0,T] \times \mathcal{O}$ for all functions $u$ such that $\overline{|u|}_{1+\alpha} \leq M$. If $g \in \overline{C}_{2+\delta}$ for some $\alpha < \delta < 1$ and $\partial_t g - \mathcal{L}_t g = f(t,x,g)$ on $(0,T) \times \partial\mathcal{O}$, then there exists a classical solution to problem* (1.35).

**Theorem 20.** *Suppose that the assumptions on $\mathcal{L}_t, (0,T) \times \partial\mathcal{O}$ and $g$ from Theorem 19 hold. Moreover, assume that $f$ is a Hölder continuous function such that for all $(t,x) \in (0,T) \times \mathcal{O}$ and $v \in \mathbb{R}$*

$$vf(t,x,v) \leq C_1 u^2 + C_2, \quad \text{with } C_1, C_2 \geq 0,$$

*and*

$$|f(t,x,v)| \leq A(|v|),$$

*with $A$ being a positive increasing function. If $\partial_t g - \mathcal{L}_t g = f(t,x,g)$ on $\{0\} \times \partial\mathcal{O}$ then there exists a classical solution to problem* (1.35).

For the proofs on Theorems 19 and 20 we refer to [59, Chapter 7, Section 4]. Without imposing any growth condition on $f$ the existence of solutions for (1.35) can only be proven in a restriction of the domain $(0,T] \times \mathcal{O}$.

## 1.2 Hamilton-Jacobi-Bellman equations

In this section we recall some of the most important results concerning Hamilton-Jacobi-Bellman equations. Most of the results presented in this section can be found in [42], [101], and [44].

### 1.2.1 Viscosity solutions for first-order equations on $[0,T] \times \mathbb{R}^d$

The notion of *viscosity solution* was first introduced by M.G. Crandall and P-L. Lions in [44]: previously, the main obstacle in the study of Hamilton-Jacobi equations was the lack of a notion of solution that had the good properties of existence and uniqueness. The study of viscosity solutions has begun by studying the following two classes of first-order problems: the stationary Dirichlet problem

$$H(x,u,Du) = 0 \text{ if } x \in \mathbb{R}^d, \tag{1.37}$$

and the time-dependent Cauchy problem

$$\begin{cases} \partial_t u + H(t,x,u,Du) = 0 & \text{for } (t,x) \in (0,T] \times \mathbb{R}^d, \\ u(0,x) = u_0(x) & \text{for } x \in \mathbb{R}^d. \end{cases} \tag{1.38}$$

In (1.37) and (1.38), $u_0$ is a given functions and $H : \mathbb{R}^d \times \mathbb{R} \times \mathbb{R}^d \to \mathbb{R}$ (respectively, $H : [0,T] \times \mathbb{R}^d \times \mathbb{R} \times \mathbb{R}^d \to \mathbb{R}$) is called the *Hamiltonian*. In (1.37) and (1.38) the Hamiltonian depends nonlinearly on the gradient $Du$ of $u$. It is well known that, in general, these problems do not have classical solutions even if, in the case of (1.38), the initial data $u_0$ is smooth. It is possible to deal with this problem by looking for a generalized solution in Sobolev spaces which satisfy the equation only *almost everywhere*. However, in this approach, the problem of uniqueness persists since it is possible to find several solutions to (1.37) and (1.38) in the generalized sense. From now on we will focus on (1.38). First, we recall the definition of *upper and lower semi-continuous functions*.

**Definition 21.** *A function $f : X \to \mathbb{R}$ is* upper semi-continuous *($USC(X)$) if $\{x \in X | f(x) < y\}$ is an open set for every $y \in \mathbb{R}$. A function $f : X \to \mathbb{R}$ is* lower semi-continuous *($LSC(X)$) if $\{x \in X | f(x) > y\}$ is an open set for every $y \in \mathbb{R}$.*

In what follows we will use the notation $BUC(X)$ to indicate the space of bounded and uniformly continuous functions $f : X \to \mathbb{R}$. We can now introduce the definition of viscosity sub- and supersolution to (1.38) (see e.g. [101]).

**Definition 22.** *A function $u \in USC([0,T] \times \mathbb{R}^d)$ is a* viscosity subsolution *to* (1.38) *if, for every $\varphi \in C^\infty([0,T] \times \mathbb{R}^d)$ such that $u - \varphi$ has a local maximum at $(t_0, x_0) \in (0,T] \times \mathbb{R}^d$, we have*

$$\partial_t \varphi(t_0,x_0) + H(t_0,x_0,D\varphi(t_0,x_0)) \leq 0 \text{ and } u(0,x) - u_0(x) \leq 0 \text{ for } x \in \mathbb{R}^d.$$

*A function $u \in LSC([0,T] \times \mathbb{R}^d)$ is a* viscosity supersolution *to* (1.38) *if, for every $\varphi \in C^\infty([0,T] \times \mathbb{R}^d)$, such that $u - \varphi$ has a local minimum at $(t_0, x_0) \in (0,T] \times \mathbb{R}^d$, we have*

$$\partial_t \varphi(t_0,x_0) + H(t_0,x_0,D\varphi(t_0,x_0)) \geq 0 \text{ and } u(0,x) - u_0(x) \geq 0 \text{ for } x \in \mathbb{R}^d.$$

*A function $u \in BUC([0,T] \times \mathbb{R}^d)$ is a* viscosity solution *to* (1.38) *if it is both a viscosity sub- and supersolution.*

From now on, we will use the notation $B_R(z)$ for the $d$-dimensional ball of radius $R > 0$ centered at $z \in \mathbb{R}^d$. In [101], the following assumptions on the data are considered.

(H0) The Hamiltonian $H \in C([0,T] \times \mathbb{R}^d \times \mathbb{R} \times \mathbb{R}^d)$ is uniformly continuous in $[0,T] \times \mathbb{R}^d \times [-R,R] \times \{x \in \mathbb{R}^d | \, |x| < R\}$, for each $R > 0$.

(H1) There is a constant $C > 0$ such that

$$C = \sup_{[0,T] \times \mathbb{R}^d} |H(t,x,0,0)| < \infty.$$

(H2) For $R > 0$ there exists $\gamma_R \in \mathbb{R}$, such that $H(t,x,r,p) - H(t,x,s,p) \geq \gamma_R(r-s)$ for $-R \leq s \leq r \leq R$, $t \in [0,T]$, $p \in \mathbb{R}^d$.

(H3) If we define

$$\Lambda_R(\alpha) = \sup\{|H(t,x,r,p) - H(t,y,r,p)| \mid |x-y| \leq \alpha, \\ |p| \leq R, |r| \leq R, t \in [0,T]\},$$

then $\lim_{\alpha \downarrow 0} \Lambda_R(\alpha) = 0$.

(H4) For $R > 0$ there exists $C_R > 0$ such that

$$|H(t,x,r,p) - H(t,y,r,p)| \leq C_R(1+|p|)\,|x-y|$$

for $t \in [0,T]$, $|r| \leq R$, and $x, y, p \in \mathbb{R}^d$.

(H5) There exists a differentiable Lipschitz function $\mu : \mathbb{R}^d \to [0,\infty)$ and a continuous function $h : [0,\infty) \times [0,\infty) \to [0,\infty)$, decreasing in both arguments, such that $h(0,\cdot) = 0$, $\lim\limits_{|x| \to \infty} \mu(x) = 0$, and

$$H(t,x,r,p) - H(t,x,r,p + \lambda D\mu(x)) \leq h(\lambda, |p|)$$

for $(t,x,r,p) \in [0,T] \times \mathbb{R}^d \times \mathbb{R} \times \mathbb{R}^d$ and $\lambda \in [0,1]$.

(H6) There exist $r_0 > 0$ and, for each $\varepsilon > 0$, a continuous function $\omega_\varepsilon : [0,T] \times \overline{\Delta} \to [0,\infty)$, where

$$\Delta = \{(x,y) \in \mathbb{R}^d \times \mathbb{R}^d \mid |x-y| < r_0\},$$

which is Lipschitz continuous and differentiable in $[0,T] \times \Delta$ and satisfies:

(i) for $r \in \mathbb{R}$ and $(t,x,y) \in [0,T] \times \Delta$

$$\partial_t \omega_\varepsilon(t,x,y) + H(t,x,r,D_x\omega_\varepsilon(x,y)) - H(t,y,r,-D_y\omega_\varepsilon(x,y)) \geq 0.$$

(ii) for $(x,y) \in [0,T] \times \partial\Delta$

$$\omega_\varepsilon(t,x,y) \leq \varepsilon \text{ for } x \in \mathbb{R}^d \text{ and } \omega_\varepsilon(t,x,y) \geq 1/\varepsilon.$$

(iii) for $0 < r \leq r_0$

$$\liminf_{\varepsilon \downarrow 0} \{\omega_\varepsilon(0,x,y) = \mid |x-y| \geq r\} = +\infty.$$

Under these assumptions it is possible to prove a comparison principle (see e.g. [42, Theorem 2]).

**Theorem 23** (Comparison principle)**.** *Assume that $H$ is continuous and that the map $r \to H(t,x,r,p)$ is nondecreasing for all $(t,x,p) \in [0,T] \times \mathbb{R}^d \times \mathbb{R}^d$. Assume that* (H5) *and* (H6) *hold. Let $u$ and $v \in C([0,T] \times \mathbb{R}^d)$ be a viscosity*

*sub- and supersolution to* (1.38)*, respectively. Assume that* $C > 0$ *is a constant such that, for* $(x, y) \in \Delta$ *and* $t \in [0, T]$ *either*

$$|u(t,x) - v(t,y)| \leq C \quad \text{or} \quad |u(t,y) - v(t,x)| \leq C \tag{1.39}$$

*holds, and that either* $u(0,\cdot)$ *or* $v(0,\cdot)$ *is uniformly continuous. If also*

$$\sup_{[0,T]\times\mathbb{R}^d} (u - v) < \infty,$$

*then*

$$\sup_{[0,T]\times\mathbb{R}^d} (u - v)^+ \leq \sup_{\mathbb{R}^d} (u(0,\cdot) - v(0,\cdot))^+. \tag{1.40}$$

Notice that (1.39) holds if $u$ or $v$ is the sum of a bounded and a uniformly continuous function.

Before presenting the results concerning the existence and uniqueness of a viscosity solution to (1.38), we would like to briefly explain the term *viscosity* solutions. If $H$ and $u_0$ are sufficiently smooth, it is possible to show that, given $\varepsilon > 0$, if we add a viscous term $-\varepsilon\Delta u$ to (1.38), then the classical solution to

$$\begin{cases} \partial_t u_\varepsilon - \varepsilon\Delta u_\varepsilon + H(t,x,u_\varepsilon,Du_\varepsilon) = 0 & \text{for } (t,x) \in (0,T]\times\mathbb{R}^d, \\ u_\varepsilon(0,x) = u_0(x) & \text{for } x \in \mathbb{R}^d \end{cases} \tag{1.41}$$

converges, as $\varepsilon \to 0$, uniformly in $[0,T]\times\mathbb{R}^d$ to a function $u$ which is the viscosity solution of (1.38).

For a function $u : [0,T]\times\mathbb{R}^d \to \mathbb{R}$ we will use the notation $C^{1,2}([0,T]\times\mathbb{R}^d)$ to indicate that $u$ is differentiable in the first variable and twice differentiable in the second variable, with continuous derivatives, and the notation $C_b^2([0,T]\times\mathbb{R}^d)$ if $u \in C^2([0,T]\times\mathbb{R}^d)$ and is bounded.

**Theorem 24.** *Assume that*

- $H \in C^2([0,T]\times\mathbb{R}^d\times\mathbb{R}\times\mathbb{R}^d)$*,*
- $H$ *is bounded,*
- $H$ *satisfies* (H1)*,* (H2)*, and* (H4)*, with* $\gamma = \gamma_R \leq 0$ *for every* $R > 0$*.*

*For* $u_0 \in C_b^2(\mathbb{R}^d)$ *and* $\varepsilon > 0$*, let* $u_\varepsilon \in BUC([0,T]\times\mathbb{R}^d) \cap C^{1,2}([0,T]\times\mathbb{R}^d)$ *be the unique solution to* (1.41)*. Then, there exists* $u \in BUC([0,T]\times\mathbb{R}^d)$ *viscosity solution to* (1.38) *such that* $u_\varepsilon \to u$ *uniformly on* $[0,T]\times\mathbb{R}^d$ *as* $\varepsilon \to 0$*. Moreover,*

$$\sup_{\tau\in[0,T]}\sup_{x\in\mathbb{R}^d} |u_\varepsilon(\tau,x) - u(\tau,x)| \leq K\sqrt{\varepsilon}, \tag{1.42}$$

*where* $K$ *is a positive constant depending on* $\sup_{x\in\mathbb{R}^d}|u_0|$ *and* $sup_{x\in\mathbb{R}^d}|Du_0|$*.*

**Remark 25.** *The strategy of finding a solution to* (1.38) *starting from the solution to* (1.41) *and then passing to the limit as* $\varepsilon \to 0$*, is called* vanishing viscosity method*. It may be natural to think of a numerical method for approximating the viscosity solution to* (1.38) *starting from the solution to* (1.41)*. However, even though such a scheme would give the desired numerical approximation, relation* (1.42) *gives an explicit estimate for the rate of convergence, which is only* $O(\sqrt{\varepsilon})$*. Therefore, a numerical scheme based on the vanishing viscosity technique would not be efficient.*

It is possible to prove the following existence result for viscosity solutions to (1.38).

**Theorem 26.** *Assume that $H : [0,T] \times \mathbb{R}^d \times \mathbb{R} \times \mathbb{R}^d \to \mathbb{R}$ satisfies* (H0), (H1), (H2)*, and either* (H3) *or* (H4)*. Then, for every $u_0 \in BUC(\mathbb{R}^d)$ there exists $T = T(\sup_{x\in\mathbb{R}^d} |u_0|) > 0$ and $u \in BUC([0,T] \times \mathbb{R}^d)$ such that $u$ is the unique viscosity solution to* (1.38) *in $[0,T] \times \mathbb{R}^d$. Moreover, if $\gamma_R$ in* (H2) *is independent of $R$, then* (1.38) *has a unique viscosity solution in $[0,T] \times \mathbb{R}^d$ for every $T > 0$.*

**Theorem 27.** *Let $u, v \in BUC([0,T] \times \mathbb{R}^d)$ be viscosity solutions to* (1.38) *with initial data $u_0$ and $v_0$, respectively. Assume that $H$ satisfies* (H0), (H2)*, and either* (H3) *or* (H4)*. Let $R_0 = \max(\sup_{x\in\mathbb{R}^d} |u_0|, \sup_{x\in\mathbb{R}^d} |v_0|)$ and $\gamma = \gamma_{R_0}$. Then, for every $t \in [0,T]$,*

$$\sup_{x\in\mathbb{R}^d} |u(t,x) - v(t,x)| \le e^{-\gamma t} \sup_{x\in\mathbb{R}^d} |u_0(x) - v_0(x)| .$$

Thorems 26 and 27, both proven in [101], imply that the viscosity solution to (1.38) exists and is unique.

We conclude this section with a stability result.

**Theorem 28.** *Let $u_n \in C([0,T] \times \mathbb{R}^d)$ be a viscosity solution to*

$$\begin{cases} \partial_t u + H_n(t,x,u,Du) = 0 & \text{for } (t,x) \in (0,T] \times \mathbb{R}^d, \\ u(0,x) = u_{0n}(x) & \text{for } x \in \mathbb{R}^d. \end{cases}$$

*Assume that $H_n \to H$ uniformly on $[0,T] \times \mathbb{R}^d \times [-R,R] \times B_R(0)$, for each $R > 0$. If $u_n \to u$ locally uniformly in $(0,T] \times \mathbb{R}^d$, then $u$ is a viscosity solution to*

$$\partial_t u + H(t,x,u,Du) = 0, \qquad \text{for } (t,x) \in (0,T] \times \mathbb{R}^d.$$

*Moreover, if $u_{0n} \to u_0$ uniformly on $\mathbb{R}^d$ and $u_n \to u$ uniformly on $[0,T] \times \mathbb{R}^d$, then $u$ is a viscosity solution to* (1.38)*.*

### 1.2.2 Viscosity solutions for second-order equations on $[0,T] \times \mathbb{R}^d$

We now present some results from the theory of viscosity solutions for parabolic equations of the form

$$\begin{cases} \partial_t u + H(t,x,u,Du,D^2u) = 0 & \text{in } (0,T] \times \mathbb{R}^d, \\ u(0,x) = u_0(x) & \text{for } x \in \mathbb{R}^d, \end{cases} \tag{1.43}$$

where $T > 0$ and $D^2u$ is the Hessian matrix of $u$. The results in this section are mainly taken from [61] and [110]. We now give the definition of viscosity sub- and supersolution to (1.43).

**Definition 29.** *A function $u \in USC([0,T] \times \mathbb{R}^d)$ is a* viscosity subsolution *to* (1.43) *if for every $\varphi \in C^{1,2}([0,T] \times \mathbb{R}^d)$ and a local maximum point $(t_0,x_0) \in (0,T] \times \mathbb{R}^d$ of $u - \varphi$:*

$$\partial_t \varphi(t_0,x_0) + H(t_0,x_0,\varphi(t_0,x_0),D\varphi(t_0,x_0),D^2\varphi(t_0,x_0)) \le 0$$

*and*

$$u(0,x) - u_0(x) \leq 0 \text{ for } x \in \mathbb{R}^d.$$

*A function* $u \in LSC([0,T] \times \mathbb{R}^d)$ *is a* viscosity supersolution *to* (1.38) *if for every* $\varphi \in \mathcal{C}^{1,2}([0,T] \times \mathbb{R}^d)$ *and a local minimum point* $(t_0, x_0) \in (0,T] \times \mathbb{R}^d$ *of* $u - \varphi$*:*

$$\partial_t \varphi(t_0,x_0) + H(t_0,x_0,\varphi(t_0,x_0), D\varphi(t_0,x_0), D^2\varphi(t_0,x_0)) \geq 0$$

*and*

$$u(0,x) - u_0(x) \geq 0 \text{ for } x \in \mathbb{R}^d.$$

*A function* $u \in BUC([0,T] \times \mathbb{R}^d)$ *is said to be a* viscosity solution *to* (1.43) *if it is both a viscosity sub- and supersolution to* (1.43)*.*

From now on, $\mathcal{S}^d$ will denote the space of symmetric matrices in $\mathbb{R}^{d\times d}$. We consider the following assumptions.

(F1) $H$ is degenerate elliptic, meaning that $H(t,x,r,p,X+Y) \leq H(t,x,r,p,X)$ in $(0,T] \times \mathbb{R}^d \times \mathbb{R} \times (\mathbb{R}^d \setminus \{0\}) \times \mathcal{S}^d$ if $Y \geq 0$.

(F2) $H : (0,T] \times \mathbb{R}^d \times \mathbb{R} \times (\mathbb{R}^d \setminus \{0\}) \times \mathcal{S}^d \to \mathbb{R}$ is continuous.

(F3) $-\infty < H_*(t,x,r,0,0) = H^*(t,x,r,0,0) < +\infty$ for all $(t,x,r) \in (0,T] \times \mathbb{R}^d \times \mathbb{R}$, where $H_*$ and $H^*$ are, respectively, the lower and upper semi-continuous envelope of $H$.

(F4) $H$ is uniformly bounded in $(t,x,r)$, locally in $X$, i.e. for every $R > 0$,

$$c_R = \sup\{|H(t,x,r,p,X)| \,|\, |p|, |X| \leq R,$$

$$(t,x,r,p,X) \in (0,T] \times \mathbb{R}^d \times \mathbb{R} \times (\mathbb{R}^d \setminus \{0\}) \times \mathcal{S}^d\} < \infty,$$

where $|X| = \max_{i,j} |X_{i,j}|$.

(F5) For every $K > 0$ there exists a constant $c_0 = c_0(d,T,K)$ such that for all $(t,x,p,X) \in (0,T] \times \mathbb{R}^d \times (\mathbb{R}^d \setminus \{0\}) \times \mathcal{S}^d$, with $|r| \leq K$, $r \to H(t,x,r,p,X) + c_0 r$ is nondecreasing.

(F6) For every $R > \rho > 0$ there is a modulus of continuity $\omega = \omega_{R_\rho}$ such that

$$|H(t,x,r,p,X) - H(t,x,r,q,Y)| \leq \omega_{R_\rho}(|p-q| + |X-Y|)$$

for all $(t,x,r) \in (0,T] \times \mathbb{R}^d \times \mathbb{R}$, $\rho \leq |p|, |q| \leq R$ and $|X|, |Y| \leq R$.

(F7) There exist $\rho_0 > 0$ and a modulus of continuity $\omega_1$ such that

$$H^*(t,x,r,p,X) - H^*(t,x,r,0,0) \leq \omega_1(|p| + |X|),$$

$$H_*(t,x,r,p,X) - H_*(t,x,r,0,0) \geq -\omega_1(|p| + |X|)$$

if $(t,x,r) \in (0,T] \times \mathbb{R}^d \times \mathbb{R}$ and $|p|, |X| \leq \rho_0$.

(F8) There exists a modulus of continuity $\omega_2$ such that

$$|H(t,x,r,p,X) - H(t,y,r,p,X)| \leq \omega_2(|x-y|\,(|p|+1))$$

for $y \in \mathbb{R}^d$, $(t,x,r,p,X) \in (0,T] \times \mathbb{R}^d \times \mathbb{R} \times (\mathbb{R}^d \setminus \{0\}) \times \mathcal{S}^d$.

The following result states that almost everywhere solutions are also viscosity solutions.

**Proposition 30.** *Assume* (F1). *If* $u \in C((0,T] \times \mathbb{R}^d)$, $u(\cdot,x) \in W^{1,d+1}((0,T])$ *for all* $x \in \mathbb{R}^d$, $u(t,\cdot) \in W^{2,d+1}(\mathbb{R}^d)$ *for all* $t \in [0,T]$ *and*

$$\partial_t u + H(t,x,u,Du,D^2u) = 0 \text{ a.e. in } (0,T] \times \mathbb{R}^d,$$

*then* $u$ *is a viscosity solution to* (1.43).

Given $u : [0,T] \times \mathbb{R}^d \to \mathbb{R}$, let us define its *upper semicontinuous envelope*

$$u^*(t,x) = \lim \sup_{\substack{(\tau,y)\to(t,x) \\ (\tau,y)\in(0,T]\times\mathbb{R}^d}} u(\tau,y), \tag{1.44}$$

and its *lower semicontinuous envelope*

$$u_*(t,x) = \lim \inf_{\substack{(\tau,y)\to(t,x) \\ (\tau,y)\in(0,T]\times\mathbb{R}^d}} u(\tau,y). \tag{1.45}$$

Notice that $u_* = -(-u)^*$, $u^* \in USC((0,T] \times \mathbb{R}^d)$, and $u_* \in LSC((0,T] \times \mathbb{R}^d)$. The following comparison principle for (1.43) holds.

**Theorem 31.** *Suppose that* $H$ *satisfies* (F1)-(F8). *Let* $u$ *and* $v$ *be viscosity sub- and supersolutions to* (1.43), *respectively. Assume that*

*(A1) there exists* $K > 0$, *independent of* $(t,x) \in (0,T] \times \mathbb{R}^d$, *such that* $u(t,x) \leq K(|x|+1)$ *and* $v(t,x) \geq -K(|x|+1)$ *for all* $(t,x) \in [0,T] \times \mathbb{R}^d$,

*(A2) there exists a modulus of continuity* $m_T$ *such that*

$$u^*(0,x) - v_*(,y) \leq m_T(|x-y|) \text{ for all } (x,y) \in \mathbb{R}^d \times \mathbb{R}^d,$$

*(A3)* $u^*(0,x) - v_*(0,y) \leq K(|x-y|+1)$ *on* $\mathbb{R}^d \times \mathbb{R}^d$ *for some* $K > 0$ *independent of* $(x,y) \in \mathbb{R}^d \times \mathbb{R}^d$.

*Then there is a modulus of continuity* $m$ *such that*

$$u^*(t,x) - v_*(t,y) \leq m(|x-y|) \text{ for } (t,x,y) \in (0,T] \times \mathbb{R}^d \times \mathbb{R}^d. \tag{1.46}$$

It is possible to prove the following stability result also for (1.43).

**Theorem 32.** *Let $Q_n$ be a sequence of sets, nondecreasing with respect to $n$, and such that $\cup_{n\to\infty} Q_n = (0,T]\times\mathbb{R}^d$.*

*Assume that $u_n \in USC(Q_n)$ is a viscosity subsolution to*

$$\partial_t u_n + H_n(t,x,u_n,Du_n,D^2u_n) = 0 \quad \text{in } Q_n,$$

*$u_n$ converges uniformly to a function $u$ on any compact subsets of $(0,T]\times\mathbb{R}^d$. Assume the existence of a function $H$ such that, for all sequences $(t,x_n,r_n,p_n,X_n) \underset{n\to\infty}{\longrightarrow} (t,x,r,p,X)$ we have*

$$\lim_{n\to\infty} H_n(t,x_n,r_n,p_n,X_n) \geq H(t,x,r,p,X).$$

*Then $u$ is a viscosity subsolution of*

$$\partial_t u + H(t,x,u,Du,D^2u) = 0 \quad \text{in } (0,T]\times\mathbb{R}^d.$$

*Assume that $u_n \in USC(Q_n)$ is a viscosity supersolution to*

$$\partial_t u_n + H_n(t,x,u_n,Du_n,D^2u_n) = 0 \quad \text{in } Q_n,$$

*$u_n$ converges uniformly to a function $u$ on any compact subsets of $(0,T]\times\mathbb{R}^d$. Assume, also, that there exists a function $H$ such that, for all sequences $(t,x_n,r_n,p_n,X_n) \underset{n\to\infty}{\longrightarrow} (t,x,r,p,X)$ we have that*

$$\lim_{n\to\infty} H_n(t,x_n,r_n,p_n,X_n) \leq H(t,x,r,p,X).$$

*Then $u$ is a viscosity supersolution to*

$$\partial_t u + H(t,x,u,Du,D^2u) = 0 \quad \text{in } (0,T]\times\mathbb{R}^d.$$

### 1.2.3 Viscosity solutions for parabolic Hamilton-Jacobi-Bellman equations on bounded domains

First, consider the following initial-boundary value problem of Dirichlet type

$$\begin{cases} \partial_t u + H(t,x,u,Du,D^2u) = 0 & \text{in } (0,T]\times\mathcal{O}, \\ u(t,x) = g(t,x) & \text{on } (0,T]\times\partial\mathcal{O}, \\ u(0,x) = u_0(x) & \text{for } x\in\overline{\mathcal{O}}. \end{cases} \tag{1.47}$$

where $\mathcal{O}\subset\mathbb{R}^d$ is an open domain, $g:[0,T]\times\partial\mathcal{O}\to\mathbb{R}$ is the boundary condition and $u_0:\overline{\mathcal{O}}\to\mathbb{R}$ is the initial data. The notion of viscosity solution for problem (1.47) can be found in [43] and is the following.

**Definition 33.** *A function $\overline{u}\in USC([0,T]\times\overline{\mathcal{O}})$ is a* viscosity subsolution *to (1.47) if, for each $\varphi\in C^{1,2}([0,T]\times\overline{\mathcal{O}})$, at each maximum point $(t_0,x_0)$ of $\overline{u}-\varphi$ we have that*

$$\partial_t\varphi(t_0,x_0) + H(t_0,x_0,\overline{u},D\varphi,D^2\varphi) \leq 0 \quad \text{if } (t_0,x_0)\in(0,T]\times\mathcal{O}, \tag{1.48}$$

$$\min\{\varphi(t_0,x_0) - g(t_0,x_0), \partial_t\varphi(t_0,x_0) + H(t_0,x_0,\overline{u},D\varphi,D^2\varphi)\} \leq 0 \quad \text{if } (t_0,x_0)\in(0,T]\times\partial\mathcal{O}, \tag{1.49}$$

$$\begin{aligned}\min\{\varphi(t_0,x_0)-u_0(x_0),\partial_t\varphi(t_0,x_0)+H(t_0,x_0,\overline{u},D\varphi,D^2\varphi),\\ \varphi(t_0,x_0)-g(t_0,x_0)\}\leq 0 \quad \text{if } (t_0,x_0)\in\{0\}\times\overline{\mathcal{O}}.\end{aligned} \tag{1.50}$$

*A function* $\underline{u}\in LSC([0,T]\times\overline{\mathcal{O}})$ *is a* viscosity supersolution *to* (1.47) *if, for each* $\varphi\in C^\infty([0,T]\times\overline{\mathcal{O}})$*, at each minimum point* $(t_0,x_0)$ *of* $\underline{u}-\varphi$ *we have that*

$$\partial_t\varphi(t_0,x_0)+H(t_0,x_0,\overline{u},D\varphi,D^2\varphi)\geq 0 \quad \text{if } (t_0,x_0)\in(0,T]\times\mathcal{O}, \tag{1.51}$$

$$\begin{aligned}\max\{\varphi(t_0,x_0)-g(t_0,x_0),\partial_t\varphi(t_0,x_0)+H(t_0,x_0,\overline{u},D\varphi,D^2\varphi)\}\geq 0\\ \text{if } (t_0,x_0)\in(0,T]\times\partial\mathcal{O},\end{aligned} \tag{1.52}$$

$$\begin{aligned}\max\{\varphi(t_0,x_0)-u_0(x_0),\partial_t\varphi(t_0,x_0)+H(t_0,x_0,\overline{u},D\varphi,D^2\varphi),\\ \varphi(t_0,x_0)-g(t_0,x_0)\}\geq 0 \quad \text{if } (t_0,x_0)\in\{0\}\times\overline{\mathcal{O}}.\end{aligned} \tag{1.53}$$

*A function* $u\in C([0,T]\times\overline{\mathcal{O}})$ *is a* viscosity solution *to* (1.47) *if it is both a viscosity sub- and supersolution.*

It is possible to prove a maximum principle for (1.47) (see e.g. [93] for a proof).

**Theorem 34** (Maximum principle)**.** *Assume that* $u_0$ *is bounded and that* $H$ *is bounded and continuous. If* $\overline{u}\in USC([0,T]\times\mathcal{O})$ *is a subsolution to* (1.47) *and* $\underline{u}\in LSC([0,T]\times\mathcal{O})$ *is a supersolution of* (1.47)*, then*

$$\overline{u}-\underline{u}\leq \sup_{(\{0\}\times\overline{\mathcal{O}})\cup((0,T]\times\partial\mathcal{O})}\{\overline{u}-\underline{u}\} \quad \text{in } [0,T]\times\overline{\mathcal{O}}.$$

It is also possible to prove a comparison principle for (1.47) (a proof can be found in [43]).

**Theorem 35** (Comparison principle)**.** *Assume that* $H$ *is continuous and proper, meaning that the inverse images of compact sets are compact. Moreover, suppose that there exists a modulus of continuity* $\omega:[0,\infty)\to[0,\infty)$ *such that for each* $t\in[0,T)$ *and for all* $\alpha>0$*,*

$$H(t,y,r,\alpha(x-y),Y)-H(t,x,r,\alpha(x-y),X)\leq\omega(\alpha\,|x-y|^2+|X-Y|)$$

*for* $x,y\in\mathcal{O}$*,* $X,Y\in\mathcal{S}^d$*, with* $X\leq Y$*. If* $\overline{u}\in USC([0,T]\times\mathcal{O})$ *is a subsolution of* (1.47) *and* $\underline{u}\in LSC([0,T]\times\mathcal{O})$ *is a supersolution to* (1.47)*, then*

$$\overline{u}\leq\underline{u} \quad \text{in } [0,T]\times\mathcal{O}.$$

Let us now state an existence and uniqueness result for (1.47). For a proof, we refer to [93].

**Theorem 36** (Existence and uniqueness)**.** *Assume that* $H$ *is smooth and bounded, that the boundary condition* $g\in C^{1,2}([0,T]\times\overline{\mathcal{O}})$ *and is compatible with* $u_0$ *at time* $t=0$*, meaning that* $\lim_{t\to0}g(t,x)=u_0(x)$ *for all* $x\in\partial\mathcal{O}$*. Then, there exists a unique continuous viscosity solution* $u$ *to* (1.47)*.*

Let us now focus on the boundary value problem of Neumann type,

$$\begin{cases} \partial_t u + H(t,x,u,Du,D^2u) = 0 & \text{in } (0,T]\times\mathcal{O}, \\ B(t,x,u,Du) = 0 & \text{on } (0,T]\times\partial\mathcal{O}, \\ u(0,x) = u_0(x) & \text{for } x\in\overline{\mathcal{O}}, \end{cases} \tag{1.54}$$

where $\mathcal{O}\subset\mathbb{R}^d$ is an open domain, $u_0 : \overline{\mathcal{O}} \to \mathbb{R}$ is the initial data and $B(t,x,u,Du) = 0$ on $(0,T]\times\partial\mathcal{O}$ is a non-linear boundary condition of the Neumann type, meaning that $B(t,x,r,p)$ is strictly increasing with respect to $p$ in the normal direction to $\partial\mathcal{O}$ at $x$. More precisely, we assume that for all $R>0$, there exists a constant $\nu_R>0$ such that

$$B(t,x,r,p+\lambda n(x)) - B(t,x,r,p) \geq \nu_r, \tag{N0}$$

for all $(t,x,r,p)\in(0,T]\times\partial\mathcal{O}\times[-R,R]\times\mathbb{R}^d$ and $\lambda>0$, where $n(x)$ denotes the outward normal to $\partial\mathcal{O}$ at point $x$. It is possible to give a definition of viscosity sub- and supersolution to (1.54) which is similar to Definition 33.

**Definition 37.** *A function $\overline{u}\in USC([0,T]\times\overline{\mathcal{O}})$ is a* viscosity subsolution *to* (1.54) *if, for each $\varphi\in C^\infty([0,T]\times\overline{\mathcal{O}})$, at each maximum point $(t_0,x_0)$ of $\overline{u}-\varphi$ we have that*

$$\partial_t\varphi(t_0,x_0) + H(t_0,x_0,\overline{u},D\varphi,D^2\varphi) \leq 0 \quad \text{if } (t_0,x_0)\in(0,T]\times\mathcal{O}, \tag{1.55}$$

$$\begin{aligned} \min\{B(t_0,x_0,\overline{u},D\varphi), \partial_t\varphi(t_0,x_0) + H(t_0,x_0,\overline{u},D\varphi,D^2\varphi)\} \leq 0 \\ \text{if } (t_0,x_0)\in(0,T]\times\partial\mathcal{O}, \end{aligned} \tag{1.56}$$

$$\begin{aligned} \min\{\varphi(t_0,x_0) - u_0(x_0), \partial_t\varphi(t_0,x_0) + H(t_0,x_0,\overline{u},D\varphi,D^2\varphi), \\ B(t_0,x_0,\overline{u},D\varphi)\} \leq 0 \quad \text{if } (t_0,x_0)\in\{0\}\times\overline{\mathcal{O}}. \end{aligned} \tag{1.57}$$

*A function $\underline{u}\in LSC([0,T]\times\overline{\mathcal{O}})$ is a* viscosity supersolution *to* (1.54) *if, for each $\varphi\in C^\infty([0,T]\times\overline{\mathcal{O}})$, at each minimum point $(t_0,x_0)$ of $\underline{u}-\varphi$ we have that*

$$\partial_t\varphi(t_0,x_0) + H(t_0,x_0,\underline{u},D\varphi,D^2\varphi) \geq 0 \quad \text{if } (t_0,x_0)\in(0,T]\times\mathcal{O}, \tag{1.58}$$

$$\begin{aligned} \max\{B(t_0,x_0,\underline{u},D\varphi), \partial_t\varphi(t_0,x_0) + H(t_0,x_0,\underline{u},D\varphi,D^2\varphi)\} \geq 0 \\ \text{if } (t_0,x_0)\in(0,T]\times\partial\mathcal{O}, \end{aligned} \tag{1.59}$$

$$\begin{aligned} \max\{\varphi(t_0,x_0) - u_0(x_0), \partial_t\varphi(t_0,x_0) + H(t_0,x_0,\underline{u},D\varphi,D^2\varphi), \\ B(t_0,x_0,\underline{u},D\varphi)\} \geq 0 \quad \text{if } (t_0,x_0)\in\{0\}\times\overline{\mathcal{O}}. \end{aligned} \tag{1.60}$$

*A function $u\in C([0,T]\times\overline{\mathcal{O}})$ is a* viscosity solution *to* (1.54) *if it is both a viscosity sub- and supersolution.*

We now present a list of properties which are needed in the main results for (1.54).

(N1) For all $R>0$, there exists a function $m_R\in C((0,\infty),\mathbb{R})$ such that $m_R(0^+)=0$ and, for $G=H$ and $B$,

$$G(t,x,r,p,X) - G(t,y,r,p,Y) \geq m_R((1+|p|)|x-y| + \alpha|x-y|^2),$$

if

$$-\alpha \begin{pmatrix} I & 0 \\ 0 & I \end{pmatrix} \leq \begin{pmatrix} X & 0 \\ 0 & -Y \end{pmatrix} \leq \alpha \begin{pmatrix} I & -I \\ -I & I \end{pmatrix},$$

for all $\alpha \geq 1$, $x, y \in \overline{\mathcal{O}}$, $|r| \leq R$, $p \in \mathbb{R}^d$, $X, Y \in \mathcal{S}^d$.

(N2) For all $R > 0$, there exists $n_R \in C((0,\infty), \mathbb{R})$, such that $n_R(0^+) = 0$ and for $G = H$ and $B$

$$|G(t,x,r,p,X) - G(t,y,r,q,Y)| \leq n_R(|p-q| + |X-Y|),$$

if $x$ is in some neighnorhood $V$ of $\partial\mathcal{O}$, $|r| \leq R$, $p, q \in \mathbb{R}^d$, $X, Y \in \mathcal{S}^d$.

(N3) for all $R > 0$, there exists $\gamma_R \geq 0$ such that for $G = H$ and $B$

$$G(t,x,r,p,X) - G(t,x,s,p,X) \geq \gamma_R(r-s),$$

for all $x \in \overline{\mathcal{O}}$, $p \in \mathbb{R}^d$, $X \in S^d$, $-R \leq s \leq r \leq R$.

Under the previous properties we have the following comparison principle for (1.54). For a proof we refer to [6].

**Theorem 38** (Comparison principle)**.** *Assume that $\partial\mathcal{O} \in W^{3,\infty}$ and that* (N0), (N1), (N2), *and* (N3) *hold. Then, if $u$ and $v$ are, respectively, a bounded u.s.c. viscosity subsolution and a bounded l.s.c. supersolution to* (1.54)*, we have*

$$u \leq v \quad \text{on } [0,T] \times \overline{\mathcal{O}}.$$

Theorem 38 is the fundamental result in order to get in the following result the existence and uniqueness of a solution to (1.54), for the proof of which we refer to [6].

**Theorem 39** (Existence and uniqueness)**.** *Assume that $\partial\mathcal{O} \in W^{3,\infty}$ and that* (N0), (N1), (N2), *and* (N3) *hold. Then there exists a unique viscosity solution to* (1.54)*.*

We conclude this section with a regularity result for the solution to (1.54), under suitable assumptions on the initial data. For the proof we refer to [6].

**Theorem 40.** *Assume that $\partial\mathcal{O} \in W^{3,\infty}$ and that* (N0), (N1), (N2), *and* (N3) *hold. Moreover, assume that $u_0 \in W^{2,\infty}(\overline{\mathcal{O}})$. Then the unique viscosity solution $u \in C([0,T] \times \overline{\mathcal{O}})$ to* (1.54) *is Lipschitz continuous.*

### 1.2.4 Deterministic and stochastic optimal control

The solution to Hamilton-Jacobi-Bellman equations is related to the theory of *optimal control*. Optimal control deals with the problem of finding a control law for a given system such that a certain optimality criterion is achieved. A control problem includes a cost functional that is a function of the state and the control variables. The most popular methods to solve optimal control problems are Pontryagin's maximum principle and dynamic programming. Let us now give a brief presentation of deterministic and stochastic optimal control problems and

show their links to parabolic equations. We refer to [5, 57, 58] and [11] for the proofs of the results in this section.

Consider the following controlled ordinary differential equation (ODE)

$$\begin{cases} \dot{x}(t) = \mu(x(t), \alpha(t)) \quad \text{if } t > 0, \\ x(0) = x_0, \end{cases} \tag{1.61}$$

where $x_0 \in \mathbb{R}^d$ and $\mu : \mathbb{R}^d \times A \to \mathbb{R}^d$, with $A \subset \mathbb{R}^m$ ($m \in \mathbb{N}$) being a non-empty set and $\alpha : [0, \infty) \to A$ is the control. The curve $x$ is the response of the system. The first problem is: given the initial point $x_0$ and a target set $S \subset \mathbb{R}^d$, is there a control that steers the system to $S$ in finite time? Given a control $\alpha$, we define its payoff by

$$P(\alpha(\cdot)) = \int_0^T r(x(t), \alpha(t))\mathrm{d}t + g(x(T)), \tag{1.62}$$

where $T > 0$, $r : \mathbb{R}^d \times A \to \mathbb{R}$ is the running payoff such that $r(x(\cdot), \alpha(\cdot)) \in L^1([0, T])$, $g : \mathbb{R}^d \to \mathbb{R}$ is the terminal payoff, and $x$ is defined by (1.61). The problem is to find a control $\alpha^*(\cdot)$ such that

$$P(\alpha^*(\cdot)) = \max_{\alpha(\cdot) \in A} P(\alpha(\cdot)). \tag{1.63}$$

**Definition 41.** *The function $H : \mathbb{R}^d \times \mathbb{R}^d \times \mathcal{A} \to \mathbb{R}$ defined by*

$$H(x, p, a) = \langle \mu(x, a), p \rangle + r(x, a), \quad \text{for } x \in \mathbb{R}^d,\ p \in \mathbb{R}^d,\ a \in A, \tag{1.64}$$

*is called the* control theory Hamiltonian.

**Theorem 42** (Pontryagin maximum principle)**.** *Assume that $\alpha^*$ is an optimal control for problem* (1.63) *and define $x^*$ as its associated trajectory. Then there exists $p^* : [0, T] \to \mathbb{R}^d$ such that*

$$\dot{x}^*(t) = D_p H(x^*(t), p^*(t), \alpha^*(t)), \tag{1.65}$$

$$\dot{p}^*(t) = -D_x H(x^*(t), p^*(t), \alpha^*(t)), \tag{1.66}$$

*and, for a.e. $0 \le t \le T$*

$$H(x^*(t), p^*(t), \alpha^*(t)) = \max_{a \in A} H(x^*(t), p^*(t), a). \tag{1.67}$$

*Moreover, the map $t \to H(x^*(t), p^*(t), \alpha^*(t))$ is constant, and we have the terminal condition $p^*(T) = \nabla g(x^*(T))$.*

We can now show the link between optimal control and the Hamilton-Jacobi-Bellman equation. First, let us parametrize (1.63) by the initial time $t$ and the initial condition $x$. Given $(t, x) \in [0, T] \times \mathbb{R}^d$, define $x$ as the solution to

$$\begin{cases} \dot{x}(s) = \mu(x(s), \alpha(s)) \quad \text{if } t \le s \le T, \\ x(t) = x. \end{cases} \tag{1.68}$$

We can define the payoff for this problem as

$$P_{t,x}(\alpha) = \int_t^T r(x(s), \alpha(s))\mathrm{d}ts + g(x(T)). \tag{1.69}$$

**Definition 43.** *For $(t,x) \in [0,T] \times \mathbb{R}^d$ the* value function $v(t,x)$ *is the greatest payoff possible starting from $x \in \mathbb{R}^d$ at time $t$, i.e.*

$$v(t,x) = \sup_{\alpha \in \mathcal{A}} P_{t,x}(\alpha), \tag{1.70}$$

*where $\mathcal{A}$ denotes the set of measurable control functions.*

Notice that, by definition, $v(T,x) = g(x)$ for $x \in \mathbb{R}^d$.

**Theorem 44.** *If $v \in C^1([0,T] \times \mathbb{R}^d)$ then $v$ solves the nonlinear partial differential equation*

$$\begin{cases} \partial_t v(t,x) + \max_{a \in A}\{\langle \mu(x,a), Dv(t,x)\rangle + r(x,a)\} = 0 & \text{if } (t,x) \in [0,T] \times \mathbb{R}^d, \\ v(T,x) = g(x) & \text{for } x \in \mathbb{R}^d. \end{cases} \tag{1.71}$$

To design optimal controls in feedback form it is possible to use the *dynamic programming method*: first, find the value function $v$ as a solution to the Hamilton-Jacobi-Bellman equation, then construct $\alpha^*$ as follows: select for each $(t,x) \in [0,T] \times \mathbb{R}^d$ an $\alpha(t,x)$ such that

$$\partial_t v(t,x) + \langle \mu(x, \alpha(t,x)), Dv(t,x)\rangle + r(x, \alpha(t,x)) = 0.$$

Next, if possible, solve the ODE

$$\begin{cases} \dot{x}^*(s) = \mu(x^*(s), \alpha(x^*(s), s)) & \text{if } t \le s \le T, \\ x(t) = x. \end{cases}$$

Then $[0,T] \ni s \to \alpha(x^*(s), s) \in A$ is optimal.

A similar theory can be developed for controlled Stochastic Differential Equations. Consider the SDE

$$\begin{cases} \mathrm{d}X(s) = \mu(X(s), \alpha(s))\mathrm{d}s + \sigma \mathrm{d}W(s) & \text{if } t \le s \le T, \\ X(t) = x, \end{cases} \tag{SDE}$$

where , $\sigma > 0$ is constant, $W$ is a $d$-dimensional Brownian motion on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and $\alpha$ is a control process adapted with respect to the Brownian filtration. The previous equation means that

$$X(s) = x + \int_t^s \mu(X(r), \alpha(r))\mathrm{d}r + \sigma[W(s) - W(t)], \quad \text{for } t \le s \le T.$$

The analogue of (1.69) in the stochastic setting is the *expected payoff functional*

$$P_{t,x}(\alpha) = \mathbb{E}\left[\int_t^T r(X(s), \alpha(s))\mathrm{d}s + g(X(T))\right]. \tag{1.72}$$

For each $(t,x) \in [0,T] \times \mathbb{R}^d$, $v(t,x)$ is defined as in (1.70). The following result relates the second order Hamilton-Jacobi-Bellman equation with the stochastic optimal control theory.

**Theorem 45.** *If the value function* $[0,T]\times\mathbb{R}^d \ni (t,x) \to v(t,x) \in \mathbb{R}$ *is regular enough, then it solves the HJB equation*

$$\begin{cases} \partial_t v(t,x) + \max\limits_{a\in A}\{\langle \mu(x,a), Dv(t,x)\rangle + r(x,a)\} + \frac{\sigma^2}{2}\Delta v(t,x) = 0, \\ v(T,x) = g(x), \end{cases} \tag{1.73}$$

*for* $t \in [0,T)$ *and* $x \in \mathbb{R}^d$.

If the value function is sufficiently regular, then an optimal feedback control can be constructed in a similar manner as in the deterministic case.

## 1.3 Fokker-Planck equations and Mean Field Games

Let $\mu : [0,T]\times\mathbb{R}^d \to \mathbb{R}$ be a given vector field. The *Fokker-Planck equation* is

$$\begin{cases} \partial_t m(t,x) - \frac{\sigma^2}{2}\Delta m(t,x) - \operatorname{div}(m(t,x)\mu(t,x)) = 0 & \text{for } (t,x)\in(0,T)\times\mathbb{R}^d, \\ m(0,x) = m_0(x) & \text{for } x\in\mathbb{R}^d, \end{cases} \tag{1.74}$$

were $\sigma > 0$, the vector field $\mu$ is continuous, uniformly Lipschitz with respect to $x \in \mathbb{R}^d$ and bounded. The main properties of solutions to (1.74) are (see [91]):

- non negativity: if $m_0 \geq 0$ for all $x \in \mathbb{R}^d$, then $m(t,\cdot) \geq 0$ for all $t \in (0,T)$.
- Mass conservation: $\int_{\mathbb{R}^d} m(t,x)\mathrm{d}x = \int_{\mathbb{R}^d} m_0(x)\mathrm{d}x$ for all $t \in (0,T)$.
- Existence of a steady state $m$:

$$-\frac{\sigma^2}{2}\Delta m(x) - \operatorname{div}(m(x)\mu(t)) = 0 \quad \text{for all } x\in\mathbb{R}^d$$

  with $m(x) > 0$.

It is possible to prove existence and uniqueness of a regular solution to (1.74).

**Theorem 46.** *Suppose that* $\mu : [0,T]\times\mathbb{R}^d \to \mathbb{R}^d$ *is bounded, with bounded continuous spatial derivatives and Hölder continuous of exponent* $\alpha$ *in* $x$ *uniformly with respect to* $t$. *Then there exists a nonnegative function* $G$ *such that*

$$G(t,x,s,y) \leq C_1(t-s)^{-d/2}\exp\left(-\frac{C_2\,|x-y|^2}{4(t-s)}\right), \quad \text{for } t,s\in(0,T),\ x,y\in\mathbb{R}^d,$$

*with* $C_1, C_2 > 0$, *and for any probability measure* $\nu_0$ *the formula*

$$m(t,x) = \int_{\mathbb{R}^d} G(t,x,0,y)\nu_0\mathrm{d}y$$

*defines the unique solution in* $C^{1,2}((0,T)\times\mathbb{R}^d)\cap C([0,T)\times\mathbb{R}^d)$ *to* (1.74).

The proof of the existence can be found in [13, Chapter 6], while we refer to [13, Chapter 9] for uniqueness.

### 1.3.1 Representation formula for the Fokker-Planck equation

Solutions to (1.74) are closely related to the solution of the following SDE

$$\begin{cases} \mathrm{d}X_t = \mu(t, X_t)\mathrm{d}t + \sigma \mathrm{d}W_t, \\ X_0 = x, \end{cases} \tag{1.75}$$

where $\mu : [0,T] \times \mathbb{R}^d \to \mathbb{R}^d$ is bounded and Lipschitz. We define

$$\mathcal{L}_t := \sum_i \mu_i(t,\cdot)\frac{\partial}{\partial x_i} + \frac{\sigma^2}{2}\sum_i \frac{\partial^2}{\partial x_i \partial x_i},$$

so that (1.74) can be written as

$$\partial_t m_t = \mathcal{L}_t^* m_t,$$

where $\mathcal{L}_t^*$ denotes the formal adjoint operator of $\mathcal{L}_t$ in $L^2(\mathbb{R}^d)$. As in [56], using Itô's formula it is possible to prove that, if $X \in L^2([0,T] \times \mathcal{O} \times \Omega)$ is a family of solutions of (1.75) and $X_0$ is distributed as $m_0$, then the measure $m_t$ defined by

$$(\forall f \in C_0(\mathbb{R}^d)) \quad \int_{\mathbb{R}}^d f(x)\mathrm{d}m_t(x) = \int_{\mathbb{R}}^d \mathbb{E}\left[f\left(X(t,x,\omega)\right)\right]\mathrm{d}m_0(x),$$

is absolutely continuous with respect to the Lebesgue measure and its density, also denoted $m$, is such that $[0,T] \times \mathbb{R}^d \ni (t,x) \to m(t,x) \in \mathbb{R}$ solves (1.74). In order to better clarify the link between (1.75) and (1.74) we first need the definition of *martingale solutions* of (1.75).

**Definition 47.** *A measure $\nu_{x,s}$ on $C([0,T];\mathbb{R}^d)$ is a* martingale solution *of (1.75) starting from $x$ at time $s$ if:*

(i) $\nu_{x,s}(\{f \in C([0,T];\mathbb{R}^d))|f(s) = x\}) = 1$.

(ii) *For any $\varphi \in C_c^\infty(\mathbb{R}^d)$, the stochastic process on $C([0,T];\mathbb{R}^d)$*

$$\varphi(f(t)) - \int_s^t (\mathcal{L}_\tau \varphi)(f(u))\mathrm{d}\tau$$

*is a $\nu_{x,s}$-martingale after time $s$.*

A martingale problem is well-posed if, for any $(s,x) \in \mathbb{R}^d$ we have existence and uniqueness of martingale solutions. Moreover, the existence and uniqueness of martingale solutions for equation (1.75) is linked to the existence of solutions to the Fokker-Planck equation. We recall the definition of measurable families of probability measures.

**Definition 48.** *A family of probability measures $\{\nu_x\}_{x\in\mathbb{R}^d}$ on a probability space $(\Omega, \mathcal{F})$ is measurable if, for every $A \in \mathcal{F}$, the real-valued map $x \to \nu_x(A)$ is measurable.*

**Lemma 49.** *Let $\mu$ be bounded and $A \subset \mathbb{R}^d$ be a Borel set. The following properties are equivalent:*

(a) *time marginals of martingale solutions of equation* (1.75) *are unique for any* $x \in A$.

(b) *Finite non-negative measure-valued solutions of equation* (1.74) *are unique for any non-negative Radon measure* $m_0$ *concentrated in* $A$.

If $\nu_x$ is a martingale solution of (1.75) starting from $x$ at time $t = 0$, for $m_0$-a.e. $x$ it is possible to give a representation formula for a non-negative solution of (1.74).

**Lemma 50.** *Let* $m_0$ *be a locally finite measure on* $\mathbb{R}^d$, *and let* $\{\nu_x\}_{x\in\mathbb{R}^d}$ *be a measurable family of probability measures on* $C([0,T];\mathbb{R}^d)$ *such that* $\nu_x$ *is a martingale solution to* (1.75) *starting from* $x$ *at time* 0, *for* $|m_0|$*-a.e.* $x$. *Define on* $C([0,T];\mathbb{R}^d)$ *the measure* $\nu := \int_{\mathbb{R}^d} \nu_x \mathrm{d}m_0(x)$ *and assume that*

$$\int_0^T \int_{\mathbb{R}^d \times C([0,T];\mathbb{R}^d)} \chi_{B_R(0)}(f(t)) \mathrm{d}\nu_x \mathrm{d}\,|m_0|\,(x)\mathrm{d}t < +\infty$$

*for all* $R > 0$. *Then the measure* $m_t^\nu$ *on* $\mathbb{R}^d$ *defined by*

$$\langle m_t^\nu, \varphi\rangle := \int_{\mathbb{R}^d \times C([0,T];\mathbb{R}^d)} \varphi(f(t)) \mathrm{d}\nu_x(f) \mathrm{d}m_0(x)$$

*for every* $\varphi \in C_c^\infty(\mathbb{R}^d)$ *solves* (1.74).

Proofs of Lemma 49 and 50 can be found in [56].

### 1.3.2 Second order MFG system

Let us analyze a control problem with infinitely many agents. The distribution of the agents is given by the function $m$ and each agent controls its own dynamic, denoted by $X_s(x)$ and defined as the solution to (SDE). At this stage $m$ is given, meaning that it is the anticipation made by the agents on their future evolution, which depends on the distribution $m$ itself. Let $\mu$ be smooth enough for the solution $(X_t)$ to exist and let $\mathcal{P}_1(\mathbb{R}^d)$ (respectively $\mathcal{P}_2(\mathbb{R}^d)$) be the space of probability measures on $\mathbb{R}^d$ with first (respectively second) bounded moments. The cost of a single player is given by

$$J(t,x,\alpha) = \mathbb{E}\left[\int_t^T \left(L(s,X_s,\alpha_s) + f(X_s,m(s))\right)\mathrm{d}s + g(X_T,m(T))\right], \quad (1.76)$$

where $T > 0$ , $L : [0,T] \times \mathbb{R}^d \times A \to \mathbb{R}$, $f : \mathbb{R}^d \times \mathcal{P}_1(\mathbb{R}^d) \to \mathbb{R}$, and $g : \mathbb{R}^d \times \mathcal{P}_1(\mathbb{R}^d) \to \mathbb{R}$ are given and continuous. We define the value function

$$u(t,x) = \inf_{\alpha\in A} J(t,x,\alpha), \quad (1.77)$$

which is the solution to the Hamilton-Jacobi-Bellman equation

$$\begin{cases} -\partial_t u + H(t,x,Du) - \frac{\sigma^2}{2}\Delta u = f(x,m(t)) & \text{in } (0,T)\times\mathbb{R}^d, \\ u(T,x) = g(x,m(T)) & \text{in } \mathbb{R}^d, \end{cases} \quad (1.78)$$

with Hamiltonian

$$H(t,x,p) = \sup_{\alpha\in A}\left\{-L(t,x,\alpha) - \langle p, \mu(t,x,\alpha)\rangle\right\}. \tag{1.79}$$

Let us introduce $\alpha^*(t,x) \in A$ as a maximum point in (1.79) when $p = Du(t,x)$. It results that $\mu(t,x,\alpha^*) = -\partial_p H(t,x,Du(t,x))$.

Let us now discuss the evolution of the population density $m$, firstly making the assumptions that all the agents control the same dynamic $X_s$ (with different starting points) and minimize the same cost $J$. It results that optimal dynamics of each player is given by

$$\mathrm{d}X_s^* = \mu(s, X_s^*, \alpha^*(s, X_s^*))\mathrm{d}s + \sigma \mathrm{d}W_s. \tag{1.80}$$

The initial distribution at time $t = 0$ is $\overline{m}_0 \in \mathcal{P}_1(\mathbb{R}^d)$ and the distribution of agents at time $t$ is given by the law of $(X_s^*)$, with $\mathrm{Law}(X_0^*) = \overline{m}_0$. Starting from (1.80), using Itô's formula and integrating by parts, we obtain that $\overline{m}$ satisfies, in the sense of distributions,

$$\begin{cases} \partial_t \overline{m} - \frac{\sigma^2}{2}\Delta \overline{m} - \mathrm{div}(\overline{m}\mu(t,x,\alpha^*)) = 0 & \text{in } (0,T)\times\mathbb{R}^d, \\ \overline{m}(0,x) = \overline{m}_0(x) & \text{for } x \in \mathbb{R}^d, \end{cases} \tag{1.81}$$

At the equilibrium we expect $\overline{m}(t,x) = m(t,x)$, so we get our MFG system

$$\begin{cases} -\partial_t u - \frac{\sigma^s}{2}\Delta u + \frac{1}{2}|Du|^2 = f(x, m(t)) & \text{in } (0,T)\times\mathbb{R}^d, \\ \partial_t m - \frac{\sigma^2}{2}\Delta m - \mathrm{div}(m(t,x)\partial_p H(t,x,Du(t,x))) = 0 & \text{in } (0,T)\times\mathbb{R}^d, \\ m(0,x) = m_0(x), \quad u(T,x) = g(x, m(T)) & \text{in } \mathbb{R}^d. \end{cases} \tag{1.82}$$

Under suitable assumptions it is possible to prove the existence of a classical solution for (1.82).

**Definition 51.** *A pair $(u,m)$ is a classical solution to* (1.82) *if $u, m \in C^{1,2}([0,T]\times\mathbb{R}^d)$ and $(u,m)$ satisfies* (1.82) *in the classical sense.*

**Definition 52.** *The* Wasserstein distance *between two probability measures $m_1, m_2 \in \mathcal{P}_1(\mathbb{R}^d)$ is defined as*

$$\mathbf{d_1}(m_1,m_2) = \inf_{\gamma\in\Pi(m_1,m_2)}\left\{\int_{\mathbb{R}^d\times\mathbb{R}^d}|x-y|\mathrm{d}\gamma(x,y)\right\},$$

*where $\Pi(m_1,m_2)$ denotes the collection of all the measures on $\mathbb{R}^d\times\mathbb{R}^d$ with marginals $m_1$ and $m_2$.*

We present an existence result for (1.82) (a proof can be found in [2]).

**Theorem 53.** *Let us assume that:*

- *$f$ and $g$ are uniformly bounded by some positive constant $C_0$ over $\mathbb{R}^d\times\mathcal{P}_1$.*
- *$f$ and $g$ are Lipschitz continuous, i.e. for each $(x_1,m_1),(x_2,m_2)\in\mathbb{R}^d\times\mathcal{P}_1$*

$$|f(x_1,m_1) - f(x_2,m_2)| \le (|x_1-x_2| + \mathbf{d_1}(m_1,m_2))$$

*and*

$$|g(x_1,m_1) - g(x_2,m_2)| \le (|x_1-x_2| + \mathbf{d_1}(m_1,m_2)).$$

- *The probability measure $m_0$ is absolutely continuous with respect to the Lebesgue measure and has a $C_{2+\alpha}$ density, still denoted by $m_0$, such that $\int_{\mathbb{R}^d} |x|^2 m_0(x)\mathrm{d}x < +\infty$.*

*Then there is at least one classical solution to* (1.82).

We conclude the section with a uniqueness result, proven in [77].

**Theorem 54.** *Assume that either $f$ and $g$ are monotone in $L^2(\mathbb{R}^d \times (0,T))$ and $H(t,x,\cdot)$ is strictly convex, or $f$ and $g$ are strictly monotone, i.e.*

$$\int_{\mathbb{R}^d} (f(x, m_1(t,x)) - f(x, m_2(t,x)))\,(m_1(t,x) - m_2(t,x))\mathrm{d}x \leq 0 \Rightarrow m_1 = m_2,$$

*for $t \in [0,T)$ and $m_1, m_2 \in \mathcal{P}_1$, and*

$$\int_{\mathbb{R}^d} (g(x, m_1(T,x)) - g(x, m_2(T,x)))\,(m_1(T,x) - m_2(T,x))\mathrm{d}x \leq 0 \Rightarrow m_1 = m_2,$$

*for $m_1, m_2 \in \mathcal{P}_1$. Then the uniqueness of solutions to* (1.82) *holds.*

# Chapter 2

# Second order fully semi-Lagrangian discretizations of advection–diffusion–reaction systems

We propose a second order, fully semi-Lagrangian method for the numerical solution of systems of advection–diffusion–reaction equations, which is based on a semi-Lagrangian approach to approximate in time both the advective and the diffusive terms. The proposed method allows to use large time steps, while avoiding the solution of large linear systems, which would be required by implicit time discretization techniques. Standard interpolation procedures are used for the space discretization on structured and unstructured meshes. A novel extrapolation technique is proposed to enforce second-order accurate Dirichlet boundary conditions. We include a theoretical analysis of the scheme, along with numerical experiments which demonstrate the effectiveness of the proposed approach and its superior efficiency with respect to more conventional explicit and implicit time discretizations.

In the present work, we present a number of improvements to the fully SL approach of [18], [19]. In particular, we show how second order accuracy in time can be achieved. An improved treatment of Dirichlet boundary conditions is also discussed and analysed. The resulting approach yields an efficient combination, which is validated on a number of classical benchmarks, on both structured and unstructured meshes. Numerical results show that the method yields good quantitative agreement with reference numerical solutions, while being superior in efficiency to standard implicit methods and to approaches in which the SL method is only used for the advection term.

## 2.1 Semi-Lagrangian schemes for linear parabolic equations

Numerically, a semi-Lagrangian method mimics the method of characteristics, tracking the foot of the characteristic passing through every node and following it. What is needed is, basically, a technique for solving SDEs and track the characteristics, then a reconstruction technique, such as an interpoplation operator, to recover pointwise values of the numerical solution. We present a first and a second order technique to approximate the solution to the Cauchy problem (1.20). First of all, to sketch the ideas behind the method, we consider the problem

$$\begin{cases} \partial_t u(t,x) + \langle Du(t,x), \mu(t,x)\rangle - \frac{\sigma^2}{2}\Delta u(t,x) + f(t,x) = 0 & (t,x) \in (0,T]\times\mathbb{R}, \\ u(0,x) = u_0(x), & x \in \mathbb{R}. \end{cases} \tag{2.1}$$

According to Theorem 13, the solution to this problem is given by the following representation formula

$$u(t,x) = \mathbb{E}\left[\int_0^t f(s, X(s;t,x))\mathrm{d}s + u_0(X(0;t,x))\right] \tag{2.2}$$

with characteristics solving

$$\begin{cases} \mathrm{d}X(s) = \mu \mathrm{d}s + \sigma \mathrm{d}W(s), \quad s \in [t,T], \\ X(0) = x. \end{cases} \tag{2.3}$$

Let us discretize the time interval $[0,T]$ using a step $\Delta t > 0$, define $N_{\Delta t} = \lfloor T/\Delta t \rfloor$, the sets $\mathcal{I}_{\Delta t} = \{0,\dots,N_{\Delta t}\}$ and $\mathcal{I}^*_{\Delta t} = \mathcal{I}_{\Delta t} \setminus \{N_{\Delta t}\}$, so that $t_k = k\Delta t$ for $k \in \mathcal{I}_{\Delta t}$. We show how to construct a SL approximation using the technique shown in [52] and [49]. First of all, an approximation of the solution of (2.3) is required, using a stochastic scheme for SDEs. Using the notation $y_k$ for the numerical approximation of $X_k$, we have that for each $k = 0,\dots,N_{\Delta t}-1$ the time-discrete approximation of the characteristic can be written as

$$\begin{cases} y_{k+1} = y_k + \rho(y_k, y_{k+1}, \Delta W_k), \\ y_0 = x, \end{cases} \tag{2.4}$$

where $\rho$ is defined according to the stochastic method implied for the approximation of (2.3). For example, using the stochastic forward Euler method, we have

$$\rho(x,y,\Delta W) = -\Delta t \mu(t,x) + \sigma \Delta W$$

wich is has first order of accuracy in time, stochastic Heun method

$$\rho(x,y,\Delta W) = -\frac{\Delta t}{2}(\mu(t,x) + \mu(t, x - \Delta t\mu(t,x) + \sigma\Delta W)) + \sigma\Delta W,$$

or stochastic Crank-Nicolson method

$$\rho(x,y,\Delta W) = -\frac{\Delta t}{2}(\mu(t,x) + \mu(t,y)) + \sigma\Delta W$$

which are second order accurate in time. $\Delta W_\ell$ is a Gaussian variable with zero mean and variance equal to $\Delta t$. In practice, the expectation with respect to $\Delta W$ is approximated by working on a finite number $s \in \mathbb{N}$ of realizations $\Delta_\ell$, each one associated to a weight $\alpha_\ell$, such that the discrete probability density is

$$\mathbb{P}\left(\Delta W_\ell = \Delta_\ell\right) = \alpha_\ell, \quad \ell = 1, \dots, s,$$

with conditions

$$\alpha_\ell \geq 0, \quad \sum_{\ell=1}^{s} \alpha_\ell = 1$$

and assumptions

$$\sum_{\ell=1}^{s} \alpha_\ell \Delta_\ell = 0, \quad \sum_{\ell=1}^{s} \alpha_\ell \Delta_\ell^2 = \Delta t,$$

since the continuous variable has mean 0 and variance $\Delta t$. This means that the expected value in (2.2), using the above discrete approximation of $\Delta W$, becomes a weighted average of the argument of $\mathbb{E}$ evaluated in the realizations of $y_k$. Using the notation $\overline{F}(x, \Delta_i)$ for the approximation of the integral of $f$ using a quadrature rule, we obtain the time discrete approximation scheme for (2.1)

$$\begin{cases} u_{k+1}(x) = \sum_{\ell=1}^{s} \alpha_\ell \left( u_k(x, \Delta_i)) + \overline{F}(x, \Delta_i) \right), & k \in \mathcal{I}^*_{\Delta t}, \\ u^0(x) = u_0(x). \end{cases} \tag{2.5}$$

where $u_k(x)$ is the time-discrete approximated solution in $x$ at time $t_k = k\Delta t$.

In order to obtain a fully discrete scheme, let us now set a grid in the computational domain: we discretize $\mathbb{R}^d$ using a space step $\Delta x > 0$, so that we get a set of nodes $x_j = j\Delta x$ for any multiindex $j \in \mathbb{Z}^d$. Given a function $g(t, x)$ we will denote by $g_{k,j}$ the approximation of $g(t_k, x_j)$ and by $g_k$ the set of nodal values at time $t_k$. Analogously, given a function $h(x)$, we will denote by $h_j$ the approximation of $h(x_j)$ and by $h$ the set of its nodal values. Let $I[\cdot]$ be a polynomial interpolation operator such that $I[v](x_j) = v_j$, and if $v \in W^{q,\infty}$, then for any $x \in \mathbb{R}^d$ we have

$$|I[v](x) - v(x)| \leq C(\Delta x)^q. \tag{2.6}$$

The fully discrete semi-Lagrangian scheme for (2.1) is

$$\begin{cases} u_{k+1,j} = \sum_{\ell=1}^{s} \alpha_\ell \left( I[u_k](x_j, \Delta_i)) + \overline{F}(x_j, \Delta_i) \right), & k \in \mathcal{I}^*_{\Delta t}, j \in \mathbb{Z}^d \\ u_{0,j} = u_0(x_j), & j \in \mathbb{Z}^d \end{cases} \tag{2.7}$$

or, compactly,

$$u_{k+1} = S_\Delta(u_k). \tag{2.8}$$

As shown in [52] it is possible to carry on a rigorous convergence analysis of such schemes, performed in normalized Hölder norms:

$$\|v\|_p := \begin{cases} \left( (\Delta x)^d \sum_j |v_j|^p \right)^{1/p}, & \text{if } p < \infty, \\ \max_j |v_j|, & \text{if } p = \infty. \end{cases} \tag{2.9}$$

The scheme in (2.7) is stable if each addendum in the right hand side is stable.

**Theorem 55.** *Let*

$$v^i_{k+1,j} = I[v_k](x_j, \Delta_i) + \overline{F}(x_j, \Delta_i) \tag{2.10}$$

*for* $i = 1, \dots, s$. *If*

$$\|v^i_{k+1}\|_p \leq (1 + C\Delta t)\|v^i_k\|_p \tag{2.11}$$

*for a positive constant* $C$ *independent of* $\Delta t, \Delta x$, *then*

$$\|u_{k+1}\|_p \leq (1 + C\Delta t)\|u_k\|_p \tag{2.12}$$

*with* $u_k$ *solution of* (2.7)

**Theorem 56.** *Let* $\mu$ *be a smooth vector field and let* $u(\cdot, \cdot)$ *be a smooth solution to* (2.1). *Assume also that* (2.7) *holds. Then, the local truncation error satisfies the bound*

$$\frac{1}{\Delta t}\|u(t_{k+1}) - S_\Delta(u(t_k))\|_p \leq C\left((\Delta t)^{q'} + \frac{(\Delta x)^q}{\Delta t}\right)$$

*where* $q'$ *is the order of accuracy of the stochastic method used for the approximation in* (2.4),

In conclusion, the following convergence result holds.

**Theorem 57.** *For* $i = 1, \dots, s$ *let* (2.10) *satisfy* (2.11). *Then, for any* $k \in \mathcal{I}_{\Delta t}$,

$$\|u_k - u(t_k)\|_p \to 0$$

*for* $\Delta t \to 0, \Delta x = o((\Delta t)^{1/q})$. *Moreover, if the order of accuracy of the stochastic method used for the approximation in* (2.4) *is* $q'$, *and if for any* $t \in [0, T]$, $u(t, \cdot) \in C^q(\mathbb{R}^d)$, *then*

$$\|u_k - u(t_k)\|_p \leq C\left((\Delta t)^{q'} + \frac{(\Delta x)^q}{\Delta t}\right). \tag{2.13}$$

## 2.2 The model problem

We consider as a model problem the advection–diffusion–reaction equation with Dirichlet boundary conditions

$$\begin{cases} \partial_t u + \langle \mu, Du \rangle - \frac{\sigma^2}{2}\Delta u = f(u) & (t, x) \in (0, T] \times \mathcal{O}, \\ u(t, x) = b(t, x) & (t, x) \in (0, T] \times \partial\mathcal{O}, \\ u(0, x) = u_0(x) & x \in \mathcal{O}. \end{cases} \tag{2.14}$$

Here, $T$ denotes the final time, $\mathcal{O} \subset \mathbb{R}^2$ is an open bounded domain, $\mu : \mathcal{O} \times [0, T] \to \mathbb{R}^2$ is a velocity field and $b : \partial\mathcal{O} \times [0, T] \to \mathbb{R}$ denotes the boundary value of the species $u$. The unknown $u : \mathcal{O} \times [0, T] \to \mathbb{R}$ can be interpreted as the concentration of a chemical species that is transported through the domain $\mathcal{O}$ by the advection and diffusion processes, while undergoing locally a nonlinear evolution determined by the source term $f(u)$, which will be assumed to be globally Lipschitz continuous, with Lipschitz constant $L_f$.

In the simpler case of homogeneous boundary conditions and time independent advection field and diffusion coefficient, equation (2.14) can be written as

$$\partial_t u = \mathcal{L}u + f(u), \tag{2.15}$$

where $\mathcal{L}$ denotes a linear differential operator. We denote by $\mathcal{E}_t$ the evolution operator determining the solution to the associated homogeneous equation

$$\partial_t \tilde{u} = \mathcal{L}\tilde{u} \tag{2.16}$$

with the same initial datum $\tilde{u}(0,x) = u_0(x)$ and boundary conditions as in (2.14), so that $\tilde{u}(t) = \mathcal{E}_t[u_0]$. By formal application of the variation of constants formula, the solution of (2.15) can then be represented as

$$u(t,x) = \mathcal{E}_t[u_0](x) + \int_0^t \mathcal{E}_{t-s}[f \circ u](x)\ \mathrm{d}s. \tag{2.17}$$

If discrete time levels $t_k, n = 0, \ldots, N$ are introduced, so that $t_k = n\Delta t$ and $\Delta t = T/N$, the same representation formula on the interval $[t_k, t_{k+1}]$ reads

$$u(t_{k+1},x) = \mathcal{E}_{\Delta t}[u(t_k,\cdot)](x) + \int_{t_k}^{t_{k+1}} \mathcal{E}_{t_{k+1}-s}[f \circ u](x)\ \mathrm{d}s. \tag{2.18}$$

The construction of the scheme relies on the application to (2.18) of the Feynman–Kac formula to represent the solution to (2.16) (see, e.g., [52]), so that

$$\mathcal{E}_{\Delta t}[u(t_k,\cdot)](x) = \mathbb{E}\left\{u(y_k, X(t_k))\right\} \tag{2.19}$$

where $\mathbb{E}$ denotes the probabilistic expectation w.r.t. the Wiener measure, and $X(t)$ is the solution of the stochastic differential equation (SDE):

$$\begin{cases} \mathrm{d}X = -\mu(s, X(s))\mathrm{d}s + \sigma \mathrm{d}W(s), \\ X(t_{k+1}) = x, \end{cases} \tag{2.20}$$

for $s \in [t_k, t_{k+1}]$, with $W(s)$ denoting a standard 2-dimensional Wiener process. Note that, for $\sigma = 0$, (2.20) reduces to a deterministic ODE and the evolution operator (2.19) can be approximated accordingly by the well-known method of characteristics for transport problems.

While the proposed numerical method will be presented in this simpler case, the target for more realistic applications are systems of coupled advection–diffusion–reaction equations of the form

$$\begin{cases} \partial_t u_n + \langle \mu, Du_n \rangle - \nabla \cdot (ADu_n) = f_n(u_1, \ldots, u_S) & (t,x) \in \mathcal{O} \times (0,T], \\ u_n(t,x) = b_n(t,x) & (t,x) \in \partial\mathcal{O} \times (0,T], \\ u_n(0,x) = u_{0,n}(x) & x \in \mathcal{O}, n = 1, \ldots, S. \end{cases} \tag{2.21}$$

Here, $A$ denotes a symmetric and positive semi-definite diffusivity tensor, possibly dependent on space and time. As remarked in the Introduction, systems of this kind, with a possibly large number of species $S$, are responsible for the largest share of the computational cost of typical environmental fluid dynamics models,

so that even minor increases in the efficiency of the discretization for this very classical problem are of great practical relevance.

Notice that, while the use of a representation formula like (2.17) may recall the procedure that it is followed to introduce exponential integrators (EI) (see e.g. the review in [66]), there are substantial differences between SL and EI methods. For example, EI are based on the approximation of a representation formula for the solutions of a spatially discretized problem, while SL methods employ a space-time representation formula. Furthermore, SL methods approximate the evolution operator by a local approach based on trajectory computation, while standard EI entail a global step for the computation of the matrix exponential, which is computationally quite demanding, see e.g. the discussion in [60].

## 2.3 Fully semi-Lagrangian methods

A numerical method for the solution of equation (2.14) on the interval $[t_k, t_{k+1}]$ can then be derived heuristically from (2.18) by discretizing the time integral using the trapezoidal rule, so that one obtains

$$u(t_{k+1}, x) \approx \mathcal{E}_{\Delta t}[u(t_k, \cdot)](x) + \frac{\Delta t}{2} \left[\mathcal{E}_{\Delta t}[f \circ u](x) + f(u(t_{k+1}, x))\right]. \tag{2.22}$$

If the diffusion term is dropped in equation (2.14), and the evolution operator is approximated by a numerical version of the flow streamline together with an interpolation at the departure point of the streamline, a numerical method based on (2.22) can be interpreted as a semi-Lagrangian extension of the trapezoidal rule with global truncation error of second order. Semi-Lagrangian methods based on this formula have been successfully used in a large number of applications (see, among many others, [15],[37],[41],[103],[104],[106]). Due to a possible stiffness of the reaction term, we might rather use a first order, off-centered version of the above formula, defined, for $\theta \in [1/2, 1]$ as

$$u(t_{k+1}, x) \approx \mathcal{E}_{\Delta t}[u(t_k, \cdot)](x) + (1-\theta)\Delta t \mathcal{E}_{\Delta t}[f \circ u](x) + \theta \Delta t f(u(t_{k+1}, x)). \tag{2.23}$$

In order to discretize (2.22) (or (2.23)), we introduce a space mesh $\mathcal{G}_{\Delta x} = \{x_i : x_i \in \mathcal{O}\}$, where $\Delta x$ denotes the mesh resolution. The mesh can be structured as well as unstructured; the only necessary restriction is that it should be possible to define a piecewise polynomial interpolation operator $I$ of degree $q$, constructed on the values of a grid function $c$ defined on $\mathcal{G}_{\Delta x}$ (we refer to [96] for a precise definition of the general setting). We denote by $I[u_k](x)$ the value at $x$ of the interpolant $I$ computed using the values of the grid function $u_k$. The vector $u_k$ collects the values $u_{k,i}$ of the numerical solution to (2.14) at the space-time mesh nodes $(t_k, x_i)$.

The discretization of (2.20), whenever aimed at approximating the expectation in (2.19), is performed via the so-called *weak schemes* for SDEs (see the classical review [70]). At a generic node $x = x_i$, weak schemes approximate the expectation in (2.19) as

$$\mathbb{E}\left\{u(t_k, X(t_k))\right\} = \sum_{\ell} \alpha_\ell u(t_k, y^\ell_{k,i}) + O(\Delta t^{q'}) \tag{2.24}$$

for a suitable definition of the points $y_{k,i}^{\ell}$ of the weights $\alpha_\ell$. For our purposes, we will consider cases in which $q' = 1, 2$, and set

$$y_{k,i}^{\ell} = x_i + \delta_{k,i}^{\ell}.$$

In the simplest case, a two-dimensional, first-order weak scheme ($q' = 1$) which generalizes the explicit Euler scheme, may be obtained for

$$\delta_{k,i}^{\ell} = -\Delta t \mu(t_{k+1}, x_i) + \sqrt{2\Delta t}\sigma e^{\ell}$$

for $\ell = 1, \ldots, 4$, with $\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = 1/4$, and

$$e^1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \quad e^2 = -\begin{pmatrix} 1 \\ 0 \end{pmatrix}, \quad e^3 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \quad e^4 = -\begin{pmatrix} 0 \\ 1 \end{pmatrix}.$$

The discrete set of displacements $\sqrt{2\Delta t}\sigma e^{\ell}$ and weights $\alpha_\ell$ ($k = 1, \ldots, 4$) is constructed (see [70]) in order to approximate the probability density of the 2-dimensional Gaussian random variable

$$\sigma \Delta W := \sigma \left( W(\Delta t) - W(0) \right)$$

with the discrete density

$$P\left( \sigma \Delta W = \sqrt{2\Delta t}\sigma e^{\ell} \right) = \alpha_\ell, \quad (\ell = 1, \ldots, 4)$$

up to a certain number of moments. More precisely, in this first-order case they coincide up to the third moment (note that odd moments are always zero by symmetry).

Introducing the space interpolation, and replacing (2.19) with its discretization (2.24), a first order in time approximation $u_k$ of the solution to (2.14) can then be defined as

$$\begin{aligned} u_{k+1,i} &= \frac{1}{4} \sum_{\ell=1}^{4} I[u_k]\left( y_{k,i}^{\ell} \right) \\ &\quad + (1-\theta)\Delta t \frac{1}{4} \sum_{\ell=1}^{4} f(I[u_k])\left( y_{k,i}^{\ell} \right) + \theta \Delta t f(u_{k+1,i}), \end{aligned} \tag{2.25}$$

Notice that, for simplicity, we neglect in (2.25) the treatment of boundary conditions. Possible approaches to handle Dirichlet boundary conditions will be discussed in Section 2.5.

It is easy to show that (2.25) has a unique solution for $\Delta t$ small enough. In fact, (2.25) is in the form of a set of decoupled fixed point equations for the unknowns $u_{k+1,i}$,

$$u_{k+1,i} = F_i(u_k) + \theta \Delta t f(u_{k+1,i}) \tag{2.26}$$

and the Lipschitz constant of the right-hand side is $\theta \Delta t L_f$. Therefore, the right-hand side is a contraction as soon as $\Delta t < \frac{1}{\theta L_f}$, regardless of the Courant number. Moreover, since

$$\begin{aligned} |f(\gamma)| &\leq |f(\gamma) - f(0)| + |f(0)| \\ &\leq L_f |\gamma| + |f(0)| \end{aligned}$$

we can obtain from (2.26)

$$|F_i(u_k) + \theta\Delta t f(\gamma)| \leq |F_i(u_k)| + \theta\Delta t(L_f|\gamma| + |f(0)|)$$

(note that the right-hand side is increasing with $|\gamma|$). In order to obtain an invariant set of the form $|\gamma| \leq R$, we should therefore satisfy the condition

$$|F_i(u_k)| + \theta\Delta t(L_f R + |f(0)|) \leq R,$$

which gives, provided $\theta\Delta t L_f < 1$,

$$R \geq \frac{|F_i(u_k)| + \theta\Delta t|f(0)|}{1 - \theta\Delta t L_f}.$$

Under this condition, the interval $[-R, R]$ is invariant, both assumptions of the Banach fixed point theorem are satisfied, and (2.25) has a unique solution $u_{k+1,i} \in [-R, R]$.

The method (2.25) will be denoted in what follows by SL1. This method inherits the same stability and convergence properties of the parent methods, as it will be discussed in Section 2.4. Notice that this approach can be extended to spatially varying diffusion coefficients and that, while only first order in time, its effective accuracy can be substantially superior to that of more standard techniques, if higher degree interpolation operators are used, as shown in [18].

In order to derive a method of second order in time, we follow the main steps of [52],[87]. Applying the implicit weak method of order 2 defined in [70] for the approximation of the stochastic streamlines (2.20) (see also [87] for a general theory of weak approximation for SDE), we define the points $y^\ell_{k,i}$ as the solutions of the nonlinear equations

$$y^\ell_{k,i} = x_i - \frac{\Delta t}{2}\left(\mu(t_{k+1}, x_i) + \mu(t_k, y^\ell_{k,i})\right) + \sqrt{3\Delta t}\sigma e^\ell. \tag{2.27}$$

Here, the symbols $e^\ell$ denote the vectors:

$$e^1 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \quad e^2 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \quad e^3 = \begin{pmatrix} 0 \\ -1 \end{pmatrix},$$

$$e^4 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \quad e^5 = \begin{pmatrix} -1 \\ 0 \end{pmatrix}, \quad e^6 = \begin{pmatrix} 1 \\ 1 \end{pmatrix},$$

$$e^7 = \begin{pmatrix} 1 \\ -1 \end{pmatrix}, \quad e^8 = \begin{pmatrix} -1 \\ 1 \end{pmatrix}, \quad e^9 = \begin{pmatrix} -1 \\ -1 \end{pmatrix}.$$

Accordingly, the weights $\alpha_\ell$ are given by

$$\alpha_1 = 4/9, \quad \alpha_2 = \alpha_3 = \alpha_4 = \alpha_5 = 1/9, \quad \alpha_6 = \alpha_7 = \alpha_8 = \alpha_9 = 1/36.$$

In this case (see [70]) the increase in the order of approximation requires that moments of the probability density of $\sigma\Delta W$ are reproduced by the discrete density up to the fifth moment. This motivates the introduction of further displacements and weights.

It is to be remarked that also method (2.27) can be rewritten in terms of the displacements $\delta_{k,i}^{\ell} = y_{k,i}^{\ell} - x_i$ as

$$\delta_{k,i}^{\ell} = -\frac{\Delta t}{2}\left(\mu(t_{k+1}, x_i) + \mu(t_k, x_i + \delta_{k,i})\right) + \sqrt{3\Delta t}\sigma e^{\ell}, \tag{2.28}$$

thus yielding an implicit method that is a natural extension to stochastic differential equations of that introduced in [98] and commonly used in meteorological applications for the computation of streamlines in SL methods.

A second order in time SL (SL2) scheme can then be defined by a Crank-Nicolson approach ((2.23) with $\theta = 0.5$) as

$$u_{k+1,i} = \sum_{\ell=1}^{9} \alpha_\ell \left( I[u_k](y_{k,i}^{\ell}) + \frac{\Delta t}{2} f(I[u_k](y_{k,i}^{\ell})) \right) + \frac{\Delta t}{2} f(u_{k+1,i}). \tag{2.29}$$

Solvability of (2.29) with respect to $u_{k+1,i}$ can be proved with the same arguments used for (2.25).

Notice that, with respect to the simpler first-order in time variant (2.25), nine interpolations at the foot of the streamlines must be computed, which clearly makes this approach substantially more expensive. In applications to systems of the form (2.21), the computational cost of scheme (2.29) can be marginally reduced by setting

$$\tilde{u}_{k,i} = \sum_{\ell=1}^{9} \alpha_\ell I[u_k](y_{k,i}^{\ell})$$

and defining

$$u_{k+1,i} = \tilde{u}_{k,i} + \frac{\Delta t}{2} f(\tilde{u}_{k,i}) + \frac{\Delta t}{2} f(u_{k+1,i}), \tag{2.30}$$

so as to reduce the number of the evaluations of a possibly costly nonlinear term. Furthermore, when the coupling of the diffusion and advection term is weak, it should be possible to decouple again the approximation of a single deterministic streamline from that of the diffusive displacements, which could be added at the end of each approximate streamline without increasing too much the error. In particular, in [18],[19] explicit Euler or Heun methods were employed to compute these approximations, coupled to a substepping approach along the lines of [36],[99]. More specifically, given a positive integer $m$, a time substep was defined as $\Delta\tau = \Delta t/m$ and, for $n = 0, \ldots, m-1$, the Euler substepping

$$\begin{cases} \hat{y}_i^{(n+1)} = \hat{y}_i^{(n)} - \Delta\tau u(t_k, \hat{y}_i^{(n)}), \\ \hat{y}_i^{(0)} = x_i \end{cases} \tag{2.31}$$

was computed, so that a $y_{k,i}^{\ell}$ in (2.25) was modified with $y_{k,i}^{\ell} = \hat{y}_i^{m} + \sqrt{2\Delta t}\sigma e^{\ell}$.

A decoupled substepping variant of (2.27) might in turn be obtained by computing, for $n = 0, \ldots, m-1$,

$$\begin{cases} \hat{y}_i^{(n+1)} = \hat{y}_i^{(n)} - \frac{\Delta\tau}{2}\left[\mu(t_{k+1} - n\Delta\tau, \hat{y}_i^{(n)}) + \mu(t_{k+1} - (n+1)\Delta\tau, \hat{y}_i^{(n+1)})\right], \\ \hat{y}_i^{0} = x_i, \end{cases} \tag{2.32}$$

and setting $y^{\ell}_{k,i} = \hat{y}^{(m)}_i + \sqrt{3\Delta t}\sigma e^{\ell}$ in (2.29). We will denote this decoupled variant with substepping by SL2s.

Notice that, in realistic problems, a major shortcoming of scheme (2.29) is the fact that the Crank–Nicolson method, while A-stable, is not L-stable, see e.g. [74]. Therefore, no damping is introduced by the method for very large values of the time step and spurious oscillations may arise, see also the discussion in [21]. In order to reduce the computational cost and to address the L-stability issue, different variants of the scheme (2.29) could also be introduced and compared, along the lines proposed in [105] for the pure advection case. However, this development goes beyond the scope of this paper and will not be pursued here.

Finally, even though achieving full second order consistency is quite complicated in the variable diffusion coefficient case, the previously introduced schemes can be nonetheless extended at least in the simpler configurations as suggested in [18] for the first order case, even though full second order accuracy is not guaranteed any more.

## 2.4 Convergence analysis

We present in this section a convergence analysis for scheme (2.29). For simplicity, we assume a one-dimensional problem defined on $[0,T]\times\mathbb{R}$, with a time-independent drift term $u$:

$$\begin{cases} \partial_t u + \mu(x)u_x - \frac{\sigma^2}{2}u_{xx} = f(u) & (t,x)\in(0,T]\times\mathbb{R}, \\ u(0,x) = u_0(x) & x\in\mathbb{R}. \end{cases} \tag{2.33}$$

The multidimensional case, as well as the time dependence of $\mu$, require only small technical adaptations. On the other hand, the convergence analysis on bounded domains is still an open problem for high-order SL schemes, therefore we will not address this problem here.
First, for $i\in\mathbb{Z}$ and $k\in\mathcal{I}^*_{\Delta t}$, we rewrite scheme (2.29) with the shorthand notation

$$u_{k+1,i} = S_\Delta\left(u_{k+1}, u_k, x_i\right), \tag{2.34}$$

where $x_i = i\Delta x$, and

$$\begin{aligned} S_\Delta\left(u_{k+1}, u_k, x_i\right) &= \alpha_+\left[I[u_k](y_+(x_i)) + \frac{\Delta t}{2}f(I[u_k](y_+(x_i))\right] \\ &+ \alpha_-\left[I[u_k](y_-(x_i)) + \frac{\Delta t}{2}f(I[u_k](y_-(x_i))\right] \\ &+ \alpha_0\left[I[u_k](y_0(x_i)) + \frac{\Delta t}{2}f(I[u_k](y_0(x_i))\right] \\ &+ \frac{\Delta t}{2}f\left(u_{k+1,i}\right). \end{aligned}$$

In one space dimension, the three discrete characteristics are defined by the

equations

$$\begin{aligned} y^+(x) &= x - \frac{\Delta t}{2}[\mu(x) + \mu(y^+(x))] + \sqrt{3\Delta t}\sigma, \\ y^-(x) &= x - \frac{\Delta t}{2}[\mu(x) + \mu(y^-(x))] - \sqrt{3\Delta t}\sigma, \\ y^0(x) &= x - \frac{\Delta t}{2}[\mu(x) + \mu(y^0(x))], \end{aligned}$$

with corresponding weights $\alpha_+ = \alpha_- = 1/6$ and $\alpha_0 = 2/3$. In what follows, we will use the symbol $K$ to denote various positive constants, which do not depend on $\Delta t, x, t$. We will also assume that:

**(H0)** there exists a unique classical solution to (2.33);

**(H1)** $f(x) \in C^4(\mathbb{R})$ with $|f^{(p)}(x)| \leq K$ for $p \leq 4$;

**(H2)** $\mu(x) \in C^2(\mathbb{R})$ with $|\mu^{(p)}(x)| \leq K$ for $p \leq 2$;

**(H3)** for any $v(x) \in C^{q+1}(\mathbb{R})$ with bounded derivatives, $I[v]$ is a piecewise polynomial interpolation operator such that for any $x \in \mathbb{R}$

$$|I[v](x) - v(x)| \leq K\Delta x^q.$$

### 2.4.1 Consistency

First, we derive a consistency result via a Taylor expansion. The same kind of result can be obtained by probabilistic arguments, see [85].

**Proposition 58.** *Assume* **(H1)**–**(H3)**, *and let* $u(t,x)$ *be a smooth solution with bounded derivatives of* (2.33). *Then, for each* $(k,i) \in \mathcal{I}^*_{\Delta t} \times \mathbb{Z}$ *the consistency error of the scheme* (2.29), *defined as*

$$\mathcal{T}_{\Delta t,\Delta x}(t_k, x_i) = \frac{1}{\Delta t}\left(u(t_{k+1}, x_i) - S_\Delta\left(u(t_k), u(t_{k+1}), x_i\right)\right)$$

*where* $u(t_k) = (u(t_k, x_i))_i$, *is such that*

$$\mathcal{T}_\Delta(t,x) = O\left(\Delta t^2 + \frac{\Delta x^q}{\Delta t}\right).$$

*Proof.* In what follows, we will omit the argument of functions computed at $(t,x)$. Consider a smooth solution $u$ of (2.33). Since assumption **(H1)** holds, by differentiating in time and space (2.33) we get that $u$ is also solution to

$$u_{tt} + \mu(x)\, u_{xt} - \frac{\sigma^2}{2} u_{xxt} = f'(u) u_t, \tag{2.35}$$

$$u_{tx} + \mu'(x) u_x + \mu(x)\, u_{xx} - \frac{\sigma^2}{2} u_{xxx} = f'(u) u_x, \tag{2.36}$$

and hence, by differentiating again in space (2.36), of

$$\begin{aligned} &u_{txx} + \mu''(x) u_x + \mu'(x) u_{xx} + \mu'(x) u_{xx} + \mu(x)\, u_{xxx} - \frac{\sigma^2}{2} u_{xxxx} \\ &= f''(u)(u_x)^2 + f'(u) u_{xx}. \end{aligned} \tag{2.37}$$

Using (2.36) and (2.37) in (2.35), we get :

$$\begin{aligned} u_{tt} &= \mu\mu' u_x + \mu^2 u_{xx} - \mu\frac{\sigma^2}{2}u_{xxx} - \frac{\sigma^2}{2}\mu'' u_x - 2\frac{\sigma^2}{2}\mu' u_{xx} \\ &\quad -\mu\frac{\sigma^2}{2}u_{xxx} + \left(\frac{\sigma^2}{2}\right)^2 u_{xxxx} \\ &= \left(\mu\mu' - \frac{\sigma^2}{2}\mu'\right)u_x + \left(\mu^2 - 2\frac{\sigma^2}{2}\mu'\right)u_{xx} - 2\mu\frac{\sigma^2}{2}u_{xxx} + \frac{\sigma^2}{2}^2 u_{xxxx} \\ &\quad +f'(u)(u_t - \mu u_x + \frac{\sigma^2}{2}u_{xx}) + \frac{\sigma^2}{2}f''(u)(u_x)^2. \end{aligned} \tag{2.38}$$

Define now $U_\pm(x) = \mu(x) + \mu(z_\pm(x))$. By a Taylor expansion of $u(t, z_\pm(x))$ in space around $(t, x)$, we obtain

$$\begin{aligned} u(t, z_\pm(x)) &= u + \left(\pm\sqrt{3\Delta t}\sigma - \frac{\Delta t}{2}U_\pm\right)u_x + \frac{1}{2}\left(\pm\sqrt{3\Delta t}\sigma - \frac{\Delta t}{2}U_\pm\right)^2 u_{xx} \\ &\quad +\frac{1}{6}\left(\pm\sqrt{3\Delta t}\sigma - \frac{\Delta t}{2}U_\pm\right)^3 u_{xxx} \\ &\quad +\frac{1}{24}\left(\pm\sqrt{3\Delta t}\sigma - \frac{\Delta t}{2}U_\pm\right)^4 u_{xxxx} \\ &\quad +\frac{1}{120}\left(\pm\sqrt{3\Delta t}\sigma - \frac{\Delta t}{2}U_\pm\right)^5 u_{xxxxx} + O(\Delta t^3) \end{aligned} \tag{2.39}$$

and, defining $U_0(x) = \mu(x) + \mu(y_0(x))$,

$$u(t, y_0(x)) = u - \frac{\Delta t}{2}U_0 u_x + \frac{1}{2}\left(-\frac{\Delta t}{2}U_0\right)^2 u_{xx} + O\left(\Delta t^3\right). \tag{2.40}$$

Using (2.38),(2.39),(2.40) and the Taylor expansion

$$u(t + \Delta t, x) = u + \Delta t u_t + \frac{\Delta t^2}{2}u_{tt} + O\left(\Delta t^3\right),$$

we obtain

$$\begin{aligned} u(t + \Delta t, x) &- \sum_\ell \alpha_\ell u(t, y_k(x)) = \Delta t(u_t + \mu(x)u_x - \frac{\sigma^2}{2}u_{xx}) \\ &+ \frac{\Delta t^2}{2}(f'(u)(u_t - \mu u_x + \frac{\sigma^2}{2}u_{xx}) + \frac{\sigma^2}{2}f''(u)(u_x)^2) \\ &+ O\left(\Delta t^3\right) \end{aligned} \tag{2.41}$$

(note that, here and in what follows, $\ell$ takes values in the set $\{+, -, 0\}$). Consider now the nonlinear reaction term. By assumption **(H3)**, we have that

$$\begin{aligned} f(u(t, y)) = f(u) + f'(u)(u(t, y) - u) + \frac{1}{2}f''(u)(u(t, y) - u)^2 \\ +\frac{1}{6}f'''(u)(u(t, y) - u))^3 + O\left((u(t, y) - u)^4\right) \end{aligned} \tag{2.42}$$

Using (2.41) in (2.42), and taking into account that $u(y_\pm, t) = u \pm \sqrt{3\Delta t}\sigma u_x + O(\Delta t)$ and $u(t, y^0(x)) = u + O(\Delta t)$, we obtain

$$\begin{aligned}\sum_\ell (\alpha_\ell f(u(y^\ell(x), t)) &= f'(u)\left(u(t+\Delta t, x) - \Delta t(u_t + \mu u_x - \frac{\sigma^2}{2}u_{xx}) - u\right) \\ & f(u) + f''(u)(\Delta t\frac{\sigma^2}{2}u_x^2) + O(\Delta t^2). \end{aligned} \tag{2.43}$$

By (2.43) and (2.41), we get the consistency error for the semi-discretization,

$$\begin{aligned} u(t+\Delta t, x) &- \sum_\ell \alpha_\ell \left(u(t, y^\ell(x)) + \frac{\Delta t}{2} f(u(t, y^\ell(x)))\right) - \frac{\Delta t}{2} f(u(t+\Delta t.x)) \\ &= \Delta t(u_t + \mu(x)u_x - \frac{\sigma^2}{2}u_{xx} - f(u)) + O\left(\Delta t^3\right). \end{aligned} \tag{2.44}$$

Introducing the interpolation error and using assumptions **(H2)** and **(H1)**, we finally prove the consistency error for the fully discrete scheme. □ □

### 2.4.2 Stability

To prove stability, it is convenient to recast (2.34) in matrix form as

$$u_{k+1} - \frac{\Delta t}{2} f(u_{k+1}) = \sum_\ell \alpha_\ell \left[B^\ell u_k + \frac{\Delta t}{2} f\left(B^\ell u_k\right)\right], \tag{2.45}$$

where $f(u)$ denotes the vector obtained by applying $f$ elementwise to the components of the vector $u$, while the matrices $B^\ell$ (which represent the operation of interpolating $u_k$ at the points $y^\ell(x_i)$) have elements $b_{ij^\ell}$ defined by

$$b_{ij}^\ell = \beta_j(y^\ell(x_i)), \tag{2.46}$$

for a suitable basis of cardinal functions $\{\beta_j\}$. The following proposition implies stability for the linear part of the scheme with respect to the 2-norm.

**Proposition 59.** *Assume* **(H2)**, *and let the matrix $B$ have elements defined by* (2.46), *with $(\beta_j)$ basis functions for odd degree symmetric Lagrange or splines interpolation. Then, for each $k$, there exists a constant $K_B > 0$ independent on $\Delta x, \Delta t$ such that*

$$\|B^\ell\|_2 \leq 1 + K_B \Delta t. \tag{2.47}$$

*Proof.* Following [53], [54], we sketch the arguments to prove (2.47) for the cases of symmetric Lagrange and splines interpolation. In these cases, the method can be interpreted as Lagrange–Galerkin schemes with area-weighting. First, we make explicit the dependence of the points $y^\ell$ on $x$ and $\Delta t$. We recall that $y^\ell(x) = x + \delta^\ell(x)$, with $\delta^\ell$ solving the equation:

$$\delta^\ell(x) = -\frac{\Delta t}{2}\left(\mu(x) + \mu(x + \delta^\ell(x))\right) + \sqrt{3\Delta t}\sigma e^\ell.$$

Expanding the term $u(x+\delta^\ell(x))$, we obtain therefore

$$\delta^\ell(x) = -\frac{\Delta t}{2}\left(\mu(x)+\mu(x)+\delta^\ell\mu'(x)+O\left((\delta^\ell)^2\right)\right)+\sqrt{3\Delta t}\sigma e^\ell,$$

and hence,

$$\delta^\ell(x)\left(1+\frac{\Delta t}{2}\mu'(x)\right) = -\Delta t\,\mu(x)+\sqrt{3\Delta t}\sigma e^\ell+O\left(\Delta t(\delta^\ell)^2\right)$$

(note that, here and in what follows, assumption **(H2)** ensures that all the remainder terms of the form $O(\cdot)$ are smooth and uniformly bounded wrt $x$). Dividing now by $1+\Delta t\,\mu'/2$, and using the fact that $\delta^\ell = O(\sqrt{\Delta t})$, we get, for $\Delta t \to 0$,

$$y^\ell(x) = x-\Delta t\,\mu(x)+\sqrt{3\Delta t}\sigma e^\ell+O\left(\Delta t^{3/2}\right). \tag{2.48}$$

Due to the term $\sqrt{3\Delta t}\sigma e^\ell$, the form of (2.48) does not coincide with that used in [53] for the points $y^\ell(x)$. However, for a generic couple of points $x_1, x_2 \in \mathcal{R}$, when considering differences $y^\ell(x_1)-y^\ell(x_2)$ this additional term is cancelled, so that

$$\begin{aligned} y^\ell(x_1)-y^\ell(x_2) &= (x_1-x_2)-\Delta t(\mu(x_1)-\mu(x_2))+O\left(|x_1-x_2|\Delta t^{3/2}\right) \\ &= (x_1-x_2)-\Delta t\,\mu'(\xi)(x_1-x_2)+O\left(|x_1-x_2|\Delta t^{3/2}\right), \end{aligned}$$

for a suitable point $\xi \in [\min(x_1,x_2),\max(x_1,x_2)]$ (note that the remainder term may be written in the form $O(|x_1-x_2|\Delta t^{3/2})$, since it comes from the difference of two remainders which have a smooth dependence on $x$). As a consequence, the form (2.48) still satisfies the relevant properties used in the proof of (2.47). In particular, using the triangle inequality in the form of a difference, we get

$$|y^\ell(x_1)-y^\ell(x_2)| \geq |x_1-x_2|-\Delta t\|\mu'\|_\infty|x_1-x_2|+O\left(|x_1-x_2|\Delta t^{3/2}\right).$$

Therefore, the condition [53, Lemma 3]

$$|y^\ell(x_1)-y^\ell(x_2)| \geq \frac{1}{2}|x_1-x_2|$$

is satisfied as soon as

$$\Delta t\|\mu'\|_\infty+O\left(\Delta t^{3/2}\right) < \frac{1}{2}.$$

On the other hand, we have

$$\begin{aligned} |y^\ell(x_1)-(x_1-x_2+y^\ell(x_2))| &\leq \Delta t\|\mu'\|_\infty|x_1-x_2|+O\left(|x_1-x_2|\Delta t^{3/2}\right) \\ &\leq \Delta t\left(\|\mu'\|_\infty+O\left(\Delta t^{1/2}\right)\right)|x_1-x_2|, \end{aligned}$$

which implies, for $\Delta t$ small enough, the condition [53, Theorem 4]

$$|y^\ell(x_1)-(x_1-x_2+y^\ell(x_2))| \leq K_X|x_1-x_2|\Delta t,$$

for a suitable positive constant $K_X$. Then, a careful replica of the arguments used in [53] provides the estimate (2.47). □ □

For a formal definition of the basis functions $\psi_j$ in the case of symmetric Lagrange and spline interpolation, we refer the reader to [53], [54]. While these two cases allow for a complete theory, at least in one space dimension, in the numerical tests with unstructured grids we will also use $\mathbb{P}_2$ interpolants, for which a first attempt of stability analysis in presented in [55].

### 2.4.3 Convergence

We now present a convergence result in the discrete 2-norm.

**Theorem 60.** *Assume* **(H0)**–**(H3)**, *and, in addition, that* (2.47) *is satisfied. Let* $u(t,x)$ *be the classical solution to* (2.33), *and* $u_k$ *be the solution to* (2.34). *Then, for any* $k$ *such that* $t_k \in [0,T]$ *and for* $(\Delta t, \Delta x) \to 0$,

$$\|u(t_k) - u_k\|_2 \leq K_T \left(\Delta t^2 + \frac{\Delta x^q}{\Delta t}\right),$$

*where* $K_T$ *is positive constant depending on the final time* $T$.

*Proof.* While a mere convergence proof could be carried out with weaker regularity assumptions, we will focus here on the error estimate above, which requires the regularity assumptions **(H0)**–**(H3)**. Define the vectors $\gamma_k$ and $\epsilon_k$, so that $\gamma_{k,i} = u(t_k, x_i)$, and $\epsilon_k = \gamma_k - u_k$. Then, by Proposition 58, we get

$$\gamma_{k+1} - \frac{\Delta t}{2} f(\gamma_{k+1}) = \sum_\ell \alpha_\ell \left[(B^\ell)^k \gamma_k + \frac{\Delta t}{2} f\left((B^\ell)^k \gamma_k\right)\right] + O(\Delta t^3 + \Delta x^q), \tag{2.49}$$

where the matrices $(B^\ell)^k$ (which now represent the interpolation of $u_k$ at the points $y^\ell_{k,i}$) have elements $(b^\ell_{ij})^k$ defined by

$$(b^\ell_{ij})^k = \beta_j(y^\ell_{k,i}).$$

Subtracting (2.34) from (2.49), using the Lipschitz continuity of $f$ and the triangle inequality, we obtain from the left-hand side:

$$\left\|\gamma_{k+1} - \frac{\Delta t}{2} f(\gamma_{k+1}) - u_{k+1} + \frac{\Delta t}{2} f(u_{k+1})\right\|_2 \geq \left(1 - \frac{L_f \Delta t}{2}\right) \|\epsilon_{k+1}\|_2 .$$

Taking into account that $\sum_\ell \alpha_\ell = 1$, along with the bound (2.47), we also have from the right-hand side:

$$\begin{aligned} &\left\|\gamma_{k+1} - \frac{\Delta t}{2} f(\gamma_{k+1}) - u_{k+1} + \frac{\Delta t}{2} f(u_{k+1})\right\|_2 \\ &\leq \left(1 + \frac{L_f \Delta t}{2}\right)(1 + K_B \Delta t) \|\epsilon_k\|_2 + O(\Delta t^3 + \Delta x^q). \end{aligned}$$

Therefore, it turns out that

$$\left(1 - \frac{L_f \Delta t}{2}\right) \|\epsilon_{k+1}\|_2 \leq \left(1 + \frac{L_f \Delta t}{2}\right)(1 + K_B \Delta t) \|\epsilon_k\|_2 + O(\Delta t^3 + \Delta x^q). \tag{2.50}$$

Now, for $\Delta t$ small enough to have $1 - L_f \Delta t/2 > \underline{C} > 0$, we have that there exists a constant $K_T > 0$ such that

$$\frac{1 + \dfrac{L_f \Delta t}{2}}{1 - \dfrac{L_f \Delta t}{2}} (1 + K_B \Delta t) \leq 1 + K_T \Delta t,$$

and hence, using this bound in (2.50),

$$\|\epsilon_{k+1}\|_2 \leq (1 + K_T \Delta t) \|\epsilon_k\|_2 + O(\Delta t^3 + \Delta x^q), \tag{2.51}$$

which, by standard arguments, implies that, for any $k$ such that $t_k \in [0, T]$,

$$\|\epsilon_k\|_2 \leq K_T \left( \Delta t^2 + \frac{\Delta x^q}{\Delta t} \right).$$

□ □

## 2.5 Boundary conditions

The treatment of Dirichlet boundary conditions (BCs) for this class of semi-Lagrangian methods has been considered in [86], where two methods are proposed. One approach has first order of consistency, but it does not seem possible to generalize it to multiple dimensions. The second approach has order of consistency 1/2. More recently, in [19], an easier treatment has been proposed for the scheme SL1 with time-independent Dirichlet boundary condition, again with order of consistency 1/2. This approach has been extended in [16] to unstructured meshes.

We propose here a new approach to obtain second order consistency for the scheme SL2 with Dirichlet boundary conditions. This technique is based on the idea of using extrapolation to reconstruct the solution at feet of characteristics falling outside $\mathcal{O}$, much in the spirit of the so-called *ghost-point* techniques, see e.g. [78]. We stress the fact that, while the emphasis in our presentation is on the treatment of BCs for the SL approximation of diffusive problem, the same technique and analysis also hold for the approximation of the pure advection problem, for which accurate BCs for SL methods are by no means easy to derive.

### 2.5.1 Construction of the extrapolation grid

In addition to the standard mesh $\mathcal{G}_{\Delta x} = \{x_i,\ x_i \in \overline{\mathcal{O}}\}$, on which the numerical solution is computed, we consider a second mesh $\mathcal{G}_h = \{\xi_i,\ \xi_i \in \overline{\mathcal{O}}\}$, used only for extrapolation, formed by a single layer of elements having their external side along the boundary of $\mathcal{O}$. This second mesh is constructed with a size parameter $h \sim \sqrt{\Delta t}$, and the degrees of freedom are chosen in order to allow a second-order interpolation. We point out that, as we will soon prove, stability reasons force the parameter $h$ to be at least of the same order of magnitude of the maximum distance of outgoing characteristics form $\mathcal{O}$. This prevents in general from performing extrapolation via the same mesh used for interpolating at interior points.

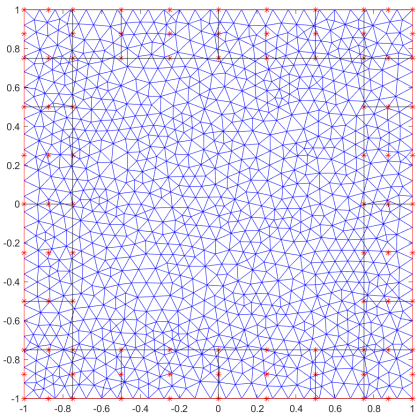



**Figure 2.1.** Unstructured computational mesh (blue triangular elements) together with boundary mesh $\mathcal{G}_h$ (black rectangular elements and red asterisks nodes)

In Fig. 2.1 we show, as an example, a square domain $\mathcal{O} = (-1, 1) \times (-1, 1)$, for which the standard mesh $\mathcal{G}_{\Delta x}$ is formed by the blue triangular elements and the mesh $\mathcal{G}_h$ is formed by the black rectangular elements. Note that, in Fig. 2.1, the elements used for the extrapolation overlap at the corners, but this does not preclude the construction of a stable extrapolation. The asterisks in red denote the nodes of $\mathcal{G}_h$, according to the standard $\mathbb{Q}_2$ element. The values of the numerical solution on the nodes $\xi_i$ are obtained by interpolation at internal nodes, and by the Dirichlet boundary condition if the nodes lie along the boundary $\partial\mathcal{O}$.

We then denote by $\mathcal{T}_{\Delta x}$ a given triangulation, with $\mathcal{G}_{\Delta x}$ the set of the vertices of the elements $K \in \mathcal{T}_{\Delta x}$ and define the polygonal domain $\mathcal{O}_{\Delta x} := \cup_{K \in \tau_{\Delta x}} K \subset \mathcal{O}$. If, for some $i$ and $\ell$, $y^\ell_{k,i} \notin \overline{\mathcal{O}}_{\Delta x}$, then its projection $P(y^\ell_{k,i})$ onto $\overline{\mathcal{O}}_{\Delta x}$ is computed, defined as the point in $\overline{\mathcal{O}}_{\Delta x}$ at minimum distance from $y^\ell_{k,i}$. The value of the numerical solution $u_k(y^\ell_{k,i})$ is then approximated by a quadratic extrapolation operator $\Psi_2$. This operator is constructed via the $\mathbb{Q}_2$ interpolant associated to the element of $\mathcal{G}_h$ to which the projection $P(y^\ell_{k,i})$ belongs:

$$u_k(y^\ell_{k,i}) \simeq \Psi_2[\hat{u}_k](y^\ell_{k,i}),$$

where $\hat{u}_k$ corresponds to

$$\hat{u}_k(\xi_i) = \begin{cases} I[u_k](\xi_i) & \text{if } \xi_i \in \mathcal{O}, \\ b(t_k, \xi_i) & \text{if } \xi_i \in \partial\mathcal{O}. \end{cases}$$

In the case of non-convex domain, the projection may not be unique and we consider as $P(y^\ell_{k,i})$ the closest projection point to the starting grid node $x_i$ with respect the euclidean distance. The method can be extended to more general domains, by considering triangular elements for $\mathcal{G}_h$. In what follows, we provide a simplified analysis for this technique only for the one-dimensional problem, while we present a numerical validation for more complex situations in Section 2.6.5.

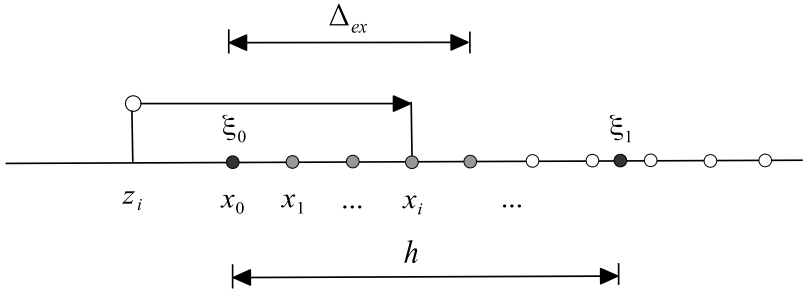



**Figure 2.2.** Boundary extrapolation: basic setting

### 2.5.2 Theoretical analysis

In order to carry out a first theoretical analysis for the extrapolated boundary conditions, we set the problem in one space dimension, use a constant space step, and assume that the computational domain is given by the positive half-line, as in Fig. 2.2. We consider a scheme in the form

$$\begin{cases} u_{k+1,i} = I[u_k](y^\ell_{k,i}) & x_i \notin [x_0, x_0 + \Delta_{ex}], \\ u_{k+1,i} = \Psi[\hat{u}_k](y^\ell_{k,i}) & x_i \in (x_0, x_0 + \Delta_{ex}], \\ u_{k+1,0} = b_{k+1}, \end{cases} \tag{2.52}$$

with $|b_k| \le M_b$, and examine in turn stability and consistency in the treatment of BCs. The form (2.52) is intended to represent a single term, for a given $\ell$, in (2.45).

**Stability** We start for simplicity by using both a first-order interpolation $I[c] = I_1[c]$ at internal points, and a first-order extrapolation $\Psi[c] = \Psi_1[c]$ at the boundary, the latter being performed between the boundary node $x_0 = \xi_0$ and an additional node $\xi_1 = \xi_0 + h$, which needs not coincide with any grid node. We also denote by $\Delta_{ex}$ the measure of the interval on which nodes has their respective feet of characteristics falling outside of the computational domain (and therefore use extrapolated values), so that

$$\Delta_{ex} = \max \left\{ \Delta_i = x_i - x_0 : x_i \in \mathcal{O}, y^\ell_{k,i} \notin \mathcal{O} \right\}.$$

In practice, we will soon show that, for the sake of stability, $h$ must be chosen as a function of $\Delta_{ex}$. In Fig. 2.2, we have marked in black the nodes used for extrapolation, in grey the nodes which require extrapolation, and in white all other nodes.

Consider a generic node $x_i \in (x_0, x_0 + \Delta_{ex})$, the corresponding value $u_{k,i}$ of the numerical solution, and the associated foot of characteristic $y^\ell_{k,i}$. Define

$$\eta = \frac{y^\ell_{k,i} - x_0}{h}$$

(note that $\eta < 0$ if and only if $x_i \in (x_0, x_0 + \Delta_{ex})$). Then, using a first-order extrapolation, we have:

$$u_{k+1,i} = \Psi_1[u_k](y^\ell_{k,i}) = \eta I[u_k](\xi_1) + (1 - \eta) b^n,$$

so that, taking absolute values in the above expression and using the nonexpansivity of $I$ and the boundedness of $b_k$,

$$|u_{k+1,i}| \leq |\eta| \|u_k\|_\infty + |1-\eta| M_b.$$

Since for all the nodes outside the interval $[x_0, x_0 + \Delta_{ex}]$ the $\infty$-norm does not increase, we get

$$\|u_{k+1}\|_\infty \leq \max\left(\|u_k\|_\infty, |\eta| \|u_k\|_\infty + |1-\eta| M_b\right),$$

which leads to a uniform bound for $\|u_k\|_\infty$ as soon as $|\eta| < 1$, that is, for

$$h > \max_i |y^\ell_{k,i} - x_i|,$$

where the maximum is taken among the nodes in the interval $[x_0, x_0 + \Delta_{ex}]$. Note that, in the case of pure advection, we would obtain $h = O(\Delta t)$, whereas, in presence of a diffusion, $h = O(\Delta t^{1/2})$. In both cases, it is natural to choose $h$ of the same order of magnitude of $\Delta_{ex}$.

At a closer look, it turns out that the value $|\eta|$, which affects the stability of the extrapolated values of the solution, is nothing but the absolute value of the Lagrange basis function associated to the node $\xi_1$. To treat a more general case, we can assume that the extrapolation is of degree $N_{ex}$, and uses $x_0$ and $N_{ex}$ more nodes at constant step $h$; in addition, we do not require the interpolation $I$ to be $L^\infty$-nonexpansive, so that possibly $\|B^\ell\|_\infty > 1$.

Then, we can prove the following result:

**Theorem 61.** *Consider the scheme* (2.52)*, and let the extrapolation* $\Psi = \Psi_{N_{ex}}$ *be performed with* $N_{ex}+1$ *evenly spaced nodes* $\xi_k$ *with step* $h$ *and with* $\xi_0 = x_0$*. Assume moreover that the corresponding values of the numerical solution are computed via a possibly high-order interpolation* $I[u_k]$*. Then, there exists a constant* $C$*, depending only on* $N_{ex}$ *and* $I$*, such that, for any* $i$ *for which* $y^\ell_{k,i} \in (x_0 - Ch, x_0]$*, the sequence* $u_{k,i}$ *remains bounded.*

*Proof.* Denote by $L_m(x)$ the Lagrange basis function associated to the extrapolation node $\xi_m$. Using the $I$-interpolated values of the numerical solution at the nodes $\xi_m$, we obtain for the extrapolated values of $u_{k+1,i}$ (second row of (2.52)):

$$\begin{aligned} u_{k+1,i} = \Psi[\hat{u}_k](y^\ell_{k,i}) &= \sum_{m=0}^{N_{ex}} I[\hat{u}_k](\xi_m) L_m(y^\ell_{k,i}) \\ &= b_k L_0(y^\ell_{k,i}) + \sum_{m=1}^{N_{ex}} I[\hat{u}_k](\xi_m) L_m(y^\ell_{k,i}). \end{aligned}$$

Following now the same ideas applied above for the first-order case, and taking into account the possible expansivity of $I$, we have

$$\|u_{k+1}\|_\infty \leq M_b |L_0(z_i)| + \|B^\ell\|_\infty \|u_k\|_\infty \sum_{m=1}^{N_{ex}} |L_m(y^\ell_{k,i})|,$$

and therefore, in order to have stability of the extrapolated values, we should require that

$$\|B^\ell\|_\infty \sum_{m=1}^{N_{ex}} |L_m(y^\ell_{k,i})| < 1. \tag{2.53}$$

On the other hand, replacing $y^\ell_{k,i}$ with the variable $x$ and using the fact that the left-hand side of (2.53) is continuous, since $L_m(\xi_0) = 0$ for all $m \neq 0$,

$$\sum_{m=1}^{N_{ex}} |L_m(\xi_0)| = 0.$$

Then, it follows that

$$\|B^\ell\|_\infty \sum_{m=1}^{N_{ex}} |L_m(x)| < 1 \tag{2.54}$$

in a suitable left neighbourhood of $\xi_0$. By similarity arguments, this neighbourhood can be written in the form $(x_0 - Ch, x_0]$ for some constant $C$ depending only on the degree $N_{ex}$ and on $\|B^\ell\|_\infty$ (that is, on the interpolation $I$). □ □

As a consequence of the previous theorem, the step $h$ should be chosen to satisfy the condition

$$h > \frac{1}{C} \max_i |y^\ell_{k,i} - x_i|, \tag{2.55}$$

where $i$ is indexing all the nodes in $(x_0, x_0 + \Delta_{ex}]$. Note that, for first-order interpolation and extrapolation, we have already obtained $C = 1$. Mixing for example a second-order extrapolation with a first-order interpolation, an easy computation based on (2.54) would provide $C = 1/3$. In the numerical tests, we will use a combination of second-order extrapolation and second-order interpolation, for which it turns out that $C \approx 0.275$.

**Consistency** In evaluating the accuracy of this technique, we should split the error in two components – one associated to internal nodes, which has already been analysed in the previous section, and one related the the treatment of BCs, which comes into play only in the interval $[x_0, x_0 + \Delta_{ex}]$. A similar analysis for the time-discrete case has been carried out in [86, Theorem 4.1] with probabilistic arguments, and we will not repeat it here. For our purposes, the central argument of this analysis is that, representing the numerical scheme as a Markov chain, the expected number of steps spent by the chain in the interval $[x_0, x_0 + \Delta_{ex}]$ is bounded from above, and therefore the error introduced by the treatment of BCs does not accumulate. In our case, this means obtaining a consistency error bounded by the maximum between the internal truncation error proved in Prop. 58 and the extrapolation error (in which the latter should also include the error in reconstructing the values $\xi_i$). Then, the form of the truncation error becomes:

$$\begin{aligned} \mathcal{T}_\Delta(t,x) &= O\left(h^{N_{ex}+1} + \Delta x^q + \Delta t^2 + \frac{\Delta x^q}{\Delta t}\right) \\ &= O\left(h^{N_{ex}+1} + \Delta t^2 + \frac{\Delta x^q}{\Delta t}\right), \end{aligned} \tag{2.56}$$

where, in the last row, we have kept only the asymptotically relevant terms. Thus, while the relationship between $h$ and $\Delta t$ is set according to the stability constraint (2.55), the degree $N_{ex}$ should be chosen to preserve the consistency rate of the scheme. The choice of $N_{ex}$ provides a specific value for the constant $C$ and ultimately, using (2.55), for $h$. We obtain then two different situations:

- *Purely hyperbolic problems ($\sigma = 0$).* In this case, according to (2.55), we have $h \sim \Delta t$, and therefore

$$\mathcal{T}_\Delta(t,x) = O\left(\Delta t^{\min(N_{ex}+1,2)} + \frac{\Delta x^q}{\Delta t}\right).$$

  In order to preserve second-order consistency wrt $\Delta t$, it suffices to enforce BCs with a linear extrapolation.

- *Parabolic problems.* Here, $h \sim \Delta t^{1/2}$ and hence

$$\mathcal{T}_\Delta(t,x) = O\left(\Delta t^{\min((N_{ex}+1)/2,2)} + \frac{\Delta x^q}{\Delta t}\right).$$

  In order to have a second-order scheme, we should therefore apply an extrapolation of degree three. Surprisingly, we will show in the numerical tests that an extrapolation of second degree suffices to retain second-order accuracy. We delay to a future work a deeper analysis of this effect, as well as of other accuracy issues.

**Remark 62.** *In the numerical tests, we will eventually use a structured grid with centered cubic Lagrange interpolation, which requires a second frame of nodes around the cell in which interpolation is performed. Although, in this situation, interpolation in cells neighbouring the boundary would in principle be performed in the "unstable" region of the interpolation stencil, we have not detected any relevant instability in the numerical tests. A complete analysis of this case is out of the scope of this paper, but we note nevertheless that the idea that errors generated at the boundary do not accumulate, used for obtaining the consistency estimate* (2.56)*, also applies to this case, and might provide a qualitative explanation for the stable behaviour of the scheme.*

## 2.6 Numerical results

A number of numerical experiments have been carried out, in order to assess the accuracy of the proposed methods on both structured and unstructured meshes. We start with a simple heat equation, and we increase the level of complexity on the next problems considering an advection- diffusion equation, a reaction- diffusione equation, an advection-diffusion-reaction system and finally an advection-diffusion equation on an non-convex domain.

In Subsections 6.1 and 6.2 , we approximate problems whose analytic solution is known and this allows to compute the errors and to perform a numerical convergence analysis. In Subsections 6.3 and 6.4, we compare the numerical

solutions with approximate solutions, obtained with higher order method. We define the errors, in the infinity and $l^2$ discrete relative norms, as

$$E_\infty = \max_{x_i \in \mathcal{G}_{\Delta x}} |u(t_N, x_i) - u_i^N| / \max_{x_i \in \mathcal{G}_{\Delta x}} |u(t_N, x_i)|,$$

$$E_2 = \left( \sum_{x_i \in \mathcal{G}_{\Delta x}} |u(t_N, x_i) - u_i^N|^2 / \sum_{x_i \in \mathcal{G}_{\Delta x}} |u(t_N, x_i)|^2 \right)^{\frac{1}{2}},$$

and we denote by $p_\infty$ and $p_2$ the corresponding convergence rates.
In the unstructured case, we have constructed a triangular mesh by the Matlab2019 function `initmesh`, with a maximum mesh edge of $\Delta x$, and used a $\mathcal{P}_2$ space reconstruction. In the structured Cartesian case, the bicubic polynomial interpolation implemented in the Matlab2019 command `interp2`, has been used. Since the goal is to evaluate the accuracy of time discretization, both choices avoid to hide the time discretization error with the error introduced by a lower order space reconstruction.

### 2.6.1 Pure diffusion

In a first, basic test, we consider equation (2.14) in the pure diffusion case, i.e., with zero advection and reaction terms, on the square domain $\mathcal{O} = (-2, 2) \times (-2, 2)$, with $T = 1$ and $\sigma^2/2 = 0.05$. Based on the test case proposed in [95], we assume a Gaussian initial datum centered in $(0, 0)$, with $\sigma_G = 0.1$, so that the exact solution in an infinite plane would be

$$u(t, x, y) = \frac{1}{1 + \sigma^2 t/\sigma_G^2} \exp\left\{ -\frac{x^2 + y^2}{2(\sigma_G^2 + \sigma^2 t)} \right\}.$$

For this test case, we only consider structured meshes with constant steps $\Delta x = 4/N$.

Following [18], we consider different time step values $\Delta t$, which correspond to different values of the parabolic stability parameter $m = \Delta t \sigma^2 / 2\Delta x^2$. We compare method SL1 (2.25) and method SL2, (2.29), and collect the results in Table 2.1. Notice that, for method SL1, the value $\theta = 0.52$. This corresponds to a typical procedure in practical applications to realistic problems, see e.g. [15], [106], in which a value of $\theta$ slightly above $1/2$ is used to minimize the amount of numerical dissipation introduced by the time discretization. It can be observed that the expected convergence rates are recovered. Furthermore, it is apparent that scheme SL2 yields a substantial accuracy improvement, without an excessive increase in computational cost. Indeed, the SL2 runs require between 30% and 60% more CPU time, depending on the resolution, while leading to corresponding error reductions between 140% and 730%. As a comparison, a standard second order discretization in space coupled to an explicit second order method in time yields at the finest resolution an error 5 times larger than that of method SL2 at approximately the same computational cost.

| Resolution | | | Relative error | | Convergence rates | |
|---|---|---|---|---|---|---|
| $\Delta x$ | $\Delta t$ | $m$ | $E_2$ | $E_\infty$ | $p_2$ | $p_\infty$ |
| 0.08 | 0.1 | 0.84 | $3.34 \cdot 10^{-2}$ | $5.10 \cdot 10^{-2}$ | - | - |
| 0.04 | 0.05 | 1.6 | $1.33 \cdot 10^{-2}$ | $2.05 \cdot 10^{-2}$ | 1.33 | 1.00 |
| 0.02 | 0.025 | 3.2 | $6.57 \cdot 10^{-3}$ | $1.03 \cdot 10^{-2}$ | 1.02 | 0.99 |

| Resolution | | | Relative error | | Convergence rates | |
|---|---|---|---|---|---|---|
| $\Delta x$ | $\Delta t$ | $m$ | $E_2$ | $E_\infty$ | $p_2$ | $p_\infty$ |
| 0.08 | 0.1 | 0.84 | $2.66 \cdot 10^{-3}$ | $4.76 \cdot 10^{-3}$ | - | - |
| 0.04 | 0.5 | 1.6 | $4.89 \cdot 10^{-4}$ | $8.24 \cdot 10^{-4}$ | 2.44 | 2.53 |
| 0.02 | 0.025 | 3.2 | $8.89 \cdot 10^{-5}$ | $1.48 \cdot 10^{-4}$ | 2.46 | 2.48 |

**Table 2.1.** Errors for the pure diffusion test, first order method SL1 (upper) and second order method SL2 (lower) on a structured mesh.

### 2.6.2 Solid body rotation

Next, we consider the advection–diffusion equation (2.14) with coefficients $\mu = (-\omega y, \omega x)$, $\omega = 2\pi$, $\sigma^2/2 = 0.05$ and $f = 0$ on the square domain $\mathcal{O} = (-2, 2) \times (-2, 2)$ and $T = 1$. Following [95], we assume a Gaussian initial datum centered at $(x_0, y_0) = (1, 0)$ with $\sigma_G = 0.05$, so that the exact solution in an infinite plane would be

$$u(t, x, y) = \frac{1}{1 + \sigma^2 t/\sigma_G^2} \exp\left\{ -\frac{(x - x(t))^2 + (y - y(t))^2}{2(\sigma_G^2 + \sigma^2 t)} \right\}, \tag{2.57}$$

where $x(t) = x_0 \cos \omega t - y_0 \sin \omega t$, $y(t) = x_0 \sin \omega t - y_0 \cos \omega t$. We first consider structured meshes with constant steps $\Delta x = 4/N$. We consider again values of $\Delta t$ corresponding to different parabolic stability parameters $m$, as well as to different Courant number $\lambda = \Delta t \max |\mu|/\Delta x$.

In the structured case, we compare method SL1, (2.25), again with $\theta = 0.52$, and Euler substepping as in (2.31), the decoupled variant SL2s of method (2.29) with Heun substepping, and method SL2 (2.29) with the fully coupling (2.28). The results are reported in Table 2.2, in which convergence rates are computed with respect to the values in the first row. Furthermore, the convergence rate estimation for the values in the last row takes into account that the time step has been reduced by a factor 4. It can be observed that the expected convergence rates with respect to the time discretization error are recovered, in the constant $\Delta x$, constant $C$ or constant $m$ convergence studies. It can also be observed that the decoupled variant SL2s, in spite of the loss of second order convergence, does indeed improve the results with respect to the SL1 method and is competitive with the full second order method SL2. As a comparison, a standard centered finite difference, second order discretization in space coupled to an explicit second order method in time yields at the finest resolution an error analogous to that of method SL2 but requires approximately three times its CPU time.

In the unstructured case, the quadratic polynomial interpolation naturally associated to $\mathbb{P}_2$ finite elements was employed and only the SL2s and SL2

| Resolution | | | | Relative error | | Convergence rates | |
|---|---|---|---|---|---|---|---|
| $\Delta x$ | $\Delta t$ | $\lambda$ | $m$ | $E_2$ | $E_\infty$ | $p_2$ | $p_\infty$ |
| 0.04 | 0.05 | 16 | 1.62 | 0.15 | 0.16 | - | - |
| 0.04 | 0.025 | 8 | 0.82 | $7.71 \cdot 10^{-2}$ | $8.13 \cdot 10^{-2}$ | 0.96 | 0.98 |
| 0.02 | 0.025 | 16 | 3.2 | $7.71 \cdot 10^{-2}$ | $8.13 \cdot 10^{-2}$ | 0.96 | 0.98 |
| 0.02 | 0.0125 | 8 | 1.6 | $3.92 \cdot 10^{-2}$ | $4.13 \cdot 10^{-2}$ | 0.97 | 0.97 |

| Resolution | | | | Relative error | | Convergence rates | |
|---|---|---|---|---|---|---|---|
| $\Delta x$ | $\Delta t$ | $\lambda$ | $m$ | $E_2$ | $E_\infty$ | $p_2$ | $p_\infty$ |
| 0.04 | 0.05 | 16 | 1.62 | $7.65 \cdot 10^{-2}$ | $7.95 \cdot 10^{-2}$ | - | - |
| 0.04 | 0.025 | 8 | 0.82 | $3.89 \cdot 10^{-2}$ | $4.02 \cdot 10^{-2}$ | 0.98 | 0.98 |
| 0.02 | 0.025 | 16 | 3.2 | $3.89 \cdot 10^{-2}$ | $4.02 \cdot 10^{-2}$ | 0.98 | 0.98 |
| 0.02 | 0.0125 | 8 | 1.6 | $1.96 \cdot 10^{-2}$ | $2.02 \cdot 10^{-2}$ | 0.98 | 0.99 |

| Resolution | | | | Relative error | | Convergence rates | |
|---|---|---|---|---|---|---|---|
| $\Delta x$ | $\Delta t$ | $\lambda$ | $m$ | $E_2$ | $E_\infty$ | $p_2$ | $p_\infty$ |
| 0.04 | 0.05 | 16 | 1.62 | 0.11 | 0.11 | - | - |
| 0.04 | 0.025 | 8 | 0.82 | $2.88 \cdot 10^{-2}$ | $2.66 \cdot 10^{-2}$ | 1.93 | 2.05 |
| 0.02 | 0.025 | 16 | 3.2 | $2.89 \cdot 10^{-2}$ | $2.67 \cdot 10^{-2}$ | 1.93 | 2.04 |
| 0.02 | 0.0125 | 8 | 1.6 | $7.35 \cdot 10^{-3}$ | $6.64 \cdot 10^{-3}$ | 1.95 | 2.03 |

**Table 2.2.** Errors for the solid body rotation test, methods SL1 (upper), SL2s (middle) and SL2 (lower) on a structured mesh.

methods were considered. The triangular mesh used was chosen with maximum triangle size $\Delta x$ approximately equal to the corresponding structured meshes. The results are reported in Table 2.3. While the behaviour of the SL2 scheme is entirely analogous to that of the structured mesh case, the SL2s method shows in this case little error reduction when the spatial resolution is kept fixed.

### 2.6.3 Reaction–diffusion equations

Following [51], we consider the Allen–Cahn equation

$$\partial_t u = \frac{\sigma^2}{2}\Delta u - u^3 + u$$

on the domain $\mathcal{O} = (0,1) \times (0,1)$, with periodic boundary conditions and for $t \in [0,2]$. As in [51], we take the initial datum $c_0(x,y) = \sin(2\pi x)\sin(2\pi y)$ and a reference solution is computed by a pseudo-spectral Fourier discretization in space, see e.g. [27], and a fourth order Runge–Kutta scheme in time with a very large number of time steps. The results are reported in Table 2.4, for the values $\sigma^2/2 = 0.01$ and $\sigma^2/2 = 0.05$ of the diffusion parameter, respectively. In this case, only the SL2 scheme on unstructured meshes was considered and the reference solution was interpolated onto the unstructured mesh nodes using a higher order interpolation procedure. Both tests show a quadratic order of convergence.

| Resolution | | | | Relative error | | Convergence rates | |
|---|---|---|---|---|---|---|---|
| $\Delta x$ | $\Delta t$ | $\lambda$ | $m$ | $E_2$ | $E_\infty$ | $p_2$ | $p_\infty$ |
| 0.04 | 0.05 | 16 | 1.62 | $2.39 \cdot 10^{-2}$ | $2.25 \cdot 10^{-2}$ | - | - |
| 0.04 | 0.025 | 8 | 0.82 | $2.72 \cdot 10^{-2}$ | $2.84 \cdot 10^{-2}$ | 0.19 | 0.34 |
| 0.02 | 0.025 | 16 | 3.2 | $7.20 \cdot 10^{-3}$ | $6.32 \cdot 10^{-3}$ | 1.73 | 1.83 |
| 0.02 | 0.0125 | 8 | 1.6 | $2.48 \cdot 10^{-3}$ | $2.59 \cdot 10^{-3}$ | 3.46 | 3.45 |

| Resolution | | | | Relative error | | Convergence rates | |
|---|---|---|---|---|---|---|---|
| $\Delta x$ | $\Delta t$ | $\lambda$ | $m$ | $E_2$ | $E_\infty$ | $p_2$ | $p_\infty$ |
| 0.04 | 0.05 | 16 | 1.62 | 0.129 | 0.139 | - | - |
| 0.04 | 0.025 | 8 | 0.82 | $4.02 \cdot 10^{-2}$ | $4.42 \cdot 10^{-2}$ | 1.68 | 1.65 |
| 0.02 | 0.025 | 16 | 3.2 | $2.88 \cdot 10^{-2}$ | $2.56 \cdot 10^{-2}$ | 2.16 | 2.44 |
| 0.02 | 0.0125 | 8 | 1.6 | $7.70 \cdot 10^{-3}$ | $8.08 \cdot 10^{-3}$ | 2.38 | 2.45 |

**Table 2.3.** Errors for the solid body rotation test, methods SL2s (upper) and SL2 (lower) on an unstructured mesh.

| Resolution | | | Relative error | | Convergence rates | |
|---|---|---|---|---|---|---|
| $\Delta x$ | $\Delta t$ | $m$ | $E_2$ | $E_\infty$ | $p_2$ | $p_\infty$ |
| 0.04 | 0.1 | 0.62 | $1.10 \cdot 10^{-3}$ | $1.31 \cdot 10^{-3}$ | - | - |
| 0.02 | 0.05 | 1.25 | $2.72 \cdot 10^{-4}$ | $2.98 \cdot 10^{-4}$ | 2.02 | 2.14 |
| 0.01 | 0.025 | 2.5 | $6.53 \cdot 10^{-5}$ | $7.06 \cdot 10^{-5}$ | 2.06 | 2.08 |

| Resolution | | | Relative error | | Convergence rates | |
|---|---|---|---|---|---|---|
| $\Delta x$ | $\Delta t$ | $m$ | $E_2$ | $E_\infty$ | $p_2$ | $p_\infty$ |
| 0.04 | 0.1 | 0.62 | $2.82 \cdot 10^{-2}$ | $4.01 \cdot 10^{-2}$ | - | - |
| 0.02 | 0.05 | 1.25 | $7.13 \cdot 10^{-3}$ | $8.47 \cdot 10^{-3}$ | 1.98 | 2.24 |
| 0.01 | 0.025 | 2.5 | $1.97 \cdot 10^{-3}$ | $2.20 \cdot 10^{-3}$ | 1.86 | 1.94 |

**Table 2.4.** Error for the Allen–Cahn test with $\sigma^2/2 = 0.01$ (upper) and $\sigma^2/2 = 0.05$ (lower), second order method SL2 on an unstructured mesh.

### 2.6.4 Advection–diffusion–reaction systems

We consider in this case a set of four coupled advection–diffusion–reaction equations of the form (2.21)

$$\frac{\partial u_k}{\partial t} + \langle \mu, Du_k \rangle - \frac{\sigma^2}{2}\Delta u_k = f_k(u_1, \ldots, u_4) \quad k = 1, \ldots, 4 \tag{2.58}$$

on the square domain $\mathcal{O} = (-5, 5) \times (-5, 5)$ and on the time interval $t \in [0, 5]$. The advection field is given by coefficients $\mu = (-\omega y, \omega x)$, $\omega = 2\pi/10$, while the diffusion coefficient is set as $\sigma^2/2 = 0.01$. The reaction terms are given by

$$\begin{aligned} f_1 &= (u_1 - u_1 u_2) - (u_1 - u_3)/5 \\ f_2 &= -2(u_2 - u_1 u_2) - (u_2 - u_4)/5 \\ f_3 &= 2(u_3 - u_3 u_4) \\ f_4 &= -4(u_4 - u_3 u_4), \end{aligned}$$

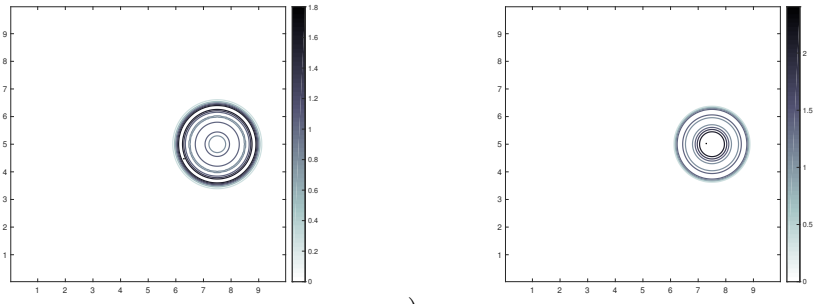
which represent two coupled Lotka–Volterra prey-predator systems. As initial datum for $u_1, u_3$, the function

$$u_0(x, y) = \begin{cases} \cos\left(2\pi[(x + 2.5)^2 + y^2)\right] & \text{for } (x + 2.5)^2 + y^2 \leq \frac{1}{4} \\ 0 & \text{for } (x + 2.5)^2 + y^2 > \frac{1}{4} \end{cases}$$

was considered, while the initial datum for $u_2, u_4$, was taken to be equal to $3u_0$. In this test, only a structured mesh was considered with constant step $\Delta x = 1/20$. A reference solution is computed by a pseudo-spectral Fourier discretization in space and a fourth order Runge–Kutta scheme in time, using a very large number of time steps. The reference solution is reported for two sample components in Figure 2.3, while the absolute error distributions obtained for the same components with the second order method SL2 (2.29) using cubic interpolation, using a timestep corresponding to $\lambda \approx 7$ and $m \approx 1/2$, are shown in Figure 2.4. As a reference, the errors for a second order finite difference approximation of (2.58) using a second order Runge–Kutta scheme in time with a time step 20 times smaller are shown in Figure 2.5, while the errors obtained using a fourth order finite difference approximation for the advection term in (2.58) with a third order Runge–Kutta scheme in time are displayed in Figure 2.6, again computed with a time step 20 times smaller than that used for the SL2 method. It can be seen that the SL2 method allows to achieve errors of the same order of magnitude as those of the third order Runge–Kutta in time, while allowing for a much larger time step without solving large algebraic systems.

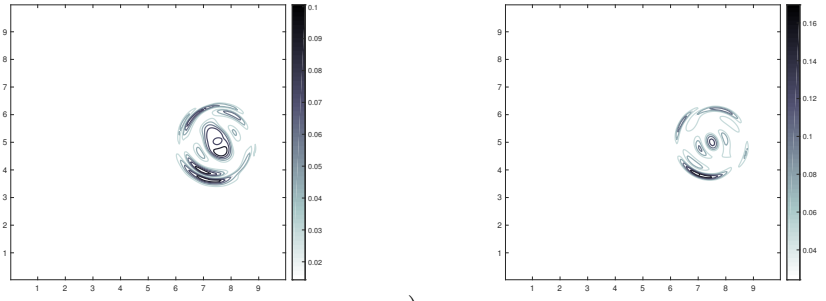
### 2.6.5 Advection–diffusion equation, nonhomogeneous boundary conditions

In this last set of numerical experiments, we consider nonhomogeneous, possibly time-dependent Dirichlet boundary conditions in four cases: pure diffusion, constant advection–diffusion, solid body rotation with diffusion and advection–diffusion on a nonconvex domain. In all these tests, we have used the SL2



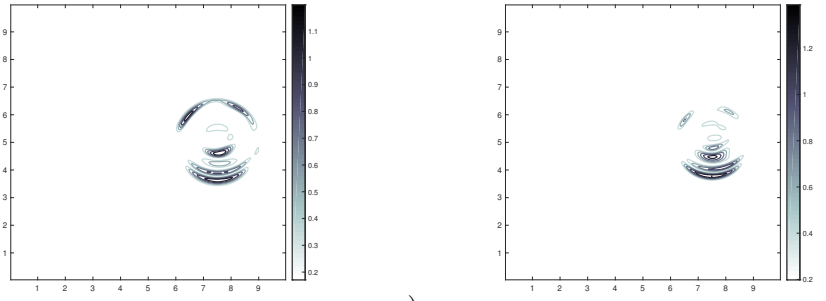
a) b)

**Figure 2.3.** Reference solutions for problem (2.58), a) component $u_3$, b) component $u_4$ at time $T = 5$.



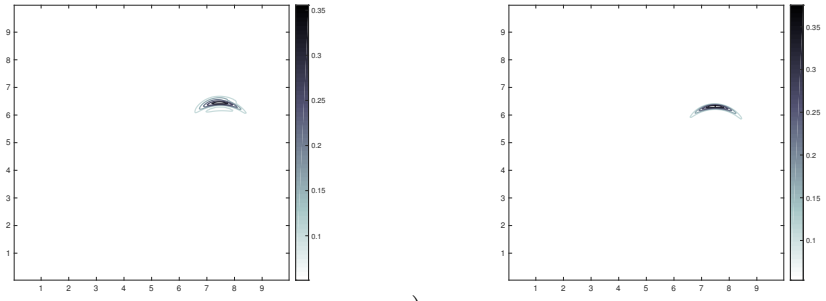
a) b)

**Figure 2.4.** Absolute errors of second order SL2 method for problem (2.58), a) component $u_3$, b) component $u_4$ at time $T = 5$.



a) b)

**Figure 2.5.** Absolute errors of second order finite difference method for problem (2.58), a) component $u_3$, b) component $u_4$ at time $T = 5$.



a) b)

**Figure 2.6.** Absolute errors of fourth order finite difference method for problem (2.58), a) component $u_3$, b) component $u_4$ at time $T = 5$.

| Resolution | | | | Relative error | | Convergence rates | |
|---|---|---|---|---|---|---|---|
| $\Delta x$ | $\Delta t$ | $m$ | $h$ | $E_2$ | $E_\infty$ | $p_2$ | $p_\infty$ |
| 0.04 | 0.05 | 1.56 | 0.5 | $4.70 \cdot 10^{-3}$ | $1.39 \cdot 10^{-2}$ | - | - |
| 0.04 | 0.025 | 0.78 | 0.5 | $3.18 \cdot 10^{-3}$ | $1.06 \cdot 10^{-2}$ | - | - |
| 0.02 | 0.025 | 3.12 | 0.33 | $3.71 \cdot 10^{-4}$ | $1.01 \cdot 10^{-3}$ | 3.66 | 3.78 |
| 0.02 | 0.0125 | 1.56 | 0.33 | $4.35 \cdot 10^{-4}$ | $9.57 \cdot 10^{-4}$ | 2.87 | 3.47 |

**Table 2.5.** Errors and convergence rates for the pure diffusion problem with nonhomogeneous Dirichlet conditions, SL2 method, unstructured mesh

| Resolution | | | | | Relative error | | Conv. rates | |
|---|---|---|---|---|---|---|---|---|
| $\Delta x$ | $\Delta t$ | $\lambda$ | $m$ | $h$ | $E_2$ | $E_\infty$ | $p_2$ | $p_\infty$ |
| 0.04 | 0.05 | 1.25 | 1.56 | 0.5 | $7.35 \cdot 10^{-3}$ | $1.18 \cdot 10^{-2}$ | - | - |
| 0.04 | 0.025 | 0.625 | 0.78 | 0.5 | $8.35 \cdot 10^{-3}$ | $1.32 \cdot 10^{-2}$ | - | - |
| 0.02 | 0.025 | 1.25 | 3.12 | 0.33 | $3.76 \cdot 10^{-4}$ | $7.59 \cdot 10^{-4}$ | 4.29 | 3.96 |
| 0.02 | 0.0125 | 0.625 | 1.56 | 0.33 | $2.64 \cdot 10^{-4}$ | $5.58 \cdot 10^{-4}$ | 4.98 | 4.56 |

**Table 2.6.** Errors and convergence rates for the advection–diffusion problem with nonhomogeneous Dirichlet conditions, SL2 method, unstructured mesh.

scheme on an unstructured mesh. In the first three cases, we consider $\mathcal{O} = (-1, 1) \times (-1, 1)$, final time $T = 1$ and an initial condition in the form of a Gaussian centered at $(x_0, y_0) = (0.5, 0)$, with $\sigma_G = 0.1$. In Fig. 2.1, we show the space meshes $\mathcal{G}_{\Delta x}$ and $\mathcal{G}_h$ corresponding to the steps $\Delta x = 0.04$, $h = 0.5$, which were used to compute the results in the first two rows of Tables 2.5-2.7. In order to have a reference solution to compare with, we compute the exact solution on the whole of $\mathcal{R}^2$ and enforce its values at the boundary as boundary conditions, so that $b(t, x, y) = u(t, x, y)$ for $(x, y) \in \partial\mathcal{O}$, $t \in [0, T]$. For all the three cases, we have set $\sigma^2/2 = 0.05$ and $T = 1$. In the second and third test, the advection field has been chosen as $\mu = (1, 0)$, and $\mu = (-2\pi y, 2\pi x)$, respectively. Tables 2.5-2.7 report the numerical errors obtained by the SL2 scheme in these tests, showing in all cases at least a quadratic convergence.

We finally consider the advection–diffusion equation with $\sigma^2/2 = 0.001$, on the domain $\mathcal{O} = ([0, 1] \times [0, 0.4]) \setminus B_{r_0}(x_0, y_0)$, where $B_{r_0}(x_0, y_0)$ denotes a circle

| Resolution | | | | | Relative error | | Conv. rates | |
|---|---|---|---|---|---|---|---|---|
| $\Delta x$ | $\Delta t$ | $\lambda$ | $m$ | $h$ | $E_2$ | $E_\infty$ | $p_2$ | $p_\infty$ |
| 0.04 | 0.05 | 7.85 | 1.56 | 0.5 | $5.62 \cdot 10^{-2}$ | $6.09 \cdot 10^{-2}$ | - | - |
| 0.04 | 0.025 | 3.92 | 0.78 | 0.5 | $1.49 \cdot 10^{-2}$ | $1.60 \cdot 10^{-2}$ | - | - |
| 0.02 | 0.025 | 7.85 | 3.12 | 0.33 | $1.49 \cdot 10^{-2}$ | $8.98 \cdot 10^{-2}$ | 1.91 | - |
| 0.02 | 0.0125 | 3.92 | 1.56 | 0.33 | $3.43 \cdot 10^{-3}$ | $3.61 \cdot 10^{-3}$ | 2.12 | 2.15 |

**Table 2.7.** Errors and convergence rates for the solid body rotation problem with nonhomogeneous Dirichlet conditions, SL2 method, unstructured mesh.

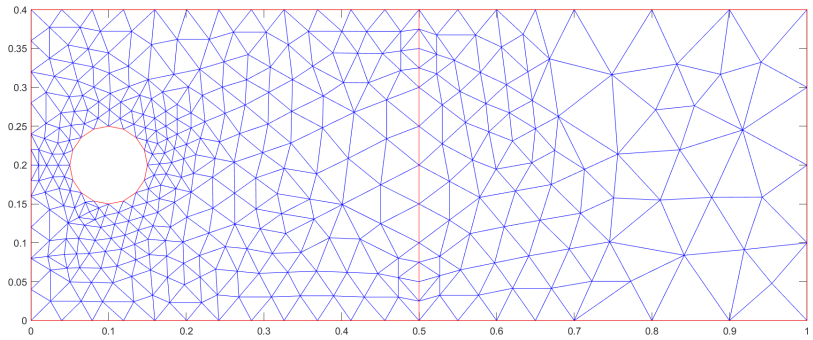



**Figure 2.7.** Unstructured mesh for the non-convex problem

with radius $r_0 = 0.05$ centered in $(x_0, y_0) = (0.1, 0.2)$. The initial datum is $u_0(x, y) = 0$ and the boundary condition

$$b(t, x, y) = \begin{cases} y(0.4 - y)\frac{4}{0.4^2} & (x, y) \in \{0\} \times [0, 0.4], t \in [0, T] \\ 1 & (x, y) \in \partial B_r(x_0), t \in [0, T] \\ 0 & \text{otherwise.} \end{cases}$$

The velocity field $\mu(x, y)$ is given by

$$\mu(x, y) = \left( \mu_0 + \frac{\mu_0 r_0^3}{2r^3} - \frac{3\mu_0 r_0^3 (x - x0)^2}{2r^5}, -\frac{3r_0^3 \mu_0 (x - x0)(y - y0)}{2r^5} \right),$$

where we set $\mu_0 = 0.2$ and $r^2 = (x - x_0)^2 + (y - y_0)^2$. In Fig. 2.7, we show the domain $\mathcal{O}$, discretized using a Delaunay mesh $\mathcal{G}_{\Delta x}$ with $\Delta x = 0.1$, refined around the circular hole. In Fig. 2.8, we show the numerical solution computed with SL2 with time step $\Delta t = 0.005$ for time $t = 0.5, 1, 2, 3$. The nonhomogeneus boundary condition are computed by extrapolation with an extra grid $\mathcal{G}_h$ with $h = 1.5\sqrt{\Delta t}$. In this case, the additional mesh $\mathcal{G}_h$ has been built around the circular hole, as well as along the external rectangular boundary. Note that, even though the wide stencil of the scheme might cause problems with discontinuous initial/boundary data (see the discussion in [52]), the boundary condition is smoothly propagated in the interior of the domain.

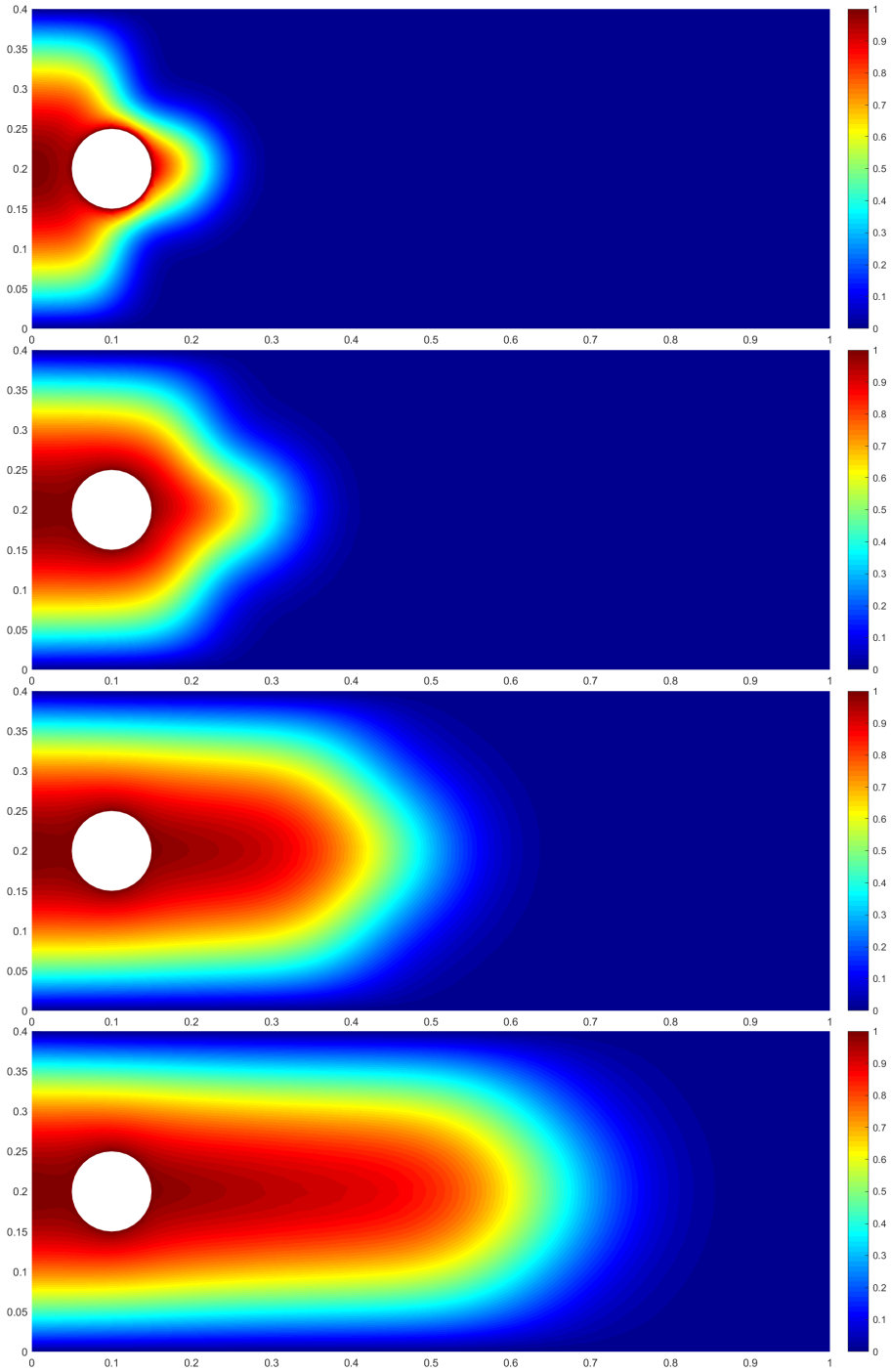


**Figure 2.8.** Numerical solution at time $t = 0.5, 1, 2, 3$

## 2.7 Conclusions

A family of fully semi-Lagrangian approaches for the discretization of advection–diffusion–reaction systems has been proposed, which extend the methods outlined in [18], [19] to full second order accuracy. A numerical treatment of Dirichlet boundary condition, also with second order accuracy, has been proposed. The stability and convergence of the basic second order method have been analyzed.The proposed methods have been validated on a number of classical benchmarks, on both structured and unstructured meshes. Numerical results show that these methods yield good quantitative agreement with reference solutions, while being superior in efficiency to standard implicit methods and to approaches in which the SL method is only used for the advection term. In future developments, the proposed method will be extended to higher order discretizations along the lines of [105] and will be applied to the development of second order fully semi-Lagrangian methods for the Navier-Stokes equations along the lines of [16], [19]. Efficiency improvement for the unstructured implementation of the scheme is currently being studied in [25].

## Acknowledgements

This work has been partly supported by PRIN 2017 project *Innovative numerical methods for evolutionary partial differential equations and applications*, and by INDAM–GNCS 2019 project *Approssimazione numerica di problemi di natura iperbolica ed applicazioni*.
We would also like to thank Stefano Micheletti for providing us MATLAB implementations of some finite element methods, and the two anonymous reviewers for their constructive and helpful comments.

## Data availability statement

Data and codes produced with this work will be made available upon reasonable request.

# Chapter 3

# A semi-Lagrangian scheme for Hamilton-Jacobi-Bellman equations with oblique boundary conditions

We investigate in this work a fully-discrete semi-Lagrangian approximation of second order possibly degenerate Hamilton-Jacobi-Bellman (HJB) equations on a bounded domain $\mathcal{O} \subset \mathbb{R}^d$ with oblique boundary conditions. These equations appear naturally in the study of optimal control of diffusion processes with oblique reflection at the boundary of the domain.

The proposed scheme is shown to satisfy a consistency type property, it is monotone and stable. Our main result is the convergence of the numerical solution towards the unique viscosity solution of the HJB equation. The convergence result holds under the same asymptotic relation between the time and space discretization steps as in the classical setting for semi-Lagrangian schemes on $\mathcal{O} = \mathbb{R}^d$. We present some numerical results that confirm the numerical convergence of the scheme.

In this chapter we deal with the numerical approximation of the following parabolic Hamilton-Jacobi-Bellman (HJB) equation

$$\begin{aligned} \partial_t u + H\left(t, x, Du, D^2u\right) &= 0 \quad \text{in } (0,T] \times \mathcal{O}, \\ L(t, x, Du) &= 0 \quad \text{on } (0,T] \times \partial\mathcal{O}, \\ u(0,x) &= \Psi(x) \quad \text{in } \overline{\mathcal{O}}. \end{aligned} \tag{3.1}$$

In the system above, $T > 0$, $\mathcal{O} \subset \mathbb{R}^d$ is a nonempty smooth bounded open set and $H$ and $L$ are nonlinear functions having the form

$$\begin{aligned} H(t,x,p,M) &= \sup_{a \in A} \left\{ -\frac{1}{2} \operatorname{Tr}\left(\sigma(t,x,a)\sigma(t,x,a)^\top M\right) - \langle \mu(t,x,a), p\rangle - \right. \\ &\qquad \left. f(t,x,a) \right\}, \end{aligned} \tag{3.2}$$

$$L(t,x,p) = \sup_{b \in B} \left\{ \langle \gamma(x,b), p\rangle - g(t,x,b) \right\}, \tag{3.3}$$

where $A \subset \mathbb{R}^{N_A}$ and $B \subset \mathbb{R}^{N_B}$ are nonempty compact sets, $\sigma : [0,T] \times \overline{\mathcal{O}} \times A \to \mathbb{R}^{N \times r}$, with $1 \leq r \leq N$, $\mu : [0,T] \times \overline{\mathcal{O}} \times A \to \mathbb{R}^d$, $f : [0,T] \times \overline{\mathcal{O}} \times A \to \mathbb{R}$, $\gamma : \partial\mathcal{O} \times \mathcal{V} \to \mathbb{R}^d$, with $\mathcal{V} \subseteq \mathbb{R}^{N_B}$ being an open set containing $B$, $g : [0,T] \times \partial\mathcal{O} \times B \to \mathbb{R}$, and $\Psi : \overline{\mathcal{O}} \to \mathbb{R}$.

If $A = \{a\}$ and $B = \{b\}$, for some $a \in \mathbb{R}^{N_A}$ and $b \in \mathbb{R}^{N_B}$, and $\gamma(x,b) = n(x)$, with $n(x)$ being the unit outward normal vector to $\overline{\mathcal{O}}$ at $x \in \partial\mathcal{O}$, then (3.1) reduces to a standard linear parabolic equation with Neumann boundary conditions. In the general case, and after a simple change of the time variable in order to write (3.1) in backward form, the HJB equation (3.1) appears in the study of optimal control of diffusion processes with controlled reflection on the boundary $\partial\mathcal{O}$ (see e.g. [81] for the first order case, i.e. $\sigma \equiv 0$, and [80, 22] for the general case). Since the HJB equation (3.1) is possibly degenerate parabolic, one cannot expect the existence of classical solutions and we have to rely on the notion of viscosity solution (see e.g. [43]). Moreover, as it has been noticed in [79, 81], in general the boundary condition in (3.1) does not hold in the pointwise sense and we have to consider a suitable weak formulation of it. We refer the reader to [81, 8] and [43, 6, 7, 68, 23], respectively, for well-posedness results for HJB equations with oblique boundary condition in the first and second order cases.

## 3.1 Preliminaries

As mentioned in the introduction, it will be simpler to describe our approximation scheme when (3.1) is written in backward form. This can be done by a simple change of the time variable and a possible modification of the time dependency of $H$. Let us set $\mathcal{O}_T := [0,T) \times \mathcal{O}$ and $\overline{\mathcal{O}}_T = [0,T] \times \overline{\mathcal{O}}$. We consider the HJB equation

$$\begin{aligned} -\partial_t u + H\left(t,x,Du,D^2u\right) &= 0 \quad \text{in } \mathcal{O}_T, \\ L(t,x,Du) &= 0 \quad \text{on } [0,T) \times \partial\mathcal{O}, \\ u(T,x) &= \Psi(x) \quad \text{in } \overline{\mathcal{O}}, \end{aligned} \tag{HJB}$$

where $H$ and $L$ are respectively given by (3.2) and (3.3).

For notational convenience, throughout this article, we will write $\gamma_b(x) = \gamma(x,b)$ for all $x \in \partial\mathcal{O}$ and $b \in B$. Our standing assumptions for the data in (HJB) are the following.

**(H1)** $\mathcal{O} \subseteq \mathbb{R}^d$ $(1 \leq N \leq 3)$ is a nonemtpy, bounded domain with boundary $\partial\mathcal{O}$ of class $C^3$.

**(H2)** The functions $\sigma$, $\mu$, $f$, $g$ and $\Psi$ are continuous. Moreover, for every $a \in A$, the functions $\sigma(\cdot,\cdot,a)$ and $\mu(\cdot,\cdot,a)$ are Lipschitz continuous, with Lipschitz constants independent of $a \in A$.

**(H3)** The function $\gamma$ is of class $C^1$. We also assume that

$$(\forall\, (x,b) \in \partial\mathcal{O} \times B) \quad |\gamma_b(x)| = 1 \quad \text{and} \quad \langle n(x), \gamma_b(x) \rangle > 0,$$

where, for every $x \in \partial\mathcal{O}$, we recall that $n(x)$ denotes the unit outward normal vector to $\overline{\mathcal{O}}$ at $x$.

We now recall the notion of viscosity solution to (HJB) (see [6]). We need first to introduce some notation. Given a bounded function $z : (0,T) \times \overline{\mathcal{O}} \to \mathbb{R}$, its upper semicontinuous (resp. lower semicontinuous) envelope is defined by

$$(\forall\,(t,x) \in \overline{\mathcal{O}}_T)\;\; z^*(t,x) := \limsup_{\substack{(s,y)\in\overline{\mathcal{O}}_T,\\ (s,y)\to(t,x)}} z(s,y) \quad \left( \text{resp.} \quad z_*(t,x) := \liminf_{\substack{(s,y)\in\overline{\mathcal{O}}_T,\\ (s,y)\to(t,x)}} z(s,y) \right). \tag{3.4}$$

**Definition 63.** [Viscosity solution] (i) *An upper semicontinuous function $u_1 : \overline{\mathcal{O}}_T \to \mathbb{R}$ is a viscosity subsolution to* (HJB) *if for any $(t,x) \in \overline{\mathcal{O}}_T$ and $\phi \in C^2(\overline{\mathcal{O}}_T)$ such that $u_1 - \phi$ has a local maximum at $(t,x)$, we have*

$$-\partial_t\phi(t,x) + H(t,x,D\phi(t,x),D^2\phi(t,x)) \le 0, \tag{3.5}$$

*if $(t,x) \in \mathcal{O}_T$,*

$$\min\Big\{-\partial_t\phi(t,x) + H(t,x,D\phi(t,x),D^2\phi(t,x)), L(t,x,D\phi(t,x))\Big\} \le 0, \tag{3.6}$$

*if $(t,x) \in [0,T) \times \partial\mathcal{O}$,*

$$u_1(t,x) \le \Psi(x), \tag{3.7}$$

*if $(t,x) \in \{T\} \times \overline{\mathcal{O}}$.*
(ii) *A lower semicontinuous function $u_2 : \overline{\mathcal{O}}_T \to \mathbb{R}$ is a viscosity supersolution to* (HJB) *if for any $(t,x) \in \overline{\mathcal{O}}_T$ and $\phi \in C^2(\overline{\mathcal{O}}_T)$ such that $u_2 - \phi$ has a local minimum at $(t,x)$, we have*

$$-\partial_t\phi(t,x) + H(t,x,D\phi(t,x),D^2\phi(t,x)) \ge 0, \tag{3.8}$$

*if $(t,x) \in \mathcal{O}_T$,*

$$\max\Big\{-\partial_t\phi(t,x) + H(t,x,D\phi(t,x),D^2\phi(t,x)), L(t,x,D\phi(t,x))\Big\} \ge 0, \tag{3.9}$$

*if $(t,x) \in [0,T) \times \partial\mathcal{O}$,*

$$u_2(t,x) \ge \Psi(x), \tag{3.10}$$

*if*

$$(t,x) \in \{T\} \times \overline{\mathcal{O}}$$

*.* (iii) *A bounded function $u : \overline{\mathcal{O}}_T \to \mathbb{R}$ is a viscosity solution to* (HJB) *if $u^*$ and $u_*$, defined in* (3.4)*, are, respectively, sub- and supersolutions to* (HJB)*.*

**Remark 64.** *As shown in [23, Proposition 6], relation* (3.7) *can be replaced by*

$$\min\Big\{-\partial_t\phi(t,x) + H(t,x,D\phi(t,x),D^2\phi(t,x)), u_1(t,x) - \Psi(x)\Big\} \le 0, \tag{3.11}$$

*if $(t,x) \in \{T\} \times \mathcal{O}$, and*

$$\begin{aligned}\min\Big\{-\partial_t\phi(t,x) + H(t,x,D\phi(t,x),D^2\phi(t,x)), L(t,x,D\phi(t,x)),\\ u_1(t,x) - \Psi(x)\} \le 0,\end{aligned} \tag{3.12}$$

*if* $(t,x) \in \{T\} \times \partial\mathcal{O}$. *Similarly, condition* (3.10) *can be replaced by*

$$\max\Big\{-\partial_t\phi(t,x) + H(t,x,D\phi(t,x),D^2\phi(t,x)), u_2(t,x) - \Psi(x)\Big\} \geq 0, \quad (3.13)$$

*if* $(t,x) \in \{T\} \times \mathcal{O}$, *and*

$$\begin{aligned}\max\Big\{-\partial_t\phi(t,x) + H(t,x,D\phi(t,x),D^2\phi(t,x)), L(t,x,D\phi(t,x)),\\ u_2(t,x) - \Psi(x)\} \geq 0,\end{aligned} \quad (3.14)$$

*if* $(t,x) \in \{T\} \times \partial\mathcal{O}$.

The following well-posedness result for (HJB) has been shown in [6, Theorem II.1] (see also [22]).

**Theorem 65.** *Assume* **(H1)-(H3)**. *Then there exists a unique viscosity solution* $u \in C(\overline{\mathcal{O}})$ *to* (HJB).

**Remark 66.** (i) [Comparison principle and uniqueness] *The existence of at most one solution to* (HJB) *follows from the following comparison principle* (*see [6, Theorem II.1] and also [22, Proposition 3.4]*). *If* $u_1 : \overline{\mathcal{O}}_T \to \mathbb{R}$ *is a bounded viscosity subsolution to* (HJB) *and* $u_2 : \overline{\mathcal{O}}_T \to \mathbb{R}$ *is a bounded viscosity supersolution to* (HJB), *then*

$$u_1 \leq u_2 \quad \text{in} \quad \overline{\mathcal{O}}_T.$$

(ii) [Existence] *Once a comparison principle has been shown, the existence of a solution to* (HJB) *follows usually from the existence of sub- and supersolutions to* (HJB) *and Perron's method. In Section 3.4, we construct sub- and supersolutions to* (HJB) *as suitable limits of solutions to the approximation scheme that we present in the next section. Together with the comparison principle, this yields an alternative existence proof of solutions to* (HJB).

*An alternative and interesting technique to show the existence of a solution to* (HJB) *is to consider a suitable stochastic optimal control problem, with controlled reflection of the state trajectory at the boundary* $\partial\mathcal{O}$, *and to show that the associated value function is a viscosity solution to* (HJB). *This strategy has been followed in [22].*

(iii) [Continuity] *The continuity of the unique viscosity solution to* (HJB) *follows directly from the comparison principle and the continuity properties required in the definition of sub- and supersolutions to* (HJB). *Notice that, as usual for parabolic problems with Neumann type boundary conditions, we do not require any compatibility condition between* $\Psi$ *and the operator* $L$ *at the boundary* $\partial\mathcal{O}$.

## 3.2 The fully discrete scheme

We introduce in this section a fully discrete SL scheme that approximates the unique viscosity solution to (HJB). Throughout this section, we assume that **(H1)-(H3)** are fulfilled.

### 3.2.1 Discretization of the space domain $\mathcal{O}$

Let us fix $\Delta x > 0$ and consider a polyhedral domain $\mathcal{O}_{\Delta x} \subseteq \mathbb{R}^d$ such that

$$d(\mathcal{O}, \mathcal{O}_{\Delta x}) = \inf \{|x - y| \,|\, x \in \mathcal{O},\, y \in \mathcal{O}_{\Delta x}\} \leq C(\Delta x)^2, \tag{3.15}$$

for some $C > 0$. A construction of such a $\mathcal{O}_{\Delta x}$ can be found in [10, Section 3] when $d = 2$ or $d = 3$, which explain the dimension constraint in **(H1)**. However, the results in the remainder of this article can be extended to $d > 3$, provided that a numerical domain $\mathcal{O}_{\Delta x}$ satisfying (3.15) exists. Let $\mathcal{T}_{\Delta x}$ be a triangulation of $\mathcal{O}_{\Delta x}$ consisting of simplicial finite elements $\mathsf{T}$ with vertices in $\mathcal{G}_{\Delta x} = \{x_i \mid i \in \{1, \dots, N_{\Delta x}\}\}$ (for some $N_{\Delta x} \in \mathbb{N}$). We assume that $\Delta x$ is the mesh size, i.e. the maximum of the diameters of $\mathsf{T} \in \mathcal{T}_{\Delta x}$, all the vertices on $\partial\mathcal{O}_{\Delta x}$ belong to $\partial\mathcal{O}$, at most one face of each element $\mathsf{T} \in \mathcal{T}_{\Delta x}$, with vertices on $\partial\mathcal{O}_{\Delta x}$, intersects $\partial\mathcal{O}_{\Delta x}$, and $\mathcal{T}_{\Delta x}$ satisfies the following regularity condition: there exists $\delta \in (0, 1)$, independent of $\Delta x$, such that each $\mathsf{T} \in \mathcal{T}_{\Delta x}$ is contained in a ball of radius $\Delta x/\delta$ and contains a ball of radius $\delta\Delta x$. As in [46], we introduce an auxiliary exact triangulation $\widehat{\mathcal{T}}_{\Delta x}$ of $\overline{\mathcal{O}}$ with vertices in $\mathcal{G}_{\Delta x}$. The boundary elements of $\widehat{\mathcal{T}}_{\Delta x}$ are allowed to be curved and we have

$$\overline{\mathcal{O}} = \bigcup_{\widehat{\mathsf{T}} \in \mathcal{T}_{\Delta x}} \widehat{\mathsf{T}}.$$

Denoting by $p_{\mathsf{T}}$ the projection on $\mathsf{T} \in \mathcal{T}_{\Delta x}$, the projection $p_{\Delta x} : \overline{\mathcal{O}} \to \overline{\mathcal{O}}_{\Delta x} \cap \overline{\mathcal{O}}$ is defined by

$$p_{\Delta x}(x) = p_{\mathsf{T}}(x) \quad \text{if } x \in \widehat{\mathsf{T}} \in \widehat{\mathcal{T}}_{\Delta x} \text{ and } \mathsf{T} \in \mathcal{T}_{\Delta x} \text{ has the same vertices than } \widehat{\mathsf{T}}.$$

Set $\mathcal{I}_{\Delta x} = \{1, \dots, N_{\Delta x}\}$ and denote by $\{\beta_i^1 \,|\, i \in \mathcal{I}_{\Delta x}\}$ the linear finite element $\mathbb{P}_1$ basis function on $\mathcal{T}_{\Delta x}$. More precisely, for each $i \in \mathcal{I}_{\Delta x}$, $\psi_i : \mathcal{O}_{\Delta x} \to \mathbb{R}$ is a continuous function, affine on each $\mathsf{T} \in \mathcal{T}_{\Delta x}$, $0 \leq \beta_i^1 \leq 1$, $\beta_i^1(x_i) = 1$, $\beta_i^1(x_j) = 0$ for all $i$, $j \in \mathcal{I}_{\Delta x}$ with $i \neq j$, and $\sum_{i=1}^{N_{\Delta x}} \beta_i^1(x) = 1$ for all $x \in \mathcal{O}_{\Delta x}$. For any $\phi : \mathcal{G}_{\Delta x} \to \mathbb{R}$ its linear interpolation $I[\phi]$ on the mesh $\widehat{\mathcal{T}}_{\Delta x}$ is defined by

$$I[\phi](x) := \sum_{i=1}^{N_{\Delta x}} \beta_i^1(p_{\Delta x}(x))\phi(x_i) \quad \text{for all } x \in \overline{\mathcal{O}}. \tag{3.16}$$

**Lemma 67.** *Let $\phi \in C^2(\overline{\mathcal{O}})$ and denote by $\phi|_{\mathcal{G}_{\Delta x}}$ its restriction to $\mathcal{G}_{\Delta x}$. Then there exists a constant $C_\phi > 0$, independent of $\Delta x$, such that*

$$\sup_{x \in \overline{\mathcal{O}}} |\phi(x) - I[\phi|_{\mathcal{G}_{\Delta x}}](x)| \leq C_\phi(\Delta x)^2. \tag{3.17}$$

*Proof.* Let $x \in \overline{\mathcal{O}}$ and let $\mathsf{T} \in \mathcal{T}_{\Delta x}$ and $\widehat{\mathsf{T}} \in \widehat{\mathcal{T}}_{\Delta x}$ be two elements having the same vertices and such that $x \in \widehat{\mathsf{T}}$. By the triangular inequality

$$|\phi(x) - I[\phi|_{\mathcal{G}_{\Delta x}}](x)| \leq |\phi(x) - \phi(p_{\mathsf{T}}(x))| + |\phi(p_{\mathsf{T}}(x)) - I[\phi|_{\mathcal{G}_{\Delta x}}](x)|.$$

Using that $\phi$ is Lipschitz, we deduce from (3.15) the existence of $C_1 > 0$, independent of $\Delta x$ and $x \in \overline{\mathcal{O}}$, such that $|\phi(x) - \phi(p_{\mathsf{T}}(x))| \leq C_1(\Delta x)^2$. In addition, by standard error estimates for $\mathbb{P}_1$ interpolation (see for instance [39]) and (3.16), there exists $C_2 > 0$, independent of $\Delta x$ and $x \in \overline{\mathcal{O}}$, such that $|\phi(p_{\mathsf{T}}(x)) - I[\phi|_{\mathcal{G}_{\Delta x}}](x)| \leq C_2(\Delta x)^2$. Relation (3.17) follows from these two estimates. □

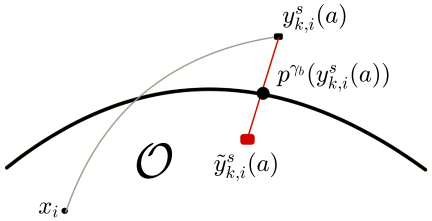



**Figure 3.1.** Reflection: reflected characteristic $\tilde{y}^s_{k,i}(a)$ (red square) starting from $x_i$ (black circle), which exits from $\mathcal{O}$ and arrives in $y^s_{k,i}(a)$ (black square). The red segment represents the oblique direction $\gamma_b$ and the black circle the projected point $p^{\gamma_b}(y^s_{k,i}(a))$.

### 3.2.2 A semi-Lagrangian scheme

Let $\Delta t > 0$, set $N_{\Delta t} := \lfloor T/\Delta t \rfloor$, $\mathcal{I}_{\Delta t} := \{0, \dots, N_{\Delta t}\}$ and $\mathcal{I}^*_{\Delta t} := \mathcal{I}_{\Delta t} \backslash \{N_T\}$. We define the time grid $\mathcal{G}_{\Delta t} := \{t_k \,|\, t_k = k\Delta t,\ k \in \mathcal{I}_{\Delta t}\}$. Given $(k,i) \in \mathcal{I}^*_{\Delta t} \times \mathcal{I}_{\Delta x}$, $a \in A$, and $\ell = 1, \dots, r$, the characteristics are approximated using a forward weak Euler method as follows. $y_k$ is approximated as

$$y_k := x + \Delta t \mu(x) + \sqrt{r\Delta t}\sigma Z$$

where $\sigma \in \mathbb{R}^{d\times r}$ and $Z$ is a vector of $r$ independent random variables, such that for $\ell = 1, \dots, r$,

$$\mathbb{P}(Z^\ell = e_\ell) = \mathbb{P}(Z^\ell = -e_\ell) = \frac{1}{2r}.$$

and

$$\mathbb{P}\left(\bigcup_{1\le \ell_1 < \ell_2 \le r} \{Z^{\ell_1} \neq 0\} \cap \{Z^{\ell_2} \neq 0\}\right) = 0.$$

Then we get $2r$ realisation of $y_k$, defined by the following two discrete fluxes

$$\begin{aligned} y_k^{+,\ell} &:= x + \Delta t \mu(t_k, x) + \sqrt{r\Delta t}\sigma^\ell(t_k, x), \\ y_k^{-,\ell} &:= x + \Delta t \mu(t_k, x) - \sqrt{r\Delta t}\sigma^\ell(t_k, x), \end{aligned} \tag{3.18}$$

for $\ell = 1, \dots, r$. Let $\mathcal{I} = \{+,-\} \times \{1, \dots, r\}$ and let $\bar{c} > 0$ be a fixed constant. For any $\delta > 0$ we set

$$(\partial\mathcal{O})_\delta := \{x \in \mathbb{R}^d \,|\, d(x, \partial\mathcal{O}) < \delta\}.$$

By Proposition 75 in the Appendix, there exist $R > 0$ and two $C^1$ functions $(\partial\mathcal{O})_R \times B \ni (x,b) \mapsto p^{\gamma_b}(x) \in \partial\mathcal{O}$ and $(\partial\mathcal{O})_R \times B \ni (x,b) \mapsto d^{\gamma_b}(x) \in \mathbb{R}$, uniquely determined, such that

$$x = p^{\gamma_b}(x) + d^{\gamma_b}(x)\gamma_b(p^{\gamma_b}(x)) \quad \text{for all } (x,b) \in (\partial\mathcal{O})_R \times B. \tag{3.19}$$

Therefore, there exists $\overline{\Delta t} > 0$ such that for all $\Delta t \in [0, \overline{\Delta t}]$, $(k,i) \in \mathcal{I}^*_{\Delta t} \times \mathcal{I}_{\Delta x}$, $a \in A$, $b \in B$, and $s \in \mathcal{I}$, the *reflected characteristic*

$$\tilde{y}^s_{k,i}(a,b) := \begin{cases} y^s_{k,i}(a) & \text{if } y^s_{k,i}(a) \in \overline{\mathcal{O}}, \\ p^{\gamma_b}(y^s_{k,i}(a)) - \bar{c}\sqrt{\Delta t}\gamma_b(p^{\gamma_b}(y^s_{k,i}(a))) & \text{otherwise} \end{cases} \tag{3.20}$$

is well-defined. In Figure 3.1 we illustrate how the reflected characteristic is computed from the projection $p^{\gamma_b}(y^s_{k,i}(a))$ of $y^s_{k,i}(a)$ onto $\partial\mathcal{O}$ parallel to $\gamma_b$. Let us also set

$$\tilde{d}^s_{k,i}(a,b) := \begin{cases} 0 & \text{if } y^s_{i,k}(a) \in \overline{\mathcal{O}}, \\ d^{\gamma_b}(y^s_{k,i}(a)) + \bar{c}\sqrt{\Delta t} & \text{otherwise,} \end{cases} \tag{3.21}$$

$$\tilde{g}^s_{k,i}(a,b) := \begin{cases} 0 & \text{if } y^s_{k,i}(a) \in \overline{\mathcal{O}}, \\ g\left(t_k, p^{\gamma_b}\left(y^s_{k,i}(a)\right), b\right) & \text{otherwise.} \end{cases} \tag{3.22}$$

Notice that if $y^s_{k,i}(a) \notin \overline{\mathcal{O}}$, then (3.19), (3.20), and (3.21) imply that

$$\tilde{y}^s_{k,i}(a,b) = y^s_{k,i}(a) - \tilde{d}^s_{k,i}(a,b)\gamma_b\left(p^{\gamma_b}(y^s_{k,i}(a))\right). \tag{3.23}$$

For $(k,i) \in \mathcal{I}^*_{\Delta t} \times \mathcal{I}_{\Delta x}$ and $\phi : \mathcal{G}_{\Delta x} \to \mathbb{R}$, let us define $\mathcal{S}_{k,i}[\phi] : A \times B \to \mathbb{R}$ by

$$\mathcal{S}_{k,i}[\phi](a,b) := \frac{1}{2r}\sum_{s\in\mathcal{I}}\left[I[\phi](\tilde{y}^s_{k,i}(a,b)) + \tilde{d}^s_{k,i}(a,b)\tilde{g}^s_{k,i}(a,b)\right] + \Delta t f(t_k, x_i, a), \tag{3.24}$$

and set

$$S_{k,i}[\phi] := \inf_{a\in A,\, b\in B} \mathcal{S}_{k,i}[\phi](a,b). \tag{3.25}$$

In the remainder of this work, we will consider the following fully discrete SL scheme to approximate the solution to (HJB).

$$\begin{aligned} u_{k,i} &= S_{k,i}[u_{k+1,(\cdot)}], \quad \text{for } (k,i) \in \mathcal{I}^*_{\Delta t} \times \mathcal{I}_{\Delta x}, \\ u_{N_{\Delta t},i} &= \Psi(x_i), \qquad \text{for } i \in \mathcal{I}_{\Delta x}. \end{aligned} \tag{$\text{HJB}_{\text{disc}}$}$$

### 3.2.3 Probabilistic interpretation of the scheme

The fully-discrete SL to approximate the solution to (HJB) in the unbounded case, i.e. $\mathcal{O} = \mathbb{R}^d$, has a natural interpretation in terms of a discrete time, finite state, Markov control process (see e.g. [26, Section 3]). We show below that a similar interpretation holds for ($\text{HJB}_{\text{disc}}$). The latter will play an important role in the stability analysis of ($\text{HJB}_{\text{disc}}$) presented in the next section. Given $k \in \mathcal{I}^*_{\Delta t}$ and $a \in A$, $b \in B$, let us define the controlled transition law

$$p_{k,i,j}(a,b) := \frac{1}{2r}\sum_{s\in\mathcal{I}} \beta^1_j(\tilde{y}^s_{k,i}(a,b)) \quad \text{for all } i,\, j \in \mathcal{I}_{\Delta x}. \tag{3.26}$$

We say that $(\pi_k)_{k\in\mathcal{I}^*_{\Delta t}}$ is a $N_{\Delta t}$-*policy* if for all $k \in \mathcal{I}^*_{\Delta t}$ we have $\pi_k : \mathcal{G}_{\Delta x} \to A \times B$. The set of $N_{\Delta t}$-policies is denoted by $\Pi_{N_{\Delta t}}$. Let us fix $k \in \mathcal{I}^*_{\Delta t}$ and, for notational convenience, set $\mathfrak{X}_k = \mathcal{G}^{N_{\Delta t}-k+1}_{\Delta x}$. Associated to $x_i \in \mathcal{G}_{\Delta x}$ and $\pi \in \Pi_{N_{\Delta t}}$, there exists a probability measure $\mathbb{P}^{k,x_i,\pi}$ on $2^{\mathfrak{X}_k}$ (the powerset of $\mathfrak{X}_k$) and a Markov chain $\{X_m \,|\, m = k, \ldots, N_{\Delta t}\}$, with state space $\mathcal{G}_{\Delta x}$, such that

$$\mathbb{P}^{k,x_i,\pi}(X_k = x_i) = 1 \quad \text{and} \quad \mathbb{P}^{k,x_i,\pi}(X_{m+1} = x_j \mid X_m = x_i) = p_{m,i,j}(\pi_m(x_i)), \tag{3.27}$$

for $m = k, \dots, N_{\Delta t} - 1$. Now, consider a family $\{\xi_{k+1}, \dots, \xi_{N_{\Delta t}}\}$ of $\mathbb{R}^r$-valued independent random variables, which are also independent of $\{X_m \,|\, m = k, \dots, N_{\Delta t}\}$, and with common distribution given by

$$\mathbb{P}(\xi_m = \pm e_\ell) = \frac{1}{2r}, \quad \text{for } m = k+1, \dots, N_{\Delta t} \text{ and } \ell = 1, \dots, r,$$

where $e_\ell$ denotes the $\ell$-th canonical vector of $\mathbb{R}^r$. By a slight abuse of notation (see (3.18)), for $m = k, \dots, N_{\Delta t} - 1$, $x_i \in \mathcal{G}_{\Delta x}$, and $a \in A$, let us set

$$y_m(x_i, a) = x_i + \Delta t \mu(t_m, x_i, a) + \sqrt{r \Delta t} \sigma(t_m, x_i, a) \xi_{m+1}. \tag{3.28}$$

For $m = k, \dots, N_{\Delta t} - 1$, $x_i \in \mathcal{G}_{\Delta x}$, $a \in A$, and $b \in B$, define the random variable

$$h(t_m, x_i, a, b) = \begin{cases} 0 & \text{if } y_m(x_i, a) \in \overline{\mathcal{O}}, \\ \left( d^{\gamma_b}(y_m(x_i, a)) + \bar{c}\sqrt{\Delta t} \right) g(t_m, p^{\gamma_b}(y_m(x_i, a)), b) & \text{otherwise.} \end{cases} \tag{3.29}$$

For all $i \in \mathcal{I}_{N_{\Delta x}}$ and $\pi \in \Pi_{N_{\Delta t}}$, let us define

$$\begin{aligned} J_{k,i}(\pi) &= \mathbb{E}_{\mathbb{P}^{k,x_i,\pi}} \left( \sum_{m=k}^{N_{\Delta t}-1} \left[ \Delta t f(t_m, X_m, \alpha_m) + h(t_m, X_m, \alpha_m, \beta_m) \right] \right. \\ &\quad \left. + \Psi(X_{N_{\Delta t}}) \right), \\ J_{N_{\Delta t},i}(\pi) &= \Psi(x_i), \end{aligned}$$

where, for notational convenience, we have denoted, respectively, by $\alpha_m$ and $\beta_m$ the first $N_A$ and the last $N_B$ coordinates of $\pi_m(X_m)$. Notice that, by construction and (3.24), we have that

$$J_{k,i}(\pi) = \mathcal{S}_{k,i}[J_{k+1,(\cdot)}(\pi)](\alpha_k, \beta_k).$$

Moreover, setting

$$\begin{aligned} \hat{U}_{k,i} &= \inf_{\pi \in \Pi_{N_{\Delta t}}} J_{k,i}(\pi), \\ \hat{U}_{N_{\Delta t},i} &= \Psi(x_i), \end{aligned}$$

for all $i \in \mathcal{G}_{\Delta x}$, the dynamic programming principle (see e.g. [65, Theorem 12.1.5]) implies that $\{\hat{U}_{k,i} \,|\, k \in \mathcal{I}_{\Delta t},\, i \in \mathcal{I}_{\Delta x}\}$ satisfies ($\mathrm{HJB}_{\mathrm{disc}}$). Since the latter has a unique solution, we deduce that $U_{k,i} = \hat{U}_{k,i}$ for all $k \in \mathcal{I}_{\Delta t}$ and $i \in \mathcal{I}_{\Delta x}$.

**Remark 68.** *Scheme* ($\mathrm{HJB}_{\mathrm{disc}}$) *can thus be interpreted as a Markov chain discretization of an stochastic control problem with oblique reflection in the boundary* (*see e.g. [22]*).

## 3.3 Properties of the fully discrete scheme

In this section, we establish some basic properties of ($\mathrm{HJB}_{\mathrm{disc}}$).

**Proposition 69.** *The following hold:*
(i) (Monotonicity) *For all* $u, v \colon \mathcal{G}_{\Delta x} \to \mathbb{R}$ *with* $u \le v$, *we have*

$$\mathcal{S}_{k,i}[u] \le \mathcal{S}_{k,i}[v] \quad \text{for } k \in \mathcal{I}^*_{\Delta t} \text{ and } i \in \mathcal{I}_{\Delta x}.$$

(ii) (Commutation by constant) *For any* $c \in \mathbb{R}$ *and* $u \colon \mathcal{G}_{\Delta x} \to \mathbb{R}$,

$$\mathcal{S}_{k,i}[u + c] = \mathcal{S}_{k,i}[u] + c \quad \text{for } k \in \mathcal{I}^*_{\Delta t} \text{ and } i \in \mathcal{I}_{\Delta x}.$$

*Proof.* Both assertions follow directly from (3.24) and ($\mathrm{HJB}_{\mathrm{disc}}$). □

We show in Proposition 70 below a consistency result for ($\mathrm{HJB}_{\mathrm{disc}}$). For this purpose, let us set

$$\begin{aligned}\mathcal{H}(t,x,p,M,a) &= -\frac{1}{2}\operatorname{Tr}\left(\sigma(t,x,a)\sigma(t,x,a)^\top M\right) - \langle \mu(t,x,a), p\rangle - f(t,x,a)\\ &\quad \text{for } (t,x,p,M,a) \in \overline{O}_T \times \mathbb{R}^d \times \mathbb{R}^{d\times r} \times A,\\ \mathcal{L}(t,x,p,b) &= \langle \gamma(x,b), p\rangle - g(t,x,b)\\ &\quad \text{for } (t,x,p,b) \in [0,T] \times \partial\mathcal{O} \times \mathbb{R}^d \times B,\end{aligned}$$

and for all $k \in \mathcal{I}^*_{\Delta t}$, $i \in \mathcal{I}_{\Delta x}$, $s \in \mathcal{I}$, $q \in \mathbb{R}^d$, $a \in A$, and $b \in B$, define

$$\tilde{\mathcal{L}}^s_{k,i}(q,a,b) := \begin{cases} 0 & \text{if } y^s_{k,i}(a) \in \overline{\mathcal{O}},\\ \mathcal{L}\left(t_k, p^{\gamma_b}(y^s_{k,i}(a)), q, b\right) & \text{otherwise.}\end{cases} \tag{3.30}$$

**Proposition 70** (Consistency)**.** *Let $\phi \in C^3\left(\overline{\mathcal{O}}\right)$ and denote by $\phi|_{\mathcal{G}_{\Delta x}}$ its restriction to $\mathcal{G}_{\Delta x}$. Then the following hold:*

(i) *For all $k \in \mathcal{I}^*_{\Delta t}$, $i \in \mathcal{I}_{\Delta x}$, $a \in A$, and $b \in B$, we have*

$$\begin{aligned}\mathcal{S}_{k,i}[\phi|_{\mathcal{G}_{\Delta x}}](a,b) - \phi(x_i) &= -\Delta t\mathcal{H}(t_k, x_i, D\phi(x_i), D^2\phi(x_i), a)\\ &\quad - \frac{1}{2r}\sum_{s\in\mathcal{I}} \tilde{d}^s_{k,i}(a,b)\left(\tilde{\mathcal{L}}^s_{k,i}(D\phi(x_i),a,b) - \sqrt{\Delta t}K^s_{k,i}(a,b)\right)\\ &\quad + O\left(\Delta t\sqrt{\Delta t} + (\Delta x)^2\right),\end{aligned} \tag{3.31}$$

*where the set of constants $\{K^s_{k,i}(a,b) \mid k \in \mathcal{I}^*_{\Delta t},\ i \in \mathcal{I}_{\Delta x},\ s \in \mathcal{I},\ a \in A,\ b \in B\}$ is bounded, independently of $(\Delta t, \Delta x)$.*

(ii) *For all $k \in \mathcal{I}^*_{\Delta t}$ and $i \in \mathcal{I}_{\Delta x}$, we have*

$$\begin{aligned}S_{k,i}[\phi|_{\mathcal{G}_{\Delta x}}] - \phi(x_i) &= -\sup_{a\in A,\, b\in B}\Big\{\Delta t\mathcal{H}(t_k, x_i, D\phi(x_i), D^2\phi(x_i), a)\\ &\quad + \frac{1}{2r}\sum_{s\in\mathcal{I}} \tilde{d}^s_{k,i}(a,b)\left(\tilde{\mathcal{L}}^s_{k,i}(D\phi(x_i),a,b) - \sqrt{\Delta t}K^s_{k,i}(a,b)\right)\Big\}\\ &\quad + O\left(\Delta t\sqrt{\Delta t} + (\Delta x)^2\right).\end{aligned} \tag{3.32}$$

*Proof.* In what follows, we denote by $C > 0$ a generic constant, which is independent of $k$, $i$, $s$ $a$, $b$, $\Delta t$ and $\Delta x$. Since assertion (ii) follows directly from (i), we only show the latter.

For every $s \in \mathcal{I}$, (3.18) and (3.21) imply that $0 \le \tilde{d}^s_{k,i}(a,b) \le C\sqrt{\Delta t}$. Thus, by (3.18), (3.23), and a second order Taylor expansion of $\phi$ around $x_i$, for every

$\ell = 1, \dots, r$, we have

$$\begin{aligned}
\phi\left(\tilde{y}^{\pm,\ell}_{k,i}(a,b)\right) &= \phi(x_i) + \Delta t \langle D\phi(x_i), \mu(t_k, x_i, a)\rangle \\
&\quad + \tfrac{r\Delta t}{2}\langle D^2\phi(x_i)\sigma^\ell(t_k,x_i,a), \sigma^\ell(t_k,x_i,a)\rangle \\
&\quad \pm\sqrt{r\Delta t}\langle D\phi(x_i), \sigma^\ell(t_k,x_i,a)\rangle \\
&\quad - \tilde{d}^{\pm,\ell}_{k,i}(a,b)\left\langle D\phi(x_i), \tilde{\gamma}^{\pm,\ell}_{k,i}(a,b)\right\rangle \\
&\quad + \tfrac{\left(\tilde{d}^{\pm,\ell}_{k,i}(a,b)\right)^2}{2}\left\langle D^2\phi(x_i)\tilde{\gamma}^{\pm,\ell}_{k,i}(a,b), \tilde{\gamma}^{\pm,\ell}_{k,i}(a,b)\right\rangle \\
&\quad \mp\sqrt{r\Delta t}\tilde{d}^{\pm,\ell}_{k,i}(a,b)\left\langle D^2\phi(x_i)\tilde{\gamma}^{\pm,\ell}_{k,i}(a,b), \sigma^\ell(t_k,x_i,a)\right\rangle \\
&\quad + O\left(\Delta t\sqrt{\Delta t}\right),
\end{aligned}$$

where, for every $s \in \mathcal{I}$,

$$\tilde{\gamma}^s_{k,i}(a,b) := \begin{cases} 0 & \text{if } y^s_{k,i}(a) \in \overline{\mathcal{O}}, \\ \gamma_b\left(p^{\gamma_b}(y^s_{k,i}(a))\right) & \text{otherwise.} \end{cases}$$

This implies that

$$\begin{aligned}
&\tfrac{1}{2}\phi\left(\tilde{y}^{+,\ell}_{k,i}(a,b)\right) + \tfrac{1}{2}\phi\left(\tilde{y}^{-,\ell}_{k,i}(a,b)\right) \\
&\qquad = \phi(x_i) + \Delta t\,\langle D\phi(x_i), \mu(t_k,x_i,a)\rangle \\
&\qquad\quad + \tfrac{r\Delta t}{2}\left\langle D^2\phi(x_i)\sigma^\ell(t_k,x_i,a), \sigma^\ell(t_k,x_i,a)\right\rangle \\
&\qquad\quad - \tilde{d}^{+,\ell}_{k,i}(a,b)\left(\left\langle D\phi(x_i), \tilde{\gamma}^{+,\ell}_{k,i}(a,b)\right\rangle - \sqrt{\Delta t}K^{+,\ell}_{k,i}(a,b)\right) \\
&\qquad\quad - \tilde{d}^{-,\ell}_{k,i}(a,b)\left(\left\langle D\phi(x_i), \tilde{\gamma}^{-,\ell}_{k,i}(a,b)\right\rangle - \sqrt{\Delta t}K^{-,\ell}_{k,i}(a,b)\right) + O\left(\Delta t\sqrt{\Delta t}\right),
\end{aligned} \tag{3.33}$$

where

$$\begin{aligned}
K^{\pm,\ell}_{k,i}(a,b) := \frac{\tilde{d}^{\pm,\ell}_{k,i}(a,b)}{2\sqrt{\Delta t}}&\langle D^2\phi(x_i)\tilde{\gamma}^{\pm,\ell}_{k,i}(a,b), \tilde{\gamma}^{\pm,\ell}_{k,i}(a,b)\rangle \\
&\mp \sqrt{r}\langle D^2\phi(x_i)\tilde{\gamma}^{\pm,\ell}_{k,i}(a,b), \sigma^\ell(t_k,x_i,a)\rangle.
\end{aligned}$$

Multiplying (3.33) by $1/r$ and taking the sum over $s \in \mathcal{I}$, we obtain

$$\begin{aligned}
\tfrac{1}{2r}\sum_{s\in\mathcal{I}}\phi(\tilde{y}^s_{k,i}(a,b)) &= \phi(x) + \Delta t\langle D\phi(x_i), \mu(t_k,x_i,a)\rangle \\
&\quad + \tfrac{\Delta t}{2}\operatorname{Tr}\left(\sigma(t_k,x_i,a)\sigma(t_k,x_i,a)^T D^2\phi(x_i)\right) \\
&\quad - \tfrac{1}{2r}\sum_{s\in\mathcal{I}}\tilde{d}^s_{k,i}(a,b)\left(\left\langle D\phi(x_i), \tilde{\gamma}^s_{k,i}(a,b)\right\rangle\right. \\
&\quad \left. - \sqrt{\Delta t}K^s_{k,i}(a,b)\right) + O\left(\Delta t\sqrt{\Delta t}\right),
\end{aligned}$$

which, by Lemma 67, yields

$$\begin{aligned}\tfrac{1}{2r}\sum_{s\in\mathcal{I}} I[\phi|_{\mathcal{G}_{\Delta x}}](\tilde{y}^s_{k,i}(a,b)) &= \phi(x) + \Delta t\langle D\phi(x_i), \mu(t_k,x_i,a)\rangle \\ &+\tfrac{\Delta t}{2}\operatorname{Tr}\left(\sigma(t_k,x_i,a)\sigma(t_k,x_i,a)^T D^2\phi(x_i)\right) \\ &-\tfrac{1}{2r}\sum_{s\in\mathcal{I}}\tilde{d}^s_{k,i}(a,b)\left(\left\langle D\phi(x_i), \tilde{\gamma}^s_{k,i}(a,b)\right\rangle\right. \\ &\left.-\sqrt{\Delta t}K^s_{k,i}(a,b)\right) + O\left(\Delta t\sqrt{\Delta t} + (\Delta x)^2\right).\end{aligned}$$

The result follows from the previous expression, (3.24), (3.30) and (3.30). □

For $k \in \mathcal{I}^*_{N_{\Delta t}}$ and $a \in A$, let us define

$$(\forall\, k \in \mathcal{I}^*_{N_{\Delta t}}, \forall\, a \in A) \quad \Gamma_k(a) := \{x_i \in \mathcal{G}_{\Delta x} \mid \exists\ s \in \mathcal{I},\ y^s_{k,i}(a) \notin \overline{\mathcal{O}}\}, \tag{3.34}$$

and recall from Section 3.2.3 that given $x_i \in \mathcal{G}_{\Delta x}$ and a policy $\pi \in \Pi_{N_{\Delta t}}$, the Markov chain $\{X_m \mid m = k, \dots, N_{\Delta t}\}$ is defined by the transition probabilities (3.27). As in Section 3.2.3, we denote by $\alpha_m$ and $\beta_m$ $(m = k, \dots, N_{\Delta t} - 1)$, respectively, the first $N_A$ and the last $N_B$ coordinates of $\pi_m(X_m)$. Finally, given $D \subset \mathbb{R}^d$, we denote by $\mathbb{I}_D$ the indicator function of $D$, i.e. $\mathbb{I}_D(x) = 1$, if $x \in \mathcal{O}$, and $\mathbb{I}_D(x) = 0$, otherwise.

The following technical result will be useful to establish the stability of $(\mathrm{HJB}_{\mathrm{disc}})$.

**Lemma 71.** *The following holds:*

$$\sup_{k\in\mathcal{I}^*_{\Delta t},\, i\in\mathcal{I}^*_{\Delta x},\pi\in\Pi_{N_{\Delta t}}} \mathbb{E}_{\mathbb{P}^{k,x_i,\pi}}\left(\sum_{m=k}^{N_T-1}\mathbb{I}_{\Gamma_m(\alpha_m)}(X_m)\right) \le \frac{C}{\sqrt{\Delta t}}, \tag{3.35}$$

*where $C > 0$ is independent of $(\Delta t, \Delta x)$ as long as $\Delta t$ is small enough and $(\Delta x)^2/\Delta t$ is bounded.*

*Proof.* The argument of the proof is inspired from [83, Lemma 1]. Let $\varepsilon > 0$, set

$$D_\varepsilon = \{x \in \overline{\mathcal{O}} \mid d(x, \partial\mathcal{O}) > \varepsilon\}, \quad \partial\mathcal{O}_\varepsilon = \{x \in \overline{\mathcal{O}} \mid d(x, \partial\mathcal{O}) = \varepsilon\},$$
$$L_\varepsilon = \{x \in \overline{\mathcal{O}} \mid d(x, \partial\mathcal{O}) \le \varepsilon\},$$

and define $\overline{\mathcal{O}} \ni x \mapsto w_\varepsilon(x) = d^2(x, D_\varepsilon) \in \mathbb{R}$. By Lemma 76(v) in the Appendix, there exists $\eta > 0$ such that $w_\eta \in C^3(\overline{\mathcal{O}} \setminus \partial\mathcal{O}_\eta)$ with bounded third order derivatives on the connected components of $\overline{\mathcal{O}} \setminus \partial\mathcal{O}_\eta$. Let us fix this $\eta$ and, for notational convenience, let us write $w = w_\eta$. Let $M > 0$ and, for any $k \in \mathcal{I}_{\Delta t}$, define

$$\overline{\mathcal{O}} \ni x \mapsto W_k(x) = \begin{cases} M(T - t_k) + w(x) & \text{if } k \in \mathcal{I}^*_{\Delta t}, \\ 0 & \text{if } k = N_{\Delta t} \end{cases} \in \mathbb{R}. \tag{3.36}$$

By (3.24), with $f \equiv 0$ and $g \equiv 0$, for all $a \in A$ and $b \in B$, we have

$$\mathcal{S}_{k,i}[W_{k+1}|_{\mathcal{G}_{\Delta x}}](a,b) - W_k(x_i) = -M\Delta t + \mathcal{S}_{k,i}[w|_{\mathcal{G}_{\Delta x}}](a,b) - w(x_i), \tag{3.37}$$
$$= -M\Delta t + \frac{1}{2r}\sum_{s\in\mathcal{I}} I[w](\tilde{y}^s_{k,i}(a,b)) - w(x_i). \tag{3.38}$$

Moreover, assumption **(H2)** implies the existence of $\overline{C} > 0$ such that

$$\sup\left\{|y^s_{k,i}(a) - x_i| \;\middle|\; k \in \mathcal{I}^*_{\Delta t},\, i \in \mathcal{I}_{\Delta x},\, a \in A,\, s \in \mathcal{I}\right\} \leq \overline{C}\sqrt{\Delta t}. \tag{3.39}$$

Now, let us fix $k \in \mathcal{I}^*_{\Delta t}$, $i \in \mathcal{I}_{\Delta x}$, $a \in A$, and $b \in B$. We have the following cases.

(i) $x_i \notin \Gamma_k(a)$ and $d(x_i, \partial\mathcal{O}_\eta) \geq \overline{C}\sqrt{\Delta t}$. The first condition implies that $y^s_{k,i}(a) \in \overline{\mathcal{O}}$, for any $s \in \mathcal{I}$, and, hence, (3.20) yields $\tilde{y}^s_{k,i}(a,b) = y^s_{k,i}(a)$. The condition $d(x_i, \partial\mathcal{O}_\eta) \geq \overline{C}\sqrt{\Delta t}$, (3.39), and standard error estimates for $\mathbb{P}_1$ interpolation (see for instance [39]), imply that

$$I[w](\tilde{y}^s_{k,i}(a,b)) = w(\tilde{y}^s_{k,i}(a,b)) + O((\Delta x)^2) = w(y^s_{k,i}(a)) + O((\Delta x)^2).$$

Since, by second order Talyor expansion, $\frac{1}{2r}\sum_{s\in\mathcal{I}} w(y^s_{k,i}(a)) - w(x_i) = O(\Delta t)$, (3.38) yields

$$\mathcal{S}_{k,i}[W_{k+1}|_{\mathcal{G}_{\Delta x}}](a,b) - W_k(x_i) = -M\Delta t + O\left(\Delta t + (\Delta x)^2\right). \tag{3.40}$$

(ii) $x_i \notin \Gamma_k(a)$ and $d(x_i, \partial\mathcal{O}_\eta) < \overline{C}\sqrt{\Delta t}$. Condition $d(x_i, \partial\mathcal{O}_\eta) < \overline{C}\sqrt{\Delta t}$ and (3.39) imply that $w(x_i) = O(\Delta t)$ and, for any $s \in \mathcal{I}$, $d^2(y^s_{k,i}(a), \partial\mathcal{O}_\eta) = O(\Delta t)$. Since the cardinality of $\mathcal{J} := \{j \in \mathcal{I}_{\Delta x} \,|\, \psi_j(y^s_{k,i}(a)) > 0\}$ is independent of $\Delta x$ and, for all $j \in \mathcal{J}$, $|y^s_{k,i}(a) - x_j| = O(\Delta x)$, we deduce that

$$\begin{aligned} I[w](y^s_{k,i}(a)) &= \textstyle\sum_{j\in\mathcal{J}} \psi_j(y^s_{k,i}(a)) w(x_j) \\ &\leq \textstyle\sum_{j\in\mathcal{J}} \psi_j(y^s_{k,i}(a)) d^2(x_j, \partial\mathcal{O}_\eta) \\ &= \textstyle\sum_{j\in\mathcal{J}} \psi_j(y^s_{k,i}(a)) d^2(y^s_{k,i}(a), \partial\mathcal{O}_\eta) + O((\Delta x)^2) \\ &= O(\Delta t + (\Delta x)^2). \end{aligned}$$

Thus, since $\tilde{y}^s_{k,i}(a,b) = y^s_{k,i}(a)$, (3.38) implies that (3.40) still holds.

(iii) $x_i \in \Gamma_k(a)$. Let $0 < \delta < \eta$. Since $\mu$ and $\sigma$ are bounded, there exists $\overline{\Delta t} > 0$, independent of $k$, $i$ and $a$, such that

$$\Gamma_k(a) \subseteq L_\delta \subset L_\eta, \tag{3.41}$$

if $\Delta t \leq \overline{\Delta t}$. By (3.37) and Proposition 70(i), with $f \equiv 0$ and $g \equiv 0$, we have

$$\begin{aligned} \mathcal{S}_{k,i}[W_{k+1}|_{\mathcal{G}_{\Delta x}}](a,b) - W_k(x_i) = \\ - M\Delta t - \frac{1}{2r}\sum_{s\in\mathcal{I}} \tilde{d}^s_{k,i}(a,b) \left\langle Dw(x_i), \gamma_b\left(p^{\gamma_b}\left(y^s_{k,i}(a)\right)\right)\right\rangle \\ + O\left(\Delta t + (\Delta x)^2\right). \end{aligned} \tag{3.42}$$

By Lemma 76(v) in the Appendix, for any $x \in L_\eta$, we have $d(x, \partial\mathcal{O}_\eta) = \eta - d(x, \partial\mathcal{O})$. Thus, Lemma 76(ii) implies that $Dd(x, \partial\mathcal{O}_\eta) = n(p_{\partial\mathcal{O}}(x))$, and hence

$$Dw(x_i) = 2d(x_i, \partial\mathcal{O}_\eta)\, Dd(x_i, \partial\mathcal{O}_\eta) = 2d(x_i, \partial\mathcal{O}_\eta)\, n(p_{\partial\mathcal{O}}(x)). \tag{3.43}$$

On the other hand, in view of [63, Proposition 1.1(v)], there exists $C > 0$ such that $|d^{\gamma_b}(x_i)| \leq Cd(x_i, \partial\mathcal{O})$. Thus,

$$\begin{aligned}|p^{\gamma_b}(x_i) - p_{\partial\mathcal{O}}(x_i)| &\leq |p^{\gamma_b}(x_i) - x_i| + |x_i - p_{\partial\mathcal{O}}(x_i)| \\ &= |d^{\gamma_b}(x_i)| + d(x_i, \partial\mathcal{O}) \leq (C+1)d(x_i, \partial\mathcal{O}).\end{aligned}$$

Since $x_i \in \Gamma_k(a)$, we have $d(x_i, \partial\mathcal{O}) = O(\sqrt{\Delta t})$ and hence $|p^{\gamma_b}(x_i) - p_{\partial\mathcal{O}}(x_i)| = O(\sqrt{\Delta t})$. Proposition 75 implies that $\gamma_b$ and $p^{\gamma_b}$ are Lipschitz. Therefore, for any $s \in \mathcal{I}$,

$$\gamma_b\left(p^{\gamma_b}\left(y^s_{k,i}(a)\right)\right) = \gamma_b\left(p^{\gamma_b}(x_i)\right) + O\left(\sqrt{\Delta t}\right) = \gamma_b\left(p_{\partial\mathcal{O}}(x_i)\right) + O\left(\sqrt{\Delta t}\right). \quad (3.44)$$

Since, for all $s \in \mathcal{I}$, $\tilde{d}^s_{k,i}(a,b) = O(\sqrt{\Delta t})$, from (3.42)-(3.44) we obtain

$$\begin{aligned}\mathcal{S}_{k,i}[W_{k+1}|_{\mathcal{G}_{\Delta x}}](a,b) - W_k(x_i) &= -M\Delta t \\ &\quad - \frac{1}{r}\sum_{s\in\mathcal{I}} d\left(x_i, \partial\mathcal{O}_\eta\right)\tilde{d}^s_{k,i}(a,b)\langle n(p_{\partial\mathcal{O}}(x_i)), \gamma_b\left(p_{\partial\mathcal{O}}(x_i)\right)\rangle \\ &\quad + O\left(\Delta t + (\Delta x)^2\right). \quad (3.45)\end{aligned}$$

Since $x_i \in \Gamma_k(a)$ there exists $\widetilde{\mathcal{I}}_{k,i} \subset \mathcal{I} \neq \emptyset$ such that $\tilde{d}^s_{k,i}(a,b) > 0$ for any $s \in \widetilde{\mathcal{I}}_{k,i}$. In addition, (3.41) implies that $d\left(x_i, \partial\mathcal{O}_\eta\right) \geq \eta - \delta > 0$. Thus, assumption **(H3)** implies that

$$\mathcal{S}_{k,i}[W_{k+1}|_{\mathcal{G}_{\Delta x}}](a,b) - W_k(x_i) \leq -M\Delta t - \frac{\sigma^2(\eta-\delta)}{2r}\sum_{s\in\widetilde{\mathcal{I}}_{k,i}} \tilde{d}^s_{k,i}(a,b) + O\left(\Delta t + (\Delta x)^2\right),$$

and hence (3.21) yields the existence of $C > 0$, independent of $k \in \mathcal{I}^*_{\Delta t}$, $i \in \mathcal{I}_{\Delta x}$, $a \in A$, and $b \in B$, such that

$$\mathcal{S}_{k,i}[W_{k+1}|_{\mathcal{G}_{\Delta x}}](a,b) - W_k(x_i) \leq -M\Delta t - C\sqrt{\Delta t} + O\left(\Delta t + (\Delta x)^2\right). \quad (3.46)$$

As long as $(\Delta x)^2/\Delta t$ is bounded, we have that $O\left(\Delta t + (\Delta x)^2\right) = O(\Delta t)$. Thus, from cases (i)-(iii) we can choose $M$ large enough such that

$$\mathcal{S}_{k,i}[W_{k+1}|_{\mathcal{G}_{\Delta x}}](a,b) - W_k(x_i) \leq -C\sqrt{\Delta t}\mathbb{I}_{\Gamma_k(a)}(x_i). \quad (3.47)$$

Now, set $q_k(x_i, a, b) = W_k(x_i) - \mathcal{S}_{k,i}[W_{k+1}|_{\mathcal{G}_{\Delta x}}](a,b)$. Then the probabilistic interpretation of the operator $\mathcal{S}_{k,i}$ (see Section 3.2.3) implies that, for any policy $\pi \in \Pi_{N_{\Delta t}}$,

$$W_k(x_i) = \mathbb{E}_{\mathbb{P}^{k,x_i,\pi}}\left(\sum_{m=k}^{N_T-1} q_m(X_m, \alpha_m, \beta_m) + w(X_{N_T})\right)$$

Since (3.47) implies that $q_k(x_i, a, b) \geq C\sqrt{\Delta t}\mathbb{I}_{\Gamma_k(a)}(x_i)$ for $k \in \mathcal{I}^*_{\Delta t}$, $i \in \mathcal{I}_{\Delta x}$, $a \in A$ and $b \in B$, we deduce that for any policy $\pi \in \Pi_{N_{\Delta t}}$ we have

$$\begin{aligned}\mathbb{E}_{\mathbb{P}^{k,x_i,\pi}}\left(\textstyle\sum_{m=k}^{N_T-1}\mathbb{I}_{\Gamma_m(\alpha_m)}(X_m)\right) &\leq \frac{1}{C\sqrt{\Delta t}}\mathbb{E}_{\mathbb{P}^{k,x_i,\pi}}\left(\sum_{m=k}^{N_T-1} q_m(X_m, \alpha_m, \beta_m)\right) \\ &= \frac{W_k(x_i) - \mathbb{E}_{\mathbb{P}^{k,x_i,\pi}}\left(w(X_{N_T})\right)}{C\sqrt{\Delta t}}.\end{aligned}$$

Finally, using that $W_k$ and $w$ are bounded, (3.35) follows. □

**Proposition 72.** (Stability) *The fully discrete scheme* ($\mathrm{HJB}_{\mathrm{disc}}$) *is stable, i.e. there exists* $C > 0$ *such that*

$$\max_{k \in \mathcal{I}^*_{\Delta t},\, i \in \mathcal{I}_{\Delta x}} |u_{k,i}| \leq C, \tag{3.48}$$

*where* $C$ *is independent of* $(\Delta t, \Delta x)$ *as long as* $\Delta t$ *is small enough and* $(\Delta x)^2/\Delta t$ *is bounded.*

*Proof.* Let us fix $k \in \mathcal{I}^*_{\Delta t}$ and $i \in \mathcal{I}_{\Delta x}$. Then the probabilistic interpretation of the scheme in Section 3.2.3 and the definition of $h$ in (3.29) imply the existence of a constant $C > 0$ such that

$$\begin{aligned} |u_{k,i}| &\leq \sup_{\pi \in \Pi_{N_{\Delta t}}} \mathbb{E}_{\mathbb{P}^{k,x_i,\pi}} \left( \sum_{m=k}^{N_{\Delta t}-1} \left[ \Delta t \left| f(t_m, X_m, \alpha_m) \right| \right. \right. \\ &\qquad \left. \left. + \left| h(t_m, X_m, \alpha_m, \beta_m) \right| \right] + \left| \Psi(X_{N_{\Delta t}}) \right| \right) \\ &\leq \|\Psi\|_\infty + T\|f\|_\infty + C\sqrt{\Delta t}\|g\|_\infty \sup_{\pi \in \Pi_{N_{\Delta t}}} \mathbb{E}_{\mathbb{P}^{k,x_i,\pi}} \left( \sum_{m=k}^{N_{\Delta t}-1} \mathbb{I}_{\Gamma_m(\alpha_m)}(X_m) \right). \end{aligned} \tag{3.49}$$

Thus, (3.48) follows from Lemma 71. □

## 3.4 Convergence analysis

In this section we provide the main result of this article which is the convergence of solutions to ($\mathrm{HJB}_{\mathrm{disc}}$) to the unique viscosity solution to (HJB). The proof is based on the half-relaxed limits technique introduced in [9] and the properties of solutions to ($\mathrm{HJB}_{\mathrm{disc}}$) investigated in Section 3.3.

Let $\Delta t > 0$, let $\Delta x > 0$ and let $(U_k)_{k=0}^{N_{\Delta t}}$ be the solution to ($\mathrm{HJB}_{\mathrm{disc}}$) associated to the discretization parameters $\Delta t$ and $\Delta x$. Let us define an extension of $(U_k)_{k=0}^{N_{\Delta t}}$ to $\overline{\mathcal{O}}_T$ by

$$(\forall\, (t,x) \in \overline{\mathcal{O}}_T) \quad u_{\Delta t, \Delta x}(t,x) := I[U_{\lfloor t/\Delta t \rfloor}](x), \tag{3.50}$$

where we recall that the interpolation operator $I[\cdot]$ is defined in (3.16). Now, let $(\Delta t_n, \Delta x_n)_{n \in \mathbb{N}} \subseteq (0, +\infty)^2$ be such that $\lim_{n\to\infty}(\Delta t_n, \Delta x_n) = (0,0)$ and the sequence $(\Delta x_n/\Delta t_n)_{n\in\mathbb{N}}$ is bounded. For every $(t,x) \in \overline{\mathcal{O}}_T$, let us define

$$\begin{aligned} \overline{u}(t,x) &:= \limsup_{\substack{n\to\infty \\ \overline{\mathcal{O}}_T \ni (s_n,y_n) \to (t,x)}} u_{\Delta t_n, \Delta x_n}(s_n, y_n), \\ \underline{u}(t,x) &:= \liminf_{\substack{n\to\infty \\ \overline{\mathcal{O}}_T \ni (s_n,y_n) \to (t,x)}} u_{\Delta t_n, \Delta x_n}(s_n, y_n). \end{aligned} \tag{3.51}$$

From Proposition 72 we deduce that $\overline{u}\colon \overline{\mathcal{O}}_T \to \mathbb{R}$ and $\underline{u}\colon \overline{\mathcal{O}}_T \to \mathbb{R}$ are well-defined and bounded. Moreover, from [5, Chapter V, Lemma 1.5], we have that $\overline{u}$ and $\underline{u}$ are, respectively, upper and lower semicontinuous functions.

**Proposition 73.** *Assume that* $(\Delta x_n)^2/\Delta t_n \to 0$*, as* $n \to \infty$*. Then* $\overline{u}$ *and* $\underline{u}$ *are, respectively, viscosity sub- and supersolutions to* (HJB)*.*

*Proof.* We only show that $\overline{u}$ is a viscosity subsolution to (HJB), the proof that $\underline{u}$ is a viscosity supersolution being similar. Let $(\bar{t}, \bar{x}) \in \overline{\mathcal{O}}_T$ and $\phi \in C^\infty(\overline{\mathcal{O}}_T)$ be such that $\overline{u}(\bar{t}, \bar{x}) = \phi(\bar{t}, \bar{x})$ and $\overline{u} - \phi$ has a maximum at $(\bar{t}, \bar{x})$. Then by [5, Chapter V, Lemma 1.6] there exists a subsequence of $(u_{\Delta t_n, \Delta x_n})_{n\in\mathbb{N}}$, which for simplicity is still labeled by $n \in \mathbb{N}$, and a sequence $(s_n, y_n)_{n\in\mathbb{N}} \subseteq \overline{\mathcal{O}}_T$ such that $(u_{\Delta t_n, \Delta x_n})_{n\in\mathbb{N}}$ is uniformly bounded, $u_{\Delta t_n, \Delta x_n} - \phi$ has a local maximum at $(s_n, y_n)$, and, as $n \to \infty$, $(s_n, y_n) \to (\bar{t}, \bar{x})$ and $u_{\Delta t_n, \Delta x_n}(s_n, y_n) \to \overline{u}(\bar{t}, \bar{x})$. Moreover, by modifying the test function $\phi$, we can assume that $u_{\Delta t_n, \Delta x_n} - \phi$ has a global maximum at $(s_n, y_n)$, i.e. setting $\xi_n := u_{\Delta t_n, \Delta x_n}(s_n, y_n) - \phi(s_n, y_n)$, we have

$$(\forall\ (t, x) \in \overline{\mathcal{O}}_T) \quad u_{\Delta t_n, \Delta x_n}(t, x) \le \phi(t, x) + \xi_n, \qquad \text{with } \xi_n \to 0. \tag{3.52}$$

We distinguish now the following cases.

**(i)** $(\bar{t}, \bar{x}) \in [0, T) \times \mathcal{O}$. In this case, for all $n$ large enough, by (3.15), we have $y_n \in \mathcal{O}_{\Delta x_n}$. Let $k : \mathbb{N} \to \mathcal{I}^*_{\Delta t_n}$ be such that $s_n \in [t_{k(n)}, t_{k(n)+1})$. As $n \to \infty$, we have $t_{k(n)} \to \bar{t}$ and, from (3.50) and (3.52), with $t = t_{k(n)+1}$, we have

$$(\forall\ x \in \overline{\mathcal{O}}) \quad I[U_{k(n)+1}](x) \le \phi(t_{k(n)+1}, x) + \xi_n. \tag{3.53}$$

From Proposition 69, we obtain

$$(\forall\ i \in \mathcal{I}_{\Delta x}) \quad S_{k_n,i}[U_{k(n)+1}] \le S_{k_n,i}[\Phi_{k(n)+1}] + \xi_n, \tag{3.54}$$

where, for all $k \in \mathcal{I}_{\Delta t}$, we have denoted $\Phi_k := \phi(t_k, \cdot)|_{\mathcal{G}_{\Delta x_n}}$. In particular, by $(\mathrm{HJB}_{\mathrm{disc}})$ we get

$$(\forall\ i \in \mathcal{I}_{\Delta x}) \quad U_{k(n),i} \le S_{k_n,i}[\Phi_{k(n)+1}] + \xi_n. \tag{3.55}$$

The monotonicity of the interpolation operator (3.16) yields

$$\left(\forall\ x \in \overline{\mathcal{O}}\right) \quad u_{\Delta t_n, \Delta x_n}(s_n, x) \le \sum_{i \in \mathcal{I}_{\Delta x_n}} \beta^1_i(p_{\Delta x_n}(x)) S_{k_n,i}[\Phi_{k(n)+1}] + \xi_n, \tag{3.56}$$

and, hence, by taking $x = y_n$ and using the definition of $\xi_n$, we get

$$\phi(s_n, y_n) \le \sum_{i \in \mathcal{I}_{\Delta x_n}} \beta^1_i(y_n) S_{k_n,i}[\Phi_{k(n)+1}]. \tag{3.57}$$

Since $(\bar{t}, \bar{x}) \in [0, T) \times \mathcal{O}$ and $A, B$ are compacts, if $n$ large enough, for all $a \in A, b \in B$ and for all $s \in \mathcal{I}$ we have $\tilde{d}^s_{k_n,i}(a, b) = 0$ for all $i \in \mathcal{I}_{\Delta x}$ such that $\beta^1_i(y_n) > 0$. Using Proposition 70(ii) and inequality (3.57), we get

$$\begin{aligned}
\phi(s_n, y_n) \quad \le \sum_{i \in \mathcal{I}_{\Delta x_n}} \beta^1_i(y_n) \Big[ & \phi(t_{k(n)+1}, x_i) \\
& - \Delta t_n \sup_{a \in A} \mathcal{H}\left(t_{k(n)}, x_i, D\phi(t_{k(n)+1}, x_i), D^2\phi(t_{k(n)+1}, x_i), a\right) \Big] \\
& + O\left(\Delta t_n \sqrt{\Delta t_n} + (\Delta x_n)^2\right).
\end{aligned}$$

Then following the same arguments than those in [30, Theorem 3.1] (see also [49, Theorem 4.22]) we conclude that

$$-\partial_t\phi(\bar{t},\bar{x}) + H(\bar{t},\bar{x},D\phi(\bar{t},\bar{x}),D^2\phi(\bar{t},\bar{x})) \leq 0, \tag{3.58}$$

and, hence, (3.5) holds.

**(ii)** $\underline{(\bar{t},\bar{x}) \in [0,T)\times\partial\mathcal{O}}$. If

$$L(\bar{t},\bar{x},D\phi(\bar{t},\bar{x})) \leq 0 \quad \text{or} \quad -\partial_t\phi(\bar{t},\bar{x}) + H(\bar{t},\bar{x},D\phi(\bar{t},\bar{x}),D^2\phi(\bar{t},\bar{x})) \leq 0,$$

holds, then (3.6) holds. Thus, let us suppose that

$$L(\bar{t},\bar{x},D\phi(\bar{t},\bar{x})) > 0 \quad \text{and} \quad -\partial_t\phi(\bar{t},\bar{x}) + H(\bar{t},\bar{x},D\phi(\bar{t},\bar{x}),D^2\phi(\bar{t},\bar{x})) > 0. \tag{3.59}$$

Letting $k:\mathbb{N}\to\{0,\dots,N_T-1\}$ as in **(i)**, we have $t_{k(n)}\to\bar{t}$, (3.56) holds true, and, hence,

$$\phi(s_n,y_n) \leq \sum_{i\in\mathcal{I}_{\Delta x_n}} \beta_i^1(p_{\Delta x_n}(y_n))S_{k_n,i}[\Phi_{k(n)+1}]. \tag{3.60}$$

On the one hand, from Proposition 70(ii) we get

$$\begin{aligned}
0 \;\leq\; & \sum_{i\in\mathcal{I}_{\Delta x_n}} \beta_i^1(p_{\Delta x_n}(y_n))\Big(\Delta t_n\partial_t\phi(t_{k(n)},x_i) \\
& -\sup_{\substack{a\in A,\\ b\in B}}\Big\{\Delta t_n\mathcal{H}(t_{k(n)},x_i,D\phi(t_{k(n)+1},x_i),D^2\phi(t_{k(n)+1},x_i),a) \\
& +\tfrac{1}{2r}\sum_{s\in\mathcal{I}}\tilde{d}^s_{k,i}(a,b)\left(\tilde{\mathcal{L}}^s_{k(n),i}(D\phi(t_{k(n)+1},x_i),a,b) - \sqrt{\Delta t}_n K^s_{k(n),i}(a,b)\right)\Big\} \\
& +O\left(\Delta t_n\sqrt{\Delta t}_n + (\Delta x_n)^2\right).
\end{aligned}$$

Therefore, for all $a\in A$ and $b\in B$, we have

$$\begin{aligned}
& \sum_{i\in\mathcal{I}_{\Delta x_n}} \beta_i^1(p_{\Delta x_n}(y_n))\Big\{ -\Delta t_n\partial_t\phi(t_{k(n)},x_i) \\
& +\Delta t_n\mathcal{H}(t_{k(n)},x_i,D\phi(t_{k(n)+1},x_i),D^2\phi(t_{k(n)+1},x_i),a) \\
& +\tfrac{1}{2r}\sum_{s\in\mathcal{I}}\tilde{d}^s_{k,i}(a,b)\left(\tilde{\mathcal{L}}^s_{k(n),i)}(D\phi(t_{k(n)+1},x_i),a,b) - \sqrt{\Delta t}_n K^s_{k(n),i}(a,b)\right)\Big\} \\
& +O\left(\Delta t_n\sqrt{\Delta t}_n + (\Delta x_n)^2\right) \leq 0.
\end{aligned} \tag{3.61}$$

On the other hand, since $A$ is compact, there exists $\bar{a}\in A$ such that

$$H(\bar{t},\bar{x},D\phi(\bar{t},\bar{x}),D^2\phi(\bar{t},\bar{x})) = \mathcal{H}(\bar{t},\bar{x},D\phi(\bar{t},\bar{x}),D^2\phi(\bar{t},\bar{x}),\bar{a})$$

and

$$\begin{aligned}
& \sum_{i\in\mathcal{I}_{\Delta x_n}} \beta_i^1(p_{\Delta x_n}(y_n))\left(-\partial_t\phi(t_{k(n)},x_i)\right. \\
& \left.+\mathcal{H}(t_{k(n)},x_i,D\phi(t_{k(n)+1},x_i),D^2\phi(t_{k(n)+1},x_i),\bar{a})\right) \\
& \to -\partial_t\phi(\bar{t},\bar{x}) + H(\bar{t},\bar{x},D\phi(\bar{t},\bar{x}),D^2\phi(\bar{t},\bar{x})), \quad \text{as } n\to\infty.
\end{aligned} \tag{3.62}$$

Let us set $\tilde{d}_n^* = \max\left\{\tilde{d}^s_{k_n,i}(\bar{a}) \mid s \in \mathcal{I},\ i \in \mathcal{I}_{\Delta x_n}\right\}$ and take $a = \bar{a}$ and an arbitrary $b \in B$ in (3.61). If there exists a subsequence, still labelled by $n$, such that $\tilde{d}_n^* = 0$, then dividing (3.61) by $\Delta t_n$, and letting $n \to \infty$, (3.62) yields

$$-\partial_t\phi(\bar{t},\bar{x}) + H(\bar{t},\bar{x},D\phi(\bar{t},\bar{x}),D^2\phi(\bar{t},\bar{x})) \le 0,$$

which contradicts (3.59). Otherwise, by (3.21), for all $n \in \mathbb{N}$, large enough, we have $\tilde{d}_n^* \ge \bar{c}\sqrt{\Delta t_n}$. Notice that the second relation in (3.59) and (3.62) imply that, for $n \in \mathbb{N}$ large enough,

$$\begin{aligned} 0 < \sum_{i\in\mathcal{I}_{\Delta x_n}} \beta_i^1(p_{\Delta x_n}(y_n))\Big(&-\partial_t\phi(t_{k(n)},x_i) \\ &+\mathcal{H}(t_{k(n)},x_i,D\phi(t_{k(n)+1},x_i),D^2\phi(t_{k(n)+1},x_i),\bar{a})\Big). \end{aligned} \tag{3.63}$$

Therefore, inequality (3.61) with $a = \bar{a}$ implies that for all $b \in B$

$$\begin{aligned} \sum_{i\in\mathcal{I}_{\Delta x_n}} \beta_i^1(p_{\Delta x_n}(y_n))\Big\{\sum_{s\in\mathcal{I}} \tilde{d}^s_{k_n,i}(\bar{a},b)\Big(&\tilde{\mathcal{L}}^s_{k_n,i}(D\phi(t_{k(n)+1},x_i),\bar{a},b) \\ &-\sqrt{\Delta t}_n K^s_{k(n),i}(\bar{a},b)\Big)\Big\} + O\left(\Delta t_n\sqrt{\Delta t}_n + (\Delta x_n)^2\right) < 0. \end{aligned} \tag{3.64}$$

Since the set $\mathcal{I} = \{+,-\}\times\{1,\dots,d\}$ is finite, there exist $\hat{s} \in \mathcal{I}$, $\{q^s \mid s \in \mathcal{I}\setminus\{\hat{s}\}\} \subseteq [0,1]$, and $i(n) \in \mathcal{I}_{\Delta x_n}$ such that, up to some subsequence, $\tilde{d}_n^* = \tilde{d}^{\hat{s}}_{k(n),i(n)}(\bar{a})$ and, for all $s \in \mathcal{I}\setminus\{\hat{s}\}$, $\tilde{d}^s_{k(n),i(n)}(\bar{a})/\tilde{d}_n^* \to q^s$. Recall that $\tilde{d}_n^* \ge \bar{c}\sqrt{\Delta t_n}$ and $(\Delta x_n)^2/\Delta t_n \to 0$ as $n \to \infty$. Dividing (3.64) by $\tilde{d}_n^*$ and taking the limit $n \to \infty$ yields

$$(\forall\, b \in B) \quad \left(\sum_{s\in\mathcal{I}\setminus\{\hat{s}\}} q^s + 1\right)\mathcal{L}(\bar{t},\bar{x},D\phi(\bar{t},\bar{x}),b) \le 0$$

and hence

$$(\forall\, b \in B) \quad \mathcal{L}(\bar{t},\bar{x},D\phi(\bar{t},\bar{x}),b) \le 0.$$

Thus, $L(\bar{t},\bar{x},D\phi(\bar{t},\bar{x})) \le 0$, which contradicts (3.59).

**(iii)** $(\bar{t},\bar{x}) \in \{T\}\times\overline{\mathcal{O}}$. Let us first assume that $(\bar{t},\bar{x}) \in \{T\}\times\mathcal{O}$. Thus, for $n \in \mathbb{N}$ large enough, we have $y_n \in \mathcal{O}$. By taking a subsequence, if necessary, it suffices to consider the cases $s_n \in [0,T)$, for all $n \in \mathbb{N}$, and $s_n = T$ for all $n \in \mathbb{N}$. In the first case, proceeding as in **(i)**, we get

$$-\partial_t\phi(\bar{t},\bar{x}) + H(\bar{t},\bar{x},D\phi(\bar{t},\bar{x}),D^2\phi(\bar{t},\bar{x})) \le 0. \tag{3.65}$$

In the second case, (3.50) implies that $u_{\Delta t_n,\Delta x_n}(s_n,y_n) = I[\Psi|_{\mathcal{G}_{\Delta x}}](y_n)$ and hence letting $n \to \infty$ we get

$$\overline{u}(\bar{t},\bar{x}) = \Psi(\bar{x}). \tag{3.66}$$

Now, assume that $(\bar{t},\bar{x}) \in \{T\}\times\partial\mathcal{O}$. As before, it suffices to consider the cases $s_n \in [0,T)$, for all $n \in \mathbb{N}$, and $s_n = T$ for all $n \in \mathbb{N}$. If $s_n \in [0,T)$, then, proceeding as in **(ii)**, we get

$$L(\bar{t},\bar{x},D\phi(\bar{t},\bar{x})) \le 0 \quad \text{or} \quad -\partial_t\phi(\bar{t},\bar{x}) + H(\bar{t},\bar{x},D\phi(\bar{t},\bar{x}),D^2\phi(\bar{t},\bar{x})) \le 0. \tag{3.67}$$

Finally, if $s_n = T$ for all $n \in \mathbb{N}$, we have $u_{\Delta t_n, \Delta x_n}(s_n, y_n) = I[\Psi|_{\mathcal{G}_{\Delta x}}](y_n)$ and hence (3.66) holds.

Altogether, (3.65) and (3.66) imply that (3.11) holds if $(\bar{t}, \bar{x}) \in \{T\} \times \mathcal{O}$, and (3.67) and (3.66) imply that (3.12) holds if $(\bar{t}, \bar{x}) \in \{T\} \times \partial\mathcal{O}$.

Thus, from cases **(i)-(iii)** and Remark 64 we obtain that $\overline{u}$ is a subsolution to (HJB). □

**Theorem 74.** *Assume* **(H1)-(H3)** *and that* $(\Delta x_n)^2/\Delta t_n \to 0$, *as* $n \to \infty$. *Then*

$$u_{\Delta t_n, \Delta x_n} \to u \quad \textit{uniformly in } \overline{\mathcal{O}}_T,$$

*where* $u$ *is the unique continuous viscosity solution to* (HJB).

*Proof.* By (3.51) we have $\underline{u} \leq \overline{u}$ in $\overline{\mathcal{O}}_T$ and, by Proposition 73 and the comparison principle for sub- and super solutions to (HJB) (see Remark 66(i)), we obtain that $\underline{u} \geq \overline{u}$ in $\overline{\mathcal{O}}_T$. Thus, $u = \underline{u} = \overline{u}$ and the result follows from [5, Chapter V, Lemma 1.9]. □

## 3.5 Numerical results

In this section, we present some numerical experiments in order to show the performance of the scheme. We consider first a one-dimensional linear parabolic equation, with homogeneous Neumann boundary conditions, and both the first and second order cases. In the former, the boundary conditions are not satisfied in the pointwise sense at every point in the boundary, but they hold in the viscosity sense (see Definition 63). The second example deals with a degenerate second order nonlinear equation on a smooth two-dimensional domain. We consider both non-homogeneous Neumann and oblique boundary conditions. In the last example, we approximate the solution to a non-degenerate second order nonlinear equation with mixed Dirichlet and homogeneous Neumann boundary conditions on a non-smooth domain. Because of the presence of Dirichlet boundary conditions and corners, the scheme has to be modified and the convergence result in Section 3.3 does not apply. However, the scheme can be successfully applied to solve numerically the problem.

The problems in the first two tests have known analytic solutions. This will allow to compute the errors of solutions to the scheme and to perform a numerical convergence analysis. In the examples dealing with two-dimensional domains, we have considered unstructured triangular meshes, constructed with the Matlab2019 function `initmesh`.

In the simulations we have chosen time and space steps satisfying $\Delta t = \Delta x$ or $\Delta t = \Delta x/2$, which are in agreement with the assumption in Theorem 74.

### 3.5.1 One-dimensional linear problem

Let $\varepsilon > 0$, set $\lambda_\varepsilon^\pm = (1 \pm \sqrt{1+4\varepsilon})/2\varepsilon$, and define

$$f(t,x) = \frac{3-t}{2}\left(1 + \frac{e^{\lambda_\varepsilon^+ x}\left(e^{\lambda_\varepsilon^-}-1\right)}{e^{\lambda_\varepsilon^+}-e^{\lambda_\varepsilon^-}}\left(1-\varepsilon\lambda_\varepsilon^+\right) + \frac{e^{\lambda_\varepsilon^- x}\left(1-e^{\lambda_\varepsilon^+}\right)}{e^{\lambda_\varepsilon^+}-e^{\lambda_\varepsilon^-}}\left(1-\varepsilon\lambda_\varepsilon^-\right)\right)$$
$$+\frac{1}{2}\left(x + \frac{e^{\lambda_\varepsilon^+ x}\left(e^{\lambda_\varepsilon^-}-1\right)}{e^{\lambda_\varepsilon^+}-e^{\lambda_\varepsilon^-}} + \frac{e^{\lambda_\varepsilon^- x}\left(1-e^{\lambda_\varepsilon^+}\right)}{e^{\lambda_\varepsilon^+}-e^{\lambda_\varepsilon^-}}\right),$$
$$u_\varepsilon(t,x) = \frac{3-t}{2}\left(x + \frac{e^{\lambda_\varepsilon^-}-1}{\lambda_\varepsilon^+\left(e^{\lambda_\varepsilon^+}-e^{\lambda_\varepsilon^-}\right)}e^{\lambda_\varepsilon^+ x} + \frac{1-e^{\lambda_\varepsilon^+}}{\lambda_\varepsilon^-\left(e^{\lambda_\varepsilon^+}-e^{\lambda_\varepsilon^-}\right)}e^{\lambda_\varepsilon^- x}\right),$$

for $(t,x) \in [0,1]^2$. Then $u_\varepsilon$ is the unique classical solution to

$$\begin{aligned} -\partial_t u - \varepsilon\partial_x^2 u + \partial_x u &= f \quad \text{in } [0,1)\times(0,1),\\ \partial_x u(\cdot,0) = \partial_x u(\cdot,1) &= 0 \quad \text{in } [0,1),\\ u(1,\cdot) &= u_\varepsilon(1,\cdot) \quad \text{in } [0,1]. \end{aligned} \tag{3.68}$$

Similarly to [43, Example 7.3], we have

$$u_\varepsilon(t,x) \underset{\varepsilon\to 0}{\longrightarrow} u_0(t,x) := \frac{3-t}{2}\left(x + e^{-x}\right), \quad \text{uniformly on } [0,1]^2$$

and $u_0$ is the unique viscosity solution to

$$\begin{aligned} -\partial_t u + \partial_x u &= f \quad \text{in } [0,1)\times(0,1),\\ \partial_x u(\cdot,0) = \partial_x u(\cdot,1) &= 0 \quad \text{in } [0,1),\\ u(1,\cdot) &= u_0(1,\cdot) \quad \text{in } [0,1]. \end{aligned} \tag{3.69}$$

Notice that for $t \in [0,1]$ we have $-\partial_t u(t,1) + \partial_x u(t,1) - f(t,1) \le 0$ and $\partial_x u(t,1) > 0$. Thus, at $(t,1)$ the boundary condition is satisfied in the viscosity sense but not in the pointwise sense.

Using ($\mathrm{HJB}_{\mathrm{disc}}$), we approximate $u_\varepsilon$ for $\varepsilon = 0.05$, $\varepsilon = 0.03$, and $\varepsilon = 0$. For these choices, we plot in Figure 3.2 respectively the approximations of $u_\varepsilon(1,\cdot)$ and $u_\varepsilon(0,\cdot)$, computed with the steps sizes $\Delta x = 3.125\cdot 10^{-3}$ and $\Delta t = \Delta x/2$.

We show in Tables 1 and 2 the errors

$$E_\infty = \max_{i\in\mathcal{I}_{\Delta x}} |U_{0,i} - u(0,x_i)|, \quad E_1 = \Delta x \sum_{i\in\mathcal{I}_{\Delta x}} |U_{0,i} - u(0,x_i)|,$$

and the corresponding convergence rates $p_\infty$ and $p_1$, for $\varepsilon = 0.05$ and $\varepsilon = 0$, respectively. In all cases, an order of convergence close to 1 is obtained.

In the simulations, we have chosen $\bar{c} := 0.025 + \sigma/2$, where $\sigma = \sqrt{2\varepsilon}$ is the diffusion parameter. With this choice, the larger the value of $\sigma$, the more the characteristics are reflected further into $\mathcal{O}$.

| $\Delta x$ | $\Delta t = \Delta x$ | | | |
|---|---|---|---|---|
| | $E_\infty$ | $E_1$ | $p_\infty$ | $p_1$ |
| $5.00 \cdot 10^{-2}$ | $3.99 \cdot 10^{-2}$ | $2.57 \cdot 10^{-2}$ | - | - |
| $2.50 \cdot 10^{-2}$ | $2.25 \cdot 10^{-2}$ | $1.06 \cdot 10^{-2}$ | 0.83 | 1.28 |
| $1.25 \cdot 10^{-2}$ | $1.17 \cdot 10^{-2}$ | $6.13 \cdot 10^{-3}$ | 0.94 | 0.79 |
| $6.25 \cdot 10^{-3}$ | $5.38 \cdot 10^{-3}$ | $2.49 \cdot 10^{-3}$ | 1.12 | 1.30 |
| $3.125 \cdot 10^{-3}$ | $2.15 \cdot 10^{-3}$ | $1.77 \cdot 10^{-3}$ | 1.32 | 0.49 |
| $\Delta x$ | $\Delta t = \Delta x/2$ | | | |
| | $E_\infty$ | $E_1$ | $p_\infty$ | $p_1$ |
| $5.00 \cdot 10^{-2}$ | $2.16 \cdot 10^{-2}$ | $2.03 \cdot 10^{-2}$ | - | - |
| $2.50 \cdot 10^{-2}$ | $1.26 \cdot 10^{-2}$ | $6.22 \cdot 10^{-3}$ | 0.78 | 1.71 |
| $1.25 \cdot 10^{-2}$ | $5.87 \cdot 10^{-3}$ | $5.64 \cdot 10^{-3}$ | 1.10 | 0.14 |
| $6.25 \cdot 10^{-3}$ | $3.17 \cdot 10^{-3}$ | $2.95 \cdot 10^{-3}$ | 0.89 | 0.93 |
| $3.125 \cdot 10^{-3}$ | $1.62 \cdot 10^{-3}$ | $1.50 \cdot 10^{-3}$ | 0.97 | 0.98 |

**Table 3.1.** Errors and convergence rates for problem (3.68) with $\varepsilon = 0.05$.

| $\Delta x$ | $\Delta t = \Delta x$ | | | |
|---|---|---|---|---|
| | $E_\infty$ | $E_1$ | $p_\infty$ | $p_1$ |
| $5.00 \cdot 10^{-2}$ | $2.83 \cdot 10^{-2}$ | $1.95 \cdot 10^{-2}$ | - | - |
| $2.50 \cdot 10^{-2}$ | $1.42 \cdot 10^{-2}$ | $1.01 \cdot 10^{-2}$ | 0.99 | 0.95 |
| $1.25 \cdot 10^{-2}$ | $7.08 \cdot 10^{-3}$ | $5.39 \cdot 10^{-3}$ | 1.00 | 0.91 |
| $6.25 \cdot 10^{-3}$ | $3.54 \cdot 10^{-3}$ | $2.91 \cdot 10^{-3}$ | 1.00 | 0.89 |
| $3.125 \cdot 10^{-3}$ | $1.77 \cdot 10^{-3}$ | $1.59 \cdot 10^{-3}$ | 1.00 | 0.87 |
| $\Delta x$ | $\Delta t = \Delta x/2$ | | | |
| | $E_\infty$ | $E_1$ | $p_\infty$ | $p_1$ |
| $5.00 \cdot 10^{-2}$ | $2.26 \cdot 10^{-2}$ | $1.86 \cdot 10^{-2}$ | - | - |
| $2.50 \cdot 10^{-2}$ | $1.15 \cdot 10^{-2}$ | $9.97 \cdot 10^{-3}$ | 0.97 | 0.90 |
| $1.25 \cdot 10^{-2}$ | $5.88 \cdot 10^{-3}$ | $5.42 \cdot 10^{-3}$ | 0.97 | 0.88 |
| $6.25 \cdot 10^{-3}$ | $3.04 \cdot 10^{-3}$ | $2.97 \cdot 10^{-3}$ | 0.95 | 0.87 |
| $3.125 \cdot 10^{-3}$ | $1.68 \cdot 10^{-3}$ | $1.63 \cdot 10^{-3}$ | 0.86 | 0.87 |

**Table 3.2.** Errors and convergence rates for problem (3.68) with $\varepsilon = 0$.

### 3.5.2 Nonlinear problem on a circular domain

Let $T = 1$, $\mathcal{O} = \{x = (x_1, x_2) \in \mathbb{R}^2 \mid |x| < 1\}$, $\sigma(t,x) = \sqrt{2}(\sin(x_1+x_2), \cos(x_1+x_2))$, and

$$\begin{aligned} f(t,x) &= \left(\tfrac{1}{2} - t\right)\sin(x_1)\sin(x_2) \\ &\quad + \left(\tfrac{3}{2} - t\right)\Big(\sqrt{\cos^2(x_1)\sin^2(x_2) + \sin^2(x_1)\cos^2(x_2)} \\ &\quad -2\sin(x_1+x_2)\cos(x_1+x_2)\cos(x_1)\cos(x_2)\Big), \\ g(t,x) &= \left(\tfrac{3}{2} - t\right)\left(x_1\cos(x_1)\sin(x_2) + x_2\sin(x_1)\cos(x_2)\right). \end{aligned}$$

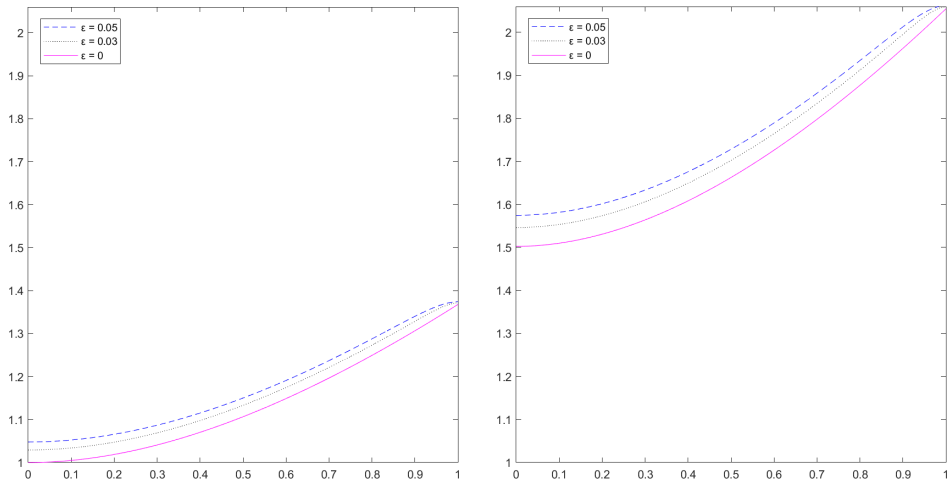



**Figure 3.2.** Exact final condition $u_\varepsilon(1,\cdot)$ (left) and numerical approximations of $u_\varepsilon(0,\cdot)$ (right) for $\varepsilon = 0.05$, $\varepsilon = 0.03$, and $\varepsilon = 0$, with step sizes $\Delta x = 6.25 \times 10^{-3}$ and $\Delta t = \Delta x/2$.

Then $\overline{\mathcal{O}}_T \ni (t, x_1, x_2) \mapsto \bar{u}(t, x_1, x_2) = \left(\frac{3}{2} - t\right) \sin(x_1) \sin(x_2)$ is the unique classical solution to

$$\begin{aligned} \partial_t u - \tfrac{1}{2}\mathrm{Tr}(\sigma\sigma^\top D^2 u) + |Du| &= f \quad \text{in } \mathcal{O}_T, \\ \langle n, Du \rangle &= g \quad \text{in } [0,T) \times \partial\mathcal{O}, \\ u(0,x) &= \bar{u}(0,x) \quad \text{in } x \in \overline{\mathcal{O}}. \end{aligned} \tag{3.70}$$

In Figure 3.3, we show the numerical solution at the final time $T = 1$ computed on an unstructured triangular mesh $\mathcal{G}_{\Delta x}$ with mesh size $\Delta x = 1.25 \cdot 10^{-1}$. On the left, we plot the result together with the contour lines. On the right, we plot the approximation together with the mesh used to compute it.

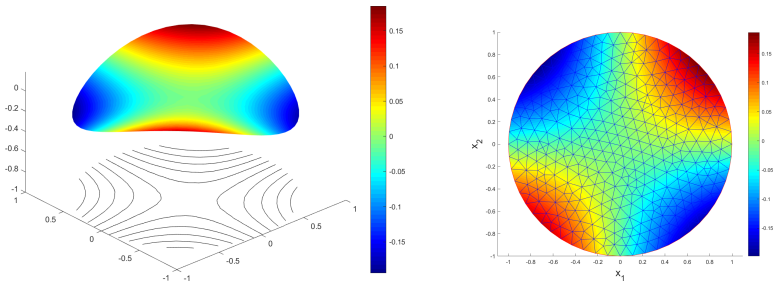


**Figure 3.3.** Numerical solution at time $T = 1$ of problem in subsect.3.5.2 with Neumann boundary condition, computed with $\Delta x = 0.125$ and $\Delta t = \Delta x/2$.

Given an element $\hat{T}$ of the triangulation, we denote by $x_{\hat{T}}$ its barycenter and by $|\hat{T}|$ its area. We show in Tables 3 and 4 the errors

$$E_\infty = \max_{i \in \mathcal{I}_{\Delta x}} |U_{N_T,i} - \bar{u}(t_{N_T}, x_i)|, \quad E_1 = \sum_{\hat{T} \in \mathcal{T}_{\Delta x}} |\hat{T}| |I[U_{N_T,(\cdot)}](x_{\hat{T}}) - \bar{u}(t_{N_T}, x_{\hat{T}})|, \tag{3.71}$$

and the corresponding convergence rates $p_\infty$ and $p_1$. In each table, we specify in the first column the mesh size $\Delta x$. To obtain the results shown in Tables 3 and 4, we have chosen $\bar{c}$ in (3.20) and (3.21) as $\bar{c} = 0.25$ and $\bar{c} = 0.5$, respectively. For both choices of $\bar{c}$, we observe similar errors and an analogue behavior of the convergence rates. As in the previous example, an order of convergence close to 1 is obtained.

| $\Delta x$ | $\Delta t = \Delta x$ | | | |
|---|---|---|---|---|
| | $E_\infty$ | $E_1$ | $p_\infty$ | $p_1$ |
| $2.50 \cdot 10^{-1}$ | $2.73 \cdot 10^{-1}$ | $2.95 \cdot 10^{-1}$ | - | - |
| $1.25 \cdot 10^{-1}$ | $1.24 \cdot 10^{-1}$ | $1.12 \cdot 10^{-1}$ | 1.14 | 1.40 |
| $6.25 \cdot 10^{-2}$ | $5.55 \cdot 10^{-2}$ | $4.72 \cdot 10^{-2}$ | 1.16 | 1.24 |
| $3.125 \cdot 10^{-2}$ | $2.49 \cdot 10^{-2}$ | $2.16 \cdot 10^{-2}$ | 1.16 | 1.13 |
| $\Delta x$ | $\Delta t = \Delta x/2$ | | | |
| | $E_\infty$ | $E_1$ | $p_\infty$ | $p_1$ |
| $2.50 \cdot 10^{-1}$ | $1.22 \cdot 10^{-1}$ | $1.07 \cdot 10^{-1}$ | - | - |
| $1.25 \cdot 10^{-1}$ | $5.54 \cdot 10^{-2}$ | $4.57 \cdot 10^{-2}$ | 1.14 | 1.24 |
| $6.25 \cdot 10^{-2}$ | $2.39 \cdot 10^{-2}$ | $2.11 \cdot 10^{-2}$ | 1.21 | 1.11 |
| $3.125 \cdot 10^{-2}$ | $1.22 \cdot 10^{-2}$ | $1.10 \cdot 10^{-2}$ | 0.97 | 0.94 |

**Table 3.3.** Errors and convergence rates for the approximation of (3.70) with $\bar{c} = 0.25$.

| $\Delta x$ | $\Delta t = \Delta x$ | | | |
|---|---|---|---|---|
| | $E_\infty$ | $E_1$ | $p_\infty$ | $p_1$ |
| $2.50 \cdot 10^{-1}$ | $2.65 \cdot 10^{-1}$ | $2.55 \cdot 10^{-1}$ | - | - |
| $1.25 \cdot 10^{-1}$ | $1.23 \cdot 10^{-1}$ | $1.12 \cdot 10^{-1}$ | 1.11 | 1.19 |
| $6.25 \cdot 10^{-2}$ | $5.74 \cdot 10^{-2}$ | $5.06 \cdot 10^{-2}$ | 1.10 | 1.15 |
| $3.125 \cdot 10^{-2}$ | $2.70 \cdot 10^{-2}$ | $2.39 \cdot 10^{-2}$ | 1.09 | 1.08 |
| $\Delta x$ | $\Delta t = \Delta x/2$ | | | |
| | $E_\infty$ | $E_1$ | $p_\infty$ | $p_1$ |
| $2.50 \cdot 10^{-1}$ | $1.18 \cdot 10^{-1}$ | $1.02 \cdot 10^{-1}$ | - | - |
| $1.25 \cdot 10^{-1}$ | $5.60 \cdot 10^{-2}$ | $4.72 \cdot 10^{-2}$ | 1.08 | 1.11 |
| $6.25 \cdot 10^{-2}$ | $2.64 \cdot 10^{-2}$ | $2.27 \cdot 10^{-2}$ | 1.08 | 1.06 |
| $3.125 \cdot 10^{-2}$ | $1.22 \cdot 10^{-2}$ | $1.10 \cdot 10^{-2}$ | 1.11 | 1.05 |

**Table 3.4.** Errors and convergence rates for the approximation of (3.70) with $\bar{c} = 0.5$.

Next, we consider the same problem but with oblique boundary conditions. More precisely, for $x = (x_1, x_2) \in \partial\mathcal{O}$ we set

$$\gamma(x) = (x_1 \cos(\pi/6) + x_2 \sin(\pi/6), x_2 \cos(\pi/6) - x_1 \sin(\pi/6))$$

and

$$\begin{aligned} \tilde{g}(t,x) \;=\; & \left(\tfrac{3}{2} - t\right) \big[ \, (x_1 \cos(\pi/6) + x_2 \sin(\pi/6)) \cos(x_1) \sin(x_2) \\ & + (x_2 \cos(\pi/6) - x_1 \sin(\pi/6)) \sin(x_1) \cos(x_2) \big] \quad \text{in } [0,T) \times \partial\mathcal{O}. \end{aligned}$$

Then $\bar{u}$ is the unique classical solution to the Neumann boundary condition in (3.70) is changed to

$$\begin{aligned} \partial_t u - \tfrac{1}{2}\mathrm{Tr}(\sigma\sigma^\top D^2 u) + |Du| &= f \quad \text{in } \mathcal{O}_T, \\ \langle \gamma, Du \rangle &= \tilde{g} \quad \text{in } [0,T) \times \partial\mathcal{O}, \\ u(0,x) &= \bar{u}(0,x) \quad \text{in } x \in \overline{\mathcal{O}}. \end{aligned} \tag{3.72}$$

The solution $\bar{u}$ is approximated by using the same unstructured meshes as in the previous case. We show in Tables 3.5 and 3.6 the errors (3.71) computed with $\bar{c} = 0.25$ and $\bar{c} = 0.5$, respectively. As in the previous case, we observe similar errors and an analogue behavior of the convergence rates for both choices of $\bar{c}$. We also observe a slight degradation of the errors and the convergence rates in the more complicated case of oblique boundary conditions.

| $\Delta x$ | $\Delta t = \Delta x$ | | | |
|---|---|---|---|---|
| | $E_\infty$ | $E_1$ | $p_\infty$ | $p_1$ |
| $2.50 \cdot 10^{-1}$ | $3.06 \cdot 10^{-1}$ | $4.38 \cdot 10^{-1}$ | - | - |
| $1.25 \cdot 10^{-1}$ | $1.56 \cdot 10^{-1}$ | $2.25 \cdot 10^{-1}$ | 0.97 | 0.96 |
| $6.25 \cdot 10^{-2}$ | $8.10 \cdot 10^{-2}$ | $1.21 \cdot 10^{-1}$ | 0.95 | 0.89 |
| $3.125 \cdot 10^{-2}$ | $4.47 \cdot 10^{-2}$ | $7.17 \cdot 10^{-2}$ | 0.86 | 0.75 |
| $\Delta x$ | $\Delta t = \Delta x/2$ | | | |
| | $E_\infty$ | $E_1$ | $p_\infty$ | $p_1$ |
| $2.50 \cdot 10^{-1}$ | $1.50 \cdot 10^{-1}$ | $2.08 \cdot 10^{-1}$ | - | - |
| $1.25 \cdot 10^{-1}$ | $7.96 \cdot 10^{-2}$ | $1.17 \cdot 10^{-1}$ | 0.91 | 0.83 |
| $6.25 \cdot 10^{-2}$ | $4.36 \cdot 10^{-2}$ | $6.84 \cdot 10^{-2}$ | 0.88 | 0.77 |
| $3.125 \cdot 10^{-2}$ | $2.58 \cdot 10^{-2}$ | $4.26 \cdot 10^{-2}$ | 0.76 | 0.68 |

**Table 3.5.** Errors and convergence rates for the approximation of (3.72) with $\bar{c} = 0.25$

| $\Delta x$ | $\Delta t = \Delta x$ | | | |
|---|---|---|---|---|
| | $E_\infty$ | $E_1$ | $p_\infty$ | $p_1$ |
| $2.50 \cdot 10^{-1}$ | $2.94 \cdot 10^{-1}$ | $3.81 \cdot 10^{-1}$ | - | - |
| $1.25 \cdot 10^{-1}$ | $1.49 \cdot 10^{-1}$ | $1.88 \cdot 10^{-1}$ | 0.98 | 1.02 |
| $6.25 \cdot 10^{-2}$ | $7.55 \cdot 10^{-2}$ | $9.33 \cdot 10^{-2}$ | 0.98 | 1.01 |
| $3.125 \cdot 10^{-2}$ | $3.95 \cdot 10^{-2}$ | $5.02 \cdot 10^{-2}$ | 0.93 | 0.89 |
| $\Delta x$ | $\Delta t = \Delta x/2$ | | | |
| | $E_\infty$ | $E_1$ | $p_\infty$ | $p_1$ |
| $2.50 \cdot 10^{-1}$ | $1.42 \cdot 10^{-1}$ | $1.69 \cdot 10^{-1}$ | - | - |
| $1.25 \cdot 10^{-1}$ | $7.22 \cdot 10^{-2}$ | $8.56 \cdot 10^{-2}$ | 0.98 | 0.98 |
| $6.25 \cdot 10^{-2}$ | $3.79 \cdot 10^{-2}$ | $4.63 \cdot 10^{-2}$ | 0.93 | 0.89 |
| $3.125 \cdot 10^{-2}$ | $2.12 \cdot 10^{-2}$ | $2.75 \cdot 10^{-2}$ | 0.84 | 0.75 |

**Table 3.6.** Errors and convergence rates for the approximation of (3.72) with $\bar{c} = 0.5$.

### 3.5.3 Nonlinear problem on a non-smooth domain with mixed Dirichlet-Neumann boundary conditions

In this last example, we deal with a problem of exiting from a bounded rectangular domain with an circular obstacle inside of it. We model this problem by considering a modification of (3.1) including mixed Dirichlet-Neumann boundary conditions, with a large time horizon $T$ in order to reach a stationary solution. We consider the space domain

$$\mathcal{O} = \Big((-1,1)\times(-0.5,0.5)\Big) \setminus \{x\in\mathbb{R}^2 \,|\, |x-(-0.5,0)|\le 0.2\},$$

a control set $A=\{a\in\mathbb{R}^2\,|\,|a|=1\}$, a drift $\mu(t,x,a)=a$, a diffusion coefficient $\sigma(t,x,a)=0.1I_2$, where $I_2$ is the identity matrix of size 2, a running cost $f\equiv 1$, and an initial condition $\Psi\equiv 0$. We impose constant Dirichlet boundary conditions on some parts of $\partial\mathcal{O}$, representing the exits of the domain, in order to model some exit costs. More precisely, Dirichlet boundary conditions (or exit costs) $u=0$ and $u=0.2$ are imposed on $\partial\mathcal{O}_1=\{x=(x_1,x_2)\in\partial\mathcal{O}\,|\,x_1=-1,|x_2|\le 0.2\}$ and $\partial\mathcal{O}_2=\{x=(x_1,x_2)\in\partial\mathcal{O}\,|\,x_1=1,|x_2|\le 0.2\}$, respectively. We also consider homogeneous Neumann boundary conditions on the remaining part of the boundary.

We treat the Dirichlet boundary conditions by using an extrapolation technique. This approximation has been proposed in [17] and has been shown to be more accurate with respect to the methods proposed in [86, 19]. We show in Figure 3.4 the numerical approximation computed on an unstructured mesh with mesh size $\Delta x=0.01$, a time step $\Delta t=\Delta x$ and final time $T=3$. Figure 3.5 diplays the quiver plot of $-Du$ at time $T=3$.

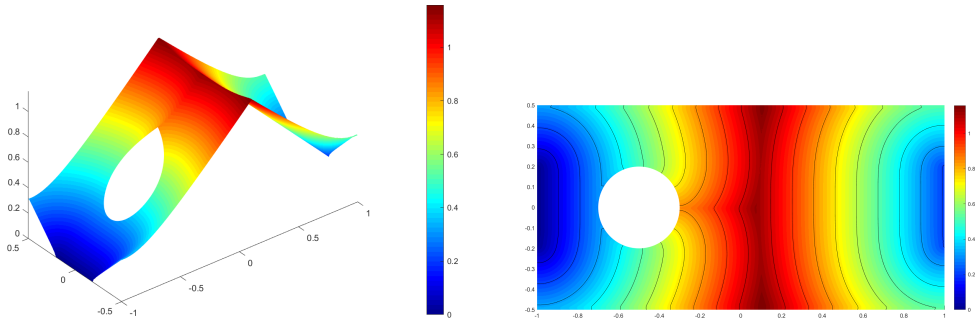


**Figure 3.4.** Solution at time $T=3$ for $\Delta x=0.01$ and for $\Delta t=\Delta x$.

## 3.6 On the existence of the oblique projection

In this section we first study the existence of the projection of $x$ onto $\partial\mathcal{O}$ parallel to $\gamma_b$ in a neighborhood of $\partial\mathcal{O}$ and for $b\in B$. These projections play an important role in the construction of our scheme in Section 3.2. The following result is an extension of a result in [63, Section 1.2] to the regularity that we assume in this paper and, more importantly, to the dependence of $\gamma$ on $b$. Recall

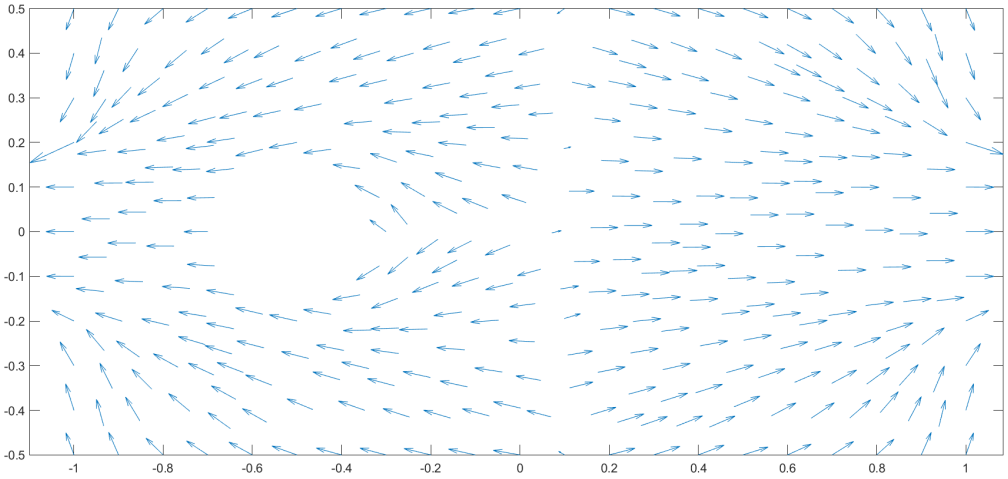



**Figure 3.5.** Quiver plot of $-Du$ at time $T = 3$.

that in **(H3)** $\partial\mathcal{O}$ is assumed to be of class $C^3$. However, the result in Proposition 75 below is also valid if $\partial\mathcal{O}$ is only of class $C^2$.

**Proposition 75.** *There exists $R > 0$ such that, for any $x \in \mathbb{R}^d$ satisfying $d(x, \partial\mathcal{O}) < R$ and for any $b \in B$, there exist a unique $p^{\gamma_b}(x) \in \partial\mathcal{O}$ and a unique $d^{\gamma_b}(x) \in \mathbb{R}$ such that*

$$x = p^{\gamma_b}(x) + d^{\gamma_b}(x)\gamma_b(p^{\gamma_b}(x)). \tag{3.73}$$

*The mappings $(x, b) \mapsto p^{\gamma_b}(x)$ and $(x, b) \mapsto d^{\gamma_b}(x)$, called respectively the projection onto $\partial\mathcal{O}$ parallel to $\gamma_b$ and the algebraic distance to $\partial\mathcal{O}$ parallel to $\gamma_b$, are of class $C^1$.*

*Proof.* We use the same outline and, as much as possible, the same notations than those in [63].

Let us fix $(s, b_0) \in \partial\mathcal{O} \times B$. Let $g^s \colon U^s \to \partial\mathcal{O}$ be a $C^2$ parameterization of $\partial\mathcal{O}$ in a neighborhood of $s$, with $U^s$ being an open subset of $\mathbb{R}^{N-1}$, $z_0 \in U^s$, and $g^s(z_0) = s$. By **(H3)** the function

$$U^s \times \mathbb{R} \times \mathcal{V} \ni (z, \lambda, b) \mapsto G^s(z, \lambda, b) = (g^s(z) + \lambda\gamma_b(g^s(z)), b) \in \mathbb{R}^d \times \mathbb{R}^{N_B}$$

is of class $C^1$. The Jacobian matrix of $G^s$ has the form

$$J^s(z, \lambda, b) = \left(\begin{array}{c|c} J_{z,\lambda}(z, \lambda, b) & J_b(z, \lambda, b) \\ \hline 0_{N_B,N} & I_{N_B} \end{array}\right),$$

where $J_{z,\lambda}(z, \lambda, b)$ coincides with $J(z, \lambda)$ of the Appendix A of [63], that is

$$J_{z,\lambda}(z, \lambda, b) =$$

$$\left(\begin{array}{c|c|c|c} \partial_{z_1} g^s(z) + \lambda\partial_{z_1}\gamma_b(g^s(z)) & \cdots & \partial_{z_{N-1}} g^s(z) + \lambda\partial_{z_{N-1}}\gamma_b(g^s(z)) & \gamma_b(g^s(z)) \end{array}\right).$$

In particular, for $\lambda = 0$,

$$J_{z,\lambda}(z,0,b) = \left( \partial_{z_1} g^s(z) \;\middle|\; \cdots \;\middle|\; \partial_{z_{N-1}} g^s(z) \;\middle|\; \gamma_b(g^s(z)) \right)$$

is invertible since its $N-1$ first columns span the tangent space to $\partial\mathcal{O}$ at $g^s(z)$ and, since

$$\langle n(g^s(z)), \gamma_b(g^s(z)) \rangle > 0,$$

its last column is non tangent to $\partial\mathcal{O}$. It follows that $J^s(z,0,b)$ is also invertible, and we can therefore apply the inverse mapping theorem to $G^s$ at $(z_0,0,b_0)$ to obtain the existence of a neighborhood $V^{s,b_0}$ of $(s,b_0)$ and $C^1$ mappings $V^{s,b_0} \ni (x,b) \mapsto p^{\gamma_b}(x) \in \partial\mathcal{O}$ and $V^{s,b_0} \ni (x,b) \mapsto d^{\gamma_b}(x)$ such that (3.73) holds for every $(x,b) \in V^{s,b_0}$. The compactness of $\partial\mathcal{O} \times B \subset \cup_{(s,b_0)\in\partial\mathcal{O}\times B} V^{s,b_0}$ enables to consider a finite number of $(s_i,(b_0)_i)$, $1 \le i \le k$, such that $\partial\mathcal{O}\times B \subset \cup_{i=1}^k V^{s_i,(b_0)_i}$. Then there exists $\bar{R} > 0$ such that $\{y \in \mathbb{R}^N \mid d(y,\partial\mathcal{O}) < \bar{R}\} \times B \subset \cup_{i=1}^k V^{s_i,(b_0)_i}$. In particular for any $x$ such that $d(x,\partial\mathcal{O}) < \bar{R}$ and any $b \in B$, there exist a least a point $p^{\gamma_b}(x)$ and a scalar $d^{\gamma_b}(x)$ such that (3.73) holds. We claim that there exists $R \in (0,\bar{R})$ such that for any $x$ satisfying $d(x,\partial\mathcal{O}) < R$ and any $b \in B$, $p^{\gamma_b}(x)$ is unique (and as a consequence $d^{\gamma_b}(x)$ is also unique). Assume that this is not the case. Then (considering for example $R = \frac{1}{k}$) one can build a sequence $(x_k,b_k)_{k\in\mathbb{N}}$ converging (after extraction a subsequence) to some point $(\hat{s},\hat{b}) \in \partial\mathcal{O} \times B$ and such that for all $k \in \mathbb{N}$, $x_k$ has two distinct projections $p_i^{\gamma_{b_k}}(x_k)$ with associated algebraic distances $d_i^{\gamma_{b_k}}(x_k)$, $i = 1,2$. At the limit point $\hat{s}$, we consider $G^{\hat{s}}$ which is a local diffeomorphism on a neighborhood of $(\hat{z},0,\hat{b})$ (with $g^{\hat{s}}(\hat{z}) = \hat{s}$). Since $x_k \to \hat{s} \in \partial\mathcal{O}$, then $p_i^{\gamma_{b_k}}(x_k) \to \hat{s}$ and $d_i^{\gamma_{b_k}}(x_k) \to 0$, $i = 1,2$. Let $z_{i,k}$ be such that $g^{\hat{s}}(z_{i,k}) = p_i^{\gamma_{b_k}}(x_k)$ and $\lambda_{i,k} = d_i^{\gamma_{b_k}}(x_k)$, $i = 1,2$. Then $(z_{i,k},\lambda_{i,k},b_k)_k$, $i = 1,2$, are distinct sequences that both converge to $(\hat{z},0,\hat{b})$ and have the same image $G^{\hat{s}}(z_{i,k},\lambda_{i,k},b_k) = (x_k,b_k)$. This contradicts that $G^{\hat{s}}$ is a local diffeomorphism on a neighborhood of $(\hat{z},0,\hat{b})$. □

For any $\varepsilon \ge 0$ let us define

$$D_\varepsilon = \{x \in \overline{\mathcal{O}} \mid d(x,\partial\mathcal{O}) > \varepsilon\}, \tag{3.74}$$

$$\partial\mathcal{O}_\varepsilon = \{x \in \overline{\mathcal{O}} \mid d(x,\partial\mathcal{O}) = \varepsilon\}, \tag{3.75}$$

$$L_\varepsilon = \{x \in \overline{\mathcal{O}} \mid d(x,\partial\mathcal{O}) \le \varepsilon\}. \tag{3.76}$$

Now we focus on the existence of projections of $x \in L_\varepsilon$ onto $\partial\mathcal{O}_\varepsilon$ and the regularity of $L_\varepsilon \ni x \mapsto d(x,D_\varepsilon) \in \mathbb{R}$. These results are important in order to show Lemma 71 which is the key to obtain the stability of the scheme in Proposition 72.

**Lemma 76.** *The following hold:*

(i) *There exists $\eta > 0$ such that on $L_\eta$, the projection $p_{\partial\mathcal{O}}$ onto $\partial\mathcal{O}$ is well-defined and $C^1$.*

(ii) *The distance function $L_\eta \ni x \mapsto d(x,\partial\mathcal{O}) \in \mathbb{R}$ is $C^3$, and $Dd(\cdot,\partial\mathcal{O})(x) = -n(p_{\partial\mathcal{O}}(x))$.*

*Let $\delta \in [0, \eta]$. Then the following hold:*

(iii) *$\partial\mathcal{O}_\delta$ is of class $C^3$ and, denoting by $n_\delta(x)$ the unit outward normal at $x \in \partial\mathcal{O}_\delta$, we have $n_\delta(x) = n(p_{\partial\mathcal{O}}(x))$.*

(iv) *For every $x \in L_\delta$, $p = p_{\partial\mathcal{O}}(x) - \delta n(p_{\partial\mathcal{O}}(x))$ is a projection of $x$ onto $\partial\mathcal{O}_\delta$.*

(v) *The function $x \mapsto d(x, \partial\mathcal{O}_\delta)$ is of class $C^3$ on $L_\delta$ and $d(x, \partial\mathcal{O}) + d(x, \partial\mathcal{O}_\delta) = \delta$ for every $x \in L_\delta$.*

*Proof.* (i)&(ii) See [62, Lemma 14.16].

(iii) This follows from (ii) and (3.75).

(iv)&(v) Let us first show that $p \in \partial\mathcal{O}_\delta$. We have $d(p, \partial\mathcal{O}) \le |p - p_{\partial\mathcal{O}}(x)| = \delta$. Thus, $p \in L_\delta$ and, by (i), $p_{\partial\mathcal{O}}(x) = p_{\partial\mathcal{O}}(p)$, which implies that $d(p, \partial\mathcal{O}) = \delta$ and hence $p \in \partial\mathcal{O}_\delta$. Since

$$x = p_{\partial\mathcal{O}}(x) - d(x, \partial\mathcal{O}) n(p_{\partial\mathcal{O}}(x)),$$

we obtain $d(x, \partial\mathcal{O}_\delta) \le |p - x| = \delta - d(x, \partial\mathcal{O})$. Assume that $d(x, \partial\mathcal{O}_\delta) < \delta - d(x, \partial\mathcal{O})$. Then there exists $p' \in \partial\mathcal{O}_\delta$ such that $|x - p'| < \delta - d(x, \partial\mathcal{O})$. This implies that

$$\delta = d(p', \partial\mathcal{O}) \le |p' - p_{\partial\mathcal{O}}(x)| \le |p' - x| + |x - p_{\partial\mathcal{O}}(x)| < \delta,$$

which is impossible. Thus

$$|p - x| = d(x, \partial\mathcal{O}_\delta) = \delta - d(x, \partial\mathcal{O}).$$

The first equality above implies that $p$ is a projection of $x$ onto $\partial\mathcal{O}_\delta$. Since $x \in L_\delta$ is arbitrary, the second equality above and (ii) imply that (v) holds. □

# Chapter 4

# A second order Lagrange-Galerkin scheme for Fokker-Planck equations and applications to MFGs

In physics, chemistry, or electrical engineering it is very important to study the microscopic qualitative changes of systems, for example in state transitions. When a transition takes place, fluctuation (modelised as random processes) have an important role, and Fokker-Planck equations can be used in order to model such problems. In general, problems wich involve a noise can be treated using Fokker-Planck equations. In this chapter, after a brief presentation of an already existing SL method for the FP equations, we present a novel LG approach for the numerical approximation of such equations. We also show how to use this method to numerically solve a Mean Field Games problem.

## 4.1 A first order semi-Lagrangian scheme for the Fokker-Planck equation

We briefly recall a first order scheme for the nonlinear Fokker-Planck equation presented in [34]. This scheme, coupled with an approximation method for the Hamilton-Jacobi equation, can be implied for a first order approximation scheme for Mean Field Games. We consider the equation

$$\begin{cases} \partial_t m - \frac{1}{2}\sum_{i,j=1}^{d} \frac{\partial^2}{\partial x_i \partial x_j}(a_{ij} m) + \operatorname{div}(\mu m) = 0 & \text{in } (0,T]\times\mathbb{R}^d \\ m(0,\cdot) = \overline{m}_0(x) & \text{in } \mathbb{R}^d, \end{cases} \tag{4.1}$$

where $\overline{m}_0 \in \mathcal{P}_2(\mathbb{R}^d)$, $a_{ij} = (\sigma(x)\sigma(x)^\top)_{i,j}$, with $\sigma : \mathbb{R}^d \to \mathbb{R}^{d\times r}$. Suppose that:

(FP1) the coefficients $b$ and $\sigma$ are Lipschitz continuous;

(FP2) the initial measure $m_0$ has a density, which we still denote by $m_0$.

Given $x \in \mathbb{R}^d$ and $s \in [0,T)$, consider the stochastic differential equation

$$\begin{cases} \mathrm{d}X(s') = \mu(X(s'))\mathrm{d}s' + \sigma(X(s'))\mathrm{d}W(s'), \\ X(s) = x, \end{cases} \tag{4.2}$$

for $s' \in (s,T)$ and denote $X^{x,s}$ its unique solution. For $t \in (s,T]$ the *flow* $\Phi_{s,t} : \mathbb{R}^d \times \mathcal{O} \to \mathbb{R}^d$ is defined as

$$\Phi_{s,t}(x,\omega) := X^{x,s}(t,\omega), \tag{4.3}$$

where $\Phi_{s,t}$ is continuous and differentiable. Given a measure $\mu$ on $\mathbb{R}^d$ and a function $\psi : \mathbb{R}^d \to \mathbb{R}^d$, for all $A \in \mathcal{B}(\mathbb{R}^d)$ we denote by $\psi\sharp\mu$ the measure given by $\psi\sharp\mu(A) = \mu(\psi^{-1}(A))$. The scheme is based on a representation formula for the solution of (4.1).

**Lemma 77.** *Under assumptions* (FP1), (FP2)*, suppose that $m$ is the unique solution to* (4.1)*. Then, for each $t \in [0,T]$,*

$$m(t)(A) = \mathbb{E}\left[\Phi_{0,t}(\cdot)\sharp m_0(A)\right] \quad \textit{for all } A \in \mathcal{B}(\mathbb{R}^d), \tag{4.4}$$

*and for $0 \le s < t \le T$*

$$m(t)(A) = \mathbb{E}\left[\Phi_{s,t}(\cdot)\sharp m(s)(A)\right] \quad \textit{for all } A \in \mathcal{B}(\mathbb{R}^d). \tag{4.5}$$

The proof can be found in [34]. First of all, the time interval $[0,T]$ is discretized using a step $\Delta t > 0$ and we set $t_k = k\Delta t$ for $k = 1, \dots, \lfloor N/\Delta t \rfloor$. Differently than Section 2.1, the diffusion here is a matrix, with possible not full rank. Then, the flow is approximated using a forward weak Euler method as in Section 3.2.2, so that for $\ell = 1, \dots, r$ there are two time discrete fluxes

$$\begin{aligned} y_k^{+,\ell} &:= x + \Delta t \mu(t_k, x) + \sqrt{r\Delta t}\sigma^\ell(t_k, x), \\ y_k^{-,\ell} &:= x + \Delta t \mu(t_k, x) - \sqrt{r\Delta t}\sigma^\ell(t_k, x). \end{aligned} \tag{4.6}$$

Given $\Delta x > 0$, define the uniform triangulation with vertices in the lattice

$$\mathcal{G}_{\Delta x} := \{x_j = j\Delta x, j \in \mathbb{Z}^d\}.$$

We define the fully-discrete characteristics as

$$\begin{aligned} y_{k,j}^{+,\ell} &:= x_j + \Delta t \mu(t_k, x_j) + \sqrt{r\Delta t}\sigma^\ell(t_k, x_j), \\ y_{k,j}^{-,\ell} &:= x_j + \Delta t \mu(t_k, x_j) - \sqrt{r\Delta t}\sigma^\ell(t_k, x_j). \end{aligned} \tag{4.7}$$

Setting

$$E_j := \left[x_j^1 - \tfrac{1}{2}\Delta x, x_j^1 + \tfrac{1}{2}\Delta x\right] \times \cdots \times \left[x_j^d - \tfrac{1}{2}\Delta x, x_j^d + \tfrac{1}{2}\Delta x\right],$$

and defining $\beta_j(x)$ the basis of $\mathbb{P}_1$ defined on the standard triangulation with vertices in $\mathcal{G}_{\Delta x}$, the scheme results in

$$\begin{cases} m_{k,i} = \frac{1}{2r} \sum\limits_{j \in \mathbb{Z}^d} \sum\limits_{\ell=1}^{r} [\beta_i(y_{k,j}^{+,\ell}) + \beta_i(y_{k,j}^{-,\ell})] m_{k,j}, \\ m_{0,i} = \int_{E_i} m_0(x)\mathrm{d}x. \end{cases} \tag{4.8}$$

Using the sequence $\{m_{k,i}\}$ computed in (4.8) it is possible to define the discrete measure

$$\mathrm{d}\tilde{m}(t_k) := \sum_{i\in\mathbb{Z}^d} m_{k,i}\beta_i(x)\mathrm{d}x.$$

Calling $\Delta = (\Delta t, \Delta x)$, we define its extension, interpolating linearly in time, as

$$m_\Delta(t) = \frac{t_{k+1}-t}{\Delta t}\tilde{m}(t_k) + \frac{t-t_k}{\Delta t}\tilde{m}(t_{k+1}), \tag{4.9}$$

which is a continuous measure. The following convergence result has been proved in [35].

**Theorem 78.** *The measure $t \in [0,T] \to m_\Delta(t)$ defined in (4.9) converges to the solution $m$ of (4.1) in the weak sense.*

**Remark 79.** *The basis functions $(\beta_i)_{i\in\mathbb{Z}^d}$ of $\mathbb{P}^1$ are nonnegative; this implies that the scheme preserves nonnegativity. In addition, since $\sum_{i\in\mathbb{Z}^d}\beta_i(x) = 1$ for all $x \in \mathbb{R}^d$, we get*

$$\sum_{i\in\mathbb{Z}^d} m_{k+1,i} = \sum_{i\in\mathbb{Z}^d} m_{k,i} = \sum_{i\in\mathbb{Z}^d} m_{0,i} = 1,$$

*meaning that the scheme is conservative.*

## 4.2 A second order Lagrange-Galerkin scheme for the Fokker-Planck equation

We propose a second order accurate numerical method for linear Fokker-Plank-Kolmogorov equations, based on the coupling of Lagrange-Galerkin techniques with semi-Lagrangian methods. The method is conservative, explicit and stable under rather large steps. We develop a convergence analysis for the exacted integrated scheme, and we propose an implementable version with non-exact integration. We consider application for time dependent Mean Field Games problems, and we show numerical simulations.

In this section, we consider the following linear FP equation

$$\begin{cases} \partial_t m - \frac{\sigma^2}{2}\Delta m + \operatorname{div}(\mu m) &= 0 \quad \text{in } (0,T)\times\mathbb{R}^d, \\ m(0,\cdot) &= \overline{m}_0 \quad \text{in } \mathbb{R}^d, \end{cases} \tag{FP}$$

where $\sigma \in \mathbb{R}\setminus\{0\}$, $\mu : [0,T]\times\mathbb{R}^d \to \mathbb{R}^d$, and $\overline{m}_0 : \mathbb{R}^d \to \mathbb{R}$.

In the following, for $p \in \mathbb{N}\cup\{\infty\}$, $C_0^p(\mathbb{R}^d)$ denotes the sets of functions $\phi$ of class $C^p$ with support $\operatorname{supp}(\phi)$ being compact.

**(H1)** We assume that:

(i) $\overline{m}_0$ is nonnegative, continuous, has compact support, and $\int_{\mathbb{R}^d}\overline{m}_0(x)\mathrm{d}x = 1$.

(ii) $\mu$ is bounded, $\mu \in C^\infty([0,T]\times\mathbb{R}^d)$, and there exists $C_\mu > 0$ such that

$$|\mu(s,x) - \mu(t,y)| \le C_\mu(|s-t| + |x-y|) \quad \text{for } s,\, t \in [0,T] \text{ and } x,\, y \in \mathbb{R}^d.$$

In the following result, we summarize some important properties of equation (**FP**).

**Theorem 80.** *Assume* **(H1)**. *Then the following hold:*

(i) *Equation* (**FP**) *admits a unique classical solution* $m^* \in C^{1,2}([0,T]\times\mathbb{R}^d)$.

(ii) $m^* \geq 0$.

(iii) $\int_{\mathbb{R}^d} m^*(t,x)\mathrm{d}x = 1$ *for all* $t \in [0,T]$.

(iv) $m^*$ *is the unique solution in* $L^2([0,T]\times\mathbb{R}^d)$ *to* (**FP**) *in the distributional sense.*

*Proof.* We refer the reader to [13, Theorem 6.6.1, Chapter 9.1] for the proofs of (i)-(ii) and to [56, Proposition 4.4 and Theorem 4.3] for the proofs of (iii)-(iv). □

Let us recall the probabilistic interpretation of the solution $m^*$ to (**FP**), which will be useful in order to construct a LG scheme. Let $W$ be a $d$-dimensional Brownian motion defined on a probability space $(\Omega,\mathcal{F},\mathbb{P})$ and let $X_0:\Omega\to\mathbb{R}^d$ be a random variable, independent of $W$, and whose distribution is absolutely continuous with respect to the Lebesgue measure in $\mathbb{R}^d$, with density given by $\overline{m}_0$. Given $(t,x)\in[0,T]\times\mathbb{R}^d$, we define $X^{t,x}$ as the unique strong solution to the stochastic differential equation

$$\begin{aligned}\mathrm{d}X(s) &= \mu(s,X(s))\mathrm{d}s + \sigma\mathrm{d}W(s) \quad \text{for } s\in(t,T)\\ X(t) &= x.\end{aligned}$$

Denote by $\mathbb{E}(Y)$ the expectation of a random variable $Y:\Omega\to\mathbb{R}$. Assumption **(H1)** implies that $X^{0,X_0}(t)$ is well defined for all $t\in[0,T]$ and its distribution is absolutely continuous with respect to the Lebesgue measure in $\mathbb{R}^d$, with density given by $m^*(t,\cdot)$ (see e.g. [56]). From the $\mathbb{P}$-a.s. equality $X^{0,X_0}(s) = X^{t,X^{0,X_0}(t)}(s)$ for every $0\leq t\leq s\leq T$, we deduce that for every continuous and bounded function $\phi:\mathbb{R}^d\to\mathbb{R}$, we have

$$\int_{\mathbb{R}^d}\phi(x)m^*(s,x)\mathrm{d}x = \int_{\mathbb{R}^d}\mathbb{E}\left(\phi(X^{t,x}(s))\right)m^*(t,x)\mathrm{d}x. \tag{4.10}$$

### 4.2.1 A space-time Lagrange-Galerkin approximation

Let focus on the numerical approximation of (**FP**). Notice that, under **(H1)**, equation (**FP**) can be written as

$$\begin{aligned}\partial_t m - \frac{\sigma^2}{2}\Delta m + \langle\mu, Dm\rangle + \mathrm{div}(\mu)m = 0 \qquad &\text{in } (0,T)\times\mathbb{R}^d,\\ m(0,\cdot) = \overline{m}_0 \qquad &\text{in } \mathbb{R}^d.\end{aligned}$$

In the form above, a semi-Lagrangian scheme can be implemented to approximate $m^*$ (see e.g. [17]). However, such a scheme is not conservative, i.e. the discrete solution does not satisfy the discrete analogous of Theorem 80(iii). The scheme that we consider, which will be built from (4.10), will allow us to preserve this property (see Theorem 59(ii) below).

Let us fix $N_{\Delta t}\in\mathbb{N}$, set $\mathcal{I}_{\Delta t} = \{0,\dots,N_{\Delta t}\}$, $\mathcal{I}^*_{\Delta t} = \mathcal{I}_{\Delta t}\setminus\{N_{\Delta t}\}$, $\Delta t = T/N_{\Delta t}$, and $t_k = k\Delta t$ ($k\in\mathcal{I}_{\Delta t}$). For $k\in\mathcal{I}^*_{\Delta t}$ and $x\in\mathbb{R}^d$, we denote by $y^{t_k,x}$ a one-step

second order *Crank-Nicolson* approximation of $X^{t_k,x}(t_{k+1})$ (see [70, Section 15.4] and also [52, Section 2]). More precisely, for all $\Delta t$ small enough, we define $y^{t_k,x}$ is the unique solution to

$$y = x + \frac{\Delta t}{2}\left(\mu(t_k, x) + \mu(t_{k+1}, y)\right) + \sqrt{\Delta t}\sigma\xi, \tag{4.11}$$

where $\xi$ is a $\mathbb{R}^d$-valued random variable with i.i.d. components such that

$$\mathbb{P}(\xi_i = 0) = 2/3 \quad \text{and} \quad \mathbb{P}(\xi_i = \pm\sqrt{3}) = 1/6 \quad \text{for} \quad i = 1, \dots, d. \tag{4.12}$$

Let $\mathcal{I}_d = \{1, \dots, 3^d\}$, define $\{e^\ell \,|\, \ell \in \mathcal{I}_d\} \subset \mathbb{R}^d$ as the set of possible values of $\xi$, set $\omega^\ell = \mathbb{P}(\xi = e^\ell)$, and denote by $y_k^\ell(x)$ the unique solution to (4.11) for $\xi = e^\ell$ ($\ell \in \mathcal{I}_d$). By standard estimates for the weak approximation of $X^{t_k,x}(t_{k+1})$ (see e.g. [70, Theorem 14.5.2]), if the space derivatives of $\mu$ up to order six have polynomial growth, for all $\phi : \mathbb{R}^d \to \mathbb{R}$ smooth enough, we have that

$$\Big|\sum_{\ell\in\mathcal{I}_d} \phi(y_k^\ell(x))\omega_\ell - \mathbb{E}\left(\phi(X^{t_k,x}(t_{k+1}))\right)\Big| = O\left((\Delta t)^3\right). \tag{4.13}$$

Thus, in order to obtain a second order scheme, it is natural to approximate (4.10) by the equation

$$\int_{\mathbb{R}^d} \phi(x) m_{k+1}(x)\mathrm{d}x = \sum_{\ell\in\mathcal{I}_d} \omega_\ell \int_{\mathbb{R}^d} \phi(y_k^\ell(x)) m_k(x)\mathrm{d}x, \tag{4.14}$$

for all $\phi$ smooth enough and $k \in \mathcal{I}_{\Delta t}$, with $m_0 = \overline{m}_0$ and unknowns $\{m_k : \mathbb{R}^d \to \mathbb{R} \,|\, k \in \mathcal{I}_{\Delta t} \setminus \{0\}\}$. Note that the boundedness of $\mu$ and (4.11) implies the existence of $L_{\Delta t} = O(1/\sqrt{\Delta t})$ such that the solution $m_{\Delta t}$ to (4.14) satisfies

$$\mathrm{supp}(m_{\Delta t,k}) \subset [-L_{\Delta t}, L_{\Delta t}]^d \quad \text{for all } k \in \mathcal{I}_{\Delta t}. \tag{4.15}$$

In order to construct a space discretization of (4.14), let us fix $p \in \mathbb{N}$, set $q := 2p + 1$, and let $\widehat{\beta} : \mathbb{R} \to \mathbb{R}$ be defined by

$$(\forall\, \xi \in [0, \infty)) \quad \widehat{\beta}(\xi) = \begin{cases} \displaystyle\prod_{k\neq 0,\, k=-p}^{p+1} \frac{\xi - k}{-k} & \text{if } \xi \in [0, 1], \\ \displaystyle\prod_{k\neq 0,\, k=-p+1}^{p+2} \frac{\xi - k}{-k} & \text{if } \xi \in (1, 2], \\ \vdots & \\ \displaystyle\prod_{k=1}^{2p+1} \frac{\xi - k}{-k} & \text{if } \xi \in (p, p+1], \\ 0 & \text{if } \xi \in (p+1, \infty), \\ \widehat{\beta}(-\xi) & \text{if } \xi \in (-\infty, 0). \end{cases} \tag{4.16}$$

Following [53], for $\Delta x \in (0,\infty)$, we consider the symmetric Lagrange interpolation basis function $\{\beta_i\}_{i\in\mathbb{Z}^d}$ defined as

$$(\forall\, z=(z_1,\dots z_d)\in\mathbb{R}^d,\ i=(i_1,\dots,i_d)\in\mathbb{Z}^d)\quad \beta_i(z)=\prod_{j=1}^{d}\widehat{\beta}\left(\frac{z_j}{\Delta x}-i_j\right).$$

For all $i\in\mathbb{Z}^d$, let us set $x_i=i\Delta x$. Notice that $\beta_i$ has compact support, $\beta_i(x_j)=1$ if $i=j$ and $\beta_i(x_j)=0$ otherwise, and, for all $x\in\mathbb{R}^d$, $\sum_{j\in\mathbb{Z}^d}\beta_j(x)=1$. Given $f\in W^{q+1,\infty}(\mathbb{R}^d)$, we define the interpolant $I[f]:\mathbb{R}^d\to\mathbb{R}$ by

$$(\forall\, x\in\mathbb{R}^d)\quad I[f](x)=\sum_{i\in\mathbb{Z}^d} f(x_i)\beta_i(x), \tag{4.17}$$

By [39, Theorem 16.1], the following estimate holds

$$(\exists\, C_I>0)\quad \sup_{x\in\mathbb{R}^d}|f(x)-I[f](x)|\le C_I(\Delta x)^{q+1}\|D^{q+1}f\|_{L^\infty}, \tag{4.18}$$

with $C_I>0$ independent of $f$ and $\Delta x$. Notice that in the one dimensional case ($d=1$), $I[f]$ restricted to a given interval $(x_i,x_{i+1})$ ($i\in\mathbb{Z}$) is the Lagrange interpolating polynomial of degree $q$ constructed on the symmetric stencil $x_{i-(q-1)/2},\dots,x_{i+1+(q-1)/2}$.

Let $L_{\Delta t}>0$ be as in (4.15), let $N_{\Delta x}\in\mathbb{N}$, and set $\mathcal{I}_{\Delta x}=\{-N_{\Delta x},\dots,N_{\Delta x}\}^d$. From now on, we assume that $\Delta x=L_{\Delta t}/N_{\Delta x}$, we set $\Delta=(\Delta t,\Delta x)$, and we consider the space domain $\mathcal{O}_\Delta=[-L_{\Delta t}-p\Delta x, L_{\Delta t}+p\Delta x]^d$. We look for an approximation $m_\Delta$ of the solution $m^*$ to (**FP**) such that, for all $k\in\mathcal{I}_{\Delta t}$,

$$m_\Delta(t_k,x)=\sum_{i\in\mathcal{I}_{\Delta x}} m_{k,i}\beta_i(x)\ \text{for } x\in\mathcal{O}_\Delta,\quad m_\Delta(t_k,x)=0\ \text{for } x\in\mathbb{R}^d\setminus\mathcal{O}_\Delta, \tag{4.19}$$

where $m_{k,i}\in\mathbb{R}$ ($k\in\mathcal{I}_{\Delta t}$, $i\in\mathcal{I}_{\Delta x}$) have to be determined. Notice that, by definition of $\mathcal{I}_{\Delta x}$, for all $k\in\mathcal{I}_{\Delta t}$ we have that $\mathrm{supp}\{m_\Delta(t_k,\cdot)\}\subset\mathcal{O}_\Delta$. Replacing $m$ by $m_\Delta$ and taking $\phi=\beta_i$ ($i\in\mathcal{I}_{\Delta x}$) in (4.14) yields the following explicit iterative scheme for the unknowns $m_{k,i}\in\mathbb{R}$ ($k\in\mathcal{I}_{\Delta t}$, $i\in\mathcal{I}_{\Delta x}$)

$$\begin{aligned}
\sum_{j\in\mathcal{I}_{\Delta x}} m_{k+1,j}\int_{\mathcal{O}_\Delta}\beta_i(x)\beta_j(x)\mathrm{d}x &= \sum_{j\in\mathcal{I}_{\Delta x}} m_{k,j}\sum_{\ell\in\mathcal{I}_d}\omega_\ell\int_{\mathcal{O}_\Delta}\beta_i(y_k^\ell(x))\beta_j(x)\mathrm{d}x\\
&\qquad\qquad \text{for } k\in\mathcal{I}^*_{\Delta t},\ i\in\mathcal{I}_{\Delta x},\\
\sum_{j\in\mathcal{I}_{\Delta x}} m_{0,j}\int_{\mathcal{O}_\Delta}\beta_i(x)\beta_j(x)\mathrm{d}x &= \int_{\mathcal{O}_\Delta}\overline{m}_0(x)\beta_i(x)\mathrm{d}x.
\end{aligned} \tag{4.20}$$

Let $A$ be the $(2N_{\Delta x}+1)^d\times(2N_{\Delta x}+1)^d$ real mass matrix with entries given by

$$A_{i,j}=\int_{\mathcal{O}_\Delta}\beta_i(x)\beta_j(x)\mathrm{d}x,\quad \text{for}\quad (i,j)\in\mathcal{I}_{\Delta x}\times\mathcal{I}_{\Delta x}. \tag{4.21}$$

For $k\in\mathcal{I}^*_{\Delta t}$ and $\ell\in\mathcal{I}_d$, let $B_k^\ell$ be the $(2N_{\Delta x}+1)^d\times(2N_{\Delta x}+1)^d$ real matrix with entries given by

$$(B_k^\ell)_{i,j}=\int_{\mathcal{O}_\Delta}\beta_i(y_k^\ell(x))\beta_j(x)\mathrm{d}x\quad \text{for}\quad (i,j)\in\mathcal{I}_{\Delta x}\times\mathcal{I}_{\Delta x}. \tag{4.22}$$

Let $m_{0,\Delta x}$ be the $(2N_{\Delta x}+1)^d$ dimensional real vector with entries

$$(m_{0,\Delta x})_i = \int_{\mathcal{O}_\Delta} \overline{m}_0(x)\beta_i(x)\mathrm{d}x \quad \text{for } i \in \mathcal{I}_{\Delta x}.$$

Calling $m_k = (m_{k,i})_{i\in\mathcal{I}_{\Delta x}}$, the scheme (4.20) can be rewritten in matrix form: find $m_k$ ($k \in \mathcal{I}_{\Delta t}$) such that

$$\begin{aligned} Am_{k+1} &= \sum_{\ell\in\mathcal{I}_d} \omega_\ell B_k^\ell m_k \quad \text{for } k \in \mathcal{I}^*_{\Delta t}, \\ A\overline{m}_0 &= m_{0,\Delta x}. \end{aligned} \tag{4.23}$$

### 4.2.2 Properties of the space-time Lagrange-Galerkin scheme

We show below some properties of the scheme (4.20) and we assume that **(H1)** are satisfied in the rest of the paper.

**Theorem 81.** *Assume that* **(H1)** *holds. For fixed* $\Delta = (\Delta t, \Delta x) \in (0,\infty)^2$, *there exists a unique solution* $(m_{k,i})_{k\in\mathcal{I}_{\Delta t}, i\in\mathcal{I}_{\Delta x}}$ *to* (4.23) *and, defining* $m_\Delta$ *as in* (4.19)*, the following assertions hold true:*

(i)[Initial condition] $\|\overline{m}_0 - m_\Delta(0,\cdot)\|_{L^2(\mathbb{R}^d)} = O((\Delta x)^{q+1})$ *if* $\overline{m}_0 \in H^{q+1}(\mathbb{R}^d)$.

(ii)[Mass conservation] $\int_{\mathcal{O}_\Delta} m_\Delta(t_k,x)\mathrm{d}x = 1$ *for* $k \in \mathcal{I}_{\Delta t}$.

(iii)[$L^2$-stability] $\max_{k\in\mathcal{I}_{\Delta t}} \|m_\Delta(t_k,\cdot)\|_{L^2}$ *is uniformly bounded with respect to* $\Delta$ *for* $\Delta t$ *small enough.*

*Proof.* The well-posedness of (4.23) follows from the positive definiteness of $A$ (see e.g. [96, Proposition 6.3.1]) and assertion (i) is a consequence of Assumption (**H1**)(i) and [96, Section 3.5]. In order to prove (ii), fix $k \in \mathcal{I}^*_{\Delta t}$ and sum over $i \in \mathbb{Z}^d$ in the first equation of (4.20) to obtain

$$\sum_{j\in\mathcal{I}_{\Delta x}} m_{k+1,j} \sum_{i\in\mathbb{Z}^d} \int_{\mathcal{O}_\Delta} \beta_j(x)\beta_i(x)\mathrm{d}x = \sum_{j\in\mathcal{I}_{\Delta x}} m_{k,j} \sum_{\ell\in\mathcal{I}} \omega_\ell \sum_{i\in\mathbb{Z}^d} \int_{\mathcal{O}_\Delta} \beta_j(x)\beta_i(y_k^\ell(x))\mathrm{d}x.$$

Recalling that, for every $y \in \mathbb{R}^d$, $\sum_{i\in\mathbb{Z}^d}\beta_i(y) = 1$, the cardinality $\{i \in \mathbb{Z}^d \,|\, \beta_i(y) \neq 0\}$ is bounded uniformly in $y$, and $\sum_{\ell\in\mathcal{I}_d}\omega_\ell = 1$, Fubini's theorem yields

$$\begin{aligned} \textstyle\int_{\mathcal{O}_\Delta} m_\Delta(t_{k+1},x)\mathrm{d}x &= \textstyle\sum_{j\in\mathcal{I}_{\Delta x}} m_{k+1,j} \int_{\mathcal{O}_\Delta} \beta_j(x)\mathrm{d}x \\ &= \textstyle\sum_{j\in\mathcal{I}_{\Delta x}} m_{k,j} \int_{\mathcal{O}_\Delta} \beta_j(x)\mathrm{d}x \\ &= \textstyle\sum_{j\in\mathcal{I}_{\Delta x}} m_{0,j} \int_{\mathcal{O}_\Delta} \beta_j(x)\mathrm{d}x \\ &= \textstyle\int_{\mathcal{O}_\Delta} m_\Delta(0,x)\mathrm{d}x. \end{aligned} \tag{4.24}$$

Analogously, using the second equation in (4.20) and summing over $i \in \mathbb{Z}^d$, we get

$$\int_{\mathcal{O}_\Delta} m_\Delta(0,x)\mathrm{d}x = \int_{\mathcal{O}_\Delta} \overline{m}_0(x)\mathrm{d}x = 1. \tag{4.25}$$

Assertion (ii) follows from (4.24) and (4.25). Finally, let us show assertion (iii). For $k = 0$, (iii) follows from Assumption **(H1)**(i) and Theorem 81(i). For $k \in \mathcal{I}^*_{\Delta t}$, (4.20) implies that

$$\begin{aligned} \|m_\Delta(t_{k+1},\cdot)\|^2_{L^2} &= \textstyle\sum_{\ell\in\mathcal{I}_d} \omega_\ell \sum_{i,j\in\mathcal{I}_{\Delta x}} m_{k+1,i} m_{k,j} \int_{\mathcal{O}_\Delta} \beta_i(x)\beta_j(y^\ell_k(x))\mathrm{d}x \\ &= \textstyle\sum_{\ell\in\mathcal{I}_d} \omega_l \int_{\mathcal{O}_\Delta} m_\Delta(t_k, y^\ell_k(x)) m_\Delta(t_{k+1},x)\mathrm{d}x, \end{aligned} \tag{4.26}$$

and hence, by the Cauchy-Schwarz inequality,

$$\|m_\Delta(t_{k+1},\cdot)\|_{L^2} \le \max_{\ell\in\mathcal{I}_d} \left(\int_{\mathcal{O}_\Delta} |m_\Delta(t_k, y^\ell_k(x))|^2 \mathrm{d}x\right)^{1/2}. \tag{4.27}$$

In order to estimate the right-hand-side above, fix $x \in \mathbb{R}^d$, $\ell \in \mathcal{I}_d$ and notice that

$$Dy^\ell_k(x) = I_d + \frac{\Delta t}{2}\Big(D\mu(t_k,x) + D\mu(t_{k+1}, y^\ell_k(x)) Dy^\ell_k(x)\Big), \tag{4.28}$$

where $I_d$ denotes the $d \times d$ identity matrix. Assumption (**H1**)(ii) implies the existence of $\overline{\Delta t}$ such that for all $k \in \mathcal{I}^*_{\Delta t}$ and $\Delta t \in [0, \overline{\Delta t}]$, $y^\ell_k$ is one-to-one, and, for all $z \in \mathbb{R}^d$, the matrix $I_d - \frac{\Delta t}{2} D\mu(t_{k+1}, z)$ is invertible. Therefore, by (4.28),

$$Dy^\ell_k(x) = \left(I_d - \frac{\Delta t}{2} D\mu(t_{k+1}, y^\ell_k(x))\right)^{-1} \left(I_d + \frac{\Delta t}{2} D\mu(t_k,x)\right), \tag{4.29}$$

from which we deduce that $Dy^\ell_k(x)$ is invertible. Then, by the change of variable formula, we get that

$$\int_{\mathcal{O}_\Delta} |m_\Delta(t_k, y^\ell_k(x))|^2 \mathrm{d}x = \int_{y^\ell_k(\mathcal{O}_\Delta)} |m_\Delta(t_k,z)|^2 \left|\det\left(Dy^\ell_k((y^\ell_k)^{-1}(z))\right)\right|^{-1} \mathrm{d}z. \tag{4.30}$$

On the other hand, by (4.29), Jacobi's formula, and (**H1**)(ii), for all $x \in \mathbb{R}^d$ we have

$$\begin{aligned} \left[\det\left(Dy^\ell_k(x)\right)\right]^{-1} &= \frac{\det\left(I_d - \frac{\Delta t}{2} D\mu(t_{k+1}, y^\ell_k(x))\right)}{\det\left(I_d + \frac{\Delta t}{2} D\mu(t_k,x)\right)} \\ &= \frac{1 - \frac{\Delta t}{2}\mathrm{Tr}\left(D\mu(t_{k+1}, y^\ell_k(x))\right) + O((\Delta t)^2)}{1 + \frac{\Delta t}{2}\mathrm{Tr}(D\mu(t_k,x)) + O((\Delta t)^2)} \\ &= \frac{1 - \frac{\Delta t}{2}\mathrm{div}\left(\bar\mu(t_{k+1}, y^\ell_k(x))\right) + O((\Delta t)^2)}{1 + \frac{\Delta t}{2}\mathrm{div}(\mu(t_k,x)) + O((\Delta t)^2)}. \end{aligned} \tag{4.31}$$

Thus, by assumption (**H1**)(ii), there exists a constant $C > 0$, independent of $x$, $k$, $\ell$, and $\Delta t$, such that

$$\left|\left[\det\left(Dy^\ell_k(x)\right)\right]^{-1}\right| \le 1 + C\Delta t. \tag{4.32}$$

Combining the previous inequality and (4.30) yields

$$\int_{\mathcal{O}_\Delta} |m_\Delta(t_k, y^\ell_k(x))|^2 \mathrm{d}x \le (1 + C\Delta t)\|m_\Delta(t_k,\cdot)\|^2_{L^2}, \tag{4.33}$$

and hence, by (4.27),

$$\|m_\Delta(t_{k+1},\cdot)\|_{L^2} \le (1+C\Delta t)^{\frac{1}{2}}\|m_\Delta(t_k,\cdot)\|_{L^2}^2.$$

Thus,

$$\|m_\Delta(t_{k+1},\cdot)\|_{L^2} \le \left(1+\frac{CT}{N_{\Delta t}}\right)^{N_{\Delta t}/2}\|m_\Delta(0,\cdot)\|_{L^2} \le e^{CT/2}\|m_\Delta(0,\cdot)\|_{L^2},$$

from which assertion (iii) follows. □

**Remark 82.** *Notice that Proposition 81*(iii) *and the Cauchy-Schwarz inequality imply that, for any compact set $K \subseteq \mathbb{R}^d$, there exists $C_K > 0$, independent of $\Delta t$ and $\Delta x$, such that*

$$\max_{k\in\mathcal{I}_{\Delta t}} \int_K |m_\Delta(t_k,x)|\mathrm{d}x \le C_K.$$

In the following, we still denote by $m_\Delta$ its extension to $[0,T]\times\mathcal{O}_\Delta$, defined as

$$m_\Delta(t,x) = \frac{t-t_k}{\Delta t}m_\Delta(t_{k+1},x) + \frac{t_{k+1}-t}{\Delta t}m_\Delta(t_k,x) \tag{4.34}$$

if $(t,x)\in[t_k,t_{k+1}]\times\mathcal{O}_\Delta$ $(k\in\mathcal{I}^*_{\Delta t})$.

Notice that (4.34) and Theorem 59(ii)-(iii) imply that

$$\int_{\mathcal{O}_\Delta} m_\Delta(t,x)\mathrm{d}x = 1 \quad \text{for all } t\in[0,T] \quad \text{and} \quad \max_{t\in[0,T]}\|m_\Delta(t,\cdot)\|_{L^2} \le C, \tag{4.35}$$

for some $C>0$, independent of $\Delta$ for $\Delta t$ small enough.

**Proposition 83.** *Under* **(H1)**, *the following assertions hold true:*
(i)[Equicontinuity] *Let $\phi\in C_0^\infty(\mathbb{R}^d)$ and $\Delta t_0>0$. Then there exists $C_\phi>0$ such that for all $\Delta=(\Delta t,\Delta x)$ satisfying $\Delta t\le\Delta t_0$ and $(\Delta x)^{q+1}\le\Delta t$, we have*

$$\left|\int_{\mathbb{R}^d}\phi(x)m_\Delta(t,x)\mathrm{d}x - \int_{\mathbb{R}^d}\phi(x)m_\Delta(s,x)\mathrm{d}x\right| \le C_\phi\Delta t \quad \textit{for all } s,\, t\in[0,T]. \tag{4.36}$$

(ii)[Consistency] *Let $\phi\in C_0^\infty(\mathbb{R}^d)$, then for any $k\in\mathcal{I}^*_{\Delta t}$ and $(\Delta t,\Delta x)\in(0,+\infty)^2$, we have*

$$\int_{\mathbb{R}^d}\phi(x)\left(m_\Delta(t_{k+1},x) - m_\Delta(t_k,x)\right)\mathrm{d}x =$$

$$\int_{t_k}^{t_{k+1}}\int_{\mathbb{R}^d}\left(\frac{\sigma^2}{2}\Delta\phi(x) + \langle\mu(s,x),D\phi(x)\rangle\right)m_\Delta(s,x)\mathrm{d}x\mathrm{d}s + O((\Delta x)^{q+1}+(\Delta t)^2). \tag{4.37}$$

*Proof.* In the proof of both assertions, we fix $\phi\in C_0^\infty(\mathbb{R}^d)$ and we will denote by $C$ a positive real number which can depend on $\phi$ but not on $\Delta t$ and $\Delta x$. We will also use the estimate

$$\left|\sum_{\ell\in\mathcal{I}_d}\omega_\ell\phi(y_k^\ell(x)) - \left[\phi(x)+\Delta t\left(\frac{\sigma^2}{2}\Delta\phi(x)+\langle\mu(x,t_k),D\phi(x)\rangle\right)\right]\right| \le C(\Delta t)^2, \tag{4.38}$$

for $x \in \mathbb{R}^d$, which follows from the definition of $y_k^\ell(x)$ and a Taylor expansion (see for instance [17]).

(i) Let us first show the assertion for $t = t_{k+1}$ and $s = t_k$ for some $k \in \mathcal{I}^*_{\Delta t}$. Set $\varepsilon := \phi - I[\phi]$ and fix $k \in \mathcal{I}^*_{\Delta t}$. Remark 82 yields the existence of $C > 0$ such that

$$\begin{aligned} &\left| \int_{\mathbb{R}^d} \phi(x) \left( m_\Delta(t_{k+1}, x) - m_\Delta(t_k, x) \right) \mathrm{d}x \right| \leq \\ &\left| \int_{\mathbb{R}^d} I[\phi](x) \left( m_\Delta(t_{k+1}, x) - m_\Delta(t_k, x) \right) \mathrm{d}x \right| + C \| \varepsilon \|_{L^\infty}. \end{aligned} \tag{4.39}$$

Recalling that $\mathrm{supp}\{m_\Delta(t_k, \cdot)\} \subset \mathcal{O}_\Delta$ and using the definition of the scheme in (4.20), we have that

$$\begin{aligned} &\int_{\mathbb{R}^d} I[\phi](x) \left( m_\Delta(t_{k+1}, x) - m_\Delta(t_k, x) \right) \mathrm{d}x \\ &= \int_{\mathcal{O}_\Delta} \sum_{i \in \mathbb{Z}^d} \phi(x_i) \beta_i(x) \left( \sum_{j \in \mathcal{I}_{\Delta x}} \left( m_{k+1,j} - m_{k,j} \right) \beta_j(x) \right) \mathrm{d}x \\ &= \sum_{i \in \mathbb{Z}^d} \phi(x_i) \left( \sum_{j \in \mathcal{I}_{\Delta x}} \left( m_{k+1,j} - m_{k,j} \right) \int_{\mathcal{O}_\Delta} \beta_i(x) \beta_j(x) \mathrm{d}x \right) \end{aligned}$$

which leads to

$$\begin{aligned} &\int_{\mathbb{R}^d} I[\phi](x) \left( m_\Delta(t_{k+1}, x) - m_\Delta(t_k, x) \right) \mathrm{d}x \\ &= \sum_{i \in \mathbb{Z}^d} \phi(x_i) \Big[ \sum_{\ell \in \mathcal{I}} \omega_\ell \sum_{j \in \mathcal{I}_{\Delta x}} m_{k,j} \Big( \int_{\mathcal{O}_\Delta} \beta_i(y_k^\ell(x)) \beta_j(x) \mathrm{d}x \int_{\mathcal{O}_\Delta} \beta_i(x) \beta_j(x) \mathrm{d}x \Big) \Big] \\ &= \sum_{\ell \in \mathcal{I}} \omega_\ell \sum_{j \in \mathcal{I}_{\Delta x}} m_{k,j} \int_{\mathcal{O}_\Delta} \Big[ I[\phi](y_k^\ell(x)) - I[\phi](x) \Big] \beta_j(x) \mathrm{d}x \\ &= \sum_{\ell \in \mathcal{I}} \omega_\ell \int_{\mathcal{O}_\Delta} \Big[ I[\phi](y_k^\ell(x)) - I[\phi](x) \Big] m_\Delta(t_k, x) \mathrm{d}x. \end{aligned} \tag{4.40}$$

On the other hand, since $\phi$ has a compact support, there exists $C > 0$ such that

$$\left\| \sum_{\ell \in \mathcal{I}} \omega_\ell \left( I[\phi](y_k^\ell(\cdot)) - \phi(y_k^\ell(\cdot)) \right) \right\|_{L^2} + \| \phi - I[\phi] \|_{L^2} \leq C \| \varepsilon \|_{L^\infty} \tag{4.41}$$

and, by (4.38) and **(H1)**(ii), there exists $C > 0$ such that

$$\left\| \sum_{\ell \in \mathcal{I}} \omega_\ell \left( \phi(y_k^\ell(\cdot)) - \phi \right) \right\|_{L^2} \leq C \Delta t. \tag{4.42}$$

Thus, by the triangular and the Cauchy-Schwarz inequalities, Theorem 81(iii), (4.39), (4.40), (4.41), and (4.42), we get the existence of $C > 0$ such that

$$\left| \int_{\mathbb{R}^d} \phi(x) \left( m_\Delta(t_{k+1}, x) - m_\Delta(t_k, x) \right) \mathrm{d}x \right| \leq C \left( \| \varepsilon \|_{L^\infty} + \Delta t \right)$$

and hence it follows from (4.18) and the condition $(\Delta x)^{q+1} \leq \Delta t$ the existence of $C > 0$ such that (4.36) holds for $t = t_{k+1}$ and $s = t_k$. Using this relation and the triangular inequality, we deduce that (4.36) holds for every $s = t_k$ and $t = t_m$ with $k$, $m \in \mathcal{I}_{\Delta t}$.

Now, let us fix $s, t \in [0, T]$ and assume, without loss of generality, that $t > s$. Let $k_1$, $k_2 \in \mathcal{I}^*_{\Delta t}$ be such that $s \in [t_{k_1}, t_{k_1+1}]$ and $t \in [t_{k_2}, t_{k_2+1}]$. By (4.34), we have that

$$\left| \int_{\mathbb{R}^d} \phi(x) \left( m_\Delta(t_{k_1+1}, x) - m_\Delta(s, x) \right) \mathrm{d}x \right| \leq \tag{4.43}$$

$$\frac{t_{k_1+1} - s}{\Delta t} \left| \int_{\mathbb{R}^d} \phi(x) \left( m_\Delta(t_{k_1+1}, x) - m_\Delta(t_{k_1}, x) \right) \mathrm{d}x \right| \leq t_{k_1+1} - s.$$

Similarly,

$$\left| \int_{\mathbb{R}^d} \phi(x) \left( m_\Delta(t_{k_2}, x) - m_\Delta(t, x) \right) \mathrm{d}x \right| \leq t - t_{k_2}. \tag{4.44}$$

Thus, (4.36) follows from the triangular inequality, (4.43), (4.44), and (4.36) with $t = t_{k_2}$ and $s = t_{k_1+1}$.

(ii) By (4.18), Remark 82, and the definition of the scheme (4.20), for each $k \in \mathcal{I}^*_{\Delta t}$ we have

$$\begin{aligned}
\int_{\mathbb{R}^d} \phi(x) m_\Delta(t_{k+1}, x) \mathrm{d}x &= \int_{\mathbb{R}^d} I[\phi](x) m_\Delta(t_{k+1}, x) \mathrm{d}x + O((\Delta x)^{q+1}) \\
&= \sum_{i \in \mathbb{Z}^d} \phi(x_i) \sum_{j \in \mathcal{I}_{\Delta x_n}} m_{k+1,j} \int_{\mathbb{R}^d} \beta_i(x) \beta_j(x) \mathrm{d}x \\
&\quad + O((\Delta x)^{q+1}) \\
&= \sum_{i \in \mathbb{Z}^d} \phi(x_i) \sum_{j \in \mathcal{I}_{\Delta x}} m_{k,j} \sum_{\ell \in \mathcal{I}_d} \omega_\ell \int_{\mathbb{R}^d} \beta_i(y^\ell_k(x)) \beta_j(x) \mathrm{d}x \\
&\quad + O((\Delta x)^{q+1}) \\
&= \sum_{j \in \mathcal{I}_{\Delta x_n}} m_{k,j} \sum_{\ell \in \mathcal{I}_d} \omega_\ell \int_{\mathbb{R}^d} I[\phi](y^\ell_k(x)) \beta_j(x) \mathrm{d}x \\
&\quad + O((\Delta x)^{q+1}) \\
&= \sum_{j \in \mathcal{I}_{\Delta x_n}} m_{k,j} \int_{\mathbb{R}^d} \left( \sum_{\ell \in \mathcal{I}_d} \omega_\ell \phi(y^\ell_k(x)) \right) \beta_j(x) \mathrm{d}x \\
&\quad + O((\Delta x)^{q+1}).
\end{aligned} \tag{4.45}$$

Using (4.38) and Remark 82, we obtain

$$\int_{\mathbb{R}^d} \phi(x) \left( m_\Delta(t_{k+1}, x) - m_\Delta(t_k, x) \right) \mathrm{d}x$$

$$= \Delta t \int_{\mathbb{R}^d} \left( \frac{\sigma^2}{2} \Delta \phi(x) + \langle \mu(t_k, x), D\phi(x) \rangle \right) m_\Delta(t_k, x) \mathrm{d}x$$

$$+ O((\Delta x)^{q+1} + (\Delta t)^2).$$

By (4.34), Assumption **(H1)**(ii), and assertion (i), for every $s \in [t_k, t_{k+1}]$, we have

$$\left| \int_{t_k}^{t_{k+1}} \int_{\mathbb{R}^d} \left( \frac{\sigma^2}{2} \Delta\phi(x) + \langle \mu(t_k, x), D\phi(x) \rangle \right) (m_\Delta(s,x) - m_\Delta(t_k,x)) \mathrm{d}x\mathrm{d}s \right| = O((\Delta t)^2). \quad (4.46)$$

Thus, by using Assumption **(H1)**(ii) again and Remark 82, we obtain (4.37). □

Let us denote by $\mathcal{D}'(\mathbb{R}^d)$ the space of distributions, which we endow with the weak$^*$ topology. In the following, for every $\Delta \in (0,\infty)^2$ and $t \in [0,T]$, we identify $m_\Delta(t,\cdot)$ with the regular distribution

$$C_0^\infty(\mathbb{R}^d) \ni \phi \mapsto \int_{\mathbb{R}^d} \phi(x) m_\Delta(t,x) \mathrm{d}x \in \mathbb{R}.$$

For every $\Delta = (\Delta t, \Delta x) \in (0,\infty)^2$, let us denote, with a slight abuse of notation, $m_\Delta$ the map $[0,T] \ni t \mapsto m_\Delta(t,\cdot) \in \mathcal{D}'(\mathbb{R}^d)$. Notice that Proposition 83(i) implies that $m_\Delta \in C([0,T]; \mathcal{D}'(\mathbb{R}^d))$.

**Lemma 84.** *There exists $\Delta t_0 > 0$ such that the family $\mathcal{M} = \{m_\Delta \,|\, \Delta t \le \Delta t_0,\ (\Delta x)^{q+1} \le \Delta t\}$ is relatively compact in $C([0,T]; \mathcal{D}'(\mathbb{R}^d))$.*

*Proof.* In view of the Arzelà-Ascoli theorem [69, Chapter 7, Theorem 18] (see also [71, Section 4]) and Proposition 83(i), it suffices to show that the family $\mathcal{M}$ is pointwise relatively compact. Let us consider the absolutely convex set $U_0 := \{\phi \in C_0^\infty(\mathbb{R}^d) \mid \|\phi\|_{L^\infty} < 1,\ \operatorname{supp}\phi \subseteq \overline{B}(0,1)\}$. This set is a neighborhood of 0 in the standard topology of $C_0^\infty(\mathbb{R}^d)$ (see e.g. [107, Chapter 10]) and, for any $t \in [0,T]$,

$$\sup_{\phi \in U_0} \left| \int_{\mathbb{R}^d} m_\Delta(t,x)\phi(x)\mathrm{d}x \right| = \sup_{\phi \in U_0} \left| \int_{\overline{B}(0,1)} m_\Delta(t,x)\phi(x)\mathrm{d}x \right|$$

$$\le \|m_\Delta(t,\cdot)\|_{L^1(\overline{B}(0,1))} \le r,$$

where $r := \sup\{\|m_\Delta(t,\cdot)\|_{L^1(\overline{B}(0,1))} \,|\, \Delta \in (0,\infty)^2\}$ belongs to $[0,+\infty)$ by (4.35). This proves that $\{m_\Delta(t,\cdot) \,|\, \Delta \in (0,\infty)^2\} \subset \left\{T \in \mathcal{D}'(\mathbb{R}^d) \mid \sup_{\phi\in U_0} |T(\phi)| \le r\right\}$ which, by the Banach-Alaoglu-Bourbaki theorem (see e.g. [82, Theorem 23.5]), is a compact subset of $\mathcal{D}'(\mathbb{R}^d)$. □

We now show a convergence result.

**Proposition 85.** *Assume that $\overline{m}_0 \in H^{q+1}(\mathbb{R}^d)$ and that* **(H1)** *holds. Consider a sequence $(\Delta_n)_{n\in\mathbb{N}} = ((\Delta t_n, \Delta x_n))_{n\in\mathbb{N}} \subseteq (0,\infty)^2$ such that, as $n \to \infty$, $(\Delta t_n, \Delta x_n) \to (0,0)$ and $(\Delta x_n)^{q+1}/\Delta t_n \to 0$. Given $n \in \mathbb{N}$, set $m^n := m_{\Delta_n}$ the solution to (4.20). Then, up to subsequence, $m^n \to m^*$ in $C([0,T]; \mathcal{D}'(\mathbb{R}^d))$ and weakly in $L^2\left([0,T]\times\mathbb{R}^d\right)$ as $n \to \infty$, where $m^*$ is the unique classical solution to* (**FP**).

*Proof.* By Theorem 81(iii), the sequence $(m^n)_{n\in\mathbb{N}}$ is bounded in $L^2([0,T]\times\mathbb{R}^d)$. Thus, there exists $\widehat{m}$ in $L^2([0,T]\times\mathbb{R}^d)$ such that, as $n\to\infty$ and up to some subsequence, $m^n$ converges weakly to $\widehat{m}$ in $L^2([0,T]\times\mathbb{R}^d)$.

Let us first show that for any $\phi\in C_0^\infty((0,T)\times\mathbb{R}^d)$, we have

$$\int_0^T\int_{\mathbb{R}^d}\left[\partial_t\phi(t,x)-\frac{\sigma^2}{2}\Delta\phi(t,x)-\langle\mu(s,x),D\phi(t,x)\rangle\right]\widehat{m}(t,x)\mathrm{d}x\mathrm{d}t=0. \quad (4.47)$$

Let $\eta\in C_0^\infty([0,T])$, $\psi\in C_0^\infty(\mathbb{R}^d)$ and define $\phi=\eta\psi\in C_0^\infty([0,T]\times\mathbb{R}^d)$. Denote by $K\subset\mathbb{R}^d$ the support of $\psi$. By (4.34) and Proposition 83(i), we have

$$\begin{aligned}\int_0^T\int_{\mathbb{R}^d}\partial_t\phi(t,x)m^n(t,x)\mathrm{d}x\mathrm{d}t&=\sum_{k=0}^{N_{\Delta t_n-1}}\int_{t_k}^{t_{k+1}}\int_K\partial_t\phi(t,x)m^n(t_k,x)\mathrm{d}x\mathrm{d}t\\&+\sum_{k=0}^{N_{\Delta t_n-1}}\int_{t_k}^{t_{k+1}}\int_K\partial_t\phi(t,x)(m^n(t_{k+1},x)-m^n(t_k,x))\frac{t-t_k}{\Delta t_n}\mathrm{d}x\mathrm{d}t \quad (4.48)\\&=\sum_{k=0}^{N_{\Delta t-1}}\int_{t_k}^{t_{k+1}}\int_K\partial_t\phi(t,x)m^n(t_k,x)\mathrm{d}x\mathrm{d}t+O(\Delta t_n).\end{aligned}$$

On the other hand, by Remark 82

$$\begin{aligned}&\sum_{k=0}^{N_{\Delta t_n-1}}\int_{t_k}^{t_{k+1}}\int_K\partial_t\phi(t,x)m^n(t_k,x)\mathrm{d}x\mathrm{d}t\\&=\sum_{k=0}^{N_{\Delta t_n-1}}\Delta t_n\int_K\partial_t\phi(t_k,x)m^n(t_k,x)\mathrm{d}x+O\left(\Delta t_n\right)\\&=\sum_{k=0}^{N_{\Delta t_n}-1}\Delta t_n\dot{\eta}(t_k)\int_K\psi(x)m^n(t_k,x)\mathrm{d}x+O(\Delta t_n) \quad (4.49)\\&=\sum_{k=0}^{N_{\Delta t_n}-1}\left(\eta(t_{k+1})-\eta(t_k)\right)\int_K\psi(x)m^n(t_k,x)\mathrm{d}x+O(\Delta t_n)\\&=\sum_{k=0}^{N_{\Delta t_n}-2}\eta(t_{k+1})\left(\int_K\psi(x)[m^n(t_k,x)-m^n(t_{k+1},x)]\mathrm{d}x\right)+O(\Delta t_n).\end{aligned}$$

By (4.48), (4.49) and using the fact that $\phi$ is equal to zero outside $K$ we get

$$\begin{aligned}&\int_0^T\int_{\mathbb{R}^d}\partial_t\phi(t,x)m^n(t,x)\mathrm{d}x\mathrm{d}t\\&=\sum_{k=0}^{N_{\Delta t_n}-2}\eta(t_{k+1})\left(\int_{\mathbb{R}^d}\psi(x)[m^n(t_k,x)-m^n(t_{k+1},x)]\mathrm{d}x\right)+O(\Delta t_n).\end{aligned} \quad (4.50)$$

Using (4.50) and Proposition 83(ii) we have

$$\int_0^T \int_{\mathbb{R}^d} \partial_t \phi(t,x) m^n(t,x) \mathrm{d}x\mathrm{d}t$$

$$= \sum_{k=0}^{N_{\Delta t_n}-1} \eta(t_{k+1}) \int_{t_k}^{t_{k+1}} \int_{\mathbb{R}^d} \left( \frac{\sigma^2}{2} \Delta\psi(x) + \langle \mu(s,x), D\psi(x) \rangle \right) m^n(s,x) \mathrm{d}x\mathrm{d}s$$

$$+ O((\Delta x_n)^{q+1}/\Delta t_n + (\Delta t_n))$$

$$= \int_0^T \int_{\mathbb{R}^d} \left( \frac{\sigma^2}{2} \Delta\phi(t,x) + \langle \mu(s,x), D\phi(t,x) \rangle \right) m^n(t,x) \mathrm{d}x\mathrm{d}t$$

$$+ O((\Delta x_n)^{q+1}/\Delta t_n + (\Delta t_n)).$$

Thus,

$$\int_0^T \int_{\mathbb{R}^d} \left[ \partial_t \phi(t,x) - \frac{\sigma^2}{2} \Delta\phi(t,x) - \langle \mu(s,x), D\phi(t,x) \rangle \right] m^n(t,x) \mathrm{d}x\mathrm{d}t$$

$$= O((\Delta x_n)^{q+1}/\Delta t_n + (\Delta t_n))$$

and hence, passing to the weak limit in $L^2([0,T] \times \mathbb{R}^d)$, we get

$$\int_0^T \int_{\mathbb{R}^d} \left[ \partial_t \phi(t,x) - \frac{\sigma^2}{2} \Delta\phi(t,x) - \langle \mu(s,x), \nabla\phi(t,x) \rangle \right] \widehat{m}(t,x) \mathrm{d}x\mathrm{d}t = 0. \quad (4.51)$$

Since the vector space spanned by $\{\eta\psi \mid \eta \in C_0^\infty((0,T)),\ \psi \in C_0^\infty(\mathbb{R}^d)\}$ is dense in $C_0^{1,2}((0,T) \times \mathbb{R}^d)$ (as in [89, Corollary 1.6.2 of the Weierstrass Approximation Theorem]), we get that (4.47) holds for any $\phi \in C_0^{1,2}((0,T) \times \mathbb{R}^d)$.

Finally, let us show that for any $\phi \in C_0(\mathbb{R}^d)$

$$\int_{\mathbb{R}^d} \phi(x)(\widehat{m}(t,x) - \overline{m}_0(x)) \mathrm{d}x \to 0 \quad \text{as } t \to 0. \quad (4.52)$$

By Lemma 84, we have that $\widehat{m} \in C([0,T]; \mathcal{D}'(\mathbb{R}^d))$. Moreover, by [56, Lemma 2.1], for any $t \in [0,T]$ and for every $\phi \in C_0(\mathbb{R}^d)$, it holds that

$$\lim_{s \to t,\, s \in [0,T]} \int_{\mathbb{R}^d} \phi(x) \widehat{m}(s,x) \mathrm{d}x = \int_{\mathbb{R}^d} \phi(x) \widehat{m}(t,x) \mathrm{d}x. \quad (4.53)$$

Since Theorem 81(i) implies that $\widehat{m}(0,\cdot) = \overline{m}_0(\cdot)$, (4.52) follows from (4.53) with $t = 0$.

The result follows from (4.47), (4.52) and [56, Theorem 4.3]. □

**Remark 86.** *The convergence of the sequence $(m^n)_{n \in \mathbb{N}}$ to $m^*$ in the previous proposition is rather weak. On the other hand, to the best of our knowledge this is the first convergence result of a high order LG scheme for equation* (**FP**). *Notice that our proof does not depend on the smoothness of $m^*$ recalled in Theorem 80*(i)*, but it can be easily adapted to deal with equations whose second order term are not uniformly elliptic (see e.g. [52, 31] and the numerical test in Section 4.4.2 below).*

## 4.3 Application to Mean Field Games

Mean Field Game problems, introduced by Lasry and Lions in [75, 76, 77], characterize Nash equilibria of symmetric stochastic differential games with an infinite number of players.

Let $(\mathcal{P}_1(\mathbb{R}^d), \mathbf{d})$ be the metric space of Borel probability measures on $\mathbb{R}^d$ with finite first order moment, endowed with the 1-Wasserstein distance $\mathbf{d}$ (see e.g. [4, Section 7.1] for the definition of $\mathbf{d}$).

In this section, we focus on the numerical approximation of the following time-dependent second order MFG with nonlocal couplings (see e.g. [77, 76]):

$$\begin{aligned} -\partial_t v - \tfrac{\sigma^2}{2}\Delta v + H(x, \nabla v) &= F(x, m(t)) \text{ in } [0,T)\times\mathbb{R}^d, \\ \partial_t m - \tfrac{\sigma^2}{2}\Delta m - \operatorname{div}\left(\partial_p H(x,\nabla v) m\right) &= 0 \text{ in } (0,T]\times\mathbb{R}^d, \\ v(T,\cdot) = G(\cdot, m(T)), \qquad & m(0,\cdot) = \overline{m}_0 \text{ in } \mathbb{R}^d, \end{aligned} \tag{MFG}$$

where $\sigma \in \mathbb{R}\setminus\{0\}$, $\mathbb{R}^d\times\mathbb{R}^d \ni (x,p) \mapsto H(x,p) \in \mathbb{R}$ is convex and differentiable with respect to $p$, $F$, $G : \mathbb{R}^d \times \mathcal{P}_1(\mathbb{R}^d) \to \mathbb{R}$, and $\overline{m}_0 : \mathbb{R}^d \to \mathbb{R}$. Notice that (**MFG**) consists of a Hamilton-Jacobi-Bellman (HJB) equation, with a terminal condition, coupled with a FP equation with an initial condition.

For the sake of simplicity, in what follows we will suppose that the *Hamiltonian* $H$ is quadratic, i.e. $H(x,p) = |p|^2/2$ for all $x$, $p \in \mathbb{R}^d$.

**(H2)** We assume that:

(i) $\overline{m}_0$ is Hölder continuous and satisfies **(H1)**(i).

(ii) $F$ and $G$ are bounded and Lipschitz continuous. Moreover, for every $m \in \mathcal{P}_1(\mathbb{R}^d)$, $F(\cdot, m)$ is of class $C^2$ and

$$\sup_{x\in\mathbb{R}^d, m\in\mathcal{P}_1(\mathbb{R}^d)} \left\{ \|DF(x,m)\|_\infty + \|D^2F(x,m)\|_\infty \right\} < \infty.$$

Under **(H2)** system (**MFG**) admits at least one classical solution (see e.g. [28, Theorem 3.1]). Moreover, if the coupling terms $F$ and $G$ satisfy a monotonicity condition with respect to $m$, then the classical solution is unique (see [77, Theorem 2.4]).

In the following, in order to obtain a second order scheme for (**MFG**), we consider a second order Semi-Lagrangian (SL) scheme for the HJB equation, which will be combined with the scheme (4.20) for the FP equation.

### 4.3.1 A semi-Lagrangian scheme for the HJB equation

Given $m \in C([0,T]; \mathcal{P}_1(\mathbb{R}^d))$, we consider the HJB equation:

$$\begin{aligned} -\partial_t v - \tfrac{\sigma^2}{2}\Delta v + \tfrac{1}{2}|\nabla v|^2 &= F(x, m(t)) \quad \text{in } (0,T)\times\mathbb{R}^d, \\ v(T,\cdot) &= G(\cdot, m(T)) \quad \text{in } \mathbb{R}^d. \end{aligned} \tag{HJB}$$

Standard results for quasilinear parabolic equations (see e.g. [73, Chapter IV and V]) yield that (**HJB**) admits a unique classical solution $v[m]$. Moreover,

using that $v[m]$ is the value function associated to a stochastic optimal control problem (see e.g. [58, Chapters IV and V]), it is easy to check that **(H2)** yields the existence of $R > 0$ such that

$$|\nabla v[m](t,x)| \le R \quad \text{for all } t \in [0,T],\, x \in \mathbb{R}^d,\, m \in C([0,T];\mathcal{P}_1(\mathbb{R}^d)).$$

We now describe a variation of the scheme in [17] to deal with the nonlinearity of the Hamiltonian in (**HJB**) with respect to $\nabla v$ (see also [85, 92] for related constructions). For a given $m \in C([0,T];\mathcal{P}_1(\mathbb{R}^d))$, let us define $\{v_{k,i} \,|\, k \in \mathcal{I}_{\Delta t},\, i \in \mathcal{I}_{\Delta x}\} \subset \mathbb{R}$ as the solution to

$$\begin{aligned} v_{k,i} &= S[m](v_{\cdot,k+1},k,i) \quad \text{for all } k \in \mathcal{I}^*_{\Delta t},\, i \in \mathcal{I}_{\Delta x}, \\ v_{N_{\Delta t},i} &= G(x_i, m(t_{N_{\Delta t}})) \quad \text{for all } i \in \mathcal{I}_{\Delta x}, \end{aligned} \tag{4.54}$$

where, for a given $f = \{f_i\}_{i\in\mathcal{I}_{\Delta x}} \subset \mathbb{R}$, $k \in \mathcal{I}^*_{\Delta t}$, and $i \in \mathcal{I}_{\Delta x}$,

$$\begin{aligned} S[m](f,k,i) = \inf_{\alpha\in A} \Bigg[ \sum_{\ell\in\mathcal{I}_d} \omega_\ell \Big( I[f](x_i - \Delta t\alpha + \sqrt{\Delta t}\sigma e^\ell) \\ + \frac{\Delta t}{2} F(x_i - \Delta t\alpha + \sqrt{\Delta t}\sigma e^\ell, m(t_{k+1})) \Big) + \frac{\Delta t}{2}|\alpha|^2 \Bigg] + \frac{\Delta t}{2} F(x_i, m(t_k)), \end{aligned} \tag{4.55}$$

with $A = \{\alpha \in \mathbb{R}^d \,|\, |\alpha| \le R\}$ and $I[f]$ being defined by (4.17). The following consistency result for $S[m]$ follows from (4.55) and **(H2)**.

**Proposition 87.** *Let $(\Delta t_n, \Delta x_n)_{n\in\mathbb{N}} \subset (0,+\infty)^2$, $(k_n)_{n\in\mathbb{N}} \subseteq \mathbb{N}$, $(i_n)_{n\in\mathbb{N}} \subset \mathbb{Z}^d$, $(m_n)_{n\in\mathbb{N}} \subset C([0,T];\mathcal{P}_1(\mathbb{R}^d))$, and $m \in C([0,T];\mathcal{P}_1(\mathbb{R}^d))$. Assume that* **(H2)**(ii) *holds and, as $n\to\infty$, $(\Delta t_n, \Delta x_n) \to (0,0)$, $(\Delta x_n)^{q+1}/\Delta t_n \to 0$, $k_n \in \mathcal{I}_{\Delta t_n}$, $i_n \in \mathcal{I}_{\Delta x_n}$, $t_{k_n} \to t$, $x_{i_n} \to x$, and $m_n \to m$. Then for every $\phi \in C^{1,3}_b\left([0,T]\times\mathbb{R}^d\right)$, satisfying $\|\nabla\phi\|_{L^\infty([0,T]\times\mathbb{R}^d)} \le R$, we have*

$$\begin{aligned} \lim_{n\to\infty} \frac{1}{\Delta t_n}\left[\phi(t_{k_n}, x_{i_n}) - S[m_n](\phi_{k_n+1}, k_n, i_n)\right] = \\ -\partial_t\phi(t,x) - \frac{\sigma^2}{2}\Delta\phi(t,x) + \frac{1}{2}|\nabla\phi(t,x)|^2 - F(x,m(t)), \end{aligned}$$

*where $\phi_k = \{\phi(t_k, x_i)\}_{i\in\mathcal{I}_{\Delta x}}$.*

*Proof.* Let $\Delta t > 0$, $\Delta x > 0$, and $\alpha \in A$. In the computations below, the big $O$ terms are uniform with respect to $\alpha \in A$. Let us apply (4.38) to $\phi(t_{k+1},\cdot)$, with $b(t,x) = -\alpha$, to obtain

$$\begin{aligned} \sum_{\ell\in\mathcal{I}_d} \omega_\ell \phi\left(t_{k+1}, x_i - \Delta t\alpha + \sqrt{\Delta t}\sigma e^\ell\right) = \\ \phi\left(t_{k+1}, x_i\right) + \Delta t\left(\frac{\sigma^2}{2}\Delta\phi(t_{k+1}, x_i) - \langle\nabla\phi(t_{k+1}, x_i), \alpha\rangle\right) + O\left((\Delta t)^2\right). \end{aligned} \tag{4.56}$$

By (**H2**)(ii) and using the first-order Taylor expansion of $F(\cdot, m(t_k))$ around $x_i$, we get

$$\begin{aligned} \frac{1}{2}\left(\sum_{\ell\in\mathcal{I}_d} \omega_\ell F(x_i - \Delta t\alpha + \sqrt{\Delta t}\sigma e^\ell, m(t_{k+1})) + F(x_i, m(t_k))\right) = \\ F(x_i, m(t_{k+1})) + O(\Delta t + \mathbf{d}(m(t_{k+1}), m(t_k))). \end{aligned} \tag{4.57}$$

Thus, by (4.55), (4.56), (4.57), and (4.18), we obtain

$$\begin{aligned} S[m](\phi_{k+1},k,i) &= \phi\left(t_{k+1},x_i\right) - \Delta t \sup_{\alpha\in A}\left[\langle\nabla\phi(t_{k+1},x_i),\alpha\rangle - \frac{|\alpha|^2}{2}\right] \\ &\quad +\Delta t\frac{\sigma^2}{2}\Delta\phi(t_{k+1},x_i) + \Delta t F(x_i,m(t_{k+1})) \\ &\quad +O\left((\Delta t)^2 + (\Delta x)^{q+1} + \Delta t\mathbf{d}(m(t_{k+1}),m(t_k))\right) \\ &= \phi\left(t_{k+1},x_i\right) - \frac{\Delta t}{2}|\nabla\phi(t_{k+1},x_i)|^2 \\ &\quad +\Delta t\tfrac{\sigma^2}{2}\Delta\phi(t_{k+1},x_i) + \Delta t F(x_i,m(t_{k+1})) \\ &\quad +O\left((\Delta t)^2 + (\Delta x)^{q+1} + \Delta t\mathbf{d}(m(t_{k+1}),m(t_k))\right). \end{aligned}$$

Finally, we get

$$\begin{aligned} \frac{1}{\Delta t}\left[\phi(t_k,x_i) - S_\Delta[m](\phi_{k+1},k,i)\right] = \\ -\partial_t\phi(t_{k+1},x_i) - \frac{\sigma^2}{2}\Delta\phi(t_{k+1},x_i) + \frac{1}{2}|\nabla\phi(t_{k+1},x_i)|^2 - F(x_i,m(t_{k+1})) \\ +O\left(\Delta t + \frac{(\Delta x)^{q+1}}{\Delta t} + \mathbf{d}(m(t_{k+1}),m(t_k))\right), \end{aligned}$$

from which the result follows. □

### 4.3.2 The scheme for MFG

For $m \in C([0,T];\mathcal{P}_1(\mathbb{R}^d))$, let us define

$$v_\Delta[m](t,x) := I[v_{[t/\Delta t]}](x) \quad \text{for all} \quad (t,x)\in[0,T]\times\mathcal{O}_\Delta, \tag{4.58}$$

where $v_{k,i}$ is given by (4.54). In order to get a differentiable function with respect to $x$, given $\varepsilon > 0$ and a non-negative function $\phi \in C^\infty(\mathbb{R}^d)$ such that $\int_{\mathbb{R}^d}\phi(x)\mathrm{d}x = 1$, let us set $\phi_\varepsilon(\cdot) = \frac{1}{\varepsilon^d}\phi(\cdot/\varepsilon)$ and define

$$v_{\Delta,\varepsilon}[m](t,\cdot) = (\phi_\varepsilon * v_\Delta[m])(t,\cdot) \quad \text{for all} \quad t\in[0,T]. \tag{4.59}$$

For $\ell \in \mathcal{I}_d$ and $k \in \mathcal{I}^*_{\Delta t}$, let us define $y^\ell_{k,\varepsilon}[m](x)$ the unique solution to

$$y = x - \frac{\Delta t}{2}(\nabla v_{\Delta,\varepsilon}[m](t_k,x) + \nabla v_{\Delta,\varepsilon}[m](t_{k+1},y) + \sqrt{\Delta t}\sigma e^\ell, \tag{4.60}$$

where $\nabla v_{\Delta,\varepsilon}[m](t,x)$ is the gradient of $v_{\Delta,\varepsilon}[m]$ with respect to $x$.

We propose the following scheme for (**MFG**): find $\{(v_{k,i},m_{k,i})\in\mathbb{R}^2 \,|\, k\in$

$\mathcal{I}_{\Delta t},\ i \in \mathcal{I}_{\Delta x}\}$ such that, for all $k \in \mathcal{I}^*_{\Delta t}$ and $i \in \mathcal{I}_{\Delta x}$,

$$\begin{aligned} v_{k,i} &= S_\Delta[m_\Delta](v_{k+1}, k, i), \\ v_{N_{\Delta t},i} &= G(x_i, m), \\ \sum_{j\in\mathcal{I}_{\Delta x}} m_{k+1,j} \int_{\mathcal{O}_\Delta} \beta_i(x)\beta_j(x)\mathrm{d}x &= \sum_{j\in\mathcal{I}_{\Delta x}} m_{k,j} \sum_{\ell\in\mathcal{I}_d} \omega_\ell \int_{\mathcal{O}_\Delta} \beta_i(y^\ell_{k,\varepsilon}[m_\Delta](x))\beta_j(x)\mathrm{d}x, \\ \sum_{j\in\mathcal{I}_{\Delta x}} m_{0,j} \int_{\mathcal{O}_\Delta} \beta_i(x)\beta_j(x)\mathrm{d}x &= \int_{\mathcal{O}_\Delta} \overline{m}_0(x)\beta_i(x)\mathrm{d}x. \end{aligned} \tag{4.61}$$

System (4.61) is solved by a fixed point method as in [32]. The iterations are stopped as soon as the $L^1$-norm, approximated by the Simpson's Rule, of the difference between two consecutive approximations of $m$ is less than a given tolerance $\tau > 0$.

## 4.4 Numerical results

In this section, we show the performance of the proposed scheme on three different problems: a linear FP equation in two spatial dimensions, a MFG with non-local couplings and a explicit solution and, finally, a MFG with local couplings and no explicit solutions. For each test, we measure the accuracy of the scheme by computing the following relative errors in the discrete uniform and $L^2$ norms

$$E_\infty = \frac{\max_{i\in\mathcal{I}_{\Delta x}} |h_\Delta(T, x_i) - h(T, x_i)|}{\max_{i\in\mathcal{I}_{\Delta x}} |h(T, x_i)|},$$

$$E_2 = \left( \frac{\mathrm{Int}_{\mathcal{O}_\Delta}(|h_\Delta(T, x) - h(T, x)|^2)}{\mathrm{Int}_{\mathcal{O}_\Delta}(|h(T, x)|^2)} \right)^{1/2},$$

where $h = m,\ v,\ h_\Delta = m_\Delta,\ v_\Delta$, and $\mathrm{Int}_{\mathcal{O}_\Delta}$ denotes the approximation of the Riemann integral on $\mathcal{O}_\Delta$ by using the Simpson's Rule. We denote by $p_\infty$ and $p_2$ the rates of convergence for $E_\infty$ and $E_2$, respectively. For these error measures, the tables show rates of convergence greater than 2 in most of the cases. For the exactly integrated scheme (4.20), the local truncation error is given by the contributions of (4.13) and (4.18), which yields a global truncation error of order $(\Delta x)^{q+1}/\Delta t + (\Delta t)^2$. As in [52], we get that the order of consistency is maximized by taking $\Delta t = O((\Delta x)^{(q+1)/3})$. With respect to the space discretization step, the previous choice suggests an order of convergence given by $2(q+1)/3$. In all the simulations we take $q = 3$, which yields an heuristic optimal rate equal to $8/3$.

### 4.4.1 An implementable version of the scheme (4.20) in dimension one

In order to obtain a implementable version of (4.23), an approximation of the integrals therein has to be introduced. For simplicity, we consider the one-dimensional case, we use Simpson's Rule on each element $[x_j, x_j + 2\Delta x]$

($j = 2m$, $m \in \mathbb{Z}$) and cubic symmetric Lagrange interpolation basis functions $\beta_j$ ($p = 1$ in (4.16)). Recalling that $\beta_j$ has support in $[x_{j-2}, x_{j+2}]$, letting $\delta_{i,j} = 1$ if $i = j$ and $\delta_{i,j} = 0$ otherwise, the entries of the mass matrix $A$ (see (4.21)) are approximated by

$$\int_{\mathcal{O}_\Delta} \beta_i(x)\beta_j(x)\mathrm{d}x = \int_{x_{j-2}}^{x_j} \beta_i(x)\beta_j(x)\mathrm{d}x + \int_{x_j}^{x_{j+2}} \beta_i(x)\beta_j(x)\mathrm{d}x \simeq \frac{2\Delta x}{3}\delta_{i,j} \tag{4.62}$$

and the entries of $B_k^\ell$ (see (4.22)) are approximated by

$$(B_k^\ell)_{i,j} = \int_{x_{j-2}}^{x_{j+2}} \beta_i(y_k^\ell(x))\beta_j(x)\mathrm{d}x \simeq \frac{2\Delta x}{3}\beta_i(y_k^\ell(x_j)). \tag{4.63}$$

We observe that, as usual in Lagrange-Galerkin methods, the integrands in (4.62) and (4.63) have not the necessary regularity in order to guarantee the standard accuracy order of the quadrature rule. This can lead to fluctuations in the order of convergence, as can be observed in some instances of the numerical tests below. However, in those tests we will see that the aforementioned quadrature rule provides an overall order of convergence close to 8/3.

Using (4.62) and (4.63), the scheme (4.23) is approximated by

$$\begin{array}{rcl} m_{k+1} & = & \sum_{\ell\in\mathcal{I}_d} \omega_\ell \widetilde{B}_k^\ell m_k \quad \text{for } k \in \mathcal{I}^*_{\Delta t}, \\ m_0 & = & \widetilde{m}_0, \end{array} \tag{4.64}$$

where $\widetilde{B}_k^\ell$ is a $(2N_{\Delta x}+1)\times(2N_{\Delta x}+1)$ matrix with entries given by

$$(\widetilde{B}_k^\ell)_{i,j} = \beta_i(y_k^\ell(x_j))$$

and $\widetilde{m}_0$ is vector of length $2N_{\Delta x}+1$ given by

$$\widetilde{m}_{0,i} = \overline{m}_0(x_i) \quad \text{for } i \in \mathcal{I}_{\Delta x}.$$

**Remark 88.** *Applied to a linearization of equation* (**HJB**)*, scheme* (4.64) *is the dual of the semi-Lagrangian scheme [52] when a Crank-Nicolson method is used to discretize the characteristic curves, together with a cubic symmetric Lagrange interpolation to reconstruct the values in the space variable. Moreover, scheme* (4.64) *is also a natural higher-order extension of the scheme proposed in [33, 31] to approximate second order MFGs.*

### 4.4.2 Linear case: damped noisy harmonic oscillator

We consider the numerical approximation of a FP equation modeling an noisy harmonic oscillator with damping coefficient $\gamma > 2$ and noise coefficient $\sigma > 0$. For $T > 0$ and an initial condition $x_0 \in \mathbb{R}^2$, the dynamics is described by the following SDE in the interval $(0, T)$

$$\begin{array}{rcl} \mathrm{d}Y_1(t) & = & Y_2(t)\mathrm{d}t \\ \mathrm{d}Y_2(t) & = & (-Y_1(t) - \gamma Y_2(t))\,\mathrm{d}t + \sigma \mathrm{d}W(t), \\ Y(0) & = & x_0. \end{array} \tag{4.65}$$

The associated (degenerated) FP equation is given by

$$\begin{aligned} \partial_t m - \frac{\sigma^2}{2}\partial^2_{x_2,x_2} m + \partial_{x_1}(x_2 m) - \partial_{x_2}((x_1 + \gamma x_2)m) &= 0 \quad \text{in } (0,T]\times\mathbb{R}^2, \\ m(0) &= \delta_{x_0} \quad \text{in } \mathbb{R}^2, \end{aligned} \tag{4.66}$$

where $\delta_{x_0}$ denotes the Dirac measure at $x_0$. It is shown in [111] that (4.66) has a unique solution $m^*$ such that, for all $t \in (0,T]$, $m^*(t)$ is absolutely continuous with respect to the Lebesgue measure, with density $m^*(t,\cdot)$ given by

$$m^*(t,x) = \frac{\nu(t,x)}{\int_{\mathbb{R}^d}\nu(t,y)\mathrm{d}y}, \quad \text{for all } x \in \mathbb{R}^d, \text{ where } \quad \nu(t,x) = \frac{e^{\gamma t - s_{x_0}(t,x)/2D(t)}}{2\pi\sqrt{D(t)}}, \tag{4.67}$$

with

$$\begin{aligned} s_{x_0}(t,x) = {} & a(t)(\psi(t,x) - \psi(0,x_0))^2 + 2H(t)\left[\psi(t,x) - \psi(0,x_0)\right]\left[\eta(t,x)\right. \\ & \left. -\eta(0,x_0))\right] + b(t)(\eta(t,x) - \eta(0,x_0))^2, \\ D(t) = {} & a(t)b(t) - H(t)^2, \end{aligned}$$

and, setting

$$\mu_1 = -\frac{\gamma}{2} + \sqrt{\frac{\gamma^2}{4} - 1}, \quad \mu_2 = -\frac{\gamma}{2} - \sqrt{\frac{\gamma^2}{4} - 1},$$

$a$, $\psi$, $H$, $\eta$, and $b$ are respectively given by

$$\begin{aligned} & a(t) = \tfrac{\sigma^2}{2\mu_1}(1 - e^{-2\mu_1 t}), \quad \psi(t,x) = (x_1\mu_1 - x_2)e^{-\mu_2 t}, \\ & H(t) = -\tfrac{\sigma^2}{\mu_1+\mu_2}(1 - e^{-(\mu_1+\mu_2)t}), \\ & \eta(t,x) = (x_1\mu_2 - x_2)e^{-\mu_1 t}, \quad \text{and} \quad b(t) = \tfrac{\sigma^2}{2\mu_2}(1 - e^{-2\mu_2 t}). \end{aligned}$$

We apply scheme (4.64) to approximate $m^*(t,\cdot)$ for $t \in [t_0, T] = [1.5, 3]$. We take $\gamma = 2.1$, two values for $\sigma^2/2$ given by 0.1 and 0.05, respectively, and $x_0 = (1,1)$. Since the SDE (4.65) is autonomous, it is sufficient to apply (4.64) to approximate (**FP**) in $[0, 1.5]$ with initial condition $\overline{m}_0(\cdot) = m^*(1.5,\cdot)$, the latter being computed by using (4.67). Since the diffusion term in (4.65) can be written as $(0,\sigma)\mathrm{d}W(t)$, the scheme (4.64) cannot be directly applied, but, as in [52], it can be simply modified by setting

$$A_\gamma = \begin{pmatrix} 0 & 1 \\ -1 & -\gamma \end{pmatrix}, \quad e^1 = \begin{pmatrix} 0 \\ -1 \end{pmatrix}, \quad e^2 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \quad e^3 = \begin{pmatrix} 0 \\ 1 \end{pmatrix},$$

and considering the discrete characteristics $y_k^\ell(x)$ ($\ell = 1, 2, 3$), defined as the unique solutions to

$$y = x + \frac{\Delta t}{2}A_\gamma(x+y) + \sqrt{\Delta t}\sigma e^\ell,$$

with corresponding weights $\omega_\ell$ ($\ell = 1, 2, 3$) given by $\omega_1 = 1/6$, $\omega_2 = 2/3$, and $\omega_3 = 1/6$. Since most of the support of the exact solution $m$ is contained in

$\mathcal{O}_\Delta = (-2,2)^2$, we consider the solution of our scheme restricted to this domain in order to obtain an implementable method. We consider homogeneous Dirichlet boundary conditions, which are approximated by taking $(\widetilde{B}_k^\ell)_{i,j} = 0$ in (4.64) if the characteristic $y_k^\ell(x_j)$ exits from $\mathcal{O}_\Delta$. Tables 4.1 and 4.2 show the errors and convergence rates in both norms. We have performed the simulations by taking $\Delta t = (\Delta x)^{4/3}/8$ in the case of $\sigma^2/2 = 0.1$ (Table 4.1), and $\Delta t = (\Delta x)^{4/3}/4$ in the case of $\sigma^2/2 = 0.05$ (Table 4.2). As for semi-Lagrangian schemes, the scheme (4.64) performs better in the hyperbolic regime case (small diffusion). In some simulations, the optimal rate of convergence 8/3 is reached.

| $\Delta x$ | Errors on the FP equation | | | |
|---|---|---|---|---|
| | $E_\infty$ | $E_2$ | $p_\infty$ | $p_2$ |
| $2.00 \cdot 10^{-1}$ | $1.83 \cdot 10^{-1}$ | $1.74 \cdot 10^{-1}$ | - | - |
| $1.00 \cdot 10^{-1}$ | $5.57 \cdot 10^{-2}$ | $3.86 \cdot 10^{-2}$ | 1.72 | 2.17 |
| $5.00 \cdot 10^{-2}$ | $8.38 \cdot 10^{-3}$ | $5.51 \cdot 10^{-3}$ | 2.73 | 2.81 |
| $2.50 \cdot 10^{-2}$ | $1.14 \cdot 10^{-3}$ | $6.67 \cdot 10^{-4}$ | 2.88 | 3.05 |
| $1.25 \cdot 10^{-2}$ | $3.18 \cdot 10^{-4}$ | $1.07 \cdot 10^{-4}$ | 1.84 | 2.64 |

**Table 4.1.** Errors and convergence rates for the approximation of (4.66) with $\sigma^2/2 = 0.1$.

| $\Delta x$ | Errors on the FP equation | | | |
|---|---|---|---|---|
| | $E_\infty$ | $E_2$ | $p_\infty$ | $p_2$ |
| $2.00 \cdot 10^{-1}$ | $3.05 \cdot 10^{-1}$ | $3.22 \cdot 10^{-1}$ | - | - |
| $1.00 \cdot 10^{-1}$ | $1.21 \cdot 10^{-1}$ | $1.04 \cdot 10^{-1}$ | 1.33 | 1.63 |
| $5.00 \cdot 10^{-2}$ | $2.54 \cdot 10^{-2}$ | $1.79 \cdot 10^{-2}$ | 2.25 | 2.54 |
| $2.50 \cdot 10^{-2}$ | $3.07 \cdot 10^{-3}$ | $2.34 \cdot 10^{-3}$ | 3.05 | 2.94 |
| $1.25 \cdot 10^{-2}$ | $6.25 \cdot 10^{-4}$ | $3.36 \cdot 10^{-4}$ | 2.30 | 2.80 |

**Table 4.2.** Errors and convergence rates for the approximation of (4.66) with $\sigma^2/2 = 0.05$.

### 4.4.3 Non local MFG with analytical solution

Consider a non-local Mean Field Game given by

$$\begin{aligned} -\partial_t v - \tfrac{\sigma^2}{2}\Delta v + \tfrac{1}{2}|Dv|^2 &= \tfrac{1}{2}\left[x - \textstyle\int_{\mathbb{R}^d}(m(t,y)y)\mathrm{d}y\right]^2 \quad \text{in } [0,T)\times\mathbb{R}^d, \\ \partial_t m - \tfrac{\sigma^2}{2}\Delta m - \operatorname{div}(Dvm) &= 0 \quad \text{in } (0,T]\times\mathbb{R}^d, \\ v(T,\cdot) = 0, \qquad & m(0,\cdot) = m_0 \quad \text{in } \mathbb{R}^d. \end{aligned} \tag{4.68}$$

The analytical solution of (4.68) can be computed as tensorial product of one dimensional solutions. In what follows, we explicitly compute the analytical formula for the couple $(v,m)$, solution of the system (4.68) in space dimension $d = 1$. Let us denote $\overline{x}(t) = \int_{\mathbb{R}^d}(m(t,y)y)\mathrm{d}y$. The analytical solution to the HJB in (4.68) has the form $v(t,x) = \frac{1}{2}\Pi(t)x^2 + s(t)x + c(t)$ with $\Pi(t), s(t)$ and $c(t)$

time dependent functions solving the following ODEs:

$$\begin{cases} -\frac{1}{2}\dot{\Pi}(t) + \frac{1}{2}\Pi^2(t) = \frac{1}{2} & t \in (0,T), \\ -\dot{s}(t) + \Pi(t)s(t) = -\overline{x}(t) & t \in (0,T), \\ -\dot{c}(t) - \frac{\sigma^2}{2}\Pi(t) + \frac{1}{2}s^2(t) = \frac{1}{2}\left(\overline{x}(t)\right)^2 & t \in (0,T), \\ \Pi(T) = 0, \quad c(T) = 0, \quad s(T) = 0. \end{cases} \tag{4.69}$$

The first one is a Riccati equation whose analytical solution has the following expression

$$\Pi(t) = \frac{e^{2T-t} - e^t}{e^{2T-t} + e^t}.$$

Then the gradient of $v$ is $Dv(t,x) = -\Pi(t)x - s(t)$ and the optimal state solves

$$\mathrm{d}x(r) = \left(-\Pi(t)x(r) - s(r)\right)\mathrm{d}r + \sigma \mathrm{d}W(r),$$

which implies

$$x(t) = x(0) + \int_0^t \left(-\Pi(r)x(r) - s(r)\right)\mathrm{d}r + \sigma W(t).$$

Taking the expectation and $x(0) = x$, we get

$$\overline{x}(t) = \mathbb{E}\left[x(t)\right] = \int_{\mathbb{R}^d} x \mathrm{d}m_0\left(x\right) + \int_0^t \left(-\Pi(r)\overline{x}(r) - s(r)\right)\mathrm{d}r,$$

which implies that $\overline{x}(t)$ solves

$$\begin{cases} \dot{\overline{x}}(t) = -\Pi(t)\overline{x}(t) - s(t) & t \in (0,T), \\ \overline{x}(0) = \int_{\mathbb{R}^d} x \mathrm{d}m_0\left(x\right). \end{cases}$$

Therefore the couple $(\overline{x}(t), s(t))$ solves the following boundary value problem

$$\begin{cases} \dot{\overline{x}}(t) = -\Pi(t)\overline{x}(t) - s(t) & t \in (0,T), \\ -\dot{s}(t) + \Pi(t)s(t) = -\overline{x}(t) & t \in (0,T), \\ \overline{x}(0) = \int_{\mathbb{R}^d} x \mathrm{d}m_0\left(x\right), \quad s(T) = 0. \end{cases} \tag{4.70}$$

The solution to (4.70) is unique (see for instance [64]) and is given by

$$\overline{x}(t) = \int_{\mathbb{R}^d} x \mathrm{d}m_0\left(x\right), \quad s(t) = -\left(\int_{\mathbb{R}^d} x \mathrm{d}m_0\left(x\right)\right)\Pi(t).$$

The last equation in (4.69) can be now explicitly solved, and $c(t)$ gets

$$c(t) = \frac{1}{2}\left(\int_{\mathbb{R}^d} x \mathrm{d}m_0\left(x\right)\right)^2 \Pi(t) - \frac{\sigma^2}{2}\log\left(\frac{2e^T}{e^{2T-t} + e^t}\right). \tag{4.71}$$

The solution $m(t,x)$ of the FP equation in (4.68) is a gaussian function with mean $\overline{x}(t)$ and variance $\mathrm{Var}\left(x(t)\right)$. To compute $\mathrm{Var}\left(x(t)\right)$, we observe that $\mathrm{Var}\left(x(t)\right) = \mathbb{E}\left(x^2(t)\right) - \overline{x}^2(t)$ and we recall Itô's formula, for a given $f : \mathbb{R} \to \mathbb{R}$

$$f(x(t)) = f(x_0) + \int_0^t f'(x(r))\mathrm{d}r + \frac{1}{2}\int_0^t f''(x(r))\sigma^2 \mathrm{d}W(r). \tag{4.72}$$

Now, choosing $f(x(t)) = x^2(t)$, we get that $x^2(t)$ solves

$$x^2(t) = x_0^2 - \int_0^t 2x(r)\left(\Pi(r)x(r) + s(r)\right)\mathrm{d}r + \int_0^t 2x(r)\sigma \mathrm{d}W(r) + \sigma^2 t. \quad (4.73)$$

Taking the expectation and $x_0 = x$, we get

$$\mathbb{E}\left(x^2(t)\right) = \int_{\mathbb{R}^d} x^2 \mathrm{d}m_0\left(x\right) - 2\int_0^t \left(\Pi(r)\mathbb{E}\left(x^2(r)\right) + s(r)\overline{x}(r)\right)\mathrm{d}r + \sigma^2 t. \quad (4.74)$$

Calling $M(t) = \mathbb{E}\left(x^2(t)\right)$ we have that $M(t)$ solves the following ODE

$$\begin{cases} \dot{M}(t) = -2\Pi(t)M(t) - 2s(t)\overline{x}(t) + \sigma^2 & t \in (0,T), \\ M(0) = \int_{\mathbb{R}^d} x^2 \mathrm{d}m_0\left(x\right). \end{cases}$$

with exact solution given by

$$M(t) = \left(e^{2T-t} + e^t\right)^2 \left(\frac{2\int_{\mathbb{R}^d} x^2 \mathrm{d}m_0\left(x\right) - 2\overline{x}^2(t) + \sigma^2\left(e^{2T}+1\right)}{2(e^{2T}+1)^2}\right) - \left(e^{2T-t} + e^t\right)^2 \left(\frac{\sigma^2}{2(e^{2T}+e^{2t})}\right) + \overline{x}^2(t).$$

Finally, we have that $\mathrm{Var}\left(x(t)\right) = M(t) - \overline{x}^2(t)$. Let us now solve system (4.68) in a bounded domain in dimension $d = 1, 2$. We choose $[0,T] \times \mathcal{O}_\Delta = [0, 0.25] \times (-2,2)^d$, with Dirichlet boundary conditions on $\partial\mathcal{O}_\Delta$, choosen equal to the exact solution of (4.68) for the HJB and homogeneous for the FP. The numerical approximation of the boundary conditions for the HJB is based on the technique proposed in [17], and for the FP we apply the same method used in the previous test. In this and the following test, to compute (4.60), we have used a fourth-order finite difference approximation of the gradient of $v_\Delta[m]$, and we have not introduced the mollifier $\phi_\varepsilon$.
For $d = 1$ we consider two cases, one with $\sigma^2/2 = 0.005$ and one with $\sigma^2/2 = 0.05$. In all the simulations we choose $\Delta t = (\Delta x)^{4/3}/4$. Tables 4.3 and 4.4 show the errors and the convergence rates for the approximation of the HJB and the FP equations. In Table 4.4 the convergence rate tends to be close to the theoretical optimal rate 8/3.
Table 4.5 shows errors and convergence rates for problem (4.68) with $d = 2$ and $\sigma^2/2 = 0.05$, and Table 4.6 shows errors and convergence rates for the approximated gradient of the value function in (4.68) with $d = 2$ and $\sigma^2/2 = 0.05$. In both Tables the order of convergence is mostly much larger than 2.
The tollerance $\tau$ for the stopping criterion is $10^{-9}$. In Fig. 4.1 we show the solution to (4.68) on $[0,T]\times\mathcal{O}_\Delta = [0, 0.25]\times(-2,2)$ with $\sigma^2/2 = 0.005$, computed with $\Delta x = 1.25 \cdot 10^{-2}$ and $\Delta t = (\Delta x)^{4/3}/4$. Fig. 4.2 displays the zoom of initial condition, numerical and exact density of (4.68), computed with $\Delta x = 6.25\cdot 10^{-3}$ and $\Delta t = (\Delta x)^{4/3}/4$.

| $\Delta x$ | Errors on the HJB equation | | | |
|---|---|---|---|---|
| | $E_\infty$ | $E_2$ | $p_\infty$ | $p_2$ |
| $2.00 \cdot 10^{-1}$ | $6.20 \cdot 10^{-5}$ | $7.40 \cdot 10^{-5}$ | - | - |
| $1.00 \cdot 10^{-1}$ | $1.09 \cdot 10^{-5}$ | $1.43 \cdot 10^{-5}$ | 2.51 | 2.37 |
| $5.00 \cdot 10^{-2}$ | $2.13 \cdot 10^{-6}$ | $3.41 \cdot 10^{-6}$ | 2.36 | 2.07 |
| $2.50 \cdot 10^{-2}$ | $5.42 \cdot 10^{-7}$ | $1.00 \cdot 10^{-6}$ | 1.97 | 1.77 |
| $1.25 \cdot 10^{-2}$ | $1.67 \cdot 10^{-7}$ | $3.20 \cdot 10^{-7}$ | 1.70 | 1.64 |
| $6.25 \cdot 10^{-3}$ | $6.45 \cdot 10^{-8}$ | $1.11 \cdot 10^{-7}$ | 1.37 | 1.53 |
| $\Delta x$ | Errors on the FP equation | | | |
| | $E_\infty$ | $E_2$ | $p_\infty$ | $p_2$ |
| $2.00 \cdot 10^{-1}$ | $2.22 \cdot 10^{-2}$ | $2.32 \cdot 10^{-2}$ | - | - |
| $1.00 \cdot 10^{-1}$ | $5.43 \cdot 10^{-3}$ | $5.10 \cdot 10^{-3}$ | 2.03 | 2.19 |
| $5.00 \cdot 10^{-2}$ | $9.32 \cdot 10^{-4}$ | $8.90 \cdot 10^{-4}$ | 2.54 | 2.52 |
| $2.50 \cdot 10^{-2}$ | $1.33 \cdot 10^{-4}$ | $1.26 \cdot 10^{-4}$ | 2.81 | 2.82 |
| $1.25 \cdot 10^{-2}$ | $1.22 \cdot 10^{-5}$ | $1.17 \cdot 10^{-5}$ | 3.45 | 3.43 |
| $6.25 \cdot 10^{-3}$ | $6.04 \cdot 10^{-7}$ | $6.08 \cdot 10^{-7}$ | 4.33 | 4.27 |

**Table 4.3.** Errors and convergence rates for problem (4.68) with $\sigma^2/2 = 0.05$.

| $\Delta x$ | Errors on the HJB equation | | | |
|---|---|---|---|---|
| | $E_\infty$ | $E_2$ | $p_\infty$ | $p_2$ |
| $2.00 \cdot 10^{-1}$ | $1.68 \cdot 10^{-4}$ | $1.70 \cdot 10^{-4}$ | - | - |
| $1.00 \cdot 10^{-1}$ | $3.56 \cdot 10^{-5}$ | $3.48 \cdot 10^{-5}$ | 2.24 | 2.29 |
| $5.00 \cdot 10^{-2}$ | $5.86 \cdot 10^{-6}$ | $5.75 \cdot 10^{-6}$ | 2.60 | 2.60 |
| $2.50 \cdot 10^{-2}$ | $1.06 \cdot 10^{-6}$ | $1.04 \cdot 10^{-6}$ | 2.47 | 2.47 |
| $1.25 \cdot 10^{-2}$ | $1.80 \cdot 10^{-7}$ | $2.13 \cdot 10^{-7}$ | 2.56 | 2.29 |
| $6.25 \cdot 10^{-3}$ | $3.75 \cdot 10^{-8}$ | $5.24 \cdot 10^{-8}$ | 2.26 | 2.02 |
| $\Delta x$ | Errors on the FP equation | | | |
| | $E_\infty$ | $E_2$ | $p_\infty$ | $p_2$ |
| $2.00 \cdot 10^{-1}$ | $8.81 \cdot 10^{-3}$ | $1.01 \cdot 10^{-2}$ | - | - |
| $1.00 \cdot 10^{-1}$ | $3.06 \cdot 10^{-3}$ | $2.53 \cdot 10^{-3}$ | 1.53 | 2.00 |
| $5.00 \cdot 10^{-2}$ | $8.01 \cdot 10^{-4}$ | $5.56 \cdot 10^{-4}$ | 1.93 | 2.19 |
| $2.50 \cdot 10^{-2}$ | $1.81 \cdot 10^{-4}$ | $1.14 \cdot 10^{-4}$ | 2.15 | 2.29 |
| $1.25 \cdot 10^{-2}$ | $3.62 \cdot 10^{-5}$ | $2.05 \cdot 10^{-5}$ | 2.32 | 2.48 |
| $6.25 \cdot 10^{-3}$ | $5.75 \cdot 10^{-6}$ | $3.27 \cdot 10^{-6}$ | 2.65 | 2.65 |

**Table 4.4.** Errors and convergence rates for problem (4.68) with $\sigma^2/2 = 0.005$.

### 4.4.4 Local MFG with reference solution

We consider a smooth problem, as in [94, Section 5.2]. We choose in (**MFG**) the following data:

$$m_0(x) = \begin{cases} 4\sin^2(2\pi(x) - \frac{1}{4}) & x \in [\frac{1}{4}, \frac{3}{4}] \\ 0 & \text{otherwise,} \end{cases}$$

$$v(T, x) = 0, \quad F(x, m(t, x)) = 3m_0(x) - \min(4, m(t, x)).$$

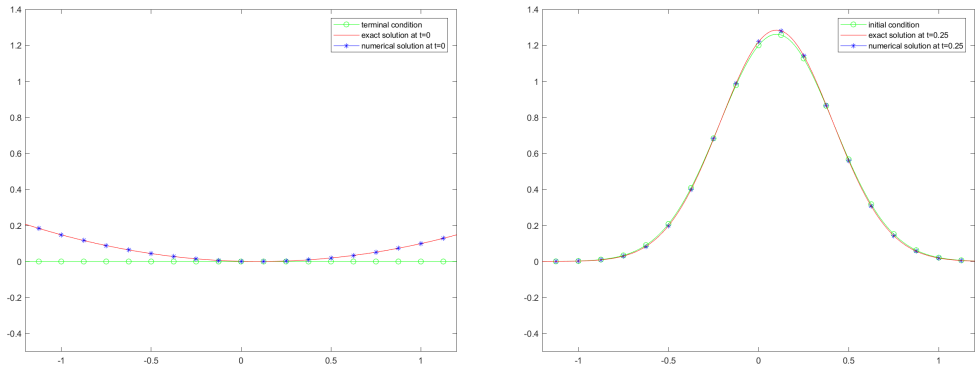



**Figure 4.1.** Solution to (4.68) on $[0,T] \times \mathcal{O}_\Delta = [0, 0.25] \times (-2, 2)$ with $\sigma^2/2 = 0.005$. Terminal condition, numerical and exact value function (left). Initial condition, numerical and exact density (right).

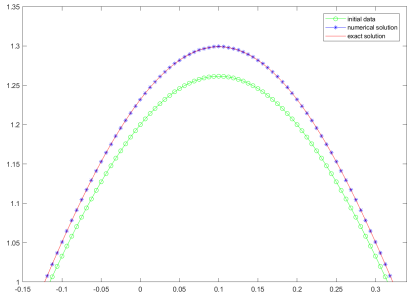



**Figure 4.2.** Zoom of initial condition, numerical and exact density of (4.68) at time $T = 0.25$.

| $\Delta x$ | Errors on the HJB equation | | | |
|---|---|---|---|---|
| | $E_\infty$ | $E_2$ | $p_\infty$ | $p_2$ |
| $2.00 \cdot 10^{-1}$ | $3.93 \cdot 10^{-2}$ | $4.85 \cdot 10^{-2}$ | - | - |
| $1.00 \cdot 10^{-1}$ | $8.62 \cdot 10^{-4}$ | $1.01 \cdot 10^{-3}$ | 5.51 | 5.59 |
| $5.00 \cdot 10^{-2}$ | $2.30 \cdot 10^{-5}$ | $3.37 \cdot 10^{-5}$ | 5.23 | 4.91 |
| $2.50 \cdot 10^{-2}$ | $3.76 \cdot 10^{-6}$ | $7.67 \cdot 10^{-6}$ | 2.61 | 2.14 |
| $\Delta x$ | Errors on the FP equation | | | |
| | $E_\infty$ | $E_2$ | $p_\infty$ | $p_2$ |
| $2.00 \cdot 10^{-1}$ | $1.20 \cdot 10^{-1}$ | $1.40 \cdot 10^{-1}$ | - | - |
| $1.00 \cdot 10^{-1}$ | $4.27 \cdot 10^{-2}$ | $3.53 \cdot 10^{-2}$ | 1.49 | 1.99 |
| $5.00 \cdot 10^{-2}$ | $8.80 \cdot 10^{-3}$ | $3.75 \cdot 10^{-3}$ | 2.28 | 2.62 |
| $2.50 \cdot 10^{-2}$ | $3.77 \cdot 10^{-4}$ | $2.53 \cdot 10^{-4}$ | 4.54 | 3.89 |

**Table 4.5.** Errors and convergence rates for problem (4.68) with $d = 2$ and $\sigma^2/2 = 0.05$.

The domain is $[0,T] \times \mathcal{O}_\Delta = [0, 0.05] \times (0, 1)$, the volatility $\sigma^2/2 = 0.05$. We suppose homogeneous Neumann boundary condition for both HJB and FP, implemented as in [29]. We compute a reference solution, using $\Delta x = 6.67 \cdot 10^{-4}$ and $\Delta t = (\Delta x)^{3/2}/3$. In Tables 4.7 and 4.8, we show errors and convergence rates, with respect to the discrete infinite and 2 norms, for the value function $v$,

| $\Delta x$ | Errors on the first component | | | |
|---|---|---|---|---|
| | $E_\infty$ | $E_2$ | $p_\infty$ | $p_2$ |
| $2.00 \cdot 10^{-1}$ | $4.64 \cdot 10^{-1}$ | $1.41 \cdot 10^{-1}$ | - | - |
| $1.00 \cdot 10^{-1}$ | $1.65 \cdot 10^{-2}$ | $3.52 \cdot 10^{-3}$ | 4.81 | 5.32 |
| $5.00 \cdot 10^{-2}$ | $4.45 \cdot 10^{-4}$ | $1.04 \cdot 10^{-4}$ | 5.21 | 5.08 |
| $2.50 \cdot 10^{-2}$ | $9.12 \cdot 10^{-5}$ | $2.32 \cdot 10^{-5}$ | 2.29 | 2.16 |
| $\Delta x$ | Errors on the second component | | | |
| | $E_\infty$ | $E_2$ | $p_\infty$ | $p_2$ |
| $2.00 \cdot 10^{-1}$ | $4.64 \cdot 10^{-1}$ | $1.41 \cdot 10^{-1}$ | - | - |
| $1.00 \cdot 10^{-1}$ | $1.65 \cdot 10^{-2}$ | $3.52 \cdot 10^{-3}$ | 4.81 | 5.32 |
| $5.00 \cdot 10^{-2}$ | $4.45 \cdot 10^{-4}$ | $1.04 \cdot 10^{-4}$ | 5.21 | 5.08 |
| $2.50 \cdot 10^{-2}$ | $9.12 \cdot 10^{-5}$ | $2.32 \cdot 10^{-5}$ | 2.29 | 2.16 |

**Table 4.6.** Errors and convergence rates for the gradient of the value function in system (4.68) with $d = 2$ and $\sigma^2/2 = 0.05$.

its gradient $\nabla v$ and the density $m$, computed with $\Delta t = (\Delta x)^{3/2}/3$. We observe an order near two in most of the cases in both norms. The mass error is of the order of $10^{-13}$ in all tests, this fact confirms that the scheme is mass preserving. Fig. 4.3 shows the numerical density at time $T$, the numerical value function and its gradient at time 0, computed with $\Delta x = 3.13 \cdot 10^{-3}$.

| $\Delta x$ | Errors on the HJB equation | | | |
|---|---|---|---|---|
| | $E_\infty$ | $E_2$ | $p_\infty$ | $p_2$ |
| $5.00 \cdot 10^{-2}$ | $5.38 \cdot 10^{-2}$ | $3.80 \cdot 10^{-2}$ | - | - |
| $2.50 \cdot 10^{-2}$ | $1.43 \cdot 10^{-2}$ | $1.29 \cdot 10^{-2}$ | 1.91 | 1.55 |
| $1.25 \cdot 10^{-2}$ | $4.25 \cdot 10^{-3}$ | $3.24 \cdot 10^{-3}$ | 1.74 | 1.99 |
| $6.25 \cdot 10^{-3}$ | $8.84 \cdot 10^{-4}$ | $7.99 \cdot 10^{-4}$ | 2.27 | 2.01 |
| $3.13 \cdot 10^{-3}$ | $3.76 \cdot 10^{-4}$ | $3.72 \cdot 10^{-4}$ | 1.23 | 1.10 |
| $1.56 \cdot 10^{-3}$ | $4.99 \cdot 10^{-5}$ | $3.60 \cdot 10^{-5}$ | 2.90 | 3.37 |
| $\Delta x$ | Errors on the gradient of $v$ | | | |
| | $E_\infty$ | $E_2$ | $p_\infty$ | $p_2$ |
| $5.00 \cdot 10^{-2}$ | $8.09 \cdot 10^{-2}$ | $4.96 \cdot 10^{-2}$ | - | - |
| $2.50 \cdot 10^{-2}$ | $1.37 \cdot 10^{-2}$ | $1.19 \cdot 10^{-2}$ | 2.53 | 2.05 |
| $1.25 \cdot 10^{-2}$ | $3.94 \cdot 10^{-3}$ | $2.79 \cdot 10^{-3}$ | 1.80 | 2.09 |
| $6.25 \cdot 10^{-3}$ | $8.34 \cdot 10^{-4}$ | $7.07 \cdot 10^{-4}$ | 2.37 | 1.98 |
| $3.13 \cdot 10^{-3}$ | $3.72 \cdot 10^{-4}$ | $3.25 \cdot 10^{-4}$ | 1.16 | 1.12 |
| $1.56 \cdot 10^{-3}$ | $5.25 \cdot 10^{-5}$ | $3.16 \cdot 10^{-5}$ | 2.82 | 3.36 |

**Table 4.7.** Errors and convergence rates for problem in Subsection 4.4.4.

| $\Delta x$ | Errors on the FP equation | | | |
|---|---|---|---|---|
| | $E_\infty$ | $E_2$ | $p_\infty$ | $p_2$ |
| $5.00 \cdot 10^{-2}$ | $9.07 \cdot 10^{-2}$ | $4.82 \cdot 10^{-2}$ | - | - |
| $2.50 \cdot 10^{-2}$ | $1.81 \cdot 10^{-2}$ | $6.79 \cdot 10^{-3}$ | 2.32 | 2.82 |
| $1.25 \cdot 10^{-2}$ | $4.81 \cdot 10^{-3}$ | $1.36 \cdot 10^{-3}$ | 1.91 | 2.32 |
| $6.25 \cdot 10^{-3}$ | $7.64 \cdot 10^{-4}$ | $2.06 \cdot 10^{-4}$ | 2.65 | 2.72 |
| $3.13 \cdot 10^{-3}$ | $1.82 \cdot 10^{-4}$ | $6.96 \cdot 10^{-5}$ | 2.07 | 1.55 |
| $1.56 \cdot 10^{-3}$ | $6.28 \cdot 10^{-5}$ | $1.24 \cdot 10^{-5}$ | 1.53 | 2.49 |

**Table 4.8.** Errors and convergence rates for problem in Subsection 4.4.4.

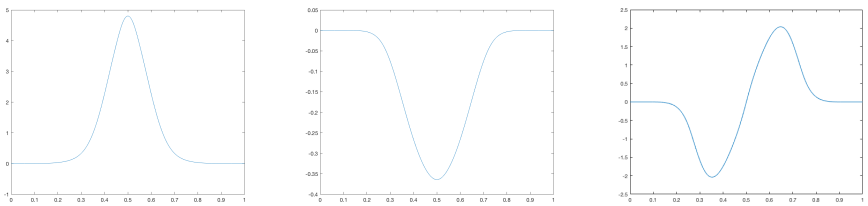


**Figure 4.3.** Density $m(T, x)$ (left), value function $v(0, x)$ (center) and its gradient $Dv(x, 0)$ (right) in the problem of Subsection 4.4.4.

# Bibliography

[1] R. Abgrall. Numerical discretization of boundary conditions for first order Hamilton-Jacobi equations. *SIAM J. Numer. Anal.*, 41(6):2233–2261, 2003.

[2] Y. Achdou, P. Cardaliaguet, F. Delarue, A. Porretta, F. Santambrogio, and Springer Nature. *Mean Field Games: Cetraro, Italy 2019.* C.I.M.E. Foundation Subseries. Springer International Publishing, 2020.

[3] Y. Achdou and M. Falcone. A semi-lagrangian scheme for mean curvature motion with nonlinear neumann conditions. *Interfaces Free Bound.*, 14(4):455–485, 2012.

[4] L. Ambrosio, N. Gigli, and G. Savaré. *Gradient flows in metric spaces and in the space of probability measures.* Second edition. Lecture notes in Mathematics ETH Zürich. Birkhäuser Verlag, Bassel, 2008.

[5] M. Bardi and I. Capuzzo-Dolcetta. *Optimal control and viscosity solutions of Hamilton-Jacobi-Bellman equations.* Birkauser, 1996.

[6] G. Barles. Fully nonlinear Neumann type boundary conditions for second-order elliptic and parabolic equations. *J. Differential Equations*, 106(1):90–106, 1993.

[7] G. Barles. Nonlinear Neumann boundary conditions for quasilinear degenerate elliptic equations and applications. *J. Differential Equations*, 154(1):191–224, 1999.

[8] G. Barles and P.-L. Lions. Fully nonlinear Neumann type boundary conditions for first-order Hamilton-Jacobi equations. *Nonlinear Anal.*, 16(2):143–153, 1991.

[9] G. Barles and P.E. Souganidis. Convergence of approximation schemes for fully nonlinear second order equations. *Asymptotic Anal.*, 4(3):271–283, 1991.

[10] J.W. Barrett and C.M. Elliott. Finite element approximation of the Dirichlet problem using the boundary penalty method. *Numer. Math.*, 49(4):343–366, 1986.

[11] E.N. Barron and R. Jensen. The pontryagin maximum principle from dynamic programming and viscosity solutions to first-order partial differential equations. *Transactions of the American Mathematical Society*, 298(2):635–641, 1986.

[12] R. Bermejo and L. Saavedra. Modified Lagrange-Galerkin methods of first and second order in time for convection-diffusion problems. *Numer. Math.*, 120(4):601–638, 2012.

[13] V.I. Bogachev, N.V. Krylov, M. Röckner, and S.V. Shaposhnikov. *Fokker-Planck-Kolmogorov equations*, volume 207 of *Mathematical Surveys and Monographs*. American Mathematical Society, Providence, RI, 2015.

[14] O. Bokanowski and G. Simarmata. Semi-Lagrangian discontinuous Galerkin schemes for some first-and second-order partial differential equations. *ESAIM: Mathematical Modelling and Numerical Analysis*, 50:1699–1730, 2016.

[15] L. Bonaventura. A semi-implicit, semi-Lagrangian scheme using the height coordinate for a nonhydrostatic and fully elastic model of atmospheric flows. *Journal of Computational Physics*, 158:186–213, 2000.

[16] L. Bonaventura, E. Calzola, E. Carlini, and R. Ferretti. A fully semi-Lagrangian method for the Navier-Stokes equations in primitive variables. In *Proceedings of the 2017 Finite Elements in Flows Conference, Rome*. Springer Verlag, 2020.

[17] L. Bonaventura, E. Calzola, E. Carlini, and R. Ferretti. Second order fully semi-Lagrangian discretizations of advection-diffusion-reaction systems. *J. Sci. Comput.*, 88(1):Paper No. 23, 29, 2021.

[18] L. Bonaventura and R. Ferretti. Semi-Lagrangian methods for parabolic problems in divergence form. *SIAM Journal of Scientific Computing*, 36:A2458 – A2477, 2014.

[19] L. Bonaventura, R. Ferretti, and L. Rocchi. A fully semi-Lagrangian discretization for the 2D Navier-Stokes equations in the vorticity–streamfunction formulation. *Applied Mathematics and Computation*, 323:132–144, 2018.

[20] L. Bonaventura, R. Redler, and R. Budich. *Earth System Modelling 2: Algorithms, Code Infrastructure and Optimisation.* Springer Verlag, New York, 2012.

[21] L. Bonaventura and A. Della Rocca. Unconditionally strong stability preserving extensions of the TR-BDF2 method. *Journal of Scientific Computing*, 70:859–895, 2017.

[22] B. Bouchard. Optimal reflection of diffusions and barrier options pricing under constraints. *SIAM J. Control Optim.*, 47(4):1785–1813, 2008.

[23] M. Bourgoing. Viscosity solutions of fully nonlinear second order parabolic equations with $L^1$ dependence in time and Neumann boundary conditions. *Discrete Contin. Dyn. Syst.*, 21(3):763–800, 2008.

[24] C.-E. Bréhier and E. Faou. Analysis of the Monte-Carlo error in a hybrid semi-Lagrangian scheme. *Applied Mathematics Research eXpress*, 2015(2):167–203, 2015.

[25] S. Cacace, L. Della Cioppa, and R. Ferretti. Efficient implementation of characteristic-based schemes on unstructured triangular grids. *in preparation.*

[26] M. Camilli and M. Falcone. An approximation scheme for the optimal control of diffusion processes. *Mathematical Modelling and Numerical Analysis*, 29:97–122, 1995.

[27] C. Canuto, M.Y. Hussaini, A. Quarteroni, and T.A. Zang. *Spectral methods: Fundamentals in single domains.* Springer Science & Business Media, 2006.

[28] P. Cardaliaguet. Notes on Mean Field Games: from P.-L. Lions' lectures at Collège de France. *Lecture Notes given at Tor Vergata*, 2010.

[29] E. Carlini, A. Calzola, Dupuis X., and F.J. Silva. A semi-lagrangian scheme for Hamilton-Jacobi-Bellman equations with oblique boundary conditions. *Preprint*, pages 1–21, 2021.

[30] E. Carlini, M. Falcone, and R. Ferretti. Convergence of a large time-step scheme for mean curvature motion. *Interfaces Free Bound.*, 12(4):409–441, 2010.

[31] E. Carlini and F. J. Silva. On the discretization of some nonlinear Fokker-Planck-Kolmogorov equations and applications. *SIAM J. Numer. Anal.*, 56(4):2148–2177, 2018.

[32] E. Carlini and F.J. Silva. A fully discrete semi-Lagrangian scheme for a first order mean field game problem. *SIAM J. Numer. Anal.*, 52(1):45–67, 2014.

[33] E. Carlini and F.J. Silva. A semi-Lagrangian scheme for a degenerate second order mean field game system. *Discrete and Continuous Dynamical Systems*, 35(9):4269–4292, 2015.

[34] E. Carlini and F.J. Silva. A semi-Lagrangian scheme for the Fokker-Planck equation. *IFAC-PapersOnLine*, 49(8):272–277, 2016. 2nd IFAC Workshop on Control of Systems Governed by Partial Differential Equations CPDE 2016.

[35] E. Carlini and F.J. Silva. On the discretization of some nonlinear Fokker-Planck-Kolmogorov equations and applications. working paper or preprint, August 2017.

[36] V. Casulli. Semi-implicit finite difference methods for the two dimensional shallow water equations. *Journal of Computational Physics*, 86:56–74, 1990.

[37] V. Casulli and E. Cattani. Stability, accuracy and efficiency of a semi-implicit method for three-dimensional shallow water flow. *Computers and Mathematics with Applications*, 27(4):99–112, 1994.

[38] J.S. Chang and G. Cooper. A practical difference scheme for Fokker-Planck equations. *Journal of Computational Physics*, 6:1–16, 1970.

[39] P.G. Ciarlet and J.-L. Lions, editors. *Handbook of numerical analysis. Vol. II.* Handbook of Numerical Analysis, II. North-Holland, Amsterdam, 1991. Finite element methods. Part 1.

[40] C. Costantini, B. Pacchiarotti, and F. Sartoretto. Numerical approximation for functionals of reflecting diffusion processes. *SIAM Journal on Applied Mathematics*, 58(1):73–102, 1998.

[41] J. Coté and A. Staniforth. A two time level semi-Lagrangian semi-implicit scheme for spectral models. *Monthly Weather Review*, 116:2003–2012, 1988.

[42] M.G. Crandall, H. Ishii, and P.-L. Lions. Uniqueness of viscosity solutions of hamilton-jacobi equations revisited. *Journal of the Mathematical Society of Japan*, 39(4):581–596, 1987.

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

[44] M.G. Crandall and P.-L. Lions. Viscosity solutions of hamilton-jacobi equations. *Transactions of the American Mathematical Society*, 277(1):1–42, 1983.

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

[46] K. Deckelnick and M. Hinze. Convergence of a finite element approximation to a state-constrained elliptic control problem. *SIAM J. Numer. Anal.*, 45(5):1937–1953, 2007.

[47] C. Erath, P.H. Lauritzen, J.H. Garcia, and H.M. Tufo. Integrating a scalable and efficient semi-Lagrangian multi-tracer transport scheme in HOMME. *Procedia Computer Science*, 9:994–1003, 2012.

[48] C. Erath, M.A. Taylor, and R.D. Nair. Two conservative multi-tracer efficient semi-Lagrangian schemes for multiple processor systems integrated in a spectral element (climate) dynamical core. *Communications in Applied and Industrial Mathematics*, 7:74–98, 2016.

[49] M. Falcone and R. Ferretti. *Semi-Lagrangian Approximation Schemes for Linear and Hamilton-Jacobi Equations*. MOS-SIAM Series on Optimization, 2013.

[50] X. Feng, R. Glowinski, and M. Neilan. Recent developments in numerical methods for fully nonlinear second order partial differential equations. *SIAM Rev.*, 55(2):205–267, 2013.

[51] X. Feng, H. Song, T. Tang, and J. Yang. Nonlinear stability of the implicit-explicit methods for the Allen-Cahn equation. *Inverse Problems and Imaging*, 7:679–695, 2013.

[52] R. Ferretti. A technique for high-order treatment of diffusion terms in semi-Lagrangian schemes. *Commun. Comput. Phys.*, 8(2):445–470, 2010.

[53] R. Ferretti. On the relationship between semi-Lagrangian and Lagrange-Galerkin schemes. *Numer. Math.*, 124(1):31–56, 2013.

[54] R. Ferretti and M. Mehrenberger. Stability of semi-Lagrangian schemes of arbitrary odd degree under constant and variable advection speed. *Math. Comp.*, 89(324):1783–1805, 2020.

[55] R. Ferretti and G. Perrone. On the stability of semi-Lagrangian advection schemes under finite element interpolations. In *Applied And Industrial Mathematics In Italy II*, pages 339–350. World Scientific, 2007.

[56] A. Figalli. Existence and uniqueness of martingale solutions for sdes with rough or degenerate coefficients. *J. Funct. Anal.*, 253:109–153, 2008.

[57] W.H. Fleming and R.W. Rishel. *Deterministic and Stochastic Optimal Control*. Applications of mathematics. Springer-Verlag, 1975.

[58] W.H. Fleming and H.M. Soner. *Controlled Markov processes and viscosity solutions*, volume 25 of *Stochastic Modelling and Applied Probability*. Springer, New York, second edition, 2006.

[59] A. Friedman. *Partial Differential Equations of Parabolic Type*. R.E. Krieger Publishing Company, 1983.

[60] F. Garcia, L. Bonaventura, M. Net, and J. Sánchez. Exponential versus IMEX high-order time integrators for thermal convection in rotating spherical shells. *Journal of Computational Physics*, 264:41–54, 2014.

[61] Y. Giga, S. Goto, H. Ishii, and M.-H. Sato. Comparison principle and convexity preserving properties for singular degenerate parabolic equations on unbounded domains. *Indiana University Mathematics Journal*, 40(2):443–470, 1991.

[62] D. Gilbarg and N.S. Trudinger. *Elliptic partial differential equations of second order*. Classics in Mathematics. Springer-Verlag, Berlin, 2001. Reprint of the 1998 edition.

[63] E. Gobet. Efficient schemes for the weak approximation of reflected diffusions. volume 7, pages 193–202. 2001. Monte Carlo and probabilistic methods for partial differential equations (Monte Carlo, 2000).

[64] M.T. Heath. *Scientific Computing: An Introductory Survey*. McGraw-Hill Education, 2005.

[65] K. Hinderer, U. Rieder, and M. Stieglitz. *Dynamic optimization*. Universitext. Springer, Cham, 2016. Deterministic and stochastic models.

[66] M. Hochbruck and A. Ostermann. Exponential integrators. *Acta Numerica*, pages 209–286, 2010.

[67] M. Huang, R.P. Malhamé, and P.E. Caines. Large population stochastic dynamic games: closed-loop McKean-Vlasov systems and the Nash certainty equivalence principle. *Commun. Inf. Syst.*, 6(3):221–251, 2006.

[68] H. Ishii and M.-H. Sato. Nonlinear oblique derivative problems for singular degenerate parabolic equations on a general domain. *Nonlinear Anal.*, 57(7-8):1077–1098, 2004.

[69] J. L. Kelley. *General topology*. D. Van Nostrand Co., Inc., Toronto-New York-London, 1955.

[70] P.E. Kloeden and E. Platen. *Numerical solution of stochastic differential equations*, volume 23 of *Applications of Mathematics (New York)*. Springer-Verlag, Berlin, 1992.

[71] M. Krukowski. Arzelà-Ascoli's theorem in uniform spaces. *Discrete Contin. Dyn. Syst. Ser. B*, 23(1):283–294, 2018.

[72] H. Kushner and P.G. Dupuis. *Numerical methods for stochastic control problems in continuous time*, volume 24. Springer Science & Business Media, 2013.

[73] O. A. Ladyvzenskaja, V. A. Solonnikov, and N. N. Uralceva. *Linear and quasilinear equations of parabolic type*. Translations of Mathematical Monographs, Vol. 23. American Mathematical Society, Providence, R.I., 1968. Translated from the Russian by S. Smith.

[74] J.D. Lambert. *Numerical methods for ordinary differential systems*. Wiley, 1991.

[75] J.-M. Lasry and P.-L. Lions. Jeux à champ moyen I. Le cas stationnaire. *C. R. Math. Acad. Sci. Paris*, 343:619–625, 2006.

[76] J.-M. Lasry and P.-L. Lions. Jeux à champ moyen II. Horizon fini et contrôle optimal. *C. R. Math. Acad. Sci. Paris*, 343:679–684, 2006.

[77] J.-M. Lasry and P.-L. Lions. Mean field games. *Jpn. J. Math.*, 2:229–260, 2007.

[78] R. J. Leveque. *Finite Volume Methods for Hyperbolic Problems*. Cambridge University Press, Cambridge, UK, 2002.

[79] P.-L. Lions. *Generalized solutions of Hamilton-Jacobi equations*, volume 69 of *Research Notes in Mathematics*. Pitman (Advanced Publishing Program), Boston, Mass.-London, 1982.

[80] P.-L. 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.

[81] P.-L. Lions. Neumann type boundary conditions for Hamilton-Jacobi equations. *Duke Math. J.*, 52(4):793–820, 1985.

[82] R. Meise and D. Vogt. *Introduction to functional analysis*, volume 2 of *Oxford Graduate Texts in Mathematics*. The Clarendon Press, Oxford University Press, New York, 1997. Translated from the German by M. S. Ramanujan.

[83] G. N. Milstein. Application of the numerical integration of stochastic equations for the solution of boundary value problems with Neumann boundary conditions. *Teor. Veroyatnost. i Primenen.*, 41(1):210–218, 1996.

[84] G.N. Milstein. The probability approach to numerical solution of nonlinear parabolic equations. *Numerical Methods for Partial Differential Equations*, 18:490–522, 2002.

[85] G.N. Milstein and M.V. Tretyakov. Numerical algorithms for semilinear parabolic equations with small parameter based on approximation of stochastic equations. *Mathematics of Computation*, 69:237–567, 2000.

[86] G.N. Milstein and M.V. Tretyakov. Numerical solution of the Dirichlet problem for nonlinear parabolic equations by a probabilistic approach. *IMA Journal of Numerical Analysis*, 21:887–917, 2001.

[87] G.N. Milstein and M.V. Tretyakov. *Stochastic numerics for mathematical physics*. Springer Science & Business Media, 2013.

[88] K.W. Morton, A. Priestley, and E. Süli. Stability of the Lagrange-Galerkin method with nonexact integration. *RAIRO Modél. Math. Anal. Numér.*, 22(4):625–653, 1988.

[89] R. Narasimhan. *Analysis on Real and Complex Manifolds*. Advanced studies in pure mathematics. Masson, 1973.

[90] M. Neilan, A. J. Salgado, and W. Zhang. Numerical analysis of strongly nonlinear PDEs. *Acta Numer.*, 26:137–303, 2017.

[91] B. Perthame. *Transport equations in biology*. Frontiers in Mathematics. Birkhäuser Verlag, Basel, 2007.

[92] A. Picarelli and C. Reisinger. Probabilistic error analysis for some approximation schemes to optimal control problems. *Systems Control Lett.*, 137:104619, 11, 2020.

[93] A. Picarelli, C. Reisinger, and J. Arto. Some regularity and convergence results for parabolic hamilton-jacobi-bellman equations in bounded domains. *Journal of Differential Equations*, 10 2017.

[94] B. Popov and V. Tomov. Central schemes for mean field games. *Commun. Math. Sci.*, 13(8):2177–2194, 2015.

[95] J. Pudykiewicz and A. Staniforth. Some properties and comparative performance of the semi-Lagrangian method of Robert in the solution of the advection diffusion equation. *Atmosphere-Ocean*, 22:283–304, 1984.

[96] A. Quarteroni and A. Valli. *Numerical approximation of partial differential equations*. Springer Verlag, 1994.

[97] H. Risken. *The Fokker-Planck equation*, volume 18 of *Springer Series in Synergetics*. Springer-Verlag, Berlin, 1984. Methods of solution and applications.

[98] A. Robert. A semi-Lagrangian and semi-implicit numerical integration scheme for the primitive meteorological equations. *Journal of the Meteorological Society of Japan*, 60:319–325, 1982.

[99] G. Rosatti, D. Cesari, and L. Bonaventura. Semi-implicit, semi-Lagrangian environmental modelling on Cartesian grids with cut cells. *Journal of Computational Physics*, 204:353–377, 2005.

[100] E. Rouy. Numerical approximation of viscosity solutions of first-order Hamilton-Jacobi equations with Neumann type boundary conditions. *Math. Models Methods Appl. Sci.*, 2(3):357–374, 1992.

[101] P.E. Souganidis. Existence of viscosity solutions of hamilton-jacobi equations. *Journal of Differential Equations*, 56(3):345–390, 1985.

[102] A. Staniforth and J. Coté. Semi-Lagrangian integration schemes for atmospheric models-a review. *Monthly Weather Review*, 119:2206–2223, 1991.

[103] C. Temperton, M. Hortal, and A. Simmons. A two-time-level semi-Lagrangian global spectral model. *Quarterly Journal of the Royal Meteorological Society*, 127:111–127, 2001.

[104] C. Temperton and A. Staniforth. An efficient two-time-level semi-Lagrangian semi-implicit integration scheme. *Quarterly Journal of the Royal Meteorological Society*, 113:1025–1039, 1987.

[105] G. Tumolo and L. Bonaventura. A semi-implicit, semi-Lagrangian, DG framework for adaptive numerical weather prediction. *Quarterly Journal of the Royal Meteorological Society*, 141:2582–2601, 2015.

[106] G. Tumolo, L. Bonaventura, and M. Restelli. A semi-implicit, semi-Lagrangian, $p-$adaptive discontinuous Galerkin method for the shallow water equations. *Journal of Computational Physics*, 232:46–67, 2013.

[107] J. Voigt. *A course on topological vector spaces*. Compact Textbooks in Mathematics. Birkhäuser/Springer, Cham, [2020] ©2020.

[108] N.P. Wedi, P. Bauer, M. Diamantakis, M. Hamrud, S. Malardel, K. Mogensen, G. Mozdzynski, and P.K. Smolarkiewicz. The modelling infrastructure of the Integrated Forecasting System: Recent advances and future challenges. Technical Report 760, ECMWF, 2015.

[109] J. Yong and X.Y. Zhou. *Stochastic Controls: Hamiltonian Systems and HJB Equations*. Stochastic Modelling and Applied Probability. Springer New York, 1999.

[110] Y. Zhan. *Viscosity Solutions of Nonlinear Degenerate Parabolic Equations and Several Applications [microform]*. Canadian theses. Thesis (Ph.D.)– University of Toronto, 1999.

[111] M.P. Zorzano, H. Mais, and L. Vazquez. Numerical solution of two-dimensional Fokker-Planck equations. *Appl. Math. Comput.*, 98(2-3):109–117, 1999.