Weierstraß-Institut
für Angewandte Analysis und Stochastik

Leibniz-Institut im Forschungsverbund Berlin e. V.

Preprint ISSN 2198-5855

Continuous time stochastic optimal control under discrete time partial observations

Christian Bayer[1], Boualem Djehiche[2], Eliza Rezvanova[3], Raúl Tempone[3,4]

submitted: January 16, 2025

[1] Weierstrass Institute
Mohrenstr. 39
10117 Berlin
Germany
E-Mail: christian.bayer@wias-berlin.de

[2] KTH
School of Engineering Sciences (SCI)
Mathematics (Dept.), Mathematical Statistics
100 44 Stockholm
Sweden
E-Mail: boualem@math.kth.se

[3] KAUST
Computer, Electrical and Mathematical
Sciences & Engineering Division (CEMSE)
Thuwal 23955 – 6900
Saudi Arabia
E-Mail: eliza.rezvanova@kaust.edu.sa
raul.tempone@kaust.edu.sa

[4] RWTH Aachen University
Chair of Mathematics for Uncertainty Quantification
Pontdriesch 14-16
52062 Aachen
Germany
E-Mail: tempone@uq.rwth-aachen.de

No. 3168
Berlin 2025

2020 *Mathematics Subject Classification.* 60H10, 60H35, 93E20, 49L20, 93E11, 60G35.

Key words and phrases. Markov decision process, stochastic optimal control, filtering, partially observed state, Bayesian updates, measure-valued process.

Ch. Bayer acknowledges seed support for DFG CRC/TRR 388 “Rough Analysis, Stochastic Dynamics and Related Topics. This work was supported by the KAUST Office of Sponsored Research (OSR) under Award No. URF/1/2584-01-01 and the Alexander von Humboldt Foundation. E. Rezvanova, and R. Tempone are members of the KAUST SRI Center for Uncertainty Quantification in Computational Science and Engineering.

Edited by
Weierstraß-Institut für Angewandte Analysis und Stochastik (WIAS)
Leibniz-Institut im Forschungsverbund Berlin e. V.
Mohrenstraße 39
10117 Berlin
Germany

Fax: +49 30 20372-303
E-Mail: preprint@wias-berlin.de
World Wide Web: http://www.wias-berlin.de/

Continuous time stochastic optimal control under discrete time partial observations

Christian Bayer, Boualem Djehiche, Eliza Rezvanova, Raúl Tempone

Abstract

This work addresses stochastic optimal control problems where the unknown state evolves in continuous time while partial, noisy, and possibly controllable measurements are only available in discrete time. We develop a framework for controlling such systems, focusing on the measure-valued process of the system's state and the control actions that depend on noisy and incomplete data. Our approach uses a stochastic optimal control framework with a probability measure-valued state, which accommodates noisy measurements and integrates them into control decisions through a Bayesian update mechanism. We characterize the control optimality in terms of a sequence of interlaced Hamilton Jacobi Bellman (HJB) equations coupled with controlled impulse steps at the measurement times. For the case of Gaussian-controlled processes, we derive an equivalent HJB equation whose state variable is finite-dimensional, namely the state's mean and covariance. We demonstrate the effectiveness of our methods through numerical examples. These include control under perfect observations, control under no observations, and control under noisy observations. Our numerical results highlight significant differences in the control strategies and their performance, emphasizing the challenges and computational demands of dealing with uncertainty in state observation.

1 Introduction

This work studies *Stochastic Optimal Control* (SOC) problems where the state evolves in continuous time but observations are available only in discrete time. The focus is on probability measure-valued processes and control actions dependent on noisy and incomplete data. SOC is a vital area of study in both theoretical and applied mathematics, finding applications in finance, engineering, and various sciences. This field deals with the challenge of making optimal decisions in systems whose dynamics are driven by random processes.

In optimal control, we aim to determine a control strategy for a dynamical system that minimizes a given cumulative running cost function over time, plus a terminal cost, within a finite time horizon. As general references, we refer to books [FS93] (providing a comprehensive introduction to the theory of SOC, focusing on Hamiltonian systems and Hamilton-Jacobi-Bellman (HJB) equations) and [BCD97] (exploring the connection between Optimal Control and the viscosity solutions of HJB equations).

In many applications which motivate our work, the system's state evolves continuously over time, but the current state of the controlled system is not perfectly known. This discrepancy between the continuous evolution of the state and imperfect observations presents a significant challenge for control strategies. The books [Ben92; Mao06; Kri16] address control problems where the system state is only partially observable, using a filtering framework and applying it to problems of control under uncertainty.

In this work, we concentrate on the important special case that

- the state process evolves in continuous time on a finite time interval, but is observed in discrete time;
- the control on the state process acts in continuous time, incurring running as well as terminal costs to be minimized;
- the (noisy, partial) observations incur costs, as well, which are taken into account in the minimization problem.

DOI 10.20347/WIAS.PREPRINT.3168

As discussed below, the literature on SOC under partial observation mainly concentrates on cases where both control and observations happen in continuous or discrete time. Moreover, the observation process is usually not considered to incur costs. The mixed case studied in this work is highly relevant in many applications where the "observation" corresponds to an actual, physical measurement, especially one requiring manual intervention.

As a practical application of such stochastic control problems, consider the treatment of a disease guided by regular diagnostic tests. The disease evolves continuously, and the medical doctor must decide when and how to treat the patient and when and which kinds of medical tests to prescribe. Each step incurs costs that should be accounted for when optimizing the patient's health. Other applications include automated trading systems in finance, where the market state evolves continuously, but traders only have discrete and noisy observations of market indicators. The goal is to optimize trading strategies to maximize profit or minimize risk. In autonomous vehicle navigation within robotics, the vehicle's position and environment evolve continuously, but sensors provide discrete, incomplete, and noisy measurements. The control strategy aims to navigate the vehicle safely and efficiently. In epidemiology, disease spreads continuously, but health officials only obtain periodic and potentially noisy data through tests of a relatively small number of individuals. The objective is to control the spread by optimizing costly interventions like vaccination or quarantine. Lastly, in industrial process control, the state of a process evolves continuously, but observations from sensors are discrete and noisy. The goal is to control the process to ensure product quality while minimizing control and data acquisition costs.

Literature review

We start by describing the main frameworks for analyzing classical SOC problems. In the 1960s, Ronald A. Howard [How60] popularized the term "Markov Decision Processes" (MDPs) and developed the policy iteration method, which is a fundamental technique for solving MDPs. Throughout the 1970s and 1980s, MDPs addressed more complex settings, including continuous-time processes and infinite-horizon problems, leading to the development of methods like value iteration and Q-learning. The integration of MDPs with machine learning, particularly reinforcement learning (RL), in the late 1980s and 1990s, see [BT96; Sze10; SB18], among others, has led to significant advancements. This includes the development of algorithms like Temporal Difference (TD) learning and the popularization of Q-learning and deep reinforcement learning. Today, MDPs are central to many applications in artificial intelligence, robotics, economics, and operations research. The Kalman filter [Kal60] became an essential tool for dealing with linear systems with Gaussian noise, while the need for handling non-linear systems led to the development of the Extended Kalman Filter [WB95], Ensemble Kalman Filter [Eve94] and, later, the Particle Filter in the 1990s [DFG01]. The concept of separation principle in control theory, which suggests that control and filtering can be separated in some cases, was a significant result for linear systems but proved challenging in non-linear settings, see [AM79]. This book develops the separation principle in the context of linear systems and discusses its extension to non-linear systems through various filtering techniques.

The main alternative approach is based on the Hamilton-Jacobi-Bellman (HJB) equation. It describes the optimal cost function's evolution in dynamic programming terms. This approach was extensively developed during the 1970s and 1980s. The application of viscosity solutions to HJB equations, cf. [Lio82; CL83; CL84] by Michael Crandall and Pierre-Louis Lions in the 1980s provided critical mathematical tools for dealing with the challenges posed by the non-linearity and high dimensionality of realistic control problems. Modern research focuses on bridging the gap between theoretical optimality and practical computability, especially in high-dimensional spaces where traditional methods are computationally infeasible. Applications now span complex systems in finance, engineering, and networked systems, where uncertainty and partial observability are key concerns.

The research in MDPs and stochastic control of *partially observed* systems continues to be a vibrant field, driven by both theoretical interests and practical applications.

To fix notations, let us assume that we are controlling a system X_t in continuous time $t \in [0,T]$, which is given as the solution of a stochastic differential equation (SDE), symbolically

$$\mathrm{d}X_t = b(X_t;\alpha)\mathrm{d}t + \sigma(X_t;\alpha)\mathrm{d}W_t, \tag{1.1}$$

driven by m-dimensional Brownian motion W, taking values in $\mathbb{R}^d$. Here, α denotes the control, taking values in a suitable set – and being progressively measurable w.r.t. a suitable filtration. Suppose now that rather than X_t, we observe a process Y_t satisfying

$$\mathrm{d}Y_t = h(X_t)\mathrm{d}t + \zeta\mathrm{d}B_t,$$

driven by another Brownian motion B. (In this context, X is often referred to as the *signal* process and Y as the *observation* process.) Given the observations $(Y_s)_{s\in[0,t]}$ up to time t, we can first compute the *conditional distribution* μ_t of the state process X_t at time t, i.e., we solve the *filtering problem*. Note that μ_t is a measure-valued stochastic process, which is adapted to the filtration generated by the observation process (Y_t). The control problem can now be re-expressed as a control problem for the conditional distribution, i.e., a measure-valued stochastic optimal control problem. However, the analysis and numerics for such stochastic optimal control problems is much less understood compared to the standard, finite-dimensional situation.

An important technical tool required for deriving dynamic programming principles or Hamilton-Jacobi-Bellman equations for measure-valued stochastic processes is an appropriate Itô formula. There has been renewed interest in this problem, mainly coming from mean-field games and control of McKean–Vlasov equations. A very general Itô formula for measure-valued semi-martingales has recently been derived in [GPW23] and references therein. More related works in the context of mean-field optimal control or optimal stopping, we refer to [TTZ23a; TTZ23b; GPW22].

The SOC problem with partial state observation is better understood when controls are *relaxed* i.e., α_t is replaced by a measure, see, for instance, [FN84; EKNJP88] for existence results for relaxed optimal controls. The classical work by Fleming [Fle80] introduced measure-valued processes for partially observed control problems, providing a theoretical foundation for analyzing stochastic control problems where the state is not fully observable. Fleming's insights are essential for our analysis of measure-valued processes. Recent works have made significant strides in approximating SOC problems under partial observation. Tan and Yang [TY23] discuss discrete-time approximation of continuous-time stochastic control problems under continuous time, partial observation. Their methods' convergence properties support our work's theoretical foundation, especially in the context of the measure-valued control framework we discuss. This recent work deals with approximating schemes for filtered problems, which may be relevant to our problem, which is not a standard filtering-control problem and only uses discrete time, partial observations.

The situation changes quite drastically when we consider SOC problems formulated within the field of MDPs. Indeed, there is a classical and rich literature on so-called *partially observed Markov decision processes* (POMDPs), which, as the name suggests, is concerned with MDPs where only partial, noisy observation of the controlled state process are available. We refer to [Kri16] for a monograph and [BR17] for an interesting recent application to an economics problem. As discussed earlier, the strategy is to lift the POMDP by considering the conditional distribution of the unobserved full state. The resulting control problem for measure-valued (hence, in general, infinite-dimensional) processes is the seen to be a standard MDP, and can be analysed by standard methods. Note that, in contrast to the stochastic optimal control literature, the MDP literature is almost exclusively concerned with discrete time problems, and this also extends to the partially-observed case.

The problem of SOC under noisy observation is also related to the problem of reinforcement learning [SB18]. Indeed, reinforcement learning operates under even less information, since not even knowledge of the driving dynamics of the controlled system is given, but rather has to be learned while controlling the system. Of course, stationarity of the system is usually assumed. Reinforcement learning is, again, generically formulated in discrete time, even though some attempts of continuous time extensions have recently been made, see, e.g., [WZZ20].

Our contribution

In our work, we consider a different problem, arising from the need to develop more efficient and reliable control strategies for systems with discrete, partial and noisy observations. Consider a continuous time SOC (1.1) with *partial* and *noisy* observations Y_{t_i} available at (fixed) discrete times $0 < t_1 < t_2 < \cdots t_n < T$. As before, decisions have to be based on the conditional distribution μ_t of the (unobserved) state process X_t given all the observations already available to us, i.e., $\{ Y_{t_i} : t_i \le t \}$. The dynamics of μ_t can now be described as follows:

1 Between observation times, μ_t follows the Fokker-Planck equation associated to the process (1.1) – a *deterministic* dynamics.

2 At an observation time $t = t_i$, we *update* the conditional distribution with the (random) new information Y_{t_i}, leading to a random jump $\mu_{t_i} = K_\epsilon(\mu_{t_i^-}, Y_{t_i})\mu_{t_i^-}$, where K_ϵ denotes the Radon–Nikodym derivative of the updated distribution w.r.t. the distribution prior to the update, and ϵ indicates the level of the noise in our observation.

We refer to (3.7) for the precise dynamics in the controlled case. Note that in a statistical sense, the second step can be interpreted as a *Bayesian update* of the *prior distribution* $\mu_{t_i^-}$ to a *posterior distribution* μ_{t_i} at time t_i taking into account our data Y_{t_i}.

As in the case of continuous time observation, we replace the control problem in the unobserved state X by the corresponding control problem in its conditional distribution μ_t. The dynamics of μ_t is, however, quite different from the continuous-time case, at least on a formal level. Rather than solving a Zakai SPDE, we need to solve a deterministic PDE with finitely many stochastic jumps. Of course, this problem is still infinite-dimensional.

The above setup invites to include a second control into our problem. As already indicated above, measurements are usually noisy (say, with standard deviation ϵ), and may not convey information about the full state X_t, anyway, – think of noisy observations of a function of X_t, e.g., just a single component. However, in many cases, *more precise measurement* methods may be available, albeit at a higher *cost*. Hence, in addition to the control α guiding the state process X_t, we may consider a second control β for the measurement method, as well as an associated cost term. In the statistical literature, this is also known as *optimal experimental design*, see, for instance, [SKA18]. Mathematically, this means that we obtain Bayesian updates $\mu_{t_i} = K_{\beta_{t_i}}(\mu_{t_i^-}, Y_{t_i})\mu_{t_i^-}$ which are directly influenced by the control, not only indirectly via μ and Y. Of course, this principle may also be applied to the observation times themselves, which could, more generally, be chosen by the controller.

From a computational perspective, we are still faced with an inherently infinite-dimensional SOC, since the conditional distribution μ_t takes values in the set of probability measures on the underlying state space. We propose to solve the stochastic optimal control problem numerically under the assumption that μ_t can be accurately characterized by the expectations $\int \varphi_j \mathrm{d}\mu_t$ of finitely many test functions φ_j, $j \in \mathcal{J}$ – either exactly, or in an approximate sense. In this case, the HJB equation for our SOC can be reduced to an HJB equation in $|\mathcal{J}|$ space variables and one time variable.

As an example, consider a drift-controlled Ornstein–Uhlenbeck process, in which case all conditional distributions μ_t are Gaussian, and, hence, can be characterized by their means m_t and co-variances σ_t^2. More generally, assume that the conditional distribution μ_t can be approximated by Gaussians $\mathcal{N}(m_t, \sigma_t^2)$. We can still solve the corresponding HJB equation, projecting both the dynamics of μ_t between observation times as well as the Bayesian updates to the set of Gaussian distributions as we go. This method can be interpreted as a generalization of the Kalman filter.

Outline of this work

We start with a motivating example of a SOC problem in continuous time with discrete time noisy observations in Section 2. In the following Section 3 we provide a formal setup for the problem, prove that the dynamic programming principle holds and derive the HJB equations under suitable regularity conditions. We then discuss the specific example of a drift-controlled one-dimensional Ornstein–Uhlenbeck process under observations with additive, independent Gaussian noise and quadratic costs, see Section 4. We observe that the problem cannot be reduced to solving a system of Riccati ODEs as usual, but we derive the associated HJB equation in time, mean and variance of the underlying Gaussian distribution μ_t. We then propose an appropriate finite-difference solver for the HJB equation, analyze its behavior and provide numerical examples in Section 5. Finally, in Section 6 we generalize the HJB equation to the multi-dimensional (approximately) Gaussian case, and make the link to the Kalman filter.

2 A motivating example

Consider an example based on a controlled Ornstein-Uhlenbeck process, with variance controlled Gaussian noisy measurements at times t_i, $i = 1, \dots, n$. More precisely, for $t \geq 0$, we consider the following controlled SDE representing the unobserved dynamics of the real-valued process X:

$$dX_t = (-\theta X_t + \alpha_t)\, dt + b\, dW_t, \tag{2.1}$$

where $\alpha_t \in \mathbb{R}$ is the control and $X_0 \sim \mu_0$ is the initial condition.

Let μ_t^α denote the random evolution of the conditional law of the unobserved controlled process X^α under the control process α_t. The natural conditioning event corresponds to all observations made up to time t, i.e., $t_i \leq t$.

Our goal is to minimize the expected cost given by a running cost, a final cost, and a cost associated to each of the measurements, namely

$$J(\alpha) \coloneqq \mathbb{E}\left[\int_0^T \left(\int_{\mathbb{R}} x^2 \mu_t^\alpha(\mathrm{d}x) + C\alpha_t^2\right) dt + \int_{\mathbb{R}} x^2 \mu_T^\alpha(\mathrm{d}x) + \sum_{i=1}^n \frac{1}{\beta_i}\right]. \tag{2.2}$$

The last sum term in the above corresponds to the cost of the measurements. More precisely, at each time t_i, we observe a noisy version of X_{t_i},

$$Y_i \coloneqq X_{t_i} + \beta_i Z_i,$$

where $Z_i \sim \mathcal{N}(0,1)$ are independent of each other and all other sources of randomness, and $X_{t_i} \sim \mu_{t_i^-}^\alpha$. We assume that the noise level $\beta_i > 0$ of the measurement at time t_i is also a control parameter. Therefore, at each time t_i, we have to update the conditional distribution μ^α according to the Bayesian update:

$$\mu_{t_i}^\alpha(\mathrm{d}y) \propto K_{\beta_i}(y; \mu_{t_i^-}^\alpha, Y_i)\mu_{t_i^-}^\alpha(\mathrm{d}y), \tag{2.3}$$

where $K_\beta(y; \mu_{t_i^-}^\alpha, Y_i)$ denotes the Gaussian likelihood corresponding to the measurement Y_i. It is intuitive to see that smaller values of β_i will provide better measurements and thus reduce the values of the first two terms in the objective functional (2.2). However, these smaller β_i values will make the last term in (2.2) larger, indicating that there is a non-trivial tradeoff to address when solving this SOC problem.

Motivated by this problem, we will state in the next section our SOC problem entirely in terms of the evolution of the conditional distribution, which is an interlaced sequence of controlled time evolutions, governed by the control α_t during intervals (t_i, t_{i+1}), with discontinuous jumps at times t_i given by the Bayesian updates, which are controlled by the β_i. This evolution of μ_t will be then Markovian and will naturally take us into a sequence of interlaced HJB equations, connected by conditional expectations matching conditions. See Section 3 for more details.

3 General theory

For a fixed time horizon $T > 0$, consider a large enough filtered probability space $(\Omega, \mathcal{F}, \mathbb{F}, \mathbb{P})$. We also denote by $\mathcal{P}(\mathbb{R}^d)$ the set of probability measures on $\big(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d)\big)$, furnished with the topology of weak convergence. We construct a stochastic optimal control problem for a probability-measure-valued process $\mu_t = \mu_t(\mathrm{d}x; \omega)$ generalizing the example discussed in Section 2. In particular, we consider a continuous-time version of the problem, where the underlying, unobserved process is a diffusion process.

3.1 Formulation of the problem

To fix notation, let us first look at the dynamics of the uncontrolled process. To start out, we fix observation times $0 < t_1 < \cdots < t_n < T$, which are (for simplicity) assumed to be deterministic. Likewise, for simplicity we assume that there are no observations at times 0 and T, and we introduce $t_0 := 0$, $t_{n+1} := T$. We also introduce the notation $\lfloor t \rfloor := \max\{\, i \mid t_i < t \,\}$, with the convention that $\lfloor 0 \rfloor := 0$. Let $\mathcal{G}$ denote the infinitesimal generator of a d-dimensional diffusion process defined for $f \in C_b^2(\mathbb{R}^d)$ by

$$\mathcal{G}f(y) := \sum_{i=1}^d b_i(y)\partial_i f(y) + \frac{1}{2}\sum_{i,j=1}^d a_{ij}(y)\partial_{ij}^2 f(y). \tag{3.1}$$

Between observation times, the process μ_t satisfies the Fokker–Planck equation associated with $\mathcal{G}$, i.e., using the adjoint operator we write

$$\mathrm{d}\mu_t = \mathcal{G}^*\mu_t \mathrm{d}t, \quad t_i \le t < t_{i+1},\ i = 1, \dots, n. \tag{3.2a}$$

We note that $\mathcal{G}^*\mu_t$ can be seen as a differential operator acting on the density of μ_t – tacitly assuming the existence of such a density, as well as identifying the measure with its density. More generally, we can understand $\mathcal{G}^*$ as an operator acting on (signed) measures defined in a weak sense.

At each observation time t_i, we obtain a noisy observation Y_i, and update the conditional measure μ according to the *Bayes rule*

$$\mu_{t_i}(\mathrm{d}x) = K_\varepsilon(x; \mu_{t_i^-}, Y_i)\mu_{t_i^-}(\mathrm{d}x),\ i = 1, \dots, n. \tag{3.2b}$$

where the kernel K_ε depends on the noise level $\varepsilon \ge 0$ as well as the nature of the measurement procedure. The following example illustrates a typical choice of measurement procedure and the corresponding choice of the kernel.

Example 3.1 (A sequence of measurements)**.** Let the measurement at t_i is given by $Y_i = \widehat{X}_i + \varepsilon Z_i$ for $\widehat{X}_i \sim \mu_{t_i^-}$ and $Z_i \sim \mathcal{N}(0,1)$ independent random variables, where we assume, for simplicity, that $d = 1$. More specifically, let $(U_i)_{i=1}^n$, $(Z_i)_{i=1}^n$ denote two i.i.d. sequences of standard uniform and normal r.v.s, respectively. Then we define

$$\widehat{X}_i := F[\mu_{t_i^-}]^{-1}(U_i), \quad Y_i := \widehat{X}_i + \varepsilon Z_i,$$

where F maps probability measures to their c.d.f.s. With ρ_ε denoting the density of $\mathcal{N}(0, \varepsilon^2)$, the posterior distribution is given by (3.2b) with

$$K_\varepsilon(x; \mu, y) := \frac{L_\varepsilon(x; y)}{\int_{\mathbb{R}} L_\varepsilon(x; y)\mu(\mathrm{d}x)}, \quad L_\varepsilon(x; y) := \rho_\varepsilon(y - x).$$

As anticipated in Section 2, we ensure that the dynamics for μ_t satisfies the Markov property. For this reason, we impose

Assumption 3.2 (Likelihood structure)**.** The observation Y_i at time t_i is a deterministic function of the observation time t_i, the probability measure $\mu_{t_i^-}$, the noise level ε, and a random variable Γ_i belonging to an sequence of independent r.v.s $(\Gamma_i)_{i=1}^n$ which are independent of all other sources of randomness.

See the above Example 3.1 for an illustrative example of an update rule satisfying Assumption 3.2 – with $\Gamma_i = (U_i, Z_i)$. To reflect Assumption 3.2 in the notation, we may also write $Y_i = Y_i^{\varepsilon,\mu_{t_i^-}} = Y_i^{\varepsilon}$, depending on the context.

Note that the $\mathcal{P}(\mathbb{R}^d)$-valued Markov process μ_t defined above has the property that its dynamics between measurement times is deterministic. The sole stochastic effects enter through the measurement values at the measurement times. At the observation points, it jumps in a random manner according to the Bayes update (3.2b).

As such, it is easy to see that μ_t satisfies an Ïtô-formula”. Given $\Phi : \mathcal{P}(\mathbb{R}^d) \to \mathbb{R}$ “nice enough”, we associate to it its *flat derivative*, see, for instance, [Daw93, Page 18f] or [Kol10, Appendix F].

Definition 3.3 (Flat derivative)**.** A function $\Phi : \mathcal{P}(\mathbb{R}^d) \longrightarrow \mathbb{R}$ is said to be differentiable at μ with derivative $\frac{\delta\Phi}{\delta\mu} : \mathcal{P}(\mathbb{R}^d) \times \mathbb{R}^d \to \mathbb{R}$ if for any $\nu \in \mathcal{P}(\mathbb{R}^d)$,

$$\Phi(\mu) - \Phi(\nu) = \int_0^1 \langle \mu - \nu, \frac{\delta\Phi}{\delta\mu}(\nu + \lambda(\mu - \nu), \cdot)\rangle \, \mathrm{d}\lambda. \tag{3.3}$$

Moreover, assuming that Φ is actually even defined in a neighborhood of μ in the space of signed measures, we can equivalently require the limit

$$\frac{\delta\Phi}{\delta\mu}(\mu, x) := \lim_{\theta\downarrow 0} \frac{\Phi(\mu + \theta\delta_x) - \Phi(\mu)}{\theta} = \frac{\mathrm{d}}{\mathrm{d}\theta}\bigg|_{\theta=0} \Phi(\mu + \theta\delta_x) \tag{3.4}$$

to exist for every $x \in \mathbb{R}^d$.

The function Φ is continuously differentiable if furthermore the function $\mathbb{R}^d \ni x \mapsto \frac{\delta\Phi}{\delta\mu}(\mu, x)$ is continuous in $\mathbb{R}^d$.

Observe, that according to Definition 3.3, if we have $\Phi(\mu) = F(\langle \mu \, , f \rangle)$ for a differentiable function $F : \mathbb{R} \to \mathbb{R}$ and a bounded function $f : \mathbb{R}^d \to \mathbb{R}$, then the flat derivative $\frac{\delta\Phi}{\delta\mu}$ always exists and satisfies

$$\frac{\delta\Phi}{\delta\mu}(\mu, x) = F'(\langle \mu \, , f \rangle) f(x). \tag{3.5}$$

In this context, in the next lemma we state a chain rule for the following class of function.

Definition 3.4 (Class $\mathcal{S}^{1,1}(\mathcal{P}(\mathbb{R}^d))$)**.** We say that a function $\Phi \in \mathcal{S}^{1,1}(\mathcal{P}(\mathbb{R}^d))$ if there is a continuous version of the flat derivative $\frac{\delta\Phi}{\delta\mu}(\mu, x)$ such that

- the mapping $(\mu, x) \mapsto \frac{\delta\Phi}{\delta\mu}(\mu, x)$ is jointly continuous w.r.t. (μ, x),
- the mapping $x \mapsto \frac{\delta\Phi}{\delta\mu}(\mu, x)$ is twice continuously differentiable with bounded first and second order derivatives.

Lemma 3.5 (Chain rule)**.** *Assume $(\mu_t)_{t\in[0,T]}$ solves* (3.2) *and that the function $\Phi(t, \mu)$ from $[0, T] \times \mathcal{P}(\mathbb{R}^d)$ to $\mathbb{R}$ is differentiable w.r.t. the time variable t and is in $\mathcal{S}^{1,1}(\mathcal{P}(\mathbb{R}^d))$ w.r.t. μ.*

Then, we have

$$\begin{aligned}\Phi(t, \mu_t) = \Phi(0, \mu_0) + \int_0^t \left[\frac{\partial\Phi}{\partial s}(s, \mu_s) + \langle \mathcal{G}^* \mu_s, \frac{\delta\Phi}{\delta\mu}(s, \mu_s, \cdot)\rangle \right] \mathrm{d}s \\ + \sum_{t_i \le t} \left[\Phi\left(t_i, K_\epsilon(\cdot; \mu_{t_i^-}, Y_i)\mu_{t_i^-}\right) - \Phi(t_i, \mu_{t_i^-}) \right].\end{aligned} \tag{3.6}$$

Sketch of proof, following [GPW23]. The proof of the chain rule lemma is first checked for the class of functions $\Phi(\mu) := F(\langle \mu, f\rangle)$ for a differentiable function $F : \mathbb{R} \longrightarrow \mathbb{R}$ and a bounded function $f : \mathbb{R}^d \longrightarrow \mathbb{R}$. Then, the result is naturally extended to the class of functions

$$\Phi(\mu) := F(\langle \mu, f_1\rangle, \langle \mu, f_2\rangle, \dots, \langle \mu, f_n\rangle)$$

for any fixed $n \in \mathbb{N}$ and polynomial $F : \mathbb{R}^n \longrightarrow \mathbb{R}$ and polynomials $f_1, f_2, \dots, f_n : \mathbb{R}^d \longrightarrow \mathbb{R}$. Finally, apply the Stone-Weierstrass (density) theorem to conclude that the chain rule is valid for any differentiable function $\Phi(\mu)$. □

For the controlled dynamics, we consider the situation of the generator $\mathcal{G}$ depending on a control α. In general, we assume that both the drift b and the diffusion σ depend on the control α. In addition, we *may* also assume ε to be a control, which we then also denote by β – i.e., we consider controls α of the dynamics and β of the measurement we take. Hence, we consider $u := (\alpha, \beta) \in U \subset \mathbb{R}^{k_1} \times \mathbb{R}^{k_2}$. Consider the filtration $(\mathcal{F}_t)_{t\in[0,T]}$ generated by the observations, i.e.,

$$\mathcal{F}_t = \sigma\left(\{\, Y_i : t_i \le t \,\}\right).$$

Note that the filtration is constant between observation times and “jumps” at the observation times t_i – we denote $\mathcal{F}_{t_i^-} := \mathcal{F}_{t_{i-1}}$. We recall that the controls α and β play very different roles:

- α acts on the dynamics of the unobserved process in continuous time, and is adapted to the filtration $(\mathcal{F}_t)_{t\in[0,T]}$;
- β only acts as an impulse control at the observation times t_i, but can only use information available just before t_i – hence, it is predictable w.r.t. the filtration $(\mathcal{F}_t)_{t\in[0,T]}$ and not just adapted in general.

Admissible controls are processes in the following class:

Definition 3.6. Let $\mathcal{U}[0,T]$ denote the set of all controls $u = (\alpha, \beta)$ such that

1. α is adapted w.r.t. the filtration $(\mathcal{F}_t)_{t\in[0,T]}$.
2. β is a piece-wise constant process, which only jumps at observation times t_i, where it is left-continuous. As a process, it is *predictable* w.r.t. the filtration $(\mathcal{F}_t)_{t\in[0,T]}$.
3. For any t, we have $u_t = (\alpha_t, \beta_t) \in U$.

We also introduce the notation $\mathcal{U}[t,T]$, defined in the obvious way.

Note that β only acts at the observation times $t_1, \dots, t_n$. For this reason, abusing notation, we denote $\beta_i := \beta_{t_i}$ – keeping in mind that $\beta_i \in \mathcal{F}_{t_i^-}$. Furthermore, α being adapted to $(\mathcal{F}_t)_{t\in[0,T]}$ does, of course, not imply that α is piecewise-constant as well. It only implies that the dynamics of α_t between observation times has to be deterministic – conditional on all the observation made before time t.

We denote by $\mathcal{G}(\alpha)$ the *controlled generator* and by K_β the *controlled update*. The dynamics of the *controlled process* can, hence, be described by

$$\mathrm{d}\mu_t^u = \mathcal{G}^*(\alpha)\mu_t^u \mathrm{d}t, \quad t_i \le t < t_{i+1}, \tag{3.7a}$$

$$\mu_{t_{i+1}}^u(\mathrm{d}x) = K_{\beta_{i+1}}\left(x; \mu_{t_{i+1}^-}^u, Y_{i+1}^{\beta_{i+1}, \mu_{t_{i+1}^-}^u}\right)\mu_{t_{i+1}^-}^u(\mathrm{d}x), \tag{3.7b}$$

for $0 \le i < n$ and $\mu_0^u = \mu_0 \in \mathcal{P}(\mathbb{R}^d)$. The cost functional associated to the stochastic optimal control problem is given by

$$J(u) := \mathbb{E}\left[\int_0^T \ell\left(t, \mu_t^u, \alpha_t\right) \mathrm{d}t + \sum_{i=1}^n h(t_i, \mu_{t_i^-}^u, \beta_i) + g(\mu_T^u)\right], \tag{3.8}$$

where ℓ is the running cost associated to the control of the underlying process (generally only depending on α), whereas h is the cost associated to the design of the measurement (generally only depending on β). Finally, g is the terminal cost, which in most cases is linear on the measure.

The optimal control problem is to minimize (3.8) over all admissible controls u, subject to the dynamics (3.7). An optimizing control is denoted by u^* and the corresponding optimal path by μ^*_t. Under the above assumptions, we expect optimal controls to be of feedback form, i.e., $u_t = \theta(t, \mu_t)$.

3.2 Dynamic programming principle

For given $\mu \in \mathcal{P}(\mathbb{R}^d)$ and $0 \le s < T$, consider the (weak) solution of the controlled system

$$\mathrm{d}\mu^u_t = \mathcal{G}^*(\alpha)\mu^u_t \mathrm{d}t, \quad s \wedge t_i < t < t_{i+1}, \tag{3.9a}$$

$$\mu^u_{t_{i+1}}(\mathrm{d}x) = K_{\beta_{i+1}}\left(x; \mu^u_{t^-_{i+1}}, Y_{i+1}^{\beta_{i+1},\mu^u_{t^-_{i+1}}}\right)\mu^u_{t^-_{i+1}}(\mathrm{d}x), \tag{3.9b}$$

(for all i s.t. $t_i \ge s$) with $\mu^u_s = \mu$, denoted by $\mu^u_t(s,\mu)$ for $s \le t \le T$. By uniqueness of solutions, we have the *flow property* for $u \in \mathcal{U}[s,T]$ and $s \le t \le v \le T$:

$$\mu^u_v(s,\mu) = \mu^u_v(t, \mu^u_t(s,\mu)). \tag{3.10}$$

The flow property implies that the cost function associated to the control u satisfies

$$\begin{aligned} J(s,\mu; u|_{[s,T]}) \quad &:= \mathbb{E}\left[\int_s^T \ell(t, \mu^u_t(s,\mu), u_t)\mathrm{d}t + \sum_{i=\lfloor s\rfloor+1}^{n} h(t_i, \mu^u_{t^-_i}(s,\mu), \beta_i) + g(\mu^u_T)\right] \\ &= \mathbb{E}\left[\int_s^t \ell(r, \mu^u_r(s,\mu), u_r)\mathrm{d}r + \sum_{i=\lfloor s\rfloor+1}^{\lfloor t\rfloor} h(t_i, \mu^u_{t^-_i}(s,\mu), \beta_i)\right. \\ &\qquad\qquad \left. + J\left(t, \mu^u_t(s,\mu); u|_{[t,T]}\right)\right]. \end{aligned}$$

Definition 3.7. The *value function* $V : [0,T] \times \mathcal{P}(\mathbb{R}^d) \to \mathbb{R}$ is defined by $V(T,\mu) := g(\mu)$ and

$$V(s,\mu) := \inf_{u \in \mathcal{U}[s,T]} J(s,\mu;u), \quad 0 \le s < T, \quad \mu \in \mathcal{P}(\mathbb{R}^d).$$

The value function satisfies the *dynamic programming principle*:

Theorem 3.8 (Dynamic Programming Principle). *For any $0 \le s < T$, $\mu \in \mathcal{P}(\mathbb{R}^d)$, and $s \le t \le T$ we have*

$$V(s,\mu) = \inf_{u \in \mathcal{U}[s,t]} \mathbb{E}\left[\int_s^t \ell(r, \mu^u_r(s,\mu), u_r)\mathrm{d}r + \sum_{i=\lfloor s\rfloor+1}^{\lfloor t\rfloor} h(t_i, \mu^u_{t^-_i}(s,\mu), \beta_i) + V\left(t, \mu^u_t(s,\mu)\right)\right]. \tag{3.11}$$

In particular, at jump times t_i – interpreting $s = t^-_i$ and $t = t_i$ – we have

$$V(t^-_i, \mu) = \inf_{(\alpha,\beta)\in U}\left\{h(t_i,\mu,\beta) + \mathbb{E}\left[V(t_i, K_\beta(\cdot;\mu, Y_i^{\beta,\mu})\mu)\right]\right\}, \tag{3.12}$$

where the expectation is effectively taken over $Y_i^{\beta,\mu}$.

We note that the objective function in the optimization problem at the observation time t_i stated above does not, in fact, depend on α at all. Nonetheless, minimization over (α,β) is used to reflect the fact that the constraint $(\alpha,\beta) \in U$ acts on both components of the control.

Proof of Theorem (3.8)*.* The proof is standard. We recall it for convenience. Denote the right-hand side of (3.11) by $W(s,\mu)$.

For any $\epsilon > 0$, there exists a control $u^\epsilon \in \mathcal{U}[s,T]$ such that

$$\begin{aligned}
V(s,\mu) + \epsilon &\geq J(s,\mu;u^\epsilon) \\
&= \mathbb{E}\Bigg[\int_s^t \ell(r,\mu_r^{u^\epsilon}(s,\mu),u_r^\epsilon)\mathrm{d}r + \sum_{i=\lfloor s\rfloor+1}^{\lfloor t\rfloor} h(t_i,\mu_{t_i^-}^{u^\epsilon}(s,\mu),\beta_i) \\
&\qquad + J\left(t,\mu_t^{u^\epsilon}(s,\mu);u^\epsilon|_{[t,T]}\right)\Bigg] \\
&\geq \mathbb{E}\Bigg[\int_s^t \ell(r,\mu_r^{u^\epsilon}(s,\mu),u_r^\epsilon)\mathrm{d}r + \sum_{i=\lfloor s\rfloor+1}^{\lfloor t\rfloor} h(t_i,\mu_{t_i^-}^{u^\epsilon}(s,\mu),\beta_i) + V(t,\mu_t^{u^\epsilon}(s,\mu)\Bigg] \\
&\geq W(s,\mu).
\end{aligned}$$

To obtain the reverse inequality, we have for any $v \in \mathcal{U}[s,T]$

$$V(s,\mu) \leq J(s,\mu;v) = \\ \mathbb{E}\left[\int_s^t \ell(r,\mu_r^v(s,\mu),v_r)\mathrm{d}r + \sum_{i=\lfloor s\rfloor+1}^{\lfloor t\rfloor} h(t_i,\mu_{t_i^-}^v(s,\mu),\beta_i) + J(t,\mu_t^v(s,\mu);v)\right].$$

In particular, given arbitrary controls $u = (\alpha,\beta), u' = (\alpha',\beta') \in \mathcal{U}[s,T]$, by choosing

$$\nu(r) = \begin{cases} u(r), & r \in [s,t], \\ u'(r), & r \in (t,T], \end{cases}$$

we have

$$\int_s^t \ell(r,\mu_r^v(s,\mu),v_r)\mathrm{d}r + \sum_{i=\lfloor s\rfloor+1}^{\lfloor t\rfloor} h(t_i,\mu_{t_i^-}^v(s,\mu),\beta_i) = \\ \int_s^t \ell(r,\mu_r^u(s,\mu),u_r)\mathrm{d}r + \sum_{i=\lfloor s\rfloor+1}^{\lfloor t\rfloor} h(t_i,\mu_{t_i^-}^u(s,\mu),\beta_i),$$

and by the flow property (3.10), we obtain

$$J(t,\mu_t^v(s,\mu);\nu) = J(t,\mu_t^u(s,\mu;t);u').$$

Therefore, by taking the infimum over $u' \in \mathcal{U}[s,T]$ we obtain

$$V(s,\mu) \leq \mathbb{E}\left[\int_s^t \ell(r,\mu_r^u(s,\mu),u_r)\mathrm{d}r + \sum_{i=\lfloor s\rfloor+1}^{\lfloor t\rfloor} h(t_i,\mu_{t_i^-}^u(s,\mu),\beta_i) + V(t,\mu_t^u(s,\mu))\right].$$

Since u is an arbitrary admissible control, we finally obtain $V(s,\mu) \leq W(s,\mu)$. □

3.3 The HJB equation

We are now ready to derive the Hamilton–Jacobi–Bellman (HJB) equation for our control problem. We first introduce the *Hamiltonian* for the control problem outside the observation dates. For $t \in [0,T]$ and $\mu \in \mathcal{P}(\mathbb{R}^d)$ and $p \in C_b(\mathbb{R}^d)$ we set

$$\mathcal{H}(t,\mu,p) \coloneqq \inf_{(\alpha,\beta)\in U} \left\{\langle \mathcal{G}^*(\alpha)\mu\,,p\rangle + \ell(t,\mu,\alpha)\right\}. \tag{3.13}$$

Theorem 3.9 (HJB equation)**.** *Provided that the value function $V(t,\mu)$ is differentiable w.r.t. the time variable t and is in $\mathcal{S}^{1,1}(\mathcal{P}(\mathbb{R}^d))$ w.r.t. μ, it satisfies the HJB equation*

$$\frac{\partial V}{\partial t}(t,\mu) + \mathcal{H}\left(t,\mu,\frac{\delta V}{\delta \mu}(t,\mu,\cdot)\right) = 0, \quad t_i \le t < t_{i+1},\ i = 1,\dots,n, \tag{3.14a}$$

$$V(t_i^-,\mu) = \inf_{(\alpha,\beta)\in U} \left\{ h(t_i,\mu,\beta) + E\left[V(t_i, K_\beta(\cdot;\mu,Y_i^{\beta,\mu})\mu)\right]\right\}, \quad i = 1,\dots,n, \tag{3.14b}$$

$$V(T,\mu) = g(\mu). \tag{3.14c}$$

Fix a function $u^* : [0,T]\times\mathcal{P}(\mathbb{R}^d)\to U$ by

$$u^*(t,\mu) \in \begin{cases} \arg\min_{(\alpha,\beta)\in U} \left\{ \left\langle \mathcal{G}^*(\alpha)\mu\,,\, \frac{\delta V}{\delta\mu}(t,\mu,\cdot)\right\rangle + \ell(t,\mu,\alpha)\right\}, & t \notin \{\,t_1,\dots,t_n\,\}, \\ \arg\min_{(\alpha,\beta)\in U} \left\{ h(t_i,\mu,\beta) + E\left[V(t_i,K_\beta(\cdot;\mu,Y_i^{\beta,\mu})\mu)\right]\right\}, & t = t_i^-,\ i \in \{\,1,\dots,n\,\}, \end{cases} \tag{3.15}$$

provided the set of minimizers in (3.15) *is not empty. Additionally, assume that μ_t^* denotes an optimal path. Then an* optimal control *is defined by $u_t^* := u^*(t,\mu_t^*)$.*

Note that the minimization problem for the β-component of u^* is not unique for t not an observation time, which allows us to choose a piecewise-constant, predictable version, enforcing the admissibility conditions. We write $u^* = (\alpha^*,\beta^*)$ and, continuing our abuse of notation, $\beta_i^* = \beta^*_{t_i^-}$.

Proof of Theorem 3.9. First we note that at observation times t_i, (3.12) is exactly (3.14b). We only need to derive (3.14a). Consider in (3.11) a constant control $u := a = (\bar\alpha,\bar\beta)$ for some arbitrary $a \in U$. For any $t,\theta > 0$ such that, for some $i = 1,\dots,n$, $[t,t+\theta] \subset [t_i,t_{i+1})$, we have

$$V(t,\mu) \le \mathbb{E}\left[\int_t^{t+\theta} \ell(s,\mu_s^a(t,\mu),a)\,ds + V(t+\theta,\mu_{t+\theta}^a(t,\mu))\right].$$

Since V satisfies the smoothness assumptions of the chain rule (3.6), we obtain

$$\mathbb{E}\left[\int_t^{t+\theta} \frac{\partial V}{\partial s}(s,\mu_s^a(t,\mu)) + \langle \mathcal{G}^*(a)\mu_s^a(t,\mu), \frac{\delta V}{\delta\mu}(s,\mu_s^a(t,\mu),\cdot)\rangle + \ell(s,\mu_s^a(t,\mu),a)\,\mathrm{d}s\right] \quad \ge \quad 0.$$

By the mean-value theorem, dividing by θ and then sending it to 0, yields

$$\frac{\partial V}{\partial t}(t,\mu) + \langle \mathcal{G}^*(a)\mu, \frac{\delta V}{\delta\mu}(t,\mu,\cdot)\rangle + \ell(t,\mu,a) \ge 0, \quad t_i \le t < t_{i+1}, \quad i = 1,\dots,n.$$

Since this inequality is true for all $a \in U$, we obtain

$$\frac{\partial V}{\partial t}(t,\mu) + \mathcal{H}\left(t,\mu,\frac{\delta V}{\delta\mu}(t,\mu,\cdot)\right) \ge 0, \quad t_i \le t < t_{i+1}, \quad i = 1,\dots,n. \tag{3.16}$$

Now, suppose that $u^* = (\alpha^*,\beta^*)$ is an optimal control. Then the value function satisfies

$$V(t,\mu) = \mathbb{E}\left[\int_t^{t+\theta} \ell(s,\mu_s^{u^*}(t,\mu),u_s^*)\,\mathrm{d}s + \sum_{i=\lfloor t\rfloor+1}^{\lfloor t+\theta\rfloor} h(t_i,\mu^{u^*}_{t_i^-}(s,\mu),\beta_i^*) + V(t+\theta,\mu^{u^*}_{t+\theta}(t,\mu))\right].$$

In particular, at the jump times, we have

$$V(t_i^-,\mu^{u^*}_{t_i^-}) = h(t_i,\mu^{u^*}_{t_i^-},\beta_i^*) + E\left[V(t_i,K_{\beta_i^*}(\cdot;\mu^{u^*}_{t_i^-},Y_i^{\beta_i,\mu^{u^*}_{t_i^-}})\mu^{u^*}_{t_i^-})\right], \ i = 1,\dots,n. \tag{3.17}$$

Thus, by using a similar argument as above between jump times, we obtain

$$\frac{\partial V}{\partial t}(t,\mu)+\left[\langle \mathcal{G}^*(u^*)\mu, \frac{\delta V}{\delta \mu}(s,\mu)(\cdot)\rangle + \ell(s,\mu,u^*(s))\right]=0, \quad t_i \le t < t_{i+1},\ i=1,\dots,n,$$

which, in view of (3.16), suggests the value function should satisfy

$$\frac{\partial V}{\partial t}(t,\mu)+\inf_{a\in U}\left[\langle \mathcal{G}^*(a)\mu, \frac{\delta V}{\delta \mu}(s,\mu)(\cdot)\rangle + \ell(s,\mu,a)\right]=0, \quad t_i \le t < t_{i+1},\ i=1,\dots,n.$$

□

Next, we derive a verification theorem for our SOC problem.

Theorem 3.10 (Verification theorem). *Let* $W(t,\mu)$ *be a solution to the HJB equation* (3.14a)*,* (3.14b) *and* (3.14c)*. Then,*

$$W(s,\mu)=\inf_{u\in\mathcal{U}[s,T]} J(s,\mu;u), \quad 0\le s<T, \quad \mu\in\mathcal{P}(\mathbb{R}^d).$$

Furthermore, assume the function $\widehat{u}:[0,T]\times\mathcal{P}(\mathbb{R}^d)\to U$ *satisfies*

$$\widehat{u}(t,\mu)\in\begin{cases}\arg\min_{(\alpha,\beta)\in U}\left\{\langle\mathcal{G}^*(\alpha)\mu,p\rangle+\ell(t,\mu,\alpha)\right\}, & t\notin\{t_1,\dots,t_n\},\\ \arg\min_{(\alpha,\beta)\in U}\left\{h(t_i,\mu,\beta)+E\left[W(t_i,K_\beta(\cdot;\mu,Y_i^{\beta,\mu})\mu)\right]\right\}, & t=t_i^-,\ i\in\{1,\dots,n\}.\end{cases}$$

Then, the feedback control u^* *given by* $u_t^*:=\widehat{u}(t,\mu)$ *is optimal.*

Proof. In view of (3.14a), (3.14b) and (3.14c) it follows that, for each $a\in U$ and $(t,\mu)\in[0,T]\times\mathcal{P}(\mathbb{R}^d)$,

$$\begin{aligned}&\frac{\partial W}{\partial t}(t,\mu)+\langle\mathcal{G}^*(a)\mu,\frac{\delta W}{\delta\mu}(s,\mu,\cdot)\rangle+\ell(s,\mu,a)\ge 0, \quad t_i\le t<t_{i+1},\ i=1,\dots,n,\\ &W(t_i^-,\mu)=\inf_{(\alpha,\beta)\in U}\left\{h(t_i,\mu,\beta)+E\left[W(t_i,K_\beta(\cdot;\mu,Y_i^{\beta,\mu})\mu)\right]\right\}, \quad i=1,\dots,n,\end{aligned}$$

and $W(T,\mu)=g(\mu)$. Let us now replace t,a,μ by s and $u_r=(\alpha_r,\beta_r),\mu_r^u,\ s\le r\le T$. Upon conditioning on $\mu_s^u=\mu$, we get

$$W(s,\mu)\le\mathbb{E}\left[\int_s^T \ell(t,\mu_t^u(s,\mu),u_t)\mathrm{d}t+\sum_{i=\lfloor s\rfloor+1}^{n} h(t_i,\mu_{t_i^-}^u(s,\mu),\beta_{t_i})+g(\mu_T^u)\,\middle|\,\mu_s^u=\mu\right],$$

which entails that

$$W(s,\mu)=\inf_{u\in\mathcal{U}[s,T]} J(s,\mu;u), \quad 0\le s<T, \quad \mu\in\mathcal{P}(\mathbb{R}^d).$$

Now, for $u^*=\widehat{u}$, the last inequality becomes an equality. Therefore, $W(t,\mu)=J(s,\mu;u^*)$ i.e., u^* is optimal.

□

Remark 3.11 (Solving the HJB equation). The HJB equation (3.14) is a piecewise backward PDE. Given the solution $V(t_i^-,\cdot)$ at time t_i, $i=1,\dots,n+1$ (interpreted as $V(t_{n+1}^-,\mu)=g(\mu)$ at $t_{n+1}=T$), we obtain the solution on $[t_{i-1},t_i)$ by solving the PDE (3.14a) backward in time. Then, we define $V(t_{i-1}^-,\mu)$ by (3.14b), and continue as before.

Remark 3.12 (Parameterized HJB equation). Suppose that we are given a subset $\mathcal{M}\subset\mathcal{P}(\mathbb{R}^d)$, which is invariant under $\mathcal{G}^*(\alpha)$ as well as under the Bayesian update for any optimal control u, and that we start in $\mu_0\in\mathcal{M}$. Let us further assume that probability measures in $\mathcal{M}$ are uniquely characterized by the expectations of functions $\{\,\varphi_j\mid j\in\mathcal{J}\,\}$ indexed by a (finite or infinite) index set $\mathcal{J}$.

In this case, we can recast the value function as a function

$$V(t,\mu) = U\left(t, (\langle \mu\,, \varphi_j \rangle)_{j\in\mathcal{J}}\right),$$

with $\mu \in \mathcal{M}$, for some function $U(t,z)$, $z \in \mathbb{R}^{|\mathcal{J}|}$. Note that we thus have

$$\frac{\delta V}{\delta \mu}(t,\mu,\cdot) = \sum_{j\in\mathcal{J}} \frac{\partial U}{\partial z_j}\left(t, (\langle \mu\,, \varphi_j \rangle)_{j\in\mathcal{J}}\right)\varphi_j(\cdot). \tag{3.18}$$

Hence, the HJB equation (3.14) can be recasted into an HJB equation for $U(t,z)$.

Example 3.13. [Parameterized HJB equation] For a general example of Remark 3.12, consider a compact set $K \subset \mathbb{R}^d$ which is invariant under the controlled dynamics described by $\mathcal{G}^\alpha$. Let $\mathcal{M} \subset \mathcal{P}(\mathbb{R}^d)$ denote the set of probability measures supported by K, and further assume that the Bayesian update K_β leaves $\mathcal{M}$ invariant. Hence, if we start in $\mu_0 \in \mathcal{M}$, we stay in $\mathcal{M}$ for any choice of control.

Recall that any $\nu \in \mathcal{M}$ is characterized by its moments $\int \varphi_j(x)\nu(\mathrm{d}x)$, $\varphi_j(x) := x_1^{j_1}\cdots x_d^{j_d}$, $j \in \mathcal{J} = \mathbb{N}_0^d$. Hence, we are formally in the situation of Remark 3.12 and we can introduce moment coordinates in this space of measures, reducing the problem from a SOC formulation with probability measures valued states to the simpler case of a countable sequence valued state, that one may have to truncate for constructive approximation.

4 Optimal control of an Ornstein – Uhlenbeck process

Consider the linear–quadratic example based on a controlled Ornstein–Uhlenbeck process, with Gaussian noisy observations at times t_i, $i = 1,\dots,n$. More precisely, we consider the controlled generator

$$\mathcal{G}(\alpha)f(x) = (-\theta x + \alpha)\partial_x f(x) + \frac{1}{2}b^2\partial_{xx}f(x), \tag{4.1}$$

based on a control $\alpha \in \mathbb{R}$, implying that

$$\mathcal{G}(\alpha)^* p(x) = \theta p(x) + (\theta x - \alpha)\partial_x p(x) + \frac{1}{2}\partial_{xx}p(x).$$

We denote by μ_t^u the random evolution of conditional law of the unobserved controlled process X^u with a control process $u_t = (\alpha_t, \beta_t)$.

We try to minimize quadratic costs, i.e.,

$$\ell(t,\mu,\alpha) := \int_{\mathbb{R}} x^2\mu(\mathrm{d}x) + C\alpha^2, \quad g(\mu) := \int_{\mathbb{R}} x^2\mu(\mathrm{d}x), \tag{4.2}$$

with all other cost terms vanishing.

At time t_i we observe $Y_i^\beta := \widehat{X}_i^u + \beta_i Z_i$ with $Z_i \sim \mathcal{N}(0,1)$ – independent of each other and all other sources of randomness – and $\widehat{X}_i^u \sim \mu^u_{t_i^-}$, see Example 3.1. Hence, we update μ^u according to the specification

$$\mu^u_{t_i}(\mathrm{d}y) = K_{\beta_i}(y; \mu^u_{t_i^-}, Y_i^\beta)\mu^u_{t_i^-}(\mathrm{d}y). \tag{4.3}$$

For now, we assume that the noise level β is fixed, and not a control parameter. To make this clear notation-wise, we write $\varepsilon = \beta$.

Observe that the above updating rule preserves Gaussian random variables. Indeed, we have

Lemma 4.1. *Suppose that* $\mu = \mathcal{N}(m,\sigma^2)$. *Then* ν *defined by* $\nu(\mathrm{d}x) := K_\varepsilon(x;\mu,y)\mu(\mathrm{d}x)$ *is equal to* $\mathcal{N}\left(m + \frac{\sigma^2}{\sigma^2+\varepsilon^2}(y-m),\ \frac{\sigma^2\varepsilon^2}{\sigma^2+\varepsilon^2}\right)$.

Proof. Use the formula for the conditional distribution of $(X|Y = y)$ for jointly Gaussian (X, Y) and apply it with $X = \widehat{X}_i^u$ and $Y = Y_i^\varepsilon$. □

It is well–known, see [GPW23], that the value function to the stochastic optimal control problem – with no observation of the trajectories – is given as

$$V(t,\mu) = \zeta(t)\int_{\mathbb{R}} x^2\mu(\mathrm{d}x) + \eta(t)\left(\int_{\mathbb{R}} x\mu(\mathrm{d}x)\right)^2 + \xi(t) \tag{4.4}$$

in terms of the *Riccati equations*

$$\dot{\zeta}(t) - 2\theta\zeta(t) + 1 = 0, \tag{4.5a}$$

$$\dot{\eta}(t) - \frac{C-1}{C^2}\left(\zeta(t) + \eta(t)\right)^2 - 2\theta\eta(t) = 0, \tag{4.5b}$$

$$\dot{\xi}(t) + b^2\zeta(t) = 0, \tag{4.5c}$$

with terminal conditions $\zeta(T) = 1, \eta(T) = \xi(T) = 0.$

Now let us consider the case with noisy observations (in the sense of Example 3.1) at times $t_1, \dots, t_n$. It is easy to see that optimal trajectories of μ_t^u remain in the class of (random) normal distributions[1], provided that the initial distribution μ_0 is within this class. Nonetheless, the Bayesian update step (3.14b) destroys the form (4.4).

Lemma 4.2. *Assume that* $\mu = \mathcal{N}(m,\sigma^2)$ *for some* $m \in \mathbb{R}$, $\sigma^2 > 0$, *and* $V(t,\mu) = \alpha\int_{\mathbb{R}} x^2\mu(\mathrm{d}x) + \varepsilon\left(\int_{\mathbb{R}} x\mu(\mathrm{d}x)\right)^2 + \gamma = \alpha(m^2+\sigma^2) + \varepsilon m^2 + \gamma$. *For* $Y \sim \mathcal{N}(m, \sigma^2+\varepsilon^2)$, *we have*

$$E\left[V(t, K_\varepsilon(\cdot;\mu,Y)\mu)\right] = \alpha(m^2+\sigma^2) + \varepsilon\left(m^2 + \frac{\sigma^4}{\sigma^2+\varepsilon^2}\right) + \gamma,$$

which is not of the form (4.4) *for a Gaussian distribution unless* $\varepsilon = 0$.

Proof. By Lemma 4.1, we know that

$$\nu := K_\varepsilon(\cdot;\mu,Y) = \mathcal{N}\left(m + \frac{\sigma^2}{\sigma^2+\varepsilon^2}(Y-m),\ \frac{\sigma^2\varepsilon^2}{\sigma^2+\varepsilon^2}\right).$$

Hence,

$$V(t,\nu) = \alpha\left(\frac{\sigma^2\varepsilon^2}{\sigma^2+\varepsilon^2} + \left[m + \frac{\sigma^2}{\sigma^2+\varepsilon^2}(Y-m)\right]^2\right) + \varepsilon\left(m + \frac{\sigma^2}{\sigma^2+\varepsilon^2}(Y-m)\right)^2 + \gamma,$$

and we conclude by taking expectations. □

Following the parametric approach suggested in Remark 3.12, we note that normal distributions are uniquely determined by their first and second moments, or, more appropriately, by mean and variance. In order to avoid confusion with the measure μ and when taking derivatives, we use the variables m for the mean and z for the *variance* of our measures.

[1] I.e., normal distributions with possibly random mean or variance.

Lemma 4.3 (HJB, controlled O-U case). *Let $\widehat{U} = \widehat{U}(t,m,z)$ denote the value function $V(t,\mu)$ in terms of mean m and variance z of the Gaussian measure μ. The function $\widehat{U}$ satisfies the HJB equation*

$$\frac{\partial \widehat{U}}{\partial t}(t,m,z) + z + m^2 - \theta m \frac{\partial \widehat{U}}{\partial m}(t,m,z) + (b^2 - 2\theta z)\frac{\partial \widehat{U}}{\partial z}(t,m,z) \tag{4.6a}$$

$$- \frac{1}{4C}\left[\frac{\partial \widehat{U}}{\partial m}(t,m,z)\right]^2 = 0, \quad t_i \le t < t_{i+1},\ i = 1,...,n$$

$$\widehat{U}(t_i^-,m,z) = \int_{\mathbb{R}} \widehat{U}\left(t_i, m + \frac{z}{\sqrt{z+\varepsilon^2}}z, \frac{z\varepsilon^2}{z+\varepsilon^2}\right)\phi(z)\mathrm{d}z, \tag{4.6b}$$

$$\widehat{U}(T,m,z) = m^2 + z, \tag{4.6c}$$

where ϕ denotes the density of a standard normal.

Since the function $\widehat{U}$ is defined on $\mathbb{R}^+ \times \mathbb{R} \times \mathbb{R}^+$, we naturally wonder if we need to impose a value of $\widehat{U}$ on the boundary $z = 0$. Recall that equation (4.6a) - (4.6c) is a terminal value problem, which is solved backwards in time. Using the method of characteristics for first order nonlinear PDEs, it can be shown that for any point in the domain $\mathbb{R} \times \mathbb{R}^+$ at terminal time the information propagates backwards via the characteristic lines $z(\tau)$ and crosses the boundary $z = 0$ from inside the domain to the outside, and no information is entering the domain from the $z < 0$ (see Section 5). Therefore, we do not need to impose a boundary condition at $z = 0$.

Proof of Lemma 4.3. Following Remark 3.12, since we are working with Gaussian distributions in one dimension, it is enough to choose $\varphi_1(x) := x$, $\varphi_2(x) := x^2$, as well as $U(t,m,s) := V(t,\mu)$, for $m := \int_{\mathbb{R}} \varphi_1(x)\mu(\mathrm{d}x)$, $s := \int_{\mathbb{R}} \varphi_2(x)\mu(\mathrm{d}x)$. By (3.18), we have

$$\frac{\delta V}{\delta \mu}(t,\mu) = \frac{\partial U}{\partial m}(t,m,s)\varphi_1 + \frac{\partial U}{\partial s}(t,m,s)\varphi_2.$$

Note that

$$\begin{aligned}
\mathcal{H}\left(t,\mu,\frac{\delta V}{\delta \mu}(t,\mu)\right) &= \inf_{\alpha\in\mathbb{R}} \left\{\left\langle \mu\,, \mathcal{G}(\alpha)\frac{\delta V}{\delta \mu}(t,\mu)\right\rangle + \langle \mu\,, \varphi_2\rangle + C\alpha^2\right\}, \\
&= \inf_{\alpha\in\mathbb{R}} \left\{\int_{\mathbb{R}} \left[\frac{\partial U}{\partial m}(t,m,s)(-\theta x + \alpha) + 2\frac{\partial U}{\partial s}(t,m,s)x(-\theta x + \alpha) + b^2\frac{\partial U}{\partial s}(t,m,s)\right]\mu(\mathrm{d}x) + \right. \\
&\qquad \left. + \int_{\mathbb{R}^d} x^2 \mu(\mathrm{d}x) + C\alpha^2\right\} \\
&= \inf_{\alpha\in\mathbb{R}} \left\{C\alpha^2 + \left[\int_{\mathbb{R}} \left(\frac{\partial U}{\partial m}(t,m,s) + 2x\frac{\partial U}{\partial s}(t,m,s)\right)\mu(\mathrm{d}x)\right]\alpha + \right. \\
&\qquad \left. + \int_{\mathbb{R}} \left[x^2 - \theta x\frac{\partial U}{\partial m}(t,m,s) + (b^2 - 2\theta x^2)\frac{\partial U}{\partial s}(t,m,s)\right]\mu(\mathrm{d}x)\right\} \\
&= \inf_{\alpha\in\mathbb{R}} \left\{C\alpha^2 + \left(\frac{\partial U}{\partial m}(t,m,s) + 2m\frac{\partial U}{\partial s}(t,m,s)\right)u + \right. \\
&\qquad \left. + \left[s - \theta m\frac{\partial U}{\partial m}(t,m,s) + (b^2 - 2\theta s)\frac{\partial U}{\partial s}(t,m,s)\right]\right\} \\
&= s - \theta m\frac{\partial U}{\partial m}(t,m,s) + (b^2 - 2\theta s)\frac{\partial U}{\partial s}(t,m,s) \\
&\qquad - \frac{1}{4C}\left(\frac{\partial U}{\partial m}(t,m,s) + 2m\frac{\partial U}{\partial s}(t,m,s)\right)^2.
\end{aligned}$$

Changing variables $s \to z = s - m^2$, we get

$$\frac{\partial U}{\partial m}(t,m,s) = \frac{\partial \widehat{U}}{\partial m}(t,m,z) - 2m\frac{\partial \widehat{U}}{\partial z}(t,m,z),$$

whereas derivatives w.r.t. s can just be replaced by derivatives w.r.t. z. This change of variables implies that

$$\mathcal{H}\left(t,\mu,\frac{\delta V}{\delta\mu}(t,\mu)\right)=m^2+z-\theta m\frac{\partial\widehat{U}}{\partial m}(t,m,z)+(b^2-2\theta z)\frac{\partial\widehat{U}}{\partial s}(t,m,z)\\ -\frac{1}{4C}\left(\frac{\partial\widehat{U}}{\partial m}(t,m,z)\right)^2,$$

from which we immediately get (4.6a).

By Lemma 4.1, for $\mu=\mathcal{N}(m,z)$,

$$V(t_i,K_\varepsilon(\cdot;\mu,y))=U\left(t_i,m+\frac{z}{z+\varepsilon^2}(y-m),\frac{z\varepsilon^2}{z+\varepsilon^2}\right).$$

By assumption, $Y_i^\alpha=\widehat{X}_i^\alpha+\varepsilon Z_i\sim\mathcal{N}(m,z+\varepsilon^2)$, for $Z_i\sim\mathcal{N}(0,1)$, and provided that $\mu=\mathcal{N}(m,z)$. As $h\equiv 0$ and ε is considered fix (i.e., not controlled), (3.14b) implies (4.6b). □

5 Numerical approach

Recall from Lemma 4.3 the problem under consideration: find $\widehat{U}=\widehat{U}(t,m,z)$ which satisfies the HJB equation

$$\frac{\partial\widehat{U}}{\partial t}(t,m,z)+z+m^2-\theta m\frac{\partial\widehat{U}}{\partial m}(t,m,z)+(b^2-2\theta z)\frac{\partial\widehat{U}}{\partial z}(t,m,z)\\ -\frac{1}{4C}\left[\frac{\partial\widehat{U}}{\partial m}(t,m,z)\right]^2=0,\quad t_n\le t<t_{n+1},\quad n=0,...,N \tag{5.1a}$$

$$\widehat{U}(t_n^-,m,z)=\int_{\mathbb{R}}\widehat{U}\left(t_n,m+\frac{z}{\sqrt{z+\varepsilon^2}}w,\frac{z\varepsilon^2}{z+\varepsilon^2}\right)\phi(w)\mathrm{d}w, \tag{5.1b}$$

$$\widehat{U}(T,m,z)=m^2+z, \tag{5.1c}$$

where ϕ denotes the density of a standard normal.

In this section we present the numerical approach to solving this problem. Observe that this equation is defined on an unbounded domain. We show how we truncate the domain to be able to treat this problem numerically. We introduce the numerical scheme used to solve the HJB equation (5.1a) on time interval $t\in[t_n,t_{n+1})$, and we show how to transform the result at time t_i according to (5.1b). Finally, we present the results obtained testing this approach on a synthetic problem.

5.1 Choice of the computational domain and boundary conditions

As mentioned, the original problem is defined on an unbounded domain with $m\in(-\infty,\infty)$ and $z\in[0,\infty)$. However, to solve the equation numerically, one must choose the truncated domain. Furthermore, one should either prescribe the boundary conditions at the boundaries of the chosen domain, or propose the numerical scheme, which uses only the interior points to compute current values and is shown to be stable. The upwind difference scheme has the desired property as long as information propagates from the initial conditions by lines flowing outside the domain. To check if this is the case, one can compute the characteristic lines of the equation following section 3.2 of [Eva10]. In our case, the characteristic lines of (5.1a) are flowing from outside the domain, which is unwanted. However, if we restate the problem in a reversed time flow as follows

$$\frac{\partial \widehat{U}}{\partial t}(t,m,z) + H\Big(m,z,\frac{\partial \widehat{U}}{\partial m},\frac{\partial \widehat{U}}{\partial z}\Big) = 0, \quad t_i \le t < t_{i+1}, \tag{5.2a}$$

$$H\Big(m,z,\frac{\partial \widehat{U}}{\partial m},\frac{\partial \widehat{U}}{\partial z}\Big) := H_1\Big(m,z,\frac{\partial \widehat{U}}{\partial m}\Big) + H_2\Big(m,z,\frac{\partial \widehat{U}}{\partial z}\Big) \tag{5.2b}$$

$$H_1\Big(m,z,\frac{\partial \widehat{U}}{\partial m}\Big) := -m^2 + \theta m \frac{\partial \widehat{U}}{\partial m}(t,m,z) + \frac{1}{4C}\left[\frac{\partial \widehat{U}}{\partial m}(t,m,z)\right]^2, \tag{5.2c}$$

$$H_2\Big(m,z,\frac{\partial \widehat{U}}{\partial z}\Big) := -z - (b^2 - 2\theta z)\frac{\partial \widehat{U}}{\partial z}(t,m,z), \tag{5.2d}$$

we can choose a domain, such that the characteristic lines cross the boundary from inside the domain. See, for example, Figure 1, which shows the characteristic lines for $m \in [-1,1]$ and $z \in [0,1]$ for the synthetic problem with parameters given in subsection 5.5. The lines for m are symmetric around $m(0) = 0$; for z they are symmetric around $z = b^2/(2\theta) = 0.5$. Therefore, as long as the computational domain contains $m = 0$ and $z = 0.5$, the numerical solution is stable. The details on the derivation of the characteristics are presented in Appendix A. Additionally, Saldi, Basar, and Raginsky [SBR18] explored the asymptotic optimality of finite model approximations for partially observed Markov decision processes (POMDPs). Although relevant for a time discrete setting, their approximation results for POMDPs may provide theoretical support for our domain truncation numerical approach as well, connecting our work to established results in the approximation of partially observed control problems.

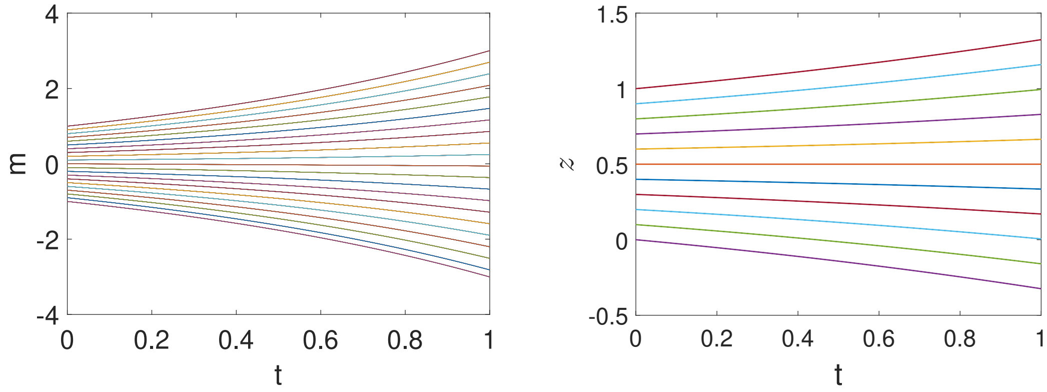


Figure 1: Characteristic lines shown for the domain of interest $m_i \in [-1,1]$, $z_j \in [0,1]$. For m, the lines are symmetric w.r.t. zero; for z they are symmetric around $z = b^2/(2\theta) = 0.5$.

5.2 Numerical scheme for HJB equation

In order to construct a convergent scheme we will follow the guidelines outlined in section 3.1 of [FF16] and [CL84]. We introduce the grid

$$\mathcal{G} := \{(m_i, z_j, t_n) : m_i = i\Delta m,\ z_j = j\Delta z,\ t_n = n\Delta t, \tag{5.3}$$

$$\forall i \in [0,\dots,I],\ j \in [0,\dots,J],\ n \in [0,\dots,N]\} \tag{5.4}$$

with uniform step sizes $\Delta m = \frac{m_I - m_0}{I}$, $\Delta z = \frac{z_J - z_0}{J}$, $\Delta t = \frac{t_N - t_0}{N}$. The boundary points m_0, m_I, z_0 and z_J are chosen to ensure the stability of the scheme as described in the previous subsection.

Let $U^n_{i,j} \approx \widehat{U}(t_n, m_i, z_j)$ denote the approximation of the solution to (5.2a). To approximate the spatial derivatives at time t_n we will use left and right finite differences denoted by $D^{\pm}(U^n_{i,j})$ and defined as follows:

$$\begin{aligned} D^-_m(U^n_{i,j}) &= \frac{U^n_{i,j} - U^n_{i-1,j}}{\Delta m}, \quad D^+_m(U^n_{i,j}) = \frac{U^n_{i+1,j} - U^n_{i,j}}{\Delta m}, \\ D^-_z(U^n_{i,j}) &= \frac{U^n_{i,j} - U^n_{i,j-1}}{\Delta z}, \quad D^+_z(U^n_{i,j}) = \frac{U^n_{i,j+1} - U^n_{i,j}}{\Delta z}. \end{aligned} \tag{5.5}$$

We aim to construct a scheme in a differenced form

$$U^{n+1}_{i,j} = G(U^n_{i-1,j}, U^n_{i,j}, U^n_{i+1,j}, U^n_{i,j-1}, U^n_{i,j+1}), \tag{5.6}$$

$$G(U^n_{i-1,j}, U^n_{i,j}, U^n_{i+1,j}, U^n_{i,j-1}, U^n_{i,j+1}) := \tag{5.7}$$

$$U^n_{i,j} - \Delta t \mathcal{H}(m, D^-_m(U^n_{i,j}), D^+_m(U^n_{i,j}); z, D^-_z(U^n_{i,j}), D^+_z(U^n_{i,j})), \tag{5.8}$$

where the function $\mathcal{H}$ is called the numerical Hamiltonian. Theorem 1 in [CL84] guarantees the convergence, provided that the scheme G is monotone and consistent.

Monotonicity: The scheme is said to be monotone, if the function G is monotone in each of its arguments as long as $|D^{\pm}_m(U^n_{i,j})|, |D^{\pm}_z(U^n_{i,j})| \le L$. To achieve that, the information about the speed of propagation of the solution $\partial H/\partial m$, $\partial H/\partial z$ is used. For instance, $\frac{\partial H_1}{\partial(\partial \hat{U}/\partial m)} \ge 0$, if $\frac{\partial \widehat{U}}{\partial m} \ge -2C\theta m$, and $\frac{\partial H_2}{\partial(\partial \hat{U}/\partial z)} \ge 0$, if $b^2 - 2\theta z \le 0$, thus the numerical Hamiltonians can be defined as follows:

$$\mathcal{H}_1(m, z, D^-_m, D^+_m) = \tag{5.9}$$

$$\begin{cases} H_1(m, z, D^-_m), & \text{if } D^-_m, D^+_m \ge -2C\theta m, \\ H_1(m, z, D^+_m) + H_1(m, z, D^-_m) - H_1(m, z, -2C\theta m), & \text{if } D^-_m \ge -2C\theta m,\ D^+_m \le -2C\theta m \\ H_1(m, z, -2C\theta m), & \text{if } D^-_m \le -2C\theta m,\ D^+_m \ge -2C\theta m \\ H_1(m, z, D^+_m), & \text{if } D^-_m, D^+_m \le -2C\theta m \end{cases}$$

$$\mathcal{H}_2(m, z, D^-_z, D^+_z) = \begin{cases} H_2(m, z, D^-_z), & \text{if } b^2 - 2\theta z < 0, \\ H_2(m, z, D^+_z), & \text{if } b^2 - 2\theta z \ge 0 \end{cases}, \tag{5.10}$$

$$\mathcal{H}(m, D^-_m(U^n_{i,j}), D^+_m(U^n_{i,j}); z, D^-_z(U^n_{i,j}), D^+_z(U^n_{i,j})) = \mathcal{H}_1(m, z, D^-_m, D^+_m) + \mathcal{H}_2(m, z, D^-_z, D^+_z) \tag{5.11}$$

With Hamiltonian defined in (5.11) the scheme G is monotone if

$$1 - 2\frac{\Delta t}{\Delta m}\Big|\frac{\partial H_1}{\partial(\partial \hat{U}/\partial m)}(m, z, \alpha)\Big| - \frac{\Delta t}{\Delta z}\Big|\frac{\partial H_2}{\partial(\partial \hat{U}/\partial z)}(m, z, \gamma)\Big| \ge 0 \tag{5.12}$$

as long as $|\alpha|$, $|\gamma| \le L$, which is an a priori bound on the derivatives of the value function. The derivation of (5.12) and the discussion of the bounds L are given in Appendix B.

Consistency. It is easy to check that the consistency of the proposed scheme is trivially satisfied

$$\mathcal{H}(m, \alpha, \alpha; z, \beta, \beta) = H_1(m, \alpha) + H_2(z, \beta) = H(m, z, \alpha, \beta). \tag{5.13}$$

5.3 Computation of $\hat{U}$ at time t_n.

Once we obtained the value of $\widehat{U}(t_n, m, z)$ at time t_n, the observation is made and this information is used to correct the value function $\widehat{U}(t_n^-, m, z)$ according to (5.1b):

$$\widehat{U}(t_n^-, m, z) = \int_{\mathbb{R}} \widehat{U}\Big(t_n, m + \frac{z}{\sqrt{z + \epsilon^2}} w, \frac{z\epsilon^2}{z + \epsilon^2}\Big)\phi(w)dw, \tag{5.14}$$

where $\phi(w) = \frac{1}{\sqrt{2\pi}}e^{-\frac{w^2}{2}}$. With a change of variable $-\frac{z}{\sqrt{z+\epsilon^2}}w = \tau$ we can rewrite the integral as

$$\widehat{U}(t_n^-, m, z) = \int_{\mathbb{R}} \widehat{U}\Big(t_n, m-\tau, \frac{z\epsilon^2}{z+\epsilon^2}\Big)\hat{\phi}(\tau)d\tau, \tag{5.15}$$

where

$$\hat{\phi}(\tau) = \frac{\sqrt{z+\epsilon^2}}{z}\frac{1}{\sqrt{2\pi}}e^{-\frac{\tau^2(z+\epsilon^2)}{2z^2}} \tag{5.16}$$

The integral (5.15) is a convolution for each given z and can be computed using Fourier transform. Note that the points $(\cdot, \cdot, \frac{z\epsilon^2}{z+\epsilon^2})$ are off grid $\mathcal{G}$, therefore the value function at these points can be interpolated from the nearby points.

5.4 Control under perfect observation

We now assume that we are in the classical case, namely that the process $X(t)$ can be fully observed at all times with no noise. Consider again the linear-quadratic example based on a controlled Ornstein-Uhlenbeck process described in Section 4

$$dX(t) = (-\theta X(t) + \alpha(t))dt + bdW(t),\ t > 0 \tag{5.17}$$

$$X(0) = X_0 \tag{5.18}$$

with the cost function to minimize given by

$$\mathbb{E}\Big[\int_0^T X(t)^2 dt + C\int_0^T \alpha(t)^2 dt\Big] \to \min \tag{5.19}$$

Define the value function

$$u(t,x) := \min_\alpha \mathbb{E}\Big[\int_t^T X(t)^2 dt + C\int_t^T \alpha(t)^2 dt\Big| X(t) = x\Big] \tag{5.20}$$

Then u solves the HJB equation

$$\partial_t u + H(x, t, \partial_x u, \partial_x^2 u) = 0,\ t < T \tag{5.21}$$

$$u(T, \cdot) = 0, \tag{5.22}$$

with

$$H(x,t,\partial_x u, \partial_x^2 u) = \min_\alpha\{(-\theta x + \alpha)\partial_x u + b^2\partial_x^2 u/2 + x^2 + C\alpha^2\} \tag{5.23}$$

$$= (-\theta x - \partial_x u/(2C))\partial_x u + b^2\partial_x^2 u/2 + x^2 + \frac{(\partial_x u)^2}{4C}. \tag{5.24}$$

One can check that $u(t,x) = f(t)x^2 + g(t)$ solves the HJB equation (5.21)-(5.23) if $f(t)$ solves the following Ricatti ODE

$$1 - 2\theta f - f^2/C + f' = 0, \tag{5.25}$$

$$f(T) = 0, \tag{5.26}$$

and $g(t) = \sigma^2\int_t^T f(s)ds$. Equation (5.25) can be integrated numerically. Thus, computing $u(t,x)$ provides a lower bound on the value function for the case with noisy measurements.

5.5 Simulations.

In this section we present the numerical results obtained for a test problem with the following parameters:

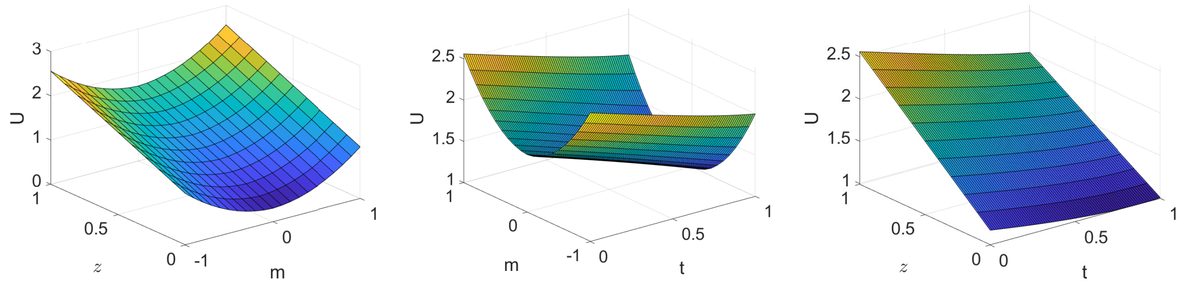
$$\theta = 0.25, b = 0.5, C = 1, \epsilon = 0.9 \tag{5.27}$$

We truncated the domain with $m \in [-1, 1]$, $z \in [0, 1]$, and $T \in [0, 1]$, as described in section 5.1. The step sizes $\Delta t = 0.0125$, $\Delta m = 0.1$, $\Delta z = 0.1$ were chosen to satisfy condition 5.12.

The tests were performed for three cases:

- Control under no observations
- Control under noisy observations (observations happen with time step $\Delta t_{obs} = 20\Delta t$)
- Control under perfect observation (no noise, full observation in continuous time)

Figure 2 (A) shows the projection of the value function onto axes of m and z at time $t = 0$, in the case of no observations. Observe that the cost is smaller for smaller values of z, because small z means more accurate information about the state. Figure 2 (B) shows the slices of the value function onto axes of t and m with $z = 1$; and (C) onto axes of t and z with $m = 1$.



(a)

(b)

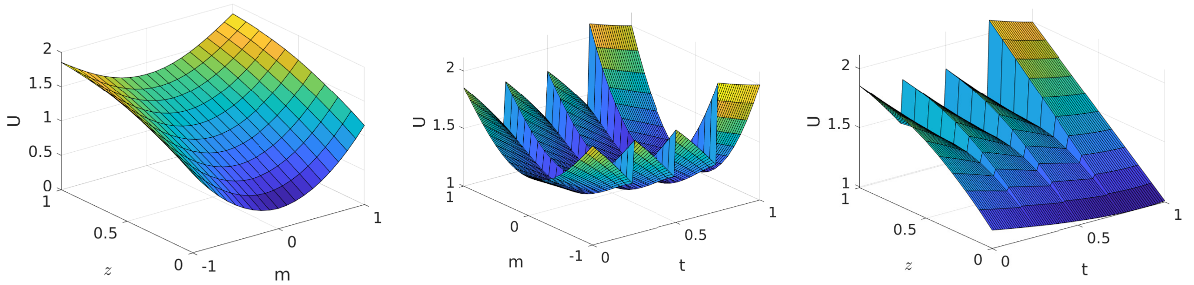
(c)

Figure 2: The value function of control under no observations problem. (A) The value function at initial time; (B) the slice of value function w.r.t. m and t, fixed $z = 1$; (C) the slice of value function w.r.t. z and t, fixed $m = 1$.

Figure 3 shows the same slices of the value function in the case with noisy observations made with time step $\Delta t_{obs} = 20\Delta t$. Overall we see that the cost is smaller compared to the case with no observations (Figure 2). Furthermore, according to (5.1b), at each observation point t_n^- and all z the value function is updated from the value function at points with significantly smaller variance $\frac{z\epsilon^2}{z+\epsilon^2}$ at time t_n, where the value function is smaller. That explains the jumps at each observation point which can be seen on Figure 3 (B) and (C). The information gained due to observation improves our knowledge and facilitates better decision making.

Figure 4 shows the optimal paths of the mean $m^*(t)$ and variance $z^*(t)$. The optimal paths are computed according to the dynamics (5.28) between the observation points and updated at the observation points using the simulated noisy observations of $X(t_{n+1})$, namely $Y(t_{n+1}) \sim \mathcal{N}(m^*(t_{n+1}^-), z^*(t_{n+1}^-) + \epsilon^2)$ as follows:

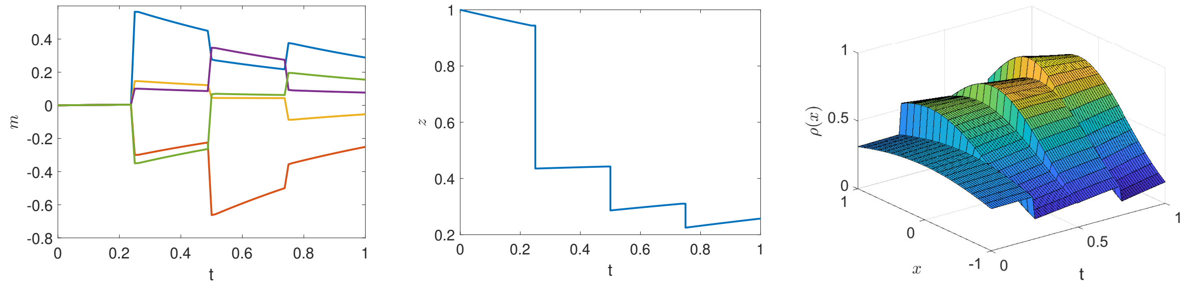
$$\begin{cases} dm^*(t) = \Big(-\theta m^*(t) + \frac{\partial \hat{U}}{\partial m}(t, m^*(t), z^*(t)) \Big) dt, & t_n \le t < t_{n+1},\ n = 0, ..., N \\ m^*(t_{n+1}) = m^*(t_{n+1}^-) + \frac{z^*(t_{n+1}^-)}{z^*(t_{n+1}^-)+\epsilon^2} (Y(t_{n+1}) - m^*(t_{n+1}^-)), & n = 0, ..., N \\ m^*(t_0) = m_0, & \\ dz^*(t) = \Big(-2\theta z^*(t) + b^2 \Big) dt, & t_n \le t < t_{n+1},\ n = 0, ..., N \\ z^*(t_{n+1}) = \frac{z^*(t_{n+1}^-)\epsilon^2}{z^*(t_{n+1}^-)+\epsilon^2}, & n = 0, ..., N \\ z^*(t_0) = z_0 & \end{cases} \tag{5.28}$$



(a) (b) (c)

Figure 3: The value function of control under noisy observations problem. (A) The value function at initial time; (B) the slice of value function w.r.t. m and t, fixed $z = 1$; (C) the slice of value function w.r.t. z and t, fixed $m = 1$.

The measurements $Y(t_{n+1})$ are independent in the sense that $Y(t_{n+1}) = m^*(t_{n+1}^-) + \sqrt{z^*(t_{n+1}^-) + \epsilon^2}\, W_{n+1}$ with $\{W_{n+1}\}_{n \geq 0}$ i.i.d. $\sim \mathcal{N}(0,1)$. Note also that m^* is updated using the obtained observation, while z^* evolves deterministically. The probability density function, $\rho(x)$, over the optimal path becomes more peaked around the mean from one observation point to the next as we collect information through Bayesian updates, as seen in Figure 4(C).



(a) (b) (c)

Figure 4: The optimal paths of the mean $m^*(t)$ and variance $z^*(t)$ of control under noisy observations problem. (A) The optimal paths of the mean $m^*(t)$ for several simulated scenarios; (B) the optimal path of the variance $z^*(t)$; (C) the probability density function, $\rho(x)$, over one of the optimal paths (the blue line on plot (A)).

Figure 5 illustrates the value function's slices w.r.t. m and t with fixed $z = 1$ for the three scenarios: (A) - perfect observations; (B) - noisy observations; (C) - no observations. Evidently, the case with perfect observations provides a lower bound for the solution in the presence of noisy observations, while the scenario with no observations gives the upper bound for the solution.

6 Multidimensional example - Kalman filters with control

In this section we generalize the linear quadratic example from Section 4 to the optimal control of a multidimensional Ornstein-Uhlenbeck process under noisy, discrete time Gaussian observations. While the multidimensional case is practically important, we initially introduced the concepts in a one-dimensional setting for didactical purposes. Here, we demonstrate that the information update step following each observation can be effectively characterized using the Kalman filter.

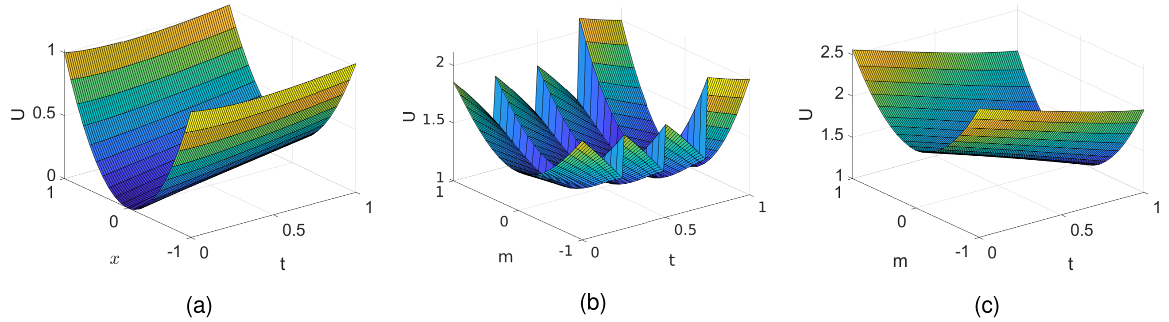


Figure 5: The comparison of the value functions at time $t = 0$ in three cases. (A) The value function of problem with full information; this case corresponds to the variance $z = 0$ in cases B and C with partially available information, and x is the exact known state which we compare to the mean m in B and C; (B) the slice of the value function of the problem with observations w.r.t. m and t, fixed $z = 1$; (C) the slice of the value function of the problem with no observations w.r.t. m and t, fixed $z = 1$.

Consider the multidimensional Ornstein-Uhlenbeck process in $\mathbb{R}^d$, with its controlled generator

$$\mathcal{G}(u)f(x) = \sum_{i=1}^{d} \Big((-\theta_i x_i + u_i)\partial_{x_i} f(x) + \sum_{j=i}^{d} \frac{b_{ij}^2}{2} \partial_{x_i x_j} f(x) \Big). \tag{6.1}$$

The sets of summation indices above are $\mathcal{I} = \{1, ..., d\}$, $\mathcal{J}_i = \{i, ..., d\}$, and we will later use the notation $\mathcal{J}_i^+ = \{i+1, ..., d\}$ as well. Following Remark 3.12, since we are working with multivariate Gaussian distributions, it is enough to choose $\varphi_{1,i}(x) = x_i$, $\varphi_{2,ij}(x) = x_i x_j$, $i, j = 1, ..., d$, and then represent the value function

$$V(t,\mu) = U(t, \{\langle \mu\,, \varphi_{1,i}\rangle\}_{i\in\mathcal{I}}, \{\langle \mu\,, \varphi_{2,ij}\rangle\}_{i\in\mathcal{I},\, j\in\mathcal{J}_i}) = U(t,m,s), \tag{6.2}$$

with the first moments $m = (m_1, ..., m_d) \in \mathbb{R}^d$, the second moments, $s = (\{s_{ij}\}_{i=1,...,d,\ j\in\mathcal{J}_i})$, and their corresponding definitions, namely $m_i = \int_{\mathbb{R}^d} \varphi_{1,i}(x)\mu(dx)$, $s_{ij} = \int_{\mathbb{R}^d} \varphi_{2,ij}(x)\mu(dx)$. Thus, $U : \mathbb{R}^+ \times \mathbb{R}^{1+d+\frac{d(d+1)}{2}} \mapsto \mathbb{R}$ is a real valued function with a particular domain, since the covariance matrix associated to μ has to be non negative definite.

Using the representation of the flat derivative (3.18), we obtain

$$\begin{aligned}\mathcal{G}(u)\frac{\delta V}{\delta \mu}(t,\mu)(x) = \sum_{i=1}^{d}\Bigg\{ (-\theta_i x_i + u_i)\Bigg(\sum_{l=1}^{d}\Bigg[\frac{\partial U}{\partial m_l}(t,m,s)\partial_{x_i}\varphi_{1,l}(x) + \sum_{k=l}^{d} \frac{\partial U}{\partial s_{lk}}(t,m,s)\partial_{x_i}\varphi_{2,lk}(x) \Bigg]\Bigg) \\ + \sum_{j=i}^{d} \frac{b_{ij}^2}{2}\Bigg(\sum_{l=1}^{d}\Bigg[\frac{\partial U}{\partial m_l}(t,m,s)\partial_{x_i x_j}\varphi_{1,l}(x) + \sum_{k=l}^{d} \frac{\partial U}{\partial s_{lk}}(t,m,s)\partial_{x_i x_j}\varphi_{2,lk}(x) \Bigg]\Bigg]\Bigg\}.\end{aligned} \tag{6.3}$$

Now we need to substitute the particular values of the partial derivatives of the φ functions. Indeed, noting right away that $\partial_{x_i x_j}\varphi_{1,l} \equiv 0$, each of the terms in (6.3) yield

$$\sum_{i=1}^{d}(-\theta_i x_i + u_i)\sum_{l=1}^{d}\frac{\partial U}{\partial m_l}(t,m,s)\partial_{x_i}\varphi_{1,l}(x) = \sum_{i=1}^{d}(-\theta_i x_i + u_i)\frac{\partial U}{\partial m_i}(t,m,s),$$

$$\begin{aligned}&\sum_{i=1}^{d}(-\theta_i x_i + u_i)\sum_{l=1}^{d}\sum_{k=l}^{d}\frac{\partial U}{\partial s_{lk}}(t,m,s)\partial_{x_i}\varphi_{2,lk}(x) = \\ &\sum_{i=1}^{d}(-\theta_i x_i + u_i)\Bigg(2\frac{\partial U}{\partial s_{ii}}(t,m,s)\varphi_{1,i}(x) + \sum_{k=i+1}^{d}\frac{\partial U}{\partial s_{ik}}(t,m,s)\varphi_{1,k}(x) + \sum_{1\le l<i}\frac{\partial U}{\partial s_{li}}(t,m,s)\varphi_{1,l}(x)\Bigg),\end{aligned}$$

and finally

$$\sum_{i=1}^{d}\sum_{j=i}^{d}\frac{b_{ij}^2}{2}\sum_{l=1}^{d}\sum_{k=l}^{d}\frac{\partial U}{\partial s_{lk}}(t,m,s)\partial_{x_i x_j}\varphi_{2,lk}(x) = \\ \sum_{i=1}^{d}\bigg(b_{ii}^2\frac{\partial U}{\partial s_{ii}}(t,m,s)+\sum_{j=i+1}^{d}\frac{b_{ij}^2}{2}\frac{\partial U}{\partial s_{ij}}(t,m,s)\bigg).$$

To compute the resulting Hamiltonian following (3.13), we need to use the value of $\mathcal{G}(u)\frac{\delta V}{\delta \mu}(t,\mu)$ from (6.3), which yields

$$\begin{aligned}
&\mathcal{H}\bigg(t,\mu,\frac{\delta V}{\delta\mu}(t,\mu)\bigg) = \inf_{u\in\mathbb{R}^d}\bigg\{\bigg\langle\mu\,,\mathcal{G}(u)\frac{\delta V}{\delta\mu}(t,\mu)\bigg\rangle+\sum_{i=1}^{d}\langle\mu\,,\varphi_{2,ii}\rangle+\sum_{i=1}^{d}C_i u_i^2\bigg\},\\
&= \inf_{u\in\mathbb{R}^d}\bigg\{\int_{\mathbb{R}^d}\bigg[\sum_{i=1}^{d}\frac{\partial U}{\partial m_i}(t,m,s)(-\theta_i x_i+u_i)\\
&\quad+\sum_{i=1}^{d}\Big(2x_i\frac{\partial U}{\partial s_{ii}}(t,m,s)+\sum_{j=i+1}^{d}x_j\frac{\partial U}{\partial s_{ij}}(t,m,s)+\sum_{1\le l<i}x_l\frac{\partial U}{\partial s_{li}}(t,m,s)\Big)(-\theta_i x_i+u_i)\\
&\quad+\sum_{i=1}^{d}\Big(b_{ii}^2\frac{\partial U}{\partial s_{ii}}(t,m,s)+\sum_{j=i+1}^{d}\frac{b_{ij}^2}{2}\frac{\partial U}{\partial s_{ij}}(t,m,s)\Big)\bigg]\mu(\mathrm{d}x)+\int_{\mathbb{R}^d}\sum_{i=1}^{d}x_i^2\mu(\mathrm{d}x)+\sum_{i=1}^{d}C_i u_i^2\bigg\}\\
&= \sum_{i=1}^{d}\bigg[s_{ii}-\theta_i m_i\frac{\partial U}{\partial m_i}(t,m,s)+(b_{ii}^2-2\theta_i s_{ii})\frac{\partial U}{\partial s_{ii}}(t,m,s)\\
&+\sum_{j=i+1}^{d}\Big(\frac{b_{ij}^2}{2}-\theta_i s_{ij}\Big)\frac{\partial U}{\partial s_{ij}}(t,m,s)-\theta_i\sum_{1\le l<i}s_{li}\frac{\partial U}{\partial s_{li}}(t,m,s)\bigg]\\
&\quad-\sum_{i=1}^{d}\frac{1}{4C_i}\left(\frac{\partial U}{\partial m_i}(t,m,s)+2m_i\frac{\partial U}{\partial s_{ii}}(t,m,s)+\sum_{j=i+1}^{d}m_j\frac{\partial U}{\partial s_{ij}}(t,m,s)+\sum_{1\le l<i}m_l\frac{\partial U}{\partial s_{li}}(t,m,s)\right)^2.
\end{aligned}$$

In the Hamiltonian above, U is a function of the first and the second moments. We can transform the input second moment variables of U into the corresponding entries of the covariance matrix of X, Σ, yielding a new function $\widehat{U}(t,m,z)$ as follows. Let $z_{ii} = s_{ii} - m_i^2$ and $z_{ij} = s_{ij} - m_i m_j$, so that $\Sigma = \Sigma(z)$ satisfies $\Sigma_{ij} = z_{ij}$ if $i \le j$ and $\Sigma_{ij} = z_{ji}$ otherwise. With this construction, Σ is only symmetric and we will have to ensure later that it is nonnegative definite by properly defining the state domain. The derivative $\frac{\partial U}{\partial m_i}(t,m,s)$, for $i \in \mathcal{I}$, can be written using the new variables, namely:

$$\frac{\partial U}{\partial m_i}(t,m,s) = \frac{\partial \widehat{U}}{\partial m_i}(t,m,z) - 2m_i\frac{\partial \widehat{U}}{\partial z_{ii}}(t,m,z) - \sum_{j=i+1}^{d} m_j\frac{\partial \widehat{U}}{\partial z_{ij}}(t,m,z). \tag{6.4}$$

Similarly, the partial derivatives of U w.r.t. s_{ii}, s_{ij} are replaced by the corresponding derivatives of $\widehat{U}$ w.r.t. z_{ii},

z_{ij}. The Hamiltonian for this problem expressed in terms of $\widehat{U}$ and the new input variables then reads:

$$\mathcal{H}\left(t, \mu, \frac{\delta V}{\delta \mu}(t,\mu)\right) = \hat{\mathcal{H}}\left(t, m, z, D\widehat{U}(t,m,z)\right)$$

$$= \sum_{i=1}^{d} \Bigg[z_{ii} + m_i^2 - \theta_i m_i \frac{\partial \widehat{U}}{\partial m_i}(t,m,z) + (b_{ii}^2 - 2\theta_i z_{ii}) \frac{\partial \widehat{U}}{\partial z_{ii}}(t,m,z) \tag{6.5}$$

$$- \theta_i \sum_{1 \le l < i} z_{li} \frac{\partial \widehat{U}}{\partial z_{li}}(t,m,z) - \theta_i \sum_{1 \le l < i} m_l m_i \frac{\partial \widehat{U}}{\partial z_{li}}(t,m,z) \tag{6.6}$$

$$+ \sum_{j=i+1}^{d} \left(\frac{b_{ij}^2}{2} - \theta_i z_{ij} \right) \frac{\partial \widehat{U}}{\partial z_{ij}}(t,m,z) - \frac{1}{4C_i} \left(\frac{\partial \widehat{U}}{\partial m_i}(t,m,z) + \sum_{1 \le l < i} m_l \frac{\partial \widehat{U}}{\partial z_{li}}(t,m,z) \right)^2 \Bigg].$$

This Hamiltonian determines the evolution of the value function $\widehat{U}$ in between observations according to the HJB equation in its natural domain, namely m in the same domain as x and z such that $\Sigma(z)$ is non negative definite.

Update after each observation:

At the observation time t_i we gather the datum $Y_i^\alpha := H\widehat{X}_i^\alpha + \varepsilon Z_i$ with $\widehat{X}_i^\alpha \in \mathbb{R}^d$, and Y_i^α, $Z_i \in \mathbb{R}^n$. Here we have $\widehat{X}_i^\alpha \sim \mu_{t_i^-} = \mathcal{N}(m_{t_i^-}, \Sigma_{t_i^-})$ and $Z_i \sim \mathcal{N}(0,1)$, mutually independent. After observation, the conditional distribution of $\widehat{X}_i^\alpha|_{Y_i^\alpha = y}$ is Gaussian and can be computed using the Kalman filter as follows:

$$\widehat{X}_i^\alpha|_{Y_i^\alpha = y} \sim \mathcal{N}\Big(m_{t_i}, \Sigma_{t_i}\Big) \tag{6.7}$$

$$\text{with } m_{t_i} = m_{t_i^-} + K_{t_i}(y - Hm_{t_i^-}), \tag{6.8}$$

$$\Sigma_{t_i} = (I - K_{t_i}H)\Sigma_{t_i^-}, \tag{6.9}$$

$$\text{and the Kalman gain matrix } K_{t_i} = \Sigma_{t_i^-} H^T \Big(H\Sigma_{t_i^-}H^T + \epsilon^2 I \Big)^{-1} \tag{6.10}$$

We now introduce the auxiliary function $\tilde{z} : \mathbb{R}^{d\times d} \to \mathbb{R}^{d+d(d+1)/2}$ with $\tilde{z}(\Sigma)_{ij} = \Sigma_{ij}$ for $i \le j$. Thus, according to (3.14b) we update $\widehat{U}(t_i^-, m_{t_i^-}, \tilde{z}(\Sigma_{t_i^-}))$ as follows:

$$\widehat{U}(t_i^-, m_{t_i^-}, \tilde{z}(\Sigma_{t_i^-})) = \int_{\mathbb{R}^n} \widehat{U}\Big(t_i, m_{t_i^-} + K_{t_i}Lw, \tilde{z}((I - K_{t_i}H)\Sigma_{t_i^-})\Big)\phi(w)\mathrm{d}w, \tag{6.11}$$

where the matrix L is such that $LL^T = H\Sigma_{t_{t_i}^-}H^T + \epsilon^2 I$ and $\phi(w)$ is a standard Gaussian density in $\mathbb{R}^n$. The matrix L can be computed using, for instance, a Cholesky decomposition.

Remark 6.1 (Observation costs)**.** We can have additional, controllable costs, from the observation as in (3.14b). In that case, (6.11) preserves the same structure, and only an additional optimization step is necessary to find the optimal data acquisition setup β.

Summing up, in this section we have characterized the different ingredients to set up HJB equation in our multivariate Gaussian process case. The Hamiltonian defined by (6.7) determines the evolution of the value function between observation times t_{i-1}, t_i, $i = 1, ..., n$, and the conditional expectation at each of the observation times t_i (6.11) using the Kalman filter equations. It is important to emphasize at this point that the state dimension of the time dependent HJB PDE is $d + \frac{d(d+1)}{2}$, which is relatively large, even for $d = 2$. This means that one should be particularly careful when choosing appropriate discretization tools, to address the ensuing curse of dimensionality.

A Details on the characteristics

As discussed in Section 5.1, it is necessary to truncate the originally unbounded domain to obtain a numerical solution for the proposed PDE (5.1a). To effectively truncate the domain and construct a convergent numerical scheme, it is crucial to understand the propagation of information within the PDE. To achieve this, we compute the characteristics of equation (5.1a) by following the methodology outlined in Section 3.2 of [Eva10]. These characteristics provide the curves along which information flows from initial points throughout the domain in first-order PDEs. By knowing these curves, we can appropriately select a domain where the characteristic lines flow from inside the domain to the outside, ensuring that the numerical scheme is well-posed and convergent.

Now we compute the characteristics of (5.1a). Letting $p = \frac{\partial U}{\partial m}$, $q = \frac{\partial U}{\partial z}$, the characteristics are given by the following system of ordinary differential equations (ODEs):

$$\begin{cases} \frac{dm}{d\tau} = \theta m(\tau) + \frac{1}{2C}p(\tau), \\ \frac{dz}{d\tau} = -(b^2 - 2\theta z(\tau)), \\ \frac{dp}{d\tau} = 2m(\tau) - \theta p(\tau), \\ \frac{dq}{d\tau} = 1 - 2\theta q(\tau). \end{cases} \tag{A.1}$$

The solution to this system is given by:

$$\begin{cases} z(\tau) = \frac{b^2}{2\theta} + C_1(m_0, z_0)e^{2\theta\tau}, \\ q(\tau) = \frac{1}{2\theta} + C_2(m_0, z_0)e^{-2\theta\tau}, \\ p(\tau) = C_3(m_0, z_0)e^{\sqrt{\theta^2+1/C}\tau} + C_4(m_0, z_0)e^{-\sqrt{\theta^2+1/C}\tau}, \\ m(\tau) = (\theta + \sqrt{\theta^2 + 1/C})\frac{C_3(m_0,z_0)}{2}e^{\sqrt{\theta^2+1/C}\tau} + (\theta - \sqrt{\theta^2 + 1/C})\frac{C_4(m_0,z_0)}{2}e^{-\sqrt{\theta^2+1/C}\tau} \end{cases} \tag{A.2}$$

where the constants C_1, C_2, C_3 and C_4 depend on the root of the characteristic line - the initial point (m_0, z_0), and can be determined from equations for $m(0)$, $z(0)$ and $p(m(0), z(0))$, $q(m(0), z(0))$. The explicit values of the constants are the following:

$$\begin{cases} C_1(m_0, z_0) = z_0 - \frac{b^2}{2\theta}, \\ C_2(m_0, z_0) = q_0(m_0, z_0) - \frac{1}{2\theta}, \\ C_3(m_0, z_0) = p(m_0, z_0) - C_4(m_0, z_0), \\ C_4(m_0, z_0) = \left(\frac{\theta+\sqrt{\theta^2+1/C}}{2}p(m_0, z_0) - m_0\right)\frac{1}{\sqrt{\theta^2+1/C}}. \end{cases} \tag{A.3}$$

Thus, the characteristic curves are given by (A.2) - (A.3). Figure 1 in Section 5.1 shows an example of the characteristics for $m \in [-1, 1]$ and $z \in [0, 1]$. Choosing a domain which contains $m = 0$ and $z = 0.5$ ensures that the information flows across the boundary from inside of the domain to the outside. A similar discussion applies to the multidimensional case.

B Derivation of the monotonicity condition

In Section 5.2 we construct the numerical scheme which is guaranteed to converge to the true solution, provided that it is consistent and monotone. In this Appendix we derive the condition which ensures that our scheme is monotone. To do that, we check directly under which conditions the function G is monotone in each of its arguments $U^n_{i-1,j}$, $U^n_{i,j}$, $U^n_{i+1,j}$, $U^n_{i,j-1}$, $U^n_{i,j+1}$.

Let's show the monotonicity w.r.t. $U^n_{i,j}$ as an example. For a grid point (m_i, z_j), let α_0 be the point where the Hamiltonian H_1 changes sign, so that $\frac{\partial H_1}{\partial(\partial \hat{U}/\partial m)}(m_i, z_j, \alpha)(\alpha - \alpha_0) \geq 0$. Let us also use the following notation for convenience:

$$\alpha = \frac{U^n_{i,j} - U^n_{i-1,j}}{\Delta m}, \quad \beta = \frac{U^n_{i+1,j} - U^n_{i,j}}{\Delta m}, \tag{B.1}$$

$$\gamma = \frac{U^n_{i,j+1} - U^n_{i,j}}{\Delta z}, \quad \delta = \frac{U^n_{i,j} - U^n_{i,j-1}}{\Delta z} \tag{B.2}$$

We write out the scheme G explicitly as follows:

$$U^{n+1}_{i,j} = U^n_{i,j} - \Delta t\Big[\mathbb{1}_{\{\beta \le \alpha_0\}}\Big(H_1\Big(m_i, z_j, \frac{U^n_{i+1,j} - U^n_{i,j}}{\Delta m}\Big) - H_1(m_i, z_j, \alpha_0)\Big) \tag{B.3}$$

$$+ (1 - \mathbb{1}_{\{\alpha \le \alpha_0\}})\Big(H_1\Big(m_i, z_j, \frac{U^n_{i,j} - U^n_{i-1,j}}{\Delta m}\Big) - H_1(m_i, z_j, \alpha_0)\Big) + H_1(m_i, z_j, \alpha_0) \tag{B.4}$$

$$+ \mathbb{1}_{\{b^2 - 2\theta z_j \ge 0\}} H_2\Big(m_i, z_j, \frac{U^n_{i,j+1} - U^n_{i,j}}{\Delta z}\Big) + \mathbb{1}_{\{b^2 - 2\theta z_j < 0\}} H_2\Big(m_i, z_j, \frac{U^n_{i,j} - U^n_{i,j-1}}{\Delta z}\Big)\Big] \tag{B.5}$$

The derivative of $U^{n+1}_{i,j}$ w.r.t. $U^n_{i,j}$ is given by the following:

$$\frac{\partial U^{n+1}_{i,j}}{\partial U^n_{i,j}} = 1 - \Delta t\Big[\underbrace{\mathbb{1}_{\{\beta \le \alpha_0\}} \frac{\partial H_1}{\partial(\partial \hat{U}/\partial m)}(m_i, z_j, \beta)\Big(-\frac{1}{\Delta m}\Big) + (1 - \mathbb{1}_{\{\alpha \le \alpha_0\}})\frac{\partial H_1}{\partial(\partial \hat{U}/\partial m)}(m_i, z_j, \alpha)\frac{1}{\Delta m}}_{:=h_1(m_i, z_j, \alpha, \beta)} \tag{B.6}$$

$$\underbrace{+\mathbb{1}_{\{b^2 - 2\theta z_j \ge 0\}} \frac{\partial H_2}{\partial(\partial \hat{U}/\partial z)}(m_i, z_j, \gamma)\Big(-\frac{1}{\Delta z}\Big) + \mathbb{1}_{\{b^2 - 2\theta z_j < 0\}} \frac{\partial H_2}{\partial(\partial \hat{U}/\partial z)}(m_i, z_j, \delta)\frac{1}{\Delta z}\Big]}_{:=h_2(m_i, z_j, \gamma, \delta)} \tag{B.7}$$

Since $U^{n+1}_{i,j}$ is an increasing function, so $\frac{\partial U^{n+1}_{i,j}}{\partial U^n_{i,j}} \ge 0$. Recalling that $\frac{\partial H_1}{\partial(\partial \hat{U}/\partial m)}(m_i, z_j, \alpha)(\alpha - \alpha_0) \ge 0$ and expanding on $h_1(m_i, \alpha, \beta)$ we get

$$h_1(m_i, z_j, \alpha, \beta) = \begin{cases} \frac{1}{\Delta m} \frac{\partial H_1}{\partial(\partial \hat{U}/\partial m)}(m_i, z_j, \beta), & \text{if } \beta, \alpha \le \alpha_0, \\ \frac{1}{\Delta m}\Big(\frac{\partial H_1}{\partial(\partial \hat{U}/\partial m)}(m_i, z_j, \beta) - \frac{\partial H_1}{\partial(\partial \hat{U}/\partial m)}(m_i, z_j, \alpha)\Big), & \text{if } \beta \le \alpha_0,\ \alpha \ge \alpha_0, \\ 0, & \text{if } \beta \ge \alpha_0,\ \alpha \le \alpha_0, \\ \frac{1}{\Delta m} \frac{\partial H_1}{\partial(\partial \hat{U}/\partial m)}(m_i, z_j, \alpha), & \text{if } \beta, \alpha \ge \alpha_0 \end{cases} \tag{B.8}$$

$$= \begin{cases} \frac{1}{\Delta m} \Big|\frac{\partial H_1}{\partial(\partial \hat{U}/\partial m)}(m_i, z_j, \beta)\Big|, & \text{if } \beta, \alpha \le \alpha_0, \\ \frac{1}{\Delta m}\Big(\Big|\frac{\partial H_1}{\partial(\partial \hat{U}/\partial m)}(m_i, z_j, \beta)\Big| + \Big|\frac{\partial H_1}{\partial(\partial \hat{U}/\partial m)}(m_i, z_j, \alpha)\Big|\Big), & \text{if } \beta \le \alpha_0,\ \alpha \ge \alpha_0, \\ 0, & \text{if } \beta \ge \alpha_0,\ \alpha \le \alpha_0, \\ \frac{1}{\Delta m} \Big|\frac{\partial H_1}{\partial(\partial \hat{U}/\partial m)}(m_i, z_j, \alpha)\Big|, & \text{if } \beta, \alpha \ge \alpha_0 \end{cases} \tag{B.9}$$

W.l.o.g., suppose that $\Big|\frac{\partial H_1}{\partial(\partial \hat{U}/\partial m)}(m_i, z_j, \alpha)\Big| \ge \Big|\frac{\partial H_1}{\partial(\partial \hat{U}/\partial m)}(m_i, z_j, \beta)\Big|$. Then

$$h_1(m_i, z_j, \alpha, \beta) \le \frac{2}{\Delta m}\Big|\frac{\partial H_1}{\partial(\partial \hat{U}/\partial m)}(m_i, z_j, \alpha)\Big| \tag{B.10}$$

Expanding on $h_2(m_i, z_j, \gamma, \delta)$ we get

$$h_2(m_i, z_j, \gamma, \delta) = \underbrace{\mathbb{1}_{\{b^2-2\theta z_j \geq 0\}} \frac{\partial H_2}{\partial(\partial \hat{U}/\partial z)}(m_i, z_j, \gamma)\Big(-\frac{1}{\Delta z}\Big)}_{\frac{\partial H_2}{\partial(\partial \hat{U}/\partial z)}(m_i,z_j,\gamma)\leq 0} + \underbrace{\mathbb{1}_{\{b^2-2\theta z_j<0\}} \frac{\partial H_2}{\partial(\partial \hat{U}/\partial z)}(z_j, \delta)\frac{1}{\Delta z}}_{\frac{\partial H_2}{\partial(\partial \hat{U}/\partial z)}(m_i,z_j,\delta)\geq 0} \tag{B.11}$$

$$= \mathbb{1}_{\{b^2-2\theta z_j \geq 0\}} \frac{1}{\Delta z}\Big|\frac{\partial H_2}{\partial(\partial \hat{U}/\partial z)}(m_i, z_j, \gamma)\Big| + \mathbb{1}_{\{b^2-2\theta z_j<0\}} \frac{1}{\Delta z}\Big|\frac{\partial H_2}{\partial(\partial \hat{U}/\partial z)}(m_i, z_j, \delta)\Big| \tag{B.12}$$

Suppose, w.l.o.g. that $\Big|\frac{\partial H_2}{\partial(\partial \hat{U}/\partial z)}(m_i, z_j, \gamma)\Big| \geq \Big|\frac{\partial H_2}{\partial(\partial \hat{U}/\partial z)}(m_i, z_j, \delta)\Big|$, then we have

$$h_2(m_i, z_j, \gamma, \delta) \leq \frac{1}{\Delta z}\Big|\frac{\partial H_2}{\partial(\partial \hat{U}/\partial z)}(m_i, z_j, \gamma)\Big| \tag{B.13}$$

Combining (B.10) and (B.13), we get the condition which guarantees the monotonicity of the scheme G:

$$1 - 2\frac{\Delta t}{\Delta m}\Big|\frac{\partial H_1}{\partial(\partial \hat{U}/\partial m)}(m_i, z_j, \alpha)\Big| - \frac{\Delta t}{\Delta z}\Big|\frac{\partial H_2}{\partial(\partial \hat{U}/\partial z)}(m_i, z_j, \gamma)\Big| \geq 0 \tag{B.14}$$

To compute the relations $\Delta t/\Delta m$, $\Delta t/\Delta z$, which satisfy condition (B.14), we must estimate $\Big|\frac{\partial H_1}{\partial(\partial \hat{U}/\partial m)}(m_i, z_j, \alpha)\Big|$ and $\Big|\frac{\partial H_2}{\partial(\partial \hat{U}/\partial z)}(m_i, z_j, \gamma)\Big|$. The derivatives are given by the following

$$\frac{\partial H_1}{\partial(\partial \hat{U}/\partial m)}(m_i, z_j, \alpha) = -\theta m_i + \frac{\alpha}{2C}, \tag{B.15}$$

$$\frac{\partial H_2}{\partial(\partial \hat{U}/\partial z)}(m_i, z_j, \gamma) = 2\theta z_j - b^2 \tag{B.16}$$

On a domain defined by $[m_0, m_I] \times [z_0, z_J]$ with $m_0 < 0$, we have the following bounds

$$\Big|\frac{\partial H_1}{\partial(\partial \hat{U}/\partial m)}(m_i, z_j, \alpha)\Big| \leq \theta|m_0| + \frac{|\alpha|}{2C}, \tag{B.17}$$

$$\Big|\frac{\partial H_2}{\partial(\partial \hat{U}/\partial z)}(m_i, z_j, \gamma)\Big| \leq 2\theta z_J - b^2 \tag{B.18}$$

Obtaining $\Big|\frac{\partial H_2}{\partial(\partial \hat{U}/\partial z)}(m_i, z_j, \gamma)\Big|$ is straightforward, while $\Big|\frac{\partial H_1}{\partial(\partial \hat{U}/\partial m)}(m_i, z_j, \alpha)\Big|$ requires an a priori upper bound on $|\alpha| = |\partial \widehat{U}/\partial m|$. For example, the Lipschitz constant can be used if known. Otherwise, condition (B.14) could be verified to hold at every point of the grid (m_i, z_j) during numerical simulation. It is easy to see that the requirement on the derivatives of $U^{n+1}_{i,j}$ (or G) w.r.t. the rest of the variables $U^n_{i-1,j}$, $U^n_{i+1,j}$, $U^n_{i,j-1}$, $U^n_{i,j+1}$ to be positive is also satisfied by (B.14).

References

[AM79] B. D. O. Anderson and J. B. Moore. *Optimal Filtering*. Prentice Hall, 1979.

[BCD97] M. Bardi and I. Capuzzo-Dolcetta. *Optimal Control and Viscosity Solutions of Hamilton-Jacobi-Bellman Equations*. Birkhäuser, 1997.

[Ben92] A. Bensoussan. *Stochastic Control of Partially Observable Systems*. Cambridge University Press, 1992.

[BR17] N. Bäuerle and U. Rieder. "Partially observable risk-sensitive Markov decision processes". In: *Math. Oper. Res.* 42.4 (2017), pp. 1180–1196. DOI: 10.1287/moor.2016.0844.

[BT96] D. P. Bertsekas and J. N. Tsitsiklis. *Neuro-Dynamic Programming*. Athena Scientific, 1996.

[CL83] M. G. Crandall and P.-L. Lions. "Viscosity solutions of Hamilton-Jacobi equations". In: *Transactions of the American Mathematical Society* 277.1 (1983), pp. 1–42.

[CL84] M. G. Crandall and P. L. Lions. "Two Approximations of Solutions of Hamilton-Jacobi Equations". In: *Mathematics of Computation* 43.167 (1984), pp. 1–19.

[Daw93] D. Dawson. "Measure-valued Markov processes". In: *École d'été de probabilités de Saint-Flour XXI-1991*. Springer, 1993, pp. 1–260.

[DFG01] A. Doucet, N. de Freitas, and N. Gordon, eds. *Sequential Monte Carlo methods in practice*. Springer-Verlag New York, 2001.

[EKNJP88] N. El Karoui, D. H. Nguyen, and M. Jeanblanc-Picqué. "Existence of an optimal Markovian filter for the control under partial observations". In: *SIAM J. Control Optim.* 26.5 (1988), pp. 1025–1061. DOI: 10.1137/0326057.

[Eva10] L. C. Evans. *Partial Differential Equations: Second Edition*. American Mathematical Society, 2010.

[Eve94] G. Evensen. "Data assimilation using an ensemble Kalman filter technique". In: *Monthly Weather Review* 122.5 (1994), pp. 895–910.

[FF16] M. Falcone and R. Ferretti. *Chapter 23 - Numerical Methods for Hamilton-Jacobi Type Equations*. Vol. 17. Elsevier, 2016, pp. 603–626.

[Fle80] W. H. Fleming. "Measure-valued processes in the control of partially-observable stochastic systems". In: *Applied Mathematics and Optimization* 6.1 (1980), pp. 271–285.

[FN84] W. H. Fleming and M. Nisio. "On stochastic relaxed control for partially observed diffusions". In: *Nagoya Math. J.* 93 (1984), pp. 71–108. DOI: 10.1017/S0027763000020742.

[FS93] W. H. Fleming and H. M. Soner. *Stochastic Control: Hamiltonian Systems and HJB Equations*. Springer-Verlag, 1993.

[GPW22] M. Germain, H. Pham, and X. Warin. "Rate of convergence for particle approximation of PDEs in Wasserstein space". In: *J. Appl. Probab.* 59.4 (2022), pp. 992–1008. DOI: 10.1017/jpr.2021.102.

[GPW23] X. Guo, H. Pham, and X. Wei. "Itô's formula for flows of measures on semimartingales". In: *Stochastic Process. Appl.* 159 (2023), pp. 350–390. DOI: 10.1016/j.spa.2023.02.004.

[How60] R. A. Howard. *Dynamic Programming and Markov Processes*. MIT Press, 1960.

[Kal60] R. E. Kalman. "A new approach to linear filtering and prediction problems". In: *Journal of Basic Engineering* 82.1 (1960), pp. 35–45.

[Kol10] V. N. Kolokoltsov. *Nonlinear Markov processes and kinetic equations*. Vol. 182. Cambridge University Press, 2010.

[Kri16] V. Krishnamurthy. *Partially observed Markov decision processes*. From filtering to controlled sensing. Cambridge University Press, Cambridge, 2016, pp. xiii+476. DOI: 10.1017/CBO9781316471104.

[Lio82] P.-L. Lions. *Generalized solutions of Hamilton-Jacobi equations*. Pitman (Advanced Publishing Program), 1982.

[Mao06] X. Mao. *Filtering and Control of Stochastic Jump Hybrid Systems*. Springer, 2006.

[SB18] R. S. Sutton and A. G. Barto. *Reinforcement learning: an introduction*. Second. Adaptive Computation and Machine Learning. MIT Press, Cambridge, MA, 2018, pp. xxii+526.

[SBR18] N. Saldi, T. Basar, and M. Raginsky. "Asymptotic optimality of finite model approximations for partially observed Markov decision processes". In: *IEEE Transactions on Automatic Control* 63.7 (2018), pp. 2031–2045.

[SKA18] B. Smucker, M. Krzywinski, and N. Altman. “Optimal experimental design”. In: *Nat. Methods* 15.8 (2018), pp. 559–560.

[Sze10] C. Szepesvári. *Algorithms for Reinforcement Learning*. Morgan and Claypool, 2010.

[TTZ23a] M. Talbi, N. Touzi, and J. Zhang. “Dynamic programming equation for the mean field optimal stopping problem”. In: *SIAM J. Control Optim.* 61.4 (2023), pp. 2140–2164. DOI: 10.1137/21M1404259.

[TTZ23b] M. Talbi, N. Touzi, and J. Zhang. “Viscosity solutions for obstacle problems on Wasserstein space”. In: *SIAM J. Control Optim.* 61.3 (2023), pp. 1712–1736. DOI: 10.1137/22M1488119.

[TY23] X. Tan and J. Yang. “Discrete-time approximation of stochastic optimal control with partial observation”. In: *arXiv preprint arXiv:2302.03329* (2023).

[WB95] G. Welch and G. Bishop. *An introduction to the Kalman filter*. University of North Carolina at Chapel Hill. 1995.

[WZZ20] H. Wang, T. Zariphopoulou, and X. Y. Zhou. “Reinforcement learning in continuous time and space: a stochastic control approach”. In: *J. Mach. Learn. Res.* 21 (2020), Paper No. 198, 34.