Actor-Critic learning for mean-field control in continuous time

Noufel Frikha, Maximilien Germain, Mathieu Laurière, Huyên Pham, Xuanye Song

Actor-Critic learning for mean-field control in continuous time

Noufel FRIKHA* Maximilien GERMAIN† Mathieu LAURIERE‡
Huyên PHAM§ Xuanye SONG¶

March 12, 2023

Abstract

We study policy gradient for mean-field control in continuous time in a reinforcement learning setting. By considering randomised policies with entropy regularisation, we derive a gradient expectation representation of the value function, which is amenable to actor-critic type algorithms, where the value functions and the policies are learnt alternately based on observation samples of the state and model-free estimation of the population state distribution, either by offline or online learning. In the linear-quadratic mean-field framework, we obtain an exact parametrisation of the actor and critic functions defined on the Wasserstein space. Finally, we illustrate the results of our algorithms with some numerical experiments on concrete examples.

Keywords: Mean-field control, reinforcement learning, policy gradient, linear-quadratic, actor-critic algorithms.

1 Introduction

Mean-field control (MFC in short), also called *McKean-Vlasov* (MKV in short) control problem is concerned with the study of large population models of interacting agents who are cooperative and act for collective welfare according to a center of decision (or social planner). It has attracted a growing interest over the last years with the emergence of mean-field game, and there is now a large literature on the theory and its various applications in economics/finance,

*CES, UMR 8174, Université Paris 1 Panthéon Sorbonne, Noufel.Frikha at univ-paris1.fr

†LPSM, Université Paris Cité, maximilien.germain at gmail.com

‡NYU Shanghai, mathieu.lauriere at nyu.edu

§LPSM, Université Paris Cité, pham at lpsm.paris; This author is supported by the BNP-PAR Chair "Futures of Quantitative Finance", and by FiME, Laboratoire de Finance des Marchés de l'Energie, and the "Finance and Sustainable Development" EDF - CACIB Chair

¶LPSM, Université Paris Cité, xsong at lpsm.paris;

population dynamics, social sciences and herd behavior. We refer to the seminal two-volume monograph [4]-[5] for a detailed treatment of the topic.

Mean-field control problems lead to infinite dimensional problems in the Wasserstein space of probability measures, and analytical solutions are rarely available. It is then crucial to design efficient numerical schemes for solving such problems, and in the past few years, several works have proposed numerical methods in a model-based setting based either on forward-backward SDE characterisation of MKV from Pontryagin maximum principle, or Master Bellman equation from dynamic programming, and often relying on suitable class of neural networks, see e.g. [7], [16], [17], [21], [27], [25].

The question of learning solutions to MFC in a model-free setting, i.e. when the environment (model coefficients) is unknown, has recently attracted attention, see [8, 9], [18], [1], and this is precisely the purpose of *Reinforcement learning* (RL): learn optimal control by trial and error, i.e., repeatedly try policy, observe the state, receive and evaluate the reward, and improve the policy. There are two main approaches in RL: (i) *Q-learning* based on dynamic programming, and (ii) *Policy gradient* based on parametrisation of policies, and a key feature in RL is the *exploration* of the unknown environment to broaden search space, which can be achieved via randomised policies. RL is a very active branch of machine learning and we refer to the second edition of the monograph [29] for an overview of this field.

Most algorithms in RL are limited to discrete-time frameworks for Markov decision processes (MDP) or mean-field MDP, and the study of RL in continuous time has been recently initiated in [30], [23], [24] for controlled diffusion processes. In line with these works, we provide in this paper a theoretical treatment of policy gradient methods for MFC in continuous time and state/action space by relying on stochastic calculus that has been recently developed for MKV equations. Our main theoretical result is to obtain a policy gradient representation for value function with randomised parametric policies and entropy regularisers for encouraging exploration. Based on this representation, we design model-free actor critic algorithms involving either the whole trajectories of the state (off-line learning), or the current and next state (online learning). In the mean-field context, a key issue is to handle the population state distribution, which is an input of the policy (actor) and value function (critic), and instead of assuming that we have at disposal a simulator of the state distribution as in [8], we estimate it in a model-free manner as in [1], which is more suitable for real-world applications. We next study the linear quadratic (LQ) case for which we derive explicit solutions, and this can be used for proposing an exact parametrisation of the critic and actor functions that is incorporated in stochastic gradient when updating the policies and value functions. The explicit solutions in the LQ setting are served as benchmarks for the numerical results of our algorithms in two examples.

The rest of the paper is organized as follows. In Section 2, we formulate the mean-field control problem in continuous-time with randomised policies and entropy regularisers, and state

the partial differential equation (PDE) characterisation of the value function in the Wasserstein space. We develop in Section 3 policy gradient methods by establishing a policy gradient representation, and its implication for actor-critic algorithms. Section 4 is devoted to the linear-quadratic setting, and we present in Section 5 numerical results on two examples to illustrate the accuracy of our algorithms. Finally, proofs of the policy gradient theorem are detailed in Appendix A, while the derivation of the explicit solution in the LQ case is shown in Appendix B.

Notations. The scalar product between two vectors x and y is denoted by $x \cdot y$, and $|\cdot|$ is the Euclidian norm. Given two matrices $M = (M_{ij})$ and $N = (N_{ij})$, we denote by $M : N = \mathrm{Tr}(M^\intercal N) = \sum_{i,j} M_{ij} N_{ij}$ its inner product, and by $|M|$ the Frobenius norm of M. Here $^\intercal$ is the transpose matrice operator. Let $\boldsymbol{M} = (\boldsymbol{M}_{i_1 i_2 i_3}) \in \mathbb{R}^{d_1 \times d_2 \times d_3}$ be a tensor of order 3. For $p = 1, 2, 3$, the p-mode product of $\boldsymbol{M}$ with a vector $b = (b_i) \in \mathbb{R}^{d_p}$, is denoted by $\boldsymbol{M} \bullet_p b$, and it is a tensor of order 2, i.e. a matrix defined elementwise as

$$(\boldsymbol{M} \bullet_1 b)_{i_2 i_3} = \sum_{i_1=1}^{d_1} M_{i_1 i_2 i_3} b_{i_1}, \ (\boldsymbol{M} \bullet_2 b)_{i_1 i_3} = \sum_{i_2=1}^{d_2} M_{i_1 i_2 i_3} b_{i_2}, \ (\boldsymbol{M} \bullet_3 b)_{i_1 i_2} = \sum_{i_3=1}^{d_3} M_{i_1 i_2 i_3} b_{i_3}.$$

The p-mode product of a 3-th order tensor $\boldsymbol{M} \in \mathbb{R}^{d_1 \times d_2 \times d_3}$ with a matrix $B = (B_{ij}) \in \mathbb{R}^{d_p \times d}$, also denoted by $\boldsymbol{M} \bullet_p B$, is a 3-th order tensor defined elementwise as

$$\begin{aligned}
&(\boldsymbol{M} \bullet_1 B)_{\ell i_2 i_3} = \sum_{i_1=1}^{d_1} M_{i_1 i_2 i_3} B_{i_1 \ell}, \quad (\boldsymbol{M} \bullet_2 B)_{i_1 \ell i_3} = \sum_{i_2=1}^{d_2} M_{i_1 i_2 i_3} B_{i_2 \ell} \\
&(\boldsymbol{M} \bullet_3 B)_{i_1 i_2 \ell} = \sum_{i_3=1}^{d_3} M_{i_1 i_2 i_3} B_{i_3 \ell}.
\end{aligned}$$

Finally, the tensor contraction (or partial trace) of a 3-th order tensor $\boldsymbol{M} \in \mathbb{R}^{d_1 \times d_2 \times d_3}$ whose dimensions d_p and d_q are equal is denoted as $\mathrm{Tr}_{p,q}\boldsymbol{M}$. This tensor contraction is a tensor of order 1, i.e. a vector, defined elementwise as

$$(\mathrm{Tr}_{1,2}\boldsymbol{M})_{i_3} = \sum_{\ell=1}^{d_1} M_{\ell \ell i_3}, \ (\mathrm{Tr}_{1,3}\boldsymbol{M})_{i_2} = \sum_{\ell=1}^{d_1} M_{\ell i_2 \ell}, \ (\mathrm{Tr}_{2,3}\boldsymbol{M})_{i_1} = \sum_{\ell=1}^{d_2} M_{i_1 \ell \ell}.$$

2 Exploratory formulation of mean-field control

Let us consider a mean-field control problem where the $\mathbb{R}^d$-valued controlled state process $X = X^\alpha$ is governed by the dynamics

$$\mathrm{d}X_s = b(X_s, \mathbb{P}_{X_s}, \alpha_s)\mathrm{d}s + \sigma(X_s, \mathbb{P}_{X_s}, \alpha_s)\mathrm{d}W_s, \quad s \geq 0, \tag{2.1}$$

with W a standard p-dimensional Brownian motion on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ equipped with the filtration $\mathbb{F} = (\mathcal{F}_t)_{t\geq 0}$ generated by W, and augmented with a σ-algebra $\mathcal{G}$ rich enough to support a uniformly distributed random variable independent of W. The control $\alpha = (\alpha_t)_t$ is an $\mathbb{F}$-progressively measurable process with α_t representing the action of the agent at time t, and valued in the action space $A \subset \mathbb{R}^q$. Here, $\mathbb{P}_{X_t}$ denotes the marginal law of X_t, $\mathcal{P}_2(\mathbb{R}^d)$ is the Wasserstein space of probability measures μ with a finite second order moment, i.e., $M_2(\mu) := \big(\int |x|^2 \mu(\mathrm{d}x)\big)^{\frac{1}{2}} < \infty$, equipped with the Wasserstein distance $\mathcal{W}_2$, and the coefficient b (resp. σ) is a measurable function from $\mathbb{R}^d \times \mathcal{P}_2(\mathbb{R}^d) \times A$ into $\mathbb{R}^d$ (resp. $\mathbb{R}^{d\times p}$).

Throughout the paper, we make the standard Lipschitz assumptions on the coefficients b and σ to ensure the existence and uniqueness of a strong solution to the stochastic differential equation (SDE in short) (2.1) given any initial condition ξ with law $\mu \in \mathcal{P}_2(\mathbb{R}^d)$.

The objective of a mean-field control problem on finite horizon $T < \infty$, is to minimize over the control α an expected total cost of the form

$$\mathbb{E}\Big[\int_0^T e^{-\beta s} f(X_s, \mathbb{P}_{X_s}, \alpha_s)\mathrm{d}s + e^{-\beta T} g(X_T, \mathbb{P}_{X_T})\Big].$$

Here f is a running cost function defined on $\mathbb{R}^d \times \mathcal{P}_2(\mathbb{R}^d) \times A$, while g is a terminal cost function on $\mathbb{R}^d \times \mathcal{P}_2(\mathbb{R}^d)$, and $\beta \in \mathbb{R}_+$ is a given discount factor. In a model-based setting, i.e., when the coefficients b, σ, and the functions f, g are known, the solution to MFC control problem can be characterised by a forward backward SDE arising from the maximum principle (see [3], or by a Master Bellman equation arsing from dynamic programming principle (see [26]). Moreover, the optimisation over $\mathbb{F}$-progressively measurable process α (open-loop control), or feedback (also called closed-loop) controls α, i.e., in the form $\alpha_t = \pi(t, X_t, \mathbb{P}_{X_t})$, $0 \leq t \leq T$, for some deterministic policy π, i.e., a measurable function $\pi : [0, T] \times \mathbb{R}^d \times \mathcal{P}_2(\mathbb{R}^d) \to A$, yields the same value function.

In a model-free reinforcement learning (RL) setting, when the coefficients are unknown, the agent can only rely on observation samples of state and reward in order to learn the optimal strategy. This is achieved by *trial and error* where the agent tries a policy, receive and evaluate the reward and then improve performance by repeating this procedure. A critical issue in reinforcement learning when the environment is unknown, is *exploration* in order to broaden search space, and a key and now common idea is to use *randomised* (or stochastic) policies: in a mean-field setting, this is defined by a probability transition kernel from $[0, T] \times \mathbb{R}^d \times \mathcal{P}_2(\mathbb{R}^d)$ into A, i.e., a measurable function $\pi : (t, x, \mu) \in [0, T] \times \mathbb{R}^d \times \mathcal{P}_2(\mathbb{R}^d) \mapsto \pi(.|t, x, \mu) \in \mathcal{P}(A)$, the set of probability measures on A. We then say that the process $\alpha = (\alpha_t)_t$ is a randomised feedback control generated from a stochastic policy π, denoted by $\alpha \sim \pi$, if at each time t, the action α_t is sampled from the probability distribution $\pi(.|t, X_t, \mathbb{P}_{X_t})$. Note that the sampling is drawn at each time from the σ-algebra $\mathcal{G}$ rich enough to support a uniformly distributed random variable independent of W. More precisely, it is defined as follows: given a probability transition kernel π, one can associate a measurable function $\phi_\pi : [0, T] \times \mathbb{R}^d \times \mathcal{P}_2(\mathbb{R}^d) \times [0, 1]$

$\to A$ such that the law of $\phi_\pi(t,x,\mu,U)$ is $\pi(.|t,x,\mu)$ where U is an uniform random variable on $[0,1]$. We would then naturally define the control process by $\alpha_t = \phi_\pi(t, X_t, \mathbb{P}_{X_t}, U_t)$, $0 \le t \le T$, for a collection of $\mathcal{G}$-measurable i.i.d. uniform random variables $(U_t)_t$, but this raises some measurability issues as $(t,\omega) \mapsto U_t(\omega)$ is not jointly measurable in the usual product space $([0,T]\times\Omega, \mathcal{B}_{[0,T]}\otimes\mathcal{G}, \mathrm{d}t\otimes\mathbb{P})$. To cope these issues, one can use the notion of Fubini extension, see [28]. We consider an atomless probability space $([0,T],\mathcal{T},\rho)$ extending the usual Lebesgue measure interval space $([0,T],\mathcal{B}_{[0,T]},\mathrm{d}t)$, and a rich Fubini extension $([0,T]\times\Omega, \mathcal{T}\boxtimes\mathcal{G}, \rho\boxtimes\mathbb{P})$ of the product space $([0,T]\times\Omega,\mathcal{T}\otimes\mathcal{G},\rho\otimes\mathbb{P})$. Then, from Theorem 1 in [28], there exists a $\mathcal{T}\boxtimes\mathcal{G}$-measurable map $\mathbb{U} : [0,T]\times\Omega \to [0,1]$ such that the random variables $U_t = \mathbb{U}(t,.)$ are essentially pairwise independent, and uniformly distributed on $[0,1]$. Denote by $\mathbb{F}$ the filtration generated by $(W,\mathbb{U})$, and consider the controlled process governed by

$$\mathrm{d}X_s = \; b(X_s,\mathbb{P}_{X_s},\alpha_s)\mathrm{d}s + \sigma(X_s,\mathbb{P}_{X_s},\alpha_s)\mathrm{d}W_s, \tag{2.2}$$

where $\alpha_t = \phi_\pi(t,X_t,\mathbb{P}_{X_t},U_t) \sim \pi(.|t,X_t,\mathbb{P}_{X_t})$, $0\le t\le T$, is $\mathbb{F}$-progressively measurable. Here, to alleviate notations, we write $\rho(\mathrm{d}t) \equiv \mathrm{d}t$.

Moreover, in order to encourage exploration of randomised policies, we shall substract entropy regularisers to the cost term, as adopted in the recent works by [30], [20], by considering the Shannon differential entropy defined as

$$\mathcal{E}(\pi(.|t,x,\mu)) \;\; := \;\; -\int_A \log p(t,x,\mu,a)\pi(\mathrm{d}a|t,x,\mu),$$

by assuming that $\pi(.|t,x,\mu)$ admits a density $p(t,x,\mu,.)$ with respect to some measure ν on A. The goal of the social planner is now to minimise over randomised policies π the cost

$$J(\pi) = \; \mathbb{E}_{\alpha\sim\pi}\Big[\int_0^T e^{-\beta s}\big[f(X_s,\mathbb{P}_{X_s},\alpha_s) - \lambda\mathcal{E}\big(\pi(.|s,X_s,\mathbb{P}_{X_s})\big)\big]\mathrm{d}s + e^{-\beta T}g(X_T,\mathbb{P}_{X_T})\Big], \tag{2.3}$$

where $\lambda \ge 0$ is a temperature parameter on exploration. Here, the notation in $\mathbb{E}_{\alpha\sim\pi}[.]$ means that the expectation operator is taken when the randomised feedback control α is generated from the stochastic policy π, and $X = X^\alpha$ is driven by the dynamics (2.2).

Let us now introduce the dynamic Markovian version of the above mean-field problem. Given a stochastic policy π, an initial time-state-distribution triple $(t,x,\mu) \in [0,T]\times\mathbb{R}^d \times \mathcal{P}_2(\mathbb{R}^d)$, and $\xi \in L^2(\mathcal{F}_t;\mathbb{R}^d)$ (the set of square-integrable $\mathcal{F}_t$-measurable random variables valued in $\mathbb{R}^d$) with distribution law μ ($\xi\sim\mu$), we consider the decoupled state processes $\{X_s^{t,\xi}, t\le s\le T\}$ and $\{X_s^{t,x,\xi}, t\le s\le T\}$ given by

$$\begin{aligned} X_s^{t,\xi} &= \xi + \int_t^s b(X_r^{t,\xi},\mathbb{P}_{X_r^{t,\xi}},\alpha_r)\mathrm{d}r + \int_t^s \sigma(X_r^{t,\xi},\mathbb{P}_{X_r^{t,\xi}},\alpha_r)\,\mathrm{d}W_r, \\ X_s^{t,x,\mu} &= x + \int_t^s b(X_r^{t,x,\mu},\mathbb{P}_{X_r^{t,\xi}},\alpha_r)\mathrm{d}r + \int_t^s \sigma(X_r^{t,x,\mu},\mathbb{P}_{X_r^{t,\xi}},\alpha_r)\,\mathrm{d}W_r, \quad t\le s\le T, \end{aligned} \tag{2.4}$$

where α is a randomised feedback control generated from π, i.e., α_s is sampled at each time s from $\pi(.|s, X_s^{t,x,\mu}, \mathbb{P}_{X_s^{t,\xi}})$ (here, to alleviate notations, we omit the dependence of $X^{t,\xi}$ and $X^{t,x,\mu}$ in $\alpha \sim \pi$). We make the standard Lipschitz regularity assumptions on the coefficients b and σ to ensure the existence and uniqueness of a strong solution to (2.4) given any initial condition t, ξ, x. By weak uniqueness, it follows that the law of the process $(X_s^{t,\xi})_{s\in[t,T]}$ given by the unique solution to the first SDE in (2.4) only depends upon ξ through its law μ. It thus makes sense to consider $(\mathbb{P}_{X_s^{t,\xi}})_{s\in[t,T]}$ as a function of μ without specifying the choice of the random variable ξ that has μ as distribution. In particular, for any $0 \le t \le s \le T$, the random variable $X_s^{t,x,\mu}$ depends on ξ only through its law μ. As a consequence, we can define the cost value function of the stochastic policy π as the function defined on $[0,T] \times \mathbb{R}^d \times \mathcal{P}_2(\mathbb{R}^d)$ by

$$V^\pi(t,x,\mu) = \mathbb{E}_{\alpha\sim\pi}\Big[\int_t^T e^{-\beta(s-t)}\big[f(X_s^{t,x,\mu}, \mathbb{P}_{X_s^{t,\xi}}, \alpha_s) - \lambda\mathcal{E}\big(\pi(.|s, X_s^{t,x,\mu}, \mathbb{P}_{X_s^{t,\xi}})\big)\big]\mathrm{d}s + e^{-\beta(T-t)}g(X_T^{t,x,\mu}, \mathbb{P}_{X_T^{t,\xi}})\Big]. \tag{2.5}$$

Since $X_s^{t,\xi,\xi} = X_s^{t,\xi}$ a.s., the initial cost value in (2.3) when starting from some initial random state $\xi \in L^2(\mathcal{G}; \mathbb{R}^d)$ with law μ is equal to $J(\pi) = \mathbb{E}_{\xi\sim\mu}[V^\pi(0,\xi,\mu)]$.

We complete this section by characterizing the cost value function V^π, for a given stochastic policy π, in terms of a linear parabolic partial differential equation (PDE) of mean-field type stated in the strip $[0,T] \times \mathbb{R}^d \times \mathcal{P}_2(\mathbb{R}^d)$. We first introduce the coefficients associated to the dynamics and the value function, given a stochastic policy π, namely

$$\begin{aligned} b_\pi(t,x,\mu) &= \int_A b(x,\mu,a)\pi(\mathrm{d}a|t,x,\mu), & \Sigma_\pi(t,x,\mu) &= \int_A (\sigma\sigma^\intercal)(x,\mu,a)\pi(\mathrm{d}a|t,x,\mu), \\ f_\pi(t,x,\mu) &= \int_A f(x,\mu,a)\pi(\mathrm{d}a|t,x,\mu), & E_\pi(t,x,\mu) &= -\int_A \log p(t,x,\mu,a)\pi(\mathrm{d}a|t,x,\mu), \end{aligned}$$

and let $\sigma_\pi := \Sigma_\pi^{1/2}$.

Before presenting the regularity assumptions, we introduce some notations regarding the Wasserstein derivative (also called L-derivative) of a real-valued smooth map U defined on $\mathcal{P}_2(\mathbb{R}^d)$. We follow the common practice of denoting by $\partial_\mu U(\mu)(v) \in \mathbb{R}^d$ the Wasserstein derivative of U with respect μ evaluated at $(\mu, v) \in \mathcal{P}_2(\mathbb{R}^d) \times \mathbb{R}^d$. Its ith coordinate is denoted by $\partial_\mu^i U(\mu)(v)$. We will also work with higher order derivatives. For a positive integer n, a multi-index λ of $\{1,\cdots,d\}$, a n-tuple of multi-indices $\boldsymbol{\gamma} = (\gamma_1,\cdots,\gamma_n)$ of $\{1,\cdots,d\}$ and $\boldsymbol{v} = (v_1,\cdots,v_n) \in (\mathbb{R}^d)^n$, we denote by $\partial_\mu^\lambda U(\mu)(\boldsymbol{v})$ the derivative $\partial_\mu^{\lambda_n}[\cdots[\partial_\mu^{\lambda_1}U(\mu)](v_1)\cdots](v_n)$. If $\boldsymbol{v} \mapsto \partial_\mu^\lambda U(\mu)(\boldsymbol{v})$ is smooth, we write $\partial_{\boldsymbol{v}}^{\boldsymbol{\gamma}}\partial_\mu^\lambda U(\mu)(\boldsymbol{v})$ for the derivative $\partial_{v_n}^{\gamma_n}\cdots\partial_{v_1}^{\gamma_1}\partial_\mu^\lambda U(\mu)(\boldsymbol{v})$.

We will often deal with maps that depend on additional time and space variables. In particular, we will work with the two spaces $\mathcal{C}^{2,2}(\mathbb{R}^d \times \mathcal{P}_2(\mathbb{R}^d))$ and $\mathcal{C}^{1,2,2}([0,T] \times \mathbb{R}^d \times \mathcal{P}_2(\mathbb{R}^d))$ and refer the reader to [4] Chapter 5 for more details.

Having these notations at hand, we make the following regularity assumptions on the coefficients b_π, σ_π, the cost functions f_π, g and the Shannon differential entropy E_π. Below, $\pi : [0,T] \times \mathbb{R}^d \times \mathcal{P}_2(\mathbb{R}^d) \to \mathcal{P}(A)$ is a fixed stochastic policy.

Assumption 2.1 *(i) For any $h \in \{b_\pi^i, \sigma_\pi^{i,j}, i = 1, \cdots, d, j = 1, \cdots, p\}$, the following derivatives*

$$\partial_x h(t,x,\mu),\ \partial_x^2 h(t,x,\mu),\ \partial_\mu h(t,x,\mu)(v),\ \partial_v[\partial_\mu h(t,x,\mu)](v),$$

exist for any $(t,x,v,\mu) \in [0,T] \times (\mathbb{R}^d)^2 \times \mathcal{P}_2(\mathbb{R}^d)$, are bounded and locally Lipschitz continuous with respect to x, μ, v uniformly in $t \in [0,T]$. Moreover, $h(t,.)$ is at most of linear growth, uniformly in $t \in [0,T]$, namely, there exists $C < \infty$ such that for all t, x, μ

$$|h(t,x,\mu)| \le C(1 + |x| + M_2(\mu)).$$

(ii) For any $t \in [0,T]$, $f_\pi(t,.)$, $E_\pi(t,.)$, $g \in \mathcal{C}^{2,2}(\mathbb{R}^d \times \mathcal{P}_2(\mathbb{R}^d))$.

(iii) There exists some constant $C < \infty$, such that for any $(t,x,v,\mu) \in [0,T] \times (\mathbb{R}^d)^2 \times \mathcal{P}_2(\mathbb{R}^d)$,

$$|f_\pi(t,x,\mu)| + |E_\pi(t,x,\mu)| + |g(x,\mu)| \le C(1 + |x|^2 + M_2(\mu)^q),$$

$$|\partial_x f_\pi(t,x,\mu)| + |\partial_x E_\pi(t,x,\mu)| + |\partial_x g(x,\mu)| \le C(1 + |x| + M_2(\mu)^q),$$

$$|\partial_\mu f_\pi(t,x,\mu)(v)| + |\partial_\mu E_\pi(t,x,\mu)(v)| + |\partial_\mu g(x,\mu)(v)| \le C(1 + |x| + |v| + M_2(\mu)^q),$$

$$\begin{aligned}|\partial_v[\partial_\mu f_\pi(t,x,\mu)](v)| + |\partial_x^2 f_\pi(t,x,\mu)| + |\partial_v[\partial_\mu E_\pi(t,x,\mu)](v)| + |\partial_x^2 E_\pi(t,x,\mu)| \\ + |\partial_v[\partial_\mu g(x,\mu)](v)| + |\partial_x^2 g(x,\mu)| \le C(1 + M_2(\mu)^q),\end{aligned}$$

for some $q \ge 0$.

Remark 2.1 *It is readily seen from the integral form of b_π, Σ_π, f_π, E_π that if for any $a \in A$, the functions $(x,\mu) \mapsto b(x,\mu,a), \sigma(x,\mu,a)$, $f(x,\mu,a)$ and the density $(x,\mu) \mapsto p(t,x,\mu,a)$ of the probability measure $\pi(\mathrm{d}a|t,x,\mu)$ are smooth with derivatives satisfying some adequate estimates then Assumption 2.1 is satisfied. In particular, this will be the case when the coefficients b, σ are linear functions and f together with g are quadratic functions of the variables of x, $\int_{\mathbb{R}^d} z\mu(\mathrm{d}z)$ and a and if p is a Gaussian density with a smooth mean and a time-dependent covariance-matrix as in the linear quadratic framework, see Section 4.*

We now have the following PDE characterisation of the cost value function V^π.

Proposition 2.1 *Under Assumption 2.1, the function V^π defined by (2.5) belongs to $C^{1,2,2}([0,T] \times \mathbb{R}^d \times \mathcal{P}_2(\mathbb{R}^d))$ and satisfies the following linear parabolic PDE*

$$\mathcal{L}_\pi[V^\pi](t,x,\mu) + (f_\pi - \lambda E_\pi)(t,x,\mu) = 0, \quad (t,x,\mu) \in [0,T) \times \mathbb{R}^d \times \mathcal{P}_2(\mathbb{R}^d), \tag{2.6}$$

with the terminal condition $V^\pi(T,x,\mu) = g(x,\mu)$, where $\mathcal{L}_\pi$ is the operator defined by

$$\begin{aligned}\mathcal{L}_\pi[\varphi](t,x,\mu) &= -\beta\varphi(t,x,\mu) + \partial_t\varphi(t,x,\mu) + b_\pi(t,x,\mu)\cdot D_x\varphi(t,x,\mu) + \frac{1}{2}\Sigma_\pi(t,x,\mu) : D_x^2\varphi(t,x,\mu)\\ &\quad + \mathbb{E}_{\xi\sim\mu}\Big[b_\pi(t,\xi,\mu)\cdot\partial_\mu\varphi(t,x,\mu)(\xi) + \frac{1}{2}\Sigma_\pi(t,\xi,\mu) : \partial_\upsilon\partial_\mu\varphi(t,x,\mu)(\xi)\Big].\end{aligned}$$

Remark 2.2 *In particular, the above result indicates that provided the coefficients b_π, Σ_π, the functions f_π, E_π and the terminal condition g are smooth with derivatives satisfying some appropriate estimates, the solution V^π to the Kolmogorov PDE* (2.6) *is smooth. In this sense, it preserves the regularity of the terminal condition.*

However, one can weaken the regularity assumption on the terminal condition (and actually of the coefficients themselves) by benefiting from the smoothness of the underlying fundamental solution (or the transition density of the associated stochastic process) under some additional non-degeneracy assumption. We refer e.g. to [13], [12], [11] in the uniformly elliptic diffusion setting and to [15] in the case of non-degenerate stable driven SDE.

Proof. See Appendix A.1 □

3 Policy gradient method

We now consider a parametric family of randomised policies π_θ, with densities p_θ, $\theta \in \Theta$, Θ being a non-empty open subset of $\mathbb{R}^D$, for some positive integer D, and denote by $\mathrm{J}(\theta) = J(\pi_\theta)$ the associated cost function, viewed as a function of the parameters θ, recalling that J is defined by (2.3). The principle of policy gradient method is to minimize over θ the function $\mathrm{J}(\theta)$ by stochastic gradient descent algorithm. In our RL setting, we aim to derive a probabilistic representation of the gradient function $\nabla_\theta\mathrm{J}(\theta)$ that does not involve model coefficients b, σ, but only observation samples of state X_t, state distribution $\mathbb{P}_{X_t}$, and rewards $f_t := f(X_t, \mathbb{P}_{X_t}, \alpha_t)$, $g_T := g(X_T, \mathbb{P}_{X_T})$ when taking decision $\alpha \sim \pi_\theta$.

3.1 Policy gradient representation

We make the following assumptions on the parametric family of randomised policy and coefficients.

Assumption 3.1 *(i) For any $h \in \{b^i_{\pi_\theta}, \sigma^{i,j}_{\pi_\theta}, f_{\pi_\theta}, E_{\pi_\theta}, g, i = 1,\cdots,d, j = 1,\cdots,p\}$, any multi-indices α, β, λ of $\{1,\cdots,d\}$ such that $0 \le |\alpha| \le 2$, $0 \le |\beta| \le 1$, λ being of length n, $0 \le n \le 2$, any n-tuple of multi-indices $\boldsymbol{\gamma} = (\gamma_1,\cdots,\gamma_n)$ with $0 \le |\gamma_1| + \cdots + |\gamma_n| \le 2$, denoting by $h_\theta(t,x,\mu)$ the value of h at (θ,t,x,μ), the following derivatives*

$$\begin{aligned}&\partial_\theta^\beta\partial_x^\alpha\partial_{\boldsymbol{v}}^{\boldsymbol{\gamma}}\partial_\mu^\lambda h_\theta(t,x,\mu)(\boldsymbol{v}),\ \partial_x^\alpha\partial_\theta^\beta\partial_{\boldsymbol{v}}^{\boldsymbol{\gamma}}\partial_\mu^\lambda h_\theta(t,x,\mu)(\boldsymbol{v}),\ \partial_x^\alpha\partial_{\boldsymbol{v}}^{\boldsymbol{\gamma}}\partial_\theta^\beta\partial_\mu^\lambda h_\theta(t,x,\mu)(\boldsymbol{v}),\ \partial_x^\alpha\partial_{\boldsymbol{v}}^{\boldsymbol{\gamma}}\partial_\mu^\lambda\partial_\theta^\beta h_\theta(t,x,\mu)(\boldsymbol{v}),\\ &\partial_{\boldsymbol{v}}^{\boldsymbol{\gamma}}\partial_x^\alpha\partial_\mu^\lambda\partial_\theta^\beta h_\theta(t,x,\mu)(\boldsymbol{v}),\partial_{\boldsymbol{v}}^{\boldsymbol{\gamma}}\partial_\mu^\lambda\partial_x^\alpha\partial_\theta^\beta h_\theta(t,x,\mu)(\boldsymbol{v}),\partial_{\boldsymbol{v}}^{\boldsymbol{\gamma}}\partial_\mu^\lambda\partial_\theta^\beta\partial_x^\alpha h_\theta(t,x,\mu)(\boldsymbol{v}),\partial_{\boldsymbol{v}}^{\boldsymbol{\gamma}}\partial_\theta^\beta\partial_\mu^\lambda\partial_x^\alpha h_\theta(t,x,\mu)(\boldsymbol{v}),\\ &\partial_\theta^\beta\partial_{\boldsymbol{v}}^{\boldsymbol{\gamma}}\partial_\mu^\lambda\partial_x^\alpha h_\theta(t,x,\mu)(\boldsymbol{v}),\partial_{\boldsymbol{v}}^{\boldsymbol{\gamma}}\partial_\theta^\beta\partial_x^\alpha\partial_\mu^\lambda h_\theta(t,x,\mu)(\boldsymbol{v}),\partial_{\boldsymbol{v}}^{\boldsymbol{\gamma}}\partial_x^\alpha\partial_\theta^\beta\partial_\mu^\lambda h_\theta(t,x,\mu)(\boldsymbol{v}),\partial_\theta^\beta\partial_x^\alpha\partial_{\boldsymbol{v}}^{\boldsymbol{\gamma}}\partial_\mu^\lambda h_\theta(t,x,\mu)(\boldsymbol{v}),\end{aligned}$$

exist for any $(t,\theta,x,\boldsymbol{v},\mu) \in [0,T]\times\Theta\times(\mathbb{R}^d)^{n+1}\times\mathcal{P}_2(\mathbb{R}^d)$ *and are locally Lipschitz continuous with respect to* $\theta,x,\mu,\boldsymbol{v}$ *uniformly in* $t\in[0,T]$[a]. *Moreover, if* $h = b^i_{\pi_\theta}$ *or* $\sigma^{i,j}_{\pi_\theta}$, *the aforementioned derivatives of order greater or equal to one are bounded.*

(ii) The estimates of Assumption 2.1(iii) are satisfied for the family of policies $\{\pi_\theta,\theta\in\Theta\}$, *locally uniformly in* θ, *i.e. for any* $\theta\in\mathcal{K}$, $\mathcal{K}$ *being any compact subset of* Θ. *Additionally, there exists some constant* $C<\infty$, *such that for any* $h\in\{f_{\pi_\theta},E_{\pi_\theta}\}$, *any* $(t,\mu,x)\in[0,T]\times\mathcal{P}_2(\mathbb{R}^d)\times\mathbb{R}^d$, *any* $\boldsymbol{v}=(v_1,v_2)\in(\mathbb{R}^d)^2$, *any* $\theta\in\mathcal{K}$, $\mathcal{K}$ *being any compact subset of* Θ, *any multi-index* λ, $|\lambda|=2$, *any multi-index* $\lambda=(\lambda_1,\lambda_2)$ *of* $\{1,\cdots,d\}$, *any couple of multi-indices* $\boldsymbol{\gamma}=(\gamma_1,\gamma_2)$

$$|\partial_\theta h_\theta(t,x,\mu)|\le C(1+|x|^2+M_2(\mu)^q),$$

$$|\partial_\theta\partial_x h_\theta(t,x,\mu)|+|\partial_x g(x,\mu)|\le C(1+|x|+M_2(\mu)^q),$$

$$|\partial_\theta\partial_\mu h_\theta(t,x,\mu)(v_1)|+|\partial_\mu\partial_x h_\theta(t,x,\mu)(v_1)|+|\partial_\mu\partial_x g(x,\mu)(v_1)|\le C(1+|x|+|v_1|+M_2(\mu)^q),$$

$$\begin{aligned}&|\partial_\theta\partial_{v_1}\partial_\mu h_\theta(t,x,\mu)(v_1)|+|\partial_\theta\partial_x^2 h_\theta(t,x,\mu)|+|\partial_\mu\partial_x^2 h_\theta(t,x,\mu)(v_1)|+|\partial_{\boldsymbol{v}}^{\boldsymbol{\gamma}}\partial_\mu^{\lambda}h_\theta(t,x,\mu)(\boldsymbol{v})|\\&+|\partial_{v_1}\partial_\mu g(x,\mu)(v_1)|+|\partial_x^2 g(x,\mu)|+|\partial_{\boldsymbol{v}}^{\boldsymbol{\gamma}}\partial_\mu^{\lambda}g(t,x,\mu)(\boldsymbol{v})|\le C(1+M_2(\mu)^q),\end{aligned}$$

for some $q\ge 0$.

As shown in Appendix A.2, Assumption 3.1 guarantees that the derivatives $(t,\theta,x,\mu,v)\mapsto\partial_\theta\partial_t\mathrm{V}_\theta(t,x,\mu)$, $\partial_\theta\mathrm{V}_\theta(t,x,\mu)$, $\partial_\theta\partial_x\mathrm{V}_\theta(t,x,\mu)$, $\partial_\theta\partial_\mu\mathrm{V}_\theta(t,x,\mu)(v)$, $\partial_\theta\partial_x^2\mathrm{V}_\theta(t,x,\mu)$, $\partial_\theta\partial_v\partial_\mu\mathrm{V}_\theta(t,x,\mu)(v)$, where $\mathrm{V}_\theta(t,x,\mu):=V^{\pi_\theta}(t,x,\mu)$ defined by (2.5) with $\pi=\pi_\theta$, exist, are continuous and satisfy suitable growth conditions.

We then let $\nabla_\theta\mathrm{J}(\theta)=\mathbb{E}[\mathrm{G}_\theta(0,\xi,\mu)]$ where $\mathrm{G}_\theta(t,x,\mu):=\nabla_\theta\mathrm{V}_\theta(t,x,\mu)$. The main result of this section provides a probabilistic representation of the gradient function G_θ.

Theorem 3.1 *Suppose that Assumption 3.1 holds. Assume moreover that for any* t,x,μ,a, *the map* $\Theta\ni\theta\mapsto p_\theta(t,x,\mu,a)$ *is differentiable with a derivative satisfying the following estimates: for some constant* $C<\infty$ *and some* $q\ge 0$, *for any* $(t,x,\mu)\in[0,T]\times\mathbb{R}^d\times\mathcal{P}_2(\mathbb{R}^d)$ *and any compact subset* $\mathcal{K}\subset\Theta$.

$$\begin{aligned}\int_A\sup_{\theta\in\mathcal{K}}\{|\nabla_\theta p_\theta(t,x,\mu,a)|(|b(x,\mu,a)|+|(\sigma\sigma)^{\intercal}(x,\mu,a)|\\+|f(x,\mu,a)|+|\log(p_\theta(t,x,\mu,a))|)\}\nu(\mathrm{d}a)<\infty,\end{aligned}\tag{3.1}$$

[a]Hence, according to Clairaut's theorem, these partial derivatives are equal.

and

$$\int_A |\nabla_\theta \log(p_\theta(t,x,\mu,a))|^2 |\sigma(x,\mu,a)|^2 \, p_\theta(t,x,\mu,a)\, \nu(\mathrm{d}a) \leq C(1+|x|^q + M_2(\mu)^q). \tag{3.2}$$

Then, it holds

$$\begin{aligned} \mathrm{G}_\theta(t,x,\mu) = \; & \mathbb{E}_{\alpha\sim\pi_\theta}\Big[\int_t^T e^{-\beta(s-t)} \nabla_\theta \log p_\theta(s, X_s^{t,x,\mu}, \mathbb{P}_{X_s^{t,\xi}}, \alpha_s) \Big\{ \mathrm{dV}_\theta(s, X_s^{t,x,\mu}, \mathbb{P}_{X_s^{t,\xi}}) \\ & + \big[f(X_s^{t,x,\mu}, \mathbb{P}_{X_s^{t,\xi}}, \alpha_s) + \lambda \log p_\theta(s, X_s^{t,x,\mu}, \mathbb{P}_{X_s^{t,\xi}}, \alpha_s) - \beta \mathrm{V}_\theta(s, X_s^{t,x,\mu}, \mathbb{P}_{X_s^{t,\xi}}) \big] \mathrm{d}s \Big\} \\ & + \int_t^T e^{-\beta(s-t)} \mathcal{H}_\theta[\mathrm{V}_\theta](s, X_s^{t,x,\mu}, \mathbb{P}_{X_s^{t,\xi}}) \mathrm{d}s \Big], \end{aligned} \tag{3.3}$$

for any $(t,x,\mu,\theta) \in [0,T] \times \mathbb{R}^d \times \mathcal{P}_2(\mathbb{R}^d) \times \Theta$ *and* $\xi \sim \mu$, *where* $\mathcal{H}_\theta$ *is the operator defined by*

$$\begin{aligned} \mathcal{H}_\theta[\varphi](t,x,\mu) = \; & \mathbb{E}_{\xi\sim\mu}\big[\nabla_\theta b_\theta(t,\xi,\mu)^\intercal \partial_\mu \varphi(t,x,\mu)(\xi) \\ & + \frac{1}{2} \mathrm{tr}_{1,2}\big(\nabla_\theta \Sigma_\theta(t,\xi,\mu) \bullet_1 \partial_v \partial_\mu \varphi(t,x,\mu)(\xi) \big) \big], \end{aligned} \tag{3.4}$$

and we set $b_\theta(t,x,\mu) = \int_A b(x,\mu,a)\, \pi_\theta(\mathrm{d}a|t,x,\mu)$, $\Sigma_\theta(t,x,\mu) = \int_A (\sigma\sigma^\intercal)(x,\mu,a)\, \pi_\theta(\mathrm{d}a|t,x,\mu)$. *Here* $\nabla_\theta \Sigma_\theta = (\frac{\partial \Sigma_\theta^{ij}}{\partial \theta_k})_{i,j,k} \in \mathbb{R}^{d\times d\times D}$ *is a tensor of order* 3, *and we used the product tensor notations* $\bullet_1$ *recalled in the introduction.*

Remark 3.1 (On the martingale property of the policy gradient) *The representation in Theorem 3.1 also means that the process*

$$\begin{aligned} & \Big\{ e^{-\beta(s-t)} \mathrm{G}_\theta(s, X_s^{t,x,\mu}, \mathbb{P}_{X_s^{t,\xi}}) + \int_t^s e^{-\beta(r-t)} \nabla_\theta \log p_\theta(r, X_r^{t,x,\mu}, \mathbb{P}_{X_r^{t,\xi}}, \alpha_r) \Big\{ \mathrm{dV}_\theta(r, X_r^{t,x,\mu}, \mathbb{P}_{X_r^{t,\xi}}) \\ & + \big[f(X_r^{t,x,\mu}, \mathbb{P}_{X_r^{t,\xi}}, \alpha_r) + \lambda \log p_\theta(r, X_r^{t,x,\mu}, \mathbb{P}_{X_r^{t,\xi}}, \alpha_r) - \beta \mathrm{V}_\theta(r, X_r^{t,x,\mu}, \mathbb{P}_{X_r^{t,\xi}}) \big] \mathrm{d}r \Big\} \\ & + \int_t^s e^{-\beta(r-t)} \mathcal{H}_\theta[\mathrm{V}_\theta](r, X_r^{t,x,\mu}, \mathbb{P}_{X_r^{t,\xi}}) \mathrm{d}r, \; t \leq s \leq T \Big\} \end{aligned}$$

is a martingale, for any given $\alpha \sim \pi_\theta$.

Proof. See Appendix A.3 □

In the next section, we show how the probabilistic representation formula of the gradient function G_θ provided by Theorem 3.1 can be used to design two actor-critic algorithms for learning optimal cost function and randomised policy by relying on samples of the actions, states and state distributions.

3.2 Actor-critic Algorithms

Actor-critic (AC) methods combine policy gradient (PG) and performance evaluation (PE). Compared to most existing works on RL for mean-field problems, mainly based on Q-learning (see e.g. [8], [14], [18]) we do not assume that the agent (the social planner) has at disposal a simulator for the state distribution, but instead will estimate the distribution of the population from the observation of the state of the representative player and by updating the distribution along repeated episodes. More precisely, for each episode $i = 1, 2, \ldots, N$, from the observation of the state $X^i_{t_k}$ of a representative player i at time t_k, we update the state distribution according to

$$\mu^i_{t_k} = (1 - \rho^i_S)\mu^{i-1}_{t_k} + \rho^i_S \delta_{X^i_{t_k}}, \tag{3.5}$$

where $(\rho^i_S)_i$ is a sequence of learning parameters in $(0, 1)$, e.g. $\rho^i_S = 1/i$. It is expected from the propagation of chaos, that when the number of episodes N goes to infinity, $\mu^N_{t_k}$ converge to the limiting distribution $\mathbb{P}_{X_{t_k}}$ of the population. Notice that a similar estimation procedure was recently proposed in [1] in the context of a MFC control problem in discrete time with finite state and action spaces over an infinite horizon.

In addition to the family of randomised policies $(t, x, \mu) \mapsto \pi_\theta(\mathrm{d}a|t, x, \mu) = p_\theta(t, x, \mu)\nu(\mathrm{d}a)$, with parameter θ, we are given a family of functions $(t, x, \mu) \mapsto \mathrm{J}^\eta(t, x, \mu)$ on $[0, T]\times\mathbb{R}^d\times\mathcal{P}_2(\mathbb{R}^d)$, with parameter η, aiming to approximate the optimal cost value function. AC algorithm is then updating alternately the two parameters to find the optimal pair (θ^*, η^*), hence determining the approximate optimal randomised policy and the associated cost value function. On the one hand, the loss function in the PE step for learning J^η, for fixed policy π_θ, is based on the martingale formulation of the process

$$\Big\{e^{-\beta t}\mathrm{J}^\eta(t, X^{x,\mu}_t, \mathbb{P}_{X^\xi_t}) + \int_0^t e^{-\beta r}\big[f(X^{x,\mu}_r, \mathbb{P}_{X^\xi_r}, \alpha_r) + \lambda \log p_\theta(r, X^{x,\mu}_r, \mathbb{P}_{X^\xi_r}, \alpha_r)\big]\mathrm{d}r,\ 0 \le t \le T\Big\},$$

and on the other hand, the objective (here a cost) function in the PG step for learning π_θ, for fixed J^η, is based on the martingale formulation of the process

$$\begin{aligned}\Big\{e^{-\beta t}\mathrm{G}_\theta(t, X^{x,\mu}_t, \mathbb{P}_{X^\xi_t}) + \int_0^t e^{-\beta r}\nabla_\theta \log p_\theta(r, X^{x,\mu}_r, \mathbb{P}_{X^\xi_r}, \alpha_r)\Big[\mathrm{d}\mathrm{J}^\eta(r, X^{x,\mu}_r, \mathbb{P}_{X^\xi_r}) \\ + \big(f(X^{x,\mu}_r, \mathbb{P}_{X^\xi_r}, \alpha_r) + \lambda \log p_\theta(r, X^{x,\mu}_r, \mathbb{P}_{X^\xi_r}, \alpha_r) - \beta \mathrm{J}^\eta(r, X^{x,\mu}_r, \mathbb{P}_{X^\xi_r})\big)\mathrm{d}r\Big] \\ + \mathcal{H}_\theta[\mathrm{J}^\eta](r, X^{x,\mu}_r, \mathbb{P}_{X^\xi_r})\mathrm{d}r,\ 0 \le t \le T\Big\}.\end{aligned} \tag{3.6}$$

Here, we denote $X^{x,\mu} = X^{0,x,\mu}$ (resp. $X^\xi = X^{0,\xi}$) when the initial time of the flow is $t = 0$. We emphasise that these loss functions are minimised by training samples of the state trajectories $X^{x_0,\xi}_t$, actions $\alpha \sim \pi_\theta$, estimation μ_t of $\mathbb{P}_{X^\xi_t}$ according to (3.5), and observation of the associated running and terminal costs.

We first develop AC algorithms in the offline setting where all state trajectories are sampled. In this case, given θ, the proposed loss function for the PE step is

$$\begin{aligned} L^{PE}(\eta) &= \mathbb{E}_{\alpha\sim\pi_\theta}\Big[\int_0^T \Big|e^{-\beta(T-t)}g(X_T,\mathbb{P}_{X_T}) \\ &\quad + \int_t^T e^{-\beta(r-t)}\big[f(X_r,\mathbb{P}_{X_r},\alpha_r)+\lambda\log p_\theta(r,X_r,\mathbb{P}_{X_r},\alpha_r)\big]\mathrm{d}r - \mathrm{J}^\eta(t,X_t,\mathbb{P}_{X_t})\Big|^2\mathrm{d}t\Big], \end{aligned}$$

which leads, after time discretisation of $[0,T]$ on the grid $\{t_k = k\Delta t, k=0,\ldots,n\}$, and by applying stochastic gradient descent (SGD) with learning rate ρ_E, to the following update rule:

$$\begin{aligned} \eta \leftarrow \eta + \rho_E \sum_{k=0}^{n-1}\Big(e^{-\beta(n-k)\Delta t}g_{t_n} + \sum_{\ell=k}^{n-1} e^{-\beta(\ell-k)\Delta t}\big[f_{t_\ell} + \lambda\log p_\theta(t_\ell,X_{t_\ell},\mu_{t_\ell},\alpha_{t_\ell})\big]\Delta t \\ - \mathrm{J}^\eta(t_k,X_{t_k},\mu_{t_k})\Big)\nabla_\eta \mathrm{J}^\eta(t_k,X_{t_k},\mu_{t_k})\Delta t, \end{aligned}$$

where we set $f_{t_\ell} = f(X_{t_\ell},\mathbb{P}_{X_{t_\ell}},\alpha_{t_\ell})$, as the output running cost at time t_ℓ, for an input state X_{t_l}, action α_{t_ℓ}, $\ell = 0,\ldots,n-1$, and $g_T = g(X_T,\mathbb{P}_{X_T})$ the terminal cost for an input X_T. Given η, the learning in the PG step relies on the gradient representation (3.3), and (after time discretisation) leads to the update rule

$$\begin{aligned} \theta &\leftarrow \theta - \rho_G\hat{G}_\theta, \\ \text{with}\quad \hat{G}_\theta &= \sum_{k=0}^{n-1} e^{-\beta t_k}\nabla_\theta\log p_\theta(t_k,X_{t_k},\mu_{t_k},\alpha_{t_k})\Big[\mathrm{J}^\eta(t_{k+1},X_{t_{k+1}},\mu_{t_{k+1}}) - \mathrm{J}^\eta(t_k,X_{t_k},\mu_{t_k}) \\ &\quad + \big(f_{t_k} + \lambda\log p_\theta(t_k,X_{t_k},\mu_{t_k},\alpha_{t_k}) - \beta\mathrm{J}^\eta(t_k,X_{t_k},\mu_{t_k})\big)\Delta t\Big] + \mathcal{H}_\theta[\mathrm{J}^\eta](t_k,X_{t_k},\mu_{t_k})\Delta t. \end{aligned}$$

The pseudo-code is described in Algorithm 1.

Algorithm 1: Offline actor-critic mean-field algorithm

Input data: Number of episodes N, number of mesh time-grid n ($\leftrightarrow$ time step $\Delta t = T/n$), learning rates ρ_S^i, ρ_E^i, ρ_G^i for the state distribution, PE and PG estimation, and function of the number of episodes i. Parameter λ for entropy regularisation.
Functional forms J^η of cost value function, p_θ of density policies.
Initialisation: μ_{t_k}: state distribution on $\mathbb{R}^d$, for $k = 0, \ldots, N$, parameters η, θ.
for *each episode* $i = 1, \ldots, N$ **do**
  Initialise state $X_0 \sim \mu_0$
  for $k = 0, \ldots, n-1$ **do**
    Update state distribution: $\mu_{t_k} \leftarrow (1 - \rho_S^i)\mu_{t_k} + \rho_S^i \delta_{X_{t_k}}$
    Generate action $\alpha_{t_k} \sim \pi_\theta(.|t_k, X_{t_k}, \mu_{t_k})$
    Observe (e.g. by environment simulator) state $X_{t_{k+1}}$ and cost f_{t_k}
    If $k = n-1$, update terminal state distribution: $\mu_{t_n} \leftarrow (1 - \rho_S)\mu_{t_n} + \rho_S \delta_{X_{t_n}}$,
     and observe terminal cost g_{t_n}
    $k \leftarrow k + 1$
  end
  Compute

$$\begin{aligned}
\Delta_\eta &= \sum_{k=0}^{n-1} \Big(e^{-\beta(n-k)\Delta t} g_{t_n} + \sum_{\ell=k}^{n-1} e^{-\beta(\ell-k)\Delta t} \big[f_{t_\ell} + \lambda \log p_\theta(t_\ell, X_{t_\ell}, \mu_{t_\ell}, \alpha_{t_\ell}) \big] \Delta t \\
&\qquad - \mathrm{J}^\eta(t_k, X_{t_k}, \mu_{t_k}) \Big) \nabla_\eta \mathrm{J}^\eta(t_k, X_{t_k}, \mu_{t_k}) \Delta t \\
\hat{G}_\theta &= \sum_{k=0}^{n-1} e^{-\beta t_k} \nabla_\theta \log p_\theta(t_k, X_{t_k}, \mu_{t_k}, \alpha_{t_k}) \Big[\mathrm{J}^\eta(t_{k+1}, X_{t_{k+1}}, \mu_{t_{k+1}}) - \mathrm{J}^\eta(t_k, X_{t_k}, \mu_{t_k}) \\
&\qquad + \big(f_{t_k} + \lambda \log p_\theta(t_k, X_{t_k}, \mu_{t_k}, \alpha_{t_k}) - \beta \mathrm{J}^\eta(t_k, X_{t_k}, \mu_{t_k}) \big) \Delta t \Big] + \mathcal{H}_\theta[\mathrm{J}^\eta](t_k, X_{t_k}, \mu_{t_k}) \Delta t.
\end{aligned}$$

  Critic Update: $\eta \leftarrow \eta + \rho_E^i \Delta_\eta$; Actor Update: $\theta \leftarrow \theta - \rho_G^i \hat{G}_\theta$
end
Return: J^η, π_θ

We next develop AC algorithm for online setting where only past sample trajectory is available, and so the parameters (θ, η) are updated in real-time incrementally. In this case, given a policy π_θ, we consider at each time step t_k, $k = 0, \ldots, n-1$, a loss function for PE given by

$$\begin{aligned}
L_{t_k}^{PE}(\eta) &= \mathbb{E}_{\alpha \sim \pi_\theta} \Big[\Big| \mathrm{J}^\eta(t_{k+1}, X_{t_{k+1}}, \mathbb{P}_{X_{t_{k+1}}}) - \mathrm{J}^\eta(t_k, X_{t_k}, \mathbb{P}_{X_{t_k}}) \\
&\qquad + \big(f(X_{t_k}, \mathbb{P}_{X_{t_k}}, \alpha_{t_k}) + \lambda \log p_\theta(t_k, X_{t_k}, \mathbb{P}_{X_{t_k}}, \alpha_{t_k}) - \beta \mathrm{J}^\eta(t_k, X_{t_k}, \mathbb{P}_{X_{t_k}}) \big) \Delta t \Big|^2 \Big].
\end{aligned}$$

Concerning PG, we note that when θ is an optimal parameter, we should have $\mathrm{G}_\theta = 0$.

Therefore, from the martingale condition in (3.6), this suggests to find θ such that at any time t_k, $k = 0, \ldots, n-1$

$$\mathbb{E}_{\alpha\sim\pi_\theta}\Big\{\nabla_\theta \log p_\theta(t_k, X_{t_k}, \mathbb{P}_{X_{t_k}}, \alpha_{t_k})\big[\mathrm{J}^\eta(t_{k+1}, X_{t_{k+1}}, \mathbb{P}_{X_{t_{k+1}}}) - \mathrm{J}^\eta(t_k, X_{t_k}, \mathbb{P}_{X_{t_k}})$$
$$+ \big(f(X_{t_k}, \mathbb{P}_{X_{t_k}}, \alpha_{t_k}) + \lambda \log p_\theta(t_k, X_{t_k}, \mathbb{P}_{X_{t_k}}, \alpha_{t_k}) - \beta \mathrm{J}^\eta(t_k, X_{t_k}, \mathbb{P}_{X_{t_k}})\big)\Delta t\big] + \mathcal{H}_\theta[\mathrm{J}^\eta](t_k, X_{t_k}, \mu_{t_k})\Delta t\Big\} = 0.$$

The pseudo-code is described in Algorithm 2.

Algorithm 2: Online actor-critic mean-field algorithm

Input data: Number of episodes N, number of mesh time-grid n ($\leftrightarrow$ time step $\Delta t = T/n$), learning rates ρ_S^i, ρ_E^i, ρ_G^i for the state distribution, PE and PG estimation, and function of the number of episodes i. Parameter λ for entropy regularisation.
Functional forms J^η of cost value function, p_θ of density policies.
Initialisation: μ_{t_k}: state distribution on $\mathbb{R}^d$, for $k = 0, \ldots, n$, parameters η, θ.
for *each episode* $i = 1, \ldots, N$ **do**
  Initialise state $X_0 \sim \mu_0$
  for $k = 0, \ldots, n-1$ **do**
    Update state distribution: $\mu_{t_k} \leftarrow (1 - \rho_S^i)\mu_{t_k} + \rho_S^i \delta_{X_{t_k}}$
    Generate action $\alpha_{t_k} \sim \pi_\theta(.|t_k, X_{t_k}, \mu_{t_k})$
    Observe (e.g. by environment simulator) state $X_{t_{k+1}}$ and cost f_{t_k}
    If $k = n-1$, update terminal state distribution: $\mu_{t_{k+1}} \leftarrow$
     $(1 - \rho_S^i)\mu_{t_{k+1}} + \rho_S^i \delta_{X_{t_{k+1}}}$, and observe terminal cost $g_{t_{k+1}}$
    Compute

$$\begin{aligned}
\delta_\eta &= \mathrm{J}^\eta(t_{k+1}, X_{t_{k+1}}, \mu_{t_{k+1}}) - \mathrm{J}^\eta(t_k, X_{t_k}, \mu_{t_k}) \\
&\qquad + \big(\, f_{t_k} + \lambda \log p_\theta(t_k, X_{t_k}, \mu_{t_k}, \alpha_{t_k}) - \beta \mathrm{J}^\eta(t_k, X_{t_k}, \mu_{t_k})\big)\Delta t \\
\Delta_\eta &= \delta_\eta \nabla_\eta \mathrm{J}^\eta(t_k, X_{t_k}, \mu_{t_k}) \\
\Delta_\theta &= \delta_\eta \nabla_\theta \log p_\theta(t_k, X_{t_k}, \mu_{t_k}, \alpha_{t_k}) + \mathcal{H}_\theta[\mathrm{J}^\eta](t_k, X_{t_k}, \mu_{t_k})\Delta t,
\end{aligned}$$

    with the constraint that when $k = n-1$, $\mathrm{J}^\eta(t_{k+1}, X_{t_{k+1}}, \mu_{t_{k+1}}) = g_{t_{k+1}}$.
    Critic Update: $\eta \leftarrow \eta + \rho_E^i \Delta_\eta$; Actor Update: $\theta \leftarrow \theta - \rho_G^i \Delta_\theta$
    $k \leftarrow k + 1$
  end
end
Return: J^η, π_θ

Remark 3.2 (About the choice of actor and critic parametric functions) *In the Actor-critic algorithms, we have to specify a parametric family of randomised policies π_θ, and a*

parametric family of critic functions J^η. *In general, for critic functions, one can consider cylindrical neural network functions in the form*

$$\mathrm{J}^\eta(t,x,\mu) = \Psi(t,x,<\varphi,\mu>), \quad (t,x,\mu) \in [0,T]\times\mathbb{R}^d\times\mathcal{P}_2(\mathbb{R}^d), \tag{3.7}$$

where Ψ *is a feedforward neural network from* $[0,T]\times\mathbb{R}^d\times\mathbb{R}^k$ *into* $\mathbb{R}$, *and* φ *is another feedforward neural network from* $\mathbb{R}^d$ *into* $\mathbb{R}^k$ *(called latent space), and we use the notation* $<\phi,\mu> := \int \phi(x)\mu(\mathrm{d}x)$. *The set of parameters* η *is the union of the parameter sets for the two neural networks* Ψ *and* φ. *This choice is motivated by the density property of the set of cylindrical functions, i.e. functions in the form* (3.7) *with continuous functions* Ψ *and* φ, *with respect to continuous functions on* $[0,T]\times\mathbb{R}^d\times\mathcal{P}_2(\mathbb{R}^d)$ *as shown in [19], and the universal approximation property of feedforward neural networks on finite-dimensional space, see [22].*

Concerning the policies, notice that when the temperature parameter for exploration λ *is zero, the optimal policy is of pure (non randomised) feedback form as a function of* (t,x,μ). *When* $\lambda > 0$, *the optimal policy is in general truly randomised, and the larger is* λ, *the larger is the exploration in the sense that the variance of the randomised policy increases. We can then take for the parametric family of randomised policies, for example Gaussian distributions:*

$$\pi_\theta(.|t,x,\mu) \;=\; \mathcal{N}\big(\mathrm{m}(t,x,\mu);\vartheta(\lambda)\big),$$

where m *is a cylindrical neural network function on* $[0,T]\times\mathbb{R}^d\times\mathcal{P}_2(\mathbb{R}^d)$ *valued in* $A\subset\mathbb{R}^m$, *and* $\vartheta(.)$ *is a given symmetric matrix-valued function, nondecreasing w.r.t.* λ, *with* $\vartheta(\lambda)$ *positive-definite for* $\lambda > 0$, *and* $\vartheta(0) = 0$.

In some particular mean-field models, we may know a priori the structural form of the optimal value function and optimal randomised policy, and this suggests alternately some specific form for the parametric family of actor and critic functions. This is typically the case of the linear quadratic model, as presented in the next section.

Remark 3.3 *The above actor-critic algorithms involve the computation of the term* $\mathcal{H}_\theta[\mathrm{J}^\eta]$ *at each time* t_k, *and along the observed state* X_{t_k} *and estimated state distribution* μ_{t_k}. *This additional term, compared to the actor-critic algorithms designed in [24] for standard stochastic control without mean-field interaction, involves the operator* $\mathcal{H}_\theta$ *defined in* (3.4). *In the separable form case, namely when the coefficients of the mean-field process are in the form*

$$b(x,\mu,a) \;=\; \mathrm{b}(x,\mu) + C(a), \qquad (\sigma\sigma^\intercal)(x,\mu,a) \;=\; \Sigma(x,\mu) + F(a),$$

where C *and* F *are known functions from* A *into* $\mathbb{R}^d$, *resp.* $\mathbb{R}^{d\times d}$, *we notice that*

$$\begin{aligned}
\nabla_\theta b_\theta(t,x,\mu) &= \nabla_\theta C_\theta(t,x,\mu), \quad \textit{with} \quad C_\theta(t,x,\mu) := \int C(a)\pi_\theta(\mathrm{d}a|t,x,\mu),\\
\nabla_\theta \Sigma_\theta(t,x,\mu) &= \nabla_\theta F_\theta(t,x,\mu), \quad \textit{with} \quad F_\theta(t,x,\mu) := \int F(a)\pi_\theta(\mathrm{d}a|t,x,\mu),
\end{aligned}$$

are known functions, and consequently also the function $\mathcal{H}_\theta[\mathrm{J}^\eta]$. Another important case where the term $\mathcal{H}_\theta[\mathrm{J}^\eta]$ is a known computable function is given in the linear quadratic framework as presented in the next section.

4 The linear quadratic case

We focus on the important class of MFC control problem with linear state dynamics and quadratic reward, namely

$$\begin{cases} b(x,\mu,a) \;=\; Bx + \bar{B}\bar{\mu} + Ca, \quad \sigma(x,\mu,a) \;=\; \gamma + Dx + \bar{D}\bar{\mu} + Fa, \\ f(x,\mu,a) \;=\; x^\intercal Q x + \bar{\mu}^\intercal \bar{Q}\bar{\mu} + a^\intercal N a + 2a^\intercal I x + 2a^\intercal \bar{I}\bar{\mu} + 2M.x + 2H.a, \\ g(x,\mu) \;=\; x^\intercal P x + \bar{\mu}^\intercal \bar{P}\bar{\mu} + 2L.x, \end{cases} \tag{4.1}$$

for $(x,\mu,a) \in \mathbb{R}^d \times \mathcal{P}_2(\mathbb{R}^d) \times \mathbb{R}^m$, where we denote by $\bar{\mu} = \int x\mu(\mathrm{d}x)$, B, $\bar{B}$, D, $\bar{D}$ are constant matrices in $\mathbb{R}^{d\times d}$, C, F are constant matrices in $\mathbb{R}^{d\times m}$, γ is a constant in $\mathbb{R}^d$, N is a symmetric matrix in $\mathbb{S}_+^m$, I, $\bar{I} \in \mathbb{R}^{m\times d}$, Q, $\bar{Q}$, P, $\bar{P}$ are symmetric matrices in $\mathbb{S}^d$, with $Q \geq 0$, $P \geq 0$, M, $L \in \mathbb{R}^d$, $H \in \mathbb{R}^m$.

In this case, the optimal value function to this LQ MFC problem with entropy regularisation when minimizing over randomised controls a functional cost as in (2.5), is given by

$$v(t,x,\mu) = \; (x-\bar{\mu})^\intercal K(t)(x-\bar{\mu}) + \bar{\mu}^\intercal \Lambda(t)\bar{\mu} + 2Y(t).x + R(t),$$

where K (valued in $\mathbb{S}^d$), Λ (valued in $\mathbb{S}^d$), Y valued in $\mathbb{R}^d$, and R valued in $\mathbb{R}$, are solutions to a system of ordinary differential equations on $[0,T]$ given in (B.1). Moreover, the optimal randomised control is of feedback form with Gaussian distribution:

$$\pi^*(.|t,x,\mu) \;\;=\;\; \mathcal{N}\Big(-S(t)^{-1}\big(U(t)x + (W(t)-U(t))\bar{\mu} + O(t)\big); \frac{\lambda}{2}S(t)^{-1}\Big),$$

where

$$\begin{aligned} S(t) \;&=\; N + F^\intercal K(t) F, \qquad & O(t) \;&=\; H + C^\intercal Y(t) + F^\intercal K(t)\gamma \\ U(t) \;&=\; I + C^\intercal K(t) + F^\intercal K(t) D, \qquad & W(t) \;&=\; I + \bar{I} + C^\intercal \Lambda(t) + F^\intercal K(t)(D+\bar{D}). \end{aligned}$$

This is an extension of the mean-field LQ control without entropy and control randomization, and the proof that adapts arguments in [2] is reported in Appendix B.

In a RL setting, the coefficients of the LQ model (4.1) are unknown, thus K, Λ, Y, and R cannot be solved from the system of ODEs, and S, O, U, and W are also unknown. We shall then employ our RL algorithms to solve the LQ problem in a model-free setting. In view of the above structure of the optimal value function and randomised policy, we parametrise the cost value function by

$$\mathrm{J}^\eta(t,x,\mu) = \; (x-\bar{\mu})^\intercal K^\eta(t)(x-\bar{\mu}) + \bar{\mu}^\intercal \Lambda^\eta(t)\bar{\mu} + 2Y^\eta(t).x + R^\eta(t), \tag{4.2}$$

for some parametric functions K^η, Λ^η, Y^η, R^η on $[0,T]$, with parameters $\eta \in \mathbb{R}^p$. On the other hand, we parametrise the randomised policies by

$$\pi_\theta(.|t,x,\mu) = \mathcal{N}\big(\phi_1^\theta(t)x + \phi_2^\theta(t)\bar{\mu} + \phi_3^\theta(t); \Sigma^\theta(t)\big), \tag{4.3}$$

for some parametric functions $\phi_1^\theta, \phi_2^\theta, \phi_3^\theta, \Sigma^\theta$ on $[0,T]$, with parameter $\theta \in \mathbb{R}^q$.

The parametric functions K^η, Λ^η, Y^η, R^η, and $\phi_1^\theta, \phi_2^\theta, \phi_3^\theta, \Sigma^\theta$, could be in general neural networks on $[0,T]$, but depending on the examples, we could take more specific forms, as discussed in the next section.

For parametrisation of the cost value function and randomised policies as in (4.2), (4.3), we see that

$$\begin{aligned}
\partial_\mu \mathrm{J}^\eta(t,x,\mu)(x') &= -2K^\eta(t)(x-\bar{\mu}) + 2\Lambda\eta\bar{\mu}, \text{ and so } \partial_{x'}\partial_\mu \mathrm{J}^\eta(t,x,\mu)(x') = 0,\\
\nabla_\theta b_\theta(t,x,\mu) &= C\nabla_\theta\phi_1^\theta(t) \bullet_2 x + C\nabla_\theta\phi_2^\theta(t) \bullet_2 \bar{\mu} + C\nabla_\theta\phi_3^\theta(t)
\end{aligned}$$

and then

$$\mathcal{H}_\theta[\mathrm{J}^\eta](t,x,\mu) = 2\big[(\nabla_\theta\phi_1^\theta(t) + \nabla_\theta\phi_2^\theta(t)) \bullet_2 \bar{\mu} + \nabla_\theta\phi_3^\theta(t)\big]^\intercal C^\intercal\big(-K^\eta(t)(x-\bar{\mu}) + \Lambda^\eta\bar{\mu}\big),$$

which only involves, up to the knowledge of C, known functions of (t,x,μ). Notice also that when $\phi_1^\theta = -\phi_2^\theta$, and $\phi_3^\theta \equiv 0$ (see below the example of mean-field systemic risk), then $\mathcal{H}_\theta[\mathrm{J}^\eta] \equiv 0$.

5 Numerical examples

5.1 Example 1: mean-field systemic risk

We consider a mean-field model of systemic risk introduced in [6]. This fits into a LQ MFC with

$$\begin{gathered}
\bar{B} = -B > 0,\ C = 1,\ \gamma > 0,\ D = \bar{D} = F = 0\\
I = -\bar{I} > 0,\ Q + \bar{Q} = 0,\ N = \frac{1}{2},\ M = H = L = 0,\ P + \bar{P} = 0,
\end{gathered}$$

and $Q \geq 2I^2$. We also take $X_0 \sim \mathcal{N}(0,1)$. In this case, the solution to the system of ODEs (B.1) yields the analytic expression:

$$\begin{aligned}
K(t) &= -\frac{1}{2}\Big[\bar{B} + 2I - \sqrt{\Delta}\frac{\sqrt{\Delta}\sinh(\sqrt{\Delta}(T-t)) + (\bar{B}+2I+2P)\cosh(\sqrt{\Delta}(T-t))}{\sqrt{\Delta}\cosh(\sqrt{\Delta}(T-t)) + (\bar{B}+2I+2P)\sinh(\sqrt{\Delta}(T-t))}\Big],\\
R(t) &= \frac{\gamma^2}{2}\ln\Big[\cosh(\sqrt{\Delta}(T-t)) + \frac{\bar{B}+2I+2P}{\sqrt{\Delta}}\sinh(\sqrt{\Delta}(T-t))\Big] - \frac{\gamma^2}{2}(\bar{B}+2I)(T-t)\\
&\quad -\frac{\lambda(T-t)}{2}\log(2\pi\lambda)
\end{aligned}$$

with $\sqrt{\Delta} = \sqrt{(\bar{B}+2I)^2+2Q-4I^2}$, and $\Lambda = Y = 0$, while the optimal randomised policy is given by

$$\hat{\pi}(.|t,x,\mu) \;=\; \mathcal{N}\big(\phi(t)(x-\bar{\mu});\lambda\big), \quad \text{with } \phi(t) \;=\; -2(K(t)+I).$$

In view of these expressions, we shall use critic function as

$$\mathrm{J}^\eta(t,x,\mu) \;=\; K^\eta(t)(x-\bar{\mu})^2 + R^\eta(t),$$

for some parametric functions K^η and R^η on $[0,T]$ with parameters η, and actor functions as

$$\begin{aligned} \pi_\theta(.|t,x,\mu) &= \mathcal{N}\big(\phi^\theta(t)(x-\bar{\mu});\lambda\big), \\ \text{i.e.} \quad \log p_\theta(t,x,\mu,a) &= -\frac{1}{2}\log(2\pi\lambda) - \frac{\big|a-\phi^\theta(t)(x-\bar{\mu})\big|^2}{2\lambda}, \end{aligned}$$

for some parametric function ϕ^θ on $[0,T]$ with parameter θ. As shown in Section 4, we notice that $\mathcal{H}_\theta[\mathrm{J}^\eta] = 0$.

We shall test with two choices of parametric functions:

1. *Exact parametrisation:*

$$\begin{cases} K^\eta(t) \;=\; -\frac{1}{2}\Big[\eta_3 - \eta_1 \frac{\sinh(\eta_1(T-t))+\eta_2\cosh(\eta_1(T-t))}{\cosh(\eta_1(T-t))+\eta_2\sinh(\eta_1(T-t))}\Big], \\ R^\eta(t) \;=\; \eta_4 \ln\Big[\cosh(\eta_1(T-t)) + \eta_2\sinh(\eta_1(T-t))\Big] - \eta_3\eta_4(T-t) - \frac{\lambda(T-t)}{2}\log(2\pi\lambda) \\ \phi^\theta(t) \;=\; \theta_3 - \theta_1 \frac{\sinh(\theta_1(T-t))+\theta_2\cosh(\theta_1(T-t))}{\cosh(\theta_1(T-t))+\theta_2\sinh(\theta_1(T-t))}, \end{cases} \tag{5.1}$$

   with parameters $\eta = (\eta_1,\eta_2,\eta_3,\eta_4) \in \mathbb{R}_+^4$, and $\theta = (\theta_1,\theta_2,\theta_3) \in \mathbb{R}_+^3$, so that the optimal solution in the model-based case corresponds to $\eta_1^* = \sqrt{\Delta}$, $\eta_2^* = \bar{B}+2I+2P$, $\eta_3^* = \bar{B}+2I$, $\eta_4^* = \gamma^2/2$, and $\theta_1^* = \sqrt{\Delta}$, $\theta_2^* = \bar{B}+2I+2P$, $\theta_3^* = \bar{B}$.

2. *Neural networks:* for K^η, R^η and ϕ^θ, with time input.

We implement our actor-critic algorithms with a simulator of X for coefficients equal to

$$T = 1,\ \gamma = 1,\ \bar{B} = -B = 0.6,\ I = 0.4,\ P = Q = 1,$$

The simulator for X is based on the real mean-field model:

$$\mathrm{d}X_t \;=\; \big(\bar{B}(\mathbb{E}[X_t]-X_t)+\alpha_t\big)\mathrm{d}t + \gamma \mathrm{d}W_t.$$

Since $\alpha \sim \pi^\theta$, we note that $\mathbb{E}[\alpha_t] = \phi^\theta(t)(\mathbb{E}[X_t]-\mathbb{E}[\bar{\mu}_t]) = 0$. We deduce that under such α, $\mathrm{d}\mathbb{E}[X_t] = 0$, hence $\mathbb{E}[X_t] = \mathbb{E}[X_0]$. From the above mean-field dynamics of X, we deduce that

$$\begin{aligned} X_{t_{k+1}} - \mathbb{E}[X_0] &= e^{-\bar{B}\Delta t}(X_{t_k}-\mathbb{E}[X_0]) + \alpha_{t_k}\big(\frac{1-e^{-\bar{B}\Delta t}}{\bar{B}}\big) + \gamma\int_{t_k}^{t_{k+1}} e^{-\bar{B}(t_{k+1}-s)}\mathrm{d}W_s \\ &\simeq e^{B\Delta t}(X_{t_k}-\mathbb{E}[X_0]) + \alpha_{t_k}\big(\frac{1-e^{-\bar{B}\Delta t}}{\bar{B}}\big) + \gamma e^{-\bar{B}\Delta t}\Delta W_{t_k}. \end{aligned}$$

The cost is simulated according to

$$f_{t_k} \;=\; Q(X_{t_k} - \mathbb{E}[X_0])^2 + \frac{1}{2}\alpha_{t_k}^2 + 2\alpha_{t_k} I(X_{t_k} - \mathbb{E}[X_0]), \quad g_T \;=\; P(X_T - \mathbb{E}[X_0])^2.$$

We first present the numerical results of our offline Algorithm 1 when using the exact parametrisation (5.1). The derivatives w.r.t. to η of K^η, R^η, hence of J^η, as well as the derivative w.r.t. θ of $\log p_\theta$ have explicit analytic expressions that are implemented in the updating rule of the actor-critic algorithm.

Here we used the following parameters: μ_{t_k} was initialized at 0; the number of episodes was $N = 2100$; the time horizon was $T = 1$ and the time step $\Delta t = 0.02$. The values of the model parameters were as described above. The learning rates (ρ_S, ρ_E, ρ_G) and λ were taken as $\rho_S = 0.2$ constant, and at iteration i,

$$\rho_E(i) = \begin{cases} (0.01, 0.1, 0.01, 0.2) \text{ if } i \leq 500 \\ (0.1, 0.1, 0.1, 0.1) \text{ if } 500 < i \leq 21000 \end{cases} \qquad \rho_G(i) = \begin{cases} (0.03, 0.05, 0.03) \text{ if } i \leq 7000 \\ (0.01, 0.01, 0.01) \text{ if } 7000 < i \leq 10000 \\ (0.005, 0.01, 0.005) \text{ if } 10000 < i \leq 14000 \\ (0.002, 0.002, 0.002) \text{ if } 17000 < i \leq 21000 \end{cases}$$

and

$$\lambda(i) = \begin{cases} 0.1 \text{ if } i \leq 8000, \\ 0.01 \text{ if } 8000 < i \leq 14000, \\ 0.001 \text{ if } 14000 < i \leq 21000 \end{cases}$$

Moreover, after $i = 14000$ iterations, we also increase the size of the minibatch from 20 to 40. In Table 1, we give the learnt parameters for the critic and actor functions, to be compared wih the exact value of the parameters.

	η_1	η_2	η_3	η_4	θ_1	θ_2	θ_3
exact	1.8221	1.8660	1.4	0.5	1.8221	1.8660	0.6
learnt	1.4197	2.0536	0.9997	0.4824	1.6204	1.9167	0.3660

Table 1: Learnt vs exact parameters of the critic and actor functions.

In Figure 1, we see that, even though the parameters η and θ (shown with full lines) are slightly different from the true optimal values (shown in dashed lines), the functions K, R and ϕ are matched almost perfectly.

We also display one realization of the control and of the cost. These are based on evaluating the control and the cumulative cost along one trajectory of the state. We first simulate 10^4 realizations of a Brownian motion. Based on this, we generate trajectories for one 10^4

population of agents using the learnt control and one population of 10^4 agents using the optimal control. For the population that uses the learnt control, the control is given by the mean of the actor, namely, $\phi^\theta(t)(x-\bar{\mu})$. In the dynamics, the cost and the control, the mean field term is replaced by the empirical mean of the corresponding population at the current time. We can see that the trajectories of control (resp. cost) are very similar.

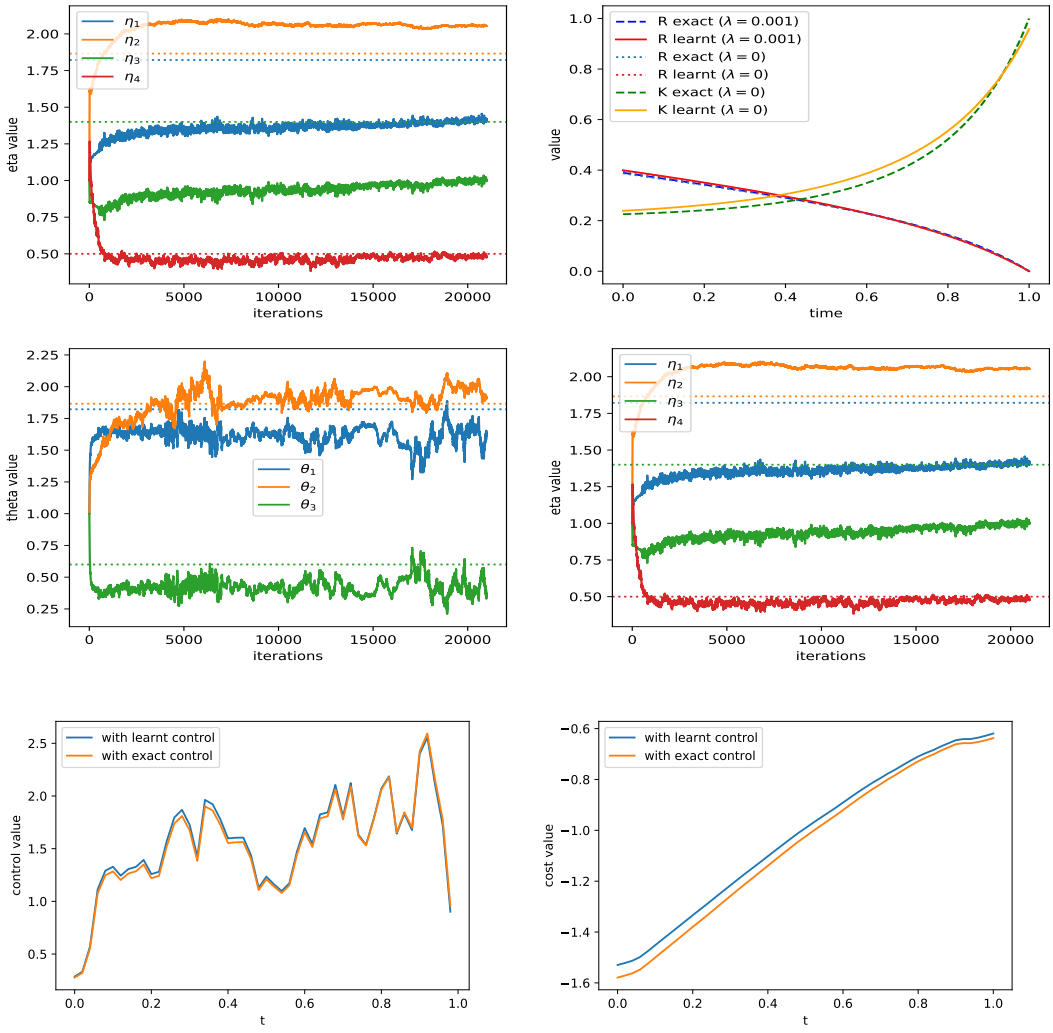

Figure 1: **Convergence of the learnt value function and policy with exact parametrisation for the offline Algorithm 1.** *Top*: learned parameters of critic (left) and associated critic functions K^η, R^η (right) vs optimal parameters and associated functions. *Middle:* learned parameters of actor (left) and associated actor function ϕ^θ vs optimal parameters and associated function. *Bottom:* one realization of the control (left) and one realization of the cost (right) vs the optimal ones, respectively along a state trajectory controlled by the learnt control and a state trajectory controlled optimally (both using the same realization of the Brownian motion).

Next, we present in Figure 2 and Figure 3 the numerical results of our online Algorithm 2 when using neural networks. In this case, the derivatives w.r.t. to η of K^η, R^η, hence of J^η, as well as the derivative w.r.t. θ of $\log p_\theta$ are computed by automatic differentiation. We use neural networks with 3 hidden layers, 10 neurons per layer and tanh activation functions. We take $n = 30$, $N = 15000$ iterations, batch size 500 (10000 for the law estimation in the simulator), constant learning rates 10^{-3}, except $\omega_S = 1$. We change λ along episodes: $\lambda = 0.1$ for the first 3334 ones, then 0.01 for the next 3333 ones, then 0.001 until the end.

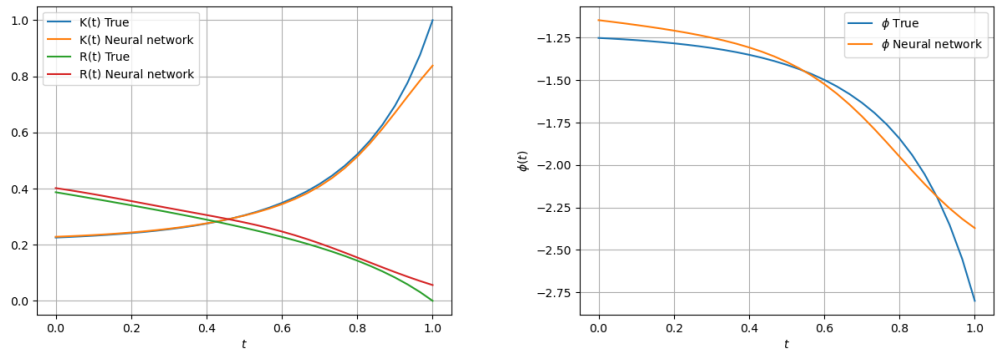

Figure 2: **Learnt critic cost function with neural networks for the online Algorithm 2**. *Left panel*: Neural network functions K^η, R^η vs optimal one. *Right panel*: Neural network function ϕ^θ vs optimal one.

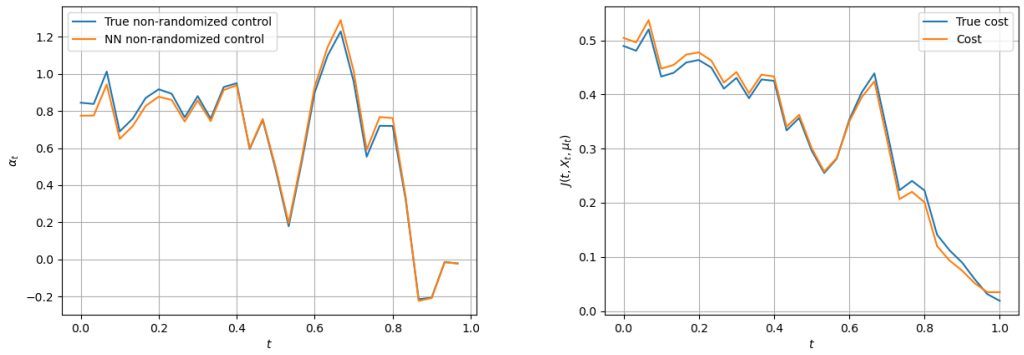

Figure 3: **Learnt actor policy function with neural networks for the online Algorithm 2**. *Left panel*: One realization of the control vs the optimal non-randomised one with $\lambda = 0$ *Right panel*: Plot of one realization of $t \mapsto \mathrm{J}^\eta(t, X_t, \mathbb{P}_{X_t})$, respectively along a state trajectory controlled by the learnt policy and a state trajectory controlled optimally (both using the same realization of the Brownian motion).

Finally, we test in Table 2 the learnt policies from the exact and NN parametrisation by computing the associated initial expected social costs. We simulate 10 populations, each

consisting of 10^4 agents. All the agents use the control function with the parameters learnt by the algorithm. For the dynamics, the cost and the control, the mean field term is replaced by the empirical mean of the corresponding population at the present time step. For each population, we compute the social cost. We then average over the 10 populations in order to get a Monte Carlo estimate of the social cost. We report in the table the value of this average social cost, the standard deviation over the 10 populations, and the relative error between the average social cost and the optimal cost computed by the formula $K_0^{\eta^*}\mathrm{Var}(X_0) + R_0^{\eta^*}$ with the optimal parameter η^*.

	Initial cost (Std dev.)	Rel. error
Learnt with exact parameterization	0.625 (0.006)	2.00%
Learnt with NN	0.632 (0.006)	3.10%
Exact value	0.613	

Table 2: Initial costs when following learnt policies vs optimal ones

5.2 Example 2: optimal trading

We consider an optimal trading problem where the inventory is governed by

$$\mathrm{d}X_t \;=\; \alpha_t \mathrm{d}t + \gamma \mathrm{d}W_t,$$

and we aim to minimize over randomised trading rate $\alpha \sim \pi$ the cost functional

$$\mathbb{E}\Big[\int_0^T \alpha_t^2 + 2H\alpha_t - \lambda\mathcal{E}(\pi_t)\mathrm{d}t + P\mathrm{Var}(X_T)\Big].$$

where $\gamma > 0$, $H > 0$ is the transaction price per trading, $P > 0$ is a risk aversion parameter, and $\lambda > 0$ is the temperature parameter. This model fits into the LQ framework, and the solution to the system of ODEs (B.1) is given by

$$K(t) \;=\; \frac{P}{1+P(T-t)}, \qquad R(t) \;=\; \gamma^2 \log(1+P(T-t)) - \big(H^2 + \frac{\lambda}{2}\log(\pi\lambda)\big)(T-t),$$

$\Lambda = Y \equiv 0$, while the optimal randomised policy is given by

$$\hat{\pi}(|t,x,\mu) \;\sim\; \mathcal{N}\Big(-K(t)(x-\bar{\mu}) - H; \frac{\lambda}{2}\Big).$$

In a RL setting, the coefficients σ, H and P are unknown, and we use critic function as

$$\mathrm{J}^\eta(t,x,\mu) \;=\; K^\eta(t)(x-\bar{\mu})^2 + R^\eta(t),$$

for some parametric functions K^η and R^η on $[0,T]$ with parameters η, and actor functions as

$$\begin{aligned} \pi_\theta(.|t,x,\mu) &= \mathcal{N}\big(\phi^\theta(t)(x-\bar{\mu})+\phi_3^\theta(t);\frac{\lambda}{2}\big), \\ \text{i.e.} \quad \log p_\theta(t,x,\mu,a) &= -\frac{1}{2}\log(\pi\lambda) - \frac{\big|a-\phi^\theta(t)(x-\bar{\mu})-\phi_3^\theta(t)\big|^2}{\lambda}, \end{aligned}$$

for some parametric functions ϕ^θ, ϕ_3^θ on $[0,T]$ with parameter θ. Given such family of parametric actor/critic functions, we have

$$\mathcal{H}_\theta[\mathrm{J}^\eta](t,x,\mu) = -2CK^\eta(t)(x-\bar{\mu})\nabla_\theta\phi_3^\theta(t).$$

We shall test with two choices of parametric functions:

1. *Exact parametrisation:*

$$\begin{cases} K^\eta(t) &= \frac{\eta_1}{1+\eta_1(T-t)} \\ R^\eta(t) &= \eta_2\log(1+\eta_1(T-t)) - \big(\eta_3+\frac{\lambda}{2}\log(\pi\lambda)\big)(T-t) \\ \phi^\theta(t) &= -\frac{\theta_1}{1+\theta_1(T-t)}, \quad \phi_3^\theta(t) = -\theta_2, \end{cases} \tag{5.2}$$

   with parameters $\eta=(\eta_1,\eta_2,\eta_3)\in(0,\infty)^3$, $\theta=(\theta_1,\theta_2)\in(0,\infty)^2$, so that the optimal solution in the model-based case corresponds to $(\eta_1^*,\eta_2^*,\eta_3^*)=(P,\gamma^2,H^2)$, and $(\theta_1^*,\theta_2^*)=(P,H)$.

2. *Neural networks:* for K^η, R^η, ϕ^θ, and ϕ_3^θ with time input. Actually, we take for ϕ_3^θ a constant function.

We first present the numerical results of our offline Algorithm 1 when using the exact parametrisation (5.2). The derivatives w.r.t. to η of K^η, R^η, hence of J^η, as well as the derivative w.r.t. θ of $\log p_\theta$, and $\mathcal{H}_\theta[\mathrm{J}^\eta]$ have explicit analytic expressions that are implemented in the updating rule of the actor-critic algorithm. Here we used the following parameters: the learning rates (ρ_S,ρ_E,ρ_G) and λ were taken as $\rho_S=0.2$ constant, and at iteration i,

$$\rho_E(i)=\begin{cases}(0.05,0.05,0.05) \text{ if } i\le 8000\\ (0.05,0.05,0.01) \text{ if } 8000<i\le 20000\end{cases} \qquad \rho_G(i)=\begin{cases}(0.005,0.005) \text{ if } i\le 8000\\ (0.001,0.001) \text{ if } 8000<i\le 13000\\ (0.0005,0.0005) \text{ if } 8000<i\le 13000\end{cases}$$

and

$$\lambda(i)=\begin{cases}0.1 \text{ if } i\le 8000,\\ 0.01 \text{ if } 8000<i\le 13000\\ 0.001 \text{ if } 13000<i\le 20000\end{cases}$$

μ_{t_k} was initialized at 0; the number of episodes was $N = 20000$; the time horizon was $T = 1$ and the time step $\Delta t = 0.02$. The values of the model parameters are: $P = 3$, $H = 2$, $\gamma = 1$, and $X_0 \sim \mathcal{N}(1, 1)$.

In Table 3, we give the learnt parameters for the critic and actor function to be compared with the exact values, when using the learnt policy with learnt empirical distribution from the algorithm.

	η_1	η_2	η_3	θ_1	θ_2
exact	3	1	4	3	2
learnt	2.9864	0.9637	3.9154	3.0161	2.0016

Table 3: Learnt vs exact parameters of the critic and actor functions.

In Figure 4, we see that the parameters and, hence, the functions K, R and ϕ are matched almost perfectly. We also display one realization of the control and of the cost. These are based on evaluating the control and the cumulative cost along one trajectory of the state. We first simulate 10^4 realizations of a Brownian motion. Based on this, we generate trajectories for one 10^4 population of agents using the learnt control and one population of 10^4 agents using the optimal control. For the population that uses the learnt control, the control is given by the mean of the actor, namely, $\phi^\theta(t)(x - \bar{\mu}) + \phi_3^\theta(t)$. In the dynamics, the cost and the control, the mean field term is replaced by the empirical mean of the the corresponding population at the current time. We can see that the trajectories of control (resp. cost) are very similar.

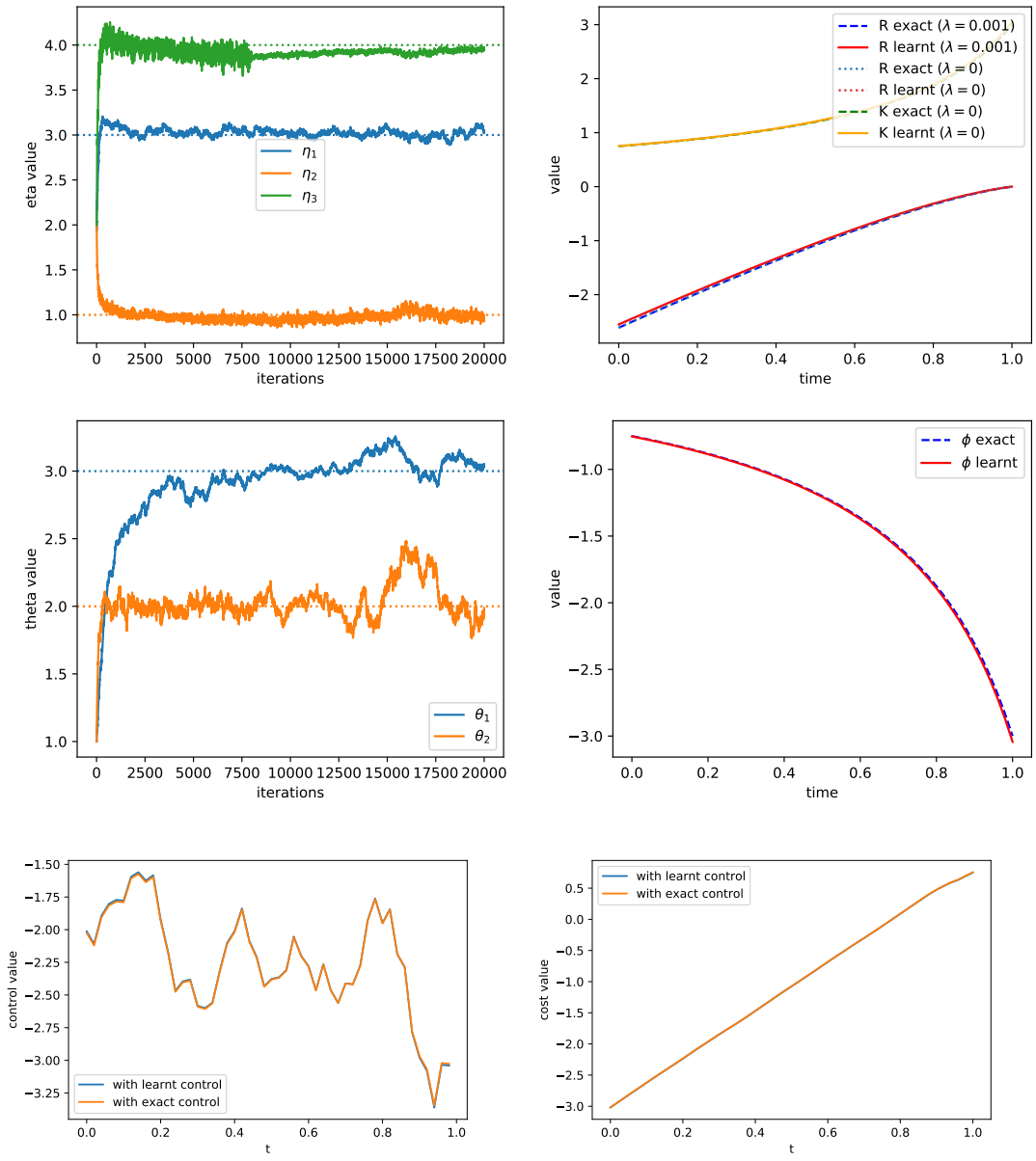

Figure 4: **Convergence of the learnt value function and policy with exact parametrisation for the offline Algorithm 1.** *Top*: learned parameters of critic (left) and associated critic functions K^η, R^η (right) vs optimal parameters and associated functions. *Middle:* learned parameters of actor (left) and associated actor function ϕ^θ vs optimal parameters and associated function. *Bottom:* one realization of the control (left) and one realization of the cost (right) vs the optimal ones, respectively along a state trajectory controlled by the learned control and a state trajectory controlled optimally (both using the same realization of the Brownian motion).

Next, we present in Figure 5 and Figure 6 the numerical results of our online Algorithm 2 when using neural networks. In this case, the derivatives w.r.t. to η of K^η, R^η, hence of J^η, as well as the derivative w.r.t. θ of $\log p_\theta$ are computed by automatic differentiation. We use neural networks with 3 hidden layers, 10 neurons per layer and tanh activation functions. We take $n = 30$, $N = 15000$ iterations, batch size 300 (10000 for the law estimation in the simulator), constant learning rates 10^{-3}, except $\omega_S = 1$. Again, we change λ along episodes: $\lambda = 0.1$ for the second 3334 ones, then 0.01 for the next 3333 ones, then 0.001 until the end.

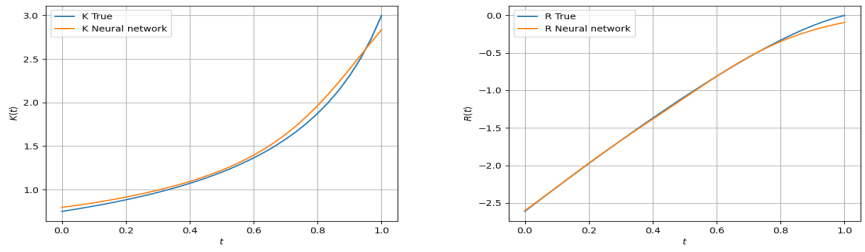

Figure 5: **Learnt critic cost function with neural networks for the online Algorithm 2**. *Left panel*: Neural network function K^η vs optimal one. *Right panel*: Neural network function R^η vs optimal one.

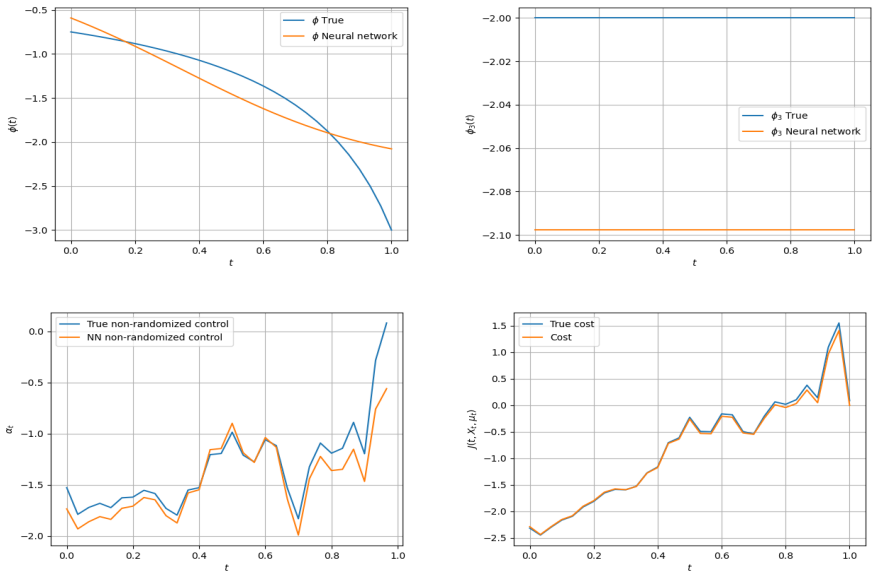

Figure 6: **Learnt actor policy function with neural networks (NN) for the online Algorithm 2**. *Left up panel*: NN ϕ^θ vs optimal one. *Right up panel*: NN ϕ^θ_3 vs optimal one. *Bottom panel*: (*left*) One realization of the control vs the optimal non-randomised one with $\lambda = 0$, and (*right*) one realization of the cost $t \mapsto \mathrm{J}^\eta(t, X_t, \mathbb{P}_{X_t})$, respectively along a state trajectory controlled by the learnt policy and a state trajectory controlled optimally (both using the same realization of the Brownian motion).

Finally, we test in Table 4 the learnt policies from the exact and NN parametrisation by computing the associated initial expected social costs. We simulate 10 populations, each consisting of 10^4 agents. All the agents use the control function with the parameters learnt by the algorithm. For the dynamics, the cost and the control, the mean field term is replaced by the empirical mean of the corresponding population at the present time step. For each population, we compute the social cost. We then average over the 10 populations in order to get a Monte Carlo estimate of the social cost. We report in the table the value of this average social cost, the standard deviation over the 10 populations, and the relative error between the average social cost and the optimal cost computed by the formula $K_0^{\eta^*}\mathrm{Var}(X_0) + R_0^{\eta^*}$ with the optimal parameter η^*.

	Social cost (Std dev.)	Rel. error
Learnt with exact parametrisation	-1.861 (0.025)	0.11%
Learnt with NN	-1.787 (0.035)	4.08%
Exact value	-1.863	

Table 4: Initial social costs when following learnt policies vs optimal one.

A Proofs of some representation results

A.1 Proof of Proposition 2.1

Step 1: For a fixed policy π, we introduce the non-linear McKean-Vlasov SDE with dynamics

$$\tilde{X}_s^{t,\xi} = \xi + \int_t^s b_\pi(r, \tilde{X}_r^{t,\xi}, \mathbb{P}_{\tilde{X}_r^{t,\xi}})\,\mathrm{d}r + \int_t^s \sigma_\pi(r, \tilde{X}_r^{t,\xi}, \mathbb{P}_{\tilde{X}_r^{t,\xi}})\,\mathrm{d}W_r, \tag{A.1}$$

recalling that $\sigma_\pi = \Sigma_\pi^{1/2}$, as well as its associated decoupled SDE with dynamics

$$\tilde{X}_s^{t,x,\mu} = x + \int_t^s b_\pi(r, \tilde{X}_r^{t,x,\mu}, \mathbb{P}_{\tilde{X}_r^{t,\xi}})\,\mathrm{d}r + \int_t^s \sigma_\pi(r, \tilde{X}_r^{t,x,\mu}, \mathbb{P}_{\tilde{X}_r^{t,\xi}})\,\mathrm{d}W_r. \tag{A.2}$$

Under Assumption 2.1(i), the coefficients b_π and σ_π are Lipschitz-continuous and with at most linear growth with respect to the variable x and μ locally uniformly in time. Hence, the SDEs (A.1)-(A.2) admit a unique strong solution.

Denoting by $\mathbb{P}$ the probability measure on $\mathcal{C}([0,\infty), \mathbb{R}^d)$ (the space of continuous functions defined on $[0,\infty)$ taking values in $\mathbb{R}^d$) induced by the unique solution to the SDE (A.1) and

by $\mathbb{P}(t)$ its marginal at time t, its infinitesimal generator is given by

$$\tilde{\mathcal{L}}_t^\pi \varphi(x) = \sum_{i=1}^d \int_A b_i(x, \mathbb{P}(t), a)\, \pi(\mathrm{d}a|t, x, \mathbb{P}(t)) \partial_{x_i}\varphi(x) + \frac{1}{2}\sum_{i,j=1}^d \int_A (\sigma\sigma^\intercal)_{i,j}(x, \mathbb{P}(t), a)\, \pi(\mathrm{d}a|t, x, \mathbb{P}(t)) \partial^2_{x_i,x_j}\varphi(x).$$

Now, coming back to the dynamics of the McKean-Vlasov SDE (2.4), we importantly point out that since at each time s, the action α_s is sampled from the probability distribution $\pi(.|s, X_s^{t,\xi}, \mathbb{P}_{X_s^{t,\xi}})$ independently of W, the infinitesimal generator at time t of (2.4) is exactly given by $\tilde{\mathcal{L}}_t^\pi$. Hence, it follows from the uniqueness of the martingale problem associated to $\tilde{\mathcal{L}}^\pi$ that $X^{t,\xi}$ and $\tilde{X}^{t,\xi}$ have the same law [b].

We thus conclude that V^π can be written as

$$V^\pi(t,x,\mu) = \mathbb{E}\Big[\int_t^T e^{-\beta(s-t)}(f_\pi - \lambda E_\pi)(s, \tilde{X}_s^{t,x,\mu}, \mathbb{P}_{\tilde{X}_s^{t,\xi}})\,\mathrm{d}s + e^{-\beta(T-t)} g(\tilde{X}_T^{t,x,\mu}, \mathbb{P}_{\tilde{X}_T^{t,\xi}})\Big]. \tag{A.3}$$

Step 2: We know, see e.g. [13] or [10], that Assumption 2.1(i) guarantees the existence of a modification of $\tilde{X}^{t,x,\mu}$ such that:

- The map $x \mapsto \tilde{X}_s^{t,x,\xi}$ is $\mathbb{P}$-a.s. twice continuously differentiable,
- for any $x \in \mathbb{R}^d$, $0 \le t \le s$, and any $p \ge 1$, the map $\mathcal{P}_2(\mathbb{R}^d) \ni \mu \mapsto \tilde{X}_s^{t,x,\mu} \in L^p(\mathbb{P})$ is differentiable and the map $\mathbb{R}^d \ni v \mapsto \partial_\mu \tilde{X}_s^{t,x,\mu}(v) \in L^p(\mathbb{P})$ is differentiable,
- for any $p \ge 1$, the derivatives $(t,x,\mu,v) \mapsto \partial_x \tilde{X}_s^{t,x,\mu}$, $\partial_x^2 \tilde{X}_s^{t,x,\mu}$, $\partial_\mu \tilde{X}_s^{t,x,\mu}(v)$, $\partial_v[\partial_\mu \tilde{X}_s^{t,x,\mu}](v) \in L^p(\mathbb{P})$ are continuous.

Moreover, the following estimates hold for $n = 1, 2$ and any $p \ge 1$

$$\sup_{0\le t\le s\le T, (x,v)\in(\mathbb{R}^d)^2, \mu\in\mathcal{P}_2(\mathbb{R}^d)} \Big\{ \|\partial_x^n \tilde{X}_s^{t,x,\mu}\|_{L^p(\mathbb{P})} + \|\partial_\mu \tilde{X}_s^{t,x,\mu}(v)\|_{L^p(\mathbb{P})} + \|\partial_v\partial_\mu \tilde{X}_s^{t,x,\mu}(v)\|_{L^p(\mathbb{P})} \Big\} < \infty.$$

We thus deduce that the functions $x \mapsto f_\pi(s, \tilde{X}_s^{t,x,\mu}, \mathbb{P}_{\tilde{X}_s^{t,\xi}})$, $E_\pi(s, \tilde{X}_s^{t,x,\mu}, \mathbb{P}_{\tilde{X}_s^{t,\xi}})$, $g(\tilde{X}_T^{t,x,\mu}, \mathbb{P}_{\tilde{X}_T^{t,\xi}})$ are $\mathbb{P}$-a.s. twice continuously differentiable with derivatives that belong to $L^p(\mathbb{P})$, for any $p \ge 1$, uniformly in x, μ and $t \in [0,s]$. The dominated convergence theorem eventually guarantees that $x \mapsto V^\pi(t,x,\mu)$ is twice continuously differentiable with

$$\partial_{x_i} V^\pi(t,x,\mu) = \mathbb{E}\Big[\int_t^T e^{-\beta(s-t)} \sum_{k=1}^d \partial_{x_k}(f_\pi - \lambda E_\pi)(s, \tilde{X}_s^{t,x,\mu}, \mathbb{P}_{\tilde{X}_s^{t,\xi}}) \partial_{x_i}(\tilde{X}_s^{t,x,\mu})^k \mathrm{d}s + e^{-\beta(T-t)} \sum_{k=1}^d \partial_{x_k} g(\tilde{X}_T^{t,x,\mu}, \mathbb{P}_{\tilde{X}_T^{t,\xi}}) \partial_{x_i}(\tilde{X}_T^{t,x,\mu})^k \Big], \tag{A.4}$$

[b] This was formally shown by law of large numbers in [30] in the standard diffusion case.

and

$$\begin{aligned}\partial^2_{x_i,x_j}V^\pi(t,x,\mu) = \ & \mathbb{E}\Big[\int_t^T e^{-\beta(s-t)}\sum_{k,\ell=1}^d \partial^2_{x_k,x_\ell}(f_\pi-\lambda E_\pi)(s,\tilde X_s^{t,x,\mu},\mathbb{P}_{\tilde X_s^{t,\xi}})\partial_{x_i}(\tilde X_s^{t,x,\mu})^k\partial_{x_j}(\tilde X_s^{t,x,\mu})^\ell \mathrm{d}s \\ & + \ e^{-\beta(T-t)}\sum_{k,\ell=1}^d \partial^2_{x_k,x_\ell} g(\tilde X_T^{t,x,\mu},\mathbb{P}_{\tilde X_T^{t,\xi}})\partial_{x_i}(\tilde X_T^{t,x,\mu})^k\partial_{x_j}(\tilde X_T^{t,x,\mu})^\ell \\ & + \int_t^T e^{-\beta(s-t)}\sum_{k=1}^d \partial_{x_k}(f_\pi-\lambda E_\pi)(s,\tilde X_s^{t,x,\mu},\mathbb{P}_{\tilde X_s^{t,\xi}})\partial^2_{x_i,x_j}(\tilde X_s^{t,x,\mu})^k\mathrm{d}s \\ & + \ e^{-\beta(T-t)}\sum_{k=1}^d \partial_{x_k} g(\tilde X_T^{t,x,\mu},\mathbb{P}_{\tilde X_T^{t,\xi}})\partial^2_{x_i,x_j}(\tilde X_T^{t,x,\mu})^k\Big].\end{aligned}$$

It follows from the above expression and again the dominated convergence theorem that $(t,x,\mu)\mapsto \partial_{x_i}V^\pi(t,x,\mu), \partial^2_{x_i,x_j}V^\pi(t,x,\mu)$ are continuous.

Similarly, note that under the current assumption, the functions $\mu\mapsto h(t,\tilde X_s^{t,x,\mu},\mathbb{P}_{\tilde X_s^{t,\xi}})$, $g(\tilde X_T^{t,x,\mu},\mathbb{P}_{\tilde X_T^{t,\xi}})$, where $h\in\{f_\pi,E_\pi\}$, are L-differentiable with derivatives satisfying

$$\begin{aligned}\partial^i_\mu[h(s,\tilde X_s^{t,x,\mu},\mathbb{P}_{\tilde X_s^{t,\xi}})](v) = & \sum_{k=1}^d \partial_{x_k}h(s,\tilde X_s^{t,x,\mu},\mathbb{P}_{\tilde X_s^{t,\xi}})\partial^i_\mu[(\tilde X_s^{t,x,\mu})^k](v) \\ & + \widehat{\mathbb{E}}\Big[\sum_{k=1}^d \partial^k_\mu h(s,\tilde X_s^{t,x,\mu},\mathbb{P}_{\tilde X_s^{t,\xi}})(\widehat X_s^{t,v,\mu})\partial_{x_i}(\widehat X_s^{t,v,\mu})^k\Big] \\ & + \int_{\mathbb{R}^d}\widehat{\mathbb{E}}\Big[\sum_{k=1}^d \partial^k_\mu h(s,\tilde X_s^{t,x,\mu},\mathbb{P}_{\tilde X_s^{t,\xi}})(\widehat X_s^{t,x',\mu})\partial^i_\mu[(\widehat X_s^{t,x',\mu})^k](v)\Big]\,\mu(\mathrm{d}x'), \\ \partial^i_\mu[g(\tilde X_T^{t,x,\mu},\mathbb{P}_{\tilde X_T^{t,\xi}})](v) = & \sum_{k=1}^d \partial_{x_k}g(\tilde X_T^{t,x,\mu},\mathbb{P}_{\tilde X_T^{t,\xi}})\partial^i_\mu[(\tilde X_T^{t,x,\mu})^k](v) \\ & + \widehat{\mathbb{E}}[\sum_{k=1}^d \partial^k_\mu g(\tilde X_T^{t,x,\mu},\mathbb{P}_{\tilde X_T^{t,\xi}})(\widehat X_T^{t,v,\mu})\partial_{x_i}(\widehat X_T^{t,v,\mu})^k] \\ & + \int_{\mathbb{R}^d}\widehat{\mathbb{E}}[\sum_{k=1}^d \partial^k_\mu g(\tilde X_T^{t,x,\mu},\mathbb{P}_{\tilde X_T^{t,\xi}})(\widehat X_T^{t,x',\mu})\partial^i_\mu[(\widehat X_T^{t,x',\mu})^k](v)]\,\mu(\mathrm{d}x'),\end{aligned} \tag{A.5}$$

where $(\widehat X_s^{t,x,\mu})_{s\in[t,T]}$ stands for a copy of $(\tilde X_s^{t,x,\mu})_{s\in[t,T]}$ defined on a copy $(\widehat\Omega,\widehat{\mathcal{F}},\widehat{\mathbb{P}})$ of the original probability space $(\Omega,\mathcal{F},\mathbb{P})$. Under Assumption 2.1, it follows from the above identities that $(t,x,\mu,v)\mapsto \partial_\mu[h(s,\tilde X_s^{t,x,\mu},\mathbb{P}_{\tilde X_s^{t,\xi}})](v)$, $\partial_\mu[g(\tilde X_T^{t,x,\mu},\mathbb{P}_{\tilde X_T^{t,\xi}})](v)\in L^p(\mathbb{P})$, $p\geq 1$, are continuous

and satisfy

$$|\partial_\mu[h(s,\tilde{X}_s^{t,x,\mu},\mathbb{P}_{\tilde{X}_s^{t,\xi}})](v)| \leq K(1+|\tilde{X}_s^{t,x,\mu}|+|v|+M_2(\mathbb{P}_{\tilde{X}_s^{t,\xi}})^q)(1+|\partial_\mu\tilde{X}_s^{t,x,\mu}(v)|)$$
$$\leq K(1+|\tilde{X}_s^{t,x,\mu}|+|v|+M_2(\mu)^q)(1+|\partial_\mu\tilde{X}_s^{t,x,\mu}(v)|),$$

and

$$|\partial_\mu[g(\tilde{X}_T^{t,x,\mu},\mathbb{P}_{\tilde{X}_T^{t,\xi}})](v)| \leq K(1+|\tilde{X}_T^{t,x,\mu}|+|v|+M_2(\mathbb{P}_{\tilde{X}_T^{t,\xi}})^q)(1+|\partial_\mu\tilde{X}_T^{t,x,\mu}(v)|)$$
$$\leq K(1+|\tilde{X}_T^{t,x,\mu}|+|v|+M_2(\mu)^q)(1+|\partial_\mu\tilde{X}_T^{t,x,\mu}(v)|),$$

where we used the fact that $M_2(\mathbb{P}_{\tilde{X}_s^{t,\xi}}) \leq K(1+M_2(\mu))$, for any $s \in [t,T]$, for the last inequality. Similarly, it follows from (A.5) and the dominated convergence theorem that $v \mapsto \partial_\mu[h(t,\tilde{X}_s^{t,x,\mu},\mathbb{P}_{\tilde{X}_s^{t,\xi}})](v)$, $\partial_\mu[g(\tilde{X}_T^{t,x,\mu},\mathbb{P}_{\tilde{X}_T^{t,\xi}})](v)$ are continuously differentiable with derivatives being continuous with respect to their entries and satisfying

$$|\partial_v\partial_\mu[h(s,\tilde{X}_s^{t,x,\mu},\mathbb{P}_{\tilde{X}_s^{t,\xi}})](v)| \leq K(1+|\tilde{X}_s^{t,x,\mu}|+|v|+M_2(\mu)^q),$$
$$|\partial_v\partial_\mu[g(\tilde{X}_T^{t,x,\mu},\mathbb{P}_{\tilde{X}_T^{t,\xi}})](v)| \leq K(1+|\tilde{X}_T^{t,x,\mu}|+|v|+M_2(\mu)^q).$$

Coming back to (A.3) and using the above estimates together with the dominated convergence theorem allows to conclude that $\mu \mapsto V^\pi(t,x,\mu)$ is L-differentiable and that $v \mapsto \partial_\mu V^\pi(t,x,\mu)(v)$ is differentiable. Moreover, both derivatives $\partial_\mu V^\pi(t,x,\mu)(v)$, $\partial_v\partial_\mu V^\pi(t,x,\mu)(v)$ are continuous with respect to their entries and satisfy

$$\sup_{t\in[0,T]}\{|\partial_\mu V^\pi(t,x,\mu)(v)|+|\partial_v\partial_\mu V^\pi(t,x,\mu)(v)|\} \leq K(1+|x|+|v|+M_2(\mu)^q). \tag{A.6}$$

We thus conclude that $V^\pi \in \mathcal{C}^{0,2,2}([0,T]\times\mathbb{R}^d\times\mathcal{P}_2(\mathbb{R}^d))$.

Step 3: Let us now prove that $(t,x,\mu) \mapsto V^\pi(t,x,\mu) \in \mathcal{C}^{1,2,2}([0,T]\times\mathbb{R}^d\times\mathcal{P}_2(\mathbb{R}^d))$. From the Markov property satisfied by the SDE (A.1), stemming from its strong well-posedness, for any $0 \leq h \leq t$, the following relation is satisfied

$$V^\pi(t-h,x,\mu) = e^{-\beta h}\mathbb{E}\Big[\int_{t-h}^t e^{-\beta(s-t)}(f_\pi-\lambda E_\pi)(s,\tilde{X}_s^{t-h,x,\mu},\mathbb{P}_{\tilde{X}_s^{t-h,\xi}})\mathrm{d}s\Big]$$
$$+ e^{-\beta h}\mathbb{E}\Big[V^\pi(t,\tilde{X}_t^{t-h,x,\mu},\mathbb{P}_{\tilde{X}_t^{t-h,\xi}})\Big].$$

Now, combining the fact that $V^\pi(t,.) \in \mathcal{C}^{2,2}(\mathbb{R}^d\times\mathcal{P}_2(\mathbb{R}^d))$ with (A.6) guarantees that one may

apply Itô's rule, see e.g. Proposition 5.102 [4]. We thus obtain

$$
\begin{aligned}
h^{-1}&(V^\pi(t-h,x,\mu)-V^\pi(t,x,\mu)) \\
&= e^{-\beta h}h^{-1}\int_{t-h}^{t} e^{-\beta(s-t)}\mathbb{E}\Big[(f_\pi-\lambda E_\pi)(s,\tilde{X}_s^{t-h,x,\mu},\mathbb{P}_{\tilde{X}_s^{t-h,\xi}})\Big]\mathrm{d}s \\
&\quad + e^{-\beta h}h^{-1}\mathbb{E}\Big[V^\pi(t,\tilde{X}_t^{t-h,x,\mu},\mathbb{P}_{\tilde{X}_t^{t-h,\xi}})-V^\pi(t,x,\mu)\Big] \\
&\quad + h^{-1}(e^{-\beta h}-1)V^\pi(t,x,\mu) \\
&= e^{-\beta h}h^{-1}\int_{t-h}^{t} e^{-\beta(s-t)}\mathbb{E}\Big[(f_\pi-\lambda E_\pi)(s,\tilde{X}_s^{t-h,x,\mu},\mathbb{P}_{\tilde{X}_s^{t-h,\xi}})\Big]\mathrm{d}s \\
&\quad + e^{-\beta h}h^{-1}\int_{t-h}^{t} \mathbb{E}\Big[\tilde{\mathcal{L}}[V^\pi](t,\tilde{X}_s^{t-h,x,\mu},\mathbb{P}_{\tilde{X}_s^{t-h,\xi}})\Big]\mathrm{d}s \\
&\quad + h^{-1}(e^{-\beta h}-1)V^\pi(t,x,\mu),
\end{aligned}
\tag{A.7}
$$

where

$$
\begin{aligned}
\tilde{\mathcal{L}}_\pi[\varphi](t,x,\mu) &= b_\pi(t,x,\mu)\cdot D_x\varphi(t,x,\mu)+\frac{1}{2}\Sigma_\pi(t,x,\mu):D_x^2\varphi(t,x,\mu) \\
&\quad + \mathbb{E}_{\xi\sim\mu}\Big[b_\pi(t,\xi,\mu)\cdot\partial_\mu\varphi(t,x,\mu)(\xi)+\frac{1}{2}\Sigma_\pi(t,\xi,\mu):\partial_\upsilon\partial_\mu\varphi(t,x,\mu)(\xi)\Big].
\end{aligned}
$$

Letting $h\downarrow 0$ in (A.7), from the continuity and quadratic growth of f_π, E_π as well as the continuity of $\tilde{\mathcal{L}}[V^\pi](t,.)$, we deduce that $t\mapsto V^\pi(t,x,\mu)$ is left-differentiable on $(0,T)$. Still from the continuity of f_π, E_π and $\tilde{\mathcal{L}}[V^\pi]$, we eventually conclude that it is differentiable on $[0,T)$ with a derivative satisfying

$$
\partial_t V^\pi(t,x,\mu)-\beta V^\pi(t,x,\mu)+\tilde{\mathcal{L}}_\pi[V^\pi](t,x,\mu)+(f_\pi-\lambda E_\pi)(t,x,\mu)=0.
$$

The proof is now complete.

A.2 Differentiability of the parametric critic function

Under the standard assumption that the coefficients $b_{\pi_\theta}(t,.)$, $\sigma_{\pi_\theta}(t,.)$ are Lipschitz-continuous on $\mathbb{R}^d\times\mathcal{P}_2(\mathbb{R}^d)$ uniformly in $t\in[0,T]$ and $\theta\in\Theta$, the system of SDEs (2.4) admits a unique strong solution when $\alpha\sim\pi_\theta$. We will denote by $(X_s^{t,\xi}(\theta),X_s^{t,x,\xi}(\theta))$ the solution taken at time s. We will also use the more compact notation

$$
\begin{aligned}
X_s^{t,\xi}(\theta) &= \xi+\int_t^s\sum_{j=0}^p g_\theta^j(r,X_r^{t,\xi}(\theta),\mathbb{P}_{X_r^{t,\xi}(\theta)})\,\mathrm{d}W_r^j \\
X_s^{t,x,\mu}(\theta) &= x+\int_t^s\sum_{j=0}^p g_\theta^j(r,X_r^{t,x,\mu}(\theta),\mathbb{P}_{X_r^{t,\xi}(\theta)})\,\mathrm{d}W_r^j,\quad t\le s\le T,
\end{aligned}
$$

with $g_\theta^0(t,x,\mu)=b_{\pi_\theta}(t,x,\mu)$, $g_\theta^j(t,x,\mu)=\sigma_{\pi_\theta}^{.,j}(t,x,\mu)$, $\mathrm{d}W_r=(\mathrm{d}W_r^0,\cdots,\mathrm{d}W_r^p)$ with $\mathrm{d}W_r^0=\mathrm{d}r$.

Lemma A.1 *Under Assumption 3.1, the derivatives* $(t,\theta,x,\mu,v) \mapsto \partial_\theta \partial_x \tilde{X}_s^{t,x,\mu}(\theta)$, $\partial_x \partial_\theta \tilde{X}_s^{t,x,\mu}(\theta)$, $\partial_\theta \partial_x^2 \tilde{X}_s^{t,x,\mu}(\theta)$, $\partial_x^2 \partial_\theta \tilde{X}_s^{t,x,\mu}(\theta)$, $\partial_\theta[\partial_\mu \tilde{X}_s^{t,x,\mu}(\theta)](v)$, $\partial_\mu \partial_\theta \tilde{X}_s^{t,x,\mu}(\theta)(v)$, $\partial_\theta \partial_v[\partial_\mu \tilde{X}_s^{t,x,\mu}(\theta)](v)$, $\partial_v[\partial_\mu \partial_\theta \tilde{X}_s^{t,x,\mu}(\theta)](v) \in L^p(\mathbb{P})$ *exist and are locally Lipschitz continuous for all* $p \geq 1$.

Proof. The proof of the existence and continuity of the derivatives of the flow $X_s^{t,x,\xi}(\theta)$ with respect to the parameters x, μ, v and θ is rather standard but quite mechanical and actually follows similar lines of reasonings as those employed for the proof of Theorem 3.2 in [13]. We thus omit it. □

With the same notations as Lemma A.1, under Assumption 3.1, taking $h_\theta = f_{\pi_\theta}$, E_{π_θ} or $g(x,\mu)$, we deduce from the above result that the derivatives $(t,\theta,x,\mu,\boldsymbol{v}) \mapsto \partial_\theta \partial_x [h_\theta(s, \tilde{X}_s^{t,x,\mu}(\theta), \mathbb{P}_{\tilde{X}_s^{t,\xi}(\theta)})]$, $\partial_x \partial_\theta [h_\theta(s, \tilde{X}_s^{t,x,\mu}(\theta), \mathbb{P}_{\tilde{X}_s^{t,\xi}(\theta)})]$, $\partial_\theta \partial_\mu [h_\theta(s, \tilde{X}_s^{t,x,\mu}(\theta), \mathbb{P}_{\tilde{X}_s^{t,\xi}(\theta)})](v)$, $\partial_\mu \partial_\theta [h_\theta(s, \tilde{X}_s^{t,x,\mu}(\theta), \mathbb{P}_{\tilde{X}_s^{t,\xi}(\theta)})](v)$, $\partial_\theta \partial_v \partial_\mu [h_\theta(s, \tilde{X}_s^{t,x,\mu}(\theta), \mathbb{P}_{\tilde{X}_s^{t,\xi}(\theta)})](v)$, $\partial_v \partial_\mu \partial_\theta [h_\theta(s, \tilde{X}_s^{t,x,\mu}(\theta), \mathbb{P}_{\tilde{X}_s^{t,\xi}(\theta)})](v) \in L^p(\mathbb{P})$, for any $p \geq 1$ and any $0 \leq t \leq s \leq T$, exist and are continuous. For instance, standard computations give

$$
\begin{aligned}
&\partial_{\theta_l} \partial_{x_i} [h_\theta(s, \tilde{X}_s^{t,x,\mu}(\theta), \mathbb{P}_{\tilde{X}_s^{t,\xi}(\theta)})] \\
&\quad = \sum_{j=1}^d \partial_{\theta_l} \partial_{x_j} h_\theta(s, \tilde{X}_s^{t,x,\mu}(\theta), \mathbb{P}_{\tilde{X}_s^{t,\xi}(\theta)}) \partial_{x_i} (\tilde{X}_s^{t,x,\mu}(\theta))^j \\
&\qquad + \sum_{j,k=1}^d \partial^2_{x_j,x_k} h_\theta(s, \tilde{X}_s^{t,x,\mu}(\theta), \mathbb{P}_{\tilde{X}_s^{t,\xi}(\theta)}) \partial_{x_i} (\tilde{X}_s^{t,x,\xi}(\theta))^j \partial_{\theta_l} (\tilde{X}_s^{t,x,\mu}(\theta))^k \\
&\qquad + \sum_{j,k=1}^d \widehat{\mathbb{E}}\Big[[\partial_\mu \partial_{x_j} h_\theta(s, \tilde{X}_s^{t,x,\mu}(\theta), \mathbb{P}_{\tilde{X}_s^{t,\xi}(\theta)})]_k (\widehat{X}_s^{t,\mu}(\theta)) \partial_{\theta_l} (\widehat{X}_s^{t,\xi}(\theta))^k \Big] \partial_{x_i} (\tilde{X}_s^{t,x,\mu}(\theta))^j \\
&\qquad + \sum_{j=1}^d \partial_{x_j} h_\theta(s, \tilde{X}_s^{t,x,\mu}(\theta), \mathbb{P}_{\tilde{X}_s^{t,\xi}(\theta)}) \partial_{\theta_l} \partial_{x_i} (\tilde{X}_s^{t,x,\mu}(\theta))^j
\end{aligned}
$$

and

$$
\begin{aligned}
&\partial_{\theta_l}\partial_\mu^i[h_\theta(s,\tilde{X}_s^{t,x,\mu}(\theta),\mathbb{P}_{\tilde{X}_s^{t,\xi}(\theta)})](v)\\
&\quad=\sum_{j=1}^d \partial_{\theta_l}\partial_{x_j}h_\theta(s,\tilde{X}_s^{t,x,\mu}(\theta),\mathbb{P}_{\tilde{X}_s^{t,\xi}(\theta)})\partial_\mu^i(\tilde{X}_s^{t,x,\mu}(\theta))^j\\
&\qquad+\sum_{j,k=1}^d \partial^2_{x_j,x_k}h_\theta(s,\tilde{X}_s^{t,x,\mu}(\theta),\mathbb{P}_{\tilde{X}_s^{t,\xi}(\theta)})\partial_\mu^i(\tilde{X}_s^{t,x,\mu}(\theta))^j\partial_{\theta_l}(\tilde{X}_s^{t,x,\mu}(\theta))^k\\
&\qquad+\sum_{j,k=1}^d \widehat{\mathbb{E}}\Big[[\partial_\mu\partial_{x_j}h_\theta(s,\tilde{X}_s^{t,x,\mu}(\theta),\mathbb{P}_{\tilde{X}_s^{t,\xi}(\theta)})]_k(\widehat{X}_s^{t,\xi}(\theta))\partial_{\theta_l}(\widehat{X}_s^{t,\xi}(\theta))^k\Big]\partial_\mu^i(\tilde{X}_s^{t,x,\mu}(\theta))^j\\
&\qquad+\sum_{j=1}^d \partial_{x_j}h_\theta(s,\tilde{X}_s^{t,x,\mu}(\theta),\mathbb{P}_{\tilde{X}_s^{t,\xi}(\theta)})\partial_{\theta_l}\partial_\mu^i(\tilde{X}_s^{t,x,\mu}(\theta))^j\\
&\qquad+\sum_{j=1}^d \widehat{\mathbb{E}}[\partial_{\theta_l}\partial_\mu^j h_\theta(s,\tilde{X}_s^{t,x,\mu}(\theta),\mathbb{P}_{\tilde{X}_s^{t,\xi}(\theta)})(\widehat{X}_s^{t,v,\mu}(\theta))\partial_{v_i}(\widehat{X}_s^{t,v,\mu}(\theta))^j]\\
&\qquad+\sum_{j,k=1}^d \widehat{\mathbb{E}}[\partial_{x_k}\partial_\mu^j h_\theta(s,\tilde{X}_s^{t,x,\mu}(\theta),\mathbb{P}_{\tilde{X}_s^{t,\xi}(\theta)})(\widehat{X}_s^{t,v,\mu}(\theta))\partial_{v_i}(\widehat{X}_s^{t,v,\mu}(\theta))^j]\partial_{\theta_l}(\tilde{X}_s^{t,x,\mu})^k\\
&\qquad+\sum_{j,k=1}^d \widehat{\mathbb{E}}\check{\mathbb{E}}[\partial_\mu^k\partial_\mu^j h_\theta(s,\tilde{X}_s^{t,x,\mu}(\theta),\mathbb{P}_{\tilde{X}_s^{t,\xi}(\theta)})(\widehat{X}_s^{t,v,\mu}(\theta),\check{X}_s^{t,\xi}(\theta))\partial_{v_i}(\widehat{X}_s^{t,v,\mu}(\theta))^j\partial_{\theta_l}(\check{X}_s^{t,\xi}(\theta))^k]\\
&\qquad+\sum_{j,k=1}^d \widehat{\mathbb{E}}[\partial_{v_k}\partial_\mu^j h_\theta(s,\tilde{X}_s^{t,x,\mu}(\theta),\mathbb{P}_{\tilde{X}_s^{t,\xi}(\theta)})(\widehat{X}_s^{t,v,\mu}(\theta))\partial_{v_i}(\widehat{X}_s^{t,v,\mu}(\theta))^j\partial_{\theta_l}(\widehat{X}_s^{t,v,\mu}(\theta))^k]\\
&\qquad+\sum_{j=1}^d \widehat{\mathbb{E}}[\partial_\mu^j h_\theta(s,\tilde{X}_s^{t,x,\mu}(\theta),\mathbb{P}_{\tilde{X}_s^{t,\xi}(\theta)})(\widehat{X}_s^{t,v,\mu}(\theta))\partial_{\theta_l}\partial_{v_i}(\widehat{X}_s^{t,v,\mu}(\theta))^j]\\
&\qquad+\sum_{j=1}^d \int_{\mathbb{R}^d}\widehat{\mathbb{E}}[\partial_{\theta_l}\partial_\mu^j h_\theta(s,\tilde{X}_s^{t,x,\mu}(\theta),\mathbb{P}_{\tilde{X}_s^{t,\xi}(\theta)})(\widehat{X}_s^{t,x',\mu}(\theta))\partial_\mu^i(\widehat{X}_s^{t,x',\mu}(\theta))^j(v)]\,\mu(\mathrm{d}x')\\
&\qquad+\sum_{j,k=1}^d \int_{\mathbb{R}^d}\widehat{\mathbb{E}}[\partial_{x_k}\partial_\mu^j h_\theta(s,\tilde{X}_s^{t,x,\mu}(\theta),\mathbb{P}_{\tilde{X}_s^{t,\xi}(\theta)})(\widehat{X}_s^{t,x',\mu}(\theta))\partial_{\theta_l}(\tilde{X}_s^{t,x,\mu}(\theta))^k\partial_\mu^i(\widehat{X}_s^{t,x',\mu}(\theta))^j(v)]\,\mu(\mathrm{d}x')\\
&\qquad+\sum_{j,k=1}^d \int_{\mathbb{R}^d}\widehat{\mathbb{E}}\check{\mathbb{E}}[\partial_\mu^k\partial_\mu^j h_\theta(s,\tilde{X}_s^{t,x,\mu}(\theta),\mathbb{P}_{\tilde{X}_s^{t,\xi}(\theta)})(\widehat{X}_s^{t,x',\mu}(\theta),\check{X}_s^{t,\xi}(\theta))\partial_\mu^i(\widehat{X}_s^{t,x',\mu}(\theta))^j(v)\partial_{\theta_l}(\check{X}_s^{t,\xi}(\theta))^k]\,\mu(\mathrm{d}x')\\
&\qquad+\sum_{j,k=1}^d \int_{\mathbb{R}^d}\widehat{\mathbb{E}}[\partial_{v_k}\partial_\mu^j h_\theta(s,\tilde{X}_s^{t,x,\mu}(\theta),\mathbb{P}_{\tilde{X}_s^{t,\xi}(\theta)})(\widehat{X}_s^{t,x',\mu}(\theta))\partial_{\theta_l}(\widehat{X}_s^{t,x',\mu}(\theta))^k\partial_\mu^i(\widehat{X}_s^{t,x',\mu}(\theta))^j(v)]\,\mu(\mathrm{d}x')\\
&\qquad+\sum_{j=1}^d \int_{\mathbb{R}^d}\widehat{\mathbb{E}}[\partial_\mu^j h_\theta(s,\tilde{X}_s^{t,x,\mu}(\theta),\mathbb{P}_{\tilde{X}_s^{t,\xi}(\theta)})(\widehat{X}_s^{t,x',\mu}(\theta))\partial_{\theta_l}\partial_\mu^i(\widehat{X}_s^{t,x',\mu}(\theta))^j(v)]\,\mu(\mathrm{d}x').
\end{aligned}
$$

In the above identity, $\check{X}_s^{t,\xi}(\theta)$ stands for a random variable independent of $(\tilde{X}_s^{t,x,\mu},\widehat{X}_s^{t,v,\mu},\widehat{X}_s^{t,x',\mu})$

with the same law as $\tilde{X}_s^{t,\xi}$.

Then, starting from the expression of V_θ in (A.3) (with $\pi = \pi_\theta$), the dominated convergence theorem guarantees that the derivatives $(t,\theta,x,\mu,v) \mapsto \partial_\theta \partial_x \mathrm{V}_\theta(t,x,\mu)$, $\partial_x \partial_\theta \mathrm{V}_\theta(t,x,\mu)$, $\partial_\theta \partial_x^2 \mathrm{V}_\theta(t,x,\mu)$, $\partial_x^2 \partial_\theta \mathrm{V}_\theta(t,x,\mu)$, $\partial_\theta \partial_\mu \mathrm{V}_\theta(t,x,\mu)(v)$, $\partial_\mu \partial_\theta \mathrm{V}_\theta(t,x,\mu)(v)$, $\partial_\theta \partial_v \partial_\mu \mathrm{V}_\theta(t,x,\mu)(v)$, $\partial_v \partial_\mu \partial_\theta \mathrm{V}_\theta(t,x,\mu)(v)$ exist and are locally Lipschitz continuous. Hence, from Clairaut's theorem, we deduce that $\partial_\theta \partial_x \mathrm{V}_\theta(t,x,\mu) = \partial_x \partial_\theta \mathrm{V}_\theta(t,x,\mu)$, $\partial_\theta \partial_x^2 \mathrm{V}_\theta(t,x,\mu) = \partial_x^2 \partial_\theta \mathrm{V}_\theta(t,x,\mu)$, $\partial_\theta \partial_\mu \mathrm{V}_\theta(t,x,\mu)(v) = \partial_\mu \partial_\theta \mathrm{V}_\theta(t,x,\mu)(v)$ and $\partial_\theta \partial_v \partial_\mu \mathrm{V}_\theta(t,x,\mu)(v) = \partial_v \partial_\mu \partial_\theta \mathrm{V}_\theta(t,x,\mu)(v)$ for all t,x,μ,θ,v.

Moreover, from Assumption 3.1 and Lemma A.1, there exist q and C such that for any t,x,μ,v and any $\theta \in \mathcal{K}$, $\mathcal{K}$ being a compact subset of Θ

$$|\partial_\theta \mathrm{V}_\theta(t,x,\mu)| \leq C(1+|x|^2 + M_2(\mu)^q), \tag{A.8}$$

$$|\partial_\theta \partial_x \mathrm{V}_\theta(t,x,\mu)| + |\partial_\theta \partial_\mu \mathrm{V}_\theta(t,x,\mu)(v)| \leq C(1+|x|+|v|+M_2(\mu)^q), \tag{A.9}$$

and

$$|\partial_\theta \partial_x^2 \mathrm{V}_\theta(t,x,\mu)| + |\partial_\theta \partial_v \partial_\mu \mathrm{V}_\theta(t,x,\mu)(v)| \leq C(1+|v|+M_2(\mu)^q). \tag{A.10}$$

Now, differentiating with respect to θ both sides of (2.6), we deduce that $\theta \mapsto \partial_t \mathrm{V}_\theta(t,x,\mu)$ is differentiable with a derivative $\partial_\theta \partial_t \mathrm{V}_\theta(t,x,\mu)$ being continuous with respect to t,x,μ,θ. Also, taking $\pi = \pi_\theta$ and differentiating with respect to θ both sides of the identity of (A.7) (using Lemma A.1 together with the estimates (A.8), (A.9), (A.10) and the dominated convergence theorem to differentiate the right-hand side therein) and then passing to the limit as $h \downarrow 0$, we get that $t \mapsto \partial_\theta \mathrm{V}_\theta(t,x,\mu)$ is differentiable with a derivative $\partial_t \partial_\theta \mathrm{V}_\theta(t,x,\mu)$ being continuous with respect to t,x,μ,θ. We thus conclude that the two derivatives $\partial_\theta \partial_t \mathrm{V}_\theta(t,x,\mu)$ and $\partial_t \partial_\theta \mathrm{V}_\theta(t,x,\mu)$ coincide for all t,x,μ,θ.

A.3 Proof of Theorem 3.1

Step 1: We start from the PDE characterisation of V_θ in Proposition 2.1 that we write as

$$\int_A \left\{ \mathcal{L}_\theta^a[\mathrm{V}_\theta](t,x,\mu) + f(x,\mu,a) + \lambda \log p_\theta(t,x,\mu,a) \right\} \pi_\theta(\mathrm{d}a|t,x,\mu) = 0, \tag{A.11}$$

where

$$\begin{aligned}\mathcal{L}_\theta^a[\varphi](t,x,\mu) &= -\beta\varphi(t,x,\mu) + \partial_t\varphi(t,x,\mu) + b(x,\mu,a)\cdot D_x\varphi(t,x,\mu) + \frac{1}{2}\sigma\sigma^\intercal(x,\mu,a) : D_x^2\varphi(t,x,\mu) \\ &\quad + \mathbb{E}_{\xi\sim\mu}\Big[b_\theta(t,\xi,\mu)\cdot\partial_\mu\varphi(t,x,\mu)(\xi) + \frac{1}{2}\Sigma_\theta(t,\xi,\mu) : \partial_v\partial_\mu\varphi(t,x,\mu)(\xi)\Big],\end{aligned}$$

recalling that $b_\theta(t,x,\mu) = \int_A b(x,\mu,a)\,\pi_\theta(\mathrm{d}a|t,x,\mu)$, $\Sigma_\theta(t,x,\mu) = \int_A (\sigma\sigma^\intercal)(x,\mu,a)\,\pi_\theta(\mathrm{d}a|t,x,\mu)$.

For any fixed t,x,μ, we now differentiate w.r.t. $\theta \in \Theta$ both sides of (A.11) to get a new system of linear PDEs satisfied by G_θ. In particular, using the identity

$$\nabla_\theta \Big[\mathcal{L}_\theta^a[\mathrm{V}_\theta](t,x,\mu) \Big] = \mathcal{L}_\theta^a[\mathrm{G}_\theta](t,x,\mu) + \mathcal{H}_\theta[\mathrm{V}_\theta](t,x,\mu),$$

together with (3.1) and the dominated convergence theorem, we get

$$\int_A \Big\{ \mathcal{L}_\theta^a[\mathrm{G}_\theta](t,x,\mu) + \mathcal{H}_\theta[\mathrm{V}_\theta](t,x,\mu) \\ + \big[\mathcal{L}_\theta^a[\mathrm{V}_\theta](t,x,\mu) + f(x,\mu,a) + \lambda \log p_\theta(t,x,\mu,a)\big] \nabla_\theta \log p_\theta(t,x,\mu,a) \Big\} \pi_\theta(\mathrm{d}a|t,x,\mu) = 0, \tag{A.12}$$

with terminal condition $\mathrm{G}_\theta(T,x,\mu) = 0$. Note that we have used the fact that

$$\int_A \nabla_\theta \log p_\theta(t,x,\mu,a) \pi_\theta(\mathrm{d}a|t,x,\mu) \;=\; \nabla_\theta \int_A \pi_\theta(\mathrm{d}a|t,x,\mu) \;=\; 0,$$

and the above PDE is a system of D equations, where $\mathcal{L}_\theta^a[\mathrm{G}_\theta]$ denotes the operator applied to each component of the $\mathbb{R}^D$-valued function G_θ.

Step 2: Denote by

$$\tilde{F}_\theta(t,x,\mu,a) = \big\{ \mathcal{L}_\theta^a[\mathrm{V}_\theta](t,x,\mu) + f(x,\mu,a) + \lambda \log p_\theta(t,x,\mu,a) \big\} \nabla_\theta \log p_\theta(t,x,\mu,a) \\ + \; \mathcal{H}_\theta[\mathrm{V}_\theta](t,x,\mu),$$

and

$$\tilde{f}_{\pi_\theta}(t,x,\mu) \;=\; \int_A \tilde{F}_\theta(t,x,\mu,a)\, \pi_\theta(\mathrm{d}a|t,x,\mu),$$

so that the linear PDE (A.12) satisfied by G_θ now writes

$$\mathcal{L}_{\pi_\theta}[\mathrm{G}_\theta](t,x,\mu) + \tilde{f}_{\pi_\theta}(t,x,\mu) \;=\; 0,$$

with terminal condition $\mathrm{G}_\theta(T,x,\mu) = 0$. Observe that the above PDE is similar to (2.6). In order to obtain the announced probabilistic representation formula, we first apply the chain rule formula on the strip $[t,T] \times \mathbb{R}^d \times \mathcal{P}_2(\mathbb{R}^d)$, see e.g. Proposition 5.102 in [4], to $(e^{-\beta s}\mathrm{G}_\theta(s, \widetilde{X}_s^{t,x,\mu}, \mathbb{P}_{\widetilde{X}_s^{t,\xi}}))_{s\in[t,T]}$ using the estimates (A.8) and (A.10). We thus obtain

$$d(e^{-\beta s}\mathrm{G}_\theta(s, \widetilde{X}_s^{t,x,\mu}, \mathbb{P}_{\widetilde{X}_s^{t,\xi}})) = -e^{-\beta s} \tilde{f}_{\pi_\theta}(s, \widetilde{X}_s^{t,x,\mu}, \mathbb{P}_{\widetilde{X}_s^{t,\xi}})\, \mathrm{d}s \\ + e^{-\beta s} \partial_x \mathrm{G}_\theta(s, \widetilde{X}_s^{t,x,\mu}, \mathbb{P}_{\widetilde{X}_s^{t,\xi}})^\intercal \sigma_{\pi_\theta}(\widetilde{X}_s^{t,x,\mu}, \mathbb{P}_{\widetilde{X}_s^{t,\xi}})\, \mathrm{d}W_s.$$

Observe that (A.9) together with the fact that for any $\theta \in \mathbb{R}^D$, $|\sigma_{\pi_\theta}(x,\mu)| \le C(1 + |x| + M_2(\mu))$, for some constant C, directly yields that the stochastic integral is a square integrable martingale. Hence, integrating from t to T both sides of the above and using the facts that $\mathrm{G}_\theta(T,x,\mu) = 0$ and $\mathbb{P}_{\widetilde{X}_s^{t,\xi}} = \mathbb{P}_{X_s^{t,\xi}}$, $\mathbb{P}_{\widetilde{X}_s^{t,x,\mu}} = \mathbb{P}_{X_s^{t,x,\mu}}$, we eventually deduce

$$\mathrm{G}_\theta(t,x,\mu) = \; \mathbb{E}_{\alpha\sim\pi_\theta}\Big[\int_t^T e^{-\beta(s-t)} \tilde{F}_\theta(s, X_s^{t,x,\mu}, \mathbb{P}_{X_s^{t,\xi}}, \alpha_s)\, \mathrm{d}s \Big].$$

Step 3: On the other hand, applying again the chain rule formula to $\mathrm{V}_\theta(s, X_s^{t,x,\mu}, \mathbb{P}_{X_s^{t,\xi}})$, when $\alpha \sim \pi_\theta$, see e.g. Proposition 5.102 in [4], we have

$$\begin{aligned}\mathrm{d}\mathrm{V}_\theta(s, X_s^{t,x,\mu}, \mathbb{P}_{X_s^{t,\xi}}) &= \big(\mathcal{L}_\theta^{\alpha_s}[\mathrm{V}_\theta](s, X_s^{t,x,\mu}, \mathbb{P}_{X_s^{t,\xi}}) + \beta \mathrm{V}_\theta(s, X_s^{t,x,\mu}, \mathbb{P}_{X_s^{t,\xi}})\big)\,\mathrm{d}s \\ &\quad + D_x\mathrm{V}_\theta(s, X_s^{t,x,\mu}, \mathbb{P}_{X_s^{t,\xi}})^\intercal \sigma(X_s^{t,x,\mu}, \mathbb{P}_{X_s^{t,\xi}}, \alpha_s)\,\mathrm{d}W_s, \quad t \le s \le T,\end{aligned}$$

and thus by definition of $\tilde{F}_\theta$

$$\begin{aligned}&\int_t^T e^{-\beta(s-t)} \tilde{F}_\theta(s, X_s^{t,x,\mu}, \mathbb{P}_{X_s^{t,\xi}}, \alpha_s)\,\mathrm{d}s \\ &= \int_t^T e^{-\beta(s-t)} \nabla_\theta \log(p_\theta(s, X_s^{t,x,\mu}, \mathbb{P}_{X_s^{t,\xi}}, \alpha_s))\Big(\mathrm{d}\mathrm{V}_\theta(s, X_s^{t,x,\mu}, \mathbb{P}_{X_s^{t,\xi}}) - \beta \mathrm{V}_\theta(s, X_s^{t,x,\mu}, \mathbb{P}_{X_s^{t,\xi}}) \\ &\quad + \mathcal{H}_\theta[\mathrm{V}_\theta](s, X_s^{t,x,\mu}, \mathbb{P}_{X_s^{t,\xi}})\Big)\,\mathrm{d}s \\ &+ \int_t^T e^{-\beta(s-t)} \nabla_\theta \log(p_\theta(s, X_s^{t,x,\mu}, \mathbb{P}_{X_s^{t,\xi}}, \alpha_s))\Big(f(X_s^{t,x,\mu}, \mathbb{P}_{X_s^{t,\xi}}, \alpha_s) + \lambda \log(p_\theta(s, X_s^{t,x,\mu}, \mathbb{P}_{X_s^{t,\xi}}, \alpha_s))\Big)\,\mathrm{d}s \\ &- \int_t^T e^{-\beta(s-t)} \nabla_\theta \log(p_\theta(s, X_s^{t,x,\mu}, \mathbb{P}_{X_s^{t,\xi}}, \alpha_s))\, D_x\mathrm{V}_\theta(s, X_s^{t,x,\mu}, \mathbb{P}_{X_s^{t,\xi}})^\intercal \sigma(X_s^{t,x,\mu}, \mathbb{P}_{X_s^{t,\xi}}, \alpha_s)\,\mathrm{d}W_s.\end{aligned}$$

Note that (3.2) as well as the bound $|D_x V_\theta(s,x,\mu)| \le C(1 + |x| + |\mu|^q)$, for some $q \ge 0$, directly deduced from the identity (A.4) and Assumption 2.1, guarantees that the stochastic integral appearing in the right-hand side of the above identity is a square integrable martingale. Hence, taking expectation in both sides of the above identity eventually yields

$$\begin{aligned}G_\theta(t,x,\mu) := \mathbb{E}_{\alpha\sim\pi_\theta}\Big[\int_t^T e^{-\beta(s-t)} \nabla_\theta \log p_\theta(s, X_s^{t,x,\mu}, \mathbb{P}_{X_s^{t,\xi}}, \alpha_s)\Big\{\mathrm{d}\mathrm{V}_\theta(s, X_s^{t,x,\mu}, \mathbb{P}_{X_s^{t,\xi}}) \\ + \big[f(X_s^{t,x,\mu}, \mathbb{P}_{X_s^{t,\xi}}, \alpha_s) + \lambda \log p_\theta(s, X_s^{t,x,\mu}, \mathbb{P}_{X_s^{t,\xi}}, \alpha_s) - \beta \mathrm{V}_\theta(s, X_s^{t,x,\mu}, \mathbb{P}_{X_s^{t,\xi}})\big]\mathrm{d}s\Big\} \\ + \int_t^T e^{-\beta(s-t)} \mathcal{H}_\theta[\mathrm{V}_\theta](s, X_s^{t,x,\mu}, \mathbb{P}_{X_s^{t,\xi}})\mathrm{d}s\Big].\end{aligned}$$

This proves the announced probabilistic representation formula for G_θ.

B Linear quadratic mean-field control with randomised controls and entropy regularisation

A stochastic policy is a probability transition kernel from $[0,T]\times\mathbb{R}^d\times\mathcal{P}_2(\mathbb{R}^d)$ into $A = \mathbb{R}^m$, i.e., a measurable function $\pi : (t,x,\mu) \in [0,T]\times\mathbb{R}^d\times\mathcal{P}_2(\mathbb{R}^d) \mapsto \pi(.|t,x,\mu) \in \mathcal{P}(\mathbb{R}^m)$. We denote by Π the set of stochastic policies π with densities p with respect to the Lebesgue measure on $\mathbb{R}^m$: $\pi(\mathrm{d}a|t,x,\mu) = p(t,x,\mu,a)\mathrm{d}a$. We say that the process $\alpha = (\alpha_t)_t$ is a randomised

feedback control generated from a stochastic policy $\pi \in \Pi$, denoted by $\alpha \sim \pi$, if at each time t, the action α_t is sampled (according to the σ-algebra $\mathcal{G}$) from the probability distribution $\pi(.|t, X_t, \mathbb{P}_{X_t})$. The dynamics $X = X^\alpha$ follows a linear mean-field dynamics with coefficients $b(x, \mu, a) = \bar{b}(x, \bar{\mu}, a)$, $\sigma(x, \mu, a) = \bar{\sigma}(x, \bar{\mu}, a)$ in the form

$$\bar{b}(x, \bar{x}, a) \ = \ Bx + \bar{B}\bar{x} + Ca, \quad \bar{\sigma}(x, \bar{x}, a) \ = \ \gamma + Dx + \bar{D}\bar{x} + Fa,$$

for $(x, \mu, \bar{x}, a) \in \mathbb{R}^d \times \mathcal{P}_2(\mathbb{R}^d) \times \mathbb{R}^d \times \mathbb{R}^m$, where we denote by $\bar{\mu} = \int x\mu(\mathrm{d}x)$, B, $\bar{B}$, D, $\bar{D}$ are constant matrices in $\mathbb{R}^{d\times d}$, C, F are constant matrices in $\mathbb{R}^{d\times m}$, γ is a constant in $\mathbb{R}^d$.

Given a stochastic policy $\pi \in \Pi$, we consider the functional cost V^π with entropy regulariser defined in (2.5) with quadratic functions $f(x, \mu, a) = \bar{f}(x, \bar{\mu}, a)$ and $g(x, \mu) = \bar{g}(x, \bar{\mu})$:

$$\begin{aligned}\bar{f}(x, \bar{x}, a) \ &= \ x^\intercal Q x + \bar{x}^\intercal \bar{Q}\bar{x} + a^\intercal N a + 2a^\intercal I x + 2a^\intercal \bar{I}\bar{x} + 2M.x + 2H.a,\\ \bar{g}(x, \bar{x}) \ &= \ x^\intercal P x + \bar{x}^\intercal \bar{P}\bar{x} + 2L.x,\end{aligned}$$

where N is a symmetric matrix in $\mathbb{S}^m_+$, I, $\bar{I} \in \mathbb{R}^{m\times d}$, Q, $\bar{Q}$, P, $\bar{P}$ are symmetric matrices in $\mathbb{S}^d$, M, $L \in \mathbb{R}^d$, $H \in \mathbb{R}^m$, assumed to satisfy the conditions:

(H1) (i) There exists $\delta > 0$ s.t.

$$N \ \geq \ \delta I_m, \quad P \ \geq \ 0, \quad Q - I^\intercal N^{-1} I \ \geq \ 0.$$

or (ii) $n = m = 1$, $I = 0$, $F \neq 0$, $Q \geq 0$, $P > 0$.

(H2) (i) There exists $\delta > 0$ s.t.

$$N \ \geq \ \delta I_m, \quad P + \bar{P} \ \geq \ 0, \quad (Q + \bar{Q}) - (I + \bar{I})^\intercal N^{-1}(I + \bar{I}) \ \geq \ 0.$$

or (ii) $I + \bar{I} = 0$, $F \neq 0$, $Q + \bar{Q} \geq 0$, $P + \bar{P} \geq 0$, $P > 0$.

The solution to the LQ mean-field control problem with entropy regulariser is then given by the following theorem:

Theorem B.1 *Let Assumptions* **(H1)**-**(H2)** *hold. Then, the value function is equal to*

$$v(t, x, \mu) \ := \ \inf_{\pi\in\Pi} V^\pi(t, x, \mu) \ \ = \ \ (x - \bar{\mu})^\intercal K(t)(x - \bar{\mu}) + \bar{\mu}^\intercal \Lambda(t)\bar{\mu} + 2Y(t)^\intercal x + R(t),$$

for $(t, x, \mu) \in [0, T] \times \mathbb{R}^d \times \mathcal{P}_2(\mathbb{R}^d)$, *where the quadruple* (K, Λ, Y, R) *valued in* $(\mathbb{S}^d_+, \mathbb{S}^d_+, \mathbb{R}^d, \mathbb{R})$

is solution on $[0,T]$ *to the system of Riccati equations:*

$$\left\{\begin{array}{l} \dot{K}(t) - \beta K(t) + Q + K(t)B + B^\intercal K(t) + D^\intercal K(t)D \\ -(I + C^\intercal K(t) + F^\intercal K(t)D)^\intercal (N + F^\intercal K(t)F)^{-1}(I + C^\intercal K(t) + F^\intercal K(t)D) \;=\; 0, \\ \dot{\Lambda}(t) - \beta\Lambda(t) + \hat{Q} + \Lambda(t)\hat{B} + \hat{B}^\intercal\Lambda(t) + \hat{D}^\intercal K(t)\hat{D} \\ -\big(\hat{I} + C^\intercal\Lambda(t) + F^\intercal K(t)\hat{D}\big)^\intercal (N + F^\intercal K(t)F)^{-1}\big(\hat{I} + C^\intercal\Lambda(t) + F^\intercal K(t)\hat{D}\big) \;=\; 0 \\ \dot{Y}(t) - \beta Y(t) + M + \hat{B}^\intercal Y(t) + \hat{D}^\intercal K(t)\gamma \\ -(\hat{I} + C^\intercal\Lambda(t) + F^\intercal K(t)\hat{D})^\intercal (N + F^\intercal K(t)F)^{-1}(H + C^\intercal Y(t) + F^\intercal K(t)\gamma) \;=\; 0 \\ \dot{R}(t) - \beta R(t) + \gamma^\intercal K(t)\gamma + \frac{\lambda m}{2}\log(2\pi) - \frac{\lambda}{2}\log|\frac{\lambda}{2det(N+F^\intercal K(t)F)}| \\ -(H + C^\intercal Y(t) + F^\intercal K(t)\gamma)^\intercal (N + F^\intercal K(t)F)^{-1}(H + C^\intercal Y(t) + F^\intercal K(t)\gamma) \;=\; 0 \end{array}\right. \tag{B.1}$$

with the terminal condition $(K(T),\Lambda(T),Y(T),R(T)) = (P,\hat{P},L,0)$*, where we set* $\hat{I} := I + \bar{I}$*,* $\hat{B} := B + \bar{B}$*,* $\hat{D} := D + \bar{D}$*,* $\hat{Q} := Q + \bar{Q}$*,* $\hat{P} := P + \bar{P}$*.*

Moreover, the optimal stochastic policy follows a Gaussian distribution:

$$\pi^*(.|t,x,\mu) = \mathcal{N}\Big(-S(t)^{-1}\big(U(t)x + (\hat{U}(t) - U(t))\bar{\mu} + O(t)\big); \frac{\lambda}{2}S(t)^{-1}\Big), \tag{B.2}$$

where we set

$$\begin{array}{rlrl} S(t) &:=\; N + F^\intercal K(t)F, & O(t) &:=\; H + C^\intercal Y(t) + F^\intercal K(t)\gamma \\ U(t) &:=\; I + C^\intercal K(t) + F^\intercal K(t)D, & \hat{U}(t) &:=\; \hat{I} + C^\intercal\Lambda(t) + F^\intercal K(t)\hat{D}. \end{array}$$

Remark B.1 *Conditions* **(H1)** *and* **(H2)** *ensure the existence and uniqueness of a solution* (K,Λ) *to the matrix Riccati equation in* (B.1) *satisfying* $K \geq 0$*,* $\Lambda \geq 0$ *(hence* $S(t)^{-1}$ *is well-defined). Given* (K,Λ)*, the equations for* (Y,R) *are simply linear ODEs.*

Proof of Theorem B.1. We adapt the arguments in [2] to our case with randomised controls and entropy regulariser.

Step 1. Let us consider the function defined on $[0,T]\times\mathbb{R}^d\times\mathcal{P}_2(\mathbb{R}^d)$ by $w(t,x,\mu) = \bar{w}(t,x,\bar{\mu})$, where $\bar{w}$ is defined on $[0,T]\times\mathbb{R}^d\times\mathbb{R}^d$ by

$$\bar{w}(t,x,\bar{x}) \;=\; (x-\bar{x})^\intercal K(t)(x-\bar{x}) + \bar{x}^\intercal\Lambda(t)\bar{x} + 2Y(t)^\intercal x + R(t),$$

for some functions (to be determined later) K, Λ, Y and R on $[0,T]$, and valued on $\mathbb{S}^d_+$, $\mathbb{S}^d_+$, $\mathbb{R}^d$, and $\mathbb{R}$. Fix $(t_0,x_0,\mu_0)\in[0,T]\times\mathbb{R}^d\times\mathcal{P}_2(\mathbb{R}^d)$, and $\xi_0\in L^2(\mathcal{F}_{t_0};\mathbb{R}^d)\sim\mu_0$. Given $\pi\in\Pi$ with density p, and a randomised control $\alpha\sim\pi$, we consider the process

$$\begin{array}{rcl} \mathcal{S}^\alpha_t &:=& e^{-\beta(t-t_0)}\bar{w}(t,X^{t_0,x_0,\mu_0}_t,\bar{X}^{t_0,\mu_0}_t) + \displaystyle\int_{t_0}^t e^{-\beta(s-t_0)}\big[\bar{f}(X^{t_0,x_0,\mu_0}_s,\bar{X}^{t_0,\mu_0}_s,\alpha_s) \\ && \qquad + \lambda\displaystyle\int_{\mathbb{R}^m}\big(\log \mathrm{p}_t(a)\big)\mathrm{p}_t(a)\mathrm{d}a\big]\mathrm{d}s, \end{array}$$

for $t_0 \le t \le T$, where we set $\mathrm{p}_t(a) = p(t, X_t^{t_0,x_0,\mu_0}, \mathbb{P}_{X_t^{t_0,\xi_0}}, a)$, and $\bar{X}_t^{t_0,\mu_0} := \mathbb{E}_{\alpha\sim\pi}[X_t^{t_0,\xi_0}]$ which follows the dynamics:

$$d\bar{X}_t \;=\; \big(\hat{B}\bar{X}_t + C\bar{\alpha}_t)\mathrm{d}t,$$

with $\bar{\alpha}_t := \mathbb{E}_{\alpha\sim\pi}[\alpha_t]$.

Step 2. We apply Itô's formula to $\mathcal{S}_t^\alpha$ for $\alpha \sim \pi$, and take the expectation to get

$$\mathrm{d}\mathbb{E}_{\alpha\sim\pi}[\mathcal{S}_t^\alpha] = \; e^{-\beta(t-t_0)}\mathbb{E}_{\alpha\sim\pi}[\mathcal{D}_t^\alpha]\mathrm{d}t, \tag{B.3}$$

with

$$\mathcal{D}_t^\alpha \;=\; -\beta\bar{w}(t,X_t,\bar{X}_t) + \frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}_{\alpha\sim\pi}[\bar{w}(t,X_t,\bar{X}_t)] + \bar{f}(X_t,\bar{X}_t,\alpha_t) + \lambda\int_{\mathbb{R}^m}(\log \mathrm{p}_t(a))\mathrm{p}_t(a)\mathrm{d}a,$$

where we omit the dependence on t_0, x_0, μ_0 of X and $\bar{X}$ to alleviate notations. By applying Itô's formula to $\bar{w}(t, X_t, \bar{X}_t)$, recalling the quadratic forms of $\bar{w}$, $\bar{f}$, and using the linear dynamics of X and $\bar{X}$, we obtain similarly as in [2] (after careful but straightforward computations):

$$\begin{aligned}
\mathbb{E}_{\alpha\sim\pi}[\mathcal{D}_t^\alpha] \;&= \mathbb{E}_{\alpha\sim\pi}\Big[(X_t-\bar{X}_t)^\intercal\big(\dot{K}(t) - \beta K(t) + Q + K(t)B + B^\intercal K(t) + D^\intercal K(t)D\big)(X_t-\bar{X}_t)\\
&\quad+\; \bar{X}_t^\intercal\big(\dot{\Lambda}(t) - \beta\Lambda(t) + \hat{Q} + \Lambda(t)\hat{B} + \hat{B}^\intercal\Lambda(t) + \hat{D}^\intercal K(t)\hat{D}\big)\bar{X}_t\\
&\quad+\; 2\big(\dot{Y}(t) - \beta Y(t) + M + \hat{B}^\intercal Y(t) + \hat{D}^\intercal K(t)\gamma\big)^\intercal X_t + \dot{R}(t) - \beta R(t) + \gamma^\intercal K(t)\gamma\\
&\quad+\; \alpha_t^\intercal S(t)\alpha_t + 2\alpha_t^\intercal\big(U(t)(X_t-\bar{X}_t) + \hat{U}(t)\bar{X}_t + O(t)\big) + \lambda\int_{\mathbb{R}^m}(\log \mathrm{p}_t(a))\mathrm{p}_t(a)\mathrm{d}a\Big]\\
&= \mathbb{E}_{\alpha\sim\pi}\Big[(X_t-\bar{X}_t)^\intercal\big(\dot{K}(t) - \beta K(t) + Q + K(t)B + B^\intercal K(t) + D^\intercal K(t)D\big)(X_t-\bar{X}_t)\\
&\quad+\; \bar{X}_t^\intercal\big(\dot{\Lambda}(t) - \beta\Lambda(t) + \hat{Q} + \Lambda(t)\hat{B} + \hat{B}^\intercal\Lambda(t) + \hat{D}^\intercal K(t)\hat{D}\big)\bar{X}_t\\
&\quad+\; 2\big(\dot{Y}(t) - \beta Y(t) + M + \hat{B}^\intercal Y(t) + \hat{D}^\intercal K(t)\gamma\big)^\intercal X_t + \dot{R}(t) - \beta R(t) + \gamma^\intercal K(t)\gamma\\
&\quad+\; \int_{\mathbb{R}^m}[\phi_t(a) + \lambda\log \mathrm{p}_t(a)]\mathrm{p}_t(a)\mathrm{d}a\Big],
\end{aligned} \tag{B.4}$$

where we used in the last equality the fact that $\alpha \sim \pi$, and set $\phi_t(a) := a^\intercal S(t)a + 2a^\intercal\chi_t$ with $\chi_t := U(t)(X_t - \bar{X}_t) + \hat{U}(t)\bar{X}_t + O(t)$.

Step 3. Let ϕ be a quadratic function on $\mathbb{R}^m$: $\phi(a) = a^\intercal Sa + 2a^\intercal\chi$ for some positive-definite matrix $S \in \mathbb{S}_+^m$, and $\chi \in \mathbb{R}^m$, and denote by $\mathcal{D}_2(\mathbb{R}^m)$ the set of square integrable density functions on $\mathbb{R}^m$, i.e., the set of nonnegative measurable functions p on $\mathbb{R}^m$ s.t. $\int_{\mathbb{R}^m}\mathrm{p}(a)\mathrm{d}a = 1$, and $\int_{\mathbb{R}^m}|a|^2\mathrm{p}(a)\mathrm{d}a < \infty$. Let us consider the cost functional on $\mathcal{D}_2(\mathbb{R}^m)$ defined by

$$C_\phi(\mathrm{p}) \;:=\; \int_{\mathbb{R}^m}[\phi(a) + \lambda\log \mathrm{p}(a)]\mathrm{p}(a)\mathrm{d}a.$$

Then, the minimizer of C_ϕ is achieved with $\mathrm{p}^* \in \mathcal{D}_2(\mathbb{R}^m)$ given by

$$\mathrm{p}^*(a) = \frac{\exp\big(-\frac{1}{\lambda}\phi(a)\big)}{\int_{\mathbb{R}^m} \exp\big(-\frac{1}{\lambda}\phi(a)\big)\mathrm{d}a}, \quad a \in \mathbb{R}^m. \tag{B.5}$$

Indeed, by considering the Lagrangian function associated to this minimization problem

$$L_\phi(\mathrm{p}, \nu) \;=\; C_\phi(\mathrm{p}) - \nu\Big(\int_{\mathbb{R}^m} \mathrm{p}(a)\mathrm{d}a - 1\Big) \;=\; \int_{\mathbb{R}^m} \big[\phi(a) + \lambda \log \mathrm{p}(a) - \nu\big]\mathrm{p}(a)\mathrm{d}a \;+\; \nu,$$

for $(\mathrm{p}, \nu) \in \mathcal{D}_2(\mathbb{R}^m) \times \mathbb{R}$, we see that the minimization over p is obtained pointwisely, i.e. inside the integral over $a \in \mathbb{R}^m$, hence leading to the first-order equations:

$$\begin{cases} \phi(a) + \lambda \log \mathrm{p}^*(a) - \nu^* + \lambda \;=\; 0, \quad a \in \mathbb{R}^m, \\ \int_{a \in \mathbb{R}^m} \mathrm{p}^*(a)\mathrm{d}a \;=\; 1. \end{cases}$$

This yields the expression of p^* in (B.5), which is actually the density of a Gaussian distribution

$$\pi^* = \mathcal{N}\Big(-S^{-1}\chi; \frac{\lambda}{2}S^{-1}\Big). \tag{B.6}$$

The infimum of C_ϕ is then equal to

$$\inf_{\mathrm{p} \in \mathcal{D}_2(\mathbb{R}^m)} C_\phi(\mathrm{p}) = \; C_\phi(\mathrm{p}^*) \;=\; -\chi^\intercal S^{-1}\chi - \frac{\lambda m}{2}\log(2\pi) - \frac{\lambda}{2}\log|\frac{\lambda}{2\mathrm{det}(S)}|. \tag{B.7}$$

Step 4. Notice that under **(H1)**, the matrix $S(t) = N + F^\intercal K(t) F$ is positive-definite for $K \geq 0$, and $\mathrm{p}_t(.) \in \mathcal{D}_2(\mathbb{R}^m)$ a.s. for $t \in [t_0, T]$. From (B.4) and (B.7), we then have for all $\pi \in \Pi$,

$$\begin{aligned}
&\mathbb{E}_{\alpha\sim\pi}[\mathcal{D}_t^\alpha] \\
&= \mathbb{E}_{\alpha\sim\pi}\Big[(X_t - \bar{X}_t)^\intercal\big(\dot{K}(t) - \beta K(t) + Q + K(t)B + B^\intercal K(t) + D^\intercal K(t) D\big)(X_t - \bar{X}_t) \\
&\quad + \; \bar{X}_t^\intercal\big(\dot{\Lambda}(t) - \beta\Lambda(t) + \hat{Q} + \Lambda(t)\hat{B} + \hat{B}^\intercal\Lambda(t) + \hat{D}^\intercal K(t)\hat{D}\big)\bar{X}_t \\
&\quad + \; 2\big(\dot{Y}(t) - \beta Y(t) + M + \hat{B}^\intercal Y(t) + \hat{D}^\intercal K(t)\gamma\big)^\intercal X_t + \dot{R}(t) - \beta R(t) + \gamma^\intercal K(t)\gamma \\
&\quad + \; C_{\phi_t}(\mathrm{p}_t)\Big] \\
&\geq \mathbb{E}_{\alpha\sim\pi}\Big[(X_t - \bar{X}_t)^\intercal\big(\dot{K}(t) - \beta K(t) + Q + K(t)B + B^\intercal K(t) + D^\intercal K(t)D - U(t)^\intercal S(t)^{-1}U(t)\big)(X_t - \bar{X}_t) \\
&\quad + \; \bar{X}_t^\intercal\big(\dot{\Lambda}(t) - \beta\Lambda(t) + \hat{Q} + \Lambda(t)\hat{B} + \hat{B}^\intercal\Lambda(t) + \hat{D}^\intercal K(t)\hat{D} - \hat{U}(t)^\intercal S(t)^{-1}\hat{U}(t)\big)\bar{X}_t \\
&\quad + \; 2\big(\dot{Y}(t) - \beta Y(t) + M + \hat{B}^\intercal Y(t) + \hat{D}^\intercal K(t)\gamma - O(t)^\intercal S(t)^{-1}\hat{U}(t)\big)^\intercal X_t \\
&\quad + \; \dot{R}(t) - \beta R(t) + \gamma^\intercal K(t)\gamma - O(t)^\intercal S(t)^{-1}O(t) - \frac{\lambda m}{2}\log(2\pi) - \frac{\lambda}{2}\log|\frac{\lambda}{2\mathrm{det}(S(t))}|.
\end{aligned} \tag{B.8}$$

Therefore, by taking (K, Λ, Y, R) solution to (B.1), we see that the r.h.s. of (B.8) vanishes, which means that for all $\pi \in \Pi$, $\mathbb{E}_{\alpha\sim\pi}[\mathcal{D}_t^\alpha] \geq 0$. Moreover, from (B.6), the equality in (B.8) holds true for the choice of $\pi^* \in \Pi$ as defined in (B.2), and thus

$$\inf_{\pi\in\Pi} \mathbb{E}_{\alpha\sim\pi}[\mathcal{D}_t^\alpha] \;=\; \mathbb{E}_{\alpha\sim\pi^*}[\mathcal{D}_t^\alpha] \;=\; 0, \quad t \in [t_0, T].$$

From (B.3), this means that the function $t \mapsto \mathbb{E}_{\alpha\sim\pi}[\mathcal{S}_t^\alpha]$ is nondecreasing on $[t_0, T]$ for any $\pi \in \Pi$, and constant on $[t_0, T]$ for $\pi = \pi^*$. By definition of $\mathcal{S}^\alpha$, V^π, and noting that $\bar{w}(T, x, \bar{x}) = \bar{g}(x, \bar{x})$ from the terminal condition on (K, Λ, Y, R), it follows that

$$w(t_0, x_0, \mu_0) \;=\; \bar{w}(t_0, x_0, \bar{\mu}_0) \;=\; \mathbb{E}_{\alpha\sim\pi}[S_{t_0}^\alpha] \leq\; \mathbb{E}_{\alpha\sim\pi}[S_T^\alpha] \;=\; V^\pi(t_0, x_0, \mu_0), \tag{B.9}$$

for any $\pi \in \Pi$, with equality in (B.9) for $\pi = \pi^*$. We conclude that

$$\begin{aligned} \inf_{\pi\in\Pi} V^\pi(t_0, x_0, \mu_0) \;&=\; V^{\pi^*}(t_0, x_0, \mu_0) \;=\; w(t_0, x_0, \mu_0) \\ &=\; (x_0 - \bar{\mu_0})^\intercal K(t_0)(x_0 - \bar{\mu_0}) + \bar{\mu_0}^\intercal \Lambda(t_0)\bar{\mu_0} + 2Y(t_0)^\intercal x_0 + R(t_0). \end{aligned}$$

□

References

[1] A. Angiuli, J-.P. Fouque, and M. Laurière. Unified reinforcement Q-learning for mean field game and control problems. *Mathematics of Control, Signals and Systems*, 34:217–271, 2022.

[2] M. Basei and H. Pham. A Weak Martingale Approach to Linear-Quadratic McKean-Vlasov Stochastic Control Problems. *Journal of Optimization Theory and Applications*, 181(2):347–382, 2019.

[3] R. Carmona and F. Delarue. Forward Backward stochastic differential equations and controlled McKean-Vlasov dynamics. *Annals of Probability*, 43:2647–2700, 2015.

[4] R. Carmona and F. Delarue. *Probabilistic Theory of Mean Field Games: vol. I, Mean Field FBSDEs, Control, and Games, Mean Field game with common noise and Master equations.* Springer, 2018.

[5] R. Carmona and F. Delarue. *Probabilistic Theory of Mean Field Games: vol. II, Mean Field game with common noise and Master equations.* Springer, 2018.

[6] R. Carmona, J.-P. Fouque, and L. Sun. Mean field games and systemic risk. *Commun. Math. Sci.*, 13(4):911–933, 2015.

[7] R. Carmona and M. Laurière. Convergence analysis of machine learning algorithms for the numerical solution of mean-field control and games: II-the finite horizon case. *to appear in Annals of Applied Probability*, 2021.

[8] R. Carmona, M. Laurière, and Z. Tan. Model-free mean-field reinforcement learning: mean-field MDP and mean-field Q-learning. arXiv: 1910.12802v1, 2019.

[9] René Carmona, Mathieu Laurière, and Zongjun Tan. Linear-quadratic mean-field reinforcement learning: convergence of policy gradient methods. 2019.

[10] J.F. Chassagneux, D. Crisan, and F. Delarue. *A probabilistic approach to classical solutions of the master equation for large population equilibria*, volume 280. Memoirs of the AMS, 2022.

[11] P.-E. Chaudru de Raynal and N. Frikha. From the backward Kolmogorov PDE on the Wasserstein space to propagation of chaos for McKean-Vlasov SDEs. *Journal de Mathématiques Pures et Appliquées*, 156:1–124, 2021.

[12] P.-E. Chaudru de Raynal and N. Frikha. Well-posedness for some non-linear SDEs and related PDE on the Wasserstein space. *Journal de Mathématiques Pures et Appliquées*, 159:1–167, 2022.

[13] D. Crisan and E. McMurray. Smoothing properties of McKean–Vlasov SDEs. *Probability Theory and Related Fields*, 171:97–148, 2018.

[14] R. Elie, J. Perolat, M. Laurière, M. Geist, and O. Pietquin. On the convergence of model free learning in mean field games. Proceedings of AAAI, 2020.

[15] N. Frikha, V. Konakov, and S. Menozzi. Well-posedness of some non-linear stable driven sdes. *Discrete and Continuous Dynamical Systems*, 41(2):849–898, 2021.

[16] M. Germain, M. Laurière, H. Pham, and X. Warin. DeepSets and their derivative networks for solving summetric PDEs. *Journal of Scientific Computing*, 91(63), 2022.

[17] M. Germain, J. Mikael, and X. Warin. Numerical resolution of McKean-Vlasov FBSDEs using neural networks. *Methodology and Computing in Applied Probability*, 2022.

[18] H. Gu, X. Guo, X. Wei, and R. Xu. Mean field controls with Q-learning for cooperative MARL: convergence and complexity analysis. *SIAM Journal on Mathematics of Data Science*, 3(4), 2021.

[19] X. Guo, H. Pham, and X. Wei. Itô's formula for flow of measures on semimartingales. *arXiv:2010.05288, to appear in Stochastic Processes and their Applications*, 2021.

[20] X. Guo, R. Xu, and T. Zariphopoulou. Entropy regularization for mean field games with learning. *Mathematics of Operations research*, 47(4), 2022.

[21] J. Han, R. Hu, and J. Long. Learning high-dimensional McKean-Vlasov forward-backward stochastic differential equations with general distribition dependence. *arXiv: 2204.11924*, 2022.

[22] K. Hornik, M. Stinchcombe, and H. White. Multilayer feedforward networks are universal approximators. *Neural Networks*, 2:359–366, 1989.

[23] Y. Jia and X. Y. Zhou. Policy evaluation and temporal difference learning in continuous time and space: a martingale approach. *to appear in Journal of Machine learning Research*, 2021.

[24] Y. Jia and X. Y. Zhou. Policy gradient and actor critic learning in continuous time and space: theory and algorithms. *to appear in Journal of Machine learning Research*, 2021. arXiv: 2111.11232v1.

[25] H. Pham and X. Warin. Mean-field neural networks-based algorithms for McKean-Vlasov control problems. *arXiv: 2212.11518*, 2022.

[26] H. Pham and X. Wei. Dynamic programming for optimal control of stochastic McKean-Vlasov dynamics. *SIAM Journal on Control and Optimization*, 55(2):1069–1101, 2017.

[27] C. Reisinger, W. Stockinger, and Y. Zhang. A fest iterative PDE-based algorithm for feedback controls of nonsmooth mean-field control problems. *arXiv: 2108.06740*, 2021.

[28] Y. Sun. The exact law of large numbers via fubini extension and characterization of insurable risks. *Journal of Economic Theory*, 126(1):31–69, 2006.

[29] R. Sutton and A. Barto. *Reinforcement Learning: An Introduction.* Cambridge, MA:MIT, 2018.

[30] H. Wang, T. Zariphopoulou, and X. Y. Zhou. Reinforcement learning in continuous time and space: A stochastic control approach. *Journal of Machine Learning Research*, 21(198):1–34, 2020.