Linear-Quadratic Mean-Field Reinforcement Learning: Convergence of Policy Gradient Methods

René Carmona[†] **Mathieu Laurière**[†] **Zongjun Tan**[†]

Abstract

We investigate reinforcement learning for mean field control problems in discrete time, which can be viewed as Markov decision processes for a large number of exchangeable agents interacting in a mean field manner. Such problems arise, for instance when a large number of robots communicate through a central unit dispatching the optimal policy computed by minimizing the overall social cost. An approximate solution is obtained by learning the optimal policy of a generic agent interacting with the statistical distribution of the states of the other agents. We prove rigorously the convergence of exact and model-free policy gradient methods in a mean-field linear-quadratic setting. We also provide graphical evidence of the convergence based on implementations of our algorithms.

1 Introduction

Typical *reinforcement learning* (RL) applications involve the search for a procedure to learn by trial and error the optimal behavior so as to maximize a reward. While similar in spirit to optimal control applications, a key difference is that in the latter, the model is assumed to be known to the controller. This is in contrast with RL for which the environment has to be explored, and the reward cannot be predicted with certainty. Still, the RL paradigm has generated numerous theoretical developments and found plenty practical applications. As a matter of fact, bidirectional links with the optimal control literature have been unveiled as common tools lie at the heart of many studies. The family of *linear-quadratic* (LQ) models proved to be of great importance in optimal control because of its tractability and versatility. Not surprisingly, these models have also been studied from a RL viewpoint. See e.g. [26, 10]. *Mean field control* (MFC), also called *optimal control of McKean-Vlasov* (MKV) dynamics, is an extension of stochastic control which has recently attracted a surge of interest (see e.g. [5, 1]). From a theoretical standpoint, the main peculiarity of this type of problems is that the transition and reward functions do not only involve the state and the action of the controller, but also the distribution of the state and potentially of the control. Practically speaking, they appear as the asymptotic limits for the control of a large number of collaborative agents, but they can also be introduced as single agent problems whose evolution and costs depend upon the distribution of her state and her control. Such problems have found a wide range of applications in distributed robotics, energy, drone fleet management, risk management, finance, etc. Although they are *bona fide* control problems for which a dynamic programming principle can be formulated, they generally lead to Bellman equations on the infinite dimensional space of measures, which are extremely difficult to solve ([2, 22, 25]). Luckily, for mean field LQ problems, simpler optimality conditions can often be framed in terms of (algebraic) Riccati equations, analogously to the non-mean field case. Figure 1 contains a schematic diagram of the relationships between optimal control (i.e., planning with a model) and the paradigms of MFC and RL. On the one hand, RL can be viewed as a development of optimal control in which

[†]Department of Operations Research and Financial Engineering, Princeton University

the model is (partially or fully) unknown. On the other hand, MFC is a generalization of optimal control for multiple agents when the number of agents grows to infinity. Mean-field reinforcement learning (MFRL for short) lies at the intersection of these two extensions and aims at describing how a large number of agents can collectively learn the optimal solution of a control problem by trial and error.

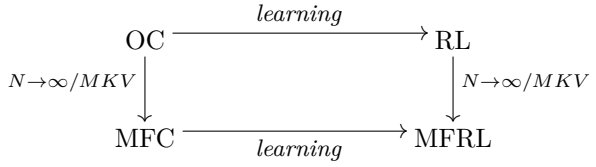


Figure 1: Diagram describing the relationship between optimal control (OC), mean-field control (MFC), reinforcement learning (RL) and mean-field reinforcement learning (MFRL). Horizontal and vertical arrows represent generalizations by model-free learning or by letting the number of agents grow to infinity.

Main contributions. From a theoretical viewpoint, we investigate reinforcement learning when the distribution of the state influences both the dynamics and the rewards. We focus on LQ models and make three main contributions. First, we identify, in a general setting which had not been studied before, the optimal control as linear in the state and its mean, which is crucial to study the link between the MFC problem and learning by a finite population of agents. We argue that it provides an approximately optimal control for the problem with a finite number of learners. We study a policy gradient (PG) method, in exact and model-free settings (see Section 3), and we prove global convergence in all cases. Notably, we show how a finite number of agents can collaborate to learn a control which is approximately optimal. The key idea is that the agents, instead of trying to learn the optimal control of their problem, try instead to learn the control which is socially optimal for the limiting MFC problem. Last, we conduct numerical experiments (see Section 4).

Our proof of convergence generalizes to the mean field setting the recent groundbreaking work of Fazel et al. [10] in which the authors have established global convergence of PG for LQ problems. One key feature of our model is the presence of a so-called *common noise*. To the best of our knowledge, it has never been considered in prior studies on mean field models and reinforcement learning. As we explain in the text, its inclusion is crucial if the model has to capture random sources of shocks which are common to all the players, and cannot average out in the asymptotic regime of large populations of learners. While its presence dramatically complicates the mathematical analysis, it can also be helpful. In the present paper, we take advantage of its impact to explore the unknown environment. Furthermore, in our model-free investigations, we study two different types of simulator. We first show how the techniques of [10] can be adapted to show convergence of PG for the mean field problem if we are provided with an (idealized) MKV simulator. We then proceed to show how to obtain convergence even when one has only access to a (more realistic) simulator for a system with a finite number of interacting agents. The proof requires mean-field type techniques relying on the law of large numbers and, more generally, the propagation of chaos. Our numerical experiments show that the method is very robust since it can be applied even if the agents are not exchangeable and have noisy dynamics.

2 Linear Quadratic Mean Field Control (LQMFC)

Notation. $\|\cdot\|$, $\|\cdot\|_F$ and $\|\cdot\|_{tr}$ denote respectively the operator norm, the Frobenius norm, and the trace norm. $\|\cdot\|_2$, or simply $\|\cdot\|$ if there is no ambiguity, denotes the Euclidean norm of a vector in $\mathbb{R}^d$. $\lambda(\cdot)$ and $\sigma_{\min}(\cdot)$ denote respectively the spectrum and the minimal singular value of a matrix. $X \succeq 0$ means that X is positive semi-definite (psd). We say that a probability measure μ on $\mathbb{R}^d$ is of order 2 if $\int \|x\|^2 \mu(dx) < \infty$. $\|\cdot\|_{L^p}$ denotes the L^p norm of a random vector. If $\xi \in \mathbb{R}^d$, $\mathrm{diag}(\xi)$ denotes the $d \times d$ diagonal matrix with the entries of ξ on the diagonal. If A, B are two matrices, $\mathrm{diag}(A, B)$ denotes the block-diagonal matrix with blocks A and B. The transpose of a row, a vector, or a matrix is denoted by $^\top$. $\|\cdot\|_{\psi_2}$ is the sub-Gaussian norm: namely for a real-valued sub-Gaussian random variable η, $\|\eta\|_{\psi_2} = \inf\left\{s > 0 : \mathbb{E}[\exp(\eta^2/s^2)] \le 2\right\}$, and for a d-dimensional sub-Gaussian random vector ξ, $\|\xi\|_{\psi_2} = \sup_{v\in\mathbb{R}^d:\|v\|=1} \|\xi^\top v\|_{\psi_2}$. See e.g. [31] for more details.

2.1 Mean field control problem

Definition of the problem. We consider the stochastic evolution of a state $x_t \in \mathbb{R}^d$: for $t = 0, 1, \dots,$

$$x_{t+1} = Ax_t + \bar{A}\bar{x}_t + Bu_t + \bar{B}\bar{u}_t + \epsilon^0_{t+1} + \epsilon^1_{t+1}, \quad (1)$$

where $u_t \in \mathbb{R}^\ell$ has the interpretation of a control and, for $t \geq 0$, the noise terms ϵ^0_{t+1} and ϵ^1_{t+1} are mean-zero $\mathbb{R}^d$-valued random vectors with a finite second moment. We shall assume that the entries in the sequences $(\epsilon^0_t)_{t\geq 1}$ and $(\epsilon^1_t)_{t\geq 1}$ are independent and identically distributed (i.i.d. for short) and they are independent of each other. They should be thought of as the sequences of increments of a common and an idiosyncratic noise respectively. We shall understand clearly this distinction when we discuss the dynamics of N agents (see § 2.2). In order to allow the initial state x_0 to be random, we assume that $x_0 = \epsilon^0_0 + \epsilon^1_0$, but while we assume that ϵ^0_0 and ϵ^1_0 are independent of each other, and independent of the ϵ^0_t and ϵ^1_t for $t \geq 1$, we shall not assume that they are mean zero themselves. Let $\tilde{\mu}^0_0$ and $\tilde{\mu}^1_1$ be their respective distributions. Here $\bar{x}_t$ and $\bar{u}_t$ are the conditional mean of x_t and u_t given $(\epsilon^0_s)_{s=0,\dots,t}$. These terms encode *mean field* (MF) interactions and explain the terminology *McKean-Vlasov* (MKV). See Section A of appendix for details. A, $\bar{A}$, B and $\bar{B}$ are fixed matrices with suitable dimensions.

We define the instantaneous MF cost at time t by:

$$c(x_t, \bar{x}_t, u_t, \bar{u}_t) = (x_t - \bar{x}_t)^\top Q(x_t - \bar{x}_t) + \bar{x}_t^\top (Q + \bar{Q})\bar{x}_t + (u_t - \bar{u}_t)^\top R(u_t - \bar{u}_t) + \bar{u}_t^\top (R + \bar{R})\bar{u}_t$$

where Q, $\bar{Q}$, R and $\bar{R}$ are real symmetric matrices of suitable dimensions (independent of time). We make the following assumption.

Assumption 1. *$Q, Q + \bar{Q}, R, R + \bar{R}$ are psd matrices.*

This assumption guarantees that the Hamiltonian of the system is convex. In fact, if any of these four matrices is positive definite, then the Hamiltonian of the system is strictly convex, ensuring uniqueness of the optimal control. The goal of the MFC problem is to minimize the expected (infinite horizon) discounted MF cost

$$J(\mathbf{u}) = \mathbb{E}\sum_{t=0}^\infty \gamma^t c(x^{\mathbf{u}}_t, \bar{x}^{\mathbf{u}}_t, u_t, \bar{u}_t) \quad (2)$$

where $\gamma \in [0, 1]$ is a discount factor, and we used the notation $(x^{\mathbf{u}}_t)$ to emphasize the fact that the state process satisfies (1) where $\mathbf{u} = (u_t)_{t=0,1,\dots}$ is an admissible control sequence, namely a sequence of random variables u_t on $(\Omega, \mathcal{F}, \mathbb{P})$, measurable with respect to the σ-field $\mathcal{F}_t$ generated by $\{\epsilon^0_1, \epsilon^1_1, \dots, \epsilon^0_t, \epsilon^1_t\}$ and which satisfy: $\mathbb{E}\sum_{t=0}^\infty \gamma^t \|u_t\|^2 < \infty$.

Characterization of the optimal control. To the best of our knowledge, the above problem has not been studied in the literature, presumably because it is set in infinite horizon and with discrete time, it includes a *common noise* ϵ^0, and the interaction is not only through the conditional mean of all the states, but also through the conditional mean of the controls. Under some suitable conditions, using techniques similar to [4, Section 3.5], it can be shown that the optimal control can be identified as a *linear* combination of x_t and $\bar{x}_t$ at each time t.

2.2 Problem with a finite number of agents

We now consider N agents interacting in a mean field manner. We denote their states at time t by $(x^n_t)_{n=1,\dots,N}$. They satisfy the state system of equations: for all $n = 1, \dots, N, t \geq 0$,

$$x^n_{t+1} = Ax^n_t + \bar{A}\bar{x}^N_t + Bu^n_t + \bar{B}\bar{u}^N_t + \epsilon^0_{t+1} + \epsilon^{1,(n)}_{t+1}, \quad (3)$$

with initial conditions $x^n_0 = \epsilon^0_0 + \epsilon^{1,(n)}_0$, where $A, \bar{A}, B, \bar{B}$ are as before and $(\epsilon^0_t)_{t\geq 1}$ and $(\epsilon^{1,(n)}_t)_{n=1,\dots,N, t\geq 1}$ are families of independent identically distributed mean-zero real-valued random variables, which are assumed to be independent of each other. We use the notations: $\bar{x}^N_t = \frac{1}{N}\sum_{n=1}^N x^n_t$ and $\bar{u}^N_t = \frac{1}{N}\sum_{n=1}^N u^n_t$ to denote the sample averages of the individual states and controls. Notice that with these notations, we can write the evolution of the sample average of the state:

$$\bar{x}^N_{t+1} = (A+\bar{A})\bar{x}^N_t + (B+\bar{B})\bar{u}^N_t + \epsilon^0_{t+1} + \tfrac{1}{N}\textstyle\sum_{n'=1}^N \epsilon^{1,(n')}_{t+1}.$$

Note that ϵ^0 affects all the agents and represents aggregate shocks e.g. in economic models (see [6, 7] for applications to systemic risk and energy management in the MFG setting). Unlike its counterpart in the MKV dynamics, $\bar{x}^N$ involves not only the common noise but also the idiosyncratic noises. Letting $X_t = [x^1_t, \dots, x^N_t]^\top$ and $U_t = [u^1_t, \dots, u^N_t]^\top$ for each integer $t \geq 0$, we have the vector dynamics

$$X_{t+1} = A^N X_t + B^N U_t + E^0_{t+1} + E^1_{t+1} \quad (4)$$

with $E^1_{t+1} = [\epsilon^{1,(1)}_{t+1}, \dots, \epsilon^{1,(N)}_{t+1}]^\top$, $E^0_{t+1} = [\epsilon^0_{t+1}, \dots, \epsilon^0_{t+1}]^\top$ and $A^N = I_N \otimes A + \frac{1}{N}\mathbf{1}_N \otimes \bar{A}$ and $B^N = I_N \otimes B + \frac{1}{N}\mathbf{1}_N \otimes \bar{B}$, where $\otimes$ denotes the Kronecker product between matrices. In other words, the matrix A^N is the sum of the block-diagonal matrix with N blocks A and of the matrix with $N \times N$ blocks $\frac{1}{N}\bar{A}$, and similarly for the matrix B^N. Throughout the paper, I_d is the $d \times d$ identity matrix, $\mathbf{1}_d$ and $\mathbb{1}_d$ are respectively the $d \times d$ matrix and the $d-$vector whose entries are all ones. We drop the subscripts when the dimension is clear from the context.

We seek to minimize the so-called population *social cost* defined as $J^N(\mathbf{U}) = \mathbb{E}\sum_{t\geq 0}\gamma^t \bar{c}^N(X_t, U_t)$ over the admissible control processes $\mathbf{U} = (U_t)_t$ where, by admissibility we mean that: 1) the control process is adapted to the filtration generated by the sources of randomness (initial condition, common noise and idiosyncratic noises), and 2) the expectation appearing in the definition of $J^N(\mathbf{U})$ is finite. Here, the instantaneous *social cost* function $\bar{c}^N$ is defined by

$$\bar{c}^N(X, U) = \frac{1}{N}\sum_{n=1}^N c^{(n)}\left(x^n, \bar{x}^N, u^n, \bar{u}^N\right) \qquad (5)$$

where $X = [x^1, \dots, x^N]^\top$ and $U = [u^1, \dots, u^N]^\top$, and for each $n \in \{1, \dots, N\}$ the function $c^{(n)}$ is the cost of agent n, defined for $x, \bar{x}, u, \bar{u} \in \mathbb{R}^d$ by:

$$c^{(n)}(x, \bar{x}, u, \bar{u}) = (x - \bar{x})^\top Q^n (x - \bar{x}) + \bar{x}^\top (Q^n + \bar{Q})\bar{x} + (u - \bar{u})^\top R(u - \bar{u}) + \bar{u}^\top (R + \bar{R})\bar{u},$$

where $Q^n = Q + \tilde{Q}^n$ for each $n = 1, \dots, N$. Here $\|\tilde{Q}^n\| \leq \tilde{h}$, with $\tilde{h} \leq \min\{\lambda_{min}(Q), \lambda_{min}(Q + \bar{Q})\}$, should be seen as a variation of Q and $\tilde{h}$ quantifies the degree of *heterogeneity* of the population.[1] We will use the notation $\tilde{Q} = [\tilde{Q}^1, \dots, \tilde{Q}^N]^\top$ for the vector of variations. We could perturb the coefficients $\bar{Q}$, R and $\bar{R}$ as well, but we chose to perturb Q only because this is enough to illustrate departures from the MKV solution which we want to emphasize. The optimization problem can be recast as an infinite-horizon discounted-cost linear-quadratic control problem where the state $\mathbf{X} = (X_t)_t$ is a stochastic process in dimension dN. In particular, the optimal control $\mathbf{U}^{*,N}$ is of the form $U^{*,N}_t = \Phi^{*,N} X_t$ for a deterministic $(dN) \times (dN)$ constant matrix $\Phi^{*,N}$.

[1] Actually, it is sufficient to assume that $\tilde{Q}^n$ is such that $Q + \tilde{Q}^n \succeq 0$ and $\bar{Q} + Q + \tilde{Q}^n \succeq 0$.

2.3 Link between the two problems

It can be shown that the optimal control for the MFC problem is given at time t by $u^*_{MKV}(t) = -K^* x_t - (L^* - K^*)\bar{x}_t$ for some constant matrices K^* and L^*. For the purpose of comparison with the matrix $\Phi^{*,N}$ defined above, we introduce the matrix $\Phi^{*,N}_{MKV} = -I_N \otimes K^* - \frac{1}{N}\mathbf{1}_N \otimes (L^* - K^*)$. It turns out that this matrix provides a control which is approximately optimal for the $N-$agent problem, i.e., if the agents use the control $U^{*,N}_{MKV}(t) = \Phi^{*,N}_{MKV} X_t = -\text{diag}(K^*, \dots, K^*)X_t - \text{diag}((L^* - K^*), \dots, (L^* - K^*))\overline{X}_t$, then the social cost is at most ϵ more than the minimum, where ϵ depends on N and the heterogeneity degree $\tilde{h}$, and vanishes as $N \to +\infty$ and $\tilde{h} \to 0$ (see Section F.2 of the appendix for numerical illustrations). In the extreme case of homogeneity, namely when $\tilde{h} = 0$, $\Phi^{*,N} = \Phi^{*,N}_{MKV}$. So in this case, the optimal strategy for N agents is the same as in the MKV case, except that the mean of the MKV process is replaced by the $N-$agent sample average. This intuition naturally leads to the notion of *decentralized* control which can be easily implemented using the solution of the MFC problem (see Remark 49 in Section F.2 of the appendix and [14]). This perfect identification was the main motivation to develop RL methods for the optimal control of MKV dynamics.

3 Policy gradient method

3.1 Exact PG for MFC

Admissible controls. We go back to the framework for the optimal control of MKV dynamics discussed in § 2.1. Our theoretical analysis suggests that we restrict our attention to controls which are linear in x and $\bar{x}$ since we know that the optimal control is in this class. Accordingly, we define Θ as the set of $\theta = (K, L) \in \mathbb{R}^{\ell \times d} \times \mathbb{R}^{\ell \times d}$ such that $\gamma\|A - BK\|^2 < 1$ and $\gamma\|A + \bar{A} - (B + \bar{B})L\|^2 < 1$. For $\theta = (K, L) \in \Theta$, we define $\tilde{u}_\theta : (x, \bar{x}) \mapsto -K(x - \bar{x}) - L\bar{x}$, and to alleviate the notations, we write x^θ_t instead of $x^{\mathbf{u}^\theta}_t$ for the state controlled by the control process $u^\theta_t = \tilde{u}_\theta(x_t, \bar{x}_t)$. We will sometimes abuse terminology and call θ a control or say that the state process is controlled by θ. We let $\mathbf{K}_\theta = \text{diag}(K, L) \in \mathbb{R}^{(2\ell)\times(2d)}$. Conversely, it will also be convenient to write $\theta(\mathbf{K}) = (K, L)$ when $\mathbf{K} = \text{diag}(K, L)$.

Re-parametrization. The goal is to recast the initial mean-field LQ problem as a LQ problem in a larger state space. Given $\theta = (K, L) \in \Theta$ and the associated controlled state process $(x_t^\theta)_{t\geq 0}$, we introduce two auxiliary processes $(y_t^\theta)_{t\geq 0}$ and $(z_t^\theta)_{t\geq 0}$ defined for $t \geq 0$ by

$$y_t^\theta = x_t^\theta - \mathbb{E}[x_t^\theta|\mathcal{F}^0], \qquad \text{and} \qquad z_t^\theta = \mathbb{E}[x_t^\theta|\mathcal{F}^0]. \quad (6)$$

Note that $y_0^\theta = y_0 = \epsilon_0^1 - \mathbb{E}[\epsilon_0^1]$ and $z_0^\theta = z_0 = \epsilon_0^0 + \mathbb{E}[\epsilon_0^1]$ are independent of θ and that

$$\begin{cases} y_{t+1}^\theta &= (A - BK)y_t^\theta + \epsilon_{t+1}^1, \\ z_{t+1}^\theta &= (A + \bar{A} - (B + \bar{B})L)z_t^\theta + \epsilon_{t+1}^0. \end{cases}$$

In particular, y_t^θ (resp. z_t^θ) depends on θ only through K (resp. L). We also introduce the state vector $\mathbf{x}_t^\theta = [y_t^\theta, z_t^\theta]^\top \in \mathbb{R}^{2d}$, and view it as a new controlled state process, with $\mathbf{x}_0^\theta = \mathbf{x}_0 = \left[\epsilon_0^1 - \mathbb{E}[\epsilon_0^1], \epsilon_0^0 + \mathbb{E}[\epsilon_0^1]\right]^\top$ and for $t \geq 0$,

$$\mathbf{x}_{t+1}^\theta = (\mathbf{A} - \mathbf{B}\mathbf{K}_\theta)\mathbf{x}_t^\theta + \boldsymbol{\epsilon}_{t+1} \quad (7)$$

where $\boldsymbol{\epsilon}_t = \left[\epsilon_t^1, \epsilon_t^0\right]^\top \in \mathbb{R}^{2d}$, $\mathbf{A} = \mathrm{diag}(A, A + \bar{A})$ and $\mathbf{B} = \mathrm{diag}(B, B)$. We introduce the new cost function $C : \mathbb{R}^{\ell\times d} \times \mathbb{R}^{\ell\times d} \ni \theta = (K, L) \mapsto C(\theta) \in \mathbb{R}$ defined by

$$C(\theta) = \mathbb{E}\sum_{t\geq 0} \gamma^t (\mathbf{x}_t^\theta)^\top \Gamma_\theta \mathbf{x}_t^\theta,$$

where $\Gamma_\theta = \mathbf{Q} + \mathbf{K}_\theta^\top \mathbf{R}\mathbf{K}_\theta$, with $\mathbf{Q} = \mathrm{diag}(Q, Q + \bar{Q})$ and $\mathbf{R} = \mathrm{diag}(R, R + \bar{R})$. Finally, we introduce the auxiliary cost functions C_y and C_z defined by:

$$C_y(K) = \mathbb{E}\sum_{t\geq 0} \gamma^t (y_t^\theta)^\top (Q + K^T R K) y_t^\theta,$$
$$C_z(L) = \mathbb{E}\sum_{t\geq 0} \gamma^t (z_t^\theta)^\top (Q + \bar{Q} + L^T (R + \bar{R}) L) z_t^\theta.$$

The following result gives the rationale for our re-parameterization of the optimization problem.

Lemma 1. *For any $\theta = (K, L) \in \Theta$, we have $J(\tilde{u}_\theta) = C(\theta) = C_y(K) + C_z(L)$, and $\inf_{\theta\in\Theta} J(\tilde{u}_\theta) = \inf_{\theta\in\Theta} C(\theta) = \inf_K C_y(K) + \inf_L C_z(L)$.*

Here we slightly abuse of notation and see J as a function of $\tilde{u}_\theta$. The proof of this lemma is deferred to Section B.1 of the appendix. Notice that, by definition of $\tilde{u}_\theta$ and $C(\theta)$, we also have $\arg\min_{\theta\in\Theta} J(\tilde{u}_\theta) = \arg\min_{\theta\in\Theta} C(\theta)$. De facto, we split the original control problem into two separate optimization problems which can be tackled in parallel.

Exact PG convergence result. Assuming full knowledge of the model, we compute the optimal parameters using a standard gradient descent algorithm. With a fixed learning rate $\eta > 0$ and initial parameter $\theta_0 = (K_0, L_0)$, we update the parameter θ from $\theta^{(k)}$ to $\theta^{(k+1)} = \theta(\mathbf{K}^{(k+1)})$ via $\mathbf{K}^{(k+1)} = \mathbf{K}^{(k)} - \eta\nabla C(\theta^{(k)})$. Convergence of the algorithm is proved under the following assumption. Let $\Sigma^1 = \mathbb{E}[\epsilon_1^1(\epsilon_1^1)^\top]$, $\Sigma^0 = \mathbb{E}[\epsilon_1^0(\epsilon_1^0)^\top]$, $\Sigma_{y_0} = \mathbb{E}\left[y_0^1(y_0^1)^\top\right]$, and $\Sigma_{z_0} = \mathbb{E}\left[z_0^1(z_0^1)^\top\right]$.

Assumption 2. $\max\{\lambda_{min}(\Sigma_{y_0}), \lambda_{min}(\Sigma^1)\} > 0$, *and* $\max\{\lambda_{min}(\Sigma_{z_0}), \lambda_{min}(\Sigma^0)\} > 0$. *Moreover $\epsilon_0^1, \epsilon_0^0$ and $\epsilon_t^1, \epsilon_t^1$ for all $t \geq 1$ are sub-Gaussian random variables whose $\|\cdot\|_{\psi_2}$ norms are bounded by a constant C_0.*

Remark 2. *The above assumption implies directly that the variances Σ_{y_0}, Σ_{z_0},Σ^1, Σ^0 are bounded in ψ_2 norm. If the initial distributions are non-degenerate with bounded supports, and if there is no noise processes (which imply Assumption 2 automatically), the arguments of [10] can be adapted in a rather straightforward way. However, in the more general setting considered here, which covers in particular non-degenerate Gaussian distributions, we need more tools and ideas to prove the convergence results.*

We prove the following convergence result. Here and thereafter, we denote by θ^* the optimal parameter, i.e. such that $C(\theta^*) = \min_{\theta\in\Theta} C(\theta)$.

Theorem 3. *Under our standing assumptions, for every $\varepsilon > 0$, if k is large enough and if $\eta > 0$ is small enough, then $C(\theta^{(k)}) - C(\theta^*) \leq \varepsilon$. The convergence rate is linear in the sense that it is sufficient to take the number of steps k of the order of $\mathcal{O}(\log(1/\varepsilon))$.*

Remark 4. *The constant in the big $\mathcal{O}$ can be expressed as a polynomial of the data of the problems. See [10, Theorem 7] or the proof of Theorem 6 below for more details.*

The proof is adapted from [10, Theorem 7] and the main differences are stressed in Section B of the appendix. The non-degeneracy of the randomness, which comes from Assumption 2, plays a crucial role in our ability to adapt the result of [10]. Note that the re-parameterization is critical to show that, during the gradient descent, the matrix $\mathbf{K}^{(k)}$ keeps a block-diagonal structure.

3.2 Model free PG for MFC with MKV simulator

Let us assume that we have access to the following (stochastic) simulator, called MKV simulator $\mathcal{S}^T_{MKV}$: given a control parameter θ, $\mathcal{S}^T_{MKV}(\theta)$ returns a sample of the mean-field cost (see § 2.1) for the MKV dynamics (1) using the control θ. In other words, it returns a realization of the social cost $\sum_{t=0}^{T-1}\gamma^t c_t$, where c_t is the instantaneous mean-field cost at time t. Notice that, although the simulator needs to know the model in order to sample the dynamics and compute the costs, Algorithm 1 below uses this simulator as a black-box (or an oracle), and hence uses only samples from the model and not the model itself. Without full knowledge of the model, we use derivative-free techniques (see Section C of the appendix) to estimate the gradient of the cost. We let $\mathbb{B}_\tau \subset \mathbb{R}^{\ell\times d}$ be the ball of radius τ centered at the origin, and $\mathbb{S}_\tau = \partial\mathbb{B}_\tau$ be its boundary. The uniform distributions on $\mathbb{B}_\tau$ and $\mathbb{S}_\tau$ are denoted by $\mu_{\mathbb{B}_\tau}$ and $\mu_{\mathbb{S}_\tau}$ respectively.

Algorithm 1 provides a (biased) estimator of the true policy gradient $\nabla C(\theta)$. This method is very much in the spirit of [10, Algorithm 1], except that here we have two components (the state and the conditional mean) playing different roles. The choice of two independent sources (v_{1i}) and (v_{2i}) of noise is important to get an estimation of the gradients w.r.t. K and L solely from observing a single cost provided by the MKV simulator $\mathcal{S}^T_{MKV}(\theta_i)$. We prove the following convergence result, which generalizes [10, Theorem 30].

Algorithm 1: Model-free MKV-Based Gradient Estimation

Data: Parameter $\theta = (K, L)$; number of perturbations M; length T; radius τ

Result: A biased estimator for the gradient $\nabla C(\theta)$

begin
  for $i = 1, 2, \dots, M$ **do**
    Sample v_{1i}, v_{2i} i.i.d. $\sim \mu_{\mathbb{S}_\tau}$
    Set $\theta_i = (K_i, L_i) = (K + v_{1i}, L + v_{2i})$
    Sample $\tilde{C}^i$ using $\mathcal{S}^T_{MKV}(\theta_i)$
  Set $\tilde{\nabla}_K C(\theta) = \frac{d}{\tau^2}\frac{1}{M}\sum_{i=1}^M \tilde{C}^i v_{1i}$, and $\tilde{\nabla}_L C(\theta) = \frac{d}{\tau^2}\frac{1}{M}\sum_{i=1}^M \tilde{C}^i v_{2i}$
  return $\tilde{\nabla} C(\theta) = \mathrm{diag}\left(\tilde{\nabla}_K C(\theta), \tilde{\nabla}_L C(\theta)\right)$

Theorem 5. *Under our standing assumptions, if we use the update rule* $\tilde{\mathbf{K}}^{(k+1)} = \tilde{\mathbf{K}}^{(k)} - \eta\tilde{\nabla}C(\theta^{(k)})$ *where* $\tilde{\nabla}C(\theta^{(k)})$ *is given by Algorithm 1, for every* $\varepsilon > 0$*, if* k*,* M*,* τ^{-1} *and* T *are large enough, then we have* $C(\theta^{(k)}) - C(\theta^*) \le \varepsilon$ *with probability at least* $1 - k(d/\varepsilon)^{-d}$*. The convergence rate is linear in the sense that it is sufficient to take* k *of the order of* $\mathcal{O}(\log(1/\varepsilon))$*.*

The proof is deferred to Section C of the appendix, but it is worth mentioning here some important steps. It relies on the reparametrization $\mathbf{x}^\theta = [y^K, z^L]^\top$ of the state process. Due to the idiosyncratic and common noises, the sampled costs are not bounded but only sub-exponentially distributed. A careful analysis is required to obtain polynomial bounds on the parameters M, τ^{-1} and T used in the algorithm. The main difficulty turns out to reduce to the study of a sequence of quadratic forms of sub-Gaussian random vectors with dependent coordinates (see [35, Proposition 2.5]). The discount term also plays an important role here. We refer to Lemma 35 in the appendix for more details on the cornerstone of the proof.

3.3 Model-free PG for MFC with population simulator

We now turn our attention to a more realistic setting where one does not have access to an oracle simulating the MKV dynamics, but only to an oracle merely capable of simulating the evolution of N agents. We then use the state sample average instead of the theoretical conditional mean, and for the social cost, the empirical average instead of the mean-field cost provided by the MKV simulator. We rely on the following population simulator $\mathcal{S}^{T,N}_{pop}$: given a control parameter θ, $\mathcal{S}^{T,N}_{pop}(\theta)$ returns a sample of the social cost obtained by sampling realizations of the N state trajectories controlled by θ and computing the associated cost for the population, see (3) and (5). In other words, it returns a realization of $\sum_{t=0}^T \gamma^t \bar{c}^N_t$, where $\bar{c}^N_t$ is the instantaneous N-agent social cost cost at time t.

Notice that this population simulator $\mathcal{S}^{T,N}_{pop}$ is arguably more realistic and less powerful than the previous MKV simulator $\mathcal{S}^T_{MKV}$, in the sense that the former has only an approximation (or a noisy version) of the true mean process. We stress that *all* the agents use the same control (and in Algorithm 2, they use the *same* perturbed version of a control). This point is in line with the idea that the problem corresponds to the one of a central planner or the one of a population trying to find the

best common feedback control rule in order to optimize a social cost (see § 2.2).

Algorithm 2: Model-free Population-Based Gradient Estimation

Data: $\theta = (K, L)$; number of agents N; number of perturbations M; length T; radius τ

Result: A biased estimator for $\nabla C(\theta)$.

begin

  for $i = 1, 2, \dots, M$ **do**

    Sample v_{1i}, v_{2i} i.i.d. $\sim \mu_{\mathbb{S}_\tau}$

    Set $\theta_i = (K_i, L_i) = (K + v_{1i}, L + v_{2i})$

    Sample $\tilde{C}^{i,N}$ using $\mathcal{S}^{T,N}_{pop}(\theta_i)$

  Set $\tilde{\nabla}_K C^N(\theta) = \frac{d}{\tau^2}\frac{1}{M}\sum_{i=1}^M \tilde{C}^{i,N} v_{1i}$, and $\tilde{\nabla}_L C^N(\theta) = \frac{d}{\tau^2}\frac{1}{M}\sum_{i=1}^M \tilde{C}^{i,N} v_{2i}$

  return $\tilde{\nabla} C^N(\theta) = \mathrm{diag}\left(\tilde{\nabla}_K C^N(\theta), \tilde{\nabla}_L C^N(\theta)\right)$

Theorem 6. *Under our standing assumptions, if $\tilde{h} = 0$, and if we use the update rule $\tilde{\mathbf{K}}^{(k+1)} = \tilde{\mathbf{K}}^{(k)} - \eta\tilde{\nabla}C^N(\theta^{(k)})$, where $\tilde{\nabla}C^N(\theta^{(k)})$ is given by Algorithm 2, then for every $\varepsilon > 0$, if k, M, τ^{-1}, T and N are large enough, then $C^N(\theta^{(k)}) - C^N(\theta^*) \le \varepsilon$ with probability at least $1 - k(d/\varepsilon)^{-d}$. The convergence rate is linear in the sense that it is sufficient to take the number of steps k of the order of $\mathcal{O}(\log(1/\varepsilon))$.*

The high-level structure of the proof follows the lines of [10, Theorem 28], with several important modifications that we highlight in the appendix. In particular, it relies on several crucial estimates controlling the bias introduced by MF interactions in the computation of the gradient of the cost. See, in the appendix, Section D for a sketch of proof and Section E for the details.

Remark 7. *Recall that when $\tilde{h} = 0$, the mean-field and the finite-population problems have the same θ^*. When $\tilde{h} > 0$ (agents are not perfectly identical), a similar convergence result could be proved, at the expense of an additional error term vanishing as $\tilde{h} \to 0$. Numerical tests conducted with heterogeneous agents in the next section support this idea.*

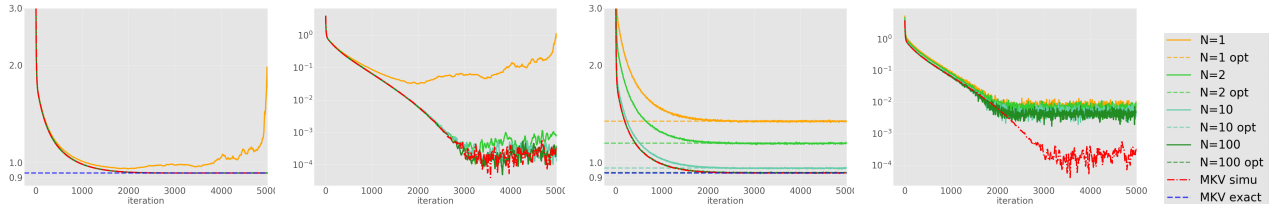
Remark 8. *Even if the common noise were degenerate, in the $N-$agent dynamics the average $\bar{x}^N$ would remain stochastic due to the term $\frac{1}{N}\sum_{n'=1}^N \epsilon_{t+1}^{1,n'}$. Hence, we expect that Theorem 6 could be proved without the non-degeneracy of the noises in Assumption 2.*

4 Numerical results

Although the main thrust of our work is theoretical, we assess our convergence results by numerical tests. To illustrate the robustness of our methods, we consider a case where the agents are heterogeneous (i.e., $\tilde{h} > 0$) and have idiosyncratic and common noise in the dynamics. For the discount, we used $\gamma = 0.9$. We consider here an example in which the state and the control are in dimension 1 because it is easier to visualize and because it allows focusing on the main difficulty of MFC problems, namely the MF interactions. For brevity, the precise setting is postponed to Section F of the appendix. The displayed curves correspond to the mean over 10 different realizations. The shaded region, if any, corresponds to the mean ± standard deviation.

For each model-free method, we look at how the control learned by PG performs both in terms of the MF cost and the $N-$agent social cost (see Figures 2). On Figure (2a) and (2b), one can see that the control learned by PG is approximately optimal for the MF cost, provided that the number of agents is strictly larger than 1. Indeed, as shown in Figure (3a), a single agent is able to learn the second component of the optimal control $\theta^* = (K^*, L^*)$ but not the first component. This can be explained by the fact that, when $N = 1$, $x^1 - \bar{x}^N$ is identically 0. So the agent does not have any estimate of the y component of the state (see (6)) and is thus unable to learn the associated control parameter. Figure (2b) displays the relative error, namely, $(C(\theta^{(k)}) - C^*)/C^*$ at iteration k, where $C^* = C(\theta^*)$ is the optimal mean field cost (i.e., the MKV cost evaluated at the optimal mean field control parameter θ^*).

As for the evaluation in terms of the population cost, Figure 2 also displays the $N-$agent social cost (2c) and the error with respect to the optimal social cost (2d), for each population size. Note that in our simulations, for a finite population, the social cost takes into account the heterogeneities of the agents's costs. In Figure (2c), the dashed line for each N corresponds to the value of $C^{*,N} = J^N(\Phi^{*,N})$, which is the optimal $N-$agent social cost (i.e., the social cost evaluated at the optimal $N-$agent control $\Phi^{*,N}$ introduced in section 2.2). Figure (2d) shows the relative error, defined as $(C^N(\theta^{(k)}) - C^{*,N})/C^{*,N}$ at iteration k. Here, we can see that the N-agents manage to learn a control which is approximately optimal for *their* own social cost (even when $N = 1$). For the sake of comparison, we also

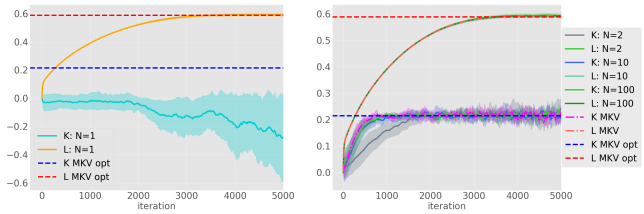


(a) (b) (c) (d)

Figure 2: Comparison of costs achieved by the controls learned with MKV and population simulator using model-free PG. (a) : mean field (MF) costs; (b) : relative error in MF cost wrt MF optimum; (c) : N-agent costs and their respective optimal value; (d) : relative error in N-agent costs wrt N-agent optimal cost.

display the performance of the model-free method with MKV simulator (measured in terms of the MKV cost), which appears in Figure (2) as well. It converges better since all the proposed methods attempt to learn the optimal control for the *mean field* problem. However, as explained in section 2.3, the difference between the $N-$agent optimal control and the mean-field optimal control vanishes as the heterogeneity degree $\tilde{h}$ goes to 0.

Besides the fact that a single agent is not able to learn K^*, we note in Figure (3b) that for $N = 2$ the convergence is almost as good with $N \geq 10$ as with with MKV simulator, but it is slower for $N = 2$. This certainly comes from propagation of chaos, and we expect some dependence on the dimension of the state, in accordance with our theoretical bounds.



(a) (b)

Figure 3: Convergence of the control parameters. (a) 1-agent simulator. (b) MKV and population simulators.

5 Conclusion and future research

In this work, we explored the central role played by MKV dynamics in multi-agent RL and showed that a relatively small number of agents are able to learn a mean field optimal control, which can then be re-used in larger systems with approximate optimality. We established convergence of policy gradient methods for linear-quadratic mean-field Markov decision problems using techniques from optimization and stochastic analysis. When a MKV simulator is available, we explained how the problem can be tackled by generalizing known results for standard LQ models. When one has only access to a population simulator, we extended existing results to the case of MF interactions. An important feature of our model is the presence of common noise, whose impact had to be controlled. The three convergence results shed light on a hierarchy of settings, namely: (1) knowledge of the mean field model, (2) access to a MKV simulator, and (3) access to a population simulator. We believe that the strategy employed here to successively prove these results (using a setting as a building block for the next one) could be useful for more general (i.e. non-linear-quadratic) mean-field RL problems.

Our results can be extended in several directions. Other RL methods (such as Actor-Critic algorithms [18]) and variance reduction techniques could be used to improve the method proposed here and generalize to non-LQ models. Our analysis and numerical implementations can be applied to other LQ problems such as mean field games or mean field control problems with several populations, with important applications to multi-agent sweeping and tracking.

Related work. Our work is at the intersection of RL and MFC. The latter has recently attracted a lot of attention, particularly since the introduction of mean field games (MFG) by Lasry and Lions [19, 20, 21]

and by Caines, Huang and Malhamé [15, 13]. MFGs correspond to the asymptotic limit of Nash equilibria for games with mean field interactions in which the number of players grows to infinity. They are inherently defined through a fixed point procedure and hence, differ both conceptually and numerically, from MFC problems which correspond to social optima. See for example [4] for details and examples. Most works on learning in the presence of mean field interactions have focused on MFGs, see e.g. [34, 3] for "learning" (or rather *solving*) MFGs based on the full knowledge of the model, and [16, 32, 33, 12, 23, 9, 28] for RL based methods. In contrast, our work focuses on MFC problems. While completing the present work, we became aware of the very recent work of [27], which also studies MFC with policy gradient methods. However, their work deals with finite state and action spaces whereas we consider continuous spaces. Furthermore, we provide rates of convergence and, finally, our results rely on assumptions that can be easily checked on the MFC model under consideration, which is not the case with the rather abstract conditions of [27].

Acknowledgments

M.L. and Z.T. thank Zhuoran Yang for fruitful discussions in the early stage of this project.

References

[1] Bensoussan, A., Frehse, J., and Yam, S. C. P. (2013). *Mean field games and mean field type control theory*. Springer Briefs in Mathematics. Springer, New York.

[2] Bensoussan, A., Frehse, J., and Yam, S. C. P. (2015). The master equation in mean field theory. *J. Math. Pures Appl. (9)*, 103(6):1441–1474.

[3] Cardaliaguet, P. and Hadikhanloo, S. (2017). Learning in mean field games: the fictitious play. *ESAIM: Control, Optimisation and Calculus of Variations*, 23(2):569–591.

[4] Carmona, R. and Delarue, F. (2018a). *Probabilistic theory of mean field games with applications. I*, volume 83 of *Probability Theory and Stochastic Modelling*. Springer, Cham. Mean field FBSDEs, control, and games.

[5] Carmona, R. and Delarue, F. (2018b). *Probabilistic Theory of Mean Field Games with Applications I-II*. Springer.

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

[7] Clemence, A., Imen, B. T., and Anis, M. (2017). An extended mean field game for storage in smart grids. *arXiv preprint arXiv:1710.08991*.

[8] Conn, A. R., Scheinberg, K., and Vicente, L. N. (2009). *Introduction to derivative-free optimization*, volume 8 of *MPS/SIAM Series on Optimization*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA; Mathematical Programming Society (MPS), Philadelphia, PA.

[9] Elie, R., Pérolat, J., Laurière, M., Geist, M., and Pietquin, O. (2019). Approximate fictitious play for mean field games. *arXiv preprint arXiv:1907.02633*.

[10] Fazel, M., Ge, R., Kakade, S., and Mesbahi, M. (2018). Global convergence of policy gradient methods for the linear quadratic regulator. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80.

[11] Flaxman, A. D., Kalai, A. T., and McMahan, H. B. (2005). Online convex optimization in the bandit setting: gradient descent without a gradient. In *Proceedings of the Sixteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 385–394. ACM, New York.

[12] Guo, X., Hu, A., Xu, R., and Zhang, J. (2019). Learning mean-field games. *arXiv preprint arXiv:1901.09585*.

[13] Huang, M., Caines, P. E., and Malhamé, R. P. (2007). Large-population cost-coupled LQG problems with nonuniform agents: individual-mass behavior and decentralized ϵ-Nash equilibria. *IEEE Trans. Automat. Control*, 52(9):1560–1571.

[14] Huang, M., Caines, P. E., and Malhamé, R. P. (2012). Social optima in mean field LQG control: centralized and decentralized strategies. *IEEE Trans. Automat. Control*, 57(7):1736–1751.

[15] Huang, M., Malhamé, R. P., and Caines, P. E. (2006). Large population stochastic dynamic games: closed-loop McKean-Vlasov systems and the Nash certainty equivalence principle. *Commun. Inf. Syst.*, 6(3):221–251.

[16] Iyer, K., Johari, R., and Sundararajan, M. (2014). Mean field equilibria of dynamic auctions with learning. *Management Science*, 60(12):2949–2970.

[17] Karimi, H., Nutini, J., and Schmidt, M. (2016). Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 795–811. Springer.

[18] Konda, V. R. and Tsitsiklis, J. N. (2000). Actor-critic algorithms. In *Advances in neural information processing systems*, pages 1008–1014.

[19] Lasry, J.-M. and Lions, P.-L. (2006a). Jeux à champ moyen. I. Le cas stationnaire. *C. R. Math. Acad. Sci. Paris*, 343(9):619–625.

[20] Lasry, J.-M. and Lions, P.-L. (2006b). Jeux à champ moyen. II. Horizon fini et contrôle optimal. *C. R. Math. Acad. Sci. Paris*, 343(10):679–684.

[21] Lasry, J.-M. and Lions, P.-L. (2007). Mean field games. *Jpn. J. Math.*, 2(1):229–260.

[22] Laurière, M. and Pironneau, O. (2016). Dynamic programming for mean-field type control. *J. Optim. Theory Appl.*, 169(3):902–924.

[23] Mguni, D., Jennings, J., and de Cote, E. M. (2018). Decentralised learning in systems with many, many strategic agents. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

[24] Nesterov, Y. and Spokoiny, V. (2017). Random gradient-free minimization of convex functions. *Found. Comput. Math.*, 17(2):527–566.

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

[26] Recht, B. (2018). A tour of reinforcement learning: The view from continuous control. *Annual Review of Control, Robotics, and Autonomous Systems*.

[27] Subramanian, J. and Mahajan, A. (2019). Reinforcement learning in stationary mean-field games. In *Proceedings. 18th International Conference on Autonomous Agents and Multiagent Systems*.

[28] Tiwari, N., Ghosh, A., and Aggarwal, V. (2019). Reinforcement learning for mean field game. *arXiv preprint arXiv:1905.13357*.

[29] Tropp, J. A. (2012). User-friendly tail bounds for sums of random matrices. *Found. Comput. Math.*, 12(4):389–434.

[30] Tropp, J. A. et al. (2015). An introduction to matrix concentration inequalities. *Foundations and Trends® in Machine Learning*, 8(1-2):1–230.

[31] Vershynin, R. (2018). *High-dimensional probability*, volume 47 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge. An introduction with applications in data science, With a foreword by Sara van de Geer.

[32] Yang, J., Ye, X., Trivedi, R., Xu, H., and Zha, H. (2018a). Deep mean field games for learning optimal behavior policy of large populations. In *International Conference on Learning Representations*.

[33] Yang, Y., Luo, R., Li, M., Zhou, M., Zhang, W., and Wang, J. (2018b). Mean field multi-agent reinforcement learning. In *International Conference on Machine Learning*, pages 5567–5576.

[34] Yin, H., Mehta, P. G., Meyn, S. P., and Shanbhag, U. V. (2013). Learning in mean-field games. *IEEE Transactions on Automatic Control*, 59(3):629–644.

[35] Zajkowski, K. (2018). Bounds on tail probabilities for quadratic forms in dependent sub-gaussian random variables. *arXiv preprint arXiv:1809.08569*.

Outline of the appendix. The structure of the appendix is the following. We start with the definition of a rigorous probabilistic framework for the idiosyncratic noise and the common noise in Section A. We then explain (see Section B) the key differences with the work of [10] to prove convergence of the exact PG method. In Section C, we provide the main ingredients for the proof of convergence of approximate PG with the MKV simulator. The most technical result of the present work is the convergence of population-based approximate PG in Theorem 6, for which we provide a sketch of proof in Section D before completing the details in Section E. Last, Section F collects the parameters used in the numerical tests. The interested reader who is not familiar with the techniques introduced in [10] could start with the sketch of proof provided in Section D to have a global view of the main ideas.

A Probabilistic setup

In this section we rigorously define the model of MKV dynamics introduced in § 2.1. A convenient way to think about this model is to view the state x_t of the system at time t as a random variable defined on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ where $\Omega = \Omega^0 \times \Omega^1$, $\mathcal{F} = \mathcal{F}^0 \times \mathcal{F}^1$ and $\mathbb{P} = \mathbb{P}^0 \times \mathbb{P}^1$. In this set-up, if $\omega = (\omega^0, \omega^1)$, $\epsilon_t^0(\omega) = \tilde{\epsilon}_t^0(\omega^0)$ and $\epsilon_t^1(\omega) = \tilde{\epsilon}_t^1(\omega^1)$ where $(\tilde{\epsilon}_t^0)_{t=1,2,\dots}$ and $(\tilde{\epsilon}_t^1)_{t=1,2,\dots}$ are i.i.d. sequences of mean-zero random variables on $(\Omega^0, \mathcal{F}^0, \mathbb{P}^0)$ and $(\Omega^1, \mathcal{F}^1, \mathbb{P}^1)$ respectively, while the initial sources of randomness $\tilde{\epsilon}_{t0}^0$ and $\tilde{\epsilon}_0^1$ are random variables on $(\Omega^0, \mathcal{F}^0, \mathbb{P}^0)$ and $(\Omega^1, \mathcal{F}^1, \mathbb{P}^1)$ with distributions $\tilde{\mu}_0^0$ and $\tilde{\mu}_0^1$ respectively, which are independent of each other and independent of $(\tilde{\epsilon}_t^0)_{t=1,2,\dots}$ and $(\tilde{\epsilon}_t^1)_{t=1,2,\dots}$. We denote by $\mathcal{F}_t$ the filtration generated by the noise up until time t, that is $\mathcal{F}_t = \sigma(\epsilon_0^0, \epsilon_0^1, \epsilon_1^0, \epsilon_1^1, \dots, \epsilon_t^0, \epsilon_t^1)$.

At each time $t \geq 0$, x_t and u_t are random elements defined on $(\Omega, \mathcal{F}, \mathbb{P})$ representing the state of the system and the control exerted by a generic agent. The quantities $\bar{x}_t$ and $\bar{u}_t$ appearing in (1) are the random variables on $(\Omega, \mathcal{F}, \mathbb{P})$ defined by: for $\omega = (\omega^0, \omega^1)$,

$$\bar{x}_t(\omega) = \int_{\Omega^1} x_t(\omega)\mathbb{P}(d\omega^1), \quad \bar{u}_t(\omega) = \int_{\Omega^1} u_t(\omega)\mathbb{P}(d\omega^1).$$

Notice that both $\bar{x}_t$ and $\bar{u}_t$ depend only upon ω^0. In fact, the best way to think of $\bar{x}_t$ and $\bar{u}_t$ is to keep in mind the following fact: $\bar{x}_t = \mathbb{E}[x_t|\mathcal{F}^0]$ and $\bar{u}_t = \mathbb{E}[u_t|\mathcal{F}^0]$. These are the mean field terms appearing in the (stochastic) dynamics of the state (1).

B Proof of Theorem 3

The structure of the proof is analogous to the one presented in Section D (see also [10, Theorem 7]). For the sake of brevity, we do not repeat all the details here and we focus on the differences in the key ingredients.

B.1 Proof of Lemma 1

Starting from this section, if there is no ambiguity, we will only state results related to the process $(y_t)_{t\geq 0}$ or to the parameter K. They can be extended to process $(z_t)_{t\geq 0}$ or to parameter L by simply replacing the corresponding parameters or operators (for example, we replace A by $A+\bar{A}$, and $C_y(K)$ by $C_z(L)$, etc).

Proof of Lemma 1. Consider a given $\theta = (K, L) \in \Theta$. The control process is $\mathbf{u} = (u_t)_{t\geq 0}$ with $u_t = \tilde{u}_\theta(x_t^\theta, \bar{x}_t^\theta)$ for all $t \geq 0$. We notice that for every $t \geq 0$, $u_t = -K(x_t^\theta - \bar{x}_t^\theta) - L\bar{x}_t^\theta = -Ky_t^\theta - Lz_t^\theta$, and

$$(u_t)^\top R u_t = (\mathbf{x}_t^\theta)^\top \begin{bmatrix} K^\top RK & K^\top RL \\ L^\top RK & L^\top RL \end{bmatrix} \mathbf{x}_t^\theta.$$

We can proceed similarly for the term with $\bar{R}$ in the cost. Moreover, $\mathbb{E}[y_t^\theta|\mathcal{F}^0] = 0$ and z_t^θ is $\mathcal{F}^0$-measurable, which implies $\mathbb{E}[(y_t^\theta)^\top K^T RL z_t^\theta|\mathcal{F}^0] = 0$, so that we deduce $\forall\, t \geq 0$,

$$\mathbb{E}\left[(u_t)^\top R u_t\right] = \mathbb{E}\left[\mathbb{E}[(u_t)^\top R u_t|\mathcal{F}^0]\right] = \mathbb{E}\left[(\mathbf{x}_t^\theta)^\top \mathbf{K}_\theta^\top \mathbf{R}\mathbf{K}_\theta\, \mathbf{x}_t^\theta\right].$$

Then the result holds. □

B.2 Gradient expression and domination

Lemma 9. *Under Assumption 2, the variance matrices $\Sigma_{y_0}, \Sigma_{z_0}, \Sigma^1, \Sigma^0$ have finite operator norms, namely there exists a finite constant $C_{0,var}$ such that*

$$\max\{\|\Sigma_{y_0}\|, \|\Sigma_{z_0}\|, \|\Sigma^1\|, \|\Sigma^0\|\} \leq C_{0,var}. \tag{8}$$

The proof of Lemma 9 is not difficult. It uses the inequality between the L^p norm and the sub-Gaussian norm. One can take $C_{0,var} = dC_{sub}C_0^2$ where C_{sub} is a universal constant.

For $\theta = (K, L)$, let us define the value function when the process $(\mathbf{x}_t^\theta)_{t\geq 0}$ starts at $\mathbf{x} \in \mathbb{R}^{2d}$ and follows (7) using control θ:

$$V_\theta(\mathbf{x}) = \sum_{t\geq 0} \gamma^t (\mathbf{x}_t^\theta)^\top \Gamma_\theta \mathbf{x}_t^\theta.$$

Then $C(\theta) = \mathbb{E}[V_\theta(\mathbf{x}_0)]$. We also define the quantity $\Sigma_\theta = \mathbb{E}\sum_{t\geq 0} \gamma^t \mathbf{x}_t^\theta (\mathbf{x}_t^\theta)^\top$. The noises in the dynamics of $y_t^\theta = y_t^K$ and $z_t^\theta = z_t^L$ are independent, so $\mathbb{E}[y_t^\theta z_t^\theta] = 0$ for all $t \geq 0$. Hence we have $\Sigma_\theta = \mathrm{diag}(\Sigma_K^y, \Sigma_L^z)$, where we introduced

$$\Sigma_K^y = \mathbb{E}\sum_{t\geq 0} \gamma^t y_t^K (y_t^K)^\top, \quad \Sigma_L^z = \mathbb{E}\sum_{t\geq 0} \gamma^t z_t^L (z_t^L)^\top.$$

We also introduces the initial variance terms

$$\Sigma_{y_0} = \mathbb{E}[y_0(y_0)^\top], \quad \Sigma_{z_0} = \mathbb{E}[z_0(z_0)^\top], \tag{9}$$

$$\Sigma^1 = \mathbb{E}[\epsilon_0^1(\epsilon_0^1)^\top], \quad \Sigma^0 = \mathbb{E}[\epsilon_0^0(\epsilon_0^0)^\top]. \tag{10}$$

Let us also consider two matrices $P_K, P_L \in \mathbb{R}^{d\times d}$ which are solutions to the following two algebraic Riccati equations

$$\left\{\begin{array}{l} P_K^y = Q + K^\top RK + \gamma(A-BK)^\top P_K^y(A-BK) \\ P_L^z = Q + \overline{Q} + L^\top(R+\bar{R})L \\ \qquad +\gamma(A+\overline{A}-(B+\bar{B})L)^\top P_L^z(A+\overline{A}-(B+\bar{B})L). \end{array}\right. \tag{11}$$

We can combine these two equations into one by introducing $\mathbf{P}_\theta = \begin{bmatrix} P_K^y & 0 \\ 0 & P_L^z \end{bmatrix}$ that satisfies

$$\mathbf{P}_\theta = \mathbf{Q} + \mathbf{K}^\top \mathbf{R}\mathbf{K} + \gamma(\mathbf{A}-\mathbf{B}\mathbf{K})^\top \mathbf{P}_\theta(\mathbf{A}-\mathbf{B}\mathbf{K}).$$

Let us also introduce $\alpha_\theta = \alpha_K^y + \alpha_L^z$, with

$$\alpha_K^y = \frac{\gamma}{1-\gamma}\mathbb{E}[(\epsilon_1^1)^\top P_K^y \epsilon_1^1] \qquad \alpha_L^z = \frac{\gamma}{1-\gamma}\mathbb{E}[(\epsilon_1^0)^\top P_L^z \epsilon_1^0]$$

For all $\mathbf{x}_0 = (y_0, z_0) \in \mathbb{R}^{2d}$, the value function can be expressed as:

$$V_\theta(\mathbf{x}_0) = \mathbf{x}_0^\top \mathbf{P}_\theta \mathbf{x}_0 + \alpha_\theta = y_0^\top P_K^y y_0 + \alpha_K^y + z_0^\top P_L^\mu z_0 + \alpha_L^z. \tag{12}$$

The following result provides a crucial expression for the gradient of the cost.

Lemma 10 (Policy gradient expression)**.** *For any $\theta = (K, L) \in \Theta$, we have*

$$\nabla_K C(\theta) = \nabla_K C_y(K) = 2E_K^y \Sigma_K^y,$$

where $E_K^y = (R + \gamma B^\top P_K^y B)K - \gamma B^\top P_K^y A$.

For the sake of completeness, we provide the proof, which is analogous to the one of [10, Lemma 1], except for the fact that we have noise in the dynamics and a discount factor.

Proof. Let us consider

$$C_y(K,\tilde{y}) = \mathbb{E}\sum_{t\geq 0}\gamma^t (y_t^K)^\top (Q + K^\top RK) y_t^K = \tilde{y}^\top P_K^y \tilde{y} + \alpha_K^y,$$

when $(y_t^K)_t$ starts from $y_0 = \tilde{y}$ and is controlled by K. We have

$$C_y(K) = \mathbb{E}_{y_0}[C_y(K, y_0)]. \tag{13}$$

Here, we use a subscript to denote the fact that the expectation is taken w.r.t. y_0. We also notice that

$$\begin{aligned} &C_y(K,\tilde{y}) \\ &= \mathbb{E}\sum_{t\geq 0}\gamma^t (y_t^K)^\top (Q + K^\top RK) y_t^K \\ &= \tilde{y}^\top (Q + K^\top RK)\tilde{y} + \mathbb{E}\sum_{t\geq 1}\gamma^t (y_t^K)^\top (Q + K^\top RK) y_t^K \\ &= \tilde{y}^\top (Q + K^\top RK)\tilde{y} + \gamma\mathbb{E}\left[C_y(K, (A-BK)y_0 + \epsilon_1^1) | y_0 = \tilde{y}\right]. \end{aligned} \tag{14}$$

Here, the conditional expectation notation means that $\tilde{y}$ is fixed while taking the expectation (on ϵ_1^1). To compute the gradient with respect to K, we note that, $\nabla_{\tilde{y}} C_y(K,\tilde{y}) = 2P_K^y\tilde{y}$ (since α_K^y does not depend upon the starting point $\tilde{y}$), and hence, using (14),

$$\begin{aligned} \nabla_K C_y(K,\tilde{y}) = 2RK\tilde{y}\tilde{y}^\top - 2\gamma B^\top P_K^y(A-BK)\tilde{y}\tilde{y}^\top \\ + \gamma\mathbb{E}\left[\nabla_K C_y(K,\tilde{y}')\big|_{\tilde{y}'=(A-BK)\tilde{y}+\epsilon_1^1}\right]. \end{aligned}$$

Expanding again the gradient in the above right hand side and using recursion leads to

$$\begin{aligned} \nabla_K C_y(K,\tilde{y}) = 2\left[(RK + \gamma B^\top P_K^y BK) - \gamma B^\top P_K^y A\right] \\ \cdot\left(\tilde{y}\tilde{y}^\top + \mathbb{E}\left[\sum_{t\geq 1}\gamma^t y_t^K (y_t^K)^\top\right]\right) \end{aligned}$$

Then we can have

$$\begin{aligned} \nabla_K C(\theta) = \nabla_K\left(C_y(K) + C_z(L)\right) = \nabla_K \mathbb{E}_{y_0}[C_y(K,y_0)] \\ = 2E_K^y\Sigma_K^y. \end{aligned}$$

The gradient of $\theta = (K,L) \mapsto C(\theta)$ w.r.t. L can be computed in a similar way. □

We can then follow the line of reasoning used in the proof of [10, Theorem 7] to each component (namely, y controlled by K and z controlled by L), and conclude the proof of Theorem 3. For the sake of brevity, we omit the details here are refer the interested reader to [10]. We simply mention how to obtain a few key estimates. Let us introduce

$$\lambda_y^1 = \lambda_{min}\left(\Sigma_{y_0} + \frac{\gamma}{1-\gamma}\Sigma^1\right), \tag{15}$$

$$\lambda_z^0 = \lambda_{min}\left(\Sigma_{z_0} + \frac{\gamma}{1-\gamma}\Sigma^0\right), \tag{16}$$

where we recall that Σ^1 and Σ^0 are defined by (10) and $\Sigma_{y_0}, \Sigma_{z_0}$ by (9).

Then, we have the following result, which generalizes [10, Lemma 13].

Lemma 11. *We have the bounds*

$$\|P_K^y\| \leq \frac{C_y(K)}{\lambda_y^1} \leq \frac{C(\theta)}{\lambda_y^1}, \quad \|\Sigma_K^y\| \leq \frac{C_y(K)}{\lambda_{min}(Q)} \leq \frac{C(\theta)}{\lambda_{min}(Q)}. \tag{17}$$

We have similar upper bounds for P_L^z and Σ_L^z by changing λ_y^1 to λ_z^0, $C_y(K)$ to $C_z(L)$, and Q to $Q + \bar{Q}$.

Another important observation is that the following operators $\mathcal{F}_K^y$,$\mathcal{F}_L^z$,$\mathcal{T}_K^y$,$\mathcal{T}_L^z$ defined on the set of matrices $X \in \mathbb{R}^{d\times d}$ are bounded:

$$\mathcal{F}_K^y(X) = \gamma(A-BK)X(A-BK)^\top, \tag{18}$$

$$\mathcal{T}_K^y(X) = \sum_{t\geq 0}\gamma^t (A-BK)^t X((A-BK)^\top)^t = \sum_{t\geq 0}(\mathcal{F}_K^y)^t(X).$$

and $\mathcal{F}_L^z$ and $\mathcal{T}_L^z$ can be defined similarly. In fact, it is straightforward to check that

$$\|\mathcal{T}_K^y\| \le \frac{1}{1-\gamma_{1,K}} < \infty, \tag{19}$$

with $\gamma_{1,K} = \gamma\|A - BK\|^2 < 1$.

The following lemma helps us to express the variance matrices Σ_K^y and Σ_L^z in terms of the operators $\mathcal{T}_K^y$ and $\mathcal{T}_L^z$.

Lemma 12. *If $\theta \in \Theta$, then*

$$\Sigma_K^y = \mathcal{T}_K^y\left(\Sigma_{y_0} + \frac{\gamma}{1-\gamma}\Sigma^1\right), \tag{20}$$

$$\Sigma_L^z = \mathcal{T}_L^z\left(\Sigma_{z_0} + \frac{\gamma}{1-\gamma}\Sigma^0\right). \tag{21}$$

Proof. We start by observing that for any $t \ge 1$,

$$\begin{aligned}
&\mathbb{E}\left[\gamma^t y_t^K (y_t^K)^\top\right] \\
=&\gamma(A-BK)\mathbb{E}\left[\gamma^{t-1} y_{t-1}^K (y_{t-1}^K)^\top\right](A-BK)^\top \\
&+\gamma^t \mathbb{E}[\epsilon_t^1(\epsilon_t^1)^\top] \\
=&(\mathcal{F}_K^y)^t(\Sigma_{y_0}) + \sum_{s=0}^{t-1}(\mathcal{F}_K^y)^s\left(\gamma^{t-s}\Sigma^1\right)
\end{aligned}$$

where the second equality is justified by $\mathbb{E}[\epsilon_t^1(\epsilon_t^1)^\top] = \mathbb{E}[\epsilon_1^1(\epsilon_1^1)^\top]$. So, we have

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=0}^{T-1}\gamma^t y_t^K (y_t^K)^\top\right] \\
=&\sum_{t=0}^{T-1}(\mathcal{F}_K^y)^t(\Sigma_{y_0}) + \sum_{t=1}^{T-1}\left(\sum_{s=0}^{t-1}(\mathcal{F}_K^y)^s(\gamma^{t-s}\Sigma^1)\right) \\
=&\mathcal{T}_K^y\left(\Sigma_{y_0} + \frac{\gamma-\gamma^T}{1-\gamma}\Sigma^1\right) + \sum_{s=1}^{T-1}\left(\gamma^s\sum_{t=0}^{T-1-s}(\mathcal{F}_K^y)^t(\Sigma^1)\right)
\end{aligned}$$

where the second equality is justified by exchanging the summations. From the linearity of operators $\mathcal{F}_K^y$ and $\mathcal{T}_K^y$, and observing the fact that for any symmetric matrix X, $\mathcal{F}_K^y(\mathcal{T}_K^y(X)) = \mathcal{T}_K^y(\mathcal{F}_K^y(X))$, we have

$$\begin{aligned}
&\mathcal{T}_K^y\left(\Sigma_{y_0} + \frac{\gamma}{1-\gamma}\Sigma^1\right) - \mathbb{E}\left[\sum_{t=0}^{T-1}\gamma^t y_t^K (y_t^K)^\top\right] \\
&= \frac{\gamma^T}{1-\gamma}\mathcal{T}_K^y\left(\Sigma^1\right) + (\mathcal{F}_K^y)^T(\mathcal{T}_K^y(\Sigma_{y_0})) \\
&+ \sum_{s=1}^{T-1}\gamma^s(\mathcal{F}_K^y)^{T-s}\left(\mathcal{T}_K^y(\Sigma^1)\right),
\end{aligned} \tag{22}$$

which is a positive semi-definite matrix, because $\mathcal{F}_K^y(\Sigma)$ and $\mathcal{T}_K^y(\Sigma)$ are positive semi-definite for any positive semi-definite $\Sigma \in \mathbb{R}^{d\times d}$.

Because the operator norms of $\mathcal{T}_K^y$ and $\mathcal{F}_K^y$ are bounded by $1/(1-\gamma_{1,K})$ and $\gamma_{1,K}$ respectively, where $\gamma_{1,K} = \gamma\|A - BK\|^2 < 1$, we have

$$\begin{aligned}
&\left\|\sum_{s=1}^{T-1}\gamma^s(\mathcal{F}_K^y)^{T-s}\left(\mathcal{T}_K^y(\Sigma^1)\right)\right\| \\
&\le \|\Sigma^1\|\frac{(\gamma_{1,K})^T}{1-\gamma_{1,K}}\sum_{s=1}^{T-1}(\gamma/\gamma_{1,K})^s \xrightarrow[T\to\infty]{} 0,
\end{aligned}$$

and $\left\|(\mathcal{F}_K^y)^T(\mathcal{T}_K^y(\Sigma_{y_0}))\right\| \le \frac{(\gamma_{1,K})^T}{1-\gamma_{1,K}}\|\Sigma_{y_0}\| \to 0$. Thus, for any vector $\zeta \in \mathbb{R}^d$,

$$\begin{aligned}
&\zeta^\top\left(\mathcal{T}_K^y\left(\Sigma_{y_0} + \frac{\gamma}{1-\gamma}\Sigma^1\right) - \mathbb{E}\left[\sum_{t=0}^{T-1}\gamma^t y_t^K (y_t^K)^\top\right]\right)\zeta \\
&\le Tr(\zeta\zeta^\top)\left\|\mathcal{T}_K^y\left(\Sigma_{y_0} + \frac{\gamma}{1-\gamma}\Sigma^1\right) - \mathbb{E}\left[\sum_{t=0}^{T-1}\gamma^t y_t^K (y_t^K)^\top\right]\right\| \\
&\xrightarrow{T\to\infty} 0.
\end{aligned}$$

Meanwhile, the matrices $\mathbb{E}[\gamma^t y_t^K (y_t^K)^\top]$ for all $t \ge 0$ are also positive semi-definite , so that we have

$$\begin{aligned}
&\zeta^\top\left(\Sigma_K^y - \mathbb{E}\left[\sum_{t=0}^{T-1}\gamma^t y_t^K (y_t^K)^\top\right]\right)\zeta \\
=&Tr\left((\zeta\zeta^\top)\mathbb{E}\left[\sum_{t=T}^{\infty}\gamma^t y_t^K (y_t^K)^\top\right]\right) \\
\le&Tr(\zeta\zeta^\top)\left\|\mathcal{T}_K^y\left(\mathbb{E}\left[\gamma^T y_T^K (y_T^K)^\top\right]\right)\right\| \\
\le&Tr(\zeta\zeta^\top)\frac{1}{1-\gamma_{1,K}}\left(\gamma_{1,K}^T\|\Sigma_{y_0}\| + \|\Sigma^1\|\gamma^T\sum_{s=0}^{T-1}(\gamma_1/\gamma)^s\right) \\
&\xrightarrow{T\to\infty} 0.
\end{aligned}$$

Hence, the result follows. □

In particular, if the control parameters is admissible, then the cost is finite.

Corollary 13. *If $\theta = (K, L) \in \Theta$, namely $\gamma\|A-BK\|^2 < 1$ and $\gamma\|A + \bar{A} - (B+\bar{B})L\|^2 < 1$, then $C(\theta) < \infty$.*

The following lemma estimates the difference between the cost $C(\theta)$ and the optimal value of C by means of gradients. It has been referred to as the "gradient domination" lemma in [10], which serves as one of the cornerstones in proving the convergence of the algorithms. This result is in line with the Polyak- Lojasiewicz condition for gradient descent in convex optimization ([17]).

Lemma 14 (Gradient domination)**.** *Let $\theta = (K, L) \in \Theta$ and $\theta^* = (K^*, L^*) \in \Theta$ be the optimal parameters. Then*

$$\begin{aligned} &C_y(K) - C_y(K^*) \\ \leq &\frac{\|\Sigma^y_{K^*}\|}{4\lambda_{min}(R)\lambda_{min}(\Sigma^y_K)^2} Tr(\nabla_K C_y(K)^\top \nabla_K C_y(K)), \end{aligned} \quad (23)$$

and

$$C_y(K) - C_y(K^*) \geq \frac{\lambda_{min}(\Sigma^y_K)}{\|R + \gamma B^\top P^y_K B\|} Tr((E^y_K)^\top E^y_K). \quad (24)$$

B.3 Some results from perturbation analysis

We study the effect of perturbing the parameter $\theta = (K, L)$ on the value function $C(\theta)$, on the gradients $(\nabla_K C(\theta), \nabla_L C(\theta))$, and on the variance matrices Σ^y_K and Σ^z_L. In this section, we provide several lemmas in this direction which generalize results of [10] to our setting. Among other things, we need to deal with the sources of noise in the dynamics and the discount factor. The following lemma generalizes [10, Lemmas 18, 19 and 20]. For the sake of completeness, we present the adapted version to our model here.

Lemma 15. *For any $\theta = (K, L), \theta' = (K', L') \in \Theta$, we have $\mathcal{T}^y_K = (\mathbf{I} - \mathcal{F}^y_K)^{-1}$, and*

$$\begin{aligned} \|\mathcal{F}^y_K - \mathcal{F}^y_{K'}\| \leq \gamma \left(2\|A - BK\| + \|B\|\|K' - K\|\right) \\ \cdot \|B\|\|K' - K\|, \end{aligned} \quad (25)$$

Moreover, for any $\eta < 1$ such that $\|\mathcal{T}^y_K\|\|\mathcal{F}^y_K - \mathcal{F}^y_{K'}\| \leq \eta$. For any positive semi-definite matrix $X \in \mathbb{R}^{d\times d}$, we have

$$\begin{aligned} \|(\mathcal{T}^y_K - \mathcal{T}^y_{K'})(X)\| &\leq \frac{1}{1-\eta}\|\mathcal{T}^y_K\|\|\mathcal{F}^y_K - \mathcal{F}^y_{K'}\|\|\mathcal{T}^y_K(X)\| \\ &\leq \frac{\eta}{1-\eta}\|\mathcal{T}^y_K(X)\|, \end{aligned} \quad (26)$$

The following lemma is analogous to a key step in the proof of [10, Lemma 17].

Lemma 16. *For every matrix $X \in \mathbb{R}^{d\times d}$ satisfying $\|X\| = 1$, we have for any positive semi-definite matrix Σ:*

$$\|\mathcal{T}^y_K(X)\| \leq \|\Sigma^{-1/2} X \Sigma^{-1/2}\|\|\mathcal{T}^y_K(\Sigma)\| \quad (27)$$

If we choose $\Sigma = \Sigma_{y_0} + \frac{\gamma}{1-\gamma}\Sigma^1$ or $\Sigma = \Sigma_{z_0} + \frac{\gamma}{1-\gamma}\Sigma^0$ in Lemma 16, then together with Lemma 12, we obtain the following bounds on the operators $\mathcal{T}^y_K$ and $\mathcal{T}^z_L$, which refines [10, Lemma 17].

Corollary 17. *We have*

$$\|\mathcal{T}^y_K\| \leq \frac{C_y(K)}{\lambda^1_y \lambda_{min}(Q)} \leq \frac{C(\theta)}{\lambda^1_y \lambda_{min}(Q)} \quad (28)$$

where we recall that λ^1_y is defined by (15)*.*

The following lemma provides a bound on the perturbation of the variance matrix.

Lemma 18. *Let $\theta, \theta' \in \Theta$ and set $\Delta K = K' - K$ Let $\zeta_1 < 1/2$. If*

$$\|\Delta K\| \leq \frac{\zeta_1}{\|B\| \max\{1, \|\mathcal{T}^y_K\|\}} \frac{1}{\|A - BK\| + 1}, \quad (29)$$

then

$$\begin{aligned} &\|\Sigma^y_K - \Sigma^y_{K'}\| \\ &\leq \left(\frac{2\gamma}{1 - 2\gamma\zeta_1}\|\mathcal{T}^y_K\| \cdot \|B\|(\|A - BK\| + 1)\|\Sigma^y_K\|\right)\|\Delta K\|. \end{aligned} \quad (30)$$

Proof. From (29) we know that $\|B\|\|\Delta K\| \leq \zeta_1 \leq 1/2$, thus

$$\begin{aligned} &\|\mathcal{T}^y_K\|\|\mathcal{F}^y_K - \mathcal{F}^y_{K'}\| \\ \leq &\gamma\|\mathcal{T}^y_K\|\left(2\|A - BK\| + \zeta_1\right)\frac{\zeta_1}{\max\{1, \|\mathcal{T}^y_K\|\}}\frac{1}{\|A - BK\| + 1} \\ \leq &2\gamma\zeta_1. \end{aligned}$$

Now, from Lemma 12, and inequality (26) with $\eta = 2\gamma\zeta_1$, we obtain

$$\begin{aligned} &\|\Sigma^y_K - \Sigma^y_{K'}\| \\ = &\left\|(\mathcal{T}^y_K - \mathcal{T}^y_{K'})\left(\Sigma_{y_0} + \frac{\gamma}{1-\gamma}\Sigma^1\right)\right\| \\ \leq &\frac{1}{1 - 2\gamma\zeta_1}\|\mathcal{T}^y_K\| \cdot \|\mathcal{F}^y_K - \mathcal{F}^y_{K'}\| \cdot \left\|\mathcal{T}^y_K\left(\Sigma_{y_0} + \frac{\gamma}{1-\gamma}\Sigma^1\right)\right\| \\ \leq &\frac{2\gamma}{1 - 2\gamma\zeta_1}\|\mathcal{T}^y_K\| \cdot (\|A - BK\| + 1)\|B\|\|\Delta K\| \cdot \|\Sigma^y_K\|, \end{aligned}$$

which yields the conclusion. □

Combining Lemma 18 with Corollary 17, we deduce the following useful consequence, which generalizes [10, Lemma 16]. Let us introduce the functions

$$h^y_{cond}(\theta) := 4\left(\frac{C(\theta)}{\lambda^1_y \lambda_{min}(Q)}\right)\|B\|(\|A - BK\| + 1) \quad (31)$$

$$\begin{aligned} h^z_{cond}(\theta) := 4\left(\frac{C(\theta)}{\lambda^0_z \lambda_{min}(Q + \bar{Q})}\right)\|B + \bar{B}\| \\ \cdot (\|A + \bar{A} - (B + \bar{B})L\| + 1). \end{aligned} \quad (32)$$

Note that these functions have a most polynomial growth in $\|\theta\|$.

Corollary 19. *Let $\theta, \theta' \in \Theta$ and set $\Delta K = K' - K$. If we choose $\zeta_1 = 1/4$ in Lemma 18, and assume that*

$$\|\Delta K\| \leq 1/h^y_{cond}(\theta), \qquad \|\Delta L\| \leq 1/h^z_{cond}(\theta). \quad (33)$$

Then, together with Lemma 11 and Corollary 17, we have

$$\|\Sigma^y_K - \Sigma^y_{K'}\| \le \left(\frac{C(\theta)}{\lambda_{min}(Q)}\right) h^y_{cond}(\theta)\|\Delta K\|, \tag{34}$$

and a similar bound for $\|\Sigma^z_L - \Sigma^z_{L'}\|$.

We now show that if $\theta \in \Theta$, then θ' produced by one step of policy gradient is still in Θ. Let us introduce

$$\rho(\mathcal{F}^y_K) = \|\gamma(A-BK)^2\|, \qquad \rho(\mathcal{F}^z_L) = \|\gamma(A-(B+\bar{B})L)^2\|.$$

With this notation, saying that $\theta \in \Theta$ amounts to say that $\rho(\mathcal{F}^y_K) < 1$ and $\rho(\mathcal{F}^z_L) < 1$. We start with the following result.

Lemma 20. *If* $\rho(\mathcal{F}^y_K) < 1$, *then*

$$Tr(\Sigma^y_K) \ge \frac{\lambda^1_y}{1-\gamma\|(A-BK)\|^2}.$$

We then show that any small enough perturbation of the control parameter preserves the property of lying in Θ. Recall that h^y_{cond}, h^z_{cond} are defined by (31).

Lemma 21. *If* $\rho(\mathcal{F}^y_K) < 1$ *and* $\|K' - K\| \le 1/h^y_{cond}(\theta)$, *then* $\rho(\mathcal{F}^y_{K'}) < 1$.

The proof to Lemma 21 is by contradiction. Then the main idea is to use the continuity of spectral radius. Suppose that there exits a K' satisfying condition $\|K'-K\| \le 1/h^y_{cond}(\theta)$ but $\rho(\mathcal{F}^y_{K'}) \ge 1$. So we have $\|A - BK'\| \ge \sqrt{1/\gamma}$ and $\|A - BK\| \le \sqrt{1/\gamma\,(1-\tilde{\epsilon})}$ for some well chosen $\tilde{\epsilon} > 0$. The choice of new parameter K'' such that $\|A - BK''\| = \sqrt{1/\gamma\,(1-\tilde{\epsilon}/2)}$ will raise contradiction.

The following Lemma 22 is adapted from the proof of [10, Lemma 24].

Lemma 22. *Let* $\theta, \theta' \in \Theta$ *and set* $\Delta K = K' - K$. *Assume that the conditions* (29) *for* $\|\Delta K\|$ *in Lemma 18 hold, then*

$$\|P^y_{K'} - P^y_K\| \le \|\Delta K\|\left[\frac{\|\mathcal{T}^y_K\|}{1-2\gamma\zeta_1}\Big(\gamma\|P^y_K\|\|B\|(2\|A-BK\| + \zeta_1) + (2\|K\| + \zeta_1/\|B\|)\,\|R\|\Big)\right]. \tag{35}$$

From here, we can bound the variations in P^y_K and P^z_L respectively by the variations of the first and second components of the control, and thus obtain a bound on the variation of the cost function.

Let us define

$$h^y_{riccati}(\theta) := \frac{2C(\theta)}{\lambda_{min}(Q)\lambda^1_y}\Big(2\frac{C(\theta)}{\lambda^1_y}\|B\|(\|A-BK\|+1) + 2\|K\|\|R\| + \frac{\|R\|}{4\|B\|}\Big) \tag{36}$$

and $h^z_{riccati}(\theta)$ similarly, and

$$h_{func,0}(\theta) := Tr\left(\Sigma_{y_0} + \frac{\gamma}{1-\gamma}\Sigma^1\right) h^y_{riccati}(\theta) + Tr\left(\Sigma_{z_0} + \frac{\gamma}{1-\gamma}\Sigma^0\right) h^z_{riccati}(\theta). \tag{37}$$

Under Assumption 2 and Lemma 9, we have

$$h_{func,0}(\theta) \le \frac{dC_{0,var}}{1-\gamma}\Big(h^y_{riccati}(\theta) + h^z_{riccati}(\theta)\Big).$$

Corollary 23. *Let* $\theta, \theta' \in \Theta$ *and set* $\Delta K = K' - K$ *We assume that* $\|\Delta K\|$ *satisfies condition* (33). *Then*

$$\|P^y_{K'} - P^y_K\| \le h^y_{riccati}(\theta)\|\Delta K\|. \tag{38}$$

The proof is simple, we can choose $\zeta_1 = 1/4$ in Lemma 22 and notice that $\gamma/(1-2\gamma\zeta_1) \le 2$.

Corollary 23 implies the following 2 results.

Lemma 24 (Perturbation of mean field cost function)**.** *Let* $\theta, \theta' \in \Theta$ *and set* $\Delta K = K' - K$ *and* $\Delta L = L' - L$. *We assume that* $\|\Delta K\|$ *and* $\|\Delta L\|$ *satisfy condition* (33). *Then*

$$|C(\theta') - C(\theta)| \le h_{func,0}(\theta)\max\{\|\Delta K\|, \|\Delta L\|\}. \tag{39}$$

Let us define

$$h_{C,cond}(\theta) := \frac{h_{func,0}(\theta)}{C(\theta)} \tag{40}$$

which still has at most polynomial growth in $\|K\|, \|L\|, 1/\lambda^1_y, 1/\lambda^0_z$ and $C(\theta)$.

Corollary 25. *We assume conditions in Lemma* (24). *Moreover, if* $\max\{\|\Delta K\|, \|\Delta L\|\} \le 1/h_{C,cond}(\theta)$, *then* $|C(\theta') - C(\theta)| \le C(\theta)$.

The following lemma is related to the perturbation of gradient. It is adapted from [10, Lemma 28] with small modifications. Let us define

$$\begin{aligned} &h^y_{grad}(\theta) \\ &:= \frac{2C(\theta)}{\lambda_{min}(Q)}\Bigg[\|R\| + \gamma h^y_{riccati}(\theta)\|B\|(1+\|A-BK\|) \\ &+ (\gamma\|B\|^2 + \sqrt{\gamma}\|B\|h^y_{cond}(K)C(\theta)/\lambda_{min}(Q))\left(\frac{C(\theta)}{\lambda^1_y}\right) \\ &+ \frac{h^y_{cond}(K)}{2\sqrt{\gamma}}\frac{C(\theta)}{\lambda_{min}(Q)}\frac{\|R\|}{\|B\|}\Bigg] \end{aligned} \tag{41}$$

and similarly for $h^z_{grad}(\theta)$. Note that $h^y_{grad}(\theta)$ and $h^z_{grad}(\theta)$ have at most polynomial growth in $C(\theta), \|K\|, \|L\|, 1/\lambda^1_y, 1/\lambda^0_z$.

Lemma 26. *Let $\theta, \theta' \in \Theta$ and set $\Delta K = K' - K$. Suppose that $\|\Delta K\|$ satisfies condition (33). Then*

$$\|\nabla_K C(\theta') - \nabla_K C(\theta)\| \leq h^y_{grad}(\theta)\|\Delta K\|. \tag{42}$$

To conclude this subsection, we analyse the variation in terms of the cost function when the learning rate is small enough.

Let us define

$$h^y_{lrate}(\theta, \|\nabla\|) := \max\left\{ \frac{6C(\theta)}{\lambda_{min}(Q)}\left(\|R\| + \gamma\|B\|^2\frac{C(\theta)}{\lambda^1_y}\right), 2\|\nabla\| h^y_{cond}(\theta)\left(\frac{C(\theta)}{\lambda_{min}(Q)\lambda^1_y}\right)\right\}$$

and similarly for $h^z_{lrate}(\theta)$. We also define

$$h_{lrate}(\theta) := \max\Big\{ h^y_{lrate}(\theta, \|\nabla_K C(\theta)\|), h^z_{lrate}(\theta, \|\nabla_L C(\theta)\|)\Big\}. \tag{43}$$

B.4 Key step in the proof of Theorem 3

The following lemma is the main key step in proving Theorem 3. It generalizes [10, Lemma 21].

Lemma 27 (One step descent in exact Policy Gradient algorithm)**.** *Suppose that $\mathbf{K}' = \mathbf{K} - \eta\nabla C(\theta)$ where the learning rate $\eta > 0$ satisfies $\eta \leq 1/h_{lrate}(\theta)$. Then*

$$C(\theta') - C(\theta^*) \leq (1 - \eta h_{update})(C(\theta) - C(\theta^*)), \tag{44}$$

where h_{update} is defined by

$$h_{update} := \min\left\{\lambda_{min}(R)\frac{(\lambda^1_y)^2}{\|\Sigma^y_{K^*}\|}, \lambda_{min}(R + \bar{R})\frac{(\lambda^0_z)^2}{\|\Sigma^z_{L^*}\|}\right\}$$

Proof. The proof relies on the gradient domination inequalities (23). Similarly to [10, Lemma 21], it can be shown that

$$C_y(K') - C_y(K) \leq -4\eta h_{update}(C_y(K) - C_y(K^*)) \cdot \left(1 - \frac{\|\Sigma^y_{K'} - \Sigma^y_K\|}{\lambda^1_y} - \eta\|\Sigma_{K'}\|\|R + \gamma B^\top P^y_K B\|\right)$$

From Corollary 19, with a small enough learning rate $\eta > 0$, we have

$$\|\Sigma^y_{K'} - \Sigma^y_K\| \leq \xi^y_2\|\Sigma^y_K\|$$

where $\xi^y_2 = \eta h^y_{cond}(\theta)\|\nabla_K C(\theta)\|$. So, if we assume that

$$\eta \leq 1/h^y_{lrate}(\theta, \|\nabla_K C(\theta)\|),$$

then $\|\Sigma^y_{K'} - \Sigma^y_K\|/\lambda^1_y \leq \xi^y_2\|\Sigma^y_K\|/\lambda^1_y \leq \frac{1}{2}$ and

$$\eta\|\Sigma_{K'}\|\|R + \gamma B^\top P^y_K B\| \leq \eta(\xi^y_2 + 1)\|\Sigma^y_K\|\|R + \gamma B^\top P^y_K B\| \leq \frac{1}{4}$$

where the last inequality comes from $\xi^y_2 \leq \frac{1}{2}\lambda^1_y/\|\Sigma^y_K\| \leq \frac{1}{2}$. Thus, we have

$$C_y(K') - C_y(K^*) \leq (1 - \eta h_{update})(C_y(K) - C_y(K^*)).$$

A similar inequality holds for $C_z(L') - C_z(L^*)$. This yields the conclusion. □

Remark 28. *We can bound the operator norm of gradients $\|\nabla_K C(\theta)\|$ and $\|\nabla_L C(\theta)\|$ by polynomials in $C(\theta), 1/\lambda^1_y, 1/\lambda^0_z$ with the help of inequalities (24) (See [10, Lemma 22] for more details). By this means, the function $h_{lrate}(\theta)$ can be bounded from above by a polynomial in $C(\theta), \|K\|, \|L\|, 1/\lambda^1_y, 1/\lambda^0_z$.*

C Proof of Theorem 5

C.1 Intuition for zeroth order optimization

Derivative-free optimization (see e.g. [8, 24]) tries to optimize a function $f : x \mapsto f(x)$ using only pointwise values of f, without having access to its gradient. The basic idea is to express the gradient of f at a point x by using values of $f(\cdot)$ at a Gaussian perturbation region around point x. This can be achieved by introducing an approximation of f using a Gaussian smoothing. Formally, if we define the perturbed function for some $\sigma > 0$ as $f_{\sigma^2}(x) = \mathbb{E}[f(x + \epsilon)], \epsilon \sim \mathcal{N}(0, \sigma^2 I)$, then its gradient can be written as $\frac{d}{dx} f_{\sigma^2}(x) = \frac{1}{\sigma^2}\mathbb{E}[f(x + \epsilon)\epsilon]$.

For technical reasons, if the function f is not defined for every x (in our case $C(\theta) = \infty$ for some $\theta = (K, L)$), we can replace the Gaussian smoothing term ϵ with variance $\sigma^2 I$ by a uniform random variable U in the ball $\mathbb{B}_\tau$ with radius τ small enough.

C.2 Estimation of $\nabla C(\theta)$ in policy gradient

For any $\theta = (K, L)$ and $\tau > 0$, we introduce the following smoothed version C_τ of C defined by

$$C_\tau(\theta) = \mathbb{E}_U[C(\theta + U)] = \mathbb{E}_{(U_1,U_2)}[C_y(K + U_1) + C_z(L + U_2)] \tag{45}$$

where $U = (U_1, U_2)$ with U_1, U_2 independent random variables uniformly distributed in $\mathbb{B}_\tau$. The notation $\mathbb{E}_U$ means that the expectation is taken with respect to the random variable U.

One can prove the following result, where intuitively V_i plays the role of $\tau\frac{U_i}{\|U_i\|}$, for $i = 1, 2$.

Lemma 29. *We have*

$$\nabla_K C_\tau(\theta) = \frac{d}{\tau^2}\mathbb{E}_{V_1}[C_y(K + V_1)V_1] = \frac{d}{\tau^2}\mathbb{E}_V[C(\theta + V)V_1], \tag{46}$$

and

$$\nabla_L C_\tau(\theta) = \frac{d}{\tau^2}\mathbb{E}_{V_2}[C_z(L + V_2)V_2] = \frac{d}{\tau^2}\mathbb{E}_V[C(\theta + V)V_2], \tag{47}$$

where $V = (V_1, V_2) \in \mathbb{R}^{\ell\times d} \times \mathbb{R}^{\ell\times d}$ with V_1, V_2 i.i.d. random variables uniformly distributed on $\mathbb{S}_\tau$.

The proof is omitted for brevity. It is similar to the one of [11, Lemma 2.1] (see also [10, Lemma 26]), except that we have two components. The last equality in (46) and (47) uses the fact that V_1, V_2 are independent and have zero-mean.

Remark 30. *Although we have rephrased the initial MFC problem as a control problem in a higher dimension, see 1, the naive idea of approximating the gradient of the cost by perturbing $\mathbf{K} = \mathrm{diag}(K, L) \in \mathbb{R}^{(2\ell)\times(2d)}$ using a single smoothing random variable $U \in \mathbb{R}^{(2\ell)\times(2d)}$ would not work because we want to preserve the block-diagonal structure of the matrix during the the PG algorithm. We hence exploit the fact that the exact PG convergence result relies on the decomposition of the value function $\theta \mapsto C(\theta)$ into two auxiliary value functions $K \mapsto C_y(K)$ and $L \mapsto C_z(L)$ associated respectively to processes $(y_t^K)_{t\geq 0}$ and $(z_t^L)_{t\geq 0}$ and we then introduce two independent smoothing random variables U_1 and U_2 in $\mathbb{R}^{\ell\times d}$ corresponding to K and L respectively, instead of using a single smoothing random variable $U \in \mathbb{R}^{(2\ell)\times(2d)}$ for $\mathbf{K}$. This is also consistent with the fact that, in the model free setting, we can only access to the total value $C(\theta)$ and we have no information about the values of $C_y(K)$ and $C_z(L)$ separately.*

Now we can define the following notation

$$\nabla C_\tau(\theta) = \begin{bmatrix} \nabla_K C_\tau(\theta) & 0 \\ 0 & \nabla_L C_\tau(\theta) \end{bmatrix} \in \mathbb{R}^{2\ell\times 2d}. \tag{48}$$

The following lemma stems from the first part of [10, Lemma 27] and provides upper bounds for this perturbed estimate. Let us define

$$h_{pert,radius}(\theta,\epsilon) := \max\Big\{h^y_{cond}(\theta), h^z_{cond}(\theta), h_{C,cond}(\theta), \frac{4}{\epsilon}h^y_{grad}(\theta), \frac{4}{\epsilon}h^z_{grad}(\theta)\Big\} \tag{49}$$

where the terms $h^y_{cond}(\theta)$, $h^z_{cond}(\theta)$, $h_{C,cond}(\theta)$, $h^y_{grad}(\theta)$, $h^z_{grad}(\theta)$ have at most polynomial growth in $C(\theta)$, $\|K\|$, $\|L\|$, $1/\lambda^1_y$, $1/\lambda^0_z$ given by (31), (40), and (41). We also define

$$h_{pert,size}(\theta,d,\epsilon) := \frac{2}{(\epsilon/4)^2}\left((d+1)\log\left(\frac{d}{\epsilon}\right) + \log(4\epsilon)\right) \cdot \left(\frac{4d^2}{\tau^2}C(\theta)^2 + \frac{2dC(\theta)}{3\tau}\frac{\epsilon}{4}\right). \tag{50}$$

Lemma 31. *For every given $\theta = (K, L) \in \Theta$, assume that the number of perturbation directions M is large enough, and all perturbed parameters $\theta_i = (K_i, L_i)$ satisfy*

$$\max\{\|K_i - K\|, \|L_i - L\|\} \leq \tau$$

for some $\tau > 0$ small enough, and $V_{i1} = K_i - K, V_{i2} = L_i - L$ for $i = 1, \ldots, M$ are all independent random matrices uniformly distributed on the sphere $\mathbb{S}_\tau$. Then, with high probability, the perturbed policy gradient $\hat{\nabla}C(\theta)$ defined by

$$\hat{\nabla}C(\theta) = \frac{d}{\tau^2}\frac{1}{M}\sum_{i=1}^{M}\left(\sum_{t=0}^{\infty}\mathbb{E}\left[c^i_{t+1}\right]\right)\begin{bmatrix} K_i - K & 0 \\ 0 & L_i - L \end{bmatrix}$$

is close to the true gradient $\nabla C(\theta)$ in operator norm. Here $\sum_{t=0}^{\infty}\mathbb{E}[c^i_{t+1}] = C(\theta + V_i)$.

More precisely, for some given $\epsilon > 0$, if

$$\tau \leq \frac{1}{h_{pert,radius}(\theta,\epsilon)} \tag{51}$$

where $h_{pert,radius}(\theta,\epsilon)$ defined by (49) has at most polynomial growth in $1/\epsilon, C(\theta), \|K\|, \|L\|, 1/\lambda^1_y, 1/\lambda^0_z$, and if

$$M \geq h_{pert,size}(\theta,d,\epsilon), \tag{52}$$

then with probability at least $1 - (d/\epsilon)^{-d}$, we have

$$\|\hat{\nabla} - \nabla\| \leq \epsilon. \tag{53}$$

where $\hat{\nabla}$ and ∇ are short for $\hat{\nabla}C(\theta)$ and $\nabla C(\theta)$.

Notice that it is enough to take the number of perturbation directions M of size $\mathcal{O}\left(d\log\left(\frac{d}{\epsilon}\right)\left(\frac{dC(\theta)}{\tau\epsilon}\right)^2\right)$, which is the same as the bound obtained in [10, Lemma 27].

We should point out here that the randomness of $\hat{\nabla}$ comes from the bounded random variables $(V_{i1})_i$ and $(V_{i2})_i$ for $i = 1, \ldots, M$. The proof uses the same techniques as shown in [10, Lemma 27], and it is based on a version of Bernstein's concentration inequality applying on bounded random matrices (see [30] for more details). The main idea is to split the difference $\|\hat{\nabla}C - \nabla C\|$ with the help of Lemma 29

$$\|\hat{\nabla} - \nabla\| \leq \left\|\hat{\nabla}C(\theta) - \frac{1}{M}\sum_{i=1}^{M}\nabla C_\tau(\theta_i)\right\| + \|\nabla C_\tau(\theta) - \nabla C(\theta)\|$$

and observe that the operator norm of random matrices $C(\theta_i)V_{i1}$ and their second order moment $\mathbb{E}[C(\theta_i)^2 V_{i1}V_{i1}^\top]$ are bounded by $2\tau C(\theta)$ and $4\tau^2 C(\theta)^2$ respectively.

C.3 Approximation with finite horizon

In this section, we explain how we approximate the cost functions and the variance matrices with a finite time horizon. Because of the presence of noise processes within the underlying dynamics, the variance Σ^y_K cannot be simply expressed as the sum of a power series of operator $\mathcal{F}^y_K$ applied on the initial variance matrix Σ_{y_0} (see also Lemma 12), the arguments used in [10, Lemma 23] for finite horizon

approximation fail to work. We prove here a new lemma for choosing the truncation horizon parameter T.

We first fix some notation. For $T \geq 0$, let us define the truncated variances and the truncated costs:

$$\Sigma_K^{y,T} := \mathbb{E}\left[\sum_{t=0}^{T-1} \gamma^t (y_t^K)(y_t^K)^\top\right], \tag{54}$$

$$\Sigma_L^{z,T} := \mathbb{E}\left[\sum_{t=0}^{T-1} \gamma^t (z_t^L)(z_t^L)^\top\right]. \tag{55}$$

and

$$\begin{aligned} C_y^T(K) &:= \mathbb{E}\left[\sum_{t=0}^{T-1} \gamma^t (y_t^K)^\top (Q + K^\top R K) y_t^K\right] \\ &= Tr((Q + K^\top R K)\Sigma_K^{y,T}), \end{aligned} \tag{56}$$

$$\begin{aligned} C_z^T(L) &:= \mathbb{E}\left[\sum_{t=0}^{T-1} \gamma^t (z_t^L)^\top (Q + \bar{Q} + L^\top (R + \bar{R}) L) z_t^L\right] \\ &= Tr((Q + \bar{Q} + L^\top (R + \bar{R}) L)\Sigma_L^{z,T}). \end{aligned} \tag{57}$$

For every $\theta \in \Theta$, we define $C^T(\theta) = C_y^T(K) + C_z^T(L)$.

Assumption 2 plays a crucial role in the following result. Let us define a lower bound for the truncation horizon T:

$$h_{trunc,T}(\epsilon, \gamma_\theta) := \left(\frac{1}{\log(1/\gamma_\theta)}\left(\log\left(\frac{1}{\epsilon}\frac{C_{0,var}}{(1-\gamma_\theta)^2}\right) + 1\right)\right)^2 \tag{58}$$

where the parameter γ_θ is given by

$$\gamma_\theta = \max\{\gamma, \gamma\|A - BK\|^2, \gamma\|A + \bar{A} - (B + \bar{B})L\|^2\}. \tag{59}$$

and the constant $C_{0,var}$ can be chosen according to Lemma 9.

Lemma 32. *For any given $\epsilon > 0$, if the truncation horizon $T \geq 2$ satisfies $T \geq h_{trunc,T}(\theta, \epsilon, \gamma_\theta)$, then $\|\Sigma_K^y - \Sigma_K^{y,T}\| \leq \epsilon$ and $\|\Sigma_L^z - \Sigma_L^{z,T}\| \leq \epsilon$.*

Proof. Note that if $\theta \in \Theta$, then $\gamma_\theta < 1$. Let us recall that $\gamma_{1,K} = \gamma\|(A - BK)^2\|$. From Lemma 12 (or directly (22)), we know that

$$\begin{aligned} &\left\|\Sigma_K^y - \mathbb{E}\left[\sum_{t=0}^{T-1} \gamma^t y_t^K (y_t^K)^\top\right]\right\| \\ &= \left\|\frac{\gamma^T}{1-\gamma}\mathcal{T}_K^y\left(\Sigma^1\right) + \left(\mathcal{F}_K^y\right)^T\left(\mathcal{T}_K^y(\Sigma_{y_0})\right)\right. \\ &\qquad \left. + \sum_{s=1}^{T-1} \gamma^s \left(\mathcal{F}_K^y\right)^{T-s}\left(\mathcal{T}_K^y(\Sigma^1)\right)\right\| \\ &\leq \left(\frac{\gamma_\theta^T}{1-\gamma_\theta} + \gamma_\theta^T + (T-1)\gamma_\theta^T\right)\frac{C_{0,var}}{1-\gamma_\theta} \\ &\leq \frac{\gamma_\theta^T(T+1)}{1-\gamma_\theta}\frac{C_{0,var}}{1-\gamma_\theta}. \end{aligned} \tag{60}$$

where the second inequality used $\|\mathcal{T}_K^y\| \leq \frac{1}{1-\gamma_{1,K}} \leq \frac{1}{1-\gamma_\theta}$. We now show that $T \geq h_{trunc,T}(\epsilon, \gamma_\theta)$ implies

$$(T+1)\gamma_\theta^T \frac{C_{0,var}}{(1-\gamma_\theta)^2} \leq \epsilon. \tag{61}$$

Indeed, by taking the log on both sides of the above inequality and by noticing the fact that $\frac{\sqrt{x}}{\log(x+1)} > 1$ for all $x \geq 2$, we can then conclude the result.

□

Corollary 33. *If the conditions in Lemma 32 hold, then*

$$C(\theta) - C^T(\theta) \leq \epsilon d(\|Q\| + \|\bar{Q}\| + (\|R\| + \|\bar{R}\|)(\|K\|^2 + \|L\|^2)).$$

Remark 34. *In general, the truncation horizon T in Lemma 32 depends on the choice of parameter $\theta = (K, L)$ through the factor γ_θ. However, if we have $\|A - BK\| < 1$ and $\|A + \bar{A} - (B + \bar{B})L\| < 1$, then $\gamma_\theta = \gamma$ and the lower bounds for truncation horizon T only depends on model parameters.*

C.4 Convergence analysis

From here, we conclude the Theorem 5. The basic idea is to follow the lines of [10, Lemma 21] for each cost function $K \mapsto C_y(K)$ and $L \mapsto C_z(L)$. However, we emphasize here that the presence of sub-Gaussian noise processes makes the proof more complicated. In particular, we will use an appropriate version of Bernstein's inequality for sub-exponential random matrices (instead of just bounded matrices as in the case without Gaussian noises, see [10]), and we will need to prove that the assumptions required to apply this theorem are satisfied (see Lemma 37 below).

For any $\theta \in \Theta$, we choose M perturbed parameters $\theta_i = (K_i, L_i)$ with directions $v_i = (v_{1,i}, v_{2,i})$ where $v_{i,1} = K_i - K$ and $v_{i,2} = L_i - L$ for $i = 1, \ldots, M$ are all independent random matrices uniformly distributed on the sphere $\mathbb{S}_\tau$. We denote by

$$\tilde{\nabla}^T C(\theta, \underline{v}) = \frac{1}{M}\frac{d}{\tau^2}\sum_{i=1}^M \tilde{C}^T(\theta_i)\begin{bmatrix} K_i - K & 0 \\ 0 & L_i - L \end{bmatrix}$$

where $\underline{v} = (v_1, \ldots, v_M)$. The term $\tilde{C}^T(\theta_i) = \tilde{C}_y^T(K_i) + \tilde{C}_z^T(L_i)$ stands for the sampled truncated perturbed cost at $\theta_i = \theta + v_i$ with

$$\begin{aligned} \tilde{C}_y^T(K_i) &= \textstyle\sum_{t=0}^{T-1} \gamma^t (y_t^{K_i})^\top (Q + K_i^\top R K_i)(y_t^{K_i}), \\ \tilde{C}_z^T(L_i) &= \textstyle\sum_{t=0}^{T-1} \gamma^t (z_t^{L_i})^\top (Q + \bar{Q} + L_i^\top (R + \bar{R}) L_i)(z_t^{L_i}). \end{aligned} \tag{62}$$

We notice that by taking expectations with respect to the idiosyncratic noises and the common noises, we recover the expressions (56) and (57) from (62), namely $C_y^T(K_i) = \mathbb{E}[\tilde{C}_y^T(K_i)|v_i]$ and $C_z^T(L_i) = \mathbb{E}[\tilde{C}_z^T(L_i)|v_i]$.

The following result, which builds on Lemma 31, allows us to approximate the true gradient by the sampled truncated perturbed gradient with the help of a long truncation horizon and a large number of perturbation directions. We notice that for any matrix $X \in \mathbb{R}^{d\times d}$, $\|X\| \le \|X\|_F \le \sqrt{d}\|X\|$ and $\|X\|_F \le \|X\|_{tr} \le \sqrt{d}\|X\|_F$.

Let us define the following functions (of at most polynomial growth) related to the perturbation radius τ:

$$h_{subexp,y}(K,\tau) := 2C_2 e d^3 \tau \Big[\|Q\| + (\|K\|+\tau)^2\|R\|\Big] \cdot \frac{\sup\left\{\|y_0\|_{\psi_2}^2, \|\epsilon_1^1\|_{\psi_2}^2\right\}}{(1-\sqrt{\gamma})^2}, \tag{63}$$

$$h_{subexp,z}(L,\tau) := 2C_2 e d^3 \tau \Big[\|Q+\bar{Q}\| + (\|L\|+\tau)^2\|R+\bar{R}\|\Big] \frac{\sup\left\{\|z_0\|_{\psi_2}^2, \|\epsilon_1^0\|_{\psi_2}^2\right\}}{(1-\sqrt{\gamma})^2}, \tag{64}$$

where C_2 is a universal constant (see [31, Definition 2.7.5, Proposition 2.7.1]).

We also define a function for the number of perturbation directions M:

$$h_{trunc,pert,size}(\theta,d,T,\tau,\epsilon) := \frac{2}{\delta^2}\left((d+1)\log\left(\frac{d}{\epsilon}\right) + \log T + \log(16\epsilon)\right) \cdot \max\{h_\delta(h_{subexp,y}(K,\tau)), h_\delta(h_{subexp,z}(L,\tau))\}, \tag{65}$$

where $\delta = \frac{\epsilon\tau^2}{16Td}$ and $h_\delta(x) = x^2 + x\delta$.

let us first introduce some notations used in the following lemma. We define

$$\tilde{\nabla}_K^T := \frac{1}{M}\frac{d}{\tau^2}\sum_{i=1}^M \tilde{C}^T(\theta+v_i)v_{i,1},$$

$$\tilde{\nabla}_L^T = \frac{1}{M}\frac{d}{\tau^2}\sum_{i=1}^M \tilde{C}^T(\theta+v_i)v_{i,2},$$

where we recall that $v_{i,1} = K_i - K$ and $v_{i,2} = L_i - L$ are fixed matrices sampled from the uniform distribution on sphere $\mathbb{S}_\tau$. The expectations of $\tilde{\nabla}_K^T$ and $\tilde{\nabla}_L^T$ over the randomness coming from $(\epsilon_t^0, \epsilon_t^1)_{t=0,1,\dots,T-1}$ are denoted by $\hat{\nabla}_K^T$ and $\hat{\nabla}_K^T$. Namely, $\hat{\nabla}_K^T := \mathbb{E}[\tilde{\nabla}_K^T|\underline{v}] = \frac{1}{M}\frac{d}{\tau^2}\sum_{i=1}^M C^T(\theta_i)v_{1,i}$, and similar expression for $\hat{\nabla}_L^T$.

We use the abbreviations $\tilde{\nabla}^T$, $\hat{\nabla}^T$, $\hat{\nabla}$ and ∇ to denote respectively $\tilde{\nabla}^T C(\theta,\underline{v}) = \mathrm{diag}(\tilde{\nabla}_K^T, \tilde{\nabla}_L^T)$, $\hat{\nabla}^T C(\theta,\underline{v}) = \mathrm{diag}(\hat{\nabla}_K^T, \hat{\nabla}_L^T)$, $\hat{\nabla}C(\theta)$, and $\nabla C(\theta)$.

For any perturbation radius $\tau > 0$, let us define

$$\gamma_{\theta,\tau} := \max\{\gamma, \gamma\|A-BK\|^2 + \gamma\|B\|^2\tau^2, \gamma\|A+\bar{A}-(B+\bar{B})L\|^2 + \gamma\|B+\bar{B}\|^2\tau^2\}.$$

$$\epsilon_\tau := \frac{(\tau\epsilon)/(8d)}{d\left(\|Q\|+\|\bar{Q}\| + (\|R\|+\|\bar{R}\|)(4\tau^2 + 2\|K\|^2 + 2\|L\|^2)\right)}$$

Let us denote by $\|\eta\|_{\psi_1}$ the sub-exponential norm of a sub-exponential random variable $\eta \in \mathbb{R}$ (see [31, Definition 2.7.5]), namely

$$\|\eta\|_{\psi_1} = \inf\{s > 0 : \mathbb{E}[\exp(|\eta|/s)] \le 2\}.$$

The following result shows that if M and T are large enough and if τ and the perturbation on the control parameters are small enough, then the sampled truncated perturbed policy gradient $\tilde{\nabla}C^T(\theta,\underline{v})$ is close enough to the true gradient $\nabla C(\theta)$ in operator norm.

Lemma 35. *For every given $\epsilon > 0$, we assume that the perturbation radius τ satisfies $\tau \le 1/h_{pert,radius}(\theta,\epsilon/2)$, and also small enough so that $\gamma_{\theta,\tau} < 1$. In addition, suppose that the truncation horizon satisfies*

$$T \ge h_{trunc,T}(\epsilon_\tau, \gamma_{\theta,\tau}).$$

Moreover, suppose that the number of perturbation directions satisfies

$$M \ge \max\{h_{pert,size}(\theta,d,\epsilon/2), h_{trunc,pert,size}(\theta,d,T,\tau,\epsilon)\},$$

Then, with probability at least $1-(d/\epsilon)^{-d}$, we have

$$\|\tilde{\nabla}^T C(\theta,\underline{v}) - \nabla C(\theta)\| \le \epsilon.$$

Proof. We decompose the difference between $\tilde{\nabla}^T$ and ∇ as

$$\tilde{\nabla}^T - \nabla = (\tilde{\nabla}^T - \hat{\nabla}^T) + (\hat{\nabla}^T - \hat{\nabla}) + (\hat{\nabla} - \nabla) =: (i)+(ii)+(iii).$$

We highlight here our proof strategy. We choose in the first place a small enough perturbation radius τ and some large number M for the third term (iii). This step relies on our analysis of the exact PG convergence. Then we determine a long enough truncation horizon T, depending only on τ, to bound the second term (ii). Once T and τ are given, taking a possibly even larger number of perturbation directions M (if necessary), we establish a concentration inequality for the first term (i).

Step 1: For the third term (iii), by Lemma 31 and inequalities (51) and (52), if

$$\tau \le h_{pert,radius}(\theta,\epsilon/2), \text{ and } M \ge h_{pert,size}(\theta,d,\epsilon/2),$$

then, with probability at least $1-(\frac{d}{\epsilon/2})^{-d} \ge 1-\frac{1}{2}(d/\epsilon)^{-d}$, we have

$$\|\hat{\nabla} - \nabla\| \le \frac{\epsilon}{2}. \tag{66}$$

Step 2: For the second term (ii), by assumption on τ, we have for all $i = 1,\dots,M$, $\gamma\|A-BK_i\|^2 \le \gamma_{\theta,\tau} < 1$ and $\gamma\|A+\bar{A}-(B+\bar{B})L_i\|^2 < 1$. so that $\theta_i = (K_i, L_i)$ is admissible and $\gamma_{\theta_i} \le \gamma_{\theta,\tau} < 1$ for all $i = 1,\dots,M$. since the function $\gamma_\theta \mapsto h_{trunc,T}(\epsilon,\gamma_\theta)$ is increasing and we have $\|K_i\|^2 + \|L_i\|^2 \le 4\tau^2 + 2\|K\|^2 + 2\|L\|^2$, so that if

$T \geq h_{trunc,T}(\epsilon_\tau, \gamma_{\theta,tau})$, Lemma 32 and Corollary 33 imply $|C(\theta_i) - C^T(\theta_i)| \leq \frac{\tau\epsilon}{8d}$. Thus, we have

$$\begin{aligned}\|\hat{\nabla}^T - \hat{\nabla}\| =& \left\| \frac{1}{M}\frac{d}{\tau^2}\sum_{i=1}^{M}\left(C^T(\theta_i) - C(\theta_i)\right)\begin{bmatrix} v_{i,1} & 0 \\ 0 & v_{i,2}\end{bmatrix}\right\| \\ \leq& \frac{1}{M}\frac{d}{\tau^2}\sum_{i=1}^{M}|C^{(T)}(\theta_i) - C(\theta_i)|(\|v_{i,1}\| + \|v_{i,2}\|) \\ \leq& \frac{\epsilon}{4}. \end{aligned} \tag{67}$$

Step 3: For the first term (i), by definition we have

$$\mathbb{E}[\tilde{\nabla}^T|\underline{v}] = \hat{\nabla}^T.$$

In order to show that the sampled truncated perturbed gradient $\tilde{\nabla}^T$ is concentrated around its expectation (over the randomness of the noise processes), our strategy is to apply at each time step $t = 0, \ldots, T-1$ an appropriate version of Bernstein's inequality on a corresponding sequence of M independent random matrices, and then we use a union bound on T concentration inequalities.

We notice that

$$\|\tilde{\nabla}^T - \hat{\nabla}^T\| \leq \|\tilde{\nabla}^T_K - \hat{\nabla}^T_K\| + \|\tilde{\nabla}^T_L - \hat{\nabla}^T_L\| =: (iv) + (v), \tag{68}$$

and

$$\begin{aligned}(iv) =& \|\tilde{\nabla}^T_K - \hat{\nabla}^T_K\| \\ \leq& \left\| \frac{1}{M}\frac{d}{\tau^2}\sum_{i=1}^{M}\left(\tilde{C}^T_y(K_i) - \mathbb{E}[\tilde{C}_y(K_i)|v_{i1}]\right)v_{i,1}\right\| \\ &+ \left\| \frac{1}{M}\frac{d}{\tau^2}\sum_{i=1}^{M}\left(\tilde{C}^T_z(L_i) - \mathbb{E}[\tilde{C}^T_z(L_i)|v_{i1}]\right)v_{i,1}\right\| \\ =& (iv_y) + (iv_z), \end{aligned}$$

where $\tilde{C}^T_y(K_i)$ and $\tilde{C}^T_z(L_i)$ are defined by equation (62). Recall that in this part of the proof, $v_i = (v_{i,1}, v_{i,2})$ is fixed so the expectation symbol $\mathbb{E}$ is only for the randomness stemming from the sources of noise in the initial condition and the dynamics.

For the sake of clarity, we present in detail how to bound the term (iv_y). Similar arguments can be applied to bound the term (iv_z) as well as (v).

Let us denote the scalar η^i_t for $t = 0, \ldots, T-1$ and for $i = 1, \ldots, M$ by

$$\begin{aligned}\eta^i_t :=& (y^{K_i}_t)^\top (Q + K_i^\top R K_i) y^{K_i}_t \\ &- \mathbb{E}[(y^{K_i}_t)^\top (Q + K_i^\top R K_i) y^{K_i}_t | v_{i1}]. \end{aligned} \tag{69}$$

We notice that

$$\begin{aligned}(iv_y) =& \left\| \frac{1}{M}\frac{d}{\tau^2}\sum_{i=1}^{M}\left(\sum_{t=0}^{T-1}\gamma^t (y^{K_i}_t)^\top (Q + K_i^\top R K_i)(y^{K_i}_t) \right.\right. \\ &\left.\left. - \mathbb{E}\left[\sum_{t=0}^{T-1}\gamma^t (y^{K_i}_t)^\top (Q + K_i^\top R K_i)(y^{K_i}_t)\right]\right) v_{i,1}\right\| \\ =& \frac{d}{\tau^2}\left\| \sum_{t=0}^{T-1}\left(\frac{\gamma^t}{M}\sum_{i=1}^{M}\eta^i_t v_{i1}\right)\right\| \end{aligned}$$

Now, we fix $t \in \{0, \ldots, T-1\}$ and look at the terms $y^{K_i}_t$ for $i = 1, \ldots, M$. They are independent because they are constructed with independent noise processes $(\epsilon^{1,i}_s)_{s=0,\ldots,t}$. Moreover, from the dynamics of $(y^{K_i}_s)_{s=0,\ldots,t}$, we have

$$y^{K_i}_t = (A - BK_i)^t y^i_0 + \sum_{s=0}^{t-1}(A - BK_i)^{t-1-s}\epsilon^{1,i}_{s+1}. \tag{70}$$

where $y^i_0 = \epsilon^{1,i}_0 - \mathbb{E}[\epsilon^{1,i}_0]$. So, the term $y^{K_i}_t$ is the sum of $t+1$ independent sub-Gaussian random vectors, which implies that it is also a sub-Gaussian vector ([31, Definition 3.4.1]). Thus, the scalar η^i_t is a sub-exponential random variable [35].

We present in detail how to apply a suitable version of Matrix Bernstein's inequality for the random matrix $\sum_{i=1}^{M}\eta^i_t v_{i,1}$.

We recall that (see [31, Proposition 2.7.1])

$$\|\eta\|_{L^p} = \mathbb{E}[\|\eta\|^p]^{1/p} \leq C_2\|\eta\|_{\psi_1} p, \qquad \forall\, p \geq 1.$$

where C_2 is a universal constant. We define for each t the following quantity, which independent of the number M of perturbation directions:

$$\zeta_t := eC_2\tau \sup_{K' \in \mathbb{B}(K,\tau)} \|\eta^{K'}_t\|_{\psi_1}, \tag{71}$$

where the set $\mathbb{B}(K, \tau)$ stands for the ball centered at K with radius τ, and the term $\eta^{K'}_t$ is obtained by replacing $y^{K_i}_t$ in equation (69) by $y^{K'}_t$ defined as follows:

$$y^{K'}_t = (A - BK')^t y_0 + \sum_{s=0}^{t-1}(A - BK')^{t-1-s}\tilde{\epsilon}^1_{s+1}.$$

The noise process $(\tilde{\epsilon}^1_{s+1})^{t-1}_{s=0}$ is an independent copy of the idiosyncratic noise processes $(\epsilon^{1,i}_{s+1})^{t-1}_{s=0}$. We claim that

$$\mathbb{P}\left(\left\|\frac{1}{M}\sum_{i=1}^{M}\eta^i_t v_{i,1}\right\| \geq \delta\right) \leq 2d\exp\left(\frac{-M\delta^2/2}{\zeta^2_t + \zeta_t\delta}\right). \tag{72}$$

The proof of this claim is deferred to Lemma 37 below.

From here, by a union bound on the above concentration inequalities (72) for $t = 0, \dots, T-1$, with $\delta = \frac{\tau^2}{Td}\frac{\epsilon}{16}$, we obtain

$$\begin{aligned}\mathbb{P}\left((iv_y) \geq \frac{\epsilon}{16}\right) &\leq \mathbb{P}\left(\frac{Td}{\tau^2}\sup_{t=0,\dots,T-1}\left\|\frac{\gamma^t}{M}\sum_{i=1}^{M}\eta_t^i v_{i,1}\right\| \geq \frac{\epsilon}{16}\right)\\ &\leq 2d\sum_{t=0}^{T-1}\exp\left(\frac{-M\delta^2/2}{\gamma_t^2\zeta_t^2+\gamma_t\zeta_t\delta}\right)\\ &\leq 2dT\exp\left(\frac{-M\delta^2}{2[(\zeta_{iv_y})^2+\zeta_{iv_y}\delta]}\right),\end{aligned}$$

where

$$\zeta_{iv_y} := \sup_{t=0,\dots,T-1}\gamma^t\zeta_t. \tag{73}$$

Let us bound from above ζ_{iv_y}. We denote by

$$\xi(t,K') = (y_t^{K'})^\top(Q+(K')^\top RK')y_t^{K'}.$$

Since

$$\|\xi(t,K') - \mathbb{E}[\xi(t,K')]\|_{\psi_1} \leq 2\|\xi(t,K')\|_{\psi_1},$$

so that we have

$$\zeta_{iv_y} \leq \sup_{t=0,\dots,T-1}\sup_{K'\in\mathbb{B}(K,\tau)}\left(eC_2\tau.\gamma^t.2\left\|\xi(t,K')\right\|_{\psi_1}\right).$$

By applying [35, Proposition 2.5], we have

$$\|\xi(t,K')\|_{\psi_1} \leq \left\|Q+(K')^\top RK'\right\|_{tr}\left\|y_t^{K'}\right\|_{\psi_2}^2$$

Moreover, since $K' \in \mathbb{B}(K,\tau)$, we have

$$\begin{aligned}\left\|Q+(K')^\top RK'\right\|_{tr} &\leq d\left\|Q+(K')^\top RK'\right\|\\ &\leq d\Big[\|Q\|+\|R\|(\|K\|+\tau)^2\Big]\end{aligned}$$

and by Lemma 36 (see below),

$$\gamma^t\|y_t^{K'}\|_{\psi_2}^2 \leq \frac{d^2}{(1-\sqrt{\gamma})^2}\sup\left\{\|y_0\|_{\psi_2}^2, \|\epsilon_1^1\|_{\psi_2}^2\right\}.$$

So by definition of $\eta_t^{K'}$, we obtain

$$\zeta_{iv_y} \leq 2eC_2\tau\beta(K,\tau)\frac{d^2\sup\left\{\|y_0\|_{\psi_2}^2, \|\epsilon_1^1\|_{\psi_2}^2\right\}}{(1-\sqrt{\gamma})^2}.$$

where $\beta(K,\tau) =\leq d\Big[\|Q\|+\|R\|(\|K\|+\tau)^2\Big]$.

Thus, if our choice of M is large enough (see Lemma 37 below) such that

$$\begin{aligned}M \geq &\frac{1}{\delta^2}\left((d+1)\log\left(\frac{d}{\epsilon}\right)+\log T+\log(16\epsilon)\right)\\ &\cdot\left(2[(\zeta_{iv_y})^2+\zeta_{iv_y}\delta]\right),\end{aligned}$$

we obtain

$$\mathbb{P}\left((iv_y) \geq \frac{\epsilon}{16}\right) \leq \frac{1}{8}\left(\frac{d}{\epsilon}\right)^{-d}.$$

We can also derive a similar bound for the term (iv_z) as well as the term (v) defined in (68). Thus,

$$\mathbb{P}\left(\|\tilde{\nabla}^T - \hat{\nabla}^T\| \geq \frac{\epsilon}{4}\right) \leq \frac{1}{2}\left(\frac{d}{\epsilon}\right)^{-d}. \tag{74}$$

Conclusion: Combining inequalities (66), (67), and (74), we conclude that

$$\mathbb{P}\left(\|\tilde{\nabla}^T C(\theta,\underline{v}) - \nabla C(\theta)\| \leq \epsilon\right) \geq 1-\left(\frac{d}{\epsilon}\right)^{-d}.$$

□

The following lemma provides an upper bound on the sub-Gaussian norm of $y_t^{K'}$ for any perturbed parameter K' close enough to K.

Lemma 36. *Under Assumption 2, for any K' satisfying $\|K'-K\| \leq \tau$ with $\tau \leq 1/h_{cond}^y(\theta)$, we have*

$$\gamma^t\|y_t^{K'}\|_{\psi_2}^2 \leq \frac{d^2}{(1-\sqrt{\gamma})^2}\sup\left\{\|y_0\|_{\psi_2}^2, \|\epsilon_1^1\|_{\psi_2}^2\right\}.$$

Proof. To alleviate the notation, let us introduce $\mathbf{A}_0 = (A-BK')^t, \mathbf{A}_{s+1} = (A-BK')^{t-s-1} \in \mathbb{R}^{d\times d}$, and $h_0 = y_0$, $h_{s+1} = \tilde{\epsilon}_{s+1}^1 \in \mathbb{R}^d$, $s = 0,\dots,t-1$. Let $\mathbf{A}_{s,j} \in \mathbb{R}^d$ denote the $j-$th column of $\mathbf{A}_s$ and let $h_{s,j} \in \mathbb{R}$ denote the $j-$th coordinate of h_s, $j = 1,\dots,d$. Then,

$$\|y_t^{K'}\|_{\psi_2} = \left\|\sum_{s=0}^{t}\mathbf{A}_s h_s\right\|_{\psi_2} \leq \sum_{s=0}^{t}\sum_{j=1}^{d}\|\mathbf{A}_{s,j}h_{s,j}\|_{\psi_2}.$$

Moreover,

$$\begin{aligned}&\|\mathbf{A}_{s,j}h_{s,j}\|_{\psi_2}\\ =&\sup_{v\in\mathbb{S}^{d-1}}\inf\left\{k\in\mathbb{R} : \mathbb{E}\left[e^{(h_{s,j}\langle\mathbf{A}_{s,j},v\rangle)^2/k^2}\right] \leq 2\right\}\\ \leq&\inf\left\{k:\mathbb{R}: \sup_{v\in\mathbb{S}^{d-1}}\mathbb{E}\left[e^{(h_{s,j}\langle A_{s,j},v\rangle)^2/k^2}\right] \leq 2\right\}\\ \leq&\inf\left\{k\in\mathbb{R} : \mathbb{E}\left[\sup_{v\in\mathbb{S}^{d-1}}e^{(h_{s,j}\langle\mathbf{A}_{s,j},v\rangle)^2/k^2}\right] \leq 2\right\}\\ =&\inf\left\{k\in\mathbb{R} : \mathbb{E}\left[e^{(h_{s,j}\|\mathbf{A}_{s,j}\|_2)^2/k^2}\right] \leq 2\right\}\\ =&\|\mathbf{A}_{s,j}\|_2\|h_{s,j}\|_{\psi_2},\end{aligned}$$

where the first equality comes from the definition of sub-Gaussian random vector $\mathbf{A}_{s,j}h_{s,j}$ for $j = 1\dots,d$. So,

$$\begin{aligned}
&\gamma^t\|y_t^{K'}\|_{\psi_2}^2\\
&\leq \gamma^t\left(\sum_{s=0}^{t}\sum_{j=1}^{d}\|\mathbf{A}_{s,j}\|_2\|h_{s,j}\|_{\psi_2}\right)^2\\
&\leq \gamma^t\left(\sum_{s=0}^{t}\sqrt{d}\sqrt{\sum_{j=1}^{d}\|\mathbf{A}_s\|_2^2}\right)^2\sup_{s=0,\dots,t}\sup_{j=1,\dots,d}\|h_{s,j}\|_{\psi_2}^2\\
&\leq d\left(\sum_{s=0}^{t}\gamma^{s/2}(\gamma^{t-s}\|\mathbf{A}_s\|_F^2)^{1/2}\right)^2\sup_{s=0,\dots,t}\|h_s\|_{\psi_2}^2,
\end{aligned}$$

where the second inequality comes from the Cauchy-Schwartz inequality and the third inequality is justified by $\|h_{s,j}\|_{\psi_2} \leq \|h_s\|_{\psi_2}$. By noticing that K' is also admissible and

$$\gamma^{t-s}\|\mathbf{A}_s\|_F^2 \leq d\left(\gamma\|A - BK'\|^2\right)^{t-s} \leq d,$$

then the result follows. □

Lemma 37. *Reusing freely the notations in the proof of Lemma 35, we have*

$$\mathbb{P}\left(\left\|\frac{1}{M}\sum_{i=1}^{M}\eta_t^i v_{i,1}\right\| \geq \delta\right) \leq 2d\exp\left(\frac{-M\delta^2/2}{\zeta_t^2 + \zeta_t\delta}\right).$$

The idea of the proof is to estimate the $p-$order moment of $(\eta_t^i v_{i,1})$ for every $i = 1,\dots,M$. By Stirling's formula, we have

$$\mathbb{E}[(\eta_t^i v_{i,1})^p] = \tau^p\|\eta_t^i\|_{L^p}^p(v_{i,1}/\tau)^p \preceq \frac{p!}{\sqrt{2\pi p}}(\tau e C_2)^p\|\eta_t^i\|_{\psi_1}^p\mathbf{I}.$$

Then, we use a dilation technique to construct a symmetric matrix $\tilde{V}_{i,1} := \begin{bmatrix} 0 & v_{i,1} \\ v_{i,1}^\top & 0 \end{bmatrix}$ (see e.g. [29, Section 2.6]) and we apply Theorem 6.2 in [29] on $(\eta_t^i\tilde{V}_{i,1})_{i=1,\dots,M}$ to conclude the result.

D Sketch of proof of Theorem 6

Before providing the details of the proof of Theorem 6 in the next section, we give here an outline of the proof. We introduce first some notations:

- $\tilde{\nabla}_{M,\tau}^{N,T} = \tilde{\nabla}_{M,\tau}^{N,T}(\theta, \underline{v}, \underline{\epsilon}_0^1, \epsilon_0^0)$ is the output of Algorithm 2, with control parameter θ, perturbation directions $\underline{v}$, initial points for the N agents given by $\underline{\epsilon}_0^1$ and ϵ_0^0, and truncation horizon T;
- $\hat{\nabla}_{M,\tau}^{N,T} = \hat{\nabla}_{M,\tau}^{N,T}(\theta, \underline{v}) = \mathbb{E}_{\underline{\epsilon}_0^1,\epsilon_0^0}[\tilde{\nabla}_{M,\tau}^{N,T}(\theta, \underline{v}, \underline{\epsilon}_0^1, \epsilon_0^0)|\underline{v}]$ is the output of Algorithm 2 in expectation over the randomness of the state trajectories;
- $\hat{\nabla}_{M,\tau}^{N} = \hat{\nabla}_{M,\tau}^{N}(\theta, \underline{v})$ is the approximation of the gradient of the $N-$agents social cost at θ computed using derivative-free techniques with M perturbation directions given by $\underline{v}$, which is defined as

$$\hat{\nabla}_{M,\tau}^{N}(\theta, \underline{v}) = \frac{1}{M}\frac{d}{\tau^2}\sum_{i=1}^{M}C^N(\theta_i)\begin{bmatrix} v_{i1} & 0 \\ 0 & v_{i2}\end{bmatrix},$$

  where $C^N(\theta_i)$ is the $N-$agent social cost for the control $\theta_i = \theta + (v_{i,1}, v_{i,2}) = (K + v_{i,1}, L + v_{i,2})$;
- $\hat{\nabla}_{M,\tau} = \hat{\nabla}_{M,\tau}(\theta)$ is the approximation of the gradient of the mean field cost at θ computed using derivative-free techniques with M perturbation directions given by $\underline{v} = (v_1,\dots,v_M)$,

$$\hat{\nabla}_{M,\tau}(\theta) = \frac{1}{M}\frac{d}{\tau^2}\sum_{i=1}^{M}C(\theta_i)\begin{bmatrix} v_{i1} & 0 \\ 0 & v_{i2}\end{bmatrix},$$

  where $C(\theta_i)$ is the mean field cost for the control θ_i;
- $\nabla = \nabla(\theta) = \nabla C(\theta)$ is the gradient of the mean field cost at θ that we want to estimate.

Sketch of proof of Theorem 6. Let $\epsilon > 0$ be the target precision.

At step k, we denote the control parameter by $\theta = \theta^{(k)} \in \Theta$. The new control parameter after the Policy Gradient update procedure is denoted by $\tilde{\theta}' = \theta^{(k+1)}$. The main idea is to show that the following inequality holds until reaching the target precision ϵ:

$$C^N(\tilde{\theta}') - C^N(\theta^*) \leq \left(1 - \frac{1}{4}\eta h_{update}\right)(C^N(\theta) - C^N(\theta^*)), \tag{75}$$

for some appropriate learning rate $\eta > 0$ and a constant h_{update} defined in Lemma 27 . We introduce $\epsilon_b = \beta\epsilon, \epsilon_a = \epsilon - 2\epsilon_b = (1-2\beta)\epsilon$ for some $\beta < \frac{1}{12}$ chosen later.

We first look at the convergence in terms of the mean-field cost C.

Step 1: Take one step of the exact policy gradient update: $\mathbf{K}' = \mathbf{K} - \eta\nabla$, where $\nabla = \nabla C(\theta)$ to alleviate the notations. From Lemma 27, for a learning rate $0 < \eta \leq h_{lrate}(\theta)$ (see equation (43)),

$$C(\theta') - C(\theta^*) \leq (1 - \eta h_{update})(C(\theta) - C(\theta^*)). \tag{76}$$

Step 2: Take one step of the $N-$agent model-free policy gradient update: $\tilde{\mathbf{K}}' = \mathbf{K} - \eta\tilde{\nabla}_{M,\tau}^{N,T}$, where $\tilde{\nabla}_{M,\tau}^{N,T} = \tilde{\nabla}_{M,\tau}^{N,T}(\theta, \underline{v}, \underline{\epsilon}_0^1, \epsilon_0^0)$ is the output of Algorithm 2 (where it was denoted by $\tilde{\nabla}C^N(\theta)$. This estimate is stochastic due to the randomness of $(\underline{v} = (v_1,\dots,v_M), \underline{\epsilon}_0^1 =$

$(\epsilon_0^{1,(1)}, \dots, \epsilon_0^{1,(N)}), \epsilon_0^0)$. Let $\tilde{\theta}' = \theta(\tilde{\mathbf{K}}')$ be the new control parameter. Suppose that with high probability (at least of order $1 - (d/\epsilon_a)^{-d}$) we have that

$$|C(\tilde{\theta}') - C(\theta')| \leq \frac{1}{2}\eta\, h_{update}\, \varepsilon_a. \tag{77}$$

then, in the case when $C(\theta) - C(\theta^*) \geq \varepsilon_a$, with high probability we have,

$$C(\tilde{\theta}') - C(\theta^*) \leq \left(1 - \frac{1}{2}\eta h_{update}\right)\big(C(\theta) - C(\theta^*)\big). \tag{78}$$

Step 3: To conclude the convergence in terms of the mean-field cost C, we need to show equation (77) holds for large enough value of T, M, τ^{-1}. Lemma 24 in Section B indicates that we only need to ensure the new control parameter $\tilde{\theta}'$ is close enough to θ in the sense that

$$\max\left\{\|\tilde{K}' - K'\|, \|\tilde{L}' - L'\|\right\} \leq \frac{1}{2}\eta\, \frac{h_{update}}{h_{func,0}(\theta)}\, \varepsilon_a, \tag{79}$$

where $h_{func,0}(\theta)$ is of polynomial growth in $\|\theta\|$ (see (37)). Since

$$\max\left\{\|\tilde{K}' - K'\|, \|\tilde{L}' - L'\|\right\} \leq \|\tilde{\mathbf{K}}' - \mathbf{K}\| = \eta\|\tilde{\nabla}_{M,\tau}^{N,T} - \nabla\|,$$

and thus, equation (79) can be achieved by applying Lemma 38 (see below) on $\frac{1}{2}\frac{h_{update}}{h_{func,0}(\theta)}\,\varepsilon_a$.

Step 4: We now show the convergence in terms of the population cost C^N.

Recall that $\theta^{(0)}$ denotes the initial parameters. From Lemma 47, if $N \geq \frac{d}{\beta\epsilon} C(\theta^{(0)}) \left(\frac{1}{\lambda_y^1} + \frac{1}{\lambda_z^0}\right) \frac{C_{0,var}}{1-\gamma}$, we have for every $\hat{\theta} \in \{\theta^{(k)}, \theta^{(k+1)}, \theta^*\}$, $|C^N(\hat{\theta}) - C(\hat{\theta})| \leq \beta\epsilon = \epsilon_b$. Moreover, if $C^N(\theta) - C^N(\theta^*) \geq \epsilon$, then

$$C(\theta) - C(\theta^*) \geq \big(C^N(\theta) - C^N(\theta^*)\big) - 2\epsilon_b \geq (1 - 2\beta)\epsilon = \epsilon_a,$$

so that with high probability we have inequality (78). As a consequence, we obtain

$$\begin{aligned}
&C^N(\tilde{\theta}') - C^N(\theta^*) \\
&\leq \left|C^N(\tilde{\theta}') - C(\tilde{\theta}')\right| + (C(\tilde{\theta}') - C(\theta^*)) + \left|C(\theta^*) - C^N(\theta^*)\right| \\
&\leq \left(1 - \frac{1}{2}\eta h_{update}\right)(C(\theta) - C(\theta^*)) + 2\epsilon_b \\
&\leq \left(1 + 4\beta - \frac{1}{2}(1 + 2\beta)\eta h_{update}\right)\left(C^N(\theta) - C^N(\theta^*)\right),
\end{aligned}$$

If we choose

$$\beta \leq \frac{\eta h_{update}}{16 - 4\eta h_{update}} \leq \frac{1}{12}, \tag{80}$$

then $1 + 4\beta - \frac{1}{2}(1 + 2\beta)\eta h_{update} \leq 1 - \frac{1}{4}\eta h_{update}$. Thus, we conclude that for N large enough, if $C^N(\theta) - C^N(\theta^*) \geq \epsilon$, then with high probability, inequality (75) holds.

□

As emphasized by the above sketch of proof, the crux of the argument is the following bound for the difference between the gradient of the cost and its estimate computed by Algorithm 2.

Lemma 38. *For a fixed $\theta = (K, L)$, for any $\epsilon > 0$, if T, M, N and τ^{-1} are large enough, then with high probability, over the randomness of $(\underline{v} = (v_1, \dots, v_M), \underline{\epsilon}_0^1 = (\epsilon_0^{1,1}, \dots, \epsilon_0^{1,N}), \epsilon_0^0)$,*

$$\|\tilde{\nabla}_{M,\tau}^{N,T}(\theta, \underline{v}, \underline{\epsilon}_0^1, \epsilon_0^0) - \nabla(\theta)\| \leq \epsilon.$$

Proof. The proof relies on the expansion

$$\begin{aligned}
\tilde{\nabla}_{M,\tau}^{N,T} - \nabla =& (\tilde{\nabla}_{M,\tau}^{N,T} - \hat{\nabla}_{M,\tau}^{N,T}) + (\hat{\nabla}_{M,\tau}^{N,T} - \hat{\nabla}_{M,\tau}^{N}) \\
&+ (\hat{\nabla}_{M,\tau}^{N} - \hat{\nabla}_{M,\tau}) + (\hat{\nabla}_{M,\tau} - \nabla),
\end{aligned} \tag{81}$$

where we have omitted the arguments of the functions (i.e., $\theta, \underline{v}, \underline{\epsilon}_0^1, \epsilon_0^0$) for brevity. We then bound each term in the right hand side of (81) respectively by Lemma 44, Lemma 46, Lemma 48, and Lemma 31 (see Sections C and E). □

E Proof of Theorem 6

In this part of the proof, all the agents are supposed to be symmetric (i.e., exchangeable).

We recall that the structure of the proof of Theorem 6 is described in Section D. We now prove the technical lemmas required for this argument.

E.1 Preliminaries: initial point perturbations and positivity of minimal eigenvalues

E.1.1 Notations for population simulation

Let us recall that the dynamics for the N-agent problem are given by (3). When the control with parameter $\theta = (K, L) \in \Theta$ is taken into account, namely we look for $u_t = -K(X_t - \bar{x}_t^N) - L\bar{x}_t^N$ as suggested from the mean-field analysis, we denote by $\overline{\mu_t^{\theta,N}} = \frac{1}{N}\sum_{n=1}^N x_t^{\theta,(n)} = \bar{x}_t^N$ the empirical mean of the agents' states at time t. For every $n = 1, \dots, N$, we define the process $y_t^{\theta,(n)} = x_t^{\theta,(n)} - \overline{\mu_t^{\theta,N}}$ for every $t \geq 0$. Then the dynamics for $(y_t^{\theta,(n)})_{t\geq 0}$ and $(\overline{\mu_t^{\theta,N}})_{t\geq 0}$ can be described as: for every $t \geq 0$,

$$y_{t+1}^{\theta,(n)} = (A - BK)y_t^{\theta,(n)} + \epsilon_{t+1}^{1,(n)} - \overline{\epsilon_{t+1}^{1,N}}, \tag{82}$$

$$\overline{\mu_{t+1}^{\theta,N}} = (A + \overline{A} - (B + \bar{B})L)\overline{\mu_t^{\theta,N}} + \epsilon_{t+1}^0 + \overline{\epsilon_{t+1}^{1,N}}, \tag{83}$$

where for every $t \geq 0$, $\overline{\epsilon_t^{1,N}} = \frac{1}{N}\sum_{n'=1}^N \epsilon_t^{1,(n')}$, and the initial points are given by $\overline{\mu_0^N} = \epsilon_0^0 + \overline{\epsilon_0^{1,N}}$ and $y_0^{(n)} = \epsilon_0^{1,(n)} - \overline{\epsilon_0^{1,N}}$. We will sometimes drop the superscript θ in $\overline{\mu_0^{\theta,N}}$ and $y_0^{\theta,n}$ at time 0.

Let

$$\Sigma^N(\theta) = \begin{bmatrix} \Sigma_y^N(\theta) & 0 \\ 0 & \Sigma_\mu^N(\theta) \end{bmatrix}.$$

$\Sigma^N(\theta) = diag(\Sigma_y^N(\theta), \Sigma_\mu^N(\theta))$ where

$$\Sigma_y^N(\theta) = \frac{1}{N}\sum_{n=1}^N \mathbb{E}\sum_{t\geq 0}\gamma^t y_t^{\theta,n}(y_t^{\theta,n})^\top,$$
$$\Sigma_\mu^N(\theta) = \mathbb{E}\sum_{t\geq 0}\gamma^t \overline{\mu_t^{\theta,N}}(\overline{\mu_t^{\theta,N}})^\top$$

We also define the two auxiliary cost functions

$$C_y^N(\theta) = \frac{1}{N}\sum_{n=1}^N \mathbb{E}\left[\sum_{t\geq 0}\gamma^t\left((y_t^{\theta,n})^\top(Q+K^\top RK)y_t^{\theta,n}\right)\right]$$
$$C_\mu^N(\theta) = \mathbb{E}\left[\sum_{t\geq 0}\gamma^t \overline{\mu_t^{\theta,N}}^\top(Q+\overline{Q}+L^T(R+\overline{R})L)\overline{\mu_t^{\theta,N}}\right].$$

The total cost is then $C^N(\theta) = C_y^N(\theta) + C_\mu^N(\theta)$.

We define the following matrix, which represents the gradient of the $N-$agent social cost:

$$\nabla C^N(\theta) = \begin{bmatrix} \nabla_K C^N(\theta) & 0 \\ 0 & \nabla_L C^N(\theta) \end{bmatrix}.$$

Moreover, notice that the dynamics for $(y_t^{\theta,1})_{t\geq 0}$ and $(\overline{\mu_t^{\theta,N}})_{t\geq 0}$ in the population setting are similar to the dynamics of $(y_t^\theta)_{t\geq 0}$ and $(z_t^\theta)_{t\geq 0}$ respectively in the mean-field setting, except for the associated noise processes and the initial distributions. Based on these observations, we have the following result.

Lemma 39. *The N processes $(y_t^{\theta,(n)})_{t\geq 0}$ with $n = 1,\dots,N$ are identical in the sense of distribution, i.e., the laws of $y_t^{\theta,(n)}$ and $y_t^{\theta,(1)}$ are the same for all n and all t. As a consequence, for every $\theta = (K,L) \in \Theta$,*

$$C_y^N(\theta) = \mathbb{E}\left[\sum_{t\geq 0}\gamma^t\left(\left(y_t^{\theta,(1)}\right)^\top(Q+K^\top RK)y_t^{\theta,(1)}\right)\right],$$
$$\Sigma_y^N(\theta) = \mathbb{E}\left[\sum_{t\geq 0}\gamma^t y_t^{\theta,(1)}(y_t^{\theta,(1)})^\top\right].$$

Moreover we have

$$C_y^N(\theta) = \mathbb{E}\left[(y_0^{\theta,(1)})^\top P_K^y(y_0^{\theta,(1)})\right] + \frac{\gamma}{1-\gamma}\alpha_\theta^{y,N}, \quad (84)$$
$$C_z^N(\theta) = \mathbb{E}\left[(\overline{\mu_0^{\theta,N}})^\top P_L^z\overline{\mu_0^{\theta,N}}\right] + \frac{\gamma}{1-\gamma}\alpha_\theta^{\mu,N}, \quad (85)$$

where $\alpha_\theta^{y,N} = \mathbb{E}\left[\left(\epsilon_1^{1,(1)} - \overline{\epsilon_1^{1,N}}\right)^\top P_K^y\left((\epsilon_1^{1,(1)} - \overline{\epsilon_1^{1,N}}\right)\right]$ *and* $\alpha_\theta^{\mu,N} = \mathbb{E}\left[\left(\epsilon_1^0 + \overline{\epsilon_1^{1,N}}\right)^\top P_L^z\left(\epsilon_1^0 + \overline{\epsilon_1^{1,N}}\right)\right]$.

E.1.2 Initial point perturbations

We introduce the matrices related to the variances of initial points:

$$\Sigma_{y_0}^N = \frac{1}{N}\sum_{n=1}^N \mathbb{E}\left[y_0^{(n)}\left(y_0^{(n)}\right)^\top\right] = \mathbb{E}\left[y_0^{(1)}(y_0^{(1)})^\top\right]$$
$$\Sigma_{\mu_0}^N = \mathbb{E}\left[\overline{\mu_0^N}(\overline{\mu_0^N})^\top\right],$$

and we recall that the matrices Σ_{y_0} and Σ_{z_0}, defined by (9), are the variance matrices for y_0 and z_0 in the mean-field problem.

The following lemma builds connections for the variance at time $t = 0$ between the population setting and the mean-field setting.

Lemma 40 (Initial point perturbations)**.** *Recall we have assumed that $\epsilon_0^{1,(1)},\dots,\epsilon_0^{1,(N)}$ are i.i.d with distribution $\tilde{\mu}_0^1$ and they are also independent of ϵ_0^0. We have*

$$\Sigma_{y_0}^N = \left(1-\frac{1}{N}\right)\Sigma_{y_0}, \quad \Sigma_{\mu_0}^N = \Sigma_{z_0} + \frac{1}{N}\Sigma_{y_0}. \quad (86)$$

The proof are simple calculations using the fact that

$$\Delta\overline{\mu}_0 := \overline{\mu_0^N} - \overline{\mu_0} = \frac{1}{N}\sum_{n=1}^N(\epsilon_0^{1,(n)} - \mathbb{E}[\epsilon_0^{1,(n)}]).$$

We also denote the variances for one time-step noise processes by

$$\Sigma^{1,N} = \mathbb{E}\left[\left(\epsilon_1^{1,(1)} - \overline{\epsilon_1^{1,N}}\right)\left(\epsilon_1^{1,(1)} - \overline{\epsilon_1^{1,N}}\right)^\top\right],$$
$$\Sigma^{0,N} = \mathbb{E}\left[\left(\epsilon_1^0 + \overline{\epsilon_1^{1,N}}\right)\left(\epsilon_1^0 + \overline{\epsilon_1^{1,N}}\right)^\top\right].$$

Lemma 41. *We have*

$$\Sigma^{1,N} = \left(1-\frac{1}{N}\right)\Sigma^1, \quad \Sigma^{0,N} = \Sigma^0 + \frac{1}{N}\Sigma^1.$$

Lemma 42. *Under Assumption 2, Lemma 40 and Lemma 41 imply that for every $n = 1,\dots,N$,*

$$\|y_0^{(n)}\|_{\psi_2} \leq 2\|y_0\|_{\psi_2} \leq 4C_0,$$
$$\|\overline{\mu_0^N}\|_{\psi_2} \leq \|y_0\|_{\psi_2} + \|z_0\|_{\psi_2} \leq 4C_0,$$

and also for every $t \geq 1$,

$$\left\|\epsilon_t^{1,(n)} - \overline{\epsilon_t^{1,N}}\right\|_{\psi_2} \leq 2\left(1-\frac{1}{N}\right)\|\epsilon_1^1\|_{\psi_2} \leq 2C_0,$$
$$\left\|\epsilon_t^0 + \overline{\epsilon_t^{1,N}}\right\|_{\psi_2} \leq \|\epsilon_1^0\|_{\psi_2} + \|\epsilon_1^1\|_{\psi_2} \leq 2C_0.$$

E.1.3 Positivity of minimal eigenvalues

Let us denote by $\lambda^N_{y,min}$ and $\lambda^N_{\mu,min}$ the smallest eigenvalues of respectively $\mathbb{E}[y_0^1(y_0^1)^\top]$ and $\mathbb{E}[\overline{\mu_0^N}(\overline{\mu_0^N})^\top]$. We also introduce

$$\lambda_y^{1,N} = \lambda_{min}\left(\Sigma^N_{y_0} + \frac{\gamma}{1-\gamma}\Sigma^{1,N}\right),$$
$$\lambda_\mu^{0,N} = \lambda_{min}\left(\Sigma^N_{\mu_0} + \frac{\gamma}{1-\gamma}\Sigma^{0,N}\right),$$

and we recall that the corresponding eigenvalues λ_y^1, λ_z^0 for the mean field problem are defined by (15).

Lemma 43 (Positivity of minimal eigenvalues)**.** *If*

$$N > \frac{1}{\lambda_z^0}\left(\|\Sigma_{y_0}\| + \frac{\gamma}{1-\gamma}\|\Sigma^1\|\right), \tag{87}$$

then we have

$$\lambda_y^{1,N} = \left(1-\frac{1}{N}\right)\lambda_y^1 > 0,$$
$$\lambda_\mu^{0,N} \geq \lambda_z^0 - \frac{1}{N}\left\|\Sigma_{y_0} + \frac{\gamma}{1-\gamma}\Sigma^1\right\| > 0.$$

The proof is based on the Weyl's theorem.

E.2 Bounding $\|\tilde{\nabla}^{N,T}_{M,\tau} - \hat{\nabla}^{N,T}_{M,\tau}\|$ (Sampling error)

The following lemma estimates the difference between the sampled and expected versions of the truncated perturbed gradient, provided M is large enough. It relies on an application of Matrix Bernstein's inequality.

By Lemma 42 and Lemma 36, we introduce a function which has at most polynomial growth in $\|\theta\|$ and τ:

$$h^N_{subexp}(\theta,\tau) := 2C_2 e d^3 \tau \frac{16C_0^2}{(1-\sqrt{\gamma})^2}$$
$$\cdot\left[\|Q\| + \|\bar{Q}\| + (\|L\| + \|K\| + \tau)^2(\|R\| + \|\bar{R}\|)\right], \tag{88}$$

and also a function for the number of perturbation direction M:

$$h^N_{trunc,pert,size}(\theta,d,T,\tau,\epsilon) \tag{89}$$
$$:= \frac{2}{\delta^2}\left(\log\left(\frac{d}{\epsilon}\right)^{d+1} + \log T + \log(16\epsilon)\right) h_\delta(h^N_{subexp}(\theta,\tau))$$

where $\delta = \frac{\epsilon\tau^2}{16Td}$ and $h_\delta(x) = x^2 + x\delta$.

Let us define the estimated gradient term by

$$\tilde{\nabla}^{N,T}_{M,\tau} = \frac{1}{M}\frac{d}{\tau^2}\sum_{i=1}^M \tilde{C}^{N,T}(\theta_i)\begin{bmatrix} v_{i1} & 0 \\ 0 & v_{i2}\end{bmatrix},$$

where $\theta_i = \theta + v_i$ with $v_i = (v_{i1}, v_{i2})$, and the sampled truncated cost for N agent is denoted by

$$\tilde{C}^{N,T}(\theta_i) = \tilde{C}_y^{N,T}(K_i) + \tilde{C}_\mu^{N,T}(L_i)$$

such that

$$\tilde{C}_y^{N,T}(K_i) = \frac{1}{N}\sum_{n=1}^N\sum_{t=0}^{T-1}\gamma^t (y_t^{\theta_i,(n)})^\top (Q + K_i^\top R K_i) y_t^{\theta_i,(n)},$$
$$\tilde{C}_\mu^{N,T}(L_i) = \sum_{t=0}^{T-1}\gamma^t (\overline{\mu_t^{\theta_i,N}})^\top (Q + \bar{Q} + L_i^\top(R+\bar{R})L_i)\overline{\mu_t^{\theta_i,N}}.$$

Note that we use $\tilde{C}_y^{N,T}(K_i)$ instead of $\tilde{C}_y^{N,T}(\theta_i)$ because the dynamics of associating processes $(y_t^{\theta_i,(n)})_{t\geq 0}$ for every $n = 1,\dots,N$ depends only on K_i.

Lemma 44 (Monte Carlo error on gradient estimates)**.** *Let τ and T be given and assume that the number of perturbation directions M satisfies $M \geq h^N_{trunc,pert,size}(\theta,d,T,\tau,\epsilon)$. Then, with probability at least $1-(d/\epsilon)^{-d}$, we have*

$$\|\tilde{\nabla}^{N,T}_{M,\tau} - \hat{\nabla}^{N,T}_{M,\tau}\| \leq \epsilon.$$

Proof. The proof is very much alike the step 3 in Lemma 35. The difficulty is that, for a fixed perturbation direction v_i, the N random processes $y_t^{\theta_i,(n)}$ for $n = 1,\dots,N$ are correlated since their noise processes are coupled. So, the trick here is not only fixing the time $t \in \{0,\dots,T-1\}$ but also fixing the agent index $n \in \{1,\dots,N\}$ and look that the concentration phenomenon of M independent random variables $(y_t^{\theta_i,(n)})$ for $i = 1,\dots,M$. We can have a similar inequality for:

$$\left\|\frac{1}{M}\frac{d}{\tau^2}\sum_{i=1}^M\left(\tilde{C}_y^{N,T}(K_i) - \mathbb{E}[\tilde{C}_y^{N,T}(K_i)|v_{i1}]\right)v_{i,1}\right\|$$
$$\leq \frac{Td}{\tau^2}\sup_{t=0,\dots,T-1}\left\|\frac{\gamma^t}{M}\sum_{i=1}^M\left(\frac{1}{N}\sum_{n=1}^N \eta_t^{i,(n)}\right)v_{i,1}\right\|$$
$$\leq \frac{Td}{\tau^2}\sup_{n=1,\dots,N}\sup_{t=0,\dots,T-1}\left\|\frac{\gamma^t}{M}\sum_{i=1}^M \eta_t^{i,(n)} v_{i,1}\right\|$$

where $\eta_t^{i,(n)}$ for $n = 1,\dots,N$ are defined similarly as in equation 69. Since we can also establish an universal upper bound for the sub-Gaussian norm of $y_t^{\theta_i,(n)}$:

$$\gamma^t\|y_t^{\theta_i,(n)}\|^2_{\psi_2}$$
$$\leq \frac{d^2}{(1-\sqrt{\gamma})^2}\sup\left\{\|y_0^{(n)}\|^2_{\psi_2}, \left\|\epsilon_1^{1,(n),i} - \overline{\epsilon_1^{1,N,i}}\right\|^2_{\psi_2}\right\}$$
$$\leq \frac{16d^2C_0^2}{(1-\sqrt{\gamma})^2}.$$

we can conclude the result following the idea presented in 35. □

E.3 Bounding $\|\hat{\nabla}^{N,T}_{M,\tau} - \hat{\nabla}^{N}_{M,\tau}\|$ (truncation error)

Let us start with a result which estimates the difference between the truncated gradient and the true gradient in $N-$agent control problem. It is analogous to Lemma 32 in the mean field problem. It is a generalization of [10, Lemma 23] to the case with noise processes. We introduce the variance matrices for the truncated processes

$$\Sigma_y^{N,T}(\theta) = \mathbb{E}\left[\sum_{t=0}^{T-1} \gamma^t y_t^{\theta,(1)} (y_t^{\theta,(1)})^\top\right],$$
$$\Sigma_\mu^{N,T}(\theta) = \mathbb{E}\left[\sum_{t=0}^{T-1} \gamma^t \overline{\mu_t^{\theta,N}} (\overline{\mu_t^{\theta,N}})^\top\right].$$

Similarly to equation (58), we also introduce in the N agent problem a lower bound for the truncated horizon T:

$$h^N_{trunc,T}(\epsilon,\gamma_\theta) := \left(\frac{1}{\log(1/\gamma_\theta)}\left(\log\left(\frac{1}{\epsilon}\frac{(1+\frac{1}{N})C_{0,var}}{(1-\gamma_\theta)^2}\right)+1\right)\right)^2,$$

where γ_θ is defined by (59).

Lemma 45 (Truncation error on cost estimates)**.** *For every $\theta = (K,L) \in \Theta$ and for every $\epsilon > 0$, if the truncation horizon $T \geq 2$ satisfies $T \geq h^N_{trunc,T}(\epsilon,\gamma_\theta)$, then*

$$\|\Sigma_y^N(\theta) - \Sigma_y^{N,T}(\theta)\| \leq \epsilon, \quad \|\Sigma_\mu^N(\theta) - \Sigma_\mu^{N,T}(\theta)\| \leq \epsilon. \quad (90)$$

The proof is very similar to the arguments used in Lemma 32 and Lemma 12. The only difference comes from the inequality

$$\left\|\Sigma_y^N(\theta) - \Sigma_y^{N,T}(\theta)\right\| \leq \left(\frac{(\gamma_\theta)^T}{1-\gamma_\theta} + (\gamma_\theta)^T + (T-1)(\gamma_\theta)^T\right) \times \frac{1}{1-\gamma_\theta}\max\left\{\left\|(1-\frac{1}{N})\Sigma^1\right\|, \left\|(1-\frac{1}{N})\Sigma_{y_0}\right\|\right\}.$$

The following result quantifies the error due to the time horizon truncation in the estimation of the gradient. Its proof is similar to the arguments used in Step 2 of Lemma 35.

Lemma 46 (Truncation error on gradient estimates)**.** *For every $\epsilon > 0$, if $T \geq h^N_{trunc,T}(\epsilon_\tau,\gamma_{\theta,\tau})$. where $\gamma_{\theta,\tau}$ and ϵ_τ are defined as in Lemma 35. Then, we have $\|\hat{\nabla}^{N,T}_{M,\tau} - \hat{\nabla}^{N}_{M,\tau}\| \leq \epsilon$.*

E.4 Bounding $\|\hat{\nabla}^{N}_{M,\tau} - \hat{\nabla}_{M,\tau}\|$ (finite population error)

We start with a bound on the difference between the population cost and the mean-field cost.

Lemma 47. *For any $\theta \in \Theta$,*

$$|C^N(\theta) - C(\theta)| \leq \frac{d}{N}(\|P_K^y\| + \|P_L^z\|)\left(\|\Sigma_{y_0}\| + \frac{\gamma}{1-\gamma}\|\Sigma^1\|\right).$$

Proof. The difference between the cost functions for the $N-$agent control problem and the mean-field type control problem can be split into two terms:

$$C^N(\theta) - C(\theta) = \left(C_y^N(\theta) - C_y(K)\right) + \left(C_\mu^N(\theta) - C_z(L)\right).$$

For the first term, we have

$$\begin{aligned} &C_y^N(\theta) - C_y(K) \\ =&\mathbb{E}\left[(y_0^{(1)})^\top P_K^y y_0^{(1)}\right] - \mathbb{E}\left[(y_0)^\top P_K^y y_0\right] + \left(\alpha_\theta^{y,N} - \alpha_K^y\right) \\ =&Tr\left(\left[\Sigma_{y_0}^N - \Sigma_{y_0}\right]P_K^y\right) + \frac{\gamma}{1-\gamma}Tr\left(P_K^y(\Sigma^{1,N} - \Sigma^1)\right) \\ =&-\frac{1}{N}Tr\left(P_K^y\left(\Sigma_{y_0} + \frac{\gamma}{1-\gamma}\Sigma^1\right)\right), \end{aligned}$$

also, $C_\mu^N(\theta) - C_z(L) = \frac{1}{N}Tr\left(P_L^z\left(\Sigma_{y_0} + \frac{\gamma}{1-\gamma}\Sigma^1\right)\right)$, hence, we can draw the conclusion with the trace inequality. □

We then establish the following lemma, which bounds the difference between $N-$agent perturbed gradient $\hat{\nabla}^N_{M,\tau}$ and perturbed gradient for mean-field control problem $\hat{\nabla}$.

Lemma 48 (Finite population error on gradient estimates)**.** *For every given $\epsilon > 0$, if the perturbation radius τ is small enough such that whenever $\|K' - K\| \leq \tau$ and $\|L' - L\| \leq \tau$, $|C(\theta') - C(\theta)| \leq C(\theta)$, and if*

$$N \geq \frac{1}{\epsilon}\frac{4d^2C(\theta)C_{0,var}}{\tau}\left(\frac{1}{\lambda_y^1} + \frac{1}{\lambda_z^0}\right), \quad (91)$$

then for all $\underline{v} = (v_1,\ldots,v_M)$, we have $\|\hat{\nabla}^N_{M,\tau} - \hat{\nabla}_{M,\tau}\| \leq \epsilon$.

Notice that by Corollary 25, the condition on τ in the above statement is satisfied for example if $\tau \leq 1/h_{C,cond}(\theta)$.

Proof. Since for every $v_i = (v_{i1}, v_{i2})$, we have

$$\left\|\begin{bmatrix} v_{i1} & 0 \\ 0 & v_{i2}\end{bmatrix}\right\| \leq \|v_{i1}\| + \|v_{i2}\| \leq 2\tau,$$

and by our condition on τ, we have

$$C(\theta_i) = C(\theta + v_i) \leq 2C(\theta).$$

By Lemma 47, we deduce that

$$\begin{aligned}
&\|\hat{\nabla}^N_{M,\tau} - \hat{\nabla}_{M,\tau}\| \\
=&\frac{1}{M}\sum_{i=1}^{M}\frac{d}{\tau^2}(C^N(\theta_i) - C(\theta_i))\begin{bmatrix} v_{i1} & 0 \\ 0 & v_{i2}\end{bmatrix} \\
\leq&\frac{1}{N}\frac{d}{\tau^2}\frac{1}{M}\sum_{i=1}^{M}\Bigg(d(\|P^y_{K_i}\| + \|P^z_{L_i}\|) \\
&\qquad\cdot\left(\|\Sigma_{y_0}\| + \frac{\gamma}{1-\gamma}\|\Sigma^1\|\right)\left\|\begin{bmatrix} v_{i1} & 0 \\ 0 & v_{i2}\end{bmatrix}\right\|\Bigg) \\
\leq&\frac{1}{N}\frac{4d^2C(\theta)C_{0,var}}{\tau M}\left(\frac{1}{\lambda^1_y} + \frac{1}{\lambda^0_z}\right).
\end{aligned}$$

Hence, by (91), we obtain $\|\hat{\nabla}^N_{M,\tau} - \hat{\nabla}_{M,\tau}\| \leq \epsilon$. □

F Details on the numerical results

F.1 Details about the numerical results of Section 4

The numerical results presented in Section 4 were obtained with the parameters given in Table 1. Recall that the $N-$agent dynamics and the cost are given respectively by (3) and (5), with parameters denoted, in the one dimensional case, by $a, \bar{a}, b, \bar{b}$ for the dynamics, $q, \bar{q}, r, \bar{r}$ for the cost. The discount factor is γ, and $\tilde{h}$ is the degree of heterogeneity.

The randomness in the state process is given in the second part of Table 1, where $\mathcal{U}([a,b])$ stands for the uniform distribution on interval $[a,b]$ and $\mathcal{N}(\mu,\sigma^2)$ is the one dimensional Gaussian distribution with mean μ and standard deviation σ.

The parameters of the model free method are T for the truncation length, M for the number of perturbation directions, τ for the perturbation radius. We used Adam optimizer (which provided slightly smoother results than a constant learning rate η as in our theoretical analysis) initialized with initial learning rate η and exponential decay rates given by β_1 and β_2. The number of perturbation directions is M and their radius is τ. These value have been chosen based on the theoretical bounds we found and further tuning after a few tests with the exact PG method.

The computations have been run on a 2.4 GHz Skylake processor. For the parameters described here, the model free PG with MKV simulator took roughly 10 hour for 5000 iterations. For the same number of iterations, the model free PG with $N-$agent population simulator took roughly 48 hours for $N = 10$.

Table 1: Simulation parameters for numerical example as described in the text.

Model parameters									
a	$\bar{a}$	b	$\bar{b}$	q	$\bar{q}$	r	$\bar{r}$	γ	$\tilde{h}$
0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.9	0.1

Initial distribution and noise processes			
ϵ^0_0	ϵ^1_0	ϵ^0_t	ϵ^1_t
$\mathcal{U}([-1,1])$	$\mathcal{U}([-1,1])$	$\mathcal{N}(0,0.01)$	$\mathcal{N}(0,0.01)$

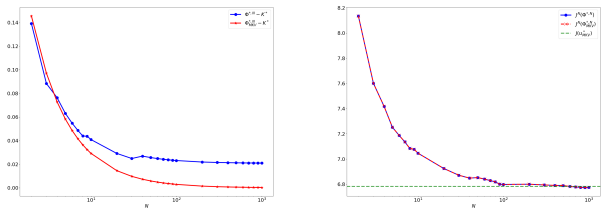
Learning parameters							
T	M	τ	η	β_1	β_2	K_0	L_0
50	10000	0.1	0.01	0.9	0.999	0.0	0.0

F.2 Approximate optimality of the mean field optimal control for the $N-$agent problem

Let us consider the general case for which the variation $\tilde{q}^n$ is not assumed to be 0. We provide numerical evidence showing that even though the optimal control $U^{*,N}_t = \Phi^{*,N}X_t$ of the N agents problem differs from the optimal control $\Phi^{*,N}_{MKV}X_t$ that can be obtained using the optimal MKV feedback, the latter provides an approximately optimal control for the N agents social cost, where the quality of the approximation depends on N and on the heterogeneity degree $\tilde{h}$ (recall that the variations $\tilde{q}^n$ are of size at most $\tilde{h}$).

When the coefficients of the model are known, the matrices $\Phi^{*,N}$ (for the $N-$agent problem) and (K^*, L^*) (for the MFC problem) can be computed by solving Riccati equations. In the $d = 1$ case, we can easily compare these controls by looking, for instance, at the diagonal coefficients of $\Phi^{*,N}$ and the value of K^*. Figure 4a shows (in blue) the graph of the function $N \mapsto \max_{i=1,\dots,N}|(\Phi^{*,N})^{i,i} - K^*|$. While this quantity decreases with N, it does not converge to 0. This transcribes the fact that the optimal control with heterogenous agents does not converge to the optimal MKV control. This figure is for one realization of the variations $\tilde{q}^n$, drawn uniformly at random in $(-\tilde{h}, \tilde{h}) = (-1, 1)$ with $q = 5$. For the sake of comparison we also show (see the red curve in Figure 4a) that the diagonal coefficients of $\Phi^{*,N}_{MKV}$ converge to K^*.

On the other hand, instead of comparing the controls, we can compare the associated costs. Indeed, once the matrices $\Phi^{*,N}$ and $\Phi^{*,N}_{MKV}$ are computed, if we use them as feedback functions, the corresponding $N-$player social cost can be readily computed using an analytical formula based on a Riccati equation, as is usual for LQ problems (see e.g. (12) below for the mean field case). Figure 4b shows the difference between the two costs as N increases. One can see that



(a) (b)

Figure 4: Comparison of the N agent optimal control and the MKV optimal control. (a): Maximum difference between diagonal terms; (b) Social cost for each control.

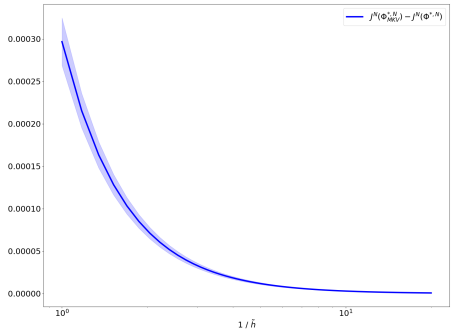


Figure 5: Influence of the heterogeneity degree on the social cost.

whether the agents use the real optimal control (blue line) or the one coming from the optimal MKV control (red line), the values of the social cost are almost the same as they seem to converge to the value of the theoretical optimal MKV social cost (green dashed line). However, a small discrepancy remains due to the heterogeneity of the population.

Figure 5 illustrates the influence of the heterogeneity of the agents. For fixed $N = 100$, as the radius $\tilde{h}$ of the variation term vanishes (i.e., $1/\tilde{h}$ increases) and the system becomes more homogeneous, the difference between the $N-$agent social cost computed with $\Phi^{*,N}$ and $\Phi^{*,N}_{MKV}$ decreases. In this figure, the curve is the averaged over 5 random realizations and the shaded region corresponds to the mean $\pm$ the standard deviation.

Remark 49. *Notice that the average of the states of all the players enters the computation of the control of player i, i.e. the i-th entry of $U_t^{*,N}$. We expect that when the number of players grows without bound, i.e. when $N \to \infty$, this average should converge to the theoretical mean $\bar{x}_t$ of the optimal state in the control of the McKean-Vlasov equation* (1)*. So if we were able to compute $\bar{x}_t$ off-line and replace in* (92) *the vector $\overline{X}_t$ of sample averages by the vector of N copies of the McKean-Vlasov true mean $\bar{x}_t$, we would get a control profile*

$$\tilde{U}_t^{*,N} = \alpha p^* X_t + [\alpha(\bar{p}^* - p^*) + \beta \bar{p}^*]\bar{x}_t \mathbb{1}, \qquad (92)$$

which is decentralized *in the sense that the control of the i-th player only depends upon the state of player i, and which is an approximation of the optimal control since their difference converges to* 0*.*