Proceedings of Machine Learning Research vol vvv:1–24, 2024

Robust Multi-Agent Reinforcement Learning: A Mean-Field Perspective

Muhammad Aneeq uz Zaman MAZAMAN2@ILLINOIS.EDU
Coordinated Science Laboratory, University of Illinois at Urbana-Champaign

Mathieu Laurière ML5197@NYU.EDU
Mathematics and Data Science, NYU Shanghai

Alec Koppel ALEC.KOPPEL@JPMCHASE.COM
Artificial Intelligence Research, JP Morgan Chase & Co.

Tamer Başar BASAR1@ILLINOIS.EDU
Coordinated Science Laboratory, University of Illinois at Urbana-Champaign

Abstract

In this paper, we study the problem of Robust Multi-agent Reinforcement Learning where a large number of cooperative agents with distributed information aim to learn policies in the presence of *stochastic* (distribution known) and *non-stochastic* (distribution unknown) uncertainties. We focus on policy optimization that accounts for both types of uncertainties, which falls into a worst-case (minimax) framework. Since this problem is intractable in general, we focus on the Linear Quadratic setting to enable derive benchmark algorithms. Since no standard theory exists for this problem (due to the distributed information structure), we utilize the mean-field paradigm (Mean-Field Type Game (MFTG)) to establish guarantees on the solution quality in the sense of achieved Nash equilibrium of the MFTG, which in turn allows us to compare its performance against the corresponding original Robust Multi-agent control problem. Then we introduce a Receding-horizon Gradient Descent Ascent (RGDA) Reinforcement Learning algorithm to find the Nash equilibrium of the MFTG and estaboish its non-asymptotic rate of convergence. Finally we provide numerical analysis to demonstrate the efficacy of our approach relative to the baseline algorithm.

1. Introduction

Reinforcement Learning (RL) has had many successes, such as autonomous driving (Sallab et al., 2017), robotics (Kober et al., 2013), and RLHF (Ziegler et al., 2019), to name a few. These successes have been focused on single-agent scenarios, but many scenarios involving, e.g., financial markets, communication networks, distributed robotics, among others, involve multiple agents. Prevailing algorithms for Multi-Agent Reinforcement Learning (MARL) (Zhang et al., 2021b; Li et al., 2021), however, do not model the distinct effects of modeled and un-modeled uncertainties on the transition dynamics, which can result in practical instability in safety-critical applications (Riley et al., 2021).

In this paper we consider a large population multi-agent setting, with stochastic and non-stochastic (un-modeled, possibly adversarial) uncertainties. These types of formulations have been studied under the guise of robust control in the single-agent case (Başar and Bernhard, 2008). The uncertainties (modeled and un-modeled) affect the performance of the system and might even lead to instability. Robust control seeks the *robust* controller which guarantees a certain level of performance for the system in the presence of these uncertainties. We employ here the Linear-Quadratic (LQ) setting in order to rigorously characterize and synthesize the solution to the robust multi-agent problem in a data-driven manner. The LQ setting entails a class of models in which the dynamics are linear and the costs are quadratic in the state and the action of the agent. This setting has been used extensively in the literature due to its tractability: the optimal decisions can be computed analytically or almost analytically, up to solving Riccati equations, when one has access to all system matrices, as in certain instances of permanent income theory (Sargent and Ljungqvist, 2000), portfolio management (Cardaliaguet and Lehalle, 2018), and wireless power control (Huang et al., 2003)), to name a few. In the absence of knowledge of system parameters, model-free RL methods have also been developed (Fazel et al., 2018; Malik et al., 2019) for single agent LQ settings. When one goes from single to multiple agents, the issue of communicating local state and control information among agents exhibit scalability problems, and in particular, practical algorithms require sharing state information that can scale exponential in the number of agents. Instead, here we consider a distributed information structure where each agent has access only to its own state and the average of states of the other agents. This distributed information structure causes the characterization of the solution to be very difficult, in that previous gradient dominance results from (Fazel et al., 2018) no longer hold. To overcome this difficulty, we utilize the mean-field paradigm, first introduced in the purely non-cooperative agent setting in (Lasry and Lions, 2006; Huang et al., 2006), which replaces individual agents by a distribution over agent types, which enables characterization and computation of the solution. In doing so, this work is the first to develop scalable algorithms for MARL that can handle model mis-specification or adversarial inputs in the sense of robust control in the very large or possibly infinite number of agents defined by the mean-field.

We start Section 2 by formulating a Robust Multi-agent control problem with stochastic and non-stochastic (un-modeled) noise. The agents have distributed information, such that they have access to their own states and the average behavior of all the agents. Solving this problem entails finding a *noise attenuation level* (noise-to-output gain) for the multi-agent system and the corresponding *robust controller*. As in the single-agent setting (Başar and Bernhard, 2008), the robust multi-agent control problem is reformulated into an equivalent zero-sum min-max game between the maximizing non-stochastic noise (which may be interpreted as an adversary) and the minimizing controller. Solving this problem is not possible in the finite agent case due to the limited information available to each agent. Thus, in Section 3 we consider the mean-field (infinite population) version of the problem, called the Robust Mean-Field Control (RMFC) problem. As in the finite-population setting, RMFC has an equivalent zero-sum min-max formulation, referred to as the 2-player zero-sum Mean-Field Type Game (ZS-MFTG) (Carmona et al., 2020, 2021), where the controller is the minimizing player and the non-stochastic disturbance is the maximizing one.

In Section 4 we propose a bi-level Reinforcement Learning algorithm to compute the Nash equilibrium for the ZS-MFTG (which equivalently yields the robust controller for the robust multi-agent problem) in the form of Receding-horizon Gradient Descent Ascent (RGDA) (Algorithm 1). The upper-level of RGDA, uses a receding-horizon approach i.e. it finds the controller parameters starting from the last timestep $T-1$ and moving backwards-in-time (à la dynamic programming). The receding-horizon policy gradient approach was introduced to Kalman filtering (Zhang et al., 2023) and LQR in (Zhang and Başar, 2023). This paper build on this approach to multi-agent problems, which helps in simplifying the complex nature of the cost landscape (known to be non-coercive (Zhang et al., 2021a)) and renders it convex-concave. The lower-level employs gradient descent-ascent to find the saddle point (Nash equilibrium) for each timestep t. The convex-concave nature of the cost (due to the receding-horizon approach) proves to be a key component in proving linear convergence of the gradient descent-ascent to the saddle point (Theorem 4). Further analysis shows that the total accumulated error in the RGDA is small given that the lower-level of RGDA has good convergence (Theorem 5). The gradient descent-ascent step requires computation of the stochastic gradient. We use a zero-order method (Fazel et al., 2018; Malik et al., 2019) which only requires access to the cost to compute stochastic gradients, and hence is *truly* model-free.

Literature Review: Robust control gained importance in the 1970s when control theorists realized the shortcomings of optimal control theory in dealing with model uncertainties (Athans et al., 1977; Harvey and Stein, 1978). The work of (Başar, 1989) was the first one to formulate the robust control problem as a zero-sum dynamic game between the controller and the uncertainty. Robust Reinforcement Learning first introduced by (Morimoto and Doya, 2005) has recently had an increase in interest in for the single agent setting, where its ability to process trajectory data without explicit knowledge of system parameters can be used to learn robust controllers to address worst-case uncertainty (Zhang et al., 2020a; Kos and Song, 2017; Zhang et al., 2021c). Some recent works consider RL in scenarios with reward uncertainties (Zhang et al., 2020b), state uncertainty (He et al., 2023) or uncertainty in other agents' policies (Sun et al., 2022). There have been some works on the intersection of RL for robust and multi-agent control (Li et al., 2019; He et al., 2023), yet there has not been any significant effort to provide (1) sufficient conditions for solvability of the multi-agent robust control problem i.e. determining the noise attenuation level of a system and (2) provable Robust multi-agent RL (RMARL) algorithms in the large population setting, as proposed in this paper.

This is made possible due to the mean-field paradigm, which considers the limiting case as the number of agents approaches infinity. This paradigm was first introduced in the context of non-cooperative game theory as Mean-Field Games (MFGs) concurrently by (Lasry and Lions, 2006; Huang et al., 2006). Since then there have been several works dealing with Reinforcement Learning for MFGs (Guo et al., 2019; Zaman et al., 2020; Xie et al., 2021), Multi-population MFGs (uz Zaman et al., 2023), graphon MFGs (Caines and Huang, 2019; Aurell et al., 2022; Cui and Koeppl, 2021) Independent RL for MFGs (Yongacoglu et al., 2022) and Oracle-free RL for MFGs (Zaman et al., 2023). The cooperative version of this paradigm is referred to as Mean-Field Control (MFC) problem (Bensoussan et al., 2013) which deals with an optimal control problem over a distribution of agents and there have been several works on Reinforcement Learning for MFC (Carmona et al., 2019a,b; Gu et al., 2021; Angiuli et al., 2022). But these works require ability to sample from the true transition model, and hence are inapplicable in the case of mis-specification or modeling errors. To address this setting, we consider the Robust MFC problem. Worth mentioning are some additional related works: the setting of Mean-Field Type Games, which contains the mixed cooperative-competitive elements (Choutri et al., 2016; Tembine, 2017; Carmona et al., 2020, 2021) in the form of a zero-sum competition between two infinitely large teams of agents. The works of (Carmona et al., 2020, 2021) also proposes a data-driven RL algorithm to compute the Nash equilibrium but provide only numerical analysis of the algorithm.

2. Formulation

In this section we introduce the robust multi-agent control problem by first defining the dynamics of the multi-agent system along with its performance and noise indices. The performance and noise indices have been introduced in the literature (Başar and Bernhard, 2008) in order to quantify the affect of the accumulated noise (referred to as noise index) on the performance of the system (called the performance index). The noise attenuation level is then defined as an upper bound on the ratio between the performance and noise indices given that the agents employ a robust controller. Hence the robust multi-agent problem is that of finding the robust controller under which a certain noise attenuation is achieved. In order to solve this problem, we reformulate it as a min-max game problem as in the single-agent setting (Başar and Bernhard, 2008).

Consider an N agent system. We let $[N] = \{1, \dots, N\}$. The i^{th} agent has dynamics which are linear in its state $x_t^i \in \mathbb{R}^m$, its action $u_t^{1,i} \in \mathbb{R}^p$, and the mean-field counterparts, $\bar{x}_t$ and $\bar{u}_t^i$. The disturbance $u_t^{2,i}$ is referred to as *non-stochastic* noise[1] since it is an un-modeled disturbance and can even be adversarial. This is similar in spirit to the works of (Simchowitz et al., 2020). Let T be a positive integer, interpreted as the horizon of the problem. The initial condition of agent i's state, $i \in [N]$, is $x_0^i = \omega^{0,i} + \bar{\omega}^0$, where $\omega^{0,i} \sim \mathcal{N}(0, \Sigma^0)$ and $\bar{\omega}^0 \sim \mathcal{N}(0, \bar{\Sigma}^0)$ are i.i.d. noises. For $t \in \{0, \dots, T-1\}$,

$$x_{t+1}^i = A_t x_t^i + \bar{A}_t \bar{x}_t + B_t u_t^{1,i} + \bar{B}_t \bar{u}_t^1 + u_t^{2,i} + \bar{u}_t^2 + \omega_t^i + \bar{\omega}_t, \forall i \in [N] \tag{1}$$

where $u_t^{1,i}$ is the control action of the i^{th} agent, $\bar{x}_t := \sum_{i=1}^N x_t^i / N$ is referred to as the state mean-field and $\bar{u}_t^j := \sum_{i=1}^N u^{j,i}/N$ for $j \in \{1, 2\}$ are the control and noise mean-fields respectively. Each agent's dynamics are perturbed by two types of noise: ω_t^i and $\bar{\omega}_t$ are referred to as stochastic noises since they are i.i.d. and their distributions are known ($\omega_t^i \sim \mathcal{N}(0, \Sigma)$ and $\bar{\omega}_t \sim \mathcal{N}(0, \bar{\Sigma})$). All of our results (excluding the finite-sample analysis of the RL Algorithm) can be readily generalized for zero-mean non-Gaussian disturbances with finite variance.

In order to define the robust control problem we define the *performance index* of the population which penalizes the deviation of the agents from their (state and control) mean-fields and also regulates the mean-fields:

$$J_N(u^1, u^2) = \frac{1}{N} \sum_{i=1}^N \mathbb{E} \sum_{t=0}^{T-1} \left[\|x_t^i - \bar{x}_t\|_{Q_t}^2 + \|\bar{x}_t\|_{\bar{Q}_t}^2 + \|u_t^{1,i} - \bar{u}_t^1\|^2 + \|\bar{u}_t^1\|^2 \right] + \|x_T^i - \bar{x}_T\|_{Q_T}^2 + \|\bar{x}_T\|_{\bar{Q}_T}^2 \tag{2}$$

where the matrices $Q_t, \bar{Q}_t > 0$ are symmetric matrices, $u^j = (u^{j,i})_{i \in [N]}$ where each $u^{j,i}$ for $j \in \{1, 2\}$ is adapted to the distribution information structure i.e. σ-algebra generated by x_t^i and $\bar{x}_t$ and $\mathcal{U}^1, \mathcal{U}^2$ represent the set of all

[1] The non-stochastic noise is assumed to have identity coefficient in the dynamics (1) for simplicity of analysis but can be easily changed to some other matrix of appropriate size.

possible u^1, u^2, respectively. We define the *noise index* of the population in a similar manner

$$\varpi_N(u^1,u^2) = \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\sum_{t=0}^{T-1}\Big[\|u_t^{2,i}-\bar{u}_t^2\|^2 + \|\bar{u}_t^2\|^2 + \|\omega_t^i\|^2 + \|\bar{\omega}_t\|^2\Big]. \tag{3}$$

The robust control problem for this N agent system is that of finding the range of noise attenuation levels $\gamma > 0$ such that:

$$\exists u^1 \in \mathscr{U}^1, \forall u^2 \in \mathscr{U}^2, \qquad J_N(u^1,u^2) \leq \gamma^2 \varpi_N(u^1,u^2) \tag{4}$$

Any γ for which the above inequality is satisfied is referred to as a *viable attenuation level* and the least among them is called the *minimum attenuation level*. The controller u^1 which ensures a particular level γ of noise attenuation is referred to as the *robust controller* corresponding to γ (or robust controller in short). Since the inequality (4) can also be reformulated as $J_N(\cdot)/\varpi_N(\cdot) \leq \gamma^2$, a viable attenuation parameter γ^2 is also an upper bound on the noise-to-output gain of the system. As outlined in (Başar and Bernhard, 2008) for a single agent problem the condition (4) is equivalent to finding the range of value of $\gamma > 0$ such that

$$\inf_{u^1}\sup_{u^2}\big(J_N(u^1,u^2) - \gamma^2\varpi_N(u^1,u^2)\big) \leq 0, \tag{5}$$

where the infimizing controller u^1 is the robust controller and the supremizing controller u^2 is the worst-case non-stochastic noise. If we define the robust N agent cost J_N^γ as follows

$$J_N^\gamma(u^1,u^2) = J_N(u^1,u^2) - \gamma^2\mathbb{E}\frac{1}{N}\sum_{i=1}^{N}\sum_{t=0}^{T-1}\big(\|u_t^{2,i}-\bar{u}_t^2\|^2 + \|\bar{u}_t^2\|^2\big),$$

then using (2) and (3), the robust N agent control problem (5) can be equivalently written as

$$\inf_{u^1}\sup_{u^2} J_N^\gamma(u^1,u^2) - \gamma^2\mathbb{E}\frac{1}{N}\sum_{i=1}^{N}\sum_{t=0}^{T-1}(\|\omega_t^i\|^2 + \|\bar{\omega}_t\|^2) \leq 0. \tag{6}$$

Due to the distributed information structure of the agents the standard theory of single-agent robust control does not apply in this setting. Hence we are unable to provide sufficient conditions for a given $\gamma > 0$ to be a viable attenuation level, and we resort to the mean-field limit as $N \to \infty$, which is of independent interest. The next section formulates the Robust Mean-Field Control (RMFC) problem and its equivalent 2-player zero-sum Mean-Field Type Game (ZS-MFTG) representation, and provides sufficient conditions for solvability of both.

3. Robust Mean-Field Control

Consider a system with infinitely many agents, where the generic agent has linear dynamics of its state x_t for a finite-horizon $t \in \{0,\dots,T-1\}$:

$$x_{t+1} = A_t x_t + \bar{A}_t\bar{x}_t + B_t u_t^1 + \bar{B}_t\bar{u}_t^1 + u_t^2 + \bar{u}_t^2 + \omega_t + \bar{\omega}_t, \tag{7}$$

where u_t^1 is the control action of the generic agent, $\bar{x}_t := \mathbb{E}[x_t|(\bar{\omega}_s)_{0\leq s\leq t-1}]$ is referred to as the state mean-field and $\bar{u}_t^j := \mathbb{E}[u_t^j|(\bar{\omega}_s)_{0\leq s\leq t-1}]$ for $j \in \{1,2\}$ are the control and noise mean-fields respectively. The initial condition of the generic agent is $x_0 = \omega^0 + \bar{\omega}^0$, where $\omega^0 \sim \mathscr{N}(0,\Sigma^0)$ and $\bar{\omega}^0 \sim \mathscr{N}(0,\bar{\Sigma}^0)$ are i.i.d. noises. The stochastic noises ω_t^i and $\bar{\omega}_t$ are i.i.d. such that $\omega_t^i \sim \mathscr{N}(0,\Sigma)$ and $\bar{\omega}_t \sim \mathscr{N}(0,\bar{\Sigma})$, whereas the non-stochastic noise u_t^2 are un-modeled uncertainties. Similar to the N agent case, we define the robust mean-field cost J^γ as follows

$$J^\gamma(u^1,u^2) = \mathbb{E}\sum_{t=0}^{T}\Big[\|x_t-\bar{x}_t\|_{Q_t}^2 + \|\bar{x}_t\|_{\bar{Q}_t}^2 + \|u_t^1-\bar{u}_t^1\|^2 + \|\bar{u}_t^1\|^2 - \gamma^2\big(\|u_t^2-\bar{u}_t^2\|^2 + \|\bar{u}_t^2\|^2\big) \tag{8}$$
$$+ \|x_T-\bar{x}_T\|_{Q_T}^2 + \|\bar{x}_T\|_{\bar{Q}_T}^2\Big].$$

Now the robust mean-field control problem which is the mean-field analog to (6) is defined as follows.

Definition 1 (Robust Mean-Field Control problem) *If for a given $\gamma > 0$ the following inequality is satisfied, then γ is a viable noise attenuation level for the robust mean-field control problem.*

$$\inf_{u^1}\sup_{u^2} J^\gamma(u^1,u^2) - \gamma^2\mathbb{E}\sum_{t=0}^{T-1}\|\omega_t\|^2 + \|\bar{\omega}_t\|^2 \leq 0. \tag{9}$$

Moreover, the infimizing controller u^1 in (9) is a robust controller (corresponding to γ).

Now, under the condition of interchangability of the inf and sup operations, the problem of finding $\inf_{u^1}\sup_{u^2} J^\gamma(u^1,u^2)$ is that of finding the Nash equilibrium (equivalently, saddle point, in this case) of the *Zero-sum 2-player Mean-Field Type Game*; see (Carmona et al., 2020, 2021) for a very similar LQ setting without the theoretical analysis of the RL algorithm. In the following section we provide sufficient conditions for existence and uniqueness of a solution to this saddle point problem along with the value of $\inf_{u^1}\sup_{u^2} J^\gamma(u^1,u^2)$.

2-player Zero-sum Mean-Field Type Games: Let us define $y_t = x_t - \bar{x}_t$, $z_t = \bar{x}_t$. The dynamics of y_t and z_t can be written as

$$y_{t+1} = A_t y_t + B_t(u^1_t - \bar{u}^1_t) + u^2_t - \bar{u}^2_t + \omega_t - \bar{\omega}_t, \quad z_{t+1} = \tilde{A}_t z_t + \tilde{B}_t\bar{u}^1_t + 2\bar{u}^2_t + 2\bar{\omega}_t,$$

where $\tilde{A}_t = A_t + \bar{A}_t$ and $\tilde{B}_t = B_t + \bar{B}_t$. The optimal controls are known to be linear (Carmona et al., 2020), hence we restrict our attention the set of linear controls in y_t and z_t,

$$u^1_t = u^1_t(x_t,\bar{x}_t) = -K^1_t(x_t - \bar{x}_t) - L^1_t\bar{x}_t, \quad u^2_t = u^2_t(x_t,\bar{x}_t) = K^2_t(x_t - \bar{x}_t) + L^2_t\bar{x}_t$$

which implies that $\bar{u}^1_t = -L^1_t\bar{x}_t$ and $\bar{u}^2_t = L^2_t\bar{x}_t$. The dynamics of the processes y_t and z_t can be re-written as

$$y_{t+1} = (A_t - B_tK^1_t + K^2_t)y_t + \omega_t - \bar{\omega}_t, \quad z_{t+1} = (\tilde{A}_t - \tilde{B}_tL^1_t + L^2_t)z_t + 2\bar{\omega}_t. \tag{10}$$

Since the dynamics of y_t and z_t are decoupled, we can decompose the cost J^γ into the following two parts:

$$\begin{aligned} J^\gamma(K,L) &= J^\gamma_y(K) + J^\gamma_z(L), \\ J^\gamma_y(K) &= \mathbb{E}\Big[\sum_{t=0}^{T-1} y_t^\top(Q_t + (K^1_t)^\top K^1_t - \gamma^2(K^2_t)^\top K^2_t)y_t + y_T^\top Q_T y_T\Big], \\ J^\gamma_z(L) &= \mathbb{E}\Big[\sum_{t=0}^{T-1} z_t^\top(\bar{Q}_t + (L^1_t)^\top L^1_t - \gamma^2(L^2_t)^\top L^2_t)z_t + z_T^\top\bar{Q}_T z_T\Big]. \end{aligned} \tag{11}$$

The 2-player MFTG (7)-(8) has been decoupled into two 2-player LQ dynamic game problems as shown below:

$$\min_{K^1_t,L^1_t}\max_{K^2_t,L^2_t} J^\gamma((K^1_t,K^2_t),(L^1_t,L^2_t)) = \min_{K^1_t}\max_{K^2_t} J^\gamma_y(K) + \min_{L^1_t}\max_{L^2_t} J^\gamma_z(L)$$

where the dynamics of y_t and z_t are defined in (10). In the following section, using results in the literature, we specify the sufficient conditions for existence and uniqueness of Nash equilibrium of the 2-player MFTG and also present the *value* (Nash cost) of the game. Building on the techniques developed in (Başar and Olsder, 1998; Carmona et al., 2020), we can prove the following result.

Theorem 2 *If for a given $\gamma > 0$,*

$$\gamma^2 I - M^\gamma_t > 0 \text{ and } \gamma^2 I - \bar{M}^\gamma_t > 0 \tag{12}$$

where M^γ_t and $\bar{M}^\gamma_t$ are positive semi-definite matrices which satisfy the Coupled Algebraic Riccati equations,

$$\begin{aligned} &M^\gamma_t = Q_t + A_t^\top M^\gamma_{t+1}\Lambda_t^{-1}A_t, \;\; \Lambda_t = I + (B_tB_t^\top - \gamma^{-2}I)M^\gamma_{t+1}, \;\; M^\gamma_T = Q_T, \\ &\bar{M}^\gamma_t = \bar{Q}_t + \tilde{A}_t^\top\bar{M}^\gamma_{t+1}\bar{\Lambda}_t^{-1}\tilde{A}_t, \;\; \bar{\Lambda}_t = I + (\tilde{B}_t\tilde{B}_t^\top - \gamma^{-2}I)\bar{M}^\gamma_{t+1}, \;\; \bar{M}^\gamma_T = \bar{Q}_T \\ &N^\gamma_t = N^\gamma_{t+1} + \mathrm{Tr}(M^\gamma_{t+1}\Sigma), \;\; N^\gamma_T = 0, \;\; \bar{N}^\gamma_t = \bar{N}^\gamma_{t+1} + \mathrm{Tr}(\bar{M}^\gamma_{t+1}\Sigma), \;\; \bar{N}^\gamma_T = 0 \end{aligned} \tag{13}$$

then, $u_t^{1*} = -K_t^{1*}(x_t - \bar{x}_t) - L_t^{1*}\bar{x}_t$ *and* $u_t^{2*} = K_t^{2*}(x_t - \bar{x}_t) + L_t^{2*}\bar{x}_t$ *(complete expressions provided in Supplementary Materials) are the unique Nash policies. Furthermore, the Nash equilibrium (equivalently, saddle point) value is*

$$\inf_{u^1} \sup_{u^2} J^\gamma(u^1, u^2) = \mathrm{Tr}(M_0^\gamma \Sigma^0) + \mathrm{Tr}(\bar{M}_0^\gamma \bar{\Sigma}^0) + N_0^\gamma + \bar{N}_0^\gamma \tag{14}$$

This result can be proved using techniques in proofs of Theorem 3.2 in (Başar and Bernhard, 2008) or Proposition 36 in (Carmona et al., 2021). We now use the Nash value of the game (14) to come up with a condition for the attenuation level γ which solves the robust mean-field control problem (9). First we simplify expression in (9) $\mathbb{E}\sum_{t=0}^{T-1}\|\omega_t\|^2 + \|\bar{\omega}_t\|^2 = T\,\mathrm{Tr}(\Sigma + \bar{\Sigma})$ using the i.i.d. stochastic nature of the noise. Combining this fact with (14), we arrive at the conclusion that (9) will be satisfied if and only if

$$\sum_{t=1}^{T} \mathrm{Tr}((M_t^\gamma - \gamma^2 I)\Sigma + (\bar{M}_t^\gamma - \gamma^2 I)\bar{\Sigma}) + \mathrm{Tr}(M_0^\gamma \Sigma^0) + \mathrm{Tr}(\bar{M}_0^\gamma \bar{\Sigma}^0) \leq 0 \tag{15}$$

Notice that the conditions (12) and (15) are different, as the first one requires positive definiteness of matrices and the second one requires a scalar inequality. Now we solve the robust N agent control problem by providing sufficient conditions for a given attenuation level γ satisfying (4).

Theorem 3 *Let* $\gamma > 0$*. If, in addition to* (12)*, we also have*

$$\sum_{t=1}^{T} \mathrm{Tr}((M_t^\gamma - \gamma^2 I)\Sigma + (\bar{M}_t^\gamma - \gamma^2 I)\bar{\Sigma}) + \mathrm{Tr}(M_0^\gamma \Sigma^0) + \mathrm{Tr}(\bar{M}_0^\gamma \bar{\Sigma}^0) \leq -\frac{CT}{N} \tag{16}$$

where C *is a constant which depends only on the model parameters and* M_t^γ *and* $\bar{M}_t^\gamma$ (13)*, then* γ *is a viable attenuation level for the Robust* N *agent control problem* (4)*. Moreover the robust controller for each agent* i *is given by* $u_t^{1,i*} = -K_t^{1*}(x_t^i - \bar{x}_t) - L_t^{1*}\bar{x}_t$*.*

The proof of this result can be found in the Supplementary Materials. The above theorem states that, if for a given γ, conditions (12) and (16) are satisfied (given that M_t^γ and $\bar{M}_t^\gamma$ are defined by (13)), then not only is γ a viable attenuation level for the original Robust multi-agent control problem (1)-(4), but the Nash equilibrium for the ZS-MFTG also yields the robust controller $u_t^{1,i*} = -K_t^{1*}(x_t^i - \bar{x}_t) - L_t^{1*}\bar{x}_t$ for the original finite-agent game. Condition (16) is strictly stronger than condition (15) but approaches (16) as $N \to \infty$.

4. Reinforcement Learning for Robust Mean-Field Control

In this section we present the Receding-horizon policy Gradient Descent Ascent (RGDA) algorithm to compute the Nash equilibrium (Theorem 2) of the 2-player MFTG (7)-(8), which will also generate the robust controller for a fixed noise attenuation level γ. For this section we assume access to only the finite-horizon costs of the agents under a set of control policies, and not the state trajectories. Under this setting the model of the agents cannot be constructed hence our approach is *truly* model free (Malik et al., 2019). Due to the non-convex non-concave (also non-coercive (Zhang et al., 2020b)) nature of the cost function J^γ in (11), instead we solve the receding-horizon problem, for each $t = \{T-1, \ldots, 1, 0\}$ backwards-in-time. This entails solving $2 \times T$ min-max problems, where each problem is convex-concave and aims at finding $(K_t, L_t) = \big((K_t^1, K_t^2), (L_t^1, L_t^2)\big)$ at time step t, given the set of *future* controllers (controllers for times greater than t), $\big((\tilde{K}_{t+1}, \tilde{L}_{t+1}), \ldots, (\tilde{K}_T, \tilde{L}_T)\big)$ are held constant. But first we must approximate the mean-field term using finitely many agents.

Approximation of mean-field terms using M agents: Since simulating infinitely many agents is impractical, in this section we outline how to use a set of $2 \leq M < \infty$ agents to approximately simulate the mean-field in a MFTG. Each of the M agents has state x_t^i at time t where $i \in [M]$. The agents follow controllers linear in their private state and empirical mean-field, x_t^i and $\tilde{z}_t$, respectively:

$$u_t^1 = -K_t^1(x_t^i - \tilde{z}_t) - L_t^1\tilde{z}_t, \qquad u_t^2 = K_t^2(x_t^i - \tilde{z}_t) + L_t^2\tilde{z}_t,$$

where the empirical mean-field is $\tilde{z}_t := \frac{1}{M}\sum_{i=1}^{M} x_t^i$. Under these control laws, the dynamics of agent $i \in [M]$ are

$$x_{t+1}^i = (A_t - B_t K_t^1 + K_t^2)(x_t^i - \tilde{z}_t) + (\tilde{A}_t - \tilde{B}_t L_t^1 + L_t^2)\tilde{z}_t + \omega_{t+1}^i + \bar{\omega}_t$$

and the dynamics of the empirical mean-field $\tilde{z}_t$ is

$$\tilde{z}_{t+1} = (\tilde{A}_t - \tilde{B}_t L_t^1 + L_t^2)\tilde{z}_t + \tilde{\omega}_{t+1}^0, \quad \text{where } \tilde{\omega}_{t+1}^0 = \bar{\omega}_t + \frac{1}{M}\sum_{i=1}^{M}\omega_{t+1}^i$$

The cost of each agent is

$$\begin{aligned}\tilde{J}^{i,\gamma}(u_1,u_2) =& \mathbb{E}\Big[\sum_{t=0}^{T-1}(x_t^i - \tilde{z}_t)^\top[Q_t + (K_t^1)^\top K_t^1 - \gamma^2 (K_t^2)^\top K_t^2](x_t^i - \tilde{z}_t) + (x_T^i - \tilde{z}_T)^\top Q_T (x_T^i - \tilde{z}_T) \\ &+ \tilde{z}_t^\top[\bar{Q}_t + (L_t^1)^\top L_t^1 - \gamma^2 (L_t^2)^\top L_t^2]\tilde{z}_t + \tilde{z}_T^\top \bar{Q}_T \tilde{z}_T\Big].\end{aligned}$$

Now, similarly to the previous section, we define $y_t^i = x_t^i - \tilde{z}_t$. The dynamics of y_t^i are

$$y_{t+1}^i = (A_t - B_t K_t^1 + K_t^2) y_t^i + \tilde{\omega}_{t+1}^i, \quad \text{where } \tilde{\omega}_{t+1}^i = \frac{M-1}{M}\omega_{t+1}^i - \frac{1}{M}\sum_{j \neq i}\omega_{t+1}^j.$$

The cost can then be decomposed in a manner similar to (11):

$$\begin{aligned}\tilde{J}^{i,\gamma}\big((K_t^1,K_t^2),(L_t^1,L_t^2)\big) &= \tilde{J}_y^{i,\gamma}(K_t^1,K_t^2) + \tilde{J}_z^{i,\gamma}(L_t^1,L_t^2), \\ \tilde{J}_y^{i,\gamma}(K_t^1,K_t^2) &= \mathbb{E}\Big[\sum_{t=0}^{T-1}(y_t^i)^\top[Q_t + (K_t^1)^\top K_t^1 - \gamma^2 (K_t^2)^\top K_t^2] y_t^i + (y_T^i)^\top Q_T y_T^i\Big], \\ \tilde{J}_z^{i,\gamma}(L_t^1,L_t^2) &= \mathbb{E}\Big[\sum_{t=0}^{T-1}\tilde{z}_t^\top[\bar{Q}_t + (L_t^1)^\top L_t^1 - \gamma^2 (L_t^2)^\top L_t^2]\tilde{z}_t + \tilde{z}_T^\top \bar{Q}_T \tilde{z}_T\Big].\end{aligned} \tag{17}$$

Receding-horizon approach: Similar to the approach in Section 2, instead of finding the optimal, K^* and L^* which optimizes $\tilde{J}$ in (17), we solve the receding-horizon problem for each $t = \{T-1,,\ldots,1,0\}$ backwards-in-time. This forms two decoupled min-max convex-concave problems of finding $(K_t, L_t) = \big((K_t^1,K_t^2),(L_t^1,L_t^2)\big)$ at each time step t, given the set of controllers for times greater than t, $\big((\tilde{K}_{t+1},\tilde{L}_{t+1}),\ldots,(\tilde{K}_T,\tilde{L}_T)\big)$

$$\begin{aligned}&\min_{(K_t^1,L_t^1)}\max_{(K_t^2,L_t^2)} \tilde{J}_t^{i,\gamma}(K_t,L_t) = \\ &\underbrace{\mathbb{E}\Big[y_t^\top(Q_t + (K_t^1)^\top K_t^1 - \gamma^2 (K_t^2)^\top K_t^2) y_t + \sum_{k=t+1}^{T} y_k^\top (Q_t + (\tilde{K}_k^1)^\top \tilde{K}_k^1 - \gamma^2 (\tilde{K}_k^2)^\top \tilde{K}_k^2) y_k\Big]}_{\tilde{J}_{y,t}^{i,\gamma}} \\ &+ \underbrace{\mathbb{E}\Big[z_t^\top(\bar{Q}_t + (L_t^1)^\top L_t^1 - \gamma^2 (L_t^2)^\top L_t^2) z_t + \sum_{k=t+1}^{T} z_k^\top (\bar{Q}_t + \tilde{L}_{1,k}^\top \tilde{L}_k^1 - \gamma^2 (\tilde{L}_k^2)^\top \tilde{L}_k^2) z_k\Big]}_{\tilde{J}_{z,t}^{i,\gamma}}.\end{aligned} \tag{18}$$

for any $i \in [M]$ and $y_t \sim \mathcal{N}(0,\Sigma_y), z_t \sim \mathcal{N}(0,\Sigma_z)$. This receding-horizon problem is solved using Receding-horizon policy Gradient Descent Ascent (RGDA) (Algorithm 1) where at each time instant t the Nash control is approached using gradient descent ascent. We anticipate a small approximation error between the optimal controller and its computed approximation $\tilde{K}_t$ (respectively $\tilde{L}_t$). However, this error is shown to be well-behaved (Theorem 5), as we progress backwards-in-time, given that the hyper-parameters of RGDA satisfy certain bounds.

Receding-horizon policy Gradient Descent Ascent (RGDA) Algorithm: The RGDA Algorithm (Algorithm 1 is a bi-level optimization algorithm where the outer loop starts at time $t = T-1$ and moves backwards-in-time, and the inner loop is a gradient descent (for control parameters (K_t^1, L_t^1)) ascent (for control policy (K_t^2, L_t^2)) update with learning rate η_k. The gradient descent ascent step entails computing an approximation of the *exact* gradients of cost $\tilde{J}_t^{i,\gamma}$ with respect to the controls variables $(K_t^1, L_t^1), (K_t^2, L_t^2)$. To obtain this approximation in a data driven manner we utilize a zero-order stochastic gradient $\tilde{\nabla}_1 \tilde{J}_t^{i,\gamma}(K_t, L_t), \tilde{\nabla}_2 \tilde{J}_t^{i,\gamma}(K_t, L_t)$ (Fazel et al., 2018; Malik et al., 2019) which requires cost computation under a given set of controllers (18) as shown below.

$$\tilde{\nabla}_1 \tilde{J}_t^{i,\gamma}(K_t, L_t) = \frac{n}{Mr^2} \sum_{j=1}^{M} \tilde{J}_t^{i,\gamma}((K_t^{j,1}, K_t^2), (L_t^{j,1}, L_t^2)) e_j, \quad \begin{pmatrix} K_t^{j,1} \\ L_t^{j,1} \end{pmatrix} = \begin{pmatrix} K_t^1 \\ L_t^1 \end{pmatrix} + e_j, \quad e_j \sim \mathbb{S}^{n-1}(r)$$

$$\tilde{\nabla}_2 \tilde{J}_t^{i,\gamma}(K_t, L_t) = \frac{n}{Mr^2} \sum_{j=1}^{M} \tilde{J}_t^{i,\gamma}((K_t^1, K_t^{j,2}), (L_t^1, L_t^{j,2})) e_j, \quad \begin{pmatrix} K_t^{j,2} \\ L_t^{j,2} \end{pmatrix} = \begin{pmatrix} K_t^2 \\ L_t^2 \end{pmatrix} + e_j, \quad e_j \sim \mathbb{S}^{n-1}(r)$$

Stochastic gradient computation entails computing the cost of N_b different *perturbed* controllers, with a perturbation magnitude if r also called the *smoothing radius*. This stochastic gradient provides us with a *biased* approximation of the exact gradient whose *bias* and *variance* can be controlled by tuning the values of N_b and r. Finally to ensure stability of the learning algorithm, we use projection Proj_D onto a D-ball such that the norm of the matrices is bounded by D, $\|(K_t, L_t)\|^2 \leq D$. The radius of the ball D is chosen such that the Nash equilibrium controllers lie within this ball.

Algorithm 1 RGDA Algorithm for 2-player MFTG

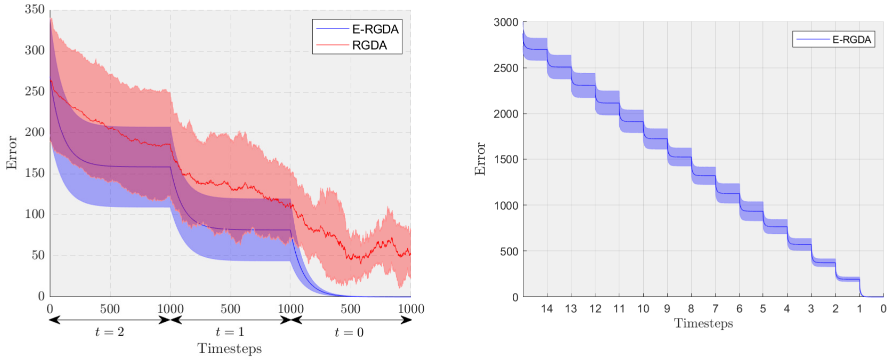
```
for $t = T-1, \ldots, 1, 0,$ do
    Initialize $K_t = (K_t^1, K_t^2) = 0, L_t = (L_t^1, L_t^2) = 0$
    for $k = 0, \ldots, K$ do
        Gradient Descent $\begin{pmatrix} K_t^1 \\ L_t^1 \end{pmatrix} \leftarrow \text{Proj}_D \left( \begin{pmatrix} K_t^1 \\ L_t^1 \end{pmatrix} - \eta_k \tilde{\nabla}_1 \tilde{J}_t^{i,\gamma}(K_t, L_t) \right),$
        Gradient Ascent $\begin{pmatrix} K_t^2 \\ L_t^2 \end{pmatrix} \leftarrow \text{Proj}_D \left( \begin{pmatrix} K_t^2 \\ L_t^2 \end{pmatrix} + \eta_k \tilde{\nabla}_2 \tilde{J}_t^{i,\gamma}(K_t, L_t) \right),$
    end for
end for
```

RGDA algorithm analysis: In this section we start by showing linear convergence of the inner loop gradient descent ascent (Theorem 4), which is made possible by the convex-concave property of the cost function under the receding horizon approach (18). Then we show that if the error accumulated in each inner loop computation is small enough, the total accumulated error is well behaved (Theorem 5).

We first define some relevant notation. We define the *joint controllers* for each timestep t as $\bar{K}_t = [(K_t^1)^\top, (K_t^2)^\top]^\top$ and $\bar{L}_t = [(L_t^1)^\top, (L_t^2)^\top]^\top$, for the sake of conciseness. For each timestep $t \in \{T-1, \ldots, 1, 0\}$ let us also define the *target* joint controllers $\tilde{\bar{K}}_t^* = (\tilde{K}_t^{1*}, \tilde{K}_t^{2*}), \tilde{\bar{L}}_t^* = (\tilde{L}_t^{1*}, \tilde{L}_t^{2*})$, as the set of policies which exactly solve the receding-horizon min-max problem (18). Notice that the set of target controllers $\tilde{\bar{K}}_t^*, \tilde{\bar{L}}_t^*$ are unique (due to convex-concave nature of (18)) but do depend on the set of future joint controllers $(\bar{K}_s, \bar{L}_s)_{t<s<T}$. On the other hand, the Nash joint controllers are denoted by $\bar{K}_t^* = (K_t^{1*}, K_t^{2*})$ and $\bar{L}_t^* = (L_t^{1*}, L_t^{2*})$. Furthermore, the target joint controllers are equal to the Nash joint controllers $(\tilde{\bar{K}}_t^*, \tilde{\bar{L}}_t^*) = (\bar{K}_t^*, \bar{L}_t^*)$ only if the future joint controllers are also Nash $(\bar{K}_s, \bar{L}_s)_{t<s<T} = (\bar{K}_s^*, \bar{L}_s^*)_{t<s<T}$.

Theorem 4 *If the learning rate η_k is smaller than a certain function of model parameters, the number of inner loop iterations $K = \Omega(\log(1/\epsilon))$, the mini-batch size $N_b = \Omega(1/\epsilon)$ and the smoothing radius $r = \mathcal{O}(\epsilon)$, then at each timestep $t \in \{T-1, \ldots, 1, 0\}$ the optimality gaps are $\|\bar{K}_t - \tilde{\bar{K}}_t^*\|_2^2 \leq \epsilon$ and $\|\bar{L}_t - \tilde{\bar{L}}_t^*\|_2^2 \leq \epsilon$.*

Closed form expressions of the bounds can be found in the proof given in the Supplementary materials. The linear rate of convergence is made possible by building upon the convergence analysis of descent ascent in (Fallah et al.,



(*a*)

(*b*)

Figure 1: Performance of RGDA Algorithm

2020) due to the convex-concave nature of the cost function (18). The proof generalizes the techniques used in (Fallah et al., 2020) to stochastic unbiased gradients by utilizing the fact that the bias in stochastic gradients $\tilde{\nabla}_j \tilde{J}_t^{i,\gamma}$ for $j \in \{1,2\}$ can be reduced by reducing the smoothing radius r. This in turn causes an increase in the variance of the stochastic gradient which is controlled by increasing the mini-batch size N_b.

Now we present the non-asymptotic convergence guarantee of the paper stating that even though each iteration of the outer loop (as timestep t moves backwards-in-time) accumulates error, if the error in each outer loop iteration is small enough, the total accumulated error will also be small enough. The proof can be found in the Supplementary Materials.

Theorem 5 *If all conditions in Theorem 4 are satisfied, then* $\max_{j\in\{1,2\}}\|K_t^j - K_t^{j*}\| = \mathcal{O}(\epsilon)$ *and* $\max_{j\in\{1,2\}}\|L_t^j - L_t^{j*}\| = \mathcal{O}(\epsilon)$ *for a small* $\epsilon > 0$ *and* $t \in \{T-1, \dots, 0\}$.

The Nash gaps at each time t, $\|K_t^j - K_t^{j*}\|$ and $\|L_t^j - L_t^{j*}\|$ for $j \in \{1,2\}$ are due to a combination of the optimality gap in the inner loop $\|\bar{K}_t - \tilde{\bar{K}}_t^*\|_2^2, \|\bar{L}_t - \tilde{\bar{L}}_t^*\|_2^2$ and the accumulated Nash gap in the future joint controllers $\|K_s^j - K_s^{j*}\|$ and $\|L_s^j - L_s^{j*}\|$ for $j \in \{1,2\}$ and $t < s < T$. The proof of Theorem 5 characterizes these two quantities and then shows that if the optimality gap at each timestep $t \in \{0, \dots, T-1\}$ never exceeds some small ϵ, then the Nash gap at any time t never exceeds ϵ scaled by a constant.

5. Numerical Analysis

We simulate the RGDA algorithm for time horizon $T = 3$, number of agents $M = 1000$ and the dimension of the state and action spaces $m = p = 2$. For each timestep $t \in \{2,1,0\}$, the number of inner-loop iterations $K = 1000$, the mini-batch size $N_b = 5 \times 10^4$ and the learning rate $\eta_k = 0.001$. In Figure 1(*a*) we compare the RGDA algorithm (Algorithm 1) with its exact version (E-RGDA) which has access to the exact policy gradients $\nabla_1 \tilde{J}_t^{i,\gamma} = \delta \tilde{J}_t^{i,\gamma} / \delta(K_t^1, L_t^1)$ and $\nabla_2 \tilde{J}_t^{i,\gamma} = \delta \tilde{J}_t^{i,\gamma} / \delta(K_t^2, L_t^2)$ at each iteration $k \in [K]$. The error plots in Figures 1(*a*) and 1(*b*) show the mean (solid lines) and standard deviation (shaded regions) of error, which is the norm of difference between iterates and Nash controllers. In Figure 1(*a*) the blue plot shows error convergence of the E-RGDA algorithm, which computes the Nash controllers for the last timestep $t = 2$ (using gradient descent ascent with exact gradients) and moves backwards in time. Since at each timestep it has good convergence to Nash policies, the convexity-concavity of cost function at the next timestep is ensured, which results in linear convergence. The red plot in Figure 1(*a*) shows the error convergence in the RGDA algorithm which uses stochastic gradients, which results in a noisy but downward trend in error. Notice that RGDA imitates E-RGDA in a noisy fashion and at each timestep the iterates only approximate the Nash controllers. This approximation can be further sharpened by increasing the mini-batch size N_b and decreasing smoothing radius r. Figure 1(*a*) shows the error convergence of E-RGDA for a ZS-MFTG with time-horizon $T = 15$ and state and action space dimensions $m = p = 2$.

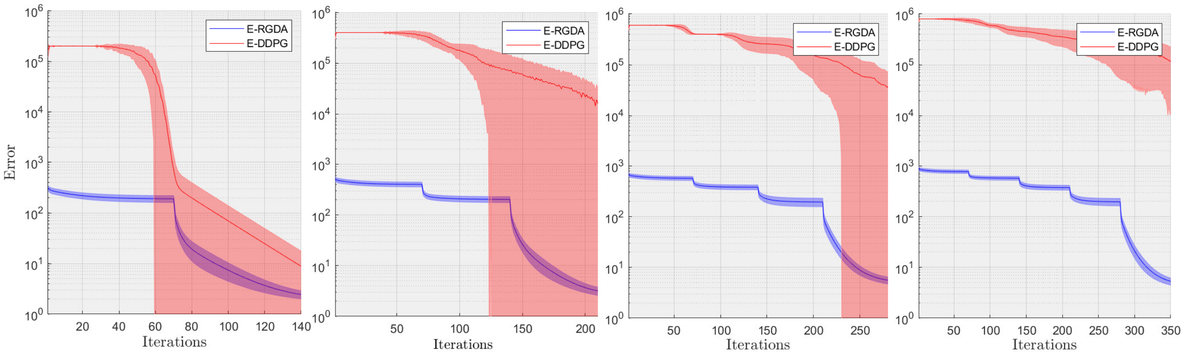


Figure 2: Comparison between E-RGDA and E-DDPG. The time-horizon is increasing from left to right with $T = 2$ (left-most), $T = 3$ (center left), $T = 4$ (center right) and $T = 5$ (right-most)

Figure 2 compares the E-RGDA algorithm with the exact 2-player zero-sum version of the MADPG algorithm (referred to as E-DDPG) (Lowe et al., 2017) which serves as a baseline as it does not use the receding-horizon approach. The number of inner-loop iterations for E-RGDA is $K = 70$ and the learning rate for both algorithms is $\eta = 0.025$. The four figures represent the comparisons for $T = \{2, 3, 4, 5\}$ and the y-axis is scaled in a logarithmic manner to best show the behavior of the algorithms. For all $T > 1$ the E-DDPG first diverges until it reaches the projection threshold then eventually starts to converge. This is due to the fact that errors in later timesteps cause the convexity-concavity condition to fail resulting in divergence in earlier timesteps. Over time the error decreases in the later timesteps, which causes the error in earlier timesteps to gradually decrease as well. But as seen from Figure 2, the convergence for E-DDPG takes significantly longer as the time-horizon increases.

6. Conclusion

In this paper, we solve a Multi-agent Reinforcement Learning problem with the objective of designing robust controllers in the presence of modeled and un-modeled uncertainties. We resort to the Linear-Quadratic (LQ) structure and mean-field paradigm to aid in tractability and to help resolve the analytical difficulty induced by the distributed information structure. This helps us obtain sufficient conditions for robustness of the problem as well as characterization of the robust control policy. We design and provide non-asymptotic analysis of a receding-horizon based Reinforcement Learning (RL) algorithm which renders the non-coercive cost as convex-concave. Through numerical analysis the receding-horizon approach is shown to ameliorate the overshooting problem observed in the performance of the vanilla algorithm. In future work we would like to explore RL problems which generalize the LQ setting and algorithms which go beyond the gradient descent-ascent updates used in this paper.

Acknowledgments

The authors would like to thank Xiangyuan Zhang (University of Illinois Urbana-Champaign) for useful discussions regarding the Receding-horizon Policy Gradient algorithm.

References

Andrea Angiuli, Jean-Pierre Fouque, and Mathieu Laurière. Unified reinforcement Q-learning for mean field game and control problems. *Mathematics of Control, Signals, and Systems*, pages 1–55, 2022.

Michael Athans, David Castanon, K-P Dunn, C Greene, Wing Lee, N Sandell, and A Willsky. The stochastic control of the f-8c aircraft using a multiple model adaptive control (mmac) method–part i: Equilibrium flight. *IEEE Transactions on Automatic Control*, 22(5):768–780, 1977.

Alexander Aurell, René Carmona, Gökçe Dayanıklı, and Mathieu Laurière. Finite state graphon games with applications to epidemics. *Dynamic Games and Applications*, 12(1):49–81, 2022.

Tamer Başar. A dynamic games approach to controller design: Disturbance rejection in discrete time. In *Proceedings of the 28th IEEE Conference on Decision and Control,*, pages 407–414. IEEE, 1989.

Tamer Başar and Pierre Bernhard. *H-infinity optimal control and related minimax design problems: a dynamic game approach*. Springer Science & Business Media, 2008.

Tamer Başar and Geert Jan Olsder. *Dynamic noncooperative game theory*. SIAM, 1998.

Alain Bensoussan, Jens Frehse, Phillip Yam, et al. *Mean field games and mean field type control theory*, volume 101. Springer, 2013.

Peter E Caines and Minyi Huang. Graphon mean field games and the GMFG equations: ε-Nash equilibria. In *2019 IEEE 58th Conference on Decision and Control (CDC)*, pages 286–292. IEEE, 2019.

Pierre Cardaliaguet and Charles-Albert Lehalle. Mean field game of controls and an application to trade crowding. *Mathematics and Financial Economics*, 12(3):335–363, 2018.

René Carmona, Mathieu Laurière, and Zongjun Tan. Linear-quadratic mean-field reinforcement learning: convergence of policy gradient methods. *arXiv preprint arXiv:1910.04295*, 2019a.

René Carmona, Mathieu Laurière, and Zongjun Tan. Model-free mean-field reinforcement learning: mean-field mdp and mean-field q-learning. *arXiv preprint arXiv:1910.12802*, 2019b.

René Carmona, Kenza Hamidouche, Mathieu Laurière, and Zongjun Tan. Policy optimization for linear-quadratic zero-sum mean-field type games. In *2020 59th IEEE Conference on Decision and Control (CDC)*, pages 1038–1043. IEEE, 2020.

René Carmona, Kenza Hamidouche, Mathieu Laurière, and Zongjun Tan. Linear-quadratic zero-sum mean-field type games: Optimality conditions and policy optimization. *Journal of Dynamics & Games*, 8(4), 2021.

Salah Eddine Choutri, Boualem Djehiche, and Hamidou Tembine. Optimal control and zero-sum games for Markov chains of mean-field type. *arXiv preprint arXiv:1606.04244*, 2016.

Kai Cui and Heinz Koeppl. Learning graphon mean field games and approximate nash equilibria. *arXiv preprint arXiv:2112.01280*, 2021.

Alireza Fallah, Asuman Ozdaglar, and Sarath Pattathil. An optimal multistage stochastic gradient method for minimax problems. In *2020 59th IEEE Conference on Decision and Control (CDC)*, pages 3573–3579. IEEE, 2020.

Maryam Fazel, Rong Ge, Sham M Kakade, and Mehran Mesbahi. Global convergence of policy gradient methods for the linear quadratic regulator. In *International Conference on Machine Learning*, pages 1467–1476, 2018.

Haotian Gu, Xin Guo, Xiaoli Wei, and Renyuan Xu. Mean-field controls with Q-learning for cooperative MARL: convergence and complexity analysis. *SIAM Journal on Mathematics of Data Science*, 3(4):1168–1196, 2021.

Xin Guo, Anran Hu, Renyuan Xu, and Junzi Zhang. Learning mean-field games. In *Advances in Neural Information Processing Systems*, 2019.

Charles Harvey and Gunter Stein. Quadratic weights for asymptotic regulator properties. *IEEE Transactions on Automatic Control*, 23(3):378–387, 1978.

Sihong He, Songyang Han, Sanbao Su, Shuo Han, Shaofeng Zou, and Fei Miao. Robust multi-agent reinforcement learning with state uncertainty. *Transactions on Machine Learning Research*, 2023.

Minyi Huang, Peter E Caines, and Roland P Malhamé. Individual and mass behaviour in large population stochastic wireless power control problems: Centralized and Nash equilibrium solutions. In *IEEE International Conference on Decision and Control*, volume 1, pages 98–103. IEEE, 2003.

Minyi Huang, Roland P Malhamé, and Peter E Caines. Large population stochastic dynamic games: Closed-loop Mckean-Vlasov systems and the Nash certainty equivalence principle. *Communications in Information & Systems*, 6(3):221–252, 2006.

Jens Kober, J Andrew Bagnell, and Jan Peters. Reinforcement learning in robotics: A survey. *The International Journal of Robotics Research*, 32(11):1238–1274, 2013.

Jernej Kos and Dawn Song. Delving into adversarial attacks on deep policies. *arXiv preprint arXiv:1705.06452*, 2017.

Jean-Michel Lasry and Pierre-Louis Lions. Jeux à champ moyen. i–le cas stationnaire. *Comptes Rendus Mathématique*, 343(9):619–625, 2006.

Shihui Li, Yi Wu, Xinyue Cui, Honghua Dong, Fei Fang, and Stuart Russell. Robust multi-agent reinforcement learning via minimax deep deterministic policy gradient. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 4213–4220, 2019.

Yingying Li, Yujie Tang, Runyu Zhang, and Na Li. Distributed reinforcement learning for decentralized linear quadratic control: A derivative-free policy optimization approach. *IEEE Transactions on Automatic Control*, 67 (12):6429–6444, 2021.

Ryan Lowe, Yi I Wu, Aviv Tamar, Jean Harb, OpenAI Pieter Abbeel, and Igor Mordatch. Multi-agent actor-critic for mixed cooperative-competitive environments. *Advances in neural information processing systems*, 30, 2017.

Dhruv Malik, Ashwin Pananjady, Kush Bhatia, Koulik Khamaru, Peter Bartlett, and Martin Wainwright. Derivative-free methods for policy optimization: Guarantees for linear quadratic systems. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2916–2925. PMLR, 2019.

Jun Morimoto and Kenji Doya. Robust reinforcement learning. *Neural computation*, 17(2):335–359, 2005.

Joshua Riley, Radu Calinescu, Colin Paterson, Daniel Kudenko, and Alec Banks. Utilising assured multi-agent reinforcement learning within safety-critical scenarios. *Procedia Computer Science*, 192:1061–1070, 2021.

Ahmad EL Sallab, Mohammed Abdou, Etienne Perot, and Senthil Yogamani. Deep reinforcement learning framework for autonomous driving. *arXiv preprint arXiv:1704.02532*, 2017.

Thomas J Sargent and Lars Ljungqvist. Recursive macroeconomic theory. *Massachusetss Institute of Technology*, 2000.

Max Simchowitz, Karan Singh, and Elad Hazan. Improper learning for non-stochastic control. In *Conference on Learning Theory*, pages 3320–3436. PMLR, 2020.

Chuangchuang Sun, Dong-Ki Kim, and Jonathan P How. Romax: Certifiably robust deep multiagent reinforcement learning via convex relaxation. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 5503–5510. IEEE, 2022.

Hamidou Tembine. Mean-field-type games. *AIMS Math*, 2(4):706–735, 2017.

Muhammad Aneeq uz Zaman, Erik Miehling, and Tamer Başar. Reinforcement learning for non-stationary discrete-time linear–quadratic mean-field games in multiple populations. *Dynamic Games and Applications*, 13(1):118–164, 2023.

Qiaomin Xie, Zhuoran Yang, Zhaoran Wang, and Andreea Minca. Learning while playing in mean-field games: Convergence and optimality. In *International Conference on Machine Learning*, pages 11436–11447. PMLR, 2021.

Bora Yongacoglu, Gürdal Arslan, and Serdar Yüksel. Independent learning and subjectivity in mean-field games. In *2022 IEEE 61st Conference on Decision and Control (CDC)*, pages 2845–2850. IEEE, 2022.

Muhammad Aneeq uz Zaman, Kaiqing Zhang, Erik Miehling, and Tamer Bașar. Reinforcement learning in non-stationary discrete-time linear-quadratic mean-field games. In *2020 59th IEEE Conference on Decision and Control (CDC)*, pages 2278–2284. IEEE, 2020.

Muhammad Aneeq Uz Zaman, Alec Koppel, Sujay Bhatt, and Tamer Başar. Oracle-free reinforcement learning in mean-field games along a single sample path. In *International Conference on Artificial Intelligence and Statistics*, pages 10178–10206. PMLR, 2023.

Huan Zhang, Hongge Chen, Chaowei Xiao, Bo Li, Mingyan Liu, Duane Boning, and Cho-Jui Hsieh. Robust deep reinforcement learning against adversarial perturbations on state observations. *Advances in Neural Information Processing Systems*, 33:21024–21037, 2020a.

Kaiqing Zhang, Tao Sun, Yunzhe Tao, Sahika Genc, Sunil Mallya, and Tamer Başar. Robust multi-agent reinforcement learning with model uncertainty. *Advances in neural information processing systems*, 33:10571–10583, 2020b.

Kaiqing Zhang, Bin Hu, and Tamer Başar. Policy optimization for H_2 linear control with H_∞ robustness guarantee: Implicit regularization and global convergence. *SIAM Journal on Control and Optimization*, 59(6):4081–4109, 2021a.

Kaiqing Zhang, Zhuoran Yang, and Tamer Başar. Multi-agent reinforcement learning: A selective overview of theories and algorithms. *Handbook of Reinforcement Learning and Control*, pages 321–384, 2021b.

Kaiqing Zhang, Xiangyuan Zhang, Bin Hu, and Tamer Başar. Derivative-free policy optimization for linear risk-sensitive and robust control design: Implicit regularization and sample complexity. *Advances in Neural Information Processing Systems*, 34:2949–2964, 2021c.

Xiangyuan Zhang and Tamer Başar. Revisiting lqr control from the perspective of receding-horizon policy gradient. *IEEE Control Systems Letters*, 2023.

Xiangyuan Zhang, Bin Hu, and Tamer Başar. Learning the kalman filter with fine-grained sample complexity. *arXiv preprint arXiv:2301.12624*, 2023.

Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*, 2019.

Supplementary Materials

Proof of Theorem 3

Proof: Central to this analysis is the quantification of the difference between the finite and infinite population costs for a given set of control policies. First we express the state and mean-field processes in terms of the noise processes, for the finite and infinite population settings. This then allows us to write the costs (in both settings) as quadratic functions of the noise process, which simplifies quantification of the difference between these two costs.

Let us first consider the finite agent setting with M number of agents. Consider the dynamics of state x^j of agent $j \in [M]$ under the NE of the MFTG (Theorem 2)

$$x^{j*,M}_{t+1} = (A_t - B_t K^{1*}_t + K^{2*}_t)x^{j*,M}_t + (\bar{A}_t - B_t(L^{1*}_t - K^{1*}_t) - \bar{B}_t L^{1*}_t + L^{2*}_t - K^{2*}_t)\bar{x}^{M*}_t + \omega^j_t + \bar{\omega}_t \tag{19}$$

where the superscript M denotes the dynamics in the finite population game (1) and $\bar{x}^{M*} = \frac{1}{M}\sum_{j\in[M]} x^{j*,M}_t$ is the empirical mean-field. We can also write the dynamics of the empirical mean-field as

$$\bar{x}^{M*}_{t+1} = \underbrace{\left(A_t + \bar{A}_t - B_t L^{1*}_t + L^{2*}_t\right)}_{\tilde{L}^*_t} \bar{x}^{M*}_t + \underbrace{\frac{1}{M}\sum_{j\in[M]} \omega^j_t + \bar{\omega}_t}_{\bar{\omega}^M_t}. \tag{20}$$

For simplicity we assume that $x^{j*,M}_0 = \omega^j_0 + \bar{\omega}_0$ which also implies that $\bar{x}^{M*}_0 = \bar{\omega}^M_0$. Using (20) we get the recursive definition of $\bar{x}^M_t$ as

$$\bar{x}^{M*}_t = \sum_{s=0}^{t} \tilde{L}^*_{[t-1,s]} \bar{\omega}^N_s, \text{ where } \tilde{L}^*_{[s,t]} := \tilde{L}^*_s \tilde{L}^*_{s-1} \ldots \tilde{L}^*_t, \text{ if } s \geq t. \text{ and } \tilde{L}^*_{[s,t]} = I \text{ otherwise.}$$

Hence $\bar{x}^{M*}_t$ can be characterized as a linear function of the noise process

$$\bar{x}^{M*}_t = \left(\bar{\Psi}^* \bar{\omega}^M\right)_t, \text{ where } \bar{\Psi}^* = \begin{pmatrix} I & 0 & 0 & \ldots \\ \tilde{L}^*_{[0,0]} & I & 0 & \ldots \\ \tilde{L}^*_{[1,0]} & \tilde{L}^*_{[1,1]} & I & \ldots \\ \tilde{L}^*_{[2,0]} & \tilde{L}^*_{[2,1]} & \tilde{L}^*_{[2,2]} & \ldots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix} \text{ and } \bar{\omega}^M = \begin{pmatrix} \bar{\omega}^M_0 \\ \bar{\omega}^M_1 \\ \bar{\omega}^M_2 \\ \vdots \\ \bar{\omega}^M_T \end{pmatrix}$$

where $(M)_t$ denotes the t^{th} block of matrix M and the covariance matrix of $\bar{\omega}^M$ is $\mathbb{E}[\bar{\omega}^M(\bar{\omega}^M)^\top] = \mathrm{diag}((\Sigma/M + \Sigma^0)_{0\leq t\leq T})$. Similarly we characterize the deviation process $x^{j*,M}_t - \bar{x}^{M*}_t$, using (19) and (20)

$$x^{j*,M}_{t+1} - \bar{x}^{M*}_{t+1} = \underbrace{(A_t - B_t K^{1*}_t + K^{2*}_t)}_{L^*_t}(x^{j*,M}_t - \bar{x}^{M*}_t) + \underbrace{\frac{M-1}{M}\omega^j_t + \frac{1}{M}\sum_{k\neq j}\omega^k_t}_{\bar{\omega}^{j,M}_t}.$$

Hence

$$\begin{aligned} x^{j*,M}_t - \bar{x}^{M*}_t &= \sum_{s=0}^{t} L^*_{[t-1,s]} \bar{\omega}^{j,M}_s \\ &= \left(\Psi^* \bar{\omega}^{j,M}\right)_t, \text{ where } \bar{\omega}^{j,M} = \frac{M-1}{M}\begin{pmatrix} \omega^j_0 \\ \vdots \\ \omega^j_{T-1} \end{pmatrix} + \frac{1}{M}\begin{pmatrix} \sum_{k\neq j}\omega^k_0 \\ \vdots \\ \sum_{k\neq j}\omega^k_{T-1} \end{pmatrix} \end{aligned}$$

where the covariance matrix of $\bar{\omega}^{j,M}$ is $\mathbb{E}[\bar{\omega}^{j,M}(\bar{\omega}^{j,M})^\top] = \operatorname{diag}(((M-1)/M \times \Sigma)_{0 \leq t \leq T})$. Similarly the infinite agent limit of this process is $x_t^* - \bar{x}_t^* = (\Psi^* \omega)_t$ where $\omega = (\omega_0^\top, \dots, \omega_{T-1}^\top)^\top$ whose covariance is $\mathbb{E}[\omega\omega^\top] = \operatorname{diag}((\Sigma)_{0 \leq t \leq T})$. Now we compute the finite agent cost in terms of the noise processes,

$$\begin{aligned}
J_M^\gamma(u^*) &= \mathbb{E}\bigg[\frac{1}{M}\sum_{j\in[M]}\sum_{t=0}^{T}\|x_t^{j*,M} - \bar{x}_t^{M*}\|_{Q_t}^2 + \|\bar{x}_t^{M*}\|_{\bar{Q}_t}^2 + \|u_t^{1,j*,M} - \bar{u}_t^{1,M*}\|^2 + \|\bar{u}_t^{1,M*}\|^2 \\
&\qquad\qquad - \gamma^2(\|u_t^{2,j*,M} - \bar{u}_t^{2,M*}\|^2 + \|\bar{u}_t^{2,M*}\|^2)\bigg] \\
&= \mathbb{E}\bigg[\frac{1}{M}\sum_{j\in[M]}\sum_{t=0}^{T}\|x_t^{j*,M} - \bar{x}_t^{M*}\|^2_{Q_t + (K_t^{1,*})^\top K_t^{1,*} - \gamma^2 (K_t^{2,*})^\top K_t^{2,*}} + \|\bar{x}_t^{M*}\|^2_{\bar{Q}_t + (L_t^{1,*})^\top L_t^{1,*} - \gamma^2 (L_t^{2,*})^\top L_t^{2,*}}\bigg] \\
&= \mathbb{E}\bigg[\frac{1}{M}\sum_{j\in[M]}(\Psi^*\bar{\omega}^{j,M})^\top\big(Q + (K^*)^\top R K^*\big)\Psi^*\bar{\omega}^{j,M} \\
&\qquad\qquad + (\bar{\Psi}^*\bar{\omega}^M)^\top\big(\bar{Q} + (\bar{K}^*)^\top \bar{R}\bar{K}^*\big)\bar{\Psi}^*\bar{\omega}^M\bigg] \\
&= \operatorname{Tr}\big((\Psi^*)^\top\big(Q + (K^*)^\top R K^*\big)\Psi^*\mathbb{E}\big[\bar{\omega}^{j,M}(\bar{\omega}^{j,M})^\top\big]\big) \\
&\qquad\qquad + \operatorname{Tr}\big((\bar{\Psi}^*)^\top\big(\bar{Q} + (\bar{K}^*)^\top \bar{R}\bar{K}^*\big)\bar{\Psi}^*\mathbb{E}\big[\bar{\omega}^M(\bar{\omega}^M)^\top\big]\big)
\end{aligned}$$

where $Q = \operatorname{diag}((Q_t)_{0\leq T})$, $R = \operatorname{diag}((\operatorname{diag}(I, -\gamma^2 I))_{0\leq T})$ and $K^* = \operatorname{diag}((K_t^{1,*}, K_t^{2,*})_{0\leq T})$ with $K_T^{1,*} = 0$ and $K_T^{2,*} = 0$. Similarly $\bar{Q} = \operatorname{diag}((\bar{Q}_t)_{0\leq T})$, $\bar{R} = \operatorname{diag}((\operatorname{diag}(I, -\gamma^2 I))_{0\leq T})$ and $\bar{K}^* = \operatorname{diag}((L_t^{1,*}, L_t^{2,*})_{0\leq T})$ with $L_T^{1,*} = 0$ and $L_T^{2,*} = 0$. Using a similar technique we can compute the infinite agent cost:

$$\begin{aligned}
J^\gamma(u^*) &= \mathbb{E}\bigg[\sum_{t=0}^{T}\|x_t^* - \bar{x}_t^*\|_{Q_t}^2 + \|\bar{x}_t^*\|_{\bar{Q}_t}^2 + \|u_t^* - \bar{u}_t^*\|^2 + \|\bar{u}_t^*\|^2\bigg] \\
&= \operatorname{Tr}\big((\Psi^*)^\top\big(Q + (K^*)^\top R K^*\big)\Psi^*\mathbb{E}\big[\omega\omega^\top\big]\big) \\
&\qquad\qquad + \operatorname{Tr}\big((\bar{\Psi}^*)^\top\big(\bar{Q} + (\bar{K}^*)^\top \bar{R}\bar{K}^*\big)\bar{\Psi}^*\mathbb{E}\big[\bar{\omega}^0(\bar{\omega}^0)^\top\big]\big).
\end{aligned}$$

Now evaluating the difference between the finite and infinite population costs:

$$\begin{aligned}
&|J_M^\gamma(u^*) - J^\gamma(u^*)| \\
&= \big|\operatorname{Tr}\big((\Psi^*)^\top\big(Q + (K^*)^\top R K^*\big)\Psi^*\big(\mathbb{E}\big[\bar{\omega}^{j,M}(\bar{\omega}^{j,M})^\top - \mathbb{E}\big[\omega\omega^\top\big]\big)\big]\big) \\
&\qquad\qquad + \operatorname{Tr}\big((\bar{\Psi}^*)^\top\big(\bar{Q} + (\bar{K}^*)^\top \bar{R}\bar{K}^*\big)\bar{\Psi}^*\big(\mathbb{E}\big[\bar{\omega}^M(\bar{\omega}^M)^\top\big] - \mathbb{E}\big[\bar{\omega}^0(\bar{\omega}^0)^\top\big]\big)\big)\big| \\
&\leq \underbrace{\Big(\|(\Psi^*)^\top\big(Q + (K^*)^\top R K^*\big)\Psi^*\|_F + \|(\bar{\Psi}^*)^\top\big(\bar{Q} + (\bar{K}^*)^\top \bar{R}\bar{K}^*\big)\bar{\Psi}^*\|_F\Big)}_{C_1^*} \\
&\qquad\qquad \operatorname{Tr}\big(\operatorname{diag}((\Sigma/M)_{0\leq t\leq T})\big) \\
&\leq C_1^*\frac{\sigma T}{M}
\end{aligned} \tag{21}$$

where $\sigma = \|\Sigma\|_F$. Now let us consider the same dynamics but under-Nash controls. The difference between finite and infinite population costs under these controls are

$$|J_M^\gamma(u) - J^\gamma(u)| \leq \underbrace{\Big(\|(\Psi)^\top\big(Q + (K)^\top R K\big)\Psi\|_F + \|(\bar{\Psi})^\top\big(\bar{Q} + (\bar{K})^\top \bar{R}\bar{K}\big)\bar{\Psi}\|_F\Big)}_{C_1}\frac{\sigma T}{M} \tag{22}$$

Using this bound along we can deduce that if,

$$\inf_{u^1} \sup_{u^2} J^\gamma(u^1, u^2) + C_1 \frac{\sigma T}{M} - \gamma^2 \mathbb{E} \sum_{t=0}^{T-1} \|\omega_t\|^2 + \|\bar{\omega}_t\|^2 \leq 0, \tag{23}$$

then the robustness condition (6) will be satisfied. Using results form Theorem 2 we can rewrite (23) into the following condition

$$\sum_{t=1}^{T} \mathrm{Tr}((M_t^\gamma - \gamma^2 I)\Sigma + (\bar{M}_t^\gamma - \gamma^2 I)\bar{\Sigma}) + \mathrm{Tr}(M_0^\gamma \Sigma^0) + \mathrm{Tr}(\bar{M}_0^\gamma \bar{\Sigma}^0) \leq -\frac{CT}{N}$$

which concludes the proof. □

Proof of Theorem 4

Proof: We start by analyzing the problem associated with the process y^i and the controller $\bar{K}_t = [(K_t^1)^\top, (K_t^2)^\top]^\top$. Using the Lyapunov equation the cost for the given set of future controllers $\big((\tilde{K}_{t+1}, \tilde{L}_{t+1}), \dots, (\tilde{K}_T, \tilde{L}_T)\big)$ can be defined in terms of symmetric matrix $\tilde{M}_{t+1}$ and covariance matrix $\tilde{\Sigma}_t^y$,

$$\min_{K_t^1} \max_{K_t^2} \tilde{J}_{y,t}^{i,\gamma}(\bar{K}_t) = \mathbb{E}_y\big[y^\top (Q_t + \bar{K}_t^\top \bar{R} \bar{K}_t) y + y^\top (A - \bar{B}_t \bar{K}_t)^\top \tilde{M}_{t+1} (A - \bar{B}_t \bar{K}_t) y\big] + \tilde{\Sigma}_t^y,$$

where $\bar{B}_t = [B_t, I]$ and $\bar{R} = \mathrm{diag}(I, -\gamma^2 I)$. The matrices $\tilde{M}_{t+1}$ and $\tilde{\Sigma}_t^y$ are fixed and can be computed offline using $\big((\tilde{K}_{t+1}, \tilde{L}_{t+1}), \dots, (\tilde{K}_T, \tilde{L}_T)\big)$ and $\mathbb{E}[\tilde{\omega}_s^i (\tilde{\omega}_s^i)^\top]$.

$$\begin{aligned} \tilde{M}_s &= Q_t + (\tilde{K}_t^1)^\top \tilde{K}_t^1 - \gamma^2 \tilde{K}_{2,t}^\top \tilde{K}_t^2 + \gamma (A_t - B_t \tilde{K}_t^1 + \tilde{K}_t^2)^T \tilde{M}_{s+1} (A_t - B_t \tilde{K}_t^1 + \tilde{K}_t^2), \\ \tilde{\Sigma}_s^y &= \mathrm{Tr}(\tilde{M}_{s+1} \mathbb{E}[\tilde{\omega}_s^i (\tilde{\omega}_s^i)^\top]) + \tilde{\Sigma}_{s+1}^y, \end{aligned}$$

for all $s \geq t+1$. For a given $t \in \{T-1, \dots, 1, 0\}$ the *exact* policy gradient of cost $\tilde{J}_{y,t}^{i,\gamma}$ with respect to K_t^1 and K_t^2 is

$$\begin{aligned} \frac{\delta \tilde{J}_{y,t}^{i,\gamma}}{\delta K_t^1} &= 2\big((I + B_t^\top \tilde{M}_{t+1} B_t) K_t^1 - B_t^\top \tilde{M}_{t+1} (A_t - K_t^2)\big) \Sigma_y, \\ -\frac{\delta \tilde{J}_{y,t}^{i,\gamma}}{\delta K_t^2} &= -2\big((-\gamma^2 I + \tilde{M}_{t+1}) K_t^2 + \tilde{M}_{t+1} (A_t - B_t K_t^1)\big) \Sigma_z \end{aligned} \tag{24}$$

First we prove smoothness and convex-concave property of the function $\tilde{J}_{y,t}^{i,\gamma}$. Due to the projection operation Proj_D the norms of K_t^1, K_t^2 are bounded by scalar D. Using Definition 2.1 in (Fallah et al., 2020) and (24), the function $\tilde{J}_{y,t}^{i,\gamma}$ is smooth with smoothness parameter

$$L = \big(\|I + B_t^\top \tilde{M}_{t+1} B_t\| + \|\tilde{M}_{t+1} - \gamma^2 I\|\big)(\|\Sigma_y\| + \|\Sigma_z\|)$$

The function $\tilde{J}_{y,t}^{i,\gamma}$ is also convex-concave since, using (24) and convexity-concavity parameters are as follows,

$$\mu_x = \sigma((I + B_t^\top \tilde{M}_{t+1} B_t)\Sigma_y) > 0 \text{ and } \mu_y = \sigma((\tilde{M}_{t+1} - \gamma^2 I)\Sigma_z) > 0 \tag{25}$$

where the second inequality is ensured when $\tilde{M}_{t+1} - \gamma^2 I > 0$. Moreover satisfying Assumption 2.3 in (Fallah et al., 2020) requires component-wise smoothness which can be satisfied with $L_x = \|I + B_t^\top \tilde{M}_{t+1} B_t\| \|\Sigma_y\|$ and $L_y = \|\tilde{M}_{t+1} - \gamma^2 I\| \|\Sigma_z\|$. Having satisfied Assumption 2.3 in (Fallah et al., 2020) now we generalize the proof of convergence for gradient descent-ascent to *biased* stochastic gradients as is the case with $\tilde{\nabla}_1 \tilde{J}_t^{i,\gamma}$ and $\tilde{\nabla}_2 \tilde{J}_t^{i,\gamma}$.

Let us first introduce a couple of notations, $\nabla_{K_{i,t}} \tilde{J}^{i,\gamma}(K_t, L_t)$ is exact gradient of cost $\tilde{J}^{i,\gamma}$ w.r.t. controller $K_{i,t}$,

$$\nabla_{K_{i,t}} \tilde{J}^{i,\gamma}(K_t, L_t) = \frac{\delta \tilde{J}^{i,\gamma}(K_t, L_t)}{\delta K_{i,t}}$$

$\tilde{\nabla}_{K_{i,t}} \tilde{J}^{i,\gamma}(K_t, L_t)$ is stochastic gradient of cost $\tilde{J}^{i,\gamma}$ w.r.t. controller $K_{i,t}$,

$$\tilde{\nabla}_{K_{i,t}} \tilde{J}^{i,\gamma}_t(K_t, L_t) = \frac{n}{Mr^2} \sum_{j=1}^{M} \tilde{J}^{i,\gamma}((K^{j,1}_t, K^2_t), L_t) e_j, \quad K^{j,1}_t = K^1_t + e_j, \quad e_j \sim \mathbb{S}^{n-1}(r),$$

$\nabla^r_{K_{i,t}} \tilde{J}^{i,\gamma}(K_t, L_t)$ is the smoothed gradient of cost $\tilde{J}^{i,\gamma}$ w.r.t. controller $K_{i,t}$,

$$\text{s.t. } \nabla^r_{K_{i,t}} \tilde{J}^{i,\gamma}(K_t, L_t) = \frac{m}{r} \mathbb{E}_e \big[\tilde{J}^{i,\gamma}((K_{i,t} + e, K_{-i,t}), L_t) e \big], \text{ where } e \sim \mathbb{S}^{m-1}(r)$$

Now we introduce some results from literature. The following result proves that the stochastic gradient is an unbiased estimator of the smoothed gradient and the bias between the smoothed gradient and the stochastic gradient is bounded by a linear function of smoothing radius r.

Lemma 6 ((Fazel et al., 2018))

$$\mathbb{E}[\tilde{\nabla}_{K_{i,t}} \tilde{J}^{i,\gamma}(K_t, L_t)] = \nabla^r_{K_{i,t}} \tilde{J}^{i,\gamma}(K_t, L_t),$$
$$\|\nabla^r_{K_{i,t}} \tilde{J}^{i,\gamma}(K_t, L_t) - \nabla_{K_{i,t}} \tilde{J}^{i,\gamma}(K_t, L_t)\|_2 \leq \phi_i r, \textit{ where } \phi_i = \|I + B_i^\top M^* B_i\|_2$$

Next we present the result that difference between the *mini-batched* stochastic gradient and the smoothed gradient can be bounded with high probability.

Lemma 7 ((Malik et al., 2019))

$$\|\tilde{\nabla}_{K_{i,t}} \tilde{J}^{i,\gamma}(K_t, L_t) - \nabla^r_{K_{i,t}} \tilde{J}^{i,\gamma}(K_t, L_t)\|_2 \leq \frac{m}{Mr} (\tilde{J}^{i,\gamma}(K_t, L_t) + \lambda r) \sqrt{\log\left(\frac{2m}{\delta}\right)}$$

with probability $1 - \delta$.

Now we compute finite sample guarantees for the Gradient Descent Ascent (GDA) update as given in Algorithm 1. For each $t \in [T]$, let us concatenate K^1_t as $\bar{K}_t = \begin{pmatrix} K^1_t \\ K^2_t \end{pmatrix}$. Let us also denote the optimal controller which optimizes $\tilde{J}^{i,\gamma}_{y,t}$ as $\bar{K}^*_t = \begin{pmatrix} K^*_{1,t} \\ K^*_{2,t} \end{pmatrix}$ such that $\tilde{J}^{i,\gamma}_{y,t}(K^*_{1,t}, K^2_t) \leq \tilde{J}^{i,\gamma}_{y,t}(K^*_{1,t}, K^*_{2,t}) \leq \tilde{J}^{i,\gamma}_{y,t}(K^1_t, K^*_{2,t})$. Since the timestep t is fixed inside the inner loop of the algorithm we discard it, instead we use the iteration index $k \in [K]$. The update rule given in Algorithm 1 is given by

$$\bar{K}_{k+1} = \bar{K}_k - \eta_k \begin{pmatrix} -\tilde{\nabla}_1 \tilde{J}^{i,\gamma}_y(\bar{K}_k) \\ \tilde{\nabla}_2 \tilde{J}^{i,\gamma}_y(\bar{K}_k) \end{pmatrix} = \bar{K}_k - \eta \tilde{\nabla} \tilde{J}^{i,\gamma}_y(\bar{K}_k) = X \bar{K}_k + Y \tilde{\nabla} \tilde{J}^{i,\gamma}_y(\bar{K}_k)$$

where $\tilde{\nabla}_1 \tilde{J}^{i,\gamma}_y(\bar{K}_k) = \tilde{\nabla}_{K^1_t} \tilde{J}^{i,\gamma}_y(\bar{K}_k), \tilde{\nabla}_2 \tilde{J}^{i,\gamma}_y(\bar{K}_k) = \tilde{\nabla}_{K^2_t} \tilde{J}^{i,\gamma}_y(\bar{K}_k), X = I$ and $Y = \eta I$. We the controller error as $\hat{K}_k = \bar{K}_k - \bar{K}^*_k$, the evolution of this error is

$$\hat{K}_{k+1} = X \hat{K}_k + Y \tilde{\nabla} \tilde{J}^{i,\gamma}(\bar{K}_k)$$

We define a Lyapunov function $V_p(\bar{K}_k) := \hat{K}_k^\top P \hat{K}_k = p\|\hat{K}_k\|_2^2$ for $P = pI$ for some $p > 0$.

$$\begin{aligned}
V_p(\bar{K}_{k+1}) - \rho^2 V_p(\bar{K}_k) = &(X\hat{K}_k + Y\tilde{\nabla}\tilde{J}^{i,\gamma}(\bar{K}_k))^\top P(X\hat{K}_k + Y\tilde{\nabla}\tilde{J}^{i,\gamma}(\bar{K}_k)) - \rho^2\hat{K}_k P\hat{K}_k, \\
\leq &(X\hat{K}_k + Y\nabla\tilde{J}^{i,\gamma}(\bar{K}_k))^\top P(X\hat{K}_k + Y\nabla\tilde{J}^{i,\gamma}(\bar{K}_k)) - \rho^2\hat{K}_k P\hat{K}_k \\
&+ (Y(\tilde{\nabla}\tilde{J}^{i,\gamma}(\bar{K}_k) - \nabla\tilde{J}^{i,\gamma}(\bar{K}_k)))^\top PX\hat{K}_k + (X\hat{K}_k)^\top P(Y(\tilde{\nabla}\tilde{J}^{i,\gamma}(\bar{K}_k) - \nabla\tilde{J}^{i,\gamma}(\bar{K}_k))) \\
&+ (\tilde{\nabla}\tilde{J}^{i,\gamma}(\bar{K}_k) - \nabla\tilde{J}^{i,\gamma}(\bar{K}_k))^\top Y^\top PY(\tilde{\nabla}\tilde{J}^{i,\gamma}(\bar{K}_k) - \nabla\tilde{J}^{i,\gamma}(\bar{K}_k)), \\
\leq &\begin{pmatrix} \hat{K}_k \\ \nabla\tilde{J}^{i,\gamma}(\bar{K}_k) \end{pmatrix}^\top \begin{bmatrix} X^T PX - \rho^2 P & X^\top PY \\ Y^\top PX & Y^\top PY \end{bmatrix} \begin{pmatrix} \hat{K}_k \\ \nabla\tilde{J}^{i,\gamma}(\bar{K}_k) \end{pmatrix} + \eta^2 p\|\tilde{\nabla}\tilde{J}^{i,\gamma}(\bar{K}_k) - \nabla\tilde{J}^{i,\gamma}\|_2^2 \\
&+ \begin{pmatrix} \hat{K}_k \\ \tilde{\nabla}\tilde{J}^{i,\gamma}(\bar{K}_k) - \nabla\tilde{J}^{i,\gamma}(\bar{K}_k) \end{pmatrix}^\top \begin{bmatrix} 0 & X^\top PY \\ Y^\top PX & 0 \end{bmatrix} \begin{pmatrix} \hat{K}_k \\ \tilde{\nabla}\tilde{J}^{i,\gamma}(\bar{K}_k) - \nabla\tilde{J}^{i,\gamma}(\bar{K}_k) \end{pmatrix}.
\end{aligned}$$

Let us enforce $\rho^2 = 1 - m\eta$, then we know due to (Fallah et al., 2020) that

$$\begin{pmatrix} \hat{K}_k \\ \nabla\tilde{J}^{i,\gamma}(\bar{K}_k) \end{pmatrix}^\top \begin{bmatrix} X^T PX - \rho^2 P & X^\top PY \\ Y^\top PX & Y^\top PY \end{bmatrix} \begin{pmatrix} \hat{K}_k \\ \nabla\tilde{J}^{i,\gamma}(\bar{K}_k) \end{pmatrix} \leq 0$$

$$\begin{aligned}
\hat{K}_{k+1}^\top \hat{K}_{k+1} - (1 - m\eta)\hat{K}_k^\top \hat{K}_k &\leq \eta\|\tilde{\nabla}\tilde{J}^{i,\gamma}(\bar{K}_k) - \nabla\tilde{J}^{i,\gamma}(\bar{K}_k)\|_2\|\hat{K}_k\|_2 + \eta^2\|\tilde{\nabla}\tilde{J}^{i,\gamma}(\bar{K}_k) - \nabla\tilde{J}^{i,\gamma}(\bar{K}_k)\|_2^2, \\
&= \sqrt{m\eta}\|\hat{K}_k\|_2 \frac{\sqrt{\eta}\|\tilde{\nabla}\tilde{J}^{i,\gamma}(\bar{K}_k) - \nabla\tilde{J}^{i,\gamma}(\bar{K}_k)\|_2}{\sqrt{m}} + \eta^2\|\tilde{\nabla}\tilde{J}^{i,\gamma}(\bar{K}_k) - \nabla\tilde{J}^{i,\gamma}(\bar{K}_k)\|_2^2
\end{aligned}$$

Using the fact $2ab \leq a^2 + b^2$ and for $\eta < 1$ we get

$$\begin{aligned}
\|\hat{K}_{k+1}\|_2^2 &\leq (1 - m\eta)\|\hat{K}_k\|_2^2 + \frac{m\eta}{2}\|\hat{K}_k\|_2 + \eta\Big(\frac{1}{m} + 1\Big)\|\tilde{\nabla}\tilde{J}^{i,\gamma}(\bar{K}_k) - \nabla\tilde{J}^{i,\gamma}(\bar{K}_k)\|_2^2, \\
&\leq \left(1 - \frac{m\eta}{2}\right)\|\hat{K}_k\|_2^2 + \eta\left(\frac{1}{m} + 1\right)\|\tilde{\nabla}\tilde{J}^{i,\gamma}(\bar{K}_k) - \nabla\tilde{J}^{i,\gamma}(\bar{K}_k)\|_2^2,
\end{aligned}$$

Hence

$$\begin{aligned}
\|\hat{K}_k\|_2^2 &\leq \left(1 - \frac{m\eta}{2}\right)^k \|\hat{K}_0\|_2^2 + \eta\left(\frac{1}{m} + 1\right)\sum_{j=0}^{k}\left(1 - \frac{m\eta}{2}\right)^j \|\tilde{\nabla}\tilde{J}^{i,\gamma}(\bar{K}_j) - \nabla\tilde{J}^{i,\gamma}(\bar{K}_j)\|_2^2, \\
&\leq \left(1 - \frac{m\eta}{2}\right)^k \|\hat{K}_0\|_2^2 + \eta\left(\frac{1}{m} + 1\right)\sum_{j=0}^{\infty}\left(1 - \frac{m\eta}{2}\right)^j \|\tilde{\nabla}\tilde{J}^{i,\gamma}(\bar{K}_j) - \nabla\tilde{J}^{i,\gamma}(\bar{K}_j)\|_2^2, \\
&\leq \left(1 - \frac{m\eta}{2}\right)^k \|\hat{K}_0\|_2^2 + \frac{2}{m}\left(\frac{1}{m} + 1\right)\left(\frac{m}{Mr}(\tilde{J}^{i,\gamma}(K_t, L_t) + \lambda r)\sqrt{\log\left(\frac{2m}{\delta}\right)} + \phi_i r\right),
\end{aligned}$$

If $k = \mathcal{O}(\log(1/\epsilon))$, $M = \mathcal{O}(1/\epsilon)$ and $r = \Omega(\epsilon)$ then $\|\hat{K}_k\|_2^2 \leq \epsilon$. □

Proof of Theorem 5

Proof: First we define a change of notation to keep the analysis concise. Let

$$B_t^1 = B_t,\ \ B_t^2 = I,\ \ Q_t^1 = Q_t,\ \ Q_t^2 = -Q_t,\ \ R_t^1 = I,\ \ R_t^2 = -\gamma^2 I,\ \ M_t^1 = M_t,\ \ M_t^2 = -M_t$$

where the sequence $(M_t)_{\forall t}$ will be defined in the following analysis. Throughout the proof we refer to the output of the inner loop of Algorithm 1 as the set of *output controllers* $(K_t^i)_{i\in[2],t\in\{0,\dots,T-1\}}$. In the proof we use two other sets of controllers as well. The first set $(K_t^{i*})_{i\in[2],t\in\{0,\dots,T-1\}}$ which denotes the NE as characterized in Theorem

2. The second set is called the *local*-NE (as in proof of Theorem 4) and is denoted by $(\tilde{K}^{i*}_t)_{i\in[2],t\in\{0,\dots,T-1\}}$. The proof quantifies the error between the output controllers $(K^i_t)_{i\in[2]}$ and the corresponding NE controllers $(K^{i*}_t)_{i\in[2]}$ by utilizing the intermediate local-NE controllers $(\tilde{K}^{i*}_t)_{i\in[2]}$ for each time $t \in \{T-1,\dots,0\}$. For each t the error is shown to depend on error in future controllers $(K^i_s)_{s\geq t, i\in[2]}$ and the approximation error Δ_t introduced by the gradient descent-ascent update. If $\Delta_t = \mathcal{O}(\epsilon)$, then the error between the output and NE controllers is shown to be $\mathcal{O}(\epsilon)$.

Let us start by denoting the NE value function matrices for agent $i \in [2]$ at time $t \in \{0,1,\dots,T\}$, under the NE control matrices $(K^{i*}_s)_{i\in[2],s\in\{t+1,\dots,T-1\}}$ by M^{i*}_t. Using (24) the NE control matrices can be characterized as:

$$\begin{aligned} K^{i*}_t &:= \operatorname*{argmin}_{K^i_t} \tilde{J}^i_{y,t}((K^i_t, K^{i*}_{t+1},\dots,K^{i*}_{T-1}), K^{-i*}) \\ &= (R^i_t + (B^i_t)^\top M^{i*}_{t+1} B^i_t)^{-1} (B^i_t)^\top M^{i*}_{t+1} A^{i*}_t \end{aligned} \tag{26}$$

where $A^{i*}_t := A_t + \sum_{j\neq i} B^j_t K^{j*}_t$. The NE value function matrices are be defined recursively using the Lyapunov equation and NE controllers as

$$\begin{aligned} M^{i*}_t &= Q^i_t + (A^{i*}_t)^\top M^{i*}_{t+1} A^{i*}_t - (A^{i*}_t)^\top M^{i*}_{t+1} B^i_t (R^i_t + (B^i_t)^\top M^{i*}_{t+1} B^i_t)^{-1} (B^i_t)^\top M^{i*}_{t+1} A^{i*}_t \\ &= Q^i_t + (A^{i*}_t)^\top M^{i*}_{t+1} (A^{i*}_t + B^i_t K^{i*}_t), \\ M^{i*}_T &= Q^i_T \end{aligned} \tag{27}$$

The sufficient condition for existence and uniqueness of the set of matrices K^{i*}_t and M^{i*}_t is shown in Theorem 2. Let us now define the perturbed values matrices M^i_t (resulting from the control matrices $(K^i_s)_{i\in[2],s\in\{t+1,\dots,T-1\}}$). At time t let us assume the existence and uniqueness of the *target* control matrices $(\tilde{K}^{i*}_t)_{i\in[2]}$ such that

$$\begin{aligned} \tilde{K}^{i*}_t &:= \operatorname*{argmin}_{K^i_t} \tilde{J}^i_{y,t}(K^i, \tilde{K}^{-i*}) \\ &= (R^i_t + (B^i_t)^\top M^i_{t+1} B^i_t)^{-1} (B^i_t)^\top M^i_{t+1} \tilde{A}^i_t \end{aligned} \tag{28}$$

where $\tilde{K}^{j*} = (\tilde{K}^{j*}_t, K^j_{t+1},\dots,K^j_{T-1})$ and $\tilde{A}^i_t := A_t + \sum_{j\neq i} B^j_t \tilde{K}^{j*}_t$. This is obtained using (24). We will show the existence and uniqueness of the target control matrices given that the inner-loop convergence in Algorithm 1 (Theorem 4) is close enough. Using these matrices we define the value function matrices $(\tilde{M}^i_t)_{i\in[2]}$ using the Lyapunov equations as follows

$$\begin{aligned} \tilde{M}^i_t &= Q^i_t + (\tilde{K}^{i*}_t)^\top R^i_t \tilde{K}^{i*}_t + \left(A_t + \sum_{j=1}^N B^j_t \tilde{K}^{i*}_t\right)^\top M^i_{t+1} \left(A_t + \sum_{j=1}^2 B^i_t \tilde{K}^{j*}_t\right) \\ &= Q^i_t + (\tilde{A}^i_t)^\top M^i_{t+1} \tilde{A}^i_t - (\tilde{A}^i_t)^\top M^i_{t+1} B^i_t (R^i_t + (B^i_t)^\top M^i_{t+1} B^i_t)^{-1} (B^i_t)^\top M^i_{t+1} \tilde{A}^i_t \\ &= Q^i_t + (\tilde{A}^i_t)^\top M^i_{t+1} (\tilde{A}^i_t + B^i_t \tilde{K}^{i*}_t) \\ \tilde{M}^i_T &= Q^i_T \end{aligned} \tag{29}$$

Finally we define the perturbed value function matrices M^i_t which result from the perturbed matrices $(K^i_s)_{i\in[2],s\in\{t+1,\dots,T-1\}}$ obtained using the gradient descent-ascent in Algorithm 1:

$$M^i_t = Q^i_t + (K^i_t)^\top R^i_t K^i_t + \left(A_t + \sum_{j=1}^2 B^j_t K^j_t\right)^\top M^i_{t+1} \left(A_t + \sum_{j=1}^2 B^j_t K^j_t\right) \tag{30}$$

Throughout this proof we assume that the output of the inner loop in Algorithm 1, also called the *output matrices* K^i_t, are Δ_t away from the target matrices $\tilde{K}^{i*}_t$, such that $\|K^i_t - \tilde{K}^{i*}_t\| \leq \Delta_t$. We know that,

$$\|K^i_t - K^{i*}_t\| \leq \|K^i_t - \tilde{K}^{i*}_t\| + \|\tilde{K}^{i*}_t - K^{i*}_t\| \leq \Delta_t + \|\tilde{K}^{i*}_t - K^{i*}_t\|.$$

Now we obtain an upper bound on the term $\|\tilde{K}_t^{i*} - K_t^{i*}\|$ using (26) and (28):

$$
\begin{aligned}
& K_t^{i*} - \tilde{K}_t^{i*} \\
& = -(R_t^i + (B_t^i)^\top M_{t+1}^i B_t^i)^{-1} (B_t^i)^\top M_{t+1}^i \tilde{A}_t^i + (R_t^i + (B_t^i)^\top M_{t+1}^{i*} B_t^i)^{-1} (B_t^i)^\top M_{t+1}^{i*} A_t^{i*} \\
& = -(R_t^i + (B_t^i)^\top M_{t+1}^i B_t^i)^{-1} (B_t^i)^\top M_{t+1}^i (\tilde{A}_t^i - A_t^{i*}) \\
& \qquad\qquad - (R_t^i + (B_t^i)^\top M_{t+1}^i B_t^i)^{-1} (B_t^i)^\top (M_{t+1}^i - M_{t+1}^{i*}) A_t^{i*} \\
& \qquad\qquad - \big((R_t^i + (B_t^i)^\top M_{t+1}^i B_t^i)^{-1} - (R_t^i + (B_t^i)^\top M_{t+1}^{i*} B_t^i)^{-1} \big) (B_t^i)^\top M_{t+1}^{i*} A_t^{i*} \\
& = -(R_t^i + (B_t^i)^\top M_{t+1}^i B_t^i)^{-1} (B_t^i)^\top M_{t+1}^i \sum_{j \neq i} B_t^j (\tilde{K}_t^{j*} - K_t^{j*}) \\
& \qquad\qquad - (R_t^i + (B_t^i)^\top M_{t+1}^i B_t^i)^{-1} (B_t^i)^\top (M_{t+1}^i - M_{t+1}^{i*}) A_t^{i*} \\
& \qquad\qquad - \big((R_t^i + (B_t^i)^\top M_{t+1}^i B_t^i)^{-1} - (R_t^i + (B_t^i)^\top M_{t+1}^{i*} B_t^i)^{-1} \big) (B_t^i)^\top M_{t+1}^{i*} A_t^{i*} \\
& = -(R_t^i + (B_t^i)^\top M_{t+1}^i B_t^i)^{-1} (B_t^i)^\top M_{t+1}^i \sum_{j \neq i} B_t^j (\tilde{K}_t^{j*} - K_t^{j*}) \\
& \qquad\qquad - (R_t^i + (B_t^i)^\top M_{t+1}^i B_t^i)^{-1} (B_t^i)^\top (M_{t+1}^i - M_{t+1}^{i*}) A_t^{i*} \\
& \quad + (R_t^i + (B_t^i)^\top M_{t+1}^i B_t^i)^{-1} (B_t^i)^\top \big(M_{t+1}^i - M_{t+1}^{i*} \big) \\
& \qquad\qquad\qquad\qquad B_t^i (R_t^i + (B_t^i)^\top M_{t+1}^{i*} B_t^i)^{-1} (B_t^i)^{-1} (B_t^i)^\top M_{t+1}^{i*} A_t^{i*}
\end{aligned}
\tag{31}
$$

where the last equality is due to

$$
\begin{aligned}
& - (R_t^i + (B_t^i)^\top M_{t+1}^i B_t^i)^{-1} + (R_t^i + (B_t^i)^\top M_{t+1}^{i*} B_t^i)^{-1} \\
& = -(R_t^i + (B_t^i)^\top M_{t+1}^i B_t^i)^{-1} (R_t^i + (B_t^i)^\top M_{t+1}^{i*} B_t^i) (R_t^i + (B_t^i)^\top M_{t+1}^{i*} B_t^i)^{-1} \\
& \qquad\qquad + (R_t^i + (B_t^i)^\top M_{t+1}^i B_t^i)^{-1} (R_t^i + (B_t^i)^\top M_{t+1}^i B_t^i) (R_t^i + (B_t^i)^\top M_{t+1}^{i*} B_t^i)^{-1} \\
& = (R_t^i + (B_t^i)^\top M_{t+1}^i B_t^i)^{-1} (B_t^i)^\top \big(M_{t+1}^i - M_{t+1}^{i*} \big) B_t^i (R_t^i + (B_t^i)^\top M_{t+1}^{i*} B_t^i)^{-1} (B_t^i)^{-1}
\end{aligned}
$$

Further analyzing (31) we get

$$
\begin{aligned}
& (R_t^i + (B_t^i)^\top M_{t+1}^i B_t^i)(\tilde{K}_t^{i*} - K_t^{i*}) + (B_t^i)^\top M_{t+1}^i \sum_{j \neq i} B_t^j (\tilde{K}_t^{j*} - K_t^{j*}) \\
& \qquad\qquad\qquad\qquad\qquad\qquad = -(B_t^i)^\top (M_{t+1}^i - M_{t+1}^{i*})(A_t^{i*} + B_t^i K_t^{i*}) \\
& \underbrace{\begin{pmatrix} R_t^1 + (B_t^1)^\top M_{t+1}^1 B_t^1 & (B_t^1)^\top M_{t+1}^1 B_t^2 \\ (B_t^2)^\top M_{t+1}^2 B_t^1 & R_t^2 + (B_t^2)^\top M_{t+1}^2 B_t^2 \end{pmatrix}}_{\tilde{\Phi}_{t+1}} \begin{pmatrix} \tilde{K}_t^{1*} - K_t^{1*} \\ \tilde{K}_t^{2*} - K_t^{2*} \end{pmatrix} \\
& \qquad\qquad\qquad = \begin{pmatrix} (B_t^1)^\top (M_{t+1}^{1*} - M_{t+1}^1)(A_t^{1*} + B_t^1 K_t^{1*}) \\ (B_t^2)^\top (M_{t+1}^{2*} - M_{t+1}^2)(A_t^{2*} + B_t^2 K_t^{2*}) \end{pmatrix}
\end{aligned}
\tag{32}
$$

where $\tilde{\Phi}_{t+1}^{-1}$ is guaranteed to exist as shown below. Using (32) we now obtain an upper bound on $\max_{i \in [2]} \|\tilde{K}_t^{i*} - K_t^{i*}\|$:

$$
\max_{j \in [2]} \|\tilde{K}_t^{i*} - K_t^{i*}\| \leq \|\tilde{\Phi}_{t+1}^{-1}\| \|A_t^*\| \max_{j \in [2]} \|B_t^j\|_\infty \|M_{t+1}^{j*} - M_{t+1}^j\|_\infty \tag{33}
$$

We also define

$$
\hat{\Phi}_{t+1} := \tilde{\Phi}_{t+1} - \Phi_{t+1}^* = \begin{pmatrix} (B_t^1)^\top (M_{t+1}^1 - M_{t+1}^{1*}) \\ (B_t^2)^\top (M_{t+1}^2 - M_{t+1}^{2*}) \end{pmatrix} \begin{pmatrix} B_t^1 & B_t^2 \end{pmatrix}
$$

Now we characterize $\|\tilde{\Phi}_{t+1}^{-1}\|$:

$$\begin{aligned}\|\tilde{\Phi}_{t+1}^{-1}\| &= \|(I + (\Phi_{t+1}^*)^{-1}\hat{\Phi}_{t+1})^{-1}(\Phi_{t+1}^*)^{-1}\| \\ &\leq \|(\Phi_{t+1}^*)^{-1}\| \sum_{k=0}^{\infty} \|\|(\Phi_{t+1}^*)^{-1}\|^k \big(\max_{j\in[2]} \|B_t^j\|^2 \|M_{t+1}^i - M_{t+1}^{j*}\|\big)^k \\ &\leq 2\|(\Phi_{t+1}^*)^{-1}\| \leq 2c_\Phi^{(-1)} \end{aligned} \tag{34}$$

where the last inequality is possible due to the fact $\max_{j\in[2]}\|M_{t+1}^j - M_{t+1}^{j*}\| \leq 1/(2c_\Phi^{(-1)}c_B^2)$ where $c_\Phi^{(-1)} = \max_{t\in\{0,\ldots,T-1\}}\|(\Phi_{t+1}^*)^{-1}\|$ and $c_B = \max_{i\in[2],t\in\{0,\ldots,T-1\}}\|B_t^i\|$. Hence combining (33)-(34),

$$\max_{j\in[2]}\|\tilde{K}_t^{i*} - K_t^{i*}\| \leq \bar{c}_1 \max_{j\in[2]}\|M_{t+1}^{j*} - M_{t+1}^j\| \tag{35}$$

where $\bar{c}_1 := 2c_\Phi^{(-1)}c_B c_A^*$ and $c_A^* = \|A_t^*\|$. Now we characterize the difference $M_t^i - M_t^{i*} = M_t^i - \tilde{M}_t^i + \tilde{M}_t^i - M_t^{i*}$. First we can characterize $\tilde{M}_t^i - M_t^{i*}$ using (27) and (29):

$$\begin{aligned} \tilde{M}_t^i - M_t^{i*} &= (\tilde{A}_t^i)^\top M_{t+1}^i(\tilde{A}_t^i + B_t^i\tilde{K}_t^{i*}) - (A_t^{i*})^\top M_{t+1}^{i*}(A_t^{i*} + B_t^iK_t^{i*}) \\ &= (\tilde{A}_t^i - A_t^{i*})^\top M_{t+1}^i(\tilde{A}_t^i + B_t^i\tilde{K}_t^{i*}) + (A_t^{i*})^\top(M_{t+1}^i - M_{t+1}^{i*})(\tilde{A}_t^i + B_t^i\tilde{K}_t^{i*}) \\ &\qquad + (A_t^{i*})^\top M_{t+1}^{i*}(\tilde{A}_t^i + B_t^i\tilde{K}_t^{i*} - A_t^{i*} - B_t^iK_t^{i*}) \\ &= \left(\sum_{j\neq i} B_t^j(\tilde{K}_t^{j*} - K_t^{j*})\right)^\top M_{t+1}^i(\tilde{A}_t^i + B_t^i\tilde{K}_t^{i*}) + (A_t^{i*})^\top(M_{t+1}^i - M_{t+1}^{i*})(\tilde{A}_t^i + B_t^i\tilde{K}_t^{i*}) \\ &\qquad + (A_t^{i*})^\top M_{t+1}^{i*}\sum_{j=1}^{2} B_t^j(\tilde{K}_t^{j*} - K_t^{j*}) \end{aligned} \tag{36}$$

Using this characterization we bound $\|\tilde{M}_t^i - M_t^{i*}\|$ using the AM-GM inequality

$$\begin{aligned} &\|\tilde{M}_t^i - M_t^{i*}\| \\ &\leq 2\|A_t^*\|\|M_{t+1}^{i*}\|c_B\sum_{j=1}^{2}\|\tilde{K}_t^{j*} - K_t^{j*}\| \\ &\quad + \big(\|A_t^*\|/2 + \|M_{t+1}^{i*}\|c_B + \|A_t^{i*}\|/2\big)c_B\left(\sum_{j=1}^{2}\|\tilde{K}_t^{j*} - K_t^{j*}\|\right)^2 + \frac{c_B^4}{2}\left(\sum_{j=1}^{2}\|\tilde{K}_t^{j*} - K_t^{j*}\|\right)^4 \\ &\quad + \big(\|A_t^*\|/2 + 1/2 + \|A_t^{i*}\|/2\big)\|M_{t+1}^i - M_{t+1}^{i*}\|^2 + \|A_t^{i*}\|\|A_t^*\|\|M_{t+1}^i - M_{t+1}^{i*}\| \\ &\leq \underbrace{\big(2c_A^*c_P^*c_B + (c_A^*/2 + c_P^*c_B + c_A^{i*}/2)c_B + c_B^4/2\big)}_{\bar{c}_2}\sum_{j=1}^{2}\|\tilde{K}_t^{j*} - K_t^{j*}\| \\ &\qquad\qquad + \underbrace{c_A^*/2 + 1/2 + c_A^{i*}/2 + c_A^{i*}c_A^*}_{\bar{c}_3}\|M_{t+1}^i - M_{t+1}^{i*}\| \\ &\leq \bar{c}_2 N\max_{j\in[2]}\|\tilde{K}_t^{j*} - K_t^{j*}\| + \bar{c}_3\|M_{t+1}^i - M_{t+1}^{i*}\| \end{aligned} \tag{37}$$

where $c_A^{i*} := \max_{t\in\{0,\ldots,T-1\}}\|A_t^{i*}\|, c_P^* := \max_{i\in[2],t\in\{0,\ldots,T-1\}}\|M_{t+1}^{i*}\|$, and the last inequality is possible due to the fact that $\|M_{t+1}^i - M_{t+1}^{i*}\|, \|\tilde{K}_t^{j*} - K_t^{j*}\| \leq 1/N$. Similarly $M_t^i - \tilde{M}_t^i$ can be decomposed using (29) and (30):

$$\begin{aligned} M_t^i - \tilde{M}_t^i = (K_t^i)^\top R_t^i K_t^i + \left(A_t + \sum_{j=1}^{2} B_t^jK_t^j\right)^\top M_{t+1}^i\left(A_t + \sum_{j=1}^{2} B_t^jK_t^j\right) \\ - \left[(\tilde{K}_t^{i*})^\top R_t^i\tilde{K}_t^{i*} + \left(A_t + \sum_{j=1}^{2} B_t^j\tilde{K}_t^{i*}\right)^\top M_{t+1}^i\left(A_t + \sum_{j=1}^{2} B_t^i\tilde{K}_t^{j*}\right)\right]. \end{aligned}$$

We start by analyzing the quadratic form $x^\top M_t^i x$:

$$\begin{aligned}
x^\top M_t^i x &= x^\top \Bigg[Q_t^i + (K_t^i)^\top R_t^i K_t^i + \Big(A_t + \sum_{j=1}^{2} B_t^j K_t^j \Big)^\top M_{t+1}^i \Big(A_t + \sum_{j=1}^{2} B_t^j K_t^j \Big) \Bigg] x \\
&= x^\top \Bigg[(K_t^i)^\top \big(R_t^i + (B_t^i)^\top M_{t+1}^i B_t^i \big) K_t^i + 2 (B_t^i K_t^i)^\top M_{t+1}^i \Big(A_t + \sum_{j \neq i} B_t^j K_t^j \Big) + Q_t^i \\
&\qquad + \Big(A_t + \sum_{j \neq i} B_t^j K_t^j \Big)^\top M_{t+1}^i \Big(A_t + \sum_{j \neq i} B_t^j K_t^j \Big) \Bigg] x \\
&= x^\top \Bigg[(K_t^i)^\top \big(R_t^i + (B_t^i)^\top M_{t+1}^i B_t^i \big) K_t^i + 2 (B_t^i K_t^i)^\top M_{t+1}^i \Big(A_t + \sum_{j \neq i} B_t^j \tilde{K}_t^{j*} \Big) + Q_t^i \\
&\quad + 2 (B_t^i K_t^i)^\top M_{t+1}^i \sum_{j \neq i} B_t^j (K_t^j - \tilde{K}_t^{j*}) + \Big(A_t + \sum_{j \neq i} B_t^j \tilde{K}_t^{j*} \Big)^\top M_{t+1}^i \Big(A_t + \sum_{j \neq i} B_t^j \tilde{K}_t^{j*} \Big) \\
&\quad + \Big(\sum_{j \neq i} B_t^j (K_t^j - \tilde{K}_t^{j*}) \Big)^\top M_{t+1}^i \Big(\sum_{j \neq i} B_t^j (K_t^j - \tilde{K}_t^{j*}) \Big) \\
&\quad + 2 \Big(A_t + \sum_{j \neq i} B_t^j \tilde{K}_t^{j*} \Big) M_{t+1}^i \Big(\sum_{j \neq i} B_t^j (K_t^j - \tilde{K}_t^{j*}) \Big) \Bigg] x
\end{aligned}$$

Completing squares we get

$$\begin{aligned}
& x^\top M_t^i x \qquad (38) \\
&= x^\top \Bigg[(K_t^i - \tilde{K}_t^{i*})^\top \big(R_t^i + (B_t^i)^\top M_{t+1}^i B_t^i \big) (K_t^i - \tilde{K}_t^{i*}) \\
&\quad + \Big(A_t + \sum_{j \neq i} B_t^j \tilde{K}_t^{j*} \Big)^\top M_{t+1}^i \Big(A_t + \sum_{j \neq i} B_t^j \tilde{K}_t^{j*} \Big) + Q_t^i \\
&\quad - \Big(A_t + \sum_{j \neq i} B_t^j \tilde{K}_t^{j*} \Big)^\top M_{t+1}^i B_t^i \big(R_t^i + (B_t^i)^\top M_{t+1}^i B_t^i \big)^{-1} (B_t^i)^\top M_{t+1}^i \Big(A_t + \sum_{j \neq i} B_t^j \tilde{K}_t^{j*} \Big) \\
&\quad + \Big(2 \Big(A_t + \sum_{j=1}^{2} B_t^j \tilde{K}_t^{j*} \Big) + \sum_{j=1}^{2} B_t^j (K_t^j - \tilde{K}_t^{j*}) \Big)^\top M_{t+1}^i \Big(\sum_{j \neq i} B_t^j (K_t^j - \tilde{K}_t^{j*}) \Big) \Bigg] x.
\end{aligned}$$

Now we take a look at the quadratic $x^\top \tilde{M}_t^i x$:

$$\begin{aligned}
&x^\top \tilde{M}_t^i x \\
&= x^\top \Bigg[Q_t^i + (\tilde{K}_t^{i*})^\top R_t^i \tilde{K}_t^{i*} + \Bigg(A_t + \sum_{j=1}^{2} B_t^j \tilde{K}_t^{j*} \Bigg)^\top M_{t+1}^i \Bigg(A_t + \sum_{j=1}^{2} B_t^j \tilde{K}_t^{j*} \Bigg) \Bigg] x \\
&= x^\top \Bigg[(\tilde{K}_t^{i*})^\top \big(R_t^i + (B_t^i)^\top M_{t+1}^i B_t^i \big) \tilde{K}_t^{i*} + 2 (B_t^i \tilde{K}_t^{i*})^\top M_{t+1}^i \Bigg(A_t + \sum_{j \neq i} B_t^j \tilde{K}_t^{j*} \Bigg) + Q_t^i \\
&\quad + \Bigg(A_t + \sum_{j \neq i} B_t^j \tilde{K}_t^{j*} \Bigg)^\top M_{t+1}^i \Bigg(A_t + \sum_{j \neq i}^{2} B_t^j \tilde{K}_t^{j*} \Bigg) \Bigg] x \\
&= x^\top \Bigg[\Bigg(A_t + \sum_{j \neq i} B_t^j \tilde{K}_t^{j*} \Bigg)^\top M_{t+1}^i \Bigg(A_t + \sum_{j \neq i}^{2} B_t^j \tilde{K}_t^{j*} \Bigg) + Q_t^i \\
&\quad - \Bigg(A_t + \sum_{j \neq i} B_t^j \tilde{K}_t^{j*} \Bigg)^\top M_{t+1}^i B_t^i \big(R_t^i + (B_t^i)^\top M_{t+1}^i B_t^i \big)^{-1} (B_t^i)^\top M_{t+1}^i \Bigg(A_t + \sum_{j \neq i} B_t^j \tilde{K}_t^{j*} \Bigg) \Bigg] x.
\end{aligned} \tag{39}$$

Using (38) and (39), we get

$$\begin{aligned}
&x^\top (M_t^i - \tilde{M}_t^i) x \\
&= x^\top \Bigg[(K_t^i - \tilde{K}_t^{i*})^\top \big(R_t^i + (B_t^i)^\top M_{t+1}^i B_t^i \big) (K_t^i - \tilde{K}_t^{i*}) \\
&\quad + \Bigg(2 \Bigg(A_t + \sum_{j=1}^{2} B_t^j \tilde{K}_t^{j*} \Bigg) + \sum_{j=1}^{2} B_t^j (K_t^j - \tilde{K}_t^{j*}) \Bigg)^\top M_{t+1}^i \Bigg(\sum_{j \neq i} B_t^j (K_t^j - \tilde{K}_t^{j*}) \Bigg) \Bigg] x \\
&= x^\top \Bigg[(K_t^i - \tilde{K}_t^{i*})^\top \big(R_t^i + (B_t^i)^\top M_{t+1}^i B_t^i \big) (K_t^i - \tilde{K}_t^{i*}) \\
&\quad + \Bigg(2 \Bigg(A_t + \sum_{j=1}^{2} B_t^j K_t^{j*} \Bigg) + 2 \sum_{j=1}^{2} B_t^j (\tilde{K}_t^{j*} - K_t^{j*}) + \sum_{j=1}^{2} B_t^j (K_t^j - \tilde{K}_t^{j*}) \Bigg)^\top \\
&\qquad\qquad M_{t+1}^i \Bigg(\sum_{j \neq i} B_t^j (K_t^j - \tilde{K}_t^{j*}) \Bigg) \Bigg] x
\end{aligned}$$

Using this characterization we bound $\|M_t^i - \tilde{M}_t^i\|$:

$$\begin{aligned}
&\|M_t^i - \tilde{M}_t^i\| \\
&\leq \big(\|R_t^i + (B_t^i)^\top M_{t+1}^{i*} B_t^i\| + \|(B_t^i)^\top \big(M_{t+1}^{i*} - M_{t+1}^{i*} \big) B_t^i\| \big) \|K_t^i - \tilde{K}_t^{i*}\| \\
&\quad + (2c_A^{i*} + 2c_B \sum_{j=1}^{2} \big(\|\tilde{K}_t^{j*} - K_t^*\| + \|K_t^j - \tilde{K}_t^{j*}\| \big) \big(\|M_{t+1}^{i*}\| + \|M_{t+1}^i - M_{t+1}^{i*}\| \big) \\
&\qquad\qquad c_B \sum_{j=1}^{2} \|K_t^j - \tilde{K}_t^{j*}\|
\end{aligned}$$

As before assuming $\|M^i_{t+1} - M^{i*}_{t+1}\|, \|\tilde{K}^{j*}_t - K^{j*}_t\| \le 1/N$,

$$\begin{aligned} \|M^i_t - \tilde{M}^i_t\| &\le \underbrace{\left(\bar{c}^* + c_B^2/2 + 2c_B^2(c_P^* + 1) + 2c_A^{i*}\right)}_{\bar{c}_4} \sum_{j=1}^{2} \|K^j_t - \tilde{K}^{j*}_t\| \\ &\qquad + \underbrace{\left(c_B^2/2 + c_A^{i*} + c_B\right)}_{\bar{c}_5} \|M^i_{t+1} - M^{i*}_{t+1}\| + \underbrace{c_B(c_P^* + 1)}_{\bar{c}_6} \sum_{j=1}^{2} \|\tilde{K}^{j*}_t - K^{j*}_t\| \\ &\le \bar{c}_4 N \Delta_t + \bar{c}_5 \|M^i_{t+1} - M^{i*}_{t+1}\| + \bar{c}_6 \sum_{j=1}^{2} \|\tilde{K}^{j*}_t - K^{j*}_t\| \end{aligned} \tag{40}$$

where $\bar{c}^* := \max_{i \in [2], t \in \{0, \dots, T-1\}} \|R^i_t + (B^i_t)^\top M^{i*}_{t+1} B^i_t\|$. Let us define $e^K_t := \max_{j \in [2]} \|K^j_t - K^{j*}_t\|, e^P_t := \max_{j \in [2]} \|M^j_t - M^{j*}_t\|$. Using (35), (37) and (40) we get

$$\begin{aligned} e^K_t &\le \bar{c}_1 e^P_{t+1} + \Delta_t \\ e^P_t &\le (\bar{c}_2 + \bar{c}_6) N e^K_t + (\bar{c}_3 + \bar{c}_5) e^P_{t+1} + \bar{c}_4 N \Delta \\ &\le \underbrace{\left(\bar{c}_1(\bar{c}_2 + \bar{c}_6)N + \bar{c}_3 + \bar{c}_5\right)}_{\bar{c}_7} e^P_{t+1} + \underbrace{(\bar{c}_2 + \bar{c}_4 + \bar{c}_6)N}_{\bar{c}_8} \Delta_t \end{aligned}$$

Using this recursive definition we deduce

$$e^P_t \le \bar{c}_7^{T-t} e^P_T + \bar{c}_8 \sum_{s=0}^{T-1} \bar{c}_7^s \Delta_{t+s} = \bar{c}_8 \sum_{s=0}^{T-1} \bar{c}_7^s \Delta_{t+s}$$

Hence if $\Delta = \mathcal{O}(\epsilon)$, in particular $\Delta_t \le \epsilon/(2\bar{c}_1 \bar{c}_7^t \bar{c}_8 T)$ then $e^P_t \le \epsilon/2\bar{c}_1$ and $e^K_t \le \epsilon$ for $t \in \{0, \dots, T-1\}$. □