Journal of Machine Learning Research 24 (2023) 1-56 Submitted 7/21; Revised 2/23; Published 4/23

Policy Gradient Methods Find the Nash Equilibrium in N-player General-sum Linear-quadratic Games

Ben Hambly HAMBLY@MATHS.OX.AC.UK
Mathematical Institute
University of Oxford
Oxford, OX2 6GG, UK

Renyuan Xu RENYUANX@USC.EDU
Department of Industrial Systems and Engineering
University of Southern California
Los Angeles, CA 90089, USA

Huining Yang HY5564@PRINCETON.EDU
Department of Operations Research and Financial Engineering
Princeton University
Princeton, NJ 08540, USA

Editor: Andreas Krause

Abstract

We consider a general-sum N-player linear-quadratic game with stochastic dynamics over a finite horizon and prove the global convergence of the natural policy gradient method to the Nash equilibrium. In order to prove convergence of the method we require a certain amount of noise in the system. We give a condition, essentially a lower bound on the covariance of the noise in terms of the model parameters, in order to guarantee convergence. We illustrate our results with numerical experiments to show that even in situations where the policy gradient method may not converge in the deterministic setting, the addition of noise leads to convergence.

Keywords: Multi-agent reinforcement learning, linear-quadratic games, policy gradient methods, general-sum games, N-player games

1. Introduction

Policy optimization algorithms have achieved substantial empirical successes in addressing a variety of non-cooperative multi-agent problems, including self-driving vehicles (Shalev-Shwartz et al., 2016), real-time bidding games (Jin et al., 2018), and optimal execution in financial markets (Hambly et al., 2021). However, there have been few results from a theoretical perspective showing why such a class of reinforcement learning algorithms performs well with the presence of competition among agents. In the literature, the convergence of such algorithms is guaranteed only for specific classes of games including normal-form games, differentiable games, and linear-quadratic games. For normal-form games (in which there are no state dynamics), the policy gradient method does not converge in a general set-up (Singh et al., 2000) and theoretical guarantees for convergence have been established only for some special cases such as policy prediction in two-player two-action bimatrix games (Zhang and

Lesser, 2010; Song et al., 2019) and two-player two-action games (Bowling and Veloso, 2002). For differentiable games (where the cost function is assumed to be differentiable and, in most cases, the gradient is Lipschitz continuous with respect to the agent's policy parameters), there is a line of recent work (Balduzzi et al., 2018; Letcher et al., 2019; Foerster et al., 2018; Fiez et al., 2020) where the convergence guarantees for these algorithms are mainly developed for zero-sum games and cooperative games, but, in most cases, are limited to a subset of local Nash equilibrium points. However, the smoothness properties required in the differentiable game are very restrictive in general and they even fail to hold for LQ games (Zhang et al., 2019; Mazumdar et al., 2020b; Bu et al., 2019).

As a starting point to tackle this challenging problem, we investigate linear-quadratic (LQ) games which can be seen as a generalization of the linear-quadratic regulator (LQR) from a single agent to multiple agents. In an LQ game, all agents jointly control a linear state process, which may be in high dimensions, where the control (or action) from each individual agent has a linear impact on the state process. Each agent optimizes a quadratic cost function which depends on the state process, the control from this agent and/or the controls from the opponents.

LQ games are a relatively simple setting in which to analyze the behavior of multi-agent reinforcement learning (MARL) algorithms in continuous action and state spaces since they admit global Nash equilibria in the space of linear feedback policies. Moreover, these equilibria can be found by solving a coupled set of Ricatti equations when system parameters are given. As such LQ games are a natural benchmark problem on which to test policy gradient algorithms in multi-agent settings when system parameters are unknown. Furthermore, policy gradient methods open up the possibility to develop new scalable approaches to finding or learning solutions to control problems even with constraints. Finally, the empirical results presented in Mazumdar et al. (2020a) imply that, even in this relatively straightforward LQ case (with linear dynamics, linear feedback policies, and quadratic costs), policy gradient MARL would be unable to find the local Nash equilibrium in a non-negligible subset of problems. This further demonstrates the necessity of understanding under what circumstances policy gradient methods work for LQ games.

For LQ games with policy gradient algorithms, most of the existing literature has focused on zero-sum games with two players (Bu et al., 2019; Zhang et al., 2019, 2021b). In the setting of deterministic dynamics and infinite time horizon, Zhang et al. (2019) proposed an alternating policy update scheme with a projection step and showed sublinear convergence for the algorithm. For a similar setting, Bu et al. (2019) proposed a leader-follower type of policy gradient algorithm which is projection-free and enjoys a global sublinear convergence rate and asymptotically linear convergence rate. For the case of stochastic dynamics and finite time horizon, Zhang et al. (2021b) provided the first sample complexity result for the global convergence of the policy gradient method with an alternating policy update scheme. In addition, Gemp and Mahadevan (2018) proved the convergence of gradient descent methods for LQ Generative Adversarial Networks (GANs) using monotone operator theory (variational inequalities), where a GAN can be viewed as a zero-sum game and the learning of the neural network parameters has a parallel in the learning of a policy for LQ games.

However, little theory has been developed for the more general class of LQ games with N players and general-sum cost functions. It has been documented that the policy gradient method may *fail to converge* in such a setting with deterministic dynamics due to the lack of theoretical guidance on how to properly choose the step size and the exploration scheme (Mazumdar et al., 2020a). For a special class of N-player LQ games with homogeneous agents where agents interact with each other through a mean-field type of "deep state" which is the aggregated state position of all agents, Roudneshin et al. (2020) showed that the policy gradient method converges to the global Nash equilibrium. Up to now, as far as we are aware, providing theoretical guarantees for the convergence of the policy gradient method remains an open problem for general-sum LQ games with more than two players (Mazumdar et al., 2020b). We note that for other forms of general-sum and zero-sum games, Vlatakis-Gkaragkounis et al. (2020) and Mertikopoulos et al. (2018) also provided negative results for no-regret learning algorithms.

Our Contributions. In this work, we explore the natural policy gradient method, which can be viewed as a normalized version of the vanilla policy gradient descent method, for a class of N-player general-sum linear-quadratic games. Our main result is Theorem 6 in which we provide a global linear convergence guarantee for this approach in the setting of a finite time horizon and stochastic dynamics provided there is a certain level of noise in the system. The noise can either come from the underlying dynamics or carefully designed explorations from the agents. Intuitively speaking, the noise can help the agents escape from traps such as those observed in Mazumdar et al. (2020a) when the dynamics are deterministic. In addition, the noise effectively smooths the cost function, changing the optimization landscape, and thus helping agents find the correct descent direction for convergence to the desired Nash equilibrium point. To the best of our knowledge, this is the first result of its kind in the MARL literature showing a provable convergence result for N-player general-sum LQ games. From a technical perspective, the main difficulty is to quantify the perturbation of the individual's gradient term (see Step 3 in the proof of Lemma 19) and to control the descent of the individual's cost function (see Step 4 in the proof of Lemma 19) in the presence of competition from the other $N-1$ agents. We also note that our main result for the natural policy gradient method can be extended to the case of the vanilla policy gradient method, see Theorem 7 and Lemma 1.

We illustrate the performance of our algorithm with three examples. We first perform the natural policy gradient algorithm under the experimental set-up in Mazumdar et al. (2020a) over a finite time horizon. The algorithm converges to the Nash equilibrium for appropriate initial policies and step sizes. The second example is a toy LQ game example with synthetic data. The empirical results suggest that, in practice, the natural policy gradient algorithm can find the Nash equilibrium even if the level of system noise is lower than that required for our theoretical analysis. The third example is a three-player general-sum game, where we show the convergence of natural policy gradient methods with known and unknown parameters.

Comparison to the Literature on General-sum LQ games. With deterministic dynamics and infinite horizon, Mazumdar et al. (2020a) provided some empirical examples where the policy gradient method fails to converge to the set of Nash equilibria (which may not be unique). These empirical examples motivate an examination of the possibility of applying policy gradient methods in the multi-agent environment. Here we explain the difference between our framework and the set-up in Mazumdar et al. (2020a). In addition, we offer some explanations for why the policy gradient method works in our framework whereas it fails to converge in Mazumdar et al. (2020a).

- **Well-definedness of LQ games:** The existence and uniqueness of a Nash equilibrium is a prerequisite for the convergence of learning algorithms. In the setting of stochastic dynamics and finite horizon, the general-sum LQ game has a unique Nash equilibrium solution under mild conditions (Başar and Olsder, 1998). For the setting with an infinite horizon, the existence of Nash equilibria can be proved under some stabilizability and detectability properties. However obtaining explicit and verifiable model conditions for stabilizability and detectability seems to be a quite challenging task let alone finding conditions for the uniqueness of the equilibrium (Başar and Olsder, 1998). It is not clear if there exists a unique Nash equilibrium for the setting considered in Mazumdar et al. (2020a).
- **Self-exploration property of time-dependent policies:** For single-agent LQR problems, Basei et al. (2022) highlighted that the time-dependent optimal feedback policy enjoys a self-exploration property in the finite time horizon setting. Namely, the time-dependent optimal feedback matrices ensure that the optimal state and control processes span the entire parameter space, which enables the design of efficient exploration-free learning algorithms. By contrast the optimal feedback policy is time-invariant in the infinite time horizon setting and learning algorithms tend to have difficulty converging without efficient exploration schemes (Mania et al., 2019). We believe a similar analogy holds for the game setting as well.
- **System noise:** In our setting we need Assumption 4 which implies that a certain level of system noise is essential for the convergence of the policy gradient method. We show via numerical experiments in Section 5 that the circulating and divergence phenomenon described in Mazumdar et al. (2020a) can be avoided with the addition of system noise. The noise can either come from the original system when the dynamics are stochastic (as suggested in Assumption 4) or agents can apply Gaussian exploration (as suggested in Mania et al. (2019) for single-agent LQR problems with infinite time horizon).

In addition, Roudneshin et al. (2020) showed the global convergence of the policy gradient method for a mean-field type of LQ game in the setting of infinite horizon and stochastic dynamics. In particular, agents are assumed to be homogeneous and are only able to interact through an aggregated state and action pair. With this special formulation, the uniqueness of the Nash equilibrium could be established (see Theorem 1 in Roudneshin et al., 2020) and the proof of convergence could be reduced to the single agent case (see Theorem 2 in Roudneshin et al., 2020). In this paper, we focus on a more general LQ game with no homogeneity assumption nor any restriction on the interactions.

Organization and Notation For any matrix $Z = (Z_1, \cdots, Z_d) \in \mathbb{R}^{m\times d}$ with $Z_j \in \mathbb{R}^m$ ($j = 1, 2, \cdots, d$), we let $Z^\top \in \mathbb{R}^{d\times m}$ denote the transpose of Z, $\|Z\|$ denotes the spectral norm of the matrix Z; $\mathrm{Tr}(Z)$ denotes the trace of a square matrix Z; and $\sigma_{\min}(Z)$ denotes the minimal singular value of a square matrix Z. For a sequence of matrices $\boldsymbol{D} = (D_0, \cdots, D_T)$, we define a new norm $|||\boldsymbol{D}|||$ as $|||\boldsymbol{D}||| = \sum_{t=0}^{T} \|D_t\|$, where $D_t \in \mathbb{R}^{m\times d}$; $\gamma_D = \max_{t=0,\cdots,T} \|D_t\|$ denotes the maximum over all $\|D_t\|$. Furthermore we denote by $\mathcal{N}(\mu, \Sigma)$ the Gaussian distribution with mean $\mu \in \mathbb{R}^d$ and covariance matrix $\Sigma \in \mathbb{R}^{d\times d}$.

The rest of the paper is organized as follows. We introduce the mathematical framework and problem set-up in Section 2. The convergence analysis of the natural policy gradient method for the case of known model parameters is provided in Section 3. When parameters are unknown, the sample-based natural policy gradient method is discussed in Section 4. Finally, the algorithm is applied to three numerical examples in Section 5.

2. N-player General-Sum Linear-quadratic Games

We consider the following N-player general-sum linear-quadratic (LQ) game over a finite time horizon T. The state process evolves as

$$x_{t+1} = A_t x_t + \sum_{i=1}^{N} B_t^i u_t^i + w_t,\ t = 0, 1, \cdots, T-1, \tag{2.1}$$

where $x_t \in \mathbb{R}^d$ is the state of the system with the initial state x_0 drawn from a Gaussian distribution, $u_t^i \in \mathbb{R}^{k_i}$ is the control of player i at time t and $\{w_t\}_{t=0}^{T-1}$ are zero-mean IID Gaussian random variables which are independent of x_0. The system parameters $A_t \in \mathbb{R}^{d\times d}$, $B_t^i \in \mathbb{R}^{d\times k_i}$, for $t = 0, 1, \cdots, T-1$ are referred to as system (transition) matrices. The objective of player i ($i = 1, \cdots, N$) is to minimise their finite time horizon value function:

$$\inf_{\{u_t^i\}_{t=0}^{T-1}} \mathbb{E}\left[\sum_{t=0}^{T} c_t^i(x_t, u_t^i)\right], \tag{2.2}$$

where the cost function

$$c_t^i(x_t, u_t^i) = x_t^\top Q_t^i x_t + (u_t^i)^\top R_t^i u_t^i, \quad t = 0, 1, \cdots, T-1, \tag{2.3}$$

with $c_T^i(x_T) = x_T^\top Q_T^i x_T$, where $Q_t^i \in \mathbb{R}^{d\times d}$ and $R_t^i \in \mathbb{R}^{k_i\times k_i}$ ($i = 1, \cdots, N$) are matrices that parameterize the quadratic costs. Note that the randomness in the LQ game comes from both the initial state and the noise process in the state equation, therefore throughout the paper, unless specified otherwise, the expectation (for example in Equation 2.2) is taken with respect to both the initial state x_0 and the noise $\{w_t\}_{t=0}^{T-1}$. We also denote by $\boldsymbol{u}^i := (u_0^i, \cdots, u_{T-1}^i)$, $\boldsymbol{x} := (x_0, \cdots, x_T)$, $\boldsymbol{Q}^i := (Q_0^i, \cdots, Q_T^i)$, and $\boldsymbol{R}^i := (R_0^i, \cdots, R_{T-1}^i)$, for $i = 1, \cdots, N$.

Assumption 1 (Cost Parameter) *Assume for $i = 1, \cdots, N$, $Q_t^i \in \mathbb{R}^{d\times d}$, for $t = 0, 1, \cdots, T$, and $R_t^i \in \mathbb{R}^{k_i\times k_i}$, for $t = 0, 1, \cdots, T-1$ are symmetric positive definite matrices.*

Assumption 2 (Initial State and Noise Process) *Assume*

1. *Initial state: x_0 is Gaussian such that $\mathbb{E}[x_0 x_0^\top]$ is positive definite.*
2. *Noise: $\{w_t\}_{t=0}^{T-1}$ are IID Gaussian and independent from x_0 such that $\mathbb{E}[w_t] = 0$, and $W = \mathbb{E}[w_t w_t^\top]$ is positive definite, $\forall t = 0, 1, \cdots, T-1$.*

Assumption 3 (Existence and Uniqueness of Solution) *Assume there exists a unique solution set $\{K_t^{i*}\}_{t=0}^{T-1}$, for $i = 1, \cdots, N$ to the following set of linear matrix equations:*

$$K_t^{i*} = \left(R_t^i + (B_t^i)^\top P_{t+1}^{i*} B_t^i\right)^{-1} (B_t^i)^\top P_{t+1}^{i*} \left(A_t - \sum_{j=1, j\neq i}^{N} B_t^j K_t^{j*}\right), \tag{2.4}$$

where $\{P_t^{i*}\}_{t=0}^T$ *are obtained recursively backwards from*

$$P_t^{i*} = Q_t^i + (K_t^{i*})^\top R_t^i K_t^{i*} + \left(A_t - \sum_{j=1}^N B_t^j K_t^{j*}\right)^\top P_{t+1}^{i*} \left(A_t - \sum_{j=1}^N B_t^j K_t^{j*}\right), \tag{2.5}$$

with terminal condition $P_T^{i*} = Q_T^i$.

A similar assumption is adopted in Zhang et al. (2019) for a two-player zero-sum LQ game and in Roudneshin et al. (2020) for a homogeneous N-player game with mean-field interaction.

Remark 1 A sufficient condition for the unique solvability of (2.4) is the invertibility of the block matrix Φ_t, $t = 0, 1, \cdots, T-1$, with the ii-th block given by $R_t^i + (B_t^i)^\top P_{t+1}^{i*} B_t^i$ and the ij-th block given by $(B_t^i)^\top P_{t+1}^{i*} B_t^j$, where $i, j = 1, \cdots, N$ and $j \neq i$. See Remark 6.5 in Başar and Olsder (1998).

Lemma 2 (Nash Equilibrium. Başar and Olsder 1998, Corollary 6.4) *Assume Assumptions 1, 2, and 3 hold. Then for* $i = 1, 2, \cdots, N$,

1. *The Nash equilibrium strategy for player* i *is given by*

$$u_t^{i*} = -K_t^{i*} x_t, \quad t = 0, 1, \cdots T-1, \tag{2.6}$$

where K_t^{i*} *is defined in* (2.4).

2. *The Nash equilibrium cost for player* i *is*

$$\mathbb{E}[x_0^\top P_0^{i*} x_0 + N_0^{i*}] = \inf_{\{u_t^i\}_{t=0}^{T-1}} \mathbb{E}\left[\sum_{t=0}^T c_t^i(x_t, u_t^i)\right], \tag{2.7}$$

where $\{P_t^{i*}\}_{t=0}^T$ *are defined in* (2.5) *and*

$$N_t^{i*} = N_{t+1}^{i*} + \mathbb{E}[w_t^\top P_{t+1}^{i*} w_t] = N_{t+1}^{i*} + \mathrm{Tr}(W P_{t+1}^{i*}), \quad t = 0, 1, \cdots, T-1 \tag{2.8}$$

with terminal condition $N_T^{i*} = 0$.

To find the Nash equilibrium strategy in the linear feedback form (2.6), we only need to focus on the following class of linear admissible policies in feedback form

$$u_t^i = -K_t^i x_t, \qquad t = 0, 1, \cdots, T-1, \quad i = 1, \cdots, N$$

which can be fully characterized by $\boldsymbol{K}^i := (K_0^i, K_1^i, \cdots, K_{T-1}^i)$ with $K_t^i \in \mathbb{R}^{k_i \times d}$. We write $\boldsymbol{K} = (\boldsymbol{K}^1, \cdots, \boldsymbol{K}^N)$ for a collection of policies and

$$\boldsymbol{K}^* = (\boldsymbol{K}^{1*}, \cdots, \boldsymbol{K}^{N*}).$$

for the collection of optimal policies. We will use the notation

$$(\boldsymbol{K}^i, \boldsymbol{K}^{-i*}) = (\boldsymbol{K}^{1*}, \cdots, \boldsymbol{K}^{(i-1)*}, \boldsymbol{K}^i, \boldsymbol{K}^{(i+1)*}, \cdots, \boldsymbol{K}^{N*}),$$

for $i = 1, \cdots, N$ for player i's policy, when all other players use their optimal policies.

3. The Natural Policy Gradient Method with Known Parameters

In this section, we provide a global linear convergence guarantee for the natural policy gradient method applied to the LQ game (2.1) - (2.2). Throughout this section we assume all the parameters in the LQ game, $\{A_t\}_{t=0}^{T-1}$, $\{B_t^i\}_{t=0}^{T-1}$, $\{Q_t^i\}_{t=0}^{T}$, and $\{R_t^i\}_{t=0}^{T-1}$ $(i=1,\cdots,N)$, are known. The analysis with known parameters paves the way for learning LQ games with unknown parameters which is discussed in Section 4.

For LQ games, any (admissible) feedback policy can be fully characterized by a set of parameters $\boldsymbol{K}=(\boldsymbol{K}^1,\cdots,\boldsymbol{K}^N)$. Therefore, we can correspondingly define player i's cost induced by the joint policy $\boldsymbol{K}$ as

$$C^i(\boldsymbol{K}) = \mathbb{E}\left[\sum_{t=0}^{T-1}\Big(x_t^\top Q_t^i x_t + (K_t^i x_t)^\top R_t^i (K_t^i x_t)\Big) + x_T^\top Q_T^i x_T\right],$$

where $\{x_t\}_{t=0}^T$ is the random path from the dynamics (2.1) induced by $\boldsymbol{K}$ starting with x_0.

We start by introducing some notation which will be used throughout the analysis. We define the state covariance matrix $\Sigma_t^{\boldsymbol{K}}$, and let $\Sigma_{\boldsymbol{K}}$ be the sum of $\Sigma_t^{\boldsymbol{K}}$:

$$\Sigma_t^{\boldsymbol{K}} = \mathbb{E}[x_t^{\boldsymbol{K}}(x_t^{\boldsymbol{K}})^\top], \quad \Sigma_{\boldsymbol{K}} = \sum_{t=0}^{T}\Sigma_t^{\boldsymbol{K}}, \tag{3.1}$$

where $\{x_t^{\boldsymbol{K}}\}_{t=0}^T$ is a state trajectory generated by following a set of policies $\boldsymbol{K}$. We will write $x_t = x_t^{\boldsymbol{K}}$ when no confusion may occur. Define $\sigma_{\boldsymbol{X}}^{\boldsymbol{K}}$ to be the lower bound over all the minimum singular values of $\Sigma_t^{\boldsymbol{K}}$:

$$\sigma_{\boldsymbol{X}}^{\boldsymbol{K}} := \min_t \sigma_{\min}(\Sigma_t^{\boldsymbol{K}}),$$

We also define $\underline{\sigma}_{\boldsymbol{X}}$ as

$$\underline{\sigma}_{\boldsymbol{X}} := \min\{\sigma_{\min}(\mathbb{E}[x_0x_0^\top]), \sigma_{\min}(W)\}. \tag{3.2}$$

Similarly, we define

$$\underline{\sigma}_{\mathrm{R},i} = \min_t \sigma_{\min}(R_t^i), \quad \underline{\sigma}_{\mathrm{Q},i} = \min_t \sigma_{\min}(Q_t^i),$$

and

$$\underline{\sigma}_{\boldsymbol{R}} = \min_i\{\underline{\sigma}_{\mathrm{R},i}\}, \quad \underline{\sigma}_{\boldsymbol{Q}} = \min_i\{\underline{\sigma}_{\mathrm{Q},i}\}. \tag{3.3}$$

We further define γ_A, γ_B, and γ_R as

$$\gamma_A = \max_{t=0,\cdots,T-1}\|A_t\|, \quad \gamma_B = \max_i\{\max_{t=0,\cdots,T-1}\|B_t^i\|\}, \quad \gamma_R = \max_i\{\max_{t=0,\cdots,T-1}\|R_t^i\|\}. \tag{3.4}$$

Under Assumption 1, we have $\underline{\sigma}_{\mathrm{R},i} \geq \underline{\sigma}_{\boldsymbol{R}} > 0$ and $\underline{\sigma}_{\mathrm{Q},i} \geq \underline{\sigma}_{\boldsymbol{Q}} > 0$, for $i=1,\ldots,N$. For the well-definedness of the state covariance matrix, we have the following result.

Lemma 3 *Assume Assumption 2 holds. Then $\mathbb{E}[x_tx_t^\top]$ is positive definite for $t=0,1,\cdots,T$ under any set of policies $\boldsymbol{K}=(\boldsymbol{K}^1,\cdots,\boldsymbol{K}^N)$ and we have $\sigma_{\boldsymbol{X}}^{\boldsymbol{K}} \geq \underline{\sigma}_{\boldsymbol{X}} > 0$.*

Proof Let $\{x_t\}_{t=0}^T$ be the state trajectory induced by an arbitrary policy set $\boldsymbol{K}$. By Assumption 2 the matrix $\mathbb{E}[x_0 x_0^\top]$ is positive definite. For $t \geq 1$, we have from (2.1) and taking expectations

$$\mathbb{E}[x_t x_t^\top] = \left(A_{t-1} - \sum_{i=1}^N B_{t-1}^i K_{t-1}^i\right) \mathbb{E}[x_{t-1} x_{t-1}^\top] \left(A_{t-1} - \sum_{i=1}^N B_{t-1}^i K_{t-1}^i\right)^\top + \mathbb{E}[w_{t-1} w_{t-1}^\top].$$

Now as $(A_{t-1} - \sum_{i=1}^N B_{t-1}^i K_{t-1}^i)\mathbb{E}[x_{t-1}x_{t-1}^\top](A_{t-1} - \sum_{i=1}^N B_{t-1}^i K_{t-1}^i)^\top$ is positive semi-definite and by assumption $\mathbb{E}[w_{t-1}w_{t-1}^\top]$ is positive definite, we have $\mathbb{E}[x_t x_t^\top]$ is positive definite and as a result the statement holds. ■

We write $\mathcal{H} = \{h \,|\, h \text{ are polynomials in the model parameters}\}$ and $\mathcal{H}(.)$ when there are other dependencies. The model parameters are expressed in terms of $\frac{1}{1+\sum_i k_i}$, $\frac{1}{d+1}$, d, $\frac{1}{N+1}$, $\frac{1}{T+1}$, $\frac{1}{\|W\|+1}$, $\frac{1}{\|\Sigma_0\|+1}$, $\frac{1}{\gamma_A+1}$, $\frac{1}{\gamma_B+1}$, $\frac{1}{\gamma_R+1}$, $\frac{1}{\underline{\sigma}_{\boldsymbol{X}}+1}$, $\underline{\sigma}_{\boldsymbol{X}}$, $\frac{1}{\underline{\sigma}_{\boldsymbol{R}}+1}$, $\underline{\sigma}_{\boldsymbol{R}}$, $\frac{1}{\underline{\sigma}_{\boldsymbol{Q}}+1}$, $\underline{\sigma}_{\boldsymbol{Q}}$, and $\frac{1}{\|\boldsymbol{K}^*\|+1}$.

Natural Policy Gradient Method. We consider the following natural policy gradient updating rule for each player i ($i = 1, \cdots, N$), in a sequence of games for $m = 1, \ldots, M$,

$$K_t^{i,(m)} = K_t^{i,(m-1)} - \eta \nabla_{K_t^i} C^i(\boldsymbol{K}^{(m-1)})(\Sigma_t^{\boldsymbol{K}^{(m-1)}})^{-1}, \quad \forall\, 0 \leq t \leq T-1, \tag{3.5}$$

where $\nabla_{K_t^i} C^i(\boldsymbol{K}^{(m-1)}) = \frac{\partial C^i(\boldsymbol{K}^{(m-1)})}{\partial K_t^i}$ is the gradient of $C^i(\boldsymbol{K}^{(m-1)})$ with respect to K_t^i, and η is the step size.

Natural policy gradient methods (Kakade, 2001)—and related algorithms such as trust region policy optimization (Schulman et al., 2015) and the natural actor critic (Peters and Schaal, 2008)—are some of the most popular and effective variants of the vanilla policy gradient methods in single-agent reinforcement learning. The natural policy gradient method performs better than the vanilla policy gradient method in practice since it takes the information geometry (Bagnell and Schneider, 2003; Kakade, 2001) into consideration by normalizing the gradient term by $(\Sigma_t^{\boldsymbol{K}^{(m)}})^{-1}$ in (3.5). By doing so, the natural gradient method represents the steepest descent direction based on the underlying structure of the parameter space (Kakade, 2001). It is worth mentioning that Fazel et al. (2018) provided the first theoretical guarantee that the natural gradient method has an improved constant in the convergence rate compared to the vanilla version.

With access to the model parameters, players can directly calculate their own policy gradient and perform the natural policy gradient steps iteratively. See Algorithm 1 for the details. Unlike the nested-loop updates in some zero-sum LQ game settings (Zhang et al., 2021a, 2019) in which the inner-loop updates are sample inefficient, each player updates their policies simultaneously at each iteration in Algorithm 1. This simultaneous updating framework is a more realistic set-up for practical examples such as online auction bidding.

Equivalent Form of the Natural Policy Gradient. Following the explanation of the natural gradient method in the single-agent setting (Fazel et al., 2018), we provide an equivalent form of the natural gradient method with Fisher information in the game setting.

In our N-player game case, we assume player i's updating rule follows

$$K_t^{i\prime} \longleftarrow K_t^i - \eta\, G_{K_t^i}^{-1} \nabla_{K_t^i} C^i(\boldsymbol{K}), \tag{3.6}$$

where $G_{K_t^i}$ is the Fisher information matrix:

$$G_{K_t^i} = \mathbb{E}\Big[\nabla_{K_t^i} \log \pi_{K_t^i}^i(u_t|x_t) \nabla_{K_t^i} \log \pi_{K_t^i}^i(u_t|x_t)^\top\Big] \tag{3.7}$$

under a *linear* policy with *additive* Gaussian noise (Rajeswaran et al., 2017), in that

$$\pi_{K_t^i}^i(u_t^i = u|x_t) = \det(2\pi\sigma^2 I)^{-1/2} \exp\left(-\frac{1}{2\sigma^2}\left(u - K_t^i x_t\right)^\top \left(u - K_t^i x_t\right)\right). \tag{3.8}$$

By a similar analysis to that in Fazel et al. (2018), we can show that the Fisher information matrix of size $k_i d \times k_i d$, which is indexed as $[G_K]_{(j,q),(j',q')}$ where $j, j' \in \{1, 2, \cdots, k_i\}$ and $q, q'; \in \{1, 2, \cdots, d\}$, has a block diagonal form where the only non-zeros blocks are $[G_{K_t^i}]_{(j,\cdot),(j,\cdot)} = \Sigma_t^{\boldsymbol{K}} = \mathbb{E}[x_t^{\boldsymbol{K}}(x_t^{\boldsymbol{K}})^\top]$ (this is the block corresponding to the i-th coordinate of the action, as j ranges from 1 to k_i). Hence (3.6) is equivalent to the following updating rule $K_t^{i\prime} \longleftarrow K_t^i - \eta \nabla_{K_t^i} C^i(\boldsymbol{K})\Big(\Sigma_t^{\boldsymbol{K}}\Big)^{-1}$, which is equivalent to (3.5).

Remark 4 In Algorithm 1, during iteration m, each player first calculates the solution to the backward Riccati equation in (3.11) based on the observations from the previous iteration $m-1$, and obtains the policy gradients in (3.12). The natural gradient method is then applied in (3.13) to update the policy for that iteration. Note that (3.13) does not involve $(\Sigma_t^{\boldsymbol{K}})^{-1}$ since, as we show later in Lemma 9, we have $2E_{t,i}^{\boldsymbol{K}} = \nabla_{K_t^i} C^i(\boldsymbol{K})(\Sigma_t^{\boldsymbol{K}})^{-1}$.

A straightforward analysis shows that the computational complexity of Algorithm 1 is $\mathcal{O}(MNT \max(k_i^2 d, d^3))$ and that it requires $(Td\sum_{i=1}^N k_i)$ storage units for $\boldsymbol{K}$.

We now introduce the main assumption and the main result for the natural policy gradient method applied to the class of general-sum LQ games. To start we define ρ^* as

$$\rho^* := \max\left\{\max_{0 \le t \le T-1}\left\|A_t - \sum_{i=1}^N B_t^i K_t^{i*}\right\|, 1 + \delta\right\}, \tag{3.9}$$

for some small constant $\delta > 0$, and define $\psi := \max_i\{C^i(\boldsymbol{K}^{i,(0)}, \boldsymbol{K}^{-i*}) - C^i(\boldsymbol{K}^*)\}$. Now set

$$\bar{\rho} := \rho^* + N\gamma_B\sqrt{\frac{T\psi}{\underline{\sigma}_{\boldsymbol{X}}\, \underline{\sigma}_{\boldsymbol{R}}}} + \frac{1}{20T^2}. \tag{3.10}$$

Algorithm 1 Natural Policy Gradient Method with Known Parameters

```
Input: Number of iterations $M$, time horizon $T$, initial policies $\boldsymbol{K}^{(0)} = (\boldsymbol{K}^{1,(0)}, \cdots, \boldsymbol{K}^{N,(0)})$, step size $\eta$, model parameters $\{A_t\}_{t=0}^{T-1}$, $\{B_t^i\}_{t=0}^{T-1}$, $\{Q_t^i\}_{t=0}^{T}$, and $\{R_t^i\}_{t=0}^{T-1}$ $(i = 1, \cdots, N)$.
for $m \in \{1, \dots, M\}$ do
  for $t \in \{T-1, \dots, 0\}$ do
    for $i \in \{1, \dots, N\}$ do
      Calculate the matrix $P_{t,i}^{\boldsymbol{K}^{(m-1)}}$ by
        $P_{t,i}^{\boldsymbol{K}^{(m-1)}} = Q_t^i + (K_t^{i,(m-1)})^\top R_t^i K_t^{i,(m-1)} + \Big(A_t - \sum_{i=1}^N B_t^i K_t^{i,(m-1)}\Big)^\top P_{t+1,i}^{\boldsymbol{K}^{(m-1)}} \Big(A_t - \sum_{i=1}^N B_t^i K_t^{i,(m-1)}\Big)$   (3.11)
      with $P_{T,i}^{\boldsymbol{K}^{(m-1)}} = Q_T^i$.
      Calculate the matrix $E_t^i$ by
        $E_{t,i}^{\boldsymbol{K}^{(m-1)}} = R_t^i K_t^{i,(m-1)} - (B_t^i)^\top P_{t+1,i}^{\boldsymbol{K}^{(m-1)}} \left(A_t - \sum_{j=1}^N B_t^j K_t^{j,(m-1)}\right).$   (3.12)
      Update the policies using the natural policy gradient updating rule:
        $K_t^{i,(m)} = K_t^{i,(m-1)} - 2\eta E_{t,i}^{\boldsymbol{K}^{(m-1)}}.$   (3.13)
    end for
  end for
end for
Return the iterates $\boldsymbol{K}^{(M)} = (\boldsymbol{K}^{1,(M)}, \cdots, \boldsymbol{K}^{N,(M)})$.
```

Assumption 4 (System Noise) *The system parameters satisfy the following inequality*

$$\frac{(\underline{\sigma}_{\boldsymbol{X}})^5}{\|\Sigma_{\boldsymbol{K}^*}\|} > 20(N-1)^2\, T^2\, d\, \frac{(\gamma_B)^4(\max_i\{C^i(\boldsymbol{K}^*)\} + \psi)^4}{\underline{\sigma}_{\boldsymbol{Q}}{}^2\, \underline{\sigma}_{\boldsymbol{R}}{}^2} \left(\frac{\bar{\rho}^{2T} - 1}{\bar{\rho}^2 - 1}\right)^2. \tag{3.14}$$

Note that $\underline{\sigma}_{\boldsymbol{X}}$ appears on both sides of (3.14) (indirectly through $\bar{\rho}$ on the R.H.S., $\Sigma_{\boldsymbol{K}^*}$ on the L.H.S, and terms involving costs C^i) and when $\underline{\sigma}_{\boldsymbol{X}}$ increases, the LHS of (3.14) increases while the RHS of (3.14) decreases. Therefore, Assumption 4 requires $\underline{\sigma}_{\boldsymbol{X}}$ to be large enough such that inequality (3.14) holds. This can be interpreted as ensuring that the system needs a certain level of noise for the learning agents to find the correct direction towards the Nash equilibrium. Assumption 4 also imposes conditions on the initial policy $\boldsymbol{K}^{(0)}$ as ψ depends on the difference between $C^i(\boldsymbol{K}^{i,(0)}, \boldsymbol{K}^{-i*})$ and $C^i(\boldsymbol{K}^*)$.

Below we provide two examples such that Assumption 4 is satisfied. In Section 5 we will show that, at least in some circumstances, the natural policy gradient method leads to the Nash equilibrium even when Assumption 4 is violated. This further demonstrates the power of the natural policy gradient method in practice.

Remark 5 (Examples and Discussion of Assumption 4) 1. A two-player example where Assumption 4 is satisfied:

- Parameters:

$$A_t = \begin{bmatrix} 0.1 & -0.05 \\ -0.05 & 0.1 \end{bmatrix}, \quad B_t^1 = \begin{bmatrix} 0.04 \\ 0.03 \end{bmatrix}, \quad B_t^2 = \begin{bmatrix} 0.01 \\ -0.05 \end{bmatrix}, \quad W = \begin{bmatrix} 0.2 & 0.05 \\ 0.05 & 0.1 \end{bmatrix},$$

$$Q_T^1 = Q_T^2 = Q_t^1 = Q_t^2 = \begin{bmatrix} 0.1 & -0.01 \\ -0.01 & 0.1 \end{bmatrix}, \quad R_t^1(t) = R_t^2(t) = 0.35,$$

and $T = 2$.

- Initialization: Take $x_0 = (x_0^1, x_0^2)$ where x_0^1, x_0^2 are independent and sampled from $\mathcal{N}(0.25, 0.2)$ and $\mathcal{N}(0.4, 0.3)$ respectively. The initial policies are $\boldsymbol{K}^{1,(0)} = \boldsymbol{K}^{2,(0)} = (0.2, 0.01)$.

2. A three-player LQ game example where Assumption 4 is satisfied:

- Parameters:

$$A_t = \begin{bmatrix} 0.05 & -0.1 & 0.1 \\ 0.1 & 0.2 & -0.06 \\ -0.02 & 0.03 & 0.1 \end{bmatrix}, B_t^1 = \begin{bmatrix} 0.05 \\ 0.01 \\ -0.01 \end{bmatrix}, B_t^2 = \begin{bmatrix} 0.01 \\ -0.05 \\ -0.02 \end{bmatrix}, B_t^3 = \begin{bmatrix} -0.02 \\ 0.01 \\ 0.05 \end{bmatrix},$$

$$W = \begin{bmatrix} 0.1 & 0.01 & 0.02 \\ 0.01 & 0.2 & 0.01 \\ 0.02 & 0.01 & 0.1 \end{bmatrix}, \quad Q_T^1 = Q_T^2 = Q_t^1 = Q_t^2 = \begin{bmatrix} 0.2 & 0 & 0 \\ 0 & 0.2 & 0 \\ 0 & 0 & 0.2 \end{bmatrix},$$

$R_t^1(t) = R_t^2(t) = 0.5$, $R_t^3(t) = 0.6$, and $T = 1$.

- Initialization: Take $x_0 = (x_0^1, x_0^2, x_0^3)$ where x_0^1, x_0^2, x_0^3 are independent and sampled from $\mathcal{N}(0.3, 0.2)$ and $\mathcal{N}(0.2, 0.3)$, and $\mathcal{N}(0.3, 0.2)$ respectively. The initial policies are $\boldsymbol{K}^{1,(0)} = (0, -0.01, 0)$, $\boldsymbol{K}^{2,(0)} = (0, -0.01, 0.01)$, and $\boldsymbol{K}^{3,(0)} = (0, -0.001, 0)$.

3. In most large population games, the individual contribution to the joint dynamics scales like $\frac{1}{N}$ (Huang et al., 2006; Lasry and Lions, 2007). In this setting, we denote by $x_t^{(N)}$, the state dynamics with N agents, which follows

$$x_{t+1}^{(N)} = A_t^{(N)} x_t^{(N)} + \sum_{i=1}^{N} B_t^{(N),i} u_t^{(N),i} + w_t^{(N)}. \tag{3.15}$$

If $B_t^{(N),i} = \mathcal{O}(\frac{1}{N})$, in the sense that there exists a constant $C_B > 0$ (which does not depend on N) and $N_0 \in \mathbb{N}_+$ such that

$$\left\| B_t^{(N),i} \right\| \leq \frac{C_B}{N} \tag{3.16}$$

for all $0 \leq t \leq T$ and $N \geq N_0$, then we have

$$\gamma_B^{(N)} := \max_i \left\{ \max_{t=0,1,\cdots,T-1} \|B_t^{(N),i}\| \right\} \leq \frac{C_B}{N} \tag{3.17}$$

for $N \geq N_0$. In this case, (3.14) in Assumption 4 is reduced to the following condition:

$$\frac{(\underline{\sigma}_{\boldsymbol{X}})^5}{\|\Sigma_{\boldsymbol{K}^*}\|} > 20\frac{(N-1)^2}{N^4}T^2 d\frac{(C_B)^4(\max_i\{C^i(\boldsymbol{K}^*)\}+\psi)^4}{\underline{\sigma}_{\boldsymbol{Q}}^2\underline{\sigma}_{\boldsymbol{R}}^2}\left(\frac{\bar{\rho}^{2T}-1}{\bar{\rho}^2-1}\right)^2 = \mathcal{O}\left(\frac{1}{N^2}\right). \tag{3.18}$$

The RHS of condition (3.18) decays quadratically in N since one can check that $C^i(\boldsymbol{K}^*)$, ψ and $\bar{\rho}$ do not grow with respect to N in this case. Hence condition (3.18) is automatically satisfied in the large population regime.

4. In addition, if we focus on the set of stabilizing policies

$$\Omega := \{\boldsymbol{K} : \|A_t - B_t K_t\| \leq \rho < 1, \quad \forall t = 0, 1, \cdots, T-1\}, \tag{3.19}$$

then the term $\left(\frac{\bar{\rho}^{2T}-1}{\bar{\rho}^2-1}\right)^2$ on the RHS of (3.14) can be replaced by $\left(\frac{1}{1-\rho}\right)^2$, which is independent of the horizon T.

5. Furthermore, T^2 on the RHS of (3.14) can be removed when the dynamics and the cost parameters are time independent, and the agent is searching for time-independent linear policies, namely

$$R_t^i \equiv R^i, Q_t^i \equiv Q^i, A_t^i \equiv A^i, B_t^i \equiv B, K_t^i \equiv K^i, \quad t = 0, 1, \cdots, T-1, i = 1, 2, \cdots, N.$$

Theorem 6 (Global Convergence of the Natural Policy Gradient Method) *Assume Assumptions 1, 2, 3 hold. We also assume Assumption 4 so that*

$$\widehat{\alpha} := \frac{\underline{\sigma}_{\boldsymbol{X}}\,\underline{\sigma}_{\boldsymbol{R}}}{\|\Sigma_{\boldsymbol{K}^*}\|} - 20\,(N-1)^2\,T^2\,d\,\frac{(\gamma_B)^4(\max_i\{C^i(\boldsymbol{K}^*)\}+\psi)^4}{{\underline{\sigma}_{\boldsymbol{X}}}^4\,{\underline{\sigma}_{\boldsymbol{Q}}}^2\,\underline{\sigma}_{\boldsymbol{R}}}\left(\frac{\bar{\rho}^{2T}-1}{\bar{\rho}^2-1}\right)^2 > 0. \tag{3.20}$$

Then there exists an $0 < \eta_0 \in \mathcal{H}(\frac{1}{\sum_{i=1}^N C^i(\boldsymbol{K}^{i,(0)}, \boldsymbol{K}^{-i*})+1})$ *such that for* $\epsilon > 0, 0 < \eta < \eta_0$*, the cost function of the natural gradient method* (3.5) *satisfies*

$$\sum_{i=1}^N \Big(C^i(\boldsymbol{K}^{i,(M)}, \boldsymbol{K}^{-i*}) - C^i(\boldsymbol{K}^*)\Big) \leq \epsilon,$$

whenever $M \geq \frac{1}{\widehat{\alpha}\eta}\log(\frac{\sum_{i=1}^N (C^i(\boldsymbol{K}^{i,(0)}, \boldsymbol{K}^{-i*}) - C^i(\boldsymbol{K}^*))}{\epsilon})$.

The convergence rate developed in Theorem 6 is usually referred to as a linear convergence rate (Nocedal and Wright, 2006) as we show that $\frac{\sum_{i=1}^N (C^i(\boldsymbol{K}^{i,(m+1)}, \boldsymbol{K}^{-i*}) - C^i(\boldsymbol{K}^*))}{\sum_{i=1}^N (C^i(\boldsymbol{K}^{i,(m)}, \boldsymbol{K}^{-i*}) - C^i(\boldsymbol{K}^*))} \leq r$ for some $r \in (0,1)$. This leads to the result, seen in point 2. of the Theorem, that $\sum_{i=1}^N \big(C^i(\boldsymbol{K}^{i,(M)}, \boldsymbol{K}^{-i*}) - C^i(\boldsymbol{K}^*)\big) = O(e^{-M})$ and therefore this is also called an exponential convergence rate in the literature (Nocedal and Wright, 2006).

Compared to directly evaluating $C^i(\boldsymbol{K}^{(m)})$ along the training process $m = 1, 2, \cdots, M$, we observe that it is more natural to analyze the cost $C^i(\boldsymbol{K}^{i,(m)}, \boldsymbol{K}^{-i*})$, with player i taking policy $\boldsymbol{K}^{i,(m)}$ and (as if) the other players were using optimal policies, as conditional optimality is the essence of a Nash equilibrium. This view point is the key to establishing the global convergence result and this criterion is also used in all the key lemmas as well as the proof of the main theorem.

Note that our result can be easily generalized to the setting with agent-dependent learning rates as long as $\eta_i < \eta_0$ holds for each agent's step size η_i.

The proof of Theorem 6 relies on the regularity of the LQ game problem, some properties of the gradient descent dynamics, and the perturbation analysis of the covariance matrix of the controlled dynamics.

Vanilla Policy Gradient Method. By modifying parts of the proof of the convergence result for the natural policy gradient method, we can extend the convergence result to the case of the vanilla policy gradient method, where the updating rule for each player i ($i = 1, \cdots, N$), in a sequence of games for $m = 1, \dots, M$, is given by

$$K_t^{i,(m)} = K_t^{i,(m-1)} - \eta \nabla_{K_t^i} C^i(\boldsymbol{K}^{(m-1)}), \quad \forall\, 0 \le t \le T-1. \tag{3.21}$$

We now give the system noise assumption and the global convergence result for the vanilla policy gradient method.

Assumption 5 (System Noise for the Vanilla Policy Gradient) *The system parameters satisfy the following two inequalities*

$$\frac{(\underline{\sigma}_{\boldsymbol{X}})^7}{\|\Sigma_{\boldsymbol{K}^*}\|} > 20(N-1)^2\, T^2\, d\, \frac{(\gamma_B)^4(\max_i\{C^i(\boldsymbol{K}^*)\}+\psi)^4}{{\underline{\sigma}_{\boldsymbol{Q}}}^2\, {\underline{\sigma}_{\boldsymbol{R}}}^2} \left(\frac{\bar{\rho}^{2T}-1}{\bar{\rho}^2-1}\right)^2 \left(\bar{\rho}^{2T}\|\Sigma_0\| + (\bar{\rho}^{2T}+1)\|W\|\right)^2 \tag{3.22}$$

and

$$(\underline{\sigma}_{\boldsymbol{X}})^2 > \frac{5(\max_i\{C^i(\boldsymbol{K}^*)\}+\psi)}{\underline{\sigma}_{\boldsymbol{Q}}} \left(\frac{\max_i\{C^i(\boldsymbol{K}^*)\}+\psi}{\underline{\sigma}_{\boldsymbol{Q}}} + \bar{\rho}^{2T}\|\Sigma_0\| + (\bar{\rho}^{2T}+1)\|W\|\right). \tag{3.23}$$

Theorem 7 (Global Convergence of the Vanilla Policy Gradient Method) *Assume Assumptions 1, 2, 3 hold. We also assume Assumption 5 so that*

$$\begin{aligned}\widetilde{\alpha} \;\; := \;\; & \frac{{\underline{\sigma}_{\boldsymbol{X}}}^2\, \underline{\sigma}_{\boldsymbol{R}}}{\|\Sigma_{\boldsymbol{K}^*}\|} - 20\,(N-1)^2\, T^2\, d\, \frac{(\gamma_B)^4(\max_i\{C^i(\boldsymbol{K}^*)\}+\psi)^4}{{\underline{\sigma}_{\boldsymbol{X}}}^5\, {\underline{\sigma}_{\boldsymbol{Q}}}^2\, \underline{\sigma}_{\boldsymbol{R}}} \left(\frac{\bar{\rho}^{2T}-1}{\bar{\rho}^2-1}\right)^2 \cdot \\ & \left(\bar{\rho}^{2T}\|\Sigma_0\| + (\bar{\rho}^{2T}+1)\|W\|\right)^2 > 0.\end{aligned} \tag{3.24}$$

Then there exists an $0 < \widetilde{\eta}_0 \in \mathcal{H}(\frac{1}{\sum_{i=1}^N C^i(\boldsymbol{K}^{i,(0)}, \boldsymbol{K}^{-i*})+1})$ *such that for* $\epsilon > 0, 0 < \eta < \widetilde{\eta}_0$, *the cost function of the vanilla policy gradient method* (3.21) *satisfies*

$$\sum_{i=1}^N \left(C^i(\boldsymbol{K}^{i,(M)}, \boldsymbol{K}^{-i*}) - C^i(\boldsymbol{K}^*)\right) \le \epsilon,$$

whenever $M \ge \frac{1}{\widetilde{\alpha}\eta} \log(\frac{\sum_{i=1}^N (C^i(\boldsymbol{K}^{i,(0)}, \boldsymbol{K}^{-i*}) - C^i(\boldsymbol{K}^*))}{\epsilon})$.

Since the key step to prove the global convergence of the natural policy gradient method (Theorem 6) is the one-step contraction lemma (Lemma 19), we also establish the one-step contraction result for the vanilla version by modifying some parts of Lemma 19, see Lemma 1 and its proof in Appendix A. To prove Theorem 7, it suffices to show that Lemma 1 holds, and the rest of the proof follows the same arguments as in the proof of Theorem 6 for the natural version.

We explain the differences between the system noise assumptions required for the vanilla and the natural policy gradient methods in Remark 1 in Appendix A.

3.1 Regularity of the LQ game and Properties of the Gradient Descent Dynamics

We begin with the analysis of some properties of the N-player general-sum LQ game (2.1)-(2.3). Our aim is to establish two key results Lemma 11 and Lemma 12 which provide the gradient dominance condition and a smoothness condition on the cost function $C^i(\boldsymbol{K})$ of player i with respect to the joint policy $\boldsymbol{K}$ from all players, respectively.

In the finite horizon setting, define $P^{\boldsymbol{K}}_{t,i}$ as the solution to

$$P^{\boldsymbol{K}}_{t,i} = Q^i_t + (K^i_t)^\top R^i_t K^i_t + \left(A_t - \sum_{j=1}^N B^j_t K^j_t\right)^\top P^{\boldsymbol{K}}_{t+1,i}\left(A_t - \sum_{j=1}^N B^j_t K^j_t\right), \tag{3.25}$$

with terminal condition

$$P^{\boldsymbol{K}}_{T,i} = Q^i_T.$$

Lemma 8 *Assume Assumptions 1 holds. Then for $i = 1, \cdots, N$, $t = 0, \ldots, T$ the matrices $P^{\boldsymbol{K}}_{t,i}$ defined in* (3.25) *are positive definite.*

Proof We prove that the terms in the sequence $\{P^{\boldsymbol{K}}_{t,i}\}_{t=0}^T$ are positive definite for $i = 1, \cdots, N$ by backward induction. For $t = T$, $P^{\boldsymbol{K}}_{T,i} = Q^i_T$ is positive definite since Q^i_T is positive definite. Assume $P^{\boldsymbol{K}}_{t+1,i}$ is positive definite for some $t+1$, then take any $z \in \mathbb{R}^d$ such that $z \neq 0$,

$$z^\top P^{\boldsymbol{K}}_{t,i} z = z^\top Q^i_t z + z^\top (K^i_t)^\top R^i_t K^i_t z + z^\top \left(A_t - \sum_{j=1}^N B^j_t K^j_t\right)^\top P^{\boldsymbol{K}}_{t+1,i}\left(A_t - \sum_{j=1}^N B^j_t K^j_t\right) z > 0.$$

The last inequality holds since $z^\top Q^i_t z > 0$ and the other terms are non-negative under Assumption 1. By backward induction, we have $P^{\boldsymbol{K}}_{t,i}$ positive definite, $\forall\, t = 0, 1, \cdots, T$. ■

Player i's cost under the set of policies $\boldsymbol{K}$ can be rewritten as

$$C^i(\boldsymbol{K}) = \mathbb{E}\left[x_0^\top P^{\boldsymbol{K}}_{0,i} x_0 + N^{\boldsymbol{K}}_{0,i}\right],$$

where the expectation is taking with respect to the initial state x_0 and the $N^{\boldsymbol{K}}_{t,i}$ are defined backwards:

$$N^{\boldsymbol{K}}_{t,i} = N^{\boldsymbol{K}}_{t+1,i} + \mathrm{Tr}(W P^{\boldsymbol{K}}_{t+1,i}), \quad N^{\boldsymbol{K}}_{T,i} = 0. \tag{3.26}$$

To see this,

$$
\begin{aligned}
\mathbb{E}\left[x_0^\top P_{0,i}^{\boldsymbol{K}} x_0 + N_{0,i}^{\boldsymbol{K}}\right] &= \mathbb{E}\left[x_0^\top Q_0^i x_0 + x_0^\top (K_0^i)^\top R_0^i K_0^i x_0 + \sum_{t=0}^{T-1} w_t^\top P_{t+1,i}^{\boldsymbol{K}} w_t \right. \\
&\qquad \left. + x_0^\top (A_0 - \sum_{j=1}^N B_0^j K_0^j)^\top P_{1,i}^{\boldsymbol{K}} (A_0 - \sum_{j=1}^N B_0^j K_0^j) x_0 \right] \\
&= \mathbb{E}\left[x_0^\top Q_0^i x_0 + (u_0^i)^\top R_0^i u_0^i + x_1^\top P_{1,i}^{\boldsymbol{K}} x_1 + \sum_{t=1}^{T-1} w_t^\top P_{t+1,i}^{\boldsymbol{K}} w_t\right] \\
&= \mathbb{E}\left[c_0^i(x_0, u_0^i) + x_1^\top P_{1,i}^{\boldsymbol{K}} x_1 + \sum_{t=1}^{T-1} w_t^\top P_{t+1,i}^{\boldsymbol{K}} w_t\right] \\
&= \mathbb{E}\left[c_0^i(x_0, u_0^i) + c_1^i(x_1, u_1^i) + x_2^\top P_{2,i}^{\boldsymbol{K}} x_2 + \sum_{t=2}^{T-1} w_t^\top P_{t+1,i}^{\boldsymbol{K}} w_t\right] \\
&= \mathbb{E}\left[\sum_{t=0}^{T} c_t^i(x_t, u_t^i)\right].
\end{aligned}
$$

In addition, for $t = 0, 1, \cdots, T-1$ and $i = 1, \cdots, N$, define

$$
E_{t,i}^{\boldsymbol{K}} = R_t^i K_t^i - (B_t^i)^\top P_{t+1,i}^{\boldsymbol{K}} \left(A_t - \sum_{j=1}^N B_t^j K_t^j\right). \tag{3.27}
$$

Then we have the following representation of the gradient terms.

Lemma 9 *The policy gradients have the following representation: for $t = 0, 1, \cdots, T-1$,*

$$
\nabla_{K_t^i} C(\boldsymbol{K}) = 2\left(R_t^i K_t^i - (B_t^i)^\top P_{t+1,i}^{\boldsymbol{K}} \left(A_t - \sum_{j=1}^N B_t^j K_t^j\right)\right) \Sigma_t^{\boldsymbol{K}} = 2 E_{t,i}^{\boldsymbol{K}} \Sigma_t^{\boldsymbol{K}}, \tag{3.28}
$$

where $E_{t,i}^{\boldsymbol{K}}$ is defined in (3.27).

Proof Expressing the cost function in terms of K_t^i and suppressing the arguments of the cost c_t^i,

$$
\begin{aligned}
C^i(\boldsymbol{K}) &= \mathbb{E}\left[\sum_{s=0}^{t-1} c_s^i + c_t^i + x_{t+1}^\top P_{t+1,i}^{\boldsymbol{K}} x_{t+1} + \sum_{s=t+1}^{T-1} w_s^\top P_{s+1,i}^{\boldsymbol{K}} w_s\right] \\
&= \mathbb{E}\left[\sum_{s=0}^{t-1} c_s^i + x_t^\top (Q_t^i + (K_t^i)^\top R_t^i K_t^i) x_t + \sum_{s=t+1}^{T-1} w_s^\top P_{s+1,i}^{\boldsymbol{K}} w_s \right. \\
&\quad \left. + \left(x_t^\top \left(A_t - \sum_{j=1}^N B_t^j K_t^j\right)^\top + w_t^\top\right) P_{t+1,i}^{\boldsymbol{K}} \left(\left(A_t - \sum_{j=1}^N B_t^j K_t^j\right) x_t + w_t\right)\right].
\end{aligned}
$$

Therefore, for $i = 1, \ldots, N$, the gradients are given by

$$\nabla_{K_t^i} C^i(\boldsymbol{K}) = 2 \left(R_t^i K_t^i - (B_t^i)^\top P_{t+1,i}^{\boldsymbol{K}} \left(A_t - \sum_{j=1}^{N} B_t^j K_t^j \right) \right) \Sigma_t^{\boldsymbol{K}}.$$

■

We can write the value function for player i, $V_{\boldsymbol{K}}^i(x,t)$ for $t = 0, 1, \cdots, T-1$, as

$$V_{\boldsymbol{K}}^i(x,t) = \mathbb{E}_{\boldsymbol{w}} \left[\sum_{s=t}^{T-1} (x_s^\top Q_s^i x_s + (u_s^i)^\top R_s^i u_s^i) + x_T^\top Q_T^i x_T \middle| x_t = x \right] = x^\top P_{t,i}^{\boldsymbol{K}} x + N_{t,i}^{\boldsymbol{K}},$$

with terminal condition

$$V_{\boldsymbol{K}}^i(x,T) = x^\top Q_T^i x, \tag{3.29}$$

where $N_{t,i}^{\boldsymbol{K}}$ is defined in (3.26), and $\mathbb{E}_{\boldsymbol{w}}$ denotes the expectation over the noise $\boldsymbol{w}$. We then define the Q function, $Q_{\boldsymbol{K}}^i(x,u,t)$ for the Markovian control $u = (u^1, \cdots, u^N)$ to be

$$Q_{\boldsymbol{K}}^i(x,u,t) = x^\top Q_t^i x + (u^i)^\top R_t^i u^i + \mathbb{E}_{w_t} \left[V_{\boldsymbol{K}}^i \Big(A_t x + \sum_{j=1}^{N} B_t^j u^j + w_t, t+1 \Big) \right],$$

and the advantage function to be

$$A_{\boldsymbol{K}}^i(x,u,t) = Q_{\boldsymbol{K}}^i(x,u,t) - V_{\boldsymbol{K}}^i(x,t).$$

Here $\mathbb{E}_{w_t}$ denotes the expectation taken with respect to w_t. Note that $C^i(\boldsymbol{K}) = \mathbb{E}[V_{\boldsymbol{K}}^i(x_0,0)]$. We write

$$(\boldsymbol{K}^{i\prime}, \boldsymbol{K}^{-i}) := (\boldsymbol{K}^1, \cdots, \boldsymbol{K}^{i-1}, \boldsymbol{K}^{i\prime}, \boldsymbol{K}^{i+1}, \cdots, \boldsymbol{K}^N),$$

and the control sequences under the policy $(\boldsymbol{K}^{i\prime}, \boldsymbol{K}^{-i})$ as $\{u_t^{i\prime,-i}\}_{t=0}^{T-1}$ with

$$u_t^{i\prime} = -K_t^{i\prime} x_t^{\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i}}, \quad \text{and} \quad u_t^j = -K_t^j x_t^{\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i}}, \quad j \neq i, \tag{3.30}$$

where $\{x_t^{\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i}}\}_{t=0}^{T}$ is the state trajectory under $(\boldsymbol{K}^{i\prime}, \boldsymbol{K}^{-i})$. Then we can write the difference between the cost functions of $\boldsymbol{K} = (\boldsymbol{K}^1, \cdots, \boldsymbol{K}^N)$ and $(\boldsymbol{K}^{i\prime}, \boldsymbol{K}^{-i})$ in terms of advantage functions.

Lemma 10 (Cost Difference) *Assume $\boldsymbol{K}$ and $(\boldsymbol{K}^{i\prime}, \boldsymbol{K}^{-i})$ have finite costs. Then*

$$V_{\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i}}^i(x,0) - V_{\boldsymbol{K}}^i(x,0) = \mathbb{E}_{\boldsymbol{w}} \left[\sum_{t=0}^{T-1} A_{\boldsymbol{K}}^i(x_t^{\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i}}, u_t^{i\prime,-i}, t) \right], \tag{3.31}$$

where $\{u_t^{i\prime,-i}\}_{t=0}^{T-1}$ is defined in (3.30) *with $x_0^{\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i}} = x$. For player i,*

$$\begin{aligned} A_{\boldsymbol{K}}^i(x_t^{\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i}}, u_t^{i\prime,-i}, t) &= (x_t^{\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i}})^\top (K_t^{i\prime} - K_t^i)^\top (R_t^i + (B_t^i)^\top P_{t+1,i}^{\boldsymbol{K}} B_t^i)(K_t^{i\prime} - K_t^i)\, x_t^{\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i}} \\ &\quad + 2(x_t^{\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i}})^\top (K_t^{i\prime} - K_t^i)^\top E_{t,i}^{\boldsymbol{K}}\, x_t^{\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i}}. \end{aligned}$$

Proof Write $x'_t = x_t^{\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i}}$ for simplicity and denote by $c^{i\prime}_t(x)$ the (instantaneous) cost of player i generated by $(\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i})$ with a single trajectory starting from $x'_0 = x$. That is,

$$c^{i\prime}_t(x'_t) = (x'_t)^\top Q^i_t x'_t + (K^{i\prime}_t x^{i\prime}_t)^\top R^i_t K^{i\prime}_t x^{i\prime}_t, \quad t = 0,1,\cdots,T-1,$$

and

$$c^{i\prime}_T(x'_T) = (x'_T)^\top Q^i_T x'_T,$$

with

$$u^{i\prime}_t = -K^{i\prime}_t x'_t \quad ,x'_{t+1} = A_t x'_t + B^i_t u^{i\prime}_t + \sum_{j=1,j\neq i}^N B^j_t u^j_t + w_t, \;\; x'_0 = x.$$

Therefore

$$\begin{aligned}
V^i_{\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i}}(x,0) - V^i_{\boldsymbol{K}}(x,0) &= \mathbb{E}_{\boldsymbol{w}}\left[\sum_{t=0}^T c^{i\prime}_t(x'_t)\right] - V^i_{\boldsymbol{K}}(x,0)\\
&= \mathbb{E}_{\boldsymbol{w}}\left[\sum_{t=0}^T \left(c^{i\prime}_t(x'_t) - V^i_{\boldsymbol{K}}(x'_t,t)\right) + \sum_{t=1}^T V^i_{\boldsymbol{K}}(x'_t,t)\right]\\
&= \mathbb{E}_{\boldsymbol{w}}\left[\sum_{t=0}^{T-1} \left(c^{i\prime}_t(x'_t) + V^i_{\boldsymbol{K}}(x'_{t+1},t+1) - V^i_{\boldsymbol{K}}(x'_t,t)\right)\right]\\
&= \mathbb{E}_{\boldsymbol{w}}\left[\sum_{t=0}^{T-1} \left(Q^i_{\boldsymbol{K}}(x'_t,u^{i\prime,-i}_t,t) - V^i_{\boldsymbol{K}}(x'_t,t)\right)\middle| x_0 = x\right]\\
&= \mathbb{E}_{\boldsymbol{w}}\left[\sum_{t=0}^{T-1} A^i_{\boldsymbol{K}}(x'_t,u^{i\prime,-i}_t,t)\middle| x_0 = x\right],
\end{aligned}$$

where the third equality holds since $c^{i\prime}_T(x'_T) = V^i_{\boldsymbol{K}}(x'_T,T)$ by (3.29) with the same single trajectory.

For player i, the advantage function is given by

$$\begin{aligned}
&A^i_{\boldsymbol{K}}(x'_t,u^{i\prime,-i}_t,t) = Q^i_{\boldsymbol{K}}(x'_t,u^{i\prime,-i}_t,t) - V^i_{\boldsymbol{K}}(x'_t,t)\\
=\;& (x'_t)^\top(Q^i_t + (K^{i\prime}_t)^\top R^i_t K^{i\prime}_t)x'_t + \mathbb{E}_{w_t}\left[V^i_{\boldsymbol{K}}((A_t - B^i_t K^{i\prime}_t - \sum_{j=1,j\neq i}^N B^j_t K^j_t)x'_t + w_t, t+1)\right]\\
&-V^i_{\boldsymbol{K}}(x'_t,t)\\
=\;& (x'_t)^\top(Q^i_t + (K^{i\prime}_t)^\top R^i_t K^{i\prime}_t)x'_t + \Bigg((x'_t)^\top\Big(A_t - B^i_t K^{i\prime}_t - \sum_{j=1,j\neq i}^N B^j_t K^j_t\Big)^\top P^{\boldsymbol{K}}_{t+1,i}\\
&\Big(A_t - B^i_t K^{i\prime}_t - \sum_{j=1,j\neq i}^N B^j_t K^j_t\Big)x'_t + \mathrm{Tr}(W P^{\boldsymbol{K}}_{t+1,i}) + N^{\boldsymbol{K}}_{t+1,i}\Bigg) - \Big((x'_t)^\top P^{\boldsymbol{K}}_{t,i} x'_t + N^{\boldsymbol{K}}_{t,i}\Big)\\
=\;& (x'_t)^\top(Q^i_t + (K^{i\prime}_t - K^i_t + K^i_t)^\top R^i_t (K^{i\prime}_t - K^i_t + K^i_t))x'_t\\
&+(x'_t)^\top\Big(A_t - B^i_t(K^{i\prime}_t - K^i_t) - \sum_{j=1}^N B^j_t K^j_t\Big)^\top P^{\boldsymbol{K}}_{t+1,i}\Big(A_t - B^i_t(K^{i\prime}_t - K^i_t) - \sum_{j=1}^N B^j_t K^j_t\Big)x'_t
\end{aligned}$$

$$
\begin{aligned}
&-(x_t')^\top \left(Q_t^i + (K_t^i)^\top R_t^i K_t^i + \Big(A_t - \sum_{j=1}^N B_t^j K_t^j\Big)^\top P_{t+1,i}^{\boldsymbol{K}} \Big(A_t - \sum_{j=1}^N B_t^j K_t^j\Big)\right) x_t' \\
= \quad & (x_t')^\top (K_t^{i\prime} - K_t^i)^\top (R_t^i + (B_t^i)^\top P_{t+1,i}^{\boldsymbol{K}} B_t^i)(K_t^{i\prime} - K_t^i) x_t' + 2(x_t')^\top (K_t^{i\prime} - K_t^i)^\top E_{t,i}^{\boldsymbol{K}}\, x_t'.
\end{aligned}
$$

■

Note that the derivations of Lemmas 8, 9 and 10 are largely inspired by Fazel et al. (2018) and Hambly et al. (2021). However, the final expressions are different since our setting is different from both Fazel et al. (2018) and Hambly et al. (2021). For completeness, we provide the proofs and derivations in our context.

For the policy gradient method (Fazel et al., 2018; Hambly et al., 2021) in the single-agent setting, gradient domination and smoothness of the objective function are two key conditions to guarantee the global convergence of the gradient descent methods. This is also the case for the N-player game setting. The gradient dominance condition for each player i is proved in Lemma 11, which indicates that for a policy $\boldsymbol{K}$, the distance between $C^i(\boldsymbol{K})$ and the optimal cost $C^i(\boldsymbol{K}^*)$ is bounded by the sum of the magnitudes of the gradients $\nabla_t C^i(\boldsymbol{K})$ for $t = 0, 1, \cdots, T-1$. The smoothness condition for each player i is proved in Lemma 12 where the difference between $C^i(\boldsymbol{K}^{i\prime}, \boldsymbol{K}^{-i})$ and $C^i(\boldsymbol{K})$ can be rewritten as a function of $\boldsymbol{K}^{i\prime} - \boldsymbol{K}^i$.

Lemma 11 (Gradient Dominance) *Assume Assumptions 1, 2, and 3 hold. Then, for player i, we have*

$$
\begin{aligned}
C^i(\boldsymbol{K}^i, \boldsymbol{K}^{-i*}) - C^i(\boldsymbol{K}^*) \quad &\leq \quad \frac{\|\Sigma_{\boldsymbol{K}^*}\|}{\underline{\sigma}_{\boldsymbol{R}}} \sum_{t=0}^{T-1} \mathrm{Tr}\left((E_{t,i}^{\boldsymbol{K}^i, \boldsymbol{K}^{-i*}})^\top E_{t,i}^{\boldsymbol{K}^i, \boldsymbol{K}^{-i*}} \right) \\
&\leq \quad \frac{\|\Sigma_{\boldsymbol{K}^*}\|}{4\, \underline{\sigma}_{\boldsymbol{R}}\, (\underline{\sigma}_{\boldsymbol{X}})^2} \sum_{t=0}^{T-1} \mathrm{Tr}\left(\nabla_{K_t^i} C^i(\boldsymbol{K}^i, \boldsymbol{K}^{-i*})^\top \nabla_{K_t^i} C^i(\boldsymbol{K}^i, \boldsymbol{K}^{-i*}) \right),
\end{aligned}
$$

and

$$
\begin{aligned}
C^i(\boldsymbol{K}^i, \boldsymbol{K}^{-i*}) - C^i(\boldsymbol{K}^*) \quad &\geq \quad \underline{\sigma}_{\boldsymbol{X}} \sum_{t=0}^{T-1} \frac{1}{\|R_t^i + (B_t^i)^\top P_{t+1,i}^{\boldsymbol{K}^i, \boldsymbol{K}^{-i*}} B_t^i\|} \mathrm{Tr}\left((E_{t,i}^{\boldsymbol{K}^i, \boldsymbol{K}^{-i*}})^\top E_{t,i}^{\boldsymbol{K}^i, \boldsymbol{K}^{-i*}} \right) \\
&\geq \quad \frac{\underline{\sigma}_{\boldsymbol{X}}}{4\|\Sigma_{\boldsymbol{K}^i, \boldsymbol{K}^{-i*}}\|^2} \sum_{t=0}^{T-1} \frac{\mathrm{Tr}\left(\nabla_{K_t^i} C^i(\boldsymbol{K}^i, \boldsymbol{K}^{-i*})^\top \nabla_{K_t^i} C^i(\boldsymbol{K}^i, \boldsymbol{K}^{-i*}) \right)}{\|R_t^i + (B_t^i)^\top P_{t+1,i}^{\boldsymbol{K}^i, \boldsymbol{K}^{-i*}} B_t^i\|}.
\end{aligned}
$$

Proof By Lemma 10, we have

$$
\begin{aligned}
& A_{\boldsymbol{K}}^i(x_t^{\boldsymbol{K}^{i\prime}, \boldsymbol{K}^{-i}}, u_t^{i\prime, -i}, t) \\
= \quad & (x_t^{\boldsymbol{K}^{i\prime}, \boldsymbol{K}^{-i}})^\top (K_t^{i\prime} - K_t^i)^\top (R_t^i + (B_t^i)^\top P_{t+1,i}^{\boldsymbol{K}} B_t^i)(K_t^{i\prime} - K_t^i)\, x_t^{\boldsymbol{K}^{i\prime}, \boldsymbol{K}^{-i}} \\
& + 2(x_t^{\boldsymbol{K}^{i\prime}, \boldsymbol{K}^{-i}})^\top (K_t^{i\prime} - K_t^i)^\top E_{t,i}^{\boldsymbol{K}}\, x_t^{\boldsymbol{K}^{i\prime}, \boldsymbol{K}^{-i}} \\
= \quad & \mathrm{Tr}\left(x_t^{\boldsymbol{K}^{i\prime}, \boldsymbol{K}^{-i}} (x_t^{\boldsymbol{K}^{i\prime}, \boldsymbol{K}^{-i}})^\top (K_t^{i\prime} - K_t^i)^\top (R_t^i + (B_t^i)^\top P_{t+1,i}^{\boldsymbol{K}} B_t^i)(K_t^{i\prime} - K_t^i) \right)
\end{aligned}
$$

$$
\begin{aligned}
& + 2 \operatorname{Tr}\left(x_t^{\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i}} (x_t^{\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i}})^\top (K_t^{i\prime} - K_t^i)^\top E_{t,i}^{\boldsymbol{K}}\right) \\
= \quad & \operatorname{Tr}\left(x_t^{\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i}} (x_t^{\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i}})^\top (K_t^{i\prime} - K_t^i + (R_t^i + (B_t^i)^\top P_{t+1,i}^{\boldsymbol{K}} B_t^i)^{-1} E_{t,i}^{\boldsymbol{K}})^\top \cdot \right. \\
& \qquad \left. (R_t^i + (B_t^i)^\top P_{t+1,i}^{\boldsymbol{K}} B_t^i)(K_t^{i\prime} - K_t^i + (R_t^i + (B_t^i)^\top P_{t+1,i}^{\boldsymbol{K}} B_t^i)^{-1} E_{t,i}^{\boldsymbol{K}})\right) \\
& - \operatorname{Tr}\left(x_t^{\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i}} (x_t^{\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i}})^\top (E_{t,i}^{\boldsymbol{K}})^\top (R_t^i + (B_t^i)^\top P_{t+1,i}^{\boldsymbol{K}} B_t^i)^{-1} E_{t,i}^{\boldsymbol{K}}\right) \\
\geq \quad & - \operatorname{Tr}\left(x_t^{\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i}} (x_t^{\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i}})^\top (E_{t,i}^{\boldsymbol{K}})^\top (R_t^i + (B_t^i)^\top P_{t+1,i}^{\boldsymbol{K}} B_t^i)^{-1} E_{t,i}^{\boldsymbol{K}}\right),
\end{aligned}
$$

with equality in the last line when $K_t^{i\prime} = K_t^i - (R_t^i + (B_t^i)^\top P_{t+1,i}^{\boldsymbol{K}} B_t^i)^{-1} E_{t,i}^{\boldsymbol{K}}$. Then, by Lemma 10, letting $\{x_t^*\}_{t=0}^T$ and $u_t^* = (u_t^{1*}, \cdots, u_t^{N*})$ with $u_t^{i*} = -K_t^{i*} x_t^*$ denote the state and control sequences induced by the set of optimal policies $\boldsymbol{K}^*$, we have

$$
\begin{aligned}
& C^i(\boldsymbol{K}^i, \boldsymbol{K}^{-i*}) - C^i(\boldsymbol{K}^*) = -\mathbb{E}\left[\sum_{t=0}^{T-1} A^i_{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}}(x_t^*, u_t^*, t)\right] \\
\leq \quad & \mathbb{E}\left[\sum_{t=0}^{T-1} \operatorname{Tr}\left(x_t^*(x_t^*)^\top (E_{t,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}})^\top (R_t^i + (B_t^i)^\top P_{t+1,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}} B_t^i)^{-1} E_{t,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}}\right)\right] \\
\leq \quad & \|\Sigma_{\boldsymbol{K}^*}\| \sum_{t=0}^{T-1} \operatorname{Tr}\left((E_{t,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}})^\top (R_t^i + (B_t^i)^\top P_{t+1,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}} B_t^i)^{-1} E_{t,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}}\right) \qquad (3.32) \\
\leq \quad & \frac{\|\Sigma_{\boldsymbol{K}^*}\|}{\underline{\sigma}_{\boldsymbol{R}}} \sum_{t=0}^{T-1} \operatorname{Tr}\left((E_{t,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}})^\top E_{t,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}}\right) \\
= \quad & \frac{\|\Sigma_{\boldsymbol{K}^*}\|}{4\,\underline{\sigma}_{\boldsymbol{R}}} \sum_{t=0}^{T-1} \operatorname{Tr}\left((\Sigma_t^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}})^{-1} \nabla_{K_t^i} C^i(\boldsymbol{K}^i,\boldsymbol{K}^{-i*})^\top \nabla_{K_t^i} C^i(\boldsymbol{K}^i,\boldsymbol{K}^{-i*}) (\Sigma_t^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}})^{-1}\right) \\
\leq \quad & \frac{\|\Sigma_{\boldsymbol{K}^*}\|}{4\,\underline{\sigma}_{\boldsymbol{R}}\,(\underline{\sigma}_{\boldsymbol{X}})^2} \sum_{t=0}^{T-1} \operatorname{Tr}\left(\nabla_{K_t^i} C^i(\boldsymbol{K}^i,\boldsymbol{K}^{-i*})^\top \nabla_{K_t^i} C^i(\boldsymbol{K}^i,\boldsymbol{K}^{-i*})\right).
\end{aligned}
$$

Note that (3.32) holds as $\operatorname{Tr}(AB) \leq \sigma_{\max}(A)\operatorname{Tr}(B)$ for any matrix A and real symmetric positive semi-definite matrices B of the same size (Saniuk and Rhodes, 1987). By Lemma 1 in Wang et al. (1986), for any symmetric matrix A and any symmetric positive semi-definite matrix B, it holds that

$$
\sigma_{\min}(A)\operatorname{Tr}(B) \leq \operatorname{Tr}(AB) \leq \sigma_{\max}(A)\operatorname{Tr}(B). \qquad (3.33)
$$

These bounds will be used in several places. For the lower bound, consider $K_t^{i\prime} = K_t^i - (R_t^i + (B_t^i)^\top P_{t+1,i}^{\boldsymbol{K}} B_t^i)^{-1} E_{t,i}^{\boldsymbol{K}}$. Using $C^i(\boldsymbol{K}^{i\prime}, \boldsymbol{K}^{-i*}) \geq C^i(\boldsymbol{K}^*)$ and letting $\{u_t^{\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i*}}\}_{t=0}^{T-1}$ denote the control sequence induced by $(\boldsymbol{K}^{i\prime}, \boldsymbol{K}^{-i*})$, by Lemma 10 we have

$$
\begin{aligned}
& C^i(\boldsymbol{K}^i, \boldsymbol{K}^{-i*}) - C^i(\boldsymbol{K}^*) \\
\geq \quad & C^i(\boldsymbol{K}^i, \boldsymbol{K}^{-i*}) - C^i(\boldsymbol{K}^{i\prime}, \boldsymbol{K}^{-i*}) \\
= \quad & -\mathbb{E}\left[\sum_{t=0}^{T-1} A^i_{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}}(x_t^{\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i*}}, u_t^{\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i*}}, t)\right]
\end{aligned}
$$

$$
\begin{aligned}
&= \mathbb{E}\left[\sum_{t=0}^{T-1} \mathrm{Tr}\left(x_t^{\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i*}}(x_t^{\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i*}})^\top (E_{t,i}^{\boldsymbol{K}^{i},\boldsymbol{K}^{-i*}})^\top (R_t^i + (B_t^i)^\top P_{t+1,i}^{\boldsymbol{K}^{i},\boldsymbol{K}^{-i*}} B_t^i)^{-1} E_{t,i}^{\boldsymbol{K}^{i},\boldsymbol{K}^{-i*}}\right)\right] \\
&\geq \underline{\sigma}_{\boldsymbol{X}} \sum_{t=0}^{T-1} \frac{1}{\|R_t^i + (B_t^i)^\top P_{t+1,i}^{\boldsymbol{K}^{i},\boldsymbol{K}^{-i*}} B_t^i\|} \mathrm{Tr}\left((E_{t,i}^{\boldsymbol{K}^{i},\boldsymbol{K}^{-i*}})^\top E_{t,i}^{\boldsymbol{K}^{i},\boldsymbol{K}^{-i*}}\right) \\
&\geq \frac{\underline{\sigma}_{\boldsymbol{X}}}{4\|\Sigma_{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}}\|^2} \sum_{t=0}^{T-1} \frac{\mathrm{Tr}\left(\nabla_{K_t^i} C^i(\boldsymbol{K}^i,\boldsymbol{K}^{-i*})^\top \nabla_{K_t^i} C^i(\boldsymbol{K}^i,\boldsymbol{K}^{-i*})\right)}{\|R_t^i + (B_t^i)^\top P_{t+1,i}^{\boldsymbol{K}^{i},\boldsymbol{K}^{-i*}} B_t^i\|}.
\end{aligned}
$$

■

As in the single-agent case (Fazel et al., 2018; Hambly et al., 2021), we now provide an expression for $C^i(\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i}) - C^i(\boldsymbol{K})$, which is easier to analyze.

Lemma 12 (Almost Smoothness) *For two sets of policies $\boldsymbol{K}$ and $(\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i})$, we have that*

$$
\begin{aligned}
C^i(\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i}) - C^i(\boldsymbol{K}) &= \sum_{t=0}^{T-1}\Big[\mathrm{Tr}\left(\Sigma_t^{\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i}}(K_t^{i\prime}-K_t^i)^\top(R_t^i+(B_t^i)^\top P_{t+1,i}^{\boldsymbol{K}}B_t^i)(K_t^{i\prime}-K_t^i)\right) \\
&\quad +2\,\mathrm{Tr}\left(\Sigma_t^{\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i}}(K_t^{i\prime}-K_t^i)^\top E_{t,i}^{\boldsymbol{K}}\right)\Big].
\end{aligned}
$$

Proof By Lemma 10,

$$
\begin{aligned}
C^i(\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i}) - C^i(\boldsymbol{K}) &= \mathbb{E}\left[\sum_{t=0}^{T-1} A_{\boldsymbol{K}}^i(x_t^{\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i}}, u_t^{i\prime,-i}, t)\right] \\
&= \sum_{t=0}^{T-1}\Big[\mathrm{Tr}\left(\Sigma_t^{\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i}}(K_t^{i\prime}-K_t^i)^\top(R_t^i+(B_t^i)^\top P_{t+1,i}^{\boldsymbol{K}}B_t^i)(K_t^{i\prime}-K_t^i)\right) \\
&\quad +2\,\mathrm{Tr}\left(\Sigma_t^{\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i}}(K_t^{i\prime}-K_t^i)^\top E_{t,i}^{\boldsymbol{K}}\right)\Big].
\end{aligned}
$$

■

Lemma 13 *Assume Assumptions 1 and 2 hold. Then for $t = 0, 1, \ldots, T$ and $i = 1, \ldots, N$*

$$
\left\|P_{t,i}^{\boldsymbol{K}}\right\| \leq \frac{C^i(\boldsymbol{K})}{\underline{\sigma}_{\boldsymbol{X}}}, \;\; \|\Sigma_{\boldsymbol{K}}\| \leq \frac{C^i(\boldsymbol{K})}{\underline{\sigma}_{\boldsymbol{Q}}}.
$$

Proof By the trace inequality (3.33), it is straightforward to see that

$$
C^i(\boldsymbol{K}) \geq \mathbb{E}[x_t^\top P_{t,i}^{\boldsymbol{K}} x_t] \geq \left\|P_{t,i}^{\boldsymbol{K}}\right\| \sigma_{\min}(\mathbb{E}[x_t x_t^\top]) \geq \underline{\sigma}_{\boldsymbol{X}} \left\|P_{t,i}^{\boldsymbol{K}}\right\|
$$

and

$$
C^i(\boldsymbol{K}) \;=\; \sum_{t=0}^{T-1} \mathrm{Tr}\left(\mathbb{E}[x_t x_t^\top](Q_t^i + (K_t^i)^\top R_t^i K_t^i)\right) + \mathrm{Tr}\left(\mathbb{E}[x_T x_T^\top] Q_T^i\right)
$$

$$\geq \min_{t\in[0,T]} \sigma_{\min}(Q_t^i)\cdot \mathrm{Tr}(\Sigma_{\boldsymbol{K}})$$
$$\geq \underline{\sigma}_{\boldsymbol{Q}} \cdot \|\Sigma_{\boldsymbol{K}}\|.$$

Then the statements in Lemma 13 follow since under Assumptions 1 and 2, we have $\underline{\sigma}_{\boldsymbol{X}} > 0$ and $\underline{\sigma}_{\boldsymbol{Q}} > 0$. ■

3.2 Perturbation Analysis of the State Covariance Matrix

Our aim in this section is to provide an explicit control of the change in the state covariance matrix after a change in policy $\boldsymbol{K}^i$. We begin by defining two linear operators on symmetric matrices. For $X \in \mathbb{R}^{d\times d}$ we set

$$\mathcal{F}_{K_t}(X) := \left(A_t - \sum_{j=1}^N B_t^j K_t^j\right) X \left(A_t - \sum_{j=1}^N B_t^j K_t^j\right)^\top, \tag{3.34}$$

and

$$\mathcal{T}_{\boldsymbol{K}}(X) := X + \sum_{t=0}^{T-1} \Pi_{i=0}^t \left(A_t - \sum_{j=1}^N B_t^j K_t^j\right) X \,\Pi_{i=0}^t \left(A_t - \sum_{j=1}^N B_t^j K_t^j\right)^\top.$$

If we write $\mathcal{G}_t^{\boldsymbol{K}} = \mathcal{F}_{K_t} \circ \mathcal{F}_{K_{t-1}} \circ \cdots \circ \mathcal{F}_{K_0}$, then the following relationships hold:

$$\mathcal{G}_t^{\boldsymbol{K}}(X) = \mathcal{F}_{K_t} \circ \mathcal{G}_{t-1}^{\boldsymbol{K}}(X) = \Pi_{i=0}^t \left(A_t - \sum_{j=1}^N B_t^j K_t^j\right) X \,\Pi_{i=0}^t \left(A_t - \sum_{j=1}^N B_t^j K_t^j\right)^\top, \tag{3.35}$$

and

$$\mathcal{T}_{\boldsymbol{K}}(X) = X + \sum_{t=0}^{T-1} \mathcal{G}_t^{\boldsymbol{K}}(X). \tag{3.36}$$

When the policy $\boldsymbol{K}$ is clear we will write $\mathcal{G}_t = \mathcal{G}_t^{\boldsymbol{K}}$ and $\mathcal{F}_t = \mathcal{F}_{K_t}$.

We also define the induced norm for these operators as

$$\|T\| := \sup_X \frac{\|T(X)\|}{\|X\|}, \tag{3.37}$$

where $T = \mathcal{F}_{K_t}, \mathcal{G}_t^{\boldsymbol{K}}, \mathcal{T}_{\boldsymbol{K}}$ and the supremum is over all symmetric matrix X with non-zero spectral norm.

We first show the relationship between the operator $\mathcal{T}_{\boldsymbol{K}}$ and the quantity $\Sigma_{\boldsymbol{K}}$.

Proposition 1 *For $T \geq 2$, we have that*

$$\Sigma_{\boldsymbol{K}} = \mathcal{T}_{\boldsymbol{K}}(\Sigma_0) + \Delta_{\boldsymbol{K}}(W), \tag{3.38}$$

where

$$\Delta_{\boldsymbol{K}}(W) = \sum_{t=1}^{T-1}\sum_{s=1}^{t} D_{t,s} W D_{t,s}^\top + T\,W,$$

with $D_{t,s} = \Pi_{u=s}^t (A_u - \sum_{j=1}^N B_u^j K_u^j)$ (for $s = 1, 2, \cdots, t$), and $\Sigma_0 = \mathbb{E}\left[x_0 x_0^\top\right]$.

Proof Recall that $\Sigma_t^{\boldsymbol{K}} = \mathbb{E}\left[x_t x_t^\top\right]$ and note that

$$\begin{aligned}
\Sigma_1^{\boldsymbol{K}} &= \mathbb{E}[x_1 x_1^\top] = \mathbb{E}\left[\left(\left(A_0 - \sum_{j=1}^N B_0^j K_0^j\right) x_0 + w_0\right)\left(\left(A_0 - \sum_{j=1}^N B_0^j K_0^j\right) x_0 + w_0\right)^\top\right] \\
&= \left(A_0 - \sum_{j=1}^N B_0^j K_0^j\right) \Sigma_0 \left(A_0 - \sum_{j=1}^N B_0^j K_0^j\right)^\top + W = \mathcal{G}_0(\Sigma_0) + W.
\end{aligned}$$

We first prove that for $t = 2, 3, \cdots, T$

$$\Sigma_t^{\boldsymbol{K}} = \mathcal{G}_{t-1}(\Sigma_0) + \sum_{s=1}^{t-1} D_{t-1,s} W D_{t-1,s}^\top + W. \tag{3.39}$$

We have the result for $t = 1$, so assume (3.39) holds for $t \leq k$. Then for $t = k+1$,

$$\begin{aligned}
\mathbb{E}[x_{t+1} x_{t+1}^\top] &= \mathbb{E}\left[\left(\left(A_t - \sum_{j=1}^N B_t^j K_t^j\right) x_t + w_t\right)\left(\left(A_t - \sum_{j=1}^N B_t^j K_t^j\right) x_t + w_t\right)^\top\right] \\
&= \left(A_t - \sum_{j=1}^N B_t^j K_t^j\right) \Sigma_t^{\boldsymbol{K}} \left(A_t - \sum_{j=1}^N B_t^j K_t^j\right)^\top + W \\
&= \mathcal{G}_t(\Sigma_0) + \sum_{s=1}^{t} D_{t,s} W D_{t,s}^\top + W.
\end{aligned}$$

Therefore (3.39) holds, $\forall\, t = 1, 2, \cdots, T$. Finally,

$$\Sigma_{\boldsymbol{K}} = \sum_{t=0}^T \Sigma_t^{\boldsymbol{K}} = \Sigma_0 + \sum_{t=0}^{T-1} \mathcal{G}_t(\Sigma_0) + \sum_{t=1}^{T-1}\sum_{s=1}^{t} D_{t,s} W D_{t,s}^\top + TW = \mathcal{T}_{\boldsymbol{K}}(\Sigma_0) + \Delta_{\boldsymbol{K}}(W).$$

■

Given two policies $\boldsymbol{K}$ and $\boldsymbol{K'} = (\boldsymbol{K}^{1\prime}, \cdots, \boldsymbol{K}^{N\prime})$, let us define

$$\begin{aligned}
\rho_{\boldsymbol{K},\boldsymbol{K'}} := \max\Bigg\{ &\max_i \left\{ \max_{0\leq t\leq T-1} \left\| A_t - B_t^i K_t^i - \sum_{j=1, j\neq i}^N B_t^j K_t^{j*} \right\| \right\}, \max_{0\leq t\leq T-1} \left\| A_t - \sum_{i=1}^N B_t^i K_t^i \right\|, \\
&\max_i \left\{ \max_{0\leq t\leq T-1} \left\| A_t - B_t^i K_t^{i\prime} - \sum_{j=1, j\neq i}^N B_t^j K_t^{j*} \right\| \right\}, \\
&\max_i \left\{ \max_{0\leq t\leq T-1} \left\| A_t - B_t^i K_t^{i\prime} - \sum_{j=1, j\neq i}^N B_t^j K_t^{j} \right\| \right\}, 1 + \xi \Bigg\},
\end{aligned} \tag{3.40}$$

for some small constant $\xi > 0$.

For any given policy $\boldsymbol{K}$, $\max_{0\le t\le T-1}\left\|A_t - \sum_{j=1}^N B_t^j K_t^j\right\|$ measures the radius of the state dynamics under policy $\boldsymbol{K}$. Thus $\rho_{\boldsymbol{K},\boldsymbol{K}'}$ defined in (3.40) is the maximum radius of the policies $\boldsymbol{K}$, $(\boldsymbol{K}^i,\boldsymbol{K}^{-i*})$, $(\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i*})$, and $(\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i})$ $(i=1,2,\cdots,N)$. We will show later that the value (or the upper bound) of $\rho_{\boldsymbol{K},\boldsymbol{K}'}$ plays an essential role in the convergence analysis.

Remark 14 By the definition of $\rho_{\boldsymbol{K},\boldsymbol{K}'}$ in (3.40), we have $\rho_{\boldsymbol{K},\boldsymbol{K}'} \ge 1+\xi > 1$. This regularization term $1+\xi$ is introduced to simplify the presentation. Alternatively, if we remove this term from the definition of $\rho_{\boldsymbol{K},\boldsymbol{K}'}$, a similar analysis can still be carried out by considering the different cases: $\rho_{\boldsymbol{K},\boldsymbol{K}'} < 1$, $\rho_{\boldsymbol{K},\boldsymbol{K}'} = 1$ and $\rho_{\boldsymbol{K},\boldsymbol{K}'} > 1$.

We now provide an upper bound for $\rho_{\boldsymbol{K},\boldsymbol{K}'}$.

Lemma 15 *Assume Assumption 3 holds. Then,*

$$\rho_{\boldsymbol{K},\boldsymbol{K}'} \le \rho^* + N\gamma_B\sqrt{\frac{T}{\underline{\sigma}_{\boldsymbol{X}}\,\underline{\sigma}_{\boldsymbol{R}}}\max_i\left\{C^i(\boldsymbol{K}^i,\boldsymbol{K}^{-i*}) - C^i(\boldsymbol{K}^*)\right\}} + \gamma_B\max_i\max_t\left\{\|K_t^{i\prime} - K_t^i\|\right\} \tag{3.41}$$

where ρ^* was defined in (3.9).

Proof By Lemma 12, we have

$$\begin{aligned}
&C^i(\boldsymbol{K}^i,\boldsymbol{K}^{-i*}) - C^i(\boldsymbol{K}^*)\\
=\;& \sum_{t=0}^{T-1}\Big[\mathrm{Tr}\left(\Sigma_t^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}}(K_t^i - K_t^{i*})^\top(R_t^i + (B_t^i)^\top P_{t+1,i}^{\boldsymbol{K}^*}B_t^i)(K_t^i - K_t^{i*})\right)\\
\ge\;& \underline{\sigma}_{\boldsymbol{X}}\,\underline{\sigma}_{\boldsymbol{R}}\sum_{t=0}^{T-1}\|K_t^i - K_t^{i*}\|^2 \ge \frac{\underline{\sigma}_{\boldsymbol{X}}\,\underline{\sigma}_{\boldsymbol{R}}}{T}\left|\!\left|\!\left|\boldsymbol{K}^i - \boldsymbol{K}^{i*}\right|\!\right|\!\right|^2,
\end{aligned} \tag{3.42}$$

where (3.42) holds by the Cauchy-Schwarz inequality. Then we have

$$\begin{aligned}
\left\|A_t - \sum_{j=1}^N B_t^j K_t^j\right\| &\le \left\|A_t - \sum_{j=1}^N B_t^j K_t^{j*}\right\| + \sum_{j=1}^N \|B_t^j\|\,\|K_t^j - K_t^{j*}\|\\
&\le \left\|A_t - \sum_{j=1}^N B_t^j K_t^{j*}\right\| + \gamma_B\sum_{j=1}^N \left|\!\left|\!\left|\boldsymbol{K}^j - \boldsymbol{K}^{j*}\right|\!\right|\!\right|\\
&\le \left\|A_t - \sum_{j=1}^N B_t^j K_t^{j*}\right\| + N\gamma_B\sqrt{\frac{T}{\underline{\sigma}_{\boldsymbol{X}}\,\underline{\sigma}_{\boldsymbol{R}}}\max_j\{C^j(\boldsymbol{K}^j,\boldsymbol{K}^{-j*}) - C^j(\boldsymbol{K}^*)\}},
\end{aligned} \tag{3.43}$$

where (3.43) holds by (3.42). Also, by the triangle inequality we have

$$\left\|A_t - \sum_{j=1,j\ne i}^N B_t^j K_t^j - B_t^i K_t^{i\prime}\right\| \le \left\|A_t - \sum_{j=1}^N B_t^j K_t^j\right\| + \gamma_B\|K_t^{i\prime} - K_t^i\|. \tag{3.44}$$

and

$$\left\|A_t - \sum_{j=1,j\neq i}^{N} B_t^j K_t^{j*} - B_t^i K_t^{i\prime}\right\| \leq \left\|A_t - \sum_{j=1,j\neq i}^{N} B_t^j K_t^{j*} - B_t^i K_t^i\right\| + \gamma_B \|K_t^{i\prime} - K_t^i\|. \quad (3.45)$$

Finally, applying (3.42),

$$\begin{aligned}\left\|A_t - \sum_{j=1,j\neq i}^{N} B_t^j K_t^{j*} - B_t^i K_t^i\right\| &\leq \left\|A_t - \sum_{j=1}^{N} B_t^j K_t^{j*}\right\| + \gamma_B \|K_t^{i*} - K_t^i\| \\ &\leq \left\|A_t - \sum_{j=1}^{N} B_t^j K_t^{j*}\right\| \\ &\quad + \gamma_B \sqrt{\frac{T}{\underline{\sigma}_{\boldsymbol{X}}\,\underline{\sigma}_{\boldsymbol{R}}}\big(C^i(\boldsymbol{K}^i, \boldsymbol{K}^{-i*}) - C^i(\boldsymbol{K}^*)\big)}. \quad (3.46)\end{aligned}$$

Therefore combining (3.43)-(3.46), we obtain the statement (3.41). ■

Recall the definition of $\mathcal{F}_t = \mathcal{F}_{K_t}$ and $\mathcal{G}_t = \mathcal{G}_t^{\boldsymbol{K}}$ in (3.34) and (3.35) associated with $\boldsymbol{K}$. Similarly let us define $\mathcal{G}_t^{i\prime} = \mathcal{F}_{K_t^{i\prime}, K_t^{-i}} \circ \mathcal{F}_{K_{t-1}^{i\prime}, K_{t-1}^{-i}} \circ \cdots \circ \mathcal{F}_{K_0^{i\prime}, K_0^{-i}}$ for the set of policies $(\boldsymbol{K}^{i\prime}, \boldsymbol{K}^{-i})$ and write $\mathcal{F}_t^{i\prime} = \mathcal{F}_{K_t^{i\prime}, K_t^{-i}}$. We now establish a perturbation analysis for $\mathcal{F}_t$ and $\mathcal{G}_t$.

Lemma 16 *For all $t = 0, 1, \cdots, T-1$, we have*

$$\|\mathcal{F}_t - \mathcal{F}_t^{i\prime}\| \leq 2\rho_{\boldsymbol{K},\boldsymbol{K}'}\gamma_B \|K_t^i - K_t^{i\prime}\|. \quad (3.47)$$

Proof For player i,

$$\begin{aligned}&(\mathcal{F}_t - \mathcal{F}_t^{i\prime})(X) \\ = &\left(A_t - \sum_{j=1}^{N} B_t^j K_t^j\right) X \left(A_t - \sum_{j=1}^{N} B_t^j K_t^j\right)^\top \\ &- \left(A_t - B_t^i K_t^{i\prime} - \sum_{j=1,j\neq i}^{N} B_t^j K_t^j\right) X \left(A_t - B_t^i K_t^{i\prime} - \sum_{j=1,j\neq i}^{N} B_t^j K_t^j\right)^\top,\end{aligned}$$

By (3.37) the operator norm of $\mathcal{F}_t - \mathcal{F}_t^{i\prime}$ is the maximum possible ratio of $\|(\mathcal{F}_t - \mathcal{F}_t^{i\prime})(X)\|$ and $\|X\|$. Then letting $Y = A_t - \sum_{j=1}^N B_t^j K_t^j$ and $Z = A_t - B_t^i K_t^{i\prime} - \sum_{j=1,j\neq i}^N B_t^j K_t^j$ in

$$YXY^\top - ZXZ^\top = \frac{(Y+Z)X(Y-Z)^\top + (Y-Z)X(Y+Z)^\top}{2} \quad (3.48)$$

and using the norm bound $\|AX\| \leq \|A\|\,\|X\|$, we have

$$\Bigg\| \left(A_t - \sum_{j=1}^{N} B_t^j K_t^j\right) X \left(A_t - \sum_{j=1}^{N} B_t^j K_t^j\right)^\top$$

$$
\begin{aligned}
& - \left(A_t - B_t^i K_t^{i\prime} - \sum_{j=1, j\neq i}^{N} B_t^j K_t^j \right) X \left(A_t - B_t^i K_t^{i\prime} - \sum_{j=1, j\neq i}^{N} B_t^j K_t^j \right)^\top \Bigg\| \\
\leq \quad & 2\rho_{\boldsymbol{K},\boldsymbol{K}'} \|X\| \, \|B_t^i (K_t^i - K_t^{i'})\| \leq 2\rho_{\boldsymbol{K},\boldsymbol{K}'} \|X\| \gamma_B \|(K_t^i - K_t^{i'})\|.
\end{aligned}
$$

Therefore we obtain the statement $\|\mathcal{F}_t - \mathcal{F}_t^{i\prime}\| \leq 2\rho_{\boldsymbol{K},\boldsymbol{K}'}\gamma_B \|(K_t^i - K_t^{i'})\|$ ■

Recall that $\mathcal{F}_t$ and $\mathcal{G}_t$ are defined in equations (3.34) and (3.35). Then we have the following lemma on perturbation analysis for $\mathcal{G}_t$.

Lemma 17 (Perturbation Analysis for $\mathcal{G}_t$) *For any symmetric matrix $\Sigma \in \mathbb{R}^{d\times d}$ and $i = 1, \cdots, N$, we have that*

$$
\sum_{t=0}^{T-1} \left\| (\mathcal{G}_t - \mathcal{G}_t^{i\prime})(\Sigma) \right\| \leq \frac{\rho_{\boldsymbol{K},\boldsymbol{K}'}^{2T} - 1}{\rho_{\boldsymbol{K},\boldsymbol{K}'}^{2} - 1} \Big(\sum_{t=0}^{T-1} \|\mathcal{F}_t - \mathcal{F}_t^{i\prime}\| \Big) \|\Sigma\|.
$$

Proof By direct calculation,

$$
\|\mathcal{G}_t^{i\prime}\| \leq \rho_{\boldsymbol{K},\boldsymbol{K}'}^{2(t+1)}, \quad i = 1, \cdots, N. \tag{3.49}
$$

Then for any symmetric matrix $\Sigma \in \mathbb{R}^{d\times d}$ and $t \geq 0$,

$$
\begin{aligned}
\|(\mathcal{G}_{t+1}^{i\prime} - \mathcal{G}_{t+1})(\Sigma)\| &= \|\mathcal{F}_{t+1}^{i\prime} \circ \mathcal{G}_t^{i\prime}(\Sigma) - \mathcal{F}_{t+1} \circ \mathcal{G}_t(\Sigma)\| \\
&= \|\mathcal{F}_{t+1}^{i\prime} \circ \mathcal{G}_t^{i\prime}(\Sigma) - \mathcal{F}_{t+1}^{i\prime} \circ \mathcal{G}_t(\Sigma) + \mathcal{F}_{t+1}^{i\prime} \circ \mathcal{G}_t(\Sigma) - \mathcal{F}_{t+1} \circ \mathcal{G}_t(\Sigma)\| \\
&\leq \|\mathcal{F}_{t+1}^{i\prime} \circ \mathcal{G}_t^{i\prime}(\Sigma) - \mathcal{F}_{t+1}^{i\prime} \circ \mathcal{G}_t(\Sigma)\| + \|\mathcal{F}_{t+1}^{i\prime} \circ \mathcal{G}_t(\Sigma) - \mathcal{F}_{t+1} \circ \mathcal{G}_t(\Sigma)\| \\
&= \|\mathcal{F}_{t+1}^{i\prime} \circ (\mathcal{G}_t^{i\prime} - \mathcal{G}_t)(\Sigma)\| + \|(\mathcal{F}_{t+1}^{i\prime} - \mathcal{F}_{t+1}) \circ \mathcal{G}_t(\Sigma)\| \\
&\leq \|\mathcal{F}_{t+1}^{i\prime}\| \, \|(\mathcal{G}_t^{i\prime} - \mathcal{G}_t)(\Sigma)\| + \|\mathcal{G}_t\| \, \|\mathcal{F}_{t+1}^{i\prime} - \mathcal{F}_{t+1}\| \, \|\Sigma\| \\
&\leq \rho_{\boldsymbol{K},\boldsymbol{K}'}^{2} \|(\mathcal{G}_t^{i\prime} - \mathcal{G}_t)(\Sigma)\| + \rho_{\boldsymbol{K},\boldsymbol{K}'}^{2(t+1)} \|\mathcal{F}_{t+1}^{i\prime} - \mathcal{F}_{t+1}\| \|\Sigma\|.
\end{aligned}
$$

Therefore,

$$
\|(\mathcal{G}_{t+1}^{i\prime} - \mathcal{G}_{t+1})(\Sigma)\| \leq \rho_{\boldsymbol{K},\boldsymbol{K}'}^{2} \|(\mathcal{G}_t^{i\prime} - \mathcal{G}_t)(\Sigma)\| + \rho_{\boldsymbol{K},\boldsymbol{K}'}^{2(t+1)} \|\mathcal{F}_{t+1}^{i\prime} - \mathcal{F}_{t+1}\| \|\Sigma\|. \tag{3.50}
$$

As it is a geometric series, summing (3.50) over $t \in \{0, 1, 2, \cdots, T-2\}$ with $\|\mathcal{G}_0^{i\prime} - \mathcal{G}_0\| = \|\mathcal{F}_0^{i\prime} - \mathcal{F}_0\|$, gives

$$
\sum_{t=0}^{T-1} \left\| (\mathcal{G}_t - \mathcal{G}_t^{i\prime})(\Sigma) \right\| \leq \frac{\rho_{\boldsymbol{K},\boldsymbol{K}'}^{2T} - 1}{\rho_{\boldsymbol{K},\boldsymbol{K}'}^{2} - 1} \Big(\sum_{t=0}^{T-1} \|\mathcal{F}_t - \mathcal{F}_t^{i\prime}\| \Big) \|\Sigma\|.
$$

■

Recall that γ_A, γ_B, and γ_R are defined in (3.4). Then we have the following perturbation analysis of $\Sigma_{\boldsymbol{K}}$.

Lemma 18 (Perturbation Analysis of $\Sigma_{\boldsymbol{K}}$) *Assume Assumption 1 holds. Then*

$$\left\|\Sigma_{\boldsymbol{K}} - \Sigma_{\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i}}\right\| \leq 2\gamma_B \,\frac{\rho_{\boldsymbol{K},\boldsymbol{K}'}(\rho^{2T}_{\boldsymbol{K},\boldsymbol{K}'}-1)}{\rho^{2}_{\boldsymbol{K},\boldsymbol{K}'}-1}\left(\frac{C^i(\boldsymbol{K}^i,\boldsymbol{K}^{-i*})}{\underline{\sigma}_{\boldsymbol{Q}}} + T\|W\|\right)|||\boldsymbol{K}^i - \boldsymbol{K}^{i\prime}|||.$$

Proof Using Lemma 16,

$$\sum_{t=0}^{T-1}\|\mathcal{F}_t - \mathcal{F}_t^{i\prime}\| \;\;\leq\;\; 2\rho_{\boldsymbol{K},\boldsymbol{K}'}\gamma_B\sum_{t=0}^{T-1}\|K_t^i - K_t^{i\prime}\|.$$

Define $D^{i\prime}_{t,s} = \Pi_{u=s}^{t}(A_u - B_u^i K_u^{i\prime} - \sum_{j=1,j\neq i}^{N} B_u^j K_u^j)$ (for $s = 1, 2, \cdots, t$). Then, in a similar way to the proof of Lemma 17, we have, $\forall\, t = 1, \cdots, T-1$,

$$\sum_{s=1}^{t}\left\|D_{t,s}WD_{t,s}^{\top} - D^{i\prime}_{t,s}W(D^{i\prime}_{t,s})^{\top}\right\| \leq \frac{\rho^{2T}_{\boldsymbol{K},\boldsymbol{K}'}-1}{\rho^{2}_{\boldsymbol{K},\boldsymbol{K}'}-1}\left(\sum_{s=0}^{t}\|\mathcal{F}_s - \mathcal{F}_s^{i\prime}\|\right)\|W\|. \tag{3.51}$$

By Proposition 1, (3.36) and (3.51), we have

$$\begin{aligned}\left\|\Sigma_{\boldsymbol{K}} - \Sigma_{\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i}}\right\| &\leq \left\|(\mathcal{T}_{\boldsymbol{K}} - \mathcal{T}_{\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i}})(\Sigma_0)\right\| + \sum_{t=1}^{T-1}\sum_{s=1}^{t}\left\|D_{t,s}WD_{t,s}^{\top} - D^{i\prime}_{t,s}W(D^{i\prime}_{t,s})^{\top}\right\| \\ &\leq \frac{\rho^{2T}_{\boldsymbol{K},\boldsymbol{K}'}-1}{\rho^{2}_{\boldsymbol{K},\boldsymbol{K}'}-1}\Big(\sum_{t=0}^{T-1}\|\mathcal{F}_t - \mathcal{F}_t^{i\prime}\|\Big)\left(\|\Sigma_0\| + T\|W\|\right) \\ &\leq \frac{\rho^{2T}_{\boldsymbol{K},\boldsymbol{K}'}-1}{\rho^{2}_{\boldsymbol{K},\boldsymbol{K}'}-1}\left(\frac{C^i(\boldsymbol{K}^i,\boldsymbol{K}^{-i*})}{\underline{\sigma}_{\boldsymbol{Q}}} + T\|W\|\right)\left(2\rho_{\boldsymbol{K},\boldsymbol{K}'}\gamma_B\,|||\boldsymbol{K}^i - \boldsymbol{K}^{i\prime}|||\right).\end{aligned} \tag{3.52}$$

The last inequality holds since $\|\Sigma_0\| \leq \|\Sigma_{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}}\| \leq \frac{C^i(\boldsymbol{K}^i,\boldsymbol{K}^{-i*})}{\underline{\sigma}_{\boldsymbol{Q}}}$ by Lemma 13. ■

3.3 Convergence and Complexity Analysis

We are now in a position to provide the proof of our main Theorem 6. This will follow from two important Lemmas. First, define

$$\rho_{\boldsymbol{K}} := \rho^* + N\gamma_B\sqrt{\frac{T}{\underline{\sigma}_{\boldsymbol{X}}\,\underline{\sigma}_{\boldsymbol{R}}}\max_i\left\{C^i(\boldsymbol{K}^i,\boldsymbol{K}^{-i*}) - C^i(\boldsymbol{K}^*)\right\}} + \frac{1}{20T^2}. \tag{3.53}$$

Further define g_1 and g_2 as follows

$$g_1 := \frac{\underline{\sigma}_{\boldsymbol{R}}}{\|\Sigma_{\boldsymbol{K}^*}\|}, \tag{3.54}$$

and

$$g_2 := 20(N-1)^2\,T^2\,d\,\frac{(\gamma_B)^4\max_i\{C^i(\boldsymbol{K}^i,\boldsymbol{K}^{-i*})\}^4}{\underline{\sigma}_{\boldsymbol{Q}}{}^2\,\underline{\sigma}_{\boldsymbol{R}}}\left(\frac{\rho^{2T}_{\boldsymbol{K}}-1}{\rho^{2}_{\boldsymbol{K}}-1}\right)^2. \tag{3.55}$$

We also write $C^{i,-i*} = C^i(\boldsymbol{K}^i,\boldsymbol{K}^{-i*})$, $C^{i*} = C^i(\boldsymbol{K}^*)$ and $C^{i\prime,-i*} = C^i(\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i*})$ to simplify notation.

Lemma 19 (One-step Contraction) *Assume Assumptions 1, 2, and 3 hold, and that*

$$\underline{\sigma}_{\boldsymbol{X}}{}^{5} > \frac{g_2}{g_1}. \tag{3.56}$$

Also assume the policy update step for player i at time t is given by

$$K_t^{i\prime} = K_t^i - \eta \nabla_{K_t^i} C^i(\boldsymbol{K})(\Sigma_t^{\boldsymbol{K}})^{-1}, \tag{3.57}$$

where

$$\eta \leq \min\left\{ I_1\,, I_2\,, \frac{1}{\underline{\sigma}_{\boldsymbol{R}}} \right\} \tag{3.58}$$

with

$$\begin{aligned} I_1 = &\left\{ 20T \frac{\rho_{\boldsymbol{K}}(\rho_{\boldsymbol{K}}^{2T}-1)}{\rho_{\boldsymbol{K}}^2 - 1} \left(\sum_{i=1}^N C^{i,-i*} + \underline{\sigma}_{\boldsymbol{Q}}\, T \|W\| \right) \gamma_B \max_i \{ \max_t \{ \|\nabla_{K_t^i} C^i(\boldsymbol{K})\| \} \} + \underline{\sigma}_{\boldsymbol{Q}} (\underline{\sigma}_{\boldsymbol{X}})^2 \right. \\ &\left. + 4\big(\gamma_R\, \underline{\sigma}_{\boldsymbol{X}} + (\gamma_B)^2 \sum_{i=1}^N C^{i,-i*}\big) \sum_{i=1}^N \{ C^{i,-i*} \} \right\}^{-1} \cdot \underline{\sigma}_{\boldsymbol{Q}} (\underline{\sigma}_{\boldsymbol{X}})^2, \\ I_2 = &\left\{ \big(\max_i \{k_i\}\big) \frac{10T \sum_{i=1}^N C^{i,-i*}}{(10T-1)\, \underline{\sigma}_{\boldsymbol{Q}}} \left(\gamma_R + (\gamma_B)^2 \frac{\sum_{i=1}^N C^{i,-i*}}{\underline{\sigma}_{\boldsymbol{X}}} \right) + \frac{2d}{\underline{\sigma}_{\boldsymbol{X}}} \left(\frac{10T \sum_{i=1}^N C^{i,-i*}}{(10T-1)\, \underline{\sigma}_{\boldsymbol{Q}}} \right)^2 \cdot \right. \\ &\left. \left(\gamma_R + (\gamma_B)^2 \frac{\sum_{i=1}^N C^{i,-i*}}{\underline{\sigma}_{\boldsymbol{X}}} \right)^2 \right\}^{-1} \cdot \frac{d}{80\, \underline{\sigma}_{\boldsymbol{X}}} \left(\frac{10T \min_i \{C^{i*}\}}{(10T-1)\, \underline{\sigma}_{\boldsymbol{Q}}} \right)^2. \end{aligned}$$

Let $\alpha := \underline{\sigma}_{\boldsymbol{X}}\, g_1 - g_2 / \underline{\sigma}_{\boldsymbol{X}}{}^4 > 0$. Then, we have

1. *$\eta \in (0, \frac{1}{\alpha})$; and*
2. *the following inequality holds*

$$\sum_{i=1}^N \big(C^i(\boldsymbol{K}^{i\prime}, \boldsymbol{K}^{-i*}) - C^i(\boldsymbol{K}^*) \big) \leq (1 - \alpha\eta) \left(\sum_{i=1}^N \big(C^i(\boldsymbol{K}^i, \boldsymbol{K}^{-i*}) - C^i(\boldsymbol{K}^*) \big) \right). \tag{3.59}$$

Remark 20 (1) In the one-step contraction analysis (Lemma 19), (3.56) imposes a condition on $C(\boldsymbol{K}^i, \boldsymbol{K}^{-i*})$ which is associated with the current policy $\boldsymbol{K}$. In the analysis of the global convergence result (Theorem 6), (3.14) imposes a similar condition on $C(\boldsymbol{K}^{i,(0)}, \boldsymbol{K}^{-i*})$ which is associated with the initial policy $\boldsymbol{K}^{(0)}$. Condition (3.14) in Assumption 4 and the step size condition in Theorem 6 guarantee that, condition (3.56) holds for any $\boldsymbol{K} = \boldsymbol{K}^{(m)}$ throughout the training process ($m = 1, 2, \cdots, M$). This further ensures that the one-step contraction analysis in Lemma 19 can be applied iteratively which leads to the global convergence result as stated in Theorem 6.
(2) Note that the numbers such as 20 and 80 that appear in I_1, I_2 are not arbitrary and, although some minor improvements can be made by optimizing at various stages, they enable us to obtain reasonable bounds.

Proof We break this proof up into a series of steps.

Step 1: We first consider the consequences of the condition $\eta \leq \min\{I_1, I_2\}$. Straightforward calculations show that when condition $\eta \leq I_1$ is satisfied, the following inequalities hold:

1. $\forall i = 1, \cdots, N$,

$$\|K_t^{i\prime} - K_t^i\| = \eta \|\nabla_{K_t^i} C^i(\boldsymbol{K}) \Sigma_t^{-1}\| \leq \frac{\underline{\sigma}_{\boldsymbol{Q}}\, \underline{\sigma}_{\boldsymbol{X}}}{20T\gamma_B C^{i,-i*}}. \tag{3.60}$$

2. $\forall i = 1, \cdots, N$,

$$\begin{aligned} & \eta \left(\frac{\rho_{\boldsymbol{K}}^{2T} - 1}{\rho_{\boldsymbol{K}}^2 - 1} \left(\frac{C^{i,-i*}}{\underline{\sigma}_{\boldsymbol{Q}}} + T\|W\| \right) 2\rho_{\boldsymbol{K}} \frac{\gamma_B}{\underline{\sigma}_{\boldsymbol{X}}} \sum_{t=0}^{T-1} \|\nabla_{K_t^i} C^i(\boldsymbol{K})\| \right) \\ \leq \; & I_1 \left(\frac{\rho_{\boldsymbol{K}}^{2T} - 1}{\rho_{\boldsymbol{K}}^2 - 1} \left(\frac{\sum_{i=1}^N C^{i,-i*}}{\underline{\sigma}_{\boldsymbol{Q}}} + T\|W\| \right) 2\rho_{\boldsymbol{K}} \frac{\gamma_B}{\underline{\sigma}_{\boldsymbol{X}}} T \max_t \{\|\nabla_{K_t^i} C^i(\boldsymbol{K})\|\} \right) \\ \leq \; & \frac{\underline{\sigma}_{\boldsymbol{X}}}{10}. \end{aligned} \tag{3.61}$$

3. $\forall i = 1, \cdots, N$,

$$\eta \leq I_1 \quad \leq \quad \frac{\underline{\sigma}_{\boldsymbol{X}}}{\underline{\sigma}_{\boldsymbol{X}} + 2\frac{2C^{i,-i*}}{\underline{\sigma}_{\boldsymbol{Q}}}(\gamma_R + \gamma_B^2 \frac{C^{i,-i*}}{\underline{\sigma}_{\boldsymbol{X}}})}. \tag{3.62}$$

In the case where $\eta \leq I_2$, we have $\forall i = 1, \cdots, N$

$$\begin{aligned} & 4\eta\, k_i \frac{10TC^{i,-i*}}{(10T-1)\,\underline{\sigma}_{\boldsymbol{Q}}} \left(\gamma_R + (\gamma_B)^2 \frac{C^{i,-i*}}{\underline{\sigma}_{\boldsymbol{X}}}\right) + 8\eta \frac{d}{\underline{\sigma}_{\boldsymbol{X}}} \left(\frac{10TC^{i,-i*}}{(10T-1)\,\underline{\sigma}_{\boldsymbol{Q}}} \right)^2 \left(\gamma_R + (\gamma_B)^2 \frac{C^{i,-i*}}{\underline{\sigma}_{\boldsymbol{X}}}\right)^2 \\ \leq \; & 4I_2 \left(k_i \frac{10TC^{i,-i*}}{(10T-1)\,\underline{\sigma}_{\boldsymbol{Q}}} \left(\gamma_R + (\gamma_B)^2 \frac{C^{i,-i*}}{\underline{\sigma}_{\boldsymbol{X}}}\right) \right. \\ & \left. \qquad + 2\frac{d}{\underline{\sigma}_{\boldsymbol{X}}} \left(\frac{10TC^{i,-i*}}{(10T-1)\,\underline{\sigma}_{\boldsymbol{Q}}} \right)^2 \left(\gamma_R + (\gamma_B)^2 \frac{C^{i,-i*}}{\underline{\sigma}_{\boldsymbol{X}}}\right)^2 \right) \\ \leq \; & \frac{d}{20\,\underline{\sigma}_{\boldsymbol{X}}} \left(\frac{10T \min_i\{C^{i*}\}}{(10T-1)\,\underline{\sigma}_{\boldsymbol{Q}}} \right)^2 \leq \frac{d}{20\,\underline{\sigma}_{\boldsymbol{X}}} \left(\frac{10T \max_i\{C^{i,-i*}\}}{(10T-1)\,\underline{\sigma}_{\boldsymbol{Q}}} \right)^2. \end{aligned} \tag{3.63}$$

By (3.60) and Lemma 13 we have

$$\gamma_B \|K_t^{i\prime} - K_t^i\| \leq \frac{\underline{\sigma}_{\boldsymbol{Q}}\, \underline{\sigma}_{\boldsymbol{X}}}{20TC^{i,-i*}} \leq \frac{1}{20T^2}.$$

Therefore, we have $\rho_{\boldsymbol{K},\boldsymbol{K}'} \leq \rho_{\boldsymbol{K}}$ by Lemma 15.

Step 2: We bound the norm of the state covariance matrix. By Lemma 16,

$$\sum_{t=0}^{T-1} \|\mathcal{F}_{K_t^i, K_t^{-i*}} - \mathcal{F}_{K_t^{i\prime}, K_t^{-i*}}\| \leq 2\rho_{\boldsymbol{K}} \gamma_B \left(\sum_{t=0}^{T-1} \|K_t^i - K_t^{i\prime}\| \right),$$

and hence by Lemma 18, we have

$$\begin{aligned}\left\|\Sigma_{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}}-\Sigma_{\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i*}}\right\| &\leq \frac{\rho_{\boldsymbol{K}}^{2T}-1}{\rho_{\boldsymbol{K}}^{2}-1}\left(\sum_{t=0}^{T-1}\|\mathcal{F}_{K_t^i,K_t^{-i*}}-\mathcal{F}_{K_t^{i\prime},K_t^{-i*}}\|\right)(\|\Sigma_0\|+T\|W\|)\\ &\leq \frac{\rho_{\boldsymbol{K}}^{2T}-1}{\rho_{\boldsymbol{K}}^{2}-1}\left(\frac{C^{i,-i*}}{\underline{\sigma}_{\boldsymbol{Q}}}+T\|W\|\right)2\rho_{\boldsymbol{K}}\gamma_B\left\|\!\left|\boldsymbol{K}^i-\boldsymbol{K}^{i\prime}\right|\!\right\|\\ &\leq \frac{\rho_{\boldsymbol{K}}^{2T}-1}{\rho_{\boldsymbol{K}}^{2}-1}\left(\frac{C^{i,-i*}}{\underline{\sigma}_{\boldsymbol{Q}}}+T\|W\|\right)2\rho_{\boldsymbol{K}}\gamma_B\,\frac{\eta}{\underline{\sigma}_{\boldsymbol{X}}}\sum_{t=0}^{T-1}\|\nabla_{K_t^i}C^i(\boldsymbol{K})\|\\ &\leq \frac{\underline{\sigma}_{\boldsymbol{X}}}{10}, \qquad (3.64)\end{aligned}$$

where the last inequality holds by (3.61) when $\sum_{t=0}^{T-1}\|\nabla_{K_t^i}C^i(\boldsymbol{K}\|>0$. Note that when $\sum_{t=0}^{T-1}\|\nabla_{K_t^i}C^i(\boldsymbol{K})\|=0$, we have $\left\|\Sigma_{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}}-\Sigma_{\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i*}}\right\|=0<\frac{\underline{\sigma}_{\boldsymbol{X}}}{10}$ and hence (3.64) still holds in this case. Therefore, by Lemma 13, and noting $\underline{\sigma}_{\boldsymbol{X}}\leq\|\Sigma_{\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i*}}\|/T$,

$$\begin{aligned}\left\|\Sigma_{\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i*}}\right\| &\leq \left\|\Sigma_{\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i*}}-\Sigma_{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}}\right\|+\left\|\Sigma_{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}}\right\|\leq\frac{\underline{\sigma}_{\boldsymbol{X}}}{10}+\frac{C^{i,-i*}}{\underline{\sigma}_{\boldsymbol{Q}}}\\ &\leq \frac{\left\|\Sigma_{\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i*}}\right\|}{10T}+\frac{C^{i,-i*}}{\underline{\sigma}_{\boldsymbol{Q}}}, \qquad (3.65)\end{aligned}$$

which leads to

$$\left\|\Sigma_{\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i*}}\right\|\leq\frac{10T\,C^{i,-i*}}{(10T-1)\,\underline{\sigma}_{\boldsymbol{Q}}}. \qquad (3.66)$$

Step 3: We now bound $\|P_{t,i}^{\boldsymbol{K}}-P_{t,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}}\|$ by $\sum_{j=1,j\neq i}^{N}\left\|\!\left|\boldsymbol{K}^j-\boldsymbol{K}^{j*}\right|\!\right\|$ where $\boldsymbol{K}=(\boldsymbol{K}^i,\boldsymbol{K}^{-i})$.

$$\begin{aligned}&\left\|P_{t,i}^{\boldsymbol{K}}-P_{t,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}}\right\|\\ =&\left\|\left(A_t-\sum_{j=1}^N B_t^jK_t^j\right)^{\top}P_{t+1,i}^{\boldsymbol{K}}\left(A_t-\sum_{j=1}^N B_t^jK_t^j\right)\right.\\ &\left.-\left(A_t-B_t^iK_t^i-\sum_{j=1,j\neq i}^N B_t^jK_t^{j*}\right)^{\top}P_{t+1,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}}\left(A_t-B_t^iK_t^i-\sum_{j=1,j\neq i}^N B_t^jK_t^{j*}\right)\right\|\\ \leq&\left\|\left(A_t-\sum_{j=1}^N B_t^jK_t^j\right)^{\top}\left(P_{t+1,i}^{\boldsymbol{K}}-P_{t+1,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}}\right)\left(A_t-\sum_{j=1}^N B_t^jK_t^j\right)\right\|\\ &+\left\|\left(A_t-\sum_{j=1}^N B_t^jK_t^j\right)^{\top}P_{t+1,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}}\left(A_t-\sum_{j=1}^N B_t^jK_t^j\right)\right.\\ &\left.-\left(A_t-B_t^iK_t^i-\sum_{j=1,j\neq i}^N B_t^jK_t^{j*}\right)^{\top}P_{t+1,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}}\left(A_t-B_t^iK_t^i-\sum_{j=1,j\neq i}^N B_t^jK_t^{j*}\right)\right\|\end{aligned}$$

$$\leq \rho_{\boldsymbol{K}}^2 \left\| P_{t+1,i}^{\boldsymbol{K}} - P_{t+1,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}} \right\| + 2\rho_{\boldsymbol{K}}\gamma_B \left\| P_{t+1,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}} \right\| \sum_{j=1,j\neq i}^{N} \left\| K_t^j - K_t^{j*} \right\|, \tag{3.67}$$

where the last inequality holds by letting $X = P_{t+1,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}}$, $Y = A_t - \sum_{j=1}^N B_t^j K_t^j$, and $Z = A_t - B_t^i K_t^i - \sum_{j=1,j\neq i}^N B_t^j K_t^{j*}$ in (3.48). Since $\left\| P_{T,i}^{\boldsymbol{K}} - P_{T,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}} \right\| = 0$ holds at terminal time T, we obtain $\forall t = 0, \cdots, T-1$,

$$\left\| P_{t,i}^{\boldsymbol{K}} - P_{t,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}} \right\| \leq \frac{2\gamma_B C^{i,-i*}}{\underline{\sigma}_{\boldsymbol{X}}} \frac{\rho_{\boldsymbol{K}}(\rho_{\boldsymbol{K}}^{2(T-t)} - 1)}{\rho_{\boldsymbol{K}}^2 - 1} \sum_{s=t}^{T-1} \left(\sum_{j=1,j\neq i}^{N} \|K_s^j - K_s^{j*}\| \right). \tag{3.68}$$

Therefore, $\forall t = 0, 1, \cdots, T-1$,

$$\begin{align}
& \left\| E_{t,i}^{\boldsymbol{K}} - E_{t,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}} \right\| \notag \\
= & \left\| (B_t^i)^\top \left(P_{t+1,i}^{\boldsymbol{K}} - P_{t+1,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}} \right) \left(A_t - \sum_{j=1}^{N} B_t^j K_t^j \right) - (B_t^i)^\top P_{t+1,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}} \sum_{j=1,j\neq i}^{N} B_t^j (K_t^j - K_t^{j*}) \right\| \notag \\
& \leq \rho_{\boldsymbol{K}}\gamma_B \| P_{t+1,i}^{\boldsymbol{K}} - P_{t+1,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}} \| + (\gamma_B)^2 \frac{C^{i,-i*}}{\underline{\sigma}_{\boldsymbol{X}}} \sum_{j=1,j\neq i}^{N} \|K_t^j - K_t^{j*}\| \tag{3.69} \\
& \leq \frac{(\gamma_B)^2 C^{i,-i*}}{\underline{\sigma}_{\boldsymbol{X}}} \left(\frac{2\rho_{\boldsymbol{K}}^2(\rho_{\boldsymbol{K}}^{2(T-t-1)} - 1)}{\rho_{\boldsymbol{K}}^2 - 1} \sum_{s=t+1}^{T-1} \sum_{j=1,j\neq i}^{N} \|K_s^j - K_s^{j*}\| + \sum_{j=1,j\neq i}^{N} \|K_t^j - K_t^{j*}\| \right) \tag{3.70} \\
& \leq \frac{(\gamma_B)^2 C^{i,-i*}}{\underline{\sigma}_{\boldsymbol{X}}} \left(\frac{2(\rho_{\boldsymbol{K}}^{2(T-t)} - 1)}{\rho_{\boldsymbol{K}}^2 - 1} \sum_{s=t}^{T-1} \left(\sum_{j=1,j\neq i}^{N} \|K_s^j - K_s^{j*}\| \right) \right) \notag \\
& \leq \frac{(\gamma_B)^2 C^{i,-i*}}{\underline{\sigma}_{\boldsymbol{X}}} \left(\frac{2(\rho_{\boldsymbol{K}}^{2T} - 1)}{\rho_{\boldsymbol{K}}^2 - 1} \sum_{j=1,j\neq i}^{N} |||\boldsymbol{K}^j - \boldsymbol{K}^{j*}||| \right), \tag{3.71}
\end{align}$$

where (3.69) holds by (3.27), and (3.70) holds by (3.68).

Step 4: We can now estimate the cost difference between using $\boldsymbol{K}^i$ and the update $\boldsymbol{K}^{i\prime}$. By Lemma 12 we have

$$\begin{aligned}
& C^{i\prime,-i*} - C^{i,-i*} \\
& = \sum_{t=0}^{T-1} \Big[\operatorname{Tr}\left(\Sigma_t^{\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i*}} (K_t^{i\prime} - K_t^i)^\top (R_t^i + (B_t^i)^\top P_{t+1,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}} B_t^i)(K_t^{i\prime} - K_t^i) \right) \\
& +2 \operatorname{Tr}\left(\Sigma_t^{\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i*}} (K_t^{i\prime} - K_t^i)^\top E_{t,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}} \right) \Big].
\end{aligned}$$

Using the updating rule $K_t^{i\prime} = K_t^i - \eta \nabla_{K_t^i} C(\boldsymbol{K})(\Sigma_t^{\boldsymbol{K}})^{-1}$ and the expression for the gradient from Lemma 9

$$C^{i\prime,-i*} - C^{i,-i*}$$

$$
\begin{aligned}
&= \sum_{t=0}^{T-1} \Big[4\eta^2 \operatorname{Tr}\Big(\Sigma_t^{\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i*}} (E_{t,i}^{\boldsymbol{K}})^\top (R_t^i + (B_t^i)^\top P_{t+1,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}} B_t^i)(E_{t,i}^{\boldsymbol{K}}) \Big) \\
&\qquad - 4\eta \operatorname{Tr}\big(\Sigma_t^{\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i*}} (E_{t,i}^{\boldsymbol{K}})^\top E_{t,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}} \big) \Big] \\
&= \sum_{t=0}^{T-1} \Big[4\eta^2 \operatorname{Tr}\big(\Sigma_t^{\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i*}} (E_{t,i}^{\boldsymbol{K}} - E_{t,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}} + E_{t,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}})^\top (R_t^i + (B_t^i)^\top P_{t+1,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}} B_t^i)(E_{t,i}^{\boldsymbol{K}} \\
&\qquad - E_{t,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}} + E_{t,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}}) \big) - 4\eta \operatorname{Tr}\big(\Sigma_t^{\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i*}} (E_{t,i}^{\boldsymbol{K}} - E_{t,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}} + E_{t,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}})^\top E_{t,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}} \big) \Big] \\
&= \sum_{t=0}^{T-1} \Big[4\eta^2 \operatorname{Tr}\big(\Sigma_t^{\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i*}} (E_{t,i}^{\boldsymbol{K}} - E_{t,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}})^\top (R_t^i + (B_t^i)^\top P_{t+1,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}} B_t^i)(E_{t,i}^{\boldsymbol{K}} - E_{t,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}}) \big) \\
&\qquad + 8\eta^2 \operatorname{Tr}\big(\Sigma_t^{\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i*}} (E_{t,i}^{\boldsymbol{K}} - E_{t,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}})^\top (R_t^i + (B_t^i)^\top P_{t+1,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}} B_t^i) E_{t,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}} \big) \\
&\qquad + 4\eta^2 \operatorname{Tr}\big(\Sigma_t^{\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i*}} (E_{t,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}})^\top (R_t^i + (B_t^i)^\top P_{t+1,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}} B_t^i) E_{t,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}} \big) \\
&\qquad - 4\eta \operatorname{Tr}\big(\Sigma_t^{\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i*}} (E_{t,i}^{\boldsymbol{K}} - E_{t,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}})^\top E_{t,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}} \big) \\
&\qquad - 4\eta \operatorname{Tr}\big(\Sigma_t^{\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i*}} (E_{t,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}})^\top E_{t,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}} \big) \Big].
\end{aligned}
$$

Now, letting $\omega^2 = \frac{2}{\underline{\sigma}_{\boldsymbol{X}}}$ in

$$
2\operatorname{Tr}(A^\top B) = \operatorname{Tr}(A^\top B + B^\top A) \le \omega^2 \operatorname{Tr}(A^\top A) + \frac{1}{\omega^2} \operatorname{Tr}(B^\top B),
$$

(which holds for any matrices A and B of the same dimension) we have

$$
\begin{aligned}
&C^{i\prime,-i*} - C^{i,-i*} \\
&\le \sum_{t=0}^{T-1} \Big[4\eta^2 \operatorname{Tr}\big(\Sigma_t^{\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i*}} (E_{t,i}^{\boldsymbol{K}} - E_{t,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}})^\top (R_t^i + (B_t^i)^\top P_{t+1,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}} B_t^i)(E_{t,i}^{\boldsymbol{K}} - E_{t,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}}) \big) \\
&\qquad + 8\eta^2 \frac{\underline{\sigma}_{\boldsymbol{X}}}{4} \operatorname{Tr}\big((E_{t,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}})^\top E_{t,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}} \big) + 8\eta^2 \frac{1}{\underline{\sigma}_{\boldsymbol{X}}} \operatorname{Tr}\big(\Sigma_t^{\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i*}} (E_{t,i}^{\boldsymbol{K}} - E_{t,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}})^\top \cdot \\
&\qquad\quad (R_t^i + (B_t^i)^\top P_{t+1,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}} B_t^i)(R_t^i + (B_t^i)^\top P_{t+1,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}} B_t^i)(E_{t,i}^{\boldsymbol{K}} - E_{t,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}}) \Sigma_t^{\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i*}} \big) \\
&\qquad + 4\eta^2 \operatorname{Tr}\big(\Sigma_t^{\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i*}} (E_{t,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}})^\top (R_t^i + (B_t^i)^\top P_{t+1,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}} B_t^i) E_{t,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}} \big) \\
&\qquad + 4\eta \frac{1}{\underline{\sigma}_{\boldsymbol{X}}} \operatorname{Tr}\big(\Sigma_t^{\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i*}} (E_{t,i}^{\boldsymbol{K}} - E_{t,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}})^\top (E_{t,i}^{\boldsymbol{K}} - E_{t,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}}) \Sigma_t^{\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i*}} \big) \\
&\qquad + 4\eta \frac{\underline{\sigma}_{\boldsymbol{X}}}{4} \operatorname{Tr}\big((E_{t,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}})^\top E_{t,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}} \big) - 4\eta\, \underline{\sigma}_{\boldsymbol{X}} \operatorname{Tr}\big((E_{t,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}})^\top E_{t,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}} \big) \Big] \\
&\le \sum_{t=0}^{T-1} \Big[\Big(4\eta^2 \|\Sigma_t^{\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i*}}\| \operatorname{Tr}\big(R_t^i + (B_t^i)^\top P_{t+1,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}} B_t^i \big) + \frac{8\eta^2}{\underline{\sigma}_{\boldsymbol{X}}} \|R_t^i + (B_t^i)^\top P_{t+1,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}} B_t^i\|^2 \cdot \\
&\qquad \operatorname{Tr}\big(\Sigma_t^{\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i*}} \Sigma_t^{\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i*}} \big) + \frac{4\eta}{\underline{\sigma}_{\boldsymbol{X}}} \operatorname{Tr}\big(\Sigma_t^{\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i*}} \Sigma_t^{\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i*}} \big) \Big) \|E_{t,i}^{\boldsymbol{K}} - E_{t,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}}\|^2 \\
&\qquad + \big(2\eta^2\, \underline{\sigma}_{\boldsymbol{X}} + 4\eta^2 \|\Sigma_t^{\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i*}}\| \|R_t^i + (B_t^i)^\top P_{t+1,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}} B_t^i\| + \eta\, \underline{\sigma}_{\boldsymbol{X}} - 4\eta\, \underline{\sigma}_{\boldsymbol{X}} \big) \cdot \\
&\qquad \operatorname{Tr}\big((E_{t,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}})^\top E_{t,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}} \big) \Big]. \qquad (3.72)
\end{aligned}
$$

Now, using $\left\|\Sigma_{\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i*}}\right\| \leq \frac{2\,C^{i,-i*}}{\underline{\sigma}_{\boldsymbol{Q}}}$ by (3.66) (this is a loose approximation in order to ease the analysis and smooth out the presentation), we can bound the step size condition in (3.62) by

$$\eta \leq \frac{\underline{\sigma}_{\boldsymbol{X}}}{\underline{\sigma}_{\boldsymbol{X}} + 2\|\Sigma_t^{\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i*}}\| \|R_t^i + (B_t^i)^\top P_{t+1,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}} B_t^i\|}.$$

This gives

$$2\eta^2\,\underline{\sigma}_{\boldsymbol{X}} + 4\eta^2 \left\|\Sigma_t^{\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i*}}\right\| \left\|R_t^i + (B_t^i)^\top P_{t+1,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}} B_t^i\right\| + \eta\,\underline{\sigma}_{\boldsymbol{X}} - 4\eta\,\underline{\sigma}_{\boldsymbol{X}} \leq -\eta\,\underline{\sigma}_{\boldsymbol{X}}\,.$$

Hence, using this in (3.72), we have

$$C^{i\prime,-i*} - C^{i,-i*} \leq \eta\, h^i_{\text{diff}} \sum_{t=0}^{T-1} \|E_{t,i}^{\boldsymbol{K}} - E_{t,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}}\|^2 - \eta\,\underline{\sigma}_{\boldsymbol{X}} \sum_{t=0}^{T-1} \operatorname{Tr}\left((E_{t,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}})^\top E_{t,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}}\right), \tag{3.73}$$

where

$$h^i_{\text{diff}} := 4\eta\, k_i \frac{10TC^{i,-i*}}{(10T-1)\,\underline{\sigma}_{\boldsymbol{Q}}} \left(\gamma_R + (\gamma_B)^2 \frac{C^{i,-i*}}{\underline{\sigma}_{\boldsymbol{X}}}\right) + 8\eta \frac{d}{\underline{\sigma}_{\boldsymbol{X}}} \left(\frac{10TC^{i,-i*}}{(10T-1)\,\underline{\sigma}_{\boldsymbol{Q}}}\right)^2 \cdot \\ \left(\gamma_R + (\gamma_B)^2 \frac{C^{i,-i*}}{\underline{\sigma}_{\boldsymbol{X}}}\right)^2 + 4\frac{d}{\underline{\sigma}_{\boldsymbol{X}}} \left(\frac{10TC^{i,-i*}}{(10T-1)\,\underline{\sigma}_{\boldsymbol{Q}}}\right)^2.$$

Therefore, by (3.66), (3.71), Lemma 11, and Lemma 13,

$$\begin{aligned} C^{i\prime,-i*} - C^{i,-i*} \quad &\leq \quad \eta\, h^i_{\text{diff}}\, T \left[\frac{(\gamma_B)^2 C^{i,-i*}}{\underline{\sigma}_{\boldsymbol{X}}} \frac{2(\rho_{\boldsymbol{K}}^{2T}-1)}{\rho_{\boldsymbol{K}}^2 - 1} \sum_{j=1,j\neq i}^{N} \|\!|\boldsymbol{K}^j - \boldsymbol{K}^{j*}|\!\|\right]^2 \\ &\qquad -\eta\,\underline{\sigma}_{\boldsymbol{X}} \sum_{t=0}^{T-1} \operatorname{Tr}\left((E_{t,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}})^\top E_{t,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}}\right) \\ &\leq \quad \eta\, h_{\text{glob}} \left(\sum_{j=1,j\neq i}^{N} \|\!|\boldsymbol{K}^j - \boldsymbol{K}^{j*}|\!\|\right)^2 - \eta \frac{\underline{\sigma}_{\boldsymbol{X}}\,\underline{\sigma}_{\boldsymbol{R}}}{\|\Sigma_{\boldsymbol{K}^*}\|} (C^{i,-i*} - C^{i*}), \end{aligned} \tag{3.74}$$

where

$$h_{\text{glob}} = 4T \left[\eta\, k_i \frac{10T \max_i\{C^{i,-i*}\}}{(10T-1)\,\underline{\sigma}_{\boldsymbol{Q}}} \left(\gamma_R + (\gamma_B)^2 \frac{\max_i\{C^{i,-i*}\}}{\underline{\sigma}_{\boldsymbol{X}}}\right) + 2\eta \frac{d}{\underline{\sigma}_{\boldsymbol{X}}} \left(\frac{10T \max_i\{C^{i,-i*}\}}{(10T-1)\,\underline{\sigma}_{\boldsymbol{Q}}}\right)^2 \cdot \right. \\ \left. \left(\gamma_R + (\gamma_B)^2 \frac{\max_i\{C^{i,-i*}\}}{\underline{\sigma}_{\boldsymbol{X}}}\right)^2 + \frac{d}{\underline{\sigma}_{\boldsymbol{X}}} \left(\frac{10T \max_i\{C^{i,-i*}\}}{(10T-1)\,\underline{\sigma}_{\boldsymbol{Q}}}\right)^2\right] \cdot \\ \left[\frac{(\gamma_B)^2 \max_i\{C^{i,-i*}\}}{\underline{\sigma}_{\boldsymbol{X}}} \frac{2(\rho_{\boldsymbol{K}}^{2T}-1)}{\rho_{\boldsymbol{K}}^2 - 1}\right]^2. \tag{3.75}$$

Step 5: Finally we can establish the one-step contraction. Using (3.74), we have

$$C^{i\prime,-i*} - C^{i*} \quad = \quad C^{i\prime,-i*} - C^{i,-i*} + C^{i,-i*} - C^{i*}$$

$$\leq \left(1-\eta\frac{\underline{\sigma}_{\boldsymbol{X}}\,\underline{\sigma}_{\boldsymbol{R}}}{\|\Sigma_{\boldsymbol{K}^*}\|}\right)\left(C^{i,-i*}-C^{i*}\right)+\eta\, h_{\text{glob}}\left(\sum_{j=1,j\neq i}^{N}\left|\!\left|\!\left|\boldsymbol{K}^j-\boldsymbol{K}^{j*}\right|\!\right|\!\right|\right)^2. \tag{3.76}$$

Hence by Lemma 12 and (3.42), we have

$$\sum_{j=1,j\neq i}^{N}\left(C^{j,-j*}-C^{j*}\right)\geq\frac{\underline{\sigma}_{\boldsymbol{X}}\,\underline{\sigma}_{\boldsymbol{R}}}{T}\sum_{j=1,j\neq i}^{N}\left|\!\left|\!\left|\boldsymbol{K}^j-\boldsymbol{K}^{j*}\right|\!\right|\!\right|^2\geq\frac{\underline{\sigma}_{\boldsymbol{X}}\,\underline{\sigma}_{\boldsymbol{R}}}{T(N-1)}\left(\sum_{j=1,j\neq i}^{N}\left|\!\left|\!\left|\boldsymbol{K}^j-\boldsymbol{K}^{j*}\right|\!\right|\!\right|\right)^2,$$

and thus

$$C^{i\prime,-i*}-C^{i*}\leq\left(1-\eta\frac{\underline{\sigma}_{\boldsymbol{X}}\,\underline{\sigma}_{\boldsymbol{R}}}{\|\Sigma_{\boldsymbol{K}^*}\|}\right)\left(C^{i,-i*}-C^{i*}\right)+\ \eta\ h_{\text{glob}}\frac{T(N-1)}{\underline{\sigma}_{\boldsymbol{X}}\,\underline{\sigma}_{\boldsymbol{R}}}\left(\sum_{j=1,j\neq i}^{N}\left(C^{j,-j*}-C^{j*}\right)\right). \tag{3.77}$$

Summing up (3.77) for $i=1,\cdots,N$, we have

$$\sum_{i=1}^{N}\left(C^{i\prime,-i*}-C^{i*}\right)\leq\left(1-\eta\frac{\underline{\sigma}_{\boldsymbol{X}}\,\underline{\sigma}_{\boldsymbol{R}}}{\|\Sigma_{\boldsymbol{K}^*}\|}+\eta(N-1)h_{\text{glob}}\frac{T(N-1)}{\underline{\sigma}_{\boldsymbol{X}}\,\underline{\sigma}_{\boldsymbol{R}}}\right)\left(\sum_{i=1}^{N}(C^{i,-i*}-C^{i*})\right). \tag{3.78}$$

Since $\eta\leq I_2$, we have (3.63) and then

$$h^i_{\text{diff}}\leq(4+\frac{1}{20})\frac{d}{\underline{\sigma}_{\boldsymbol{X}}}\left(\frac{10T\max_i\{C^{i,-i*}\}}{(10T-1)\,\underline{\sigma}_{\boldsymbol{Q}}}\right)^2\leq 5\frac{d}{\underline{\sigma}_{\boldsymbol{X}}}\left(\frac{\max_i\{C^{i,-i*}\}}{\underline{\sigma}_{\boldsymbol{Q}}}\right)^2,\quad\text{and}\quad h_{\text{glob}}\leq\bar{h}_{\text{glob}},$$

where $\bar{h}_{\text{glob}}$ is given by

$$\bar{h}_{\text{glob}}=5\,T\,d\,\frac{\max_i\{C^{i,-i*}\}^4(\gamma_B)^4}{\underline{\sigma}_{\boldsymbol{Q}}{}^2\,\underline{\sigma}_{\boldsymbol{X}}{}^3}\left[\frac{2(\rho_{\boldsymbol{K}}^{2T}-1)}{\rho_{\boldsymbol{K}}^2-1}\right]^2. \tag{3.79}$$

Under condition (3.56), we have

$$\underline{\sigma}_{\boldsymbol{X}}\,g_1-\frac{g_2}{\underline{\sigma}_{\boldsymbol{X}}{}^4}=\frac{\underline{\sigma}_{\boldsymbol{X}}\,\underline{\sigma}_{\boldsymbol{R}}}{\|\Sigma_{\boldsymbol{K}^*}\|}-(N-1)^2\bar{h}_{\text{glob}}\frac{T}{\underline{\sigma}_{\boldsymbol{X}}\,\underline{\sigma}_{\boldsymbol{R}}}>0,$$

which indicates that $\alpha\eta>0$. Since $\eta\leq\frac{1}{\underline{\sigma}_{\boldsymbol{R}}}$, we have

$$\eta<\frac{\|\Sigma_{\boldsymbol{K}^*}\|}{\underline{\sigma}_{\boldsymbol{X}}\,\underline{\sigma}_{\boldsymbol{R}}}<\left(\frac{\underline{\sigma}_{\boldsymbol{X}}\,\underline{\sigma}_{\boldsymbol{R}}}{\|\Sigma_{\boldsymbol{K}^*}\|}-(N-1)^2\bar{h}_{\text{glob}}\frac{T}{\underline{\sigma}_{\boldsymbol{X}}\,\underline{\sigma}_{\boldsymbol{R}}}\right)^{-1}=\left(\underline{\sigma}_{\boldsymbol{X}}\,g_1-\frac{g_2}{\underline{\sigma}_{\boldsymbol{X}}{}^4}\right)^{-1}.$$

Recall that in the statement we define $\alpha=\underline{\sigma}_{\boldsymbol{X}}\,g_1-g_2/\underline{\sigma}_{\boldsymbol{X}}{}^4$. Therefore we have $\alpha\eta<1$. Along with (3.78), we obtain the one-step contraction (3.59). ■

Lemma 21 *Assume Assumptions 1, 2, and 3 hold. Then we have that for player i,*

$$\begin{aligned}\sum_{t=0}^{T-1} \|\nabla_{K_t^i} C^i(\boldsymbol{K})\|^2 \quad \le \quad & 8\left\{\frac{\rho_{\boldsymbol{K}}^{2(T+1)}-1}{\rho_{\boldsymbol{K}}^2-1}\|\Sigma_0\| + \frac{\rho_{\boldsymbol{K}}^{2T}-1}{\rho_{\boldsymbol{K}}^2-1}\, T\,\|W\|\right\}^2 \cdot \\ & \Bigg\{ d\frac{T^2\,(N-1)}{\underline{\sigma}_{\boldsymbol{X}}\,\underline{\sigma}_{\boldsymbol{R}}}\left[\frac{(\gamma_B)^2\sum_{i=1}^N C^i(\boldsymbol{K}^i,\boldsymbol{K}^{-i*})}{\underline{\sigma}_{\boldsymbol{X}}}\,\frac{2(\rho_{\boldsymbol{K}}^{2T}-1)}{\rho_{\boldsymbol{K}}^2-1}\right]^2 \cdot \\ & \quad \left(\sum_{j=1,j\neq i}^N (C^j(\boldsymbol{K}^j,\boldsymbol{K}^{-j*}) - C^j(\boldsymbol{K}^*))\right) \\ & + \frac{\underline{\sigma}_{\boldsymbol{X}}\,\gamma_R + (\gamma_B)^2\sum_{i=1}^N C^i(\boldsymbol{K}^i,\boldsymbol{K}^{-i*})}{{\underline{\sigma}_{\boldsymbol{X}}}^2}\left(C^i(\boldsymbol{K}^i,\boldsymbol{K}^{-i*}) - C^i(\boldsymbol{K}^*)\right)\Bigg\},\end{aligned} \tag{3.80}$$

Proof Using Lemma 9, we have

$$\sum_{t=0}^{T-1} \|\nabla_{K_t^i} C^i(\boldsymbol{K})\|^2 \le 4\sum_{t=0}^{T-1} \mathrm{Tr}\left(\Sigma_t^{\boldsymbol{K}}(E_{t,i}^{\boldsymbol{K}})^\top E_{t,i}^{\boldsymbol{K}}\Sigma_t^{\boldsymbol{K}}\right) \le 4\,(\|\Sigma_{\boldsymbol{K}}\|)^2\sum_{t=0}^{T-1}\mathrm{Tr}\left((E_{t,i}^{\boldsymbol{K}})^\top E_{t,i}^{\boldsymbol{K}}\right), \tag{3.81}$$

and

$$\begin{aligned} & \sum_{t=0}^{T-1}\mathrm{Tr}\left((E_{t,i}^{\boldsymbol{K}})^\top E_{t,i}^{\boldsymbol{K}}\right) \\ = \quad & \sum_{t=0}^{T-1}\mathrm{Tr}\left((E_{t,i}^{\boldsymbol{K}} - E_{t,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}} + E_{t,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}})^\top(E_{t,i}^{\boldsymbol{K}} - E_{t,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}} + E_{t,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}})\right) \\ \le \quad & 2\sum_{t=0}^{T-1}\mathrm{Tr}\left((E_{t,i}^{\boldsymbol{K}} - E_{t,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}})^\top(E_{t,i}^{\boldsymbol{K}} - E_{t,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}})\right) + 2\sum_{t=0}^{T-1}\mathrm{Tr}\left((E_{t,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}})^\top E_{t,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}}\right) \\ \le \quad & 2d\sum_{t=0}^{T-1}\left\|E_{t,i}^{\boldsymbol{K}} - E_{t,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}}\right\|^2 + 2\sum_{t=0}^{T-1}\mathrm{Tr}\left((E_{t,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}})^\top E_{t,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}}\right). \end{aligned} \tag{3.82}$$

By (3.42), (3.71), we have

$$\begin{aligned} & \sum_{t=0}^{T-1}\left\|E_{t,i}^{\boldsymbol{K}} - E_{t,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}}\right\|^2 \\ \le \quad & T\left[\frac{(\gamma_B)^2 C^i(\boldsymbol{K}^i,\boldsymbol{K}^{-i*})}{\underline{\sigma}_{\boldsymbol{X}}}\,\frac{2(\rho_{\boldsymbol{K}}^{2T}-1)}{\rho_{\boldsymbol{K}}^2-1}\left(\sum_{j=1,j\neq i}^N |||\boldsymbol{K}^j - \boldsymbol{K}^{j*}|||\right)\right]^2 \\ \le \quad & \frac{T^2\,(N-1)}{\underline{\sigma}_{\boldsymbol{X}}\,\underline{\sigma}_{\boldsymbol{R}}}\left[\frac{(\gamma_B)^2 C^i(\boldsymbol{K}^i,\boldsymbol{K}^{-i*})}{\underline{\sigma}_{\boldsymbol{X}}}\,\frac{2(\rho_{\boldsymbol{K}}^{2T}-1)}{\rho_{\boldsymbol{K}}^2-1}\right]^2\left(\sum_{j=1,j\neq i}^N (C^j(\boldsymbol{K}^j,\boldsymbol{K}^{-j*}) - C^j(\boldsymbol{K}^*))\right) \\ . \quad & \end{aligned} \tag{3.83}$$

By Lemma 11 and Lemma 13, we have

$$\sum_{t=0}^{T-1} \mathrm{Tr}\left((E_{t,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}})^\top E_{t,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}}\right) \le \frac{\underline{\sigma}_{\boldsymbol{X}}\,\gamma_R + (\gamma_B)^2 C^i(\boldsymbol{K}^i,\boldsymbol{K}^{-i*})}{\underline{\sigma}_{\boldsymbol{X}}{}^2}\left(C^i(\boldsymbol{K}^i,\boldsymbol{K}^{-i*}) - C^i(\boldsymbol{K}^*)\right). \tag{3.84}$$

By Proposition 1,

$$\begin{aligned}
\|\Sigma_{\boldsymbol{K}}\| &\le \|\Sigma_0\| + \sum_{t=0}^{T-1}\|\mathcal{G}_t(\Sigma_0)\| + \sum_{t=1}^{T-1}\sum_{s=1}^{t}\|D_{t,s}WD_{t,s}^\top\| + T\|W\| \\
&\le \|\Sigma_0\| + \|\Sigma_0\|\sum_{t=0}^{T-1}\rho_{\boldsymbol{K}}^{2(t+1)} + T\|W\|\sum_{s=1}^{T-1}\rho_{\boldsymbol{K}}^{2(T-s)} + T\|W\| \\
&\le \frac{\rho_{\boldsymbol{K}}^{2(T+1)}-1}{\rho_{\boldsymbol{K}}^2-1}\|\Sigma_0\| + \frac{\rho_{\boldsymbol{K}}^{2T}-1}{\rho_{\boldsymbol{K}}^2-1}\,T\,\|W\|.
\end{aligned} \tag{3.85}$$

Note that here we use $\sum_{t=1}^{T-1}\sum_{s=1}^{t}\rho_{\boldsymbol{K}}^{2(t-s)} \le T\sum_{s=1}^{T-1}\rho_{\boldsymbol{K}}^{2(T-s)}$, which is a loose bound in order to simplify the presentation. Therefore, combining (3.81)-(3.85), we obtain the statement (3.80). ■

Finally, we are ready to provide the proof for the main result based on Lemmas 19 and 21. **Proof** [Proof of Theorem 6] We first show that the total cost for the N players decreases at round $m=1$. Take $\boldsymbol{K}=\boldsymbol{K}^{(0)}$ and $\boldsymbol{K}'=\boldsymbol{K}^{(1)}$ in Lemma 19, we begin by showing that there exists a positive lower bound on the RHS of (3.58). By the Cauchy-Schwarz inequality and Lemma 21,

$$\begin{aligned}
&\sum_{t=0}^{T-1}\|\nabla_{K_t^i}C^i(\boldsymbol{K}^{(0)})\| \le \sqrt{T\cdot\sum_{t=0}^{T-1}\|\nabla_{K_t^i}C^i(\boldsymbol{K}^{(0)})\|^2} \\
\le\;& 2\sqrt{2}\left(\frac{\rho_{\boldsymbol{K}^{(0)}}^{2(T+1)}-1}{\rho_{\boldsymbol{K}^{(0)}}^2-1}\|\Sigma_0\| + \frac{\rho_{\boldsymbol{K}^{(0)}}^{2T}-1}{\rho_{\boldsymbol{K}^{(0)}}^2-1}\,T\,\|W\|\right)\cdot\left\{d\frac{T^2\,(N-1)}{\underline{\sigma}_{\boldsymbol{X}}\,\underline{\sigma}_{\boldsymbol{R}}}\left[\frac{(\gamma_B)^2\sum_{i=1}^N C^{i,-i*}}{\underline{\sigma}_{\boldsymbol{X}}}\cdot\right.\right. \\
&\left.\left.\frac{2(\rho_{\boldsymbol{K}^{(0)}}^{2T}-1)}{\rho_{\boldsymbol{K}^{(0)}}^2-1}\right]^2 + \frac{\underline{\sigma}_{\boldsymbol{X}}\,\gamma_R + (\gamma_B)^2\sum_{i=1}^N C^{i,-i*}}{\underline{\sigma}_{\boldsymbol{X}}{}^2}\right\}^{1/2}\sqrt{T\left(\sum_{i=1}^N(C^{i,-i*}-C^{i*})\right)} \\
\le\;& 2\sqrt{2}\left(\frac{\bar{\rho}^{2(T+1)}-1}{\bar{\rho}^2-1}\|\Sigma_0\| + \frac{\bar{\rho}^{2T}-1}{\bar{\rho}^2-1}\,T\,\|W\|\right)\cdot\left\{d\frac{T^2\,(N-1)}{\underline{\sigma}_{\boldsymbol{X}}\,\underline{\sigma}_{\boldsymbol{R}}}\left[\frac{(\gamma_B)^2\sum_{i=1}^N C^{i,-i*}}{\underline{\sigma}_{\boldsymbol{X}}}\cdot\right.\right. \\
&\left.\left.\frac{2(\bar{\rho}^{2T}-1)}{\bar{\rho}^2-1}\right]^2 + \frac{\underline{\sigma}_{\boldsymbol{X}}\,\gamma_R + (\gamma_B)^2\sum_{i=1}^N C^{i,-i*}}{\underline{\sigma}_{\boldsymbol{X}}{}^2}\right\}^{1/2}\sqrt{T\left(\sum_{i=1}^N(C^{i,-i*}-C^{i*})\right)},
\end{aligned}$$

where the last inequality holds since, after performing one-step natural policy gradient $K_t^{i,(1)} = K_t^{i,(0)} - \eta\nabla_{K_t^i}C^i(\boldsymbol{K}^{(0)})(\Sigma_t^{\boldsymbol{K}^{(0)}})^{-1}$, we have

$$\rho_{\boldsymbol{K}^{(0)}} \;\le\; \rho^* + N\gamma_B\sqrt{\frac{T}{\underline{\sigma}_{\boldsymbol{X}}\,\underline{\sigma}_{\boldsymbol{R}}}\psi} + \frac{1}{20T^2} = \bar{\rho} \tag{3.86}$$

where ρ^* and $\bar{\rho}$ are defined in (3.9) and (3.10), and $\psi := \max_i\{C^i(\boldsymbol{K}^{i,(0)},\boldsymbol{K}^{-i*})-C^{i*}\}$. (3.86) holds by Lemma 15.

Now we aim to show that $\frac{1}{\bar{\rho}}$ is bounded below by polynomials in some model parameters. Given that $\left\|A_t - \sum_{i=1}^N B_t^i K_t^{i*}\right\| \leq \gamma_A + \gamma_B |\!|\!|\boldsymbol{K}^*|\!|\!|$ and that Lemma 15, $\bar{\rho}$ can be bounded above by polynomials in T, N, γ_A, γ_B, $\frac{1}{\underline{\sigma}_{\boldsymbol{X}}}$, $\frac{1}{\underline{\sigma}_{\boldsymbol{R}}}$, and $|\!|\!|\boldsymbol{K}^*|\!|\!|$, or a constant $1+\xi$. Therefore, along with the fact that $\frac{1}{d} > \frac{1}{(a+1)(b+1)(c+1)}$ for $d < ab+c$ with some $a,b,c,d>0$ and $\frac{1}{a^n+1} > \frac{1}{(a+1)^n}$ for $a>0$ and $n \in \mathbb{N}^+$, $\frac{1}{\bar{\rho}}$ is bounded below by polynomials in $\frac{1}{T+1}$, $\frac{1}{N+1}$, $\frac{1}{\gamma_A+1}$, $\frac{1}{\gamma_B+1}$, $\underline{\sigma}_{\boldsymbol{X}}$, $\frac{1}{\underline{\sigma}_{\boldsymbol{X}}+1}$, $\underline{\sigma}_{\boldsymbol{R}}$, $\frac{1}{\underline{\sigma}_{\boldsymbol{R}}+1}$, and $\frac{1}{|\!|\!|\boldsymbol{K}^*|\!|\!|+1}$, or a constant $\frac{1}{1+\xi}$. Similarly, I_1 can be bounded below by polynomials in $\frac{1}{d+1}$, $\frac{1}{N+1}$, $\frac{1}{T+1}$, $\frac{1}{\sum_{i=1}^N C^{i,-i*}+1}$, $\frac{1}{\|W\|+1}$,$\frac{1}{\|\Sigma_0\|+1}$,$\frac{1}{\gamma_A+1}$, $\frac{1}{\gamma_B+1}$,$\frac{1}{\gamma_R+1}$, $\frac{1}{\underline{\sigma}_{\boldsymbol{R}}+1}$, $\underline{\sigma}_{\boldsymbol{R}}$, $\frac{1}{\underline{\sigma}_{\boldsymbol{Q}}+1}$, $\underline{\sigma}_{\boldsymbol{Q}}$, $\frac{1}{\underline{\sigma}_{\boldsymbol{X}}+1}$, $\underline{\sigma}_{\boldsymbol{X}}$, and $\frac{1}{|\!|\!|\boldsymbol{K}^*|\!|\!|+1}$; and I_2 can be bounded below by polynomials in $\frac{1}{\sum_i k_i+1}$, $\frac{1}{d+1}$, d, $\frac{1}{\sum_{i=1}^N C^{i,-i*}+1}$, $\frac{1}{\gamma_B+1}$, $\frac{1}{\underline{\sigma}_{\boldsymbol{Q}}+1}$, $\underline{\sigma}_{\boldsymbol{Q}}$, $\frac{1}{\gamma_R+1}$, $\frac{1}{\underline{\sigma}_{\boldsymbol{X}}+1}$, $\underline{\sigma}_{\boldsymbol{X}}$.

Hence, there exists $\eta_0 \in \mathcal{H}\left(\frac{1}{\sum_{i=1}^N C^i(\boldsymbol{K}^{i,(0)},\boldsymbol{K}^{-i*})+1}\right)$ as an appropriate polynomial in $\frac{1}{\sum_{i=1}^N C^i(\boldsymbol{K}^{i,(0)},\boldsymbol{K}^{-i*})+1}$, $\frac{1}{\sum_i k_i+1}$, $\frac{1}{d+1}$, d, $\frac{1}{N+1}$, $\frac{1}{T+1}$, $\frac{1}{\|W\|+1}$,$\frac{1}{\|\Sigma_0\|+1}$, $\frac{1}{\gamma_A+1}$, $\frac{1}{\gamma_B+1}$, $\frac{1}{\gamma_R+1}$, $\frac{1}{\underline{\sigma}_{\boldsymbol{X}}+1}$, $\underline{\sigma}_{\boldsymbol{X}}$, $\frac{1}{\underline{\sigma}_{\boldsymbol{R}}+1}$, $\underline{\sigma}_{\boldsymbol{R}}$, $\frac{1}{\underline{\sigma}_{\boldsymbol{Q}}+1}$, $\underline{\sigma}_{\boldsymbol{Q}}$, and $\frac{1}{|\!|\!|\boldsymbol{K}^*|\!|\!|+1}$, such that when $\eta < \eta_0$, the step size condition (3.58) is satisfied. Therefore, by Lemma 19, we have

$$\sum_{i=1}^N \Big(C^i(\boldsymbol{K}^{i,(1)},\boldsymbol{K}^{-i*}) - C^{i*}\Big) \leq (1-\widehat{\alpha}\eta)\sum_{i=1}^N \Big(C^i(\boldsymbol{K}^{i,(0)},\boldsymbol{K}^{-i*}) - C^{i*}\Big),$$

with $\widehat{\alpha}$ defined in (3.20), which implies the total cost of N players decreases at $m=1$. Proceeding inductively, assume the following facts hold at round m:

- $C^i(\boldsymbol{K}^{i,(m-1)},\boldsymbol{K}^{-i*}) - C^{i*} \leq \psi$ for $i = 1,2,\cdots,N$;
- $\rho_{\boldsymbol{K}^{(m-1)}} \leq \bar{\rho}$;
- $\sum_{i=1}^N \Big(C^i(\boldsymbol{K}^{i,(m)},\boldsymbol{K}^{-i*}) - C^{i*}\Big) \leq (1-\widehat{\alpha}\eta)\sum_{i=1}^N \Big(C^i(\boldsymbol{K}^{i,(m-1)},\boldsymbol{K}^{-i*}) - C^{i*}\Big)$.

Now we prove the above facts also hold in round $m+1$. Taking $\boldsymbol{K} = \boldsymbol{K}^{(m-1)}$ and $\boldsymbol{K}' = \boldsymbol{K}^{(m)}$ in (3.77), we have

$$\begin{aligned}
C^i(\boldsymbol{K}^{i,(m)},\boldsymbol{K}^{-i*}) - C^{i*} \quad &\leq \quad \big(1-\eta\frac{\underline{\sigma}_{\boldsymbol{X}}\,\underline{\sigma}_{\boldsymbol{R}}}{\|\Sigma_{\boldsymbol{K}^*}\|}\big)\big(C^i(\boldsymbol{K}^{i,(m-1)},\boldsymbol{K}^{-i*}) - C^{i*}\big)\\
&\qquad + \eta\ h_{\mathrm{glob}}\frac{T(N-1)}{\underline{\sigma}_{\boldsymbol{X}}\,\underline{\sigma}_{\boldsymbol{R}}}\Bigg(\sum_{j=1,j\neq i}^N \big(C^j(\boldsymbol{K}^{j,(m-1)},\boldsymbol{K}^{-j*}) - C^{j*}\big)\Bigg)\\
&\leq \quad \big(1-\eta\frac{\underline{\sigma}_{\boldsymbol{X}}\,\underline{\sigma}_{\boldsymbol{R}}}{\|\Sigma_{\boldsymbol{K}^*}\|}\big)\psi + \eta\ h_{\mathrm{glob}}\frac{T(N-1)}{\underline{\sigma}_{\boldsymbol{X}}\,\underline{\sigma}_{\boldsymbol{R}}}(N-1)\psi \leq \psi.
\end{aligned}$$

The last inequality holds since $1-\eta\frac{\underline{\sigma}_{\boldsymbol{X}}\,\underline{\sigma}_{\boldsymbol{R}}}{\|\Sigma_{\boldsymbol{K}^*}\|} + \eta\ h_{\mathrm{glob}}\frac{T(N-1)^2}{\underline{\sigma}_{\boldsymbol{X}}\,\underline{\sigma}_{\boldsymbol{R}}} < 1$ under Assumption 4. Therefore by (3.53), we have

$$\rho_{\boldsymbol{K}^{(m)}} = \rho^* + N\gamma_B\sqrt{\frac{T}{\underline{\sigma}_{\boldsymbol{X}}\,\underline{\sigma}_{\boldsymbol{R}}}\max_i\big\{C^i(\boldsymbol{K}^{i,(m)},\boldsymbol{K}^{-i*}) - C^{i*}\big\}} + \frac{1}{20T^2} \leq \bar{\rho}.$$

Thus the step size condition (3.58) is still satisfied for $\boldsymbol{K} = \boldsymbol{K}^{(m)}$ and $\boldsymbol{K}' = \boldsymbol{K}^{(m+1)}$ with $\eta < \eta_0$, since $\rho_{\boldsymbol{K}^{(m)}} \leq \bar{\rho}$ and $\sum_{i=1}^N C^i(\boldsymbol{K}^{i,(m)}, \boldsymbol{K}^{-i*}) \leq \sum_{i=1}^N C^i(\boldsymbol{K}^{i,(0)}, \boldsymbol{K}^{-i*})$. Therefore, Lemma 19 can be applied again for the update at round $m+1$ with $\boldsymbol{K} = \boldsymbol{K}^{(m)}$ and $\boldsymbol{K}' = \boldsymbol{K}^{(m+1)}$ to obtain:

$$\sum_{i=1}^N \Big(C^i(\boldsymbol{K}^{i,(m+1)}, \boldsymbol{K}^{-i*}) - C^{i*} \Big) \leq (1 - \widehat{\alpha}\eta) \sum_{i=1}^N \Big(C^i(\boldsymbol{K}^{i,(m)}, \boldsymbol{K}^{-i*}) - C^{i*} \Big).$$

For $\epsilon > 0$, provided $M \geq \frac{1}{\widehat{\alpha}\eta} \log \left(\frac{\sum_{i=1}^N (C^i(\boldsymbol{K}^{i,(0)}, \boldsymbol{K}^{-i*}) - C^{i*})}{\epsilon} \right)$, we have $\sum_{i=1}^N \Big(C^i(\boldsymbol{K}^{i,(M)}, \boldsymbol{K}^{-i*}) - C^{i*} \Big) \leq \epsilon$. ■

4. The Natural Policy Gradient Method with Unknown Parameters

Based on the update rule in (3.5), it is straightforward to develop a *model-free version* of the natural policy gradient algorithm using *sampled data*. See Algorithm 2 for the natural policy gradient method with unknown parameters. In contrast to the model-based case, where the gradient $\nabla C^i(\boldsymbol{K})$ and covariance matrix Σ_t^i can be calculated directly, these two terms can not be calculated in the model-free setting since the model parameters are unknown. Therefore, we propose to use a zeroth-order optimization method to estimate the gradient and an empirical covariance matrix to estimate the covariance matrix (see Equation 4.1). Building upon the theories in Section 3, high-probability convergence guarantees (linear convergence rate and polynomial sample complexity) for the model-free counterpart can be established in the same way as for the Linear Quadratic Regulator setting in Hambly et al. (2021).

5. Numerical Experiments

We demonstrate the performance of the natural policy gradient algorithms with three general-sum game examples. We will specifically focus on the following questions:

- In practice, how fast do the natural policy gradient algorithms converge to the true solution? How sensitive is the natural policy gradient algorithm to the step size and the initial policy?
- Do natural gradient methods converge when Assumption 4 is violated? How restrictive is Assumption 4 in practice?
- Can Theorem 6 provide any guidance on hyper-parameter tuning? In particular, does adding system noise improve the convergence?

The first example is a modified example from Mazumdar et al. (2020a), in which they show that in the setting of infinite time horizon and deterministic dynamics, the (vanilla) policy gradient algorithms have no guarantees of even local convergence to the Nash equilibria with known parameters. We will show in Section 5.1 that, under the same experimental set-up but over a finite time horizon with stochastic dynamics, the natural policy gradient

Algorithm 2 Natural Policy Gradient Method with Unknown Parameters

```
Input: Number of iterations $M$, time horizon $T$, initial policy $\boldsymbol{K}^{(0)} = (\boldsymbol{K}^{1,(0)}, \cdots, \boldsymbol{K}^{N,(0)})$, step size $\eta$, number of trajectories $L$, smoothing parameters $r_i$, dimensions $D_i = k_i \times d$.
for $m \in \{1, \dots, M\}$ do
  for $i \in \{1, \dots, N\}$ do
    for $l \in \{1, \dots, L\}$ do
      for $t \in \{0, \dots, T-1\}$ do
        Sample the (sub)-policy at time $t$: $\widehat{K}_t^{i,l} = K_t^{i,(m-1)} + U_t^{i,l}$ where $U_t^{i,l}$ is drawn uniformly at random over matrices such that $\|U_t^{i,l}\|_F = r_i$.
        Denote $\widehat{c}_t^{\,i,l}$ as the single trajectory cost of player $i$ with policy $(\widehat{\boldsymbol{K}}^{i,(m-1)}_{l,t}, \boldsymbol{K}^{-i,(m-1)})$ where $\widehat{\boldsymbol{K}}^{i,(m-1)}_{l,t} := (K_0^{i,(m-1)}, \cdots, K_{t-1}^{i,(m-1)}, \widehat{K}_t^{i,l}, K_t^{i,(m-1)}, \cdots, K_{T-1}^{i,(m-1)})$ starting from $x_0^l$.
        Denote $\widehat{\Sigma}_t^{i,l}$ as the state covariance matrix with $\widehat{\Sigma}_t^{i,l} = x_t^{i,l}(x_t^{i,l})^\top$.
      end for
    end for
  end for
  Obtain the estimates of $\nabla_{K_t^i} C^i(\boldsymbol{K}^{(m-1)})$ and $\Sigma_t^{\boldsymbol{K}^{(m-1)}}$ for each $i$ and $t$:
      $\nabla_{K_t^i} \widehat{C^i(\boldsymbol{K}^{(m-1)})} = \frac{1}{L} \sum_{l=1}^{L} \frac{D_i}{r_i^2} \, \widehat{c}_t^{\,i,l} \, U_t^{i,l}, \qquad \widehat{\Sigma}_t^i = \frac{1}{L} \sum_{l=1}^{L} \widehat{\Sigma}_t^{i,l}.$   (4.1)
  Update the policies using natural policy gradient updating rule:
      $K_t^{i,(m)} = K_t^{i,(m-1)} - \eta \nabla_{K_t^i} \widehat{C^i(\boldsymbol{K}^{(m-1)})} (\widehat{\Sigma}_t^i)^{-1}.$   (4.2)
end for
Return the iterates $\boldsymbol{K}^{(M)} = (\boldsymbol{K}^{1,(M)}, \cdots, \boldsymbol{K}^{N,(M)})$.
```

algorithm with known parameters finds the Nash equilibrium with properly chosen initial policies and step sizes. The second example is a two-player LQ game with synthetic data (see Section 5.2). We will show that the system noise helps the natural policy gradient algorithm with unknown parameters to converge to the Nash equilibrium. Finally, we investigate the algorithm's performance with known and unknown parameters for a three-player game example in Section 5.3.

Performance Measure. We use the following *normalized error* to quantify the performance of a given set of policies $\boldsymbol{K} = (\boldsymbol{K}^1, \boldsymbol{K}^2, \dots, \boldsymbol{K}^N)$: for $i = 1, 2, \dots, N$,

$$\text{Normalized error (of player } i) = \frac{C^i(\boldsymbol{K}^i, \boldsymbol{K}^{-i*}) - C^i(\boldsymbol{K}^*)}{C^i(\boldsymbol{K}^*)}.$$

5.1 Convergence of the Natural Policy Gradient Algorithm

For policy gradient MARL under the setting of N-player general-sum LQ games, some difficulties in convergence have been identified in some empirical studies. For example, Mazumdar et al. (2020a) showed by a counterexample that in the setting of infinite time horizon and deterministic dynamics, the (vanilla) policy gradient method avoids the Nash equilibria for a non-negligible subset of problems (with known parameters). In this section, we illustrate that in our setting with finite time horizon and stochastic dynamics, the natural policy gradient algorithm (with known parameters) finds the (unique) Nash equilibrium under the same experimental set-up as in a modified example given in Mazumdar et al. (2020a).

Set-up. We set up the model parameters and initialize the policies in the same way as Section 5.1 of Mazumdar et al. (2020a). Under the following set of parameters, there exists a unique Nash equilibrium since the sufficient condition in Remark 1 is satisfied.

1. Parameters: for $t = 1, \cdots, T-1$

$$A_t = \begin{bmatrix} 0.588 & 0.028 \\ 0.570 & 0.056 \end{bmatrix}, \quad B_t^1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \quad B_t^2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad W = \begin{bmatrix} \sigma^2 & 0 \\ 0 & \sigma^2 \end{bmatrix},$$

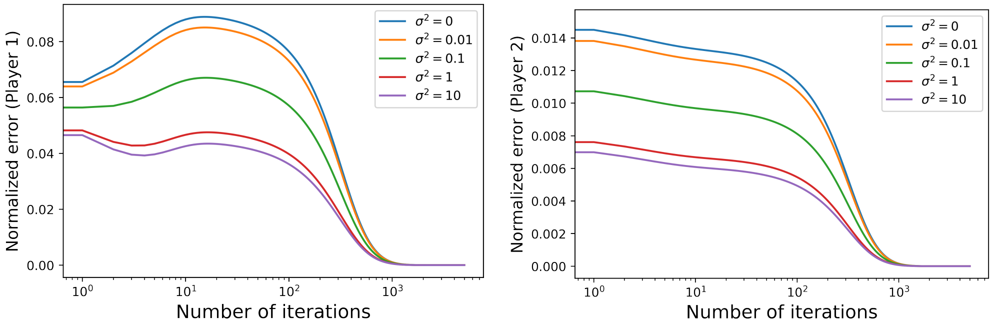
$$Q_T^1 = Q_t^1 = \begin{bmatrix} 0.01 & 0 \\ 0 & 1 \end{bmatrix}, \quad Q_T^2 = Q_t^2 = \begin{bmatrix} 1 & 0 \\ 0 & 0.147 \end{bmatrix}, \quad R_t^1(t) = R_t^2(t) = 0.01,$$

   where $\sigma \in \mathbb{R}$ and $T = 10$.

2. Initialization: we assume the initial state distribution to be $[1,1]^\top$ or $[1,1.1]^\top$ with probability 0.5 each. We initialize both players' policies $K_t^{i,(0)} = (K_{t0}^{i,(0)}, K_{t1}^{i,(0)})$ such that $(K_{t0}^{i,(0)} - K_{t0}^{i*})^2 + (K_{t1}^{i,(0)} - K_{t1}^{i*})^2 \leq r^2$, where $K_t^{i*} = (K_{t0}^{i*}, K_{t1}^{i*})$ denotes the Nash equilibrium, and r is the radius of the ball centered at K_t^{i*} in which we initialize the policies.

Convergence. The natural policy gradient algorithm shows a reasonable level of accuracy within 1000 iterations (that is, the normalized error is less than 0.5%) for both players under different levels of system noise σ^2, which ranges from 0 (deterministic dynamics) to 10. See Figure 1 for the case where $r = 0.25$ and Figure 2 for the case where $r = 0.30$. We observe that when we initialize the policies in a larger neighborhood of the Nash equilibrium, it takes the natural policy gradient algorithm (with the same step size) more iterations to converge. There is a sharp peak in the normalized error for player 2 and this peak diminishes when the noise level increases.

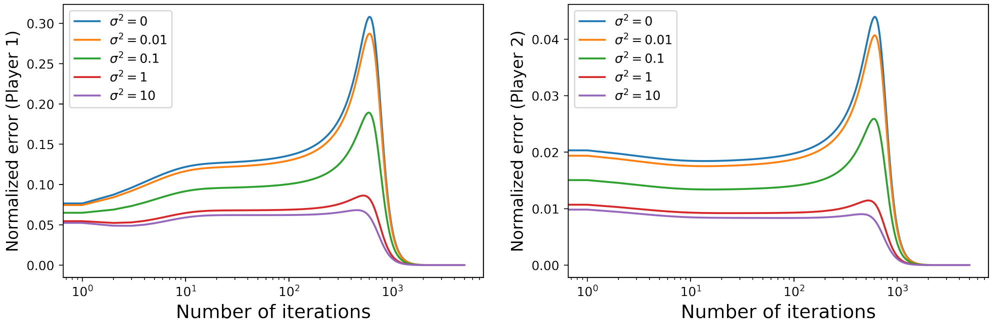
In Figure 3, we show the normalized error and the corresponding trajectories of learned policies near the peak observed in Figure 2 under $\sigma = 0$. We denote by $K_0^i = (K_{00}^i, K_{01}^i)$ the learned policy at $t = 0$ of player i. The peak period (iterations 300-900) is indicated in grey in Figures 3a-3c, and the trajectories of learned policy K_0^i for the rest of the whole 10000 iterations are indicated in red in Figures 3b-3c. The natural policy gradient algorithm overshoots in the first few iterations but detects the right direction after about 500 iterations and eventually converges to the Nash equilibrium (see the blue star in Figures 3b-3c).



(a) Player 1.

(b) Player 2.

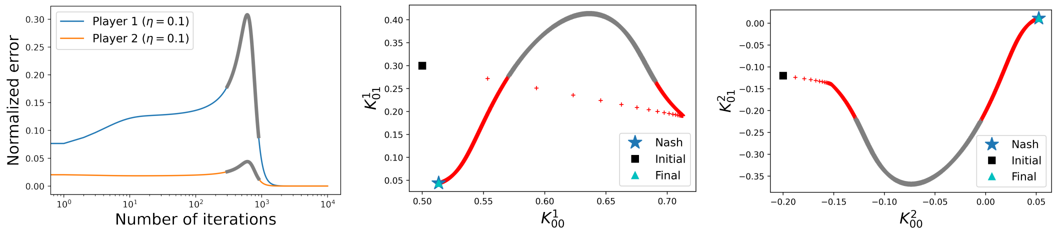
Figure 1: Normalized error under different σ^2 when $r = 0.25$ ($\eta_1 = \eta_2 = 0.1$ and $M = 5000$).



(a) Player 1.

(b) Player 2.

Figure 2: Normalized error under different σ^2 when $r = 0.3$ ($\eta_1 = \eta_2 = 0.1$ and $M = 5000$).

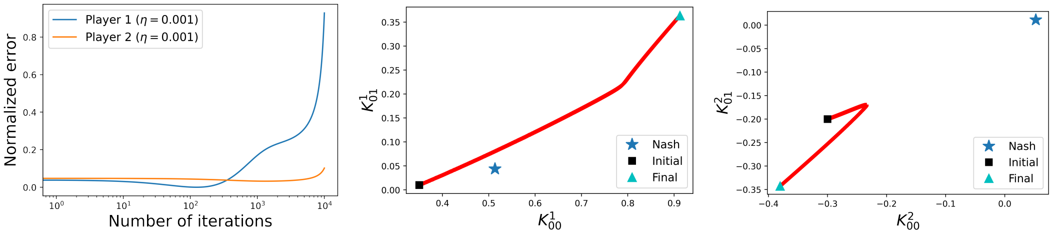


(a) Normalized error.

(b) Trajectory of K^1_0.

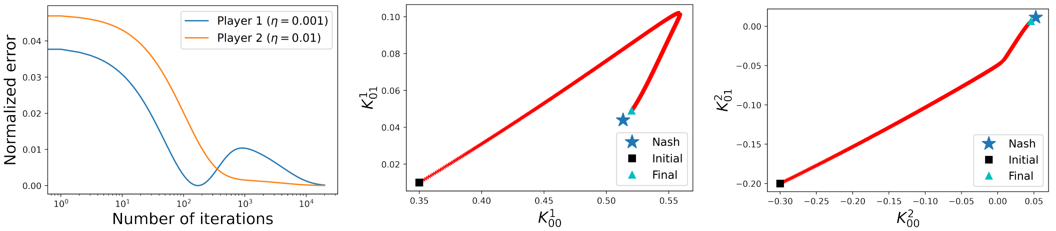
(c) Trajectory of K^2_0.

Figure 3: Normalized error and trajectories of learned policies with $\sigma^2 = 0$ and $r = 0.3$ ($\eta_1 = \eta_2 = 0.1$, $M = 10000$). The peak period between iterations 300-900 is indicated in grey.



(a) Normalized error.
(b) Trajectory of K^1_0.
(c) Trajectory of K^2_0.

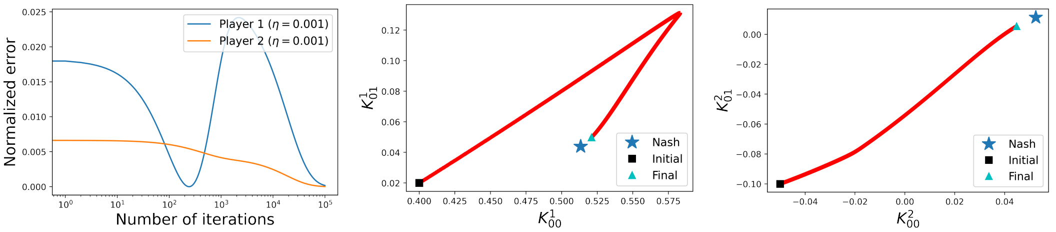
Figure 4: Performance of the natural policy gradient algorithm with $r = 0.42$ and $\eta_1 = \eta_2 = 0.001$ ($M = 10000$).



(a) Normalized error.
(b) Trajectory of K^1_0.
(c) Trajectory of K^2_0.

Figure 5: Performance of the natural policy gradient algorithm with $r = 0.42$, $\eta_1 = 0.001$, and $\eta_2 = 0.01$ ($M = 20000$).

Performance under Deterministic Dynamics. We observe that under carefully chosen initial policies and step sizes, the natural policy gradient converges to the Nash equilibrium even with deterministic state dynamics ($\sigma = 0$). We first show the case when the natural policy gradient algorithm diverges with $r = 0.42$ and $\eta_1 = \eta_2 = 0.001$ in Figure 4 (a trajectory of 10000 iterations is indicated in red). However, either by adjusting the step size to $\eta_2 = 0.01$ (see Figure 5), or by initializing the policies from a smaller neighbourhood around the Nash equilibrium (see Figure 6), the natural policy gradient method converges to the Nash equilibrium. This further demonstrates that the theoretical result, along with its assumptions, in Theorem 6 could provide insightful guidance on how to tune the hyper-parameters for practical examples.



(a) Normalized error.
(b) Trajectory of K^1_0.
(c) Trajectory of K^2_0.

Figure 6: Performance of the natural policy gradient algorithm with $r = 0.16$ and $\eta_1 = \eta_2 = 0.001$ ($M = 100000$).

5.2 Effect of the System Noise

As illustrated in the theoretical analysis in Section 3, the system noise plays an important role in the convergence guarantee of the natural policy gradient algorithm. To test the sensitivity of the performance of this algorithm to the level of system noise, we apply the natural policy gradient algorithm with unknown parameters to a two-player LQ game example with synthetic data consisting of a two-dimensional state variable and a one-dimensional control variable. The model parameters (except the level of noise σ^2 in W which we will discuss later) are randomly picked such that the conditions for our LQ game framework are satisfied.

Set-up. We perform the natural policy gradient algorithm with synthetic data given as follows.

1. Parameters: for $t = 1, \cdots, T-1$,

$$A_t = \begin{bmatrix} 0.588 & 0.28 \\ 0.57 & 0.56 \end{bmatrix}, \quad B_t^1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \quad B_t^2 = \begin{bmatrix} 0.5 \\ 1 \end{bmatrix}, \quad W = \begin{bmatrix} \sigma^2 & 0 \\ 0 & \sigma^2 \end{bmatrix},$$

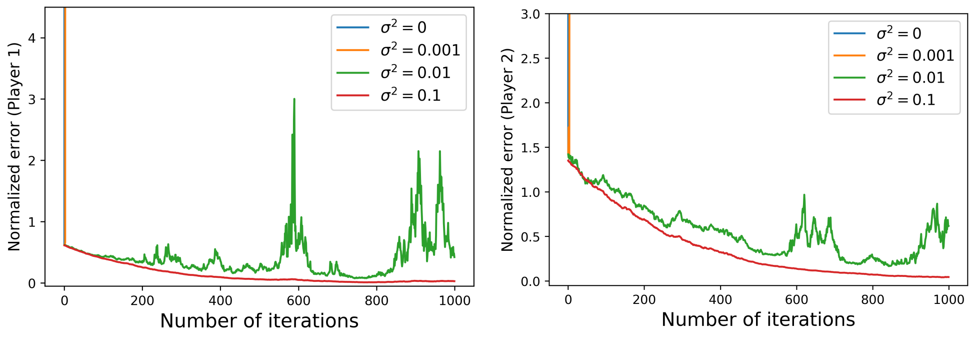
$$Q_T^1 = \begin{bmatrix} 0.5 & 0 \\ 0 & 1 \end{bmatrix}, \quad Q_T^2 = \begin{bmatrix} 1 & 0 \\ 0 & 0.3 \end{bmatrix}, \quad Q_t^1 = Q_t^2 = \begin{bmatrix} 0.001 & 0 \\ 0 & 0.001 \end{bmatrix}, \quad R_t^1 = R_t^2 = 1,$$

   where $\sigma \in \mathbb{R}_+$ and $T = 5$. The smoothing parameter is $r_1 = r_2 = 0.5$ and the number of trajectories is $L_1 = L_2 = 200$.

2. Initialization: we assume $x_0 = (x_0^1, x_0^2)$ with x_0^2 and x_0^2 independent and sampled from $\mathcal{N}(10, 2)$ and $\mathcal{N}(12, 3)$ respectively. The initial policy for player 1: $\boldsymbol{K}^1 \in \mathbb{R}^{1\times 10} = (K_0^{1,(0)}, \cdots, K_{T-1}^{1,(0)})$ with $K_t^{1,(0)} = [0.3, 0.15]$ for all t. The initial policy for player 2: $\boldsymbol{K}^2 \in \mathbb{R}^{1\times 10} = (K_0^{2,(0)}, \cdots, K_{T-1}^{2,(0)})$ with $K_t^{2,(0)} = [0.1, 0.05]$ for all t.

Convergence. To show that (even a low level of) the system noise can indeed help the natural policy gradient algorithm to find the Nash equilibrium, we vary the value of σ^2 from 0 to 0.1 and show the normalized error for different values of σ^2 in Figure 7. The natural policy gradient algorithm diverges when $\sigma^2 \leq 0.001$ for both players, and it starts to converge with large fluctuations when $\sigma^2 = 0.01$. When $\sigma^2 = 0.1$, the algorithm shows a reasonable accuracy within 300 iterations (that is the normalized error is less than 5%) for both players without fluctuations. It is worth pointing out that in fact, although Assumption 4 is violated when $\sigma^2 = 0.1$, the natural policy gradient algorithm converges to the Nash equilibrium. Finally, it is possible to make the algorithm converge when $\sigma^2 \leq 0.001$ by adjusting the initial policy and the step size, as shown in Section 5.1. In practice we may not be able to change the variance of w_t directly, however the level of system noise can also be increased by adding (Gaussian) explorations to the agents' policies. See Houthooft et al. (2016); Wang et al. (2020) for more discussion of Gaussian exploration.

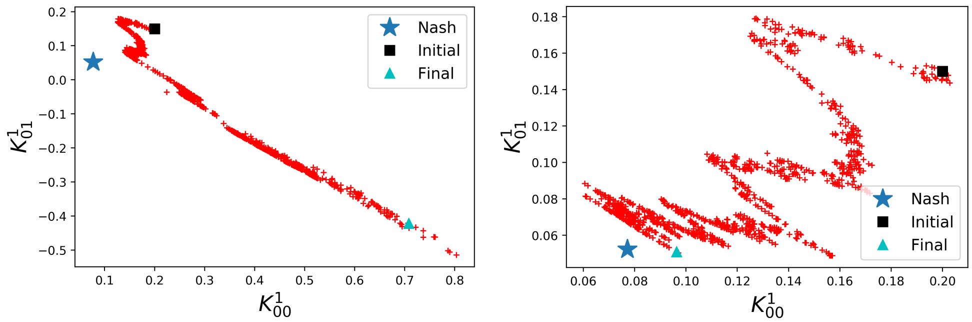
Trajectories of Learned Policies. We show the trajectories of the learned policy at $t = 0$ of player 1 by performing the natural policy gradient algorithm with unknown parameters under $\sigma^2 = 0.01$ and $\sigma^2 = 0.1$ in Figure 8 (with 1000 iterations). In the case of a lower level of system noise ($\sigma^2 = 0.01$), the natural policy gradient does not converge to Nash equilibrium with $\eta_1 = \eta_2 = 0.001$ in Figure 8a, whereas when the level of noise is increased to $\sigma^2 = 0.1$, the learned policy approaches the target within 1000 iterations with the same step size.



(a) Player 1.

(b) Player 2.

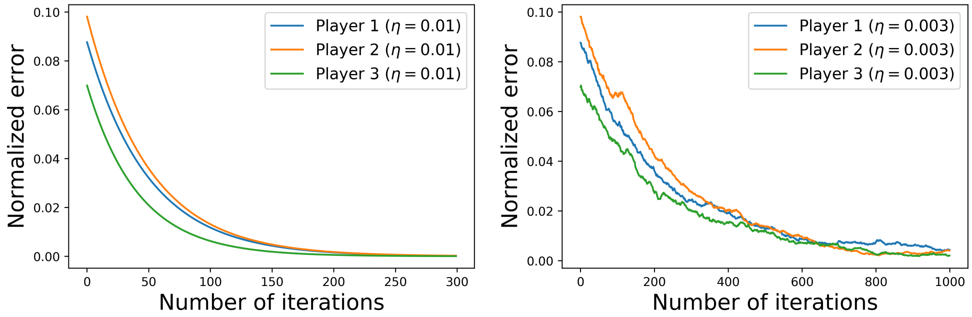
Figure 7: Performance of the natural policy gradient algorithm with unknown parameters ($\eta_1 = \eta_2 = 0.001$).



(a) $\sigma^2 = 0.01$

(b) $\sigma^2 = 0.1$.

Figure 8: Trajectory (indicated in red) of learned policy $K^1_0 = (K^1_{00}, K^1_{01})$ ($\eta_1 = \eta_2 = 0.001$, $M = 1000$).



(a) Known parameters.

(b) Unknown parameters.

Figure 9: Performance of the natural policy gradient algorithm with known and unknown parameters.

5.3 Convergence in a Three-player Game

In this section, we perform the the natural policy gradient method with known and unknown parameters in the following three-player general-sum game example. We show the convergence of the algorithms in Figure 9.

Set-up. We set-up the model parameters and initial policies as follows

1. Parameters:

$$A_t = \begin{bmatrix} 0.05 & -0.1 & 0.1 \\ 0.1 & 0.2 & -0.06 \\ -0.02 & 0.03 & 0.1 \end{bmatrix}, \quad B_t^1 = \begin{bmatrix} 0.05 \\ 0.01 \\ -0.01 \end{bmatrix}, \quad B_t^2 = \begin{bmatrix} 0.01 \\ -0.05 \\ -0.02 \end{bmatrix}, \quad B_t^3 = \begin{bmatrix} -0.02 \\ 0.01 \\ 0.05 \end{bmatrix},$$

$$W = \begin{bmatrix} 0.1 & 0.01 & 0.02 \\ 0.01 & 0.2 & 0.01 \\ 0.02 & 0.01 & 0.1 \end{bmatrix}, \quad Q_T^1 = Q_T^2 = Q_T^3 = Q_t^1 = Q_t^2 = Q_t^3 = \begin{bmatrix} 0.2 & 0 & 0 \\ 0 & 0.2 & 0 \\ 0 & 0 & 0.2 \end{bmatrix},$$

$R_t^1(t) = R_t^2(t) = 0.5$, $R_t^3(t) = 0.6$, and $T = 5$.

2. Initialization: Take $x_0 = (x_0^1, x_0^2, x_0^3)$ where x_0^1, x_0^2, x_0^3 are independent and sampled from $\mathcal{N}(0.3, 0.2)$ and $\mathcal{N}(0.2, 0.3)$, and $\mathcal{N}(0.3, 0.2)$ respectively. The initial policies are $\boldsymbol{K}^{1,(0)} = (0.35, 0.01, 0.1)$, $\boldsymbol{K}^{2,(0)} = (-0.3, -0.2, 0)$, and $\boldsymbol{K}^{3,(0)} = (-0.3, 0.1, 0)$.

Convergence. We plot the normalized error for each player in the case of known parameters and also unknown parameters in Figure 9. We can see that with three players, the algorithms still have a reasonably fast speed of convergence in practice.

Acknowledgments

We thank Paul Barnes and Amr El Zanfally at bp plc for providing part of the motivation for this work and for their continued support. Huining Yang was supported by the EPSRC Centre for Doctoral Training in Industrially Focused Mathematical Modelling (EP/L015803/1) in collaboration with bp plc.

Appendix A. The One-step Contraction Lemma for the Vanilla Policy Gradient Method

The convergence result for the natural policy gradient method can be extended to the case of the vanilla policy gradient method. The key step is to prove the one-step contraction (Lemma 19) for the vanilla version. This can be done by modifying some parts of the current Lemma 19. We first recall the definition of g_1 and g_2 as follows:

$$g_1 := \frac{\underline{\sigma}_{\boldsymbol{R}}}{\|\Sigma_{\boldsymbol{K}^*}\|},$$

and

$$g_2 := 20(N-1)^2\, T^2\, d\, \frac{(\gamma_B)^4 \max_i\{C^i(\boldsymbol{K}^i,\boldsymbol{K}^{-i*})\}^4}{\underline{\sigma}_{\boldsymbol{Q}}{}^2\, \underline{\sigma}_{\boldsymbol{R}}} \left(\frac{\rho_{\boldsymbol{K}}^{2T}-1}{\rho_{\boldsymbol{K}}^{2}-1}\right)^2.$$

We further define $\widetilde{g}_2$ as

$$\begin{aligned}\widetilde{g}_2 &:= g_2 \left(\rho_{\boldsymbol{K}}^{2T}\|\Sigma_0\| + \left(\rho_{\boldsymbol{K}}^{2T}+1\right)\|W\|\right)^2 \\ &= 20(N-1)^2\, T^2\, d\, \frac{(\gamma_B)^4 \max_i\{C^i(\boldsymbol{K}^i,\boldsymbol{K}^{-i*})\}^4}{\underline{\sigma}_{\boldsymbol{Q}}{}^2\, \underline{\sigma}_{\boldsymbol{R}}} \left(\frac{\rho_{\boldsymbol{K}}^{2T}-1}{\rho_{\boldsymbol{K}}^{2}-1}\right)^2 \left(\rho_{\boldsymbol{K}}^{2T}\|\Sigma_0\| + \left(\rho_{\boldsymbol{K}}^{2T}+1\right)\|W\|\right)^2,\end{aligned} \tag{A.1}$$

and g_3 as

$$g_3 := \frac{5\max_i\{C^i(\boldsymbol{K}^i,\boldsymbol{K}^{-i*})\}}{\underline{\sigma}_{\boldsymbol{Q}}}\left(\frac{\max_i\{C^i(\boldsymbol{K}^i,\boldsymbol{K}^{-i*})\}}{\underline{\sigma}_{\boldsymbol{Q}}} + \rho_{\boldsymbol{K}}^{2T}\|\Sigma_0\| + \left(\rho_{\boldsymbol{K}}^{2T}+1\right)\|W\|\right). \tag{A.2}$$

We also write $C^{i,-i*} = C^i(\boldsymbol{K}^i,\boldsymbol{K}^{-i*})$, $C^{i*} = C^i(\boldsymbol{K}^*)$ and $C^{i\prime,-i*} = C^i(\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i*})$ to simplify notation.

Lemma 1 (One-step Contraction for Vanilla Policy Gradient) *Assume Assumptions 1, 2, and 3 hold, and that*

$$\underline{\sigma}_{\boldsymbol{X}}{}^7 > \max\left\{\frac{\widetilde{g}_2}{g_1}, g_3^{\frac{7}{2}}\right\}. \tag{A.3}$$

Also assume the policy update step for player i at time t is given by

$$K_t^{i\prime} = K_t^i - \eta\nabla_{K_t^i} C^i(\boldsymbol{K}), \tag{A.4}$$

where

$$\eta \le \min\left\{I_3\,, I_4\,, \frac{\|\Sigma_{\boldsymbol{K}^*}\|}{\underline{\sigma}_{\boldsymbol{X}}{}^2\, \underline{\sigma}_{\boldsymbol{R}}}\right\} \tag{A.5}$$

with

$$I_3 = \left\{ \frac{20T\,\underline{\sigma}_{\boldsymbol{X}}\,\rho_{\boldsymbol{K}}(\rho_{\boldsymbol{K}}^{2T}-1)\gamma_B}{\rho_{\boldsymbol{K}}^2-1} \Big(\sum_{i=1}^N C^{i,-i*} + \underline{\sigma}_{\boldsymbol{Q}}\,T\|W\|\Big) \max_i\{\max_t\{\|\nabla_{K_t^i} C^i(\boldsymbol{K})\|\}\} + 2\,\underline{\sigma}_{\boldsymbol{Q}}\,\underline{\sigma}_{\boldsymbol{X}} \right.$$
$$\left. + 8\big(\gamma_R + \frac{(\gamma_B)^2}{\underline{\sigma}_{\boldsymbol{X}}}\sum_{i=1}^N C^{i,-i*}\big)\,\big(\rho_{\boldsymbol{K}}^{2T}\|\Sigma_0\| + \big(\rho_{\boldsymbol{K}}^{2T}+1\big)\|W\|\big)^2 \sum_{i=1}^N\{C^{i,-i*}\} \right\}^{-1} \cdot \underline{\sigma}_{\boldsymbol{Q}}(\underline{\sigma}_{\boldsymbol{X}})^2,$$

$$I_4 = \left\{ \big(\max_i\{k_i\}\big)\frac{10T\sum_{i=1}^N C^{i,-i*}}{(10T-1)\,\underline{\sigma}_{\boldsymbol{Q}}}\Big(\gamma_R + (\gamma_B)^2\frac{\sum_{i=1}^N C^{i,-i*}}{\underline{\sigma}_{\boldsymbol{X}}}\Big) + \frac{2d}{\underline{\sigma}_{\boldsymbol{X}}}\Big(\frac{10T\sum_{i=1}^N C^{i,-i*}}{(10T-1)\,\underline{\sigma}_{\boldsymbol{Q}}}\Big)^2 \cdot \right.$$
$$\left. \Big(\gamma_R + (\gamma_B)^2\frac{\sum_{i=1}^N C^{i,-i*}}{\underline{\sigma}_{\boldsymbol{X}}}\Big)^2\big(\rho_{\boldsymbol{K}}^{2T}\|\Sigma_0\| + \big(\rho_{\boldsymbol{K}}^{2T}+1\big)\|W\|\big)^2 \right\}^{-1} \cdot \frac{d}{80\,\underline{\sigma}_{\boldsymbol{X}}{}^2}\Big(\frac{10T\min_i\{C^{i*}\}}{(10T-1)\,\underline{\sigma}_{\boldsymbol{Q}}}\Big)^2.$$

Let $\alpha := \underline{\sigma}_{\boldsymbol{X}}{}^2\,g_1 - \widetilde{g}_2/\,\underline{\sigma}_{\boldsymbol{X}}{}^5 > 0$. *Then, we have*

1. $\eta \in (0, \frac{1}{\alpha})$; *and*

2. *the following inequality holds*

$$\sum_{i=1}^N \big(C^i(\boldsymbol{K}^{i\prime}, \boldsymbol{K}^{-i*}) - C^i(\boldsymbol{K}^*)\big) \leq (1-\alpha\eta)\left(\sum_{i=1}^N \big(C^i(\boldsymbol{K}^i, \boldsymbol{K}^{-i*}) - C^i(\boldsymbol{K}^*)\big)\right). \quad \text{(A.6)}$$

Remark 1 We compare the noise condition (A.3) for the vanilla policy gradient method and the condition (3.56) for the natural version in Lemma 19. We mainly focus on the orders of N, T, and $\rho_{\boldsymbol{K}}$, and ignore other constants in (A.3) and (3.56). For the natural version, we need

$$\underline{\sigma}_{\boldsymbol{X}} > \left(\frac{g_2}{g_1}\right)^{1/5} = \mathcal{O}\left((N-1)^{2/5}\,T^{2/5}\,\rho_{\boldsymbol{K}}^{4T/5}\right), \quad \text{(A.7)}$$

and for the vanilla version, we need

$$\underline{\sigma}_{\boldsymbol{X}} > \max\left\{\frac{\widetilde{g}_2}{g_1}, g_3{}^{\frac{7}{2}}\right\}^{1/7} = \mathcal{O}\left((N-1)^{2/7}\,T^{2/7}\,\rho_{\boldsymbol{K}}^{8T/7}\right). \quad \text{(A.8)}$$

The order of $\rho_{\boldsymbol{K}}$ is higher in (A.8), and the orders of T and $(N-1)$ are slightly higher in (A.7). Thus when $\rho_{\boldsymbol{K}}^T$ is small, the vanilla method has a weaker noise assumption. This may happen when the time horizon T is small and the policy $\boldsymbol{K}$ is close to the Nash equilibrium. In contrast, when $\rho_{\boldsymbol{K}}^T$ is large, (A.7) is weaker than (A.8) and the natural method is superior to the vanilla method. Additionally, when $N-1$ is very large (and $\rho_{\boldsymbol{K}}^T$ does not blow up) , (A.8) leads to a weaker assumption.

We also note that other than the noise condition, there are also some slight differences between the step size conditions (A.5) for the vanilla method and (3.58) for the natural method, mainly in the order of $\rho_{\boldsymbol{K}}^{2T}$ appearing in the denominator of I_1, I_3, and I_4.

Proof [Proof of Lemma 1.] We break this proof up into a series of steps.

Step 1: We first consider the consequences of the condition $\eta \leq \min\{I_3, I_4\}$. Straightforward calculations show that when condition $\eta \leq I_3$ is satisfied, the following inequalities hold:

1. $\forall i = 1, \cdots, N$,

$$\|K_t^{i\prime} - K_t^i\| = \eta \|\nabla_{K_t^i} C^i(\boldsymbol{K})\| \leq \frac{\underline{\sigma}_{\boldsymbol{Q}}\, \underline{\sigma}_{\boldsymbol{X}}}{20T\gamma_B C^{i,-i*}}. \tag{A.9}$$

2. $\forall i = 1, \cdots, N$,

$$\begin{aligned} & \eta \left(\frac{\rho_{\boldsymbol{K}}^{2T} - 1}{\rho_{\boldsymbol{K}}^2 - 1} \left(\frac{C^{i,-i*}}{\underline{\sigma}_{\boldsymbol{Q}}} + T\|W\| \right) 2\rho_{\boldsymbol{K}}\, \gamma_B \sum_{t=0}^{T-1} \|\nabla_{K_t^i} C^i(\boldsymbol{K})\| \right) \\ \leq\ & I_3 \left(\frac{\rho_{\boldsymbol{K}}^{2T} - 1}{\rho_{\boldsymbol{K}}^2 - 1} \left(\frac{\sum_{i=1}^N C^{i,-i*}}{\underline{\sigma}_{\boldsymbol{Q}}} + T\|W\| \right) 2\rho_{\boldsymbol{K}}\, \gamma_B T \max_t \{\|\nabla_{K_t^i} C^i(\boldsymbol{K})\|\} \right) \\ \leq\ & \frac{\underline{\sigma}_{\boldsymbol{X}}}{10}. \end{aligned} \tag{A.10}$$

3. $\forall i = 1, \cdots, N$,

$$\eta \leq I_3 \quad \leq \quad \frac{\underline{\sigma}_{\boldsymbol{X}}{}^2}{2\,\underline{\sigma}_{\boldsymbol{X}} + 4\frac{2C^{i,-i*}}{\underline{\sigma}_{\boldsymbol{Q}}} \left(\rho_{\boldsymbol{K}}^{2T} \|\Sigma_0\| + (\rho_{\boldsymbol{K}}^{2T} + 1)\|W\| \right)^2 (\gamma_R + \gamma_B^2 \frac{C^{i,-i*}}{\underline{\sigma}_{\boldsymbol{X}}})}. \tag{A.11}$$

In the case where $\eta \leq I_4$, we have $\forall i = 1, \cdots, N$

$$\begin{aligned} & 4\eta\, k_i \frac{10TC^{i,-i*}}{(10T-1)\,\underline{\sigma}_{\boldsymbol{Q}}} \left(\gamma_R + (\gamma_B)^2 \frac{C^{i,-i*}}{\underline{\sigma}_{\boldsymbol{X}}} \right) \\ & +8\eta \frac{d}{\underline{\sigma}_{\boldsymbol{X}}} \left(\frac{10TC^{i,-i*}}{(10T-1)\,\underline{\sigma}_{\boldsymbol{Q}}} \right)^2 \left(\gamma_R + (\gamma_B)^2 \frac{C^{i,-i*}}{\underline{\sigma}_{\boldsymbol{X}}} \right)^2 \left(\rho_{\boldsymbol{K}}^{2T} \|\Sigma_0\| + (\rho_{\boldsymbol{K}}^{2T} + 1)\|W\| \right)^2 \\ \leq\ & 4I_4 \left(k_i \frac{10TC^{i,-i*}}{(10T-1)\,\underline{\sigma}_{\boldsymbol{Q}}} \left(\gamma_R + (\gamma_B)^2 \frac{C^{i,-i*}}{\underline{\sigma}_{\boldsymbol{X}}} \right) \right. \\ & \left. +2 \frac{d}{\underline{\sigma}_{\boldsymbol{X}}} \left(\frac{10TC^{i,-i*}}{(10T-1)\,\underline{\sigma}_{\boldsymbol{Q}}} \right)^2 \left(\gamma_R + (\gamma_B)^2 \frac{C^{i,-i*}}{\underline{\sigma}_{\boldsymbol{X}}} \right)^2 \left(\rho_{\boldsymbol{K}}^{2T} \|\Sigma_0\| + (\rho_{\boldsymbol{K}}^{2T} + 1)\|W\| \right)^2 \right) \\ \leq\ & \frac{d}{20\,\underline{\sigma}_{\boldsymbol{X}}{}^2} \left(\frac{10T \min_i\{C^{i*}\}}{(10T-1)\,\underline{\sigma}_{\boldsymbol{Q}}} \right)^2 \leq \frac{d}{20\,\underline{\sigma}_{\boldsymbol{X}}{}^2} \left(\frac{10T \max_i\{C^{i,-i*}\}}{(10T-1)\,\underline{\sigma}_{\boldsymbol{Q}}} \right)^2. \end{aligned} \tag{A.12}$$

By (A.9) and Lemma 13 we have

$$\gamma_B \|K_t^{i\prime} - K_t^i\| \leq \frac{\underline{\sigma}_{\boldsymbol{Q}}\, \underline{\sigma}_{\boldsymbol{X}}}{20TC^{i,-i*}} \leq \frac{1}{20T^2}.$$

Therefore, we have $\rho_{\boldsymbol{K},\boldsymbol{K}'} \leq \rho_{\boldsymbol{K}}$ by Lemma 15.

Steps 2 and 3: The results in Steps 2 and 3 in Lemma 19 for the natural policy gradient method still hold for the vanilla policy gradient method by the consequences (A.9) and (A.10). Here we omit the proof and state the following results which will be used in Steps 4 and 5:

$$\left\|\Sigma_{\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i*}}\right\| \leq \frac{10T\, C^{i,-i*}}{(10T-1)\,\underline{\sigma}_{\boldsymbol{Q}}}, \tag{A.13}$$

and

$$\left\| E_{t,i}^{\boldsymbol{K}} - E_{t,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}} \right\| \leq \frac{(\gamma_B)^2 C^{i,-i*}}{\underline{\sigma}_{\boldsymbol{X}}} \left(\frac{2(\rho_{\boldsymbol{K}}^{2T}-1)}{\rho_{\boldsymbol{K}}^2-1} \sum_{j=1,j\neq i}^{N} \left|\!\left|\!\left| \boldsymbol{K}^j - \boldsymbol{K}^{j*} \right|\!\right|\!\right| \right). \tag{A.14}$$

Step 4: We can now estimate the cost difference between using $\boldsymbol{K}^i$ and the update $\boldsymbol{K}^{i\prime}$. By By Lemma 12 we have

$$\begin{aligned}
&C^{i\prime,-i*} - C^{i,-i*} \\
&= \sum_{t=0}^{T-1} \Big[\operatorname{Tr}\big(\Sigma_t^{\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i*}} (K_t^{i\prime} - K_t^i)^\top (R_t^i + (B_t^i)^\top P_{t+1,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}} B_t^i)(K_t^{i\prime} - K_t^i)\big) \\
&+2\operatorname{Tr}\big(\Sigma_t^{\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i*}} (K_t^{i\prime} - K_t^i)^\top E_{t,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}}\big)\Big].
\end{aligned} \tag{A.15}$$

For the vanilla policy gradient method, we have the following update rule

$$K_t^{i'} = K_t^i - \eta \nabla_{K_t^i} C^i(\boldsymbol{K}) = K_t^i - 2\eta\, E_{t,i}^{\boldsymbol{K}}\, \Sigma_t^{\boldsymbol{K}}$$

by Lemma 9. Then plugging in $K_t^{i'} - K_t^i = -2\eta\, E_{t,i}^{\boldsymbol{K}}\, \Sigma_t^{\boldsymbol{K}}$ into (A.15) leads to

$$\begin{aligned}
&C^{i\prime,-i*} - C^{i,-i*} \\
= \;& \sum_{t=0}^{T-1} \Big[4\eta^2 \operatorname{Tr}\big(\Sigma_t^{\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i*}} \Sigma_t^{\boldsymbol{K}} (E_{t,i}^{\boldsymbol{K}})^\top (R_t^i + (B_t^i)^\top P_{t+1,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}} B_t^i) E_{t,i}^{\boldsymbol{K}} \Sigma_t^{\boldsymbol{K}}\big) \\
&\quad -4\eta \operatorname{Tr}\big(\Sigma_t^{\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i*}} \Sigma_t^{\boldsymbol{K}} (E_{t,i}^{\boldsymbol{K}})^\top E_{t,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}}\big)\Big] \\
= \;& \sum_{t=0}^{T-1} \Big[4\eta^2 \operatorname{Tr}\Big(\Sigma_t^{\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i*}} \Sigma_t^{\boldsymbol{K}} (E_{t,i}^{\boldsymbol{K}} - E_{t,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}} + E_{t,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}})^\top (R_t^i + (B_t^i)^\top P_{t+1,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}} B_t^i) \cdot \\
&\quad (E_{t,i}^{\boldsymbol{K}} - E_{t,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}} + E_{t,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}}) \Sigma_t^{\boldsymbol{K}}\Big) - 4\eta \operatorname{Tr}\Big(\Sigma_t^{\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i*}} (E_{t,i}^{\boldsymbol{K}} \Sigma_t^{\boldsymbol{K}} \\
&\quad -E_{t,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}} \Sigma_t^{\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i*}} + E_{t,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}} \Sigma_t^{\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i*}})^\top E_{t,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}}\Big)\Big] \\
= \;& \sum_{t=0}^{T-1} \Big[4\eta^2 \operatorname{Tr}\big(\Sigma_t^{\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i*}} \Sigma_t^{\boldsymbol{K}} (E_{t,i}^{\boldsymbol{K}} - E_{t,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}})^\top (R_t^i + (B_t^i)^\top P_{t+1,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}} B_t^i) \cdot \\
&\qquad (E_{t,i}^{\boldsymbol{K}} - E_{t,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}}) \Sigma_t^{\boldsymbol{K}}\big) + 8\eta^2 \operatorname{Tr}\big(\Sigma_t^{\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i*}} \Sigma_t^{\boldsymbol{K}} (E_{t,i}^{\boldsymbol{K}} - E_{t,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}})^\top \cdot \\
&\qquad (R_t^i + (B_t^i)^\top P_{t+1,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}} B_t^i) E_{t,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}} \Sigma_t^{\boldsymbol{K}}\big) \\
&\; +4\eta^2 \operatorname{Tr}\big(\Sigma_t^{\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i*}} \Sigma_t^{\boldsymbol{K}} (E_{t,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}})^\top (R_t^i + (B_t^i)^\top P_{t+1,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}} B_t^i) E_{t,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}} \Sigma_t^{\boldsymbol{K}}\big) \\
&\; -4\eta \operatorname{Tr}\big(\Sigma_t^{\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i*}} (E_{t,i}^{\boldsymbol{K}} \Sigma_t^{\boldsymbol{K}} - E_{t,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}} \Sigma_t^{\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i*}})^\top E_{t,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}}\big) \\
&\; -4\eta \operatorname{Tr}\big(\Sigma_t^{\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i*}} (E_{t,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}})^\top E_{t,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}} \Sigma_t^{\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i*}}\big)\Big],
\end{aligned} \tag{A.16}$$

where the first equation holds by the updating rule, the second equation holds by adding and subtracting $E_{t,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}}$ and $E_{t,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}}\Sigma_t^{\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i*}}$ terms, and the third equation holds by expanding terms. Now, letting $\omega^2 = \frac{2}{\underline{\sigma}_{\boldsymbol{X}}}$ in

$$2\operatorname{Tr}(A^\top B) = \operatorname{Tr}(A^\top B + B^\top A) \leq \omega^2 \operatorname{Tr}(A^\top A) + \frac{1}{\omega^2}\operatorname{Tr}(B^\top B), \tag{A.17}$$

(which holds for any matrices A and B of the same dimension), the second term in (A.16) can be bounded by

$$\begin{aligned}
& 8\eta^2 \operatorname{Tr}\left(\Sigma_t^{\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i*}}\Sigma_t^{\boldsymbol{K}}(E_{t,i}^{\boldsymbol{K}} - E_{t,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}})^\top (R_t^i + (B_t^i)^\top P_{t+1,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}} B_t^i) E_{t,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}}\Sigma_t^{\boldsymbol{K}}\right) \\
\leq \;& 8\eta^2 \frac{\underline{\sigma}_{\boldsymbol{X}}}{4}\operatorname{Tr}\left((E_{t,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}})^\top E_{t,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}}\right) + 8\eta^2\frac{1}{\underline{\sigma}_{\boldsymbol{X}}}\operatorname{Tr}\left(\Sigma_t^{\boldsymbol{K}}\Sigma_t^{\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i*}}\Sigma_t^{\boldsymbol{K}}(E_{t,i}^{\boldsymbol{K}} - E_{t,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}})^\top \cdot\right. \\
& \left.(R_t^i + (B_t^i)^\top P_{t+1,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}} B_t^i)(R_t^i + (B_t^i)^\top P_{t+1,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}} B_t^i)(E_{t,i}^{\boldsymbol{K}} - E_{t,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}})\Sigma_t^{\boldsymbol{K}}\Sigma_t^{\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i*}}\Sigma_t^{\boldsymbol{K}}\right) \\
.
\end{aligned} \tag{A.18}$$

Now using the fact (A.17) again with $\omega^2 = \frac{2}{\underline{\sigma}_{\boldsymbol{X}}{}^2}$, we can also bound the second last term in (A.16) as follows

$$\begin{aligned}
& -4\eta\operatorname{Tr}\left(\Sigma_t^{\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i*}}(E_{t,i}^{\boldsymbol{K}}\Sigma_t^{\boldsymbol{K}} - E_{t,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}}\Sigma_t^{\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i*}})^\top E_{t,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}}\right) \\
= \;& -4\eta\operatorname{Tr}\left(\Sigma_t^{\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i*}}\left((E_{t,i}^{\boldsymbol{K}} - E_{t,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}})\Sigma_t^{\boldsymbol{K}} + E_{t,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}}(\Sigma_t^{\boldsymbol{K}} - \Sigma_t^{\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i*}})\right)^\top E_{t,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}}\right) \\
= \;& -4\eta\operatorname{Tr}\left(\Sigma_t^{\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i*}}\Sigma_t^{\boldsymbol{K}}(E_{t,i}^{\boldsymbol{K}} - E_{t,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}})^\top E_{t,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}}\right) \\
& -4\eta\operatorname{Tr}\left(\Sigma_t^{\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i*}}(\Sigma_t^{\boldsymbol{K}} - \Sigma_t^{\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i*}})(E_{t,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}})^\top E_{t,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}}\right) \\
\leq \;& 4\eta\frac{\underline{\sigma}_{\boldsymbol{X}}{}^2}{4}\operatorname{Tr}\left((E_{t,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}})^\top E_{t,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}}\right) + \frac{4\eta}{\underline{\sigma}_{\boldsymbol{X}}{}^2}\operatorname{Tr}\left(\Sigma_t^{\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i*}}\Sigma_t^{\boldsymbol{K}}(E_{t,i}^{\boldsymbol{K}} - E_{t,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}})^\top \cdot\right. \\
& \left.(E_{t,i}^{\boldsymbol{K}} - E_{t,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}})\Sigma_t^{\boldsymbol{K}}\Sigma_t^{\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i*}}\right) + 4\eta\|\Sigma_t^{\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i*}}\|\,\|\Sigma_t^{\boldsymbol{K}} - \Sigma_t^{\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i*}}\| \cdot \\
& \operatorname{Tr}\left((E_{t,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}})^\top E_{t,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}}\right) \\
\leq \;& \eta\,\underline{\sigma}_{\boldsymbol{X}}{}^2\operatorname{Tr}\left((E_{t,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}})^\top E_{t,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}}\right) \\
& + \frac{4\eta}{\underline{\sigma}_{\boldsymbol{X}}{}^2}\|E_{t,i}^{\boldsymbol{K}} - E_{t,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}}\|^2\operatorname{Tr}\left(\Sigma_t^{\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i*}}\Sigma_t^{\boldsymbol{K}}\Sigma_t^{\boldsymbol{K}}\Sigma_t^{\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i*}}\right) \\
& + 4\eta\|\Sigma_t^{\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i*}}\|\,\|\Sigma_t^{\boldsymbol{K}} - \Sigma_t^{\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i*}}\|\operatorname{Tr}\left((E_{t,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}})^\top E_{t,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}}\right).
\end{aligned} \tag{A.19}$$

Then plugging the above bounds into (A.16) gives

$$\begin{aligned}
& C^{i\prime,-i*} - C^{i,-i*} \\
\leq \;& \sum_{t=0}^{T-1}\Big[4\eta^2\operatorname{Tr}\left(\Sigma_t^{\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i*}}\Sigma_t^{\boldsymbol{K}}(E_{t,i}^{\boldsymbol{K}} - E_{t,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}})^\top (R_t^i + (B_t^i)^\top P_{t+1,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}} B_t^i) \cdot\right. \\
& \qquad\left.(E_{t,i}^{\boldsymbol{K}} - E_{t,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}})\Sigma_t^{\boldsymbol{K}}\right) + 8\eta^2\frac{\underline{\sigma}_{\boldsymbol{X}}}{4}\operatorname{Tr}\left((E_{t,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}})^\top E_{t,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}}\right) \\
& \quad + 8\eta^2\frac{1}{\underline{\sigma}_{\boldsymbol{X}}}\operatorname{Tr}\left(\Sigma_t^{\boldsymbol{K}}\Sigma_t^{\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i*}}\Sigma_t^{\boldsymbol{K}}(E_{t,i}^{\boldsymbol{K}} - E_{t,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}})^\top (R_t^i + (B_t^i)^\top P_{t+1,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}} B_t^i) \cdot\right.
\end{aligned}$$

$$
\begin{aligned}
& \qquad (R_t^i + (B_t^i)^\top P_{t+1,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}} B_t^i)(E_{t,i}^{\boldsymbol{K}} - E_{t,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}})\Sigma_t^{\boldsymbol{K}}\Sigma_t^{\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i*}}\Sigma_t^{\boldsymbol{K}}\big) \\
& \quad +4\eta^2 \operatorname{Tr}\big(\Sigma_t^{\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i*}}\Sigma_t^{\boldsymbol{K}}(E_{t,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}})^\top(R_t^i + (B_t^i)^\top P_{t+1,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}} B_t^i)E_{t,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}}\Sigma_t^{\boldsymbol{K}}\big) \\
& \quad +\frac{4\eta}{\underline{\sigma}_{\boldsymbol{X}}{}^2}\|E_{t,i}^{\boldsymbol{K}} - E_{t,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}}\|^2 \operatorname{Tr}\big(\Sigma_t^{\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i*}}\Sigma_t^{\boldsymbol{K}}\Sigma_t^{\boldsymbol{K}}\Sigma_t^{\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i*}}\big) \\
& \quad +4\eta\|\Sigma_t^{\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i*}}\|\,\|\Sigma_t^{\boldsymbol{K}} - \Sigma_t^{\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i*}}\| \operatorname{Tr}\big((E_{t,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}})^\top E_{t,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}}\big) \\
& \quad +\eta\,\underline{\sigma}_{\boldsymbol{X}}{}^2 \operatorname{Tr}\big((E_{t,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}})^\top E_{t,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}}\big) - 4\eta\,\underline{\sigma}_{\boldsymbol{X}}{}^2 \operatorname{Tr}\big((E_{t,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}})^\top E_{t,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}}\big)\Big] \\
\leq \; & \sum_{t=0}^{T-1}\Big[\Big(4\eta^2\|\Sigma_t^{\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i*}}\|\,\|\Sigma_t^{\boldsymbol{K}}\|^2 \operatorname{Tr}\big(R_t^i + (B_t^i)^\top P_{t+1,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}} B_t^i\big) \\
& \qquad +\frac{8\eta^2}{\underline{\sigma}_{\boldsymbol{X}}}\|R_t^i + (B_t^i)^\top P_{t+1,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}} B_t^i\|^2 \operatorname{Tr}\big(\Sigma_t^{\boldsymbol{K}}\Sigma_t^{\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i*}}\Sigma_t^{\boldsymbol{K}}\Sigma_t^{\boldsymbol{K}}\Sigma_t^{\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i*}}\Sigma_t^{\boldsymbol{K}}\big) \\
& \qquad +\frac{4\eta}{\underline{\sigma}_{\boldsymbol{X}}{}^2} \operatorname{Tr}\big(\Sigma_t^{\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i*}}\Sigma_t^{\boldsymbol{K}}\Sigma_t^{\boldsymbol{K}}\Sigma_t^{\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i*}}\big)\Big) \cdot \|E_{t,i}^{\boldsymbol{K}} - E_{t,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}}\|^2 \\
& \quad +\big(2\eta^2\,\underline{\sigma}_{\boldsymbol{X}} + 4\eta^2\|\Sigma_t^{\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i*}}\|\|\Sigma_t^{\boldsymbol{K}}\|^2\|R_t^i + (B_t^i)^\top P_{t+1,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}} B_t^i\| \\
& \qquad +\eta\,\underline{\sigma}_{\boldsymbol{X}}{}^2 + 4\eta\|\Sigma_t^{\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i*}}\|\,\|\Sigma_t^{\boldsymbol{K}} - \Sigma_t^{\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i*}}\| - 4\eta\,\underline{\sigma}_{\boldsymbol{X}}{}^2\big) \cdot \\
& \qquad \operatorname{Tr}\big((E_{t,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}})^\top E_{t,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}}\big)\Big],
\end{aligned}
\tag{A.20}
$$

where the first inequality holds by (A.18) and (A.19), and the second inequality holds by the trace inequality (3.33) and rearranging terms. Now we bound the term $\|\Sigma_t^{\boldsymbol{K}}\|$. By (3.39) and (3.49) we have

$$
\begin{aligned}
\|\Sigma_t^{\boldsymbol{K}}\| &= \|\mathcal{G}_{t-1}(\Sigma_0)\| + \Big\|\sum_{s=1}^{t-1} D_{t-1,s} W D_{t-1,s}^\top\Big\| + \|W\| \\
&\leq \rho_{\boldsymbol{K}}^{2T}\|\Sigma_0\| + \big(\rho_{\boldsymbol{K}}^{2T} + 1\big)\|W\|.
\end{aligned}
\tag{A.21}
$$

By (A.3) we have $\underline{\sigma}_{\boldsymbol{X}}{}^2 \geq g_3$, which implies

$$
\begin{aligned}
\underline{\sigma}_{\boldsymbol{X}}{}^2 &\geq 4\frac{10TC^{i,-i*}}{(10T-1)\,\underline{\sigma}_{\boldsymbol{Q}}}\Big(\frac{10TC^{i,-i*}}{(10T-1)\,\underline{\sigma}_{\boldsymbol{Q}}} + \rho_{\boldsymbol{K}}^{2T}\|\Sigma_0\| + \big(\rho_{\boldsymbol{K}}^{2T} + 1\big)\|W\|\Big) \\
&\geq 4\|\Sigma_t^{\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i*}}\|\,\Big(\|\Sigma_t^{\boldsymbol{K}}\| + \|\Sigma_t^{\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i*}}\|\Big) \\
&\geq 4\|\Sigma_t^{\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i*}}\|\,\|\Sigma_t^{\boldsymbol{K}} - \Sigma_t^{\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i*}}\|.
\end{aligned}
\tag{A.22}
$$

Now, since $\big\|\Sigma_{\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i*}}\big\| \leq \frac{2\,C^{i,-i*}}{\underline{\sigma}_{\boldsymbol{Q}}}$ by (A.13), we can bound the step size condition in (A.11) by

$$
\eta \leq \frac{\underline{\sigma}_{\boldsymbol{X}}{}^2}{2\,\underline{\sigma}_{\boldsymbol{X}} + 4\|\Sigma_t^{\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i*}}\|\|\Sigma_t^{\boldsymbol{K}}\|^2\|R_t^i + (B_t^i)^\top P_{t+1,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}} B_t^i\|}.
\tag{A.23}
$$

Combining (A.22) and (A.23) gives

$$
2\eta^2\,\underline{\sigma}_{\boldsymbol{X}} + 4\eta^2\|\Sigma_t^{\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i*}}\|\|\Sigma_t^{\boldsymbol{K}}\|^2\|R_t^i + (B_t^i)^\top P_{t+1,i}^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}} B_t^i\|
$$

$$
\begin{aligned}
&+\eta\, \underline{\sigma}_{\boldsymbol{X}}{}^2 + 4\eta \|\Sigma_t^{\boldsymbol{K}^{i\prime},\boldsymbol{K}^{-i*}}\| \, \|\Sigma_t^{\boldsymbol{K}} - \Sigma_t^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}}\| - 4\eta\, \underline{\sigma}_{\boldsymbol{X}}{}^2 \\
\leq \quad & -\eta\, \underline{\sigma}_{\boldsymbol{X}}{}^2 . \qquad \text{(A.24)}
\end{aligned}
$$

Hence, using this in (A.20), we have

$$
C^{i\prime,-i*} - C^{i,-i*} \leq \eta\, h^i_{\text{diff}} \sum_{t=0}^{T-1} \|E^{\boldsymbol{K}}_{t,i} - E^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}}_{t,i}\|^2 - \eta\, \underline{\sigma}_{\boldsymbol{X}}{}^2 \sum_{t=0}^{T-1} \operatorname{Tr}\left((E^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}}_{t,i})^\top E^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}}_{t,i} \right), \tag{A.25}
$$

where

$$
\begin{aligned}
h^i_{\text{diff}} := {} & 4\eta\, k_i \frac{10T C^{i,-i*}}{(10T-1)\, \underline{\sigma}_{\boldsymbol{Q}}} \left(\rho^{2T}_{\boldsymbol{K}} \|\Sigma_0\| + \left(\rho^{2T}_{\boldsymbol{K}} + 1\right)\|W\|\right)^2 \left(\gamma_R + (\gamma_B)^2 \frac{C^{i,-i*}}{\underline{\sigma}_{\boldsymbol{X}}}\right) \\
& + 8\eta \frac{d}{\underline{\sigma}_{\boldsymbol{X}}} \left(\frac{10T C^{i,-i*}}{(10T-1)\, \underline{\sigma}_{\boldsymbol{Q}}}\right)^2 \left(\rho^{2T}_{\boldsymbol{K}} \|\Sigma_0\| + \left(\rho^{2T}_{\boldsymbol{K}} + 1\right)\|W\|\right)^4 \left(\gamma_R + (\gamma_B)^2 \frac{C^{i,-i*}}{\underline{\sigma}_{\boldsymbol{X}}}\right)^2 \\
& + 4 \frac{d}{\underline{\sigma}_{\boldsymbol{X}}{}^2} \left(\frac{10T C^{i,-i*}}{(10T-1)\, \underline{\sigma}_{\boldsymbol{Q}}}\right)^2 \left(\rho^{2T}_{\boldsymbol{K}} \|\Sigma_0\| + \left(\rho^{2T}_{\boldsymbol{K}} + 1\right)\|W\|\right)^2 .
\end{aligned}
$$

Therefore, by (A.14), (A.13), Lemma 11, and Lemma 13,

$$
\begin{aligned}
C^{i\prime,-i*} - C^{i,-i*} \quad \leq \quad & \eta\, h^i_{\text{diff}}\, T \left[\frac{(\gamma_B)^2 C^{i,-i*}}{\underline{\sigma}_{\boldsymbol{X}}} \frac{2(\rho^{2T}_{\boldsymbol{K}} - 1)}{\rho^2_{\boldsymbol{K}} - 1} \sum_{j=1, j\neq i}^{N} |||\boldsymbol{K}^j - \boldsymbol{K}^{j*}||| \right]^2 \\
& -\eta\, \underline{\sigma}_{\boldsymbol{X}}{}^2 \sum_{t=0}^{T-1} \operatorname{Tr}\left((E^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}}_{t,i})^\top E^{\boldsymbol{K}^i,\boldsymbol{K}^{-i*}}_{t,i} \right) \\
\leq \quad & \eta\, h_{\text{glob}} \left(\sum_{j=1, j\neq i}^{N} |||\boldsymbol{K}^j - \boldsymbol{K}^{j*}||| \right)^2 - \eta \frac{\underline{\sigma}_{\boldsymbol{X}}{}^2\, \underline{\sigma}_{\boldsymbol{R}}}{\|\Sigma_{\boldsymbol{K}^*}\|} \left(C^{i,-i*} - C^{i*} \right), \qquad \text{(A.26)}
\end{aligned}
$$

where

$$
\begin{aligned}
& h_{\text{glob}} \\
= {} & 4T \Bigg[\eta\, k_i \frac{10T \max_i\{C^{i,-i*}\}}{(10T-1)\, \underline{\sigma}_{\boldsymbol{Q}}} \left(\rho^{2T}_{\boldsymbol{K}} \|\Sigma_0\| + \left(\rho^{2T}_{\boldsymbol{K}} + 1\right)\|W\|\right)^2 \left(\gamma_R + (\gamma_B)^2 \frac{\max_i\{C^{i,-i*}\}}{\underline{\sigma}_{\boldsymbol{X}}}\right) \\
& + 2\eta \frac{d}{\underline{\sigma}_{\boldsymbol{X}}} \left(\frac{10T \max_i\{C^{i,-i*}\}}{(10T-1)\, \underline{\sigma}_{\boldsymbol{Q}}}\right)^2 \left(\rho^{2T}_{\boldsymbol{K}} \|\Sigma_0\| + \left(\rho^{2T}_{\boldsymbol{K}} + 1\right)\|W\|\right)^4 \\
& \left(\gamma_R + (\gamma_B)^2 \frac{\max_i\{C^{i,-i*}\}}{\underline{\sigma}_{\boldsymbol{X}}}\right)^2 + \frac{d}{\underline{\sigma}_{\boldsymbol{X}}{}^2} \left(\frac{10T \max_i\{C^{i,-i*}\}}{(10T-1)\, \underline{\sigma}_{\boldsymbol{Q}}}\right)^2 \\
& \left(\rho^{2T}_{\boldsymbol{K}} \|\Sigma_0\| + \left(\rho^{2T}_{\boldsymbol{K}} + 1\right)\|W\|\right)^2 \Bigg] \cdot \left[\frac{(\gamma_B)^2 \max_i\{C^{i,-i*}\}}{\underline{\sigma}_{\boldsymbol{X}}} \frac{2(\rho^{2T}_{\boldsymbol{K}} - 1)}{\rho^2_{\boldsymbol{K}} - 1} \right]^2 .
\end{aligned} \tag{A.27}
$$

Step 5: Finally we can establish the one step contraction. Using (A.26), we have

$$\begin{aligned} C^{i\prime,-i*} - C^{i*} &= C^{i\prime,-i*} - C^{i,-i*} + C^{i,-i*} - C^{i*} \\ &\leq \left(1 - \eta \frac{\underline{\sigma}_{\boldsymbol{X}}{}^2 \, \underline{\sigma}_{\boldsymbol{R}}}{\|\Sigma_{\boldsymbol{K}^*}\|}\right) \left(C^{i,-i*} - C^{i*}\right) + \eta \, h_{\text{glob}} \left(\sum_{j=1, j\neq i}^{N} |\!|\!| \boldsymbol{K}^j - \boldsymbol{K}^{j*} |\!|\!| \right)^2 . \end{aligned} \tag{A.28}$$

Hence by Lemma 12 and (3.42), we have

$$\sum_{j=1, j\neq i}^{N} \left(C^{j,-j*} - C^{j*}\right) \geq \frac{\underline{\sigma}_{\boldsymbol{X}} \, \underline{\sigma}_{\boldsymbol{R}}}{T} \sum_{j=1, j\neq i}^{N} |\!|\!| \boldsymbol{K}^j - \boldsymbol{K}^{j*} |\!|\!|^2 \geq \frac{\underline{\sigma}_{\boldsymbol{X}} \, \underline{\sigma}_{\boldsymbol{R}}}{T(N-1)} \left(\sum_{j=1, j\neq i}^{N} |\!|\!| \boldsymbol{K}^j - \boldsymbol{K}^{j*} |\!|\!| \right)^2 ,$$

and thus

$$C^{i\prime,-i*} - C^{i*} \leq \big(1 - \eta \frac{\underline{\sigma}_{\boldsymbol{X}}{}^2 \, \underline{\sigma}_{\boldsymbol{R}}}{\|\Sigma_{\boldsymbol{K}^*}\|}\big)\big(C^{i,-i*} - C^{i*}\big) + \eta \; h_{\text{glob}} \frac{T(N-1)}{\underline{\sigma}_{\boldsymbol{X}} \, \underline{\sigma}_{\boldsymbol{R}}} \left(\sum_{j=1, j\neq i}^{N} (C^{j,-j*} - C^{j*}) \right) . \tag{A.29}$$

Summing up (A.29) for $i = 1, \cdots, N$, we have

$$\sum_{i=1}^{N} \left(C^{i\prime,-i*} - C^{i*}\right) \leq \left(1 - \eta \frac{\underline{\sigma}_{\boldsymbol{X}}{}^2 \, \underline{\sigma}_{\boldsymbol{R}}}{\|\Sigma_{\boldsymbol{K}^*}\|} + \eta (N-1) h_{\text{glob}} \frac{T(N-1)}{\underline{\sigma}_{\boldsymbol{X}} \, \underline{\sigma}_{\boldsymbol{R}}}\right) \left(\sum_{i=1}^{N} (C^{i,-i*} - C^{i*}) \right) . \tag{A.30}$$

Since $\eta \leq I_4$, we have (A.12) and then

$$\begin{aligned} h^i_{\text{diff}} &\leq (4 + \frac{1}{20}) \frac{d}{\underline{\sigma}_{\boldsymbol{X}}{}^2} \left(\frac{10T \max_i \{C^{i,-i*}\}}{(10T-1) \, \underline{\sigma}_{\boldsymbol{Q}}} \right)^2 \left(\rho_{\boldsymbol{K}}^{2T} \|\Sigma_0\| + \big(\rho_{\boldsymbol{K}}^{2T} + 1\big)\|W\|\right)^2 \\ &\leq 5 \frac{d}{\underline{\sigma}_{\boldsymbol{X}}{}^2} \left(\frac{\max_i \{C^{i,-i*}\}}{\underline{\sigma}_{\boldsymbol{Q}}} \right)^2 \left(\rho_{\boldsymbol{K}}^{2T} \|\Sigma_0\| + \big(\rho_{\boldsymbol{K}}^{2T} + 1\big)\|W\|\right)^2 , \quad \text{and} \quad h_{\text{glob}} \leq \bar{h}_{\text{glob}}, \end{aligned}$$

where $\bar{h}_{\text{glob}}$ is given by

$$\bar{h}_{\text{glob}} = 5 \, T \, d \, \frac{\max_i \{C^{i,-i*}\}^4 (\gamma_B)^4}{\underline{\sigma}_{\boldsymbol{Q}}{}^2 \, \underline{\sigma}_{\boldsymbol{X}}{}^4} \left[\frac{2(\rho_{\boldsymbol{K}}^{2T} - 1)}{\rho_{\boldsymbol{K}}^2 - 1} \right]^2 \left(\rho_{\boldsymbol{K}}^{2T} \|\Sigma_0\| + \big(\rho_{\boldsymbol{K}}^{2T} + 1\big)\|W\|\right)^2 . \tag{A.31}$$

Under condition (A.3), we have

$$\underline{\sigma}_{\boldsymbol{X}}{}^2 \, g_1 - \frac{\widetilde{g}_2}{\underline{\sigma}_{\boldsymbol{X}}{}^5} = \frac{\underline{\sigma}_{\boldsymbol{X}}{}^2 \, \underline{\sigma}_{\boldsymbol{R}}}{\|\Sigma_{\boldsymbol{K}^*}\|} - (N-1)^2 \bar{h}_{\text{glob}} \frac{T}{\underline{\sigma}_{\boldsymbol{X}} \, \underline{\sigma}_{\boldsymbol{R}}} > 0,$$

which indicates that $\alpha\eta > 0$. Since $\eta \leq \frac{\|\Sigma_{\boldsymbol{K}^*}\|}{\underline{\sigma}_{\boldsymbol{X}}{}^2 \underline{\sigma}_{\boldsymbol{R}}}$ by assumption, we have

$$\eta \leq \frac{\|\Sigma_{\boldsymbol{K}^*}\|}{\underline{\sigma}_{\boldsymbol{X}}{}^2 \, \underline{\sigma}_{\boldsymbol{R}}} < \left(\frac{\underline{\sigma}_{\boldsymbol{X}}{}^2 \, \underline{\sigma}_{\boldsymbol{R}}}{\|\Sigma_{\boldsymbol{K}^*}\|} - (N-1)^2 \bar{h}_{\text{glob}} \frac{T}{\underline{\sigma}_{\boldsymbol{X}} \, \underline{\sigma}_{\boldsymbol{R}}} \right)^{-1} = \left(\underline{\sigma}_{\boldsymbol{X}}{}^2 \, g_1 - \frac{\widetilde{g}_2}{\underline{\sigma}_{\boldsymbol{X}}{}^5} \right)^{-1} .$$

Recall that in the statement we define $\alpha = {\underline{\sigma}_{\boldsymbol{X}}}^2 \, g_1 - \widetilde{g}_2 / {\underline{\sigma}_{\boldsymbol{X}}}^5$. Therefore we have $\alpha\eta < 1$. Along with (A.30), we obtain the one-step contraction (A.6). ■

We now explain the main differences between the above proof and the proof for the natural policy gradient method. Since for the vanilla method we have $K_t^{i'} - K_t^i = -2\eta \, E_{t,i}^{\boldsymbol{K}} \, \Sigma_t^{\boldsymbol{K}}$ rather than $K_t^{i'} - K_t^i = -2\eta \, E_{t,i}^{\boldsymbol{K}}$ in the case of natural version, the extra $\Sigma_t^{\boldsymbol{K}}$ term needs to be dealt with when calculating the individual cost difference $C^{i\prime,-i*} - C^{i,-i*}$. More precisely, the presence of $\Sigma_t^{\boldsymbol{K}}$ causes more terms which need to be bounded (see Equation A.18) when finding an upper bound on $C^{i\prime,-i*} - C^{i,-i*}$ in (A.16), and the term $\|\Sigma_t^{\boldsymbol{K}}\|$ also needs to be bounded above (see Equation A.21). Due to these extra terms, the amount of noise needed in the system has a more complex form given in (A.3), where a new g_3 is introduced to guarantee (A.24) holds. This further demonstrates that the natural policy gradient method can be considered as a normalized version of the vanilla policy gradient method.

References

James Drew Bagnell and Jeff Schneider. Covariant policy search. In *Proceedings of 18th International Joint Conference on Artificial Intelligence*, pages 1019–1024, August 2003.

David Balduzzi, Sebastien Racaniere, James Martens, Jakob Foerster, Karl Tuyls, and Thore Graepel. The mechanics of n-player differentiable games. In *International Conference on Machine Learning*, pages 354–363. PMLR, 2018.

Tamer Başar and Geert Jan Olsder. *Dynamic Non-Cooperative Game Theory*. Society for Industrial and Applied Mathematics, 1998.

Matteo Basei, Xin Guo, Anran Hu, and Yufei Zhang. Logarithmic regret for episodic continuous-time linear-quadratic reinforcement learning over a finite-time horizon. *Journal of Machine Learning Research*, 23(178):1–34, 2022.

Michael Bowling and Manuela Veloso. Multiagent learning using a variable learning rate. *Artificial Intelligence*, 136(2):215–250, 2002.

Jingjing Bu, Lillian Ratliff, and Mehran Mesbahi. Global convergence of policy gradient for sequential zero-sum linear quadratic dynamic games. *arXiv preprint arXiv:1911.04672*, 2019.

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

Tanner Fiez, Benjamin Chasnov, and Lillian Ratliff. Implicit learning dynamics in Stackelberg games: Equilibria characterization, convergence analysis, and empirical study. In *International Conference on Machine Learning*, pages 3133–3144. PMLR, 2020.

Jakob Foerster, Richard Chen, Maruan Al-Shedivat, Shimon Whiteson, Pieter Abbeel, and Igor Mordatch. Learning with opponent-learning awareness. In *Proceedings of the 17th*

International Conference on Autonomous Agents and MultiAgent Systems, pages 122–130, 2018.

Ian Gemp and Sridhar Mahadevan. Global convergence to the equilibrium of GANs using variational inequalities. *arXiv preprint arXiv:1808.01531*, 2018.

Ben Hambly, Renyuan Xu, and Huining Yang. Policy gradient methods for the noisy linear quadratic regulator over a finite horizon. *SIAM Journal on Control and Optimization*, 59 (5):3359–3391, 2021.

Rein Houthooft, Xi Chen, Yan Duan, John Schulman, Filip De Turck, and Pieter Abbeel. VIME: Variational information maximizing exploration. *Advances in neural information processing systems*, 29, 2016.

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

Junqi Jin, Chengru Song, Han Li, Kun Gai, Jun Wang, and Weinan Zhang. Real-time bidding with multi-agent reinforcement learning in display advertising. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, pages 2193–2201, 2018.

Sham Kakade. A natural policy gradient. *Advances in Neural Information Processing Systems*, 14, 2001.

Jean-Michel Lasry and Pierre-Louis Lions. Mean field games. *Japanese Journal of Mathematics*, 2(1):229–260, 2007.

Alistair Letcher, David Balduzzi, Sébastien Racaniere, James Martens, Jakob Foerster, Karl Tuyls, and Thore Graepel. Differentiable game mechanics. *Journal of Machine Learning Research*, 20(1):3032–3071, 2019.

Horia Mania, Stephen Tu, and Benjamin Recht. Certainty equivalence is efficient for linear quadratic control. In *Advances in Neural Information Processing Systems*, pages 10154–10164, 2019.

Eric Mazumdar, Lillian Ratliff, Michael Jordan, and Sosale Shankara Sastry. Policy-gradient algorithms have no guarantees of convergence in linear quadratic games. In *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems*, pages 860–868, 2020a.

Eric Mazumdar, Lillian Ratliff, and Shankar Sastry. On gradient-based learning in continuous games. *SIAM Journal on Mathematics of Data Science*, 2(1):103–131, 2020b.

Panayotis Mertikopoulos, Christos Papadimitriou, and Georgios Piliouras. Cycles in adversarial regularized learning. In *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 2703–2717. SIAM, 2018.

Jorge Nocedal and Stephen Wright. *Numerical optimization*. Springer Science & Business Media, 2006.

Jan Peters and Stefan Schaal. Natural actor-critic. *Neurocomputing*, 71(7-9):1180–1190, 2008.

Aravind Rajeswaran, Kendall Lowrey, Emanuel Todorov, and Sham Kakade. Towards generalization and simplicity in continuous control. *Advances in Neural Information Processing Systems*, 30, 2017.

Masoud Roudneshin, Jalal Arabneydi, and Amir Aghdam. Reinforcement learning in nonzero-sum linear quadratic deep structured games: Global convergence of policy optimization. In *2020 59th IEEE Conference on Decision and Control*, pages 512–517. IEEE, 2020.

Joan Saniuk and Ian Rhodes. A matrix inequality associated with bounds on solutions of algebraic Riccati and Lyapunov equations. *IEEE Transactions on Automatic Control*, 32 (8):739–740, 1987. doi: 10.1109/TAC.1987.1104700.

John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *International Conference on Machine Learning*, pages 1889–1897. PMLR, 2015.

Shai Shalev-Shwartz, Shaked Shammah, and Amnon Shashua. Safe, multi-agent, reinforcement learning for autonomous driving. *arXiv preprint arXiv:1610.03295*, 2016.

Satinder Singh, Michael Kearns, and Yishay Mansour. Nash convergence of gradient dynamics in general-sum games. In *Proceedings of the Sixteenth conference on Uncertainty in Artificial Intelligence*, pages 541–548, 2000.

Xinliang Song, Tonghan Wang, and Chongjie Zhang. Convergence of multi-agent learning with a finite step size in general-sum games. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*, pages 935–943, 2019.

Emmanouil-Vasileios Vlatakis-Gkaragkounis, Lampros Flokas, Thanasis Lianeas, Panayotis Mertikopoulos, and Georgios Piliouras. No-regret learning and mixed Nash equilibria: They do not mix. *Advances in Neural Information Processing Systems*, 33:1380–1391, 2020.

Haoran Wang, Thaleia Zariphopoulou, and Xun Yu Zhou. Reinforcement learning in continuous time and space: A stochastic control approach. *Journal of Machine Learning Research*, 21:198–1, 2020.

Sheng-De Wang, Te-Son Kuo, and Chen-Fa Hsu. Trace bounds on the solution of the algebraic matrix Riccati and Lyapunov equation. *IEEE Transactions on Automatic Control*, 31(7): 654–656, 1986.

Chongjie Zhang and Victor Lesser. Multi-agent learning with policy prediction. In *Twenty-fourth AAAI conference on artificial intelligence*, 2010.

Junzi Zhang, Jongho Kim, Brendan O'Donoghue, and Stephen Boyd. Sample efficient reinforcement learning with REINFORCE. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 10887–10895, 2021a.

Kaiqing Zhang, Zhuoran Yang, and Tamer Başar. Policy optimization provably converges to Nash equilibria in zero-sum linear quadratic games. In *Advances in Neural Information Processing Systems*, 2019.

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, 2021b.