Learning Zero-Sum Linear Quadratic Games with Improved Sample Complexity and Last-Iterate Convergence

Master Thesis

Author(s):
Wu, Jiduan

Publication date:
2023

Permanent link:
https://doi.org/https://doi.org/10.3929/ethz-b-000612636

Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Learning Zero-Sum Linear Quadratic Games with Improved Sample Complexity and Last-Iterate Convergence

Master Thesis

Jiduan Wu

Monday 15th May, 2023

Supervisors: Ilyas Fatkhullin
Dr. Anas Barakat
Prof. Dr. Niao He

Department of Computer Science, ETH Zürich

Abstract

Zero-sum Linear Quadratic (LQ) games are fundamental in optimal control and can be used (i) as a dynamic game formulation for risk-sensitive or robust control, or (ii) as a benchmark setting for multi-agent reinforcement learning with two competing agents in continuous state-control spaces. In contrast to the well-studied single-agent linear quadratic regulator problem, zero-sum LQ games entail solving a challenging nonconvex-nonconcave min-max problem with an objective function that lacks coercivity. Recently, Zhang et al. Zhang et al. [2021b] discovered an implicit regularization property of natural policy gradient methods which is crucial for safety-critical control systems since it preserves the robustness of the controller during learning. Moreover, in the model-free setting where the knowledge of model parameters is not available, Zhang et al. proposed the first polynomial sample complexity algorithm to reach an ε-neighborhood of the Nash equilibrium while maintaining the desirable implicit regularization property. In this work, we propose a simpler nested Zeroth-Order (ZO) algorithm improving sample complexity by several orders of magnitude. Our main results are two-fold: (i) our first result guarantees a $\widetilde{\mathcal{O}}(\varepsilon^{-3})$ sample complexity under the same assumptions using a single-point ZO estimator. Furthermore, when the estimator is replaced by a two-point estimator, our method enjoys even faster convergence with a $\widetilde{\mathcal{O}}(\varepsilon^{-2})$ sample complexity; (ii) secondly, to the best of our knowledge, we provide the first last-iterate convergence result for the nested algorithm that seeks NE of zero-sum LQ games in addition to the diminishing gradient norms. The complexity analyses are provided for both deterministic and stochastic cases. Our key improvements in the sample complexity rely on a more sample-efficient nested algorithm design and finer control of the ZO natural gradient estimation error. As for the last-iterate convergence results, the analysis relies on the implicit regularization property of the algorithm, and the derivation of the sample complexity in the stochastic case reuses our improvement mentioned earlier.

Contents

Chapter 1

Introduction

While policy optimization has a long history in control for unknown and parameterized system models (see for e.g., Makila and Toivonen [1987]), recent successes in reinforcement learning and continuous control tasks have renewed the interest in direct policy search thanks to its flexibility and scalability to high-dimensional problems. Despite these desirable features, theoretical guarantees for policy gradient methods have remained elusive until very recently because of the nonconvexity of the induced optimization landscape. In particular, in contrast to control-theoretic approaches which are often model-based and estimate the system dynamics first before designing optimal controllers, the computational and sample complexities of model-free policy gradient methods were only recently analyzed. We refer the interested reader to a nice recent survey about learning control policies Hu et al. [2022]. For instance, while the classic Linear Quadratic Regulator (LQR) problem induces a nonconvex optimization problem over the set of stable control gain matrices, the gradient domination property Polyak [1963] and the coercivity of the cost function respectively allow to derive global convergence to optimal policies for policy gradient methods and ensure stable feedback policies at each iteration Fazel et al. [2018]. As exact gradients are often unavailable when system dynamics are unknown, derivative-free optimization techniques using cost values have been employed to design model-free policy gradient methods to solve LQR problems Fazel et al. [2018]. Alternative approaches to solve LQR include system identification Fiechter [1997], Ljung [1998], iterative solution of Algebraic Riccati Equation Hewer [1971], Lancaster and Rodman [1995] and convex semi-definite program formulations Balakrishnan and Vandenberghe [2003]. However, such methods are not easily adaptable to the simulation-based model-free setting.

Besides the desired stability constraint, other requirements such as robustness and risk sensitivity constraints also play an important role in the design of controllers for safety-critical control systems. Indeed, system perturbations, modeling imprecision, and adversarial uncertainty are ubiquitous in control systems and may lead to severe degradation in performance Bhattacharyya and Keel [1995], Campi and James [1996]. Robustness constraints can be incorporated into control design via different approaches including using statistical models for disturbances such as for linear quadratic Gaussian design, adopting a game theory perspective via designing 'minimax' controllers and incorporating an $\mathcal{H}_\infty$ norm bound of input-output operators as in $\mathcal{H}_\infty$ control Başar and Bernhard [1995]. Classical linear models for robust control include the LQ disturbances attenuation problem and the linear exponential quadratic Gaussian problem which are well-known to be equivalent to zero-sum LQ games Başar and Bernhard [1995], Mageirou and Ho [1977], Zhang et al. [2021a]. Besides its relevance for robust control problem formulation, zero-sum LQ games also constitute a benchmark problem for multi-agent continuous control problems involving two competing agents. However, solving this problem faces (at least) two distinct challenges requiring to deal with (a) a constrained nonconvex-nonconcave problem and (b) lack of coercivity, unlike for the classic LQR problem for which descent over the objective ensures feasibility and stability of the iterates during learning.

While the formulation of zero-sum LQ games dates back at least to the seventies Mageirou and Ho [1977][*], the sample complexity analysis of model-free policy gradient algorithms solving this problem was only recently explored in the literature Zhang et al. [2021b]. More precisely, Zhang et al. Zhang et al. [2021b] showed that an ε-Nash equilibrium of finite horizon zero-sum LQ games can be learned via nested model-free Natural Policy Gradient (NPG) algorithms with polynomial sample complexity in the accuracy ε. Interestingly, the aforementioned algorithms enjoy an Implicit Regularization (IR) property which maintains the robustness of the controllers during learning Zhang et al. [2021a,b]. In particular, the iterates of the algorithms are guaranteed to stay in some feasible set where the worst-case cost is finite without using any explicit regularization or projection operation. In the present work, we show that significantly less samples are required to guarantee both the IR property and the convergence to an ε-Nash equilibrium of the zero-sum LQ games problem while only having access to ZO information. Our contributions can be summarized as follows:

Contributions. Our main result states that our derivative-free nested policy gradient algorithm requires $\widetilde{\mathcal{O}}(\varepsilon^{-3})$ samples to reach an ε-neighborhood of the Nash equilibrium (NE) of the zero-sum LQ games problem, improving over the best-known-so-far $\widetilde{\mathcal{O}}(\varepsilon^{-9})$[†] total sample complexity established in Zhang et al. [2021b]. We also show that our algorithm enjoys the IR property upon choosing adequate values for ZO estimation parameters such as the batch sizes and the perturbation radius which are less restrictive compared to prior work Zhang et al. [2021b]. Our improvement follows from (a) a simpler algorithm design reducing the number of calls to the inner-loop maximizing procedure, (b) a better sample complexity to solve the inner maximization problem and (c) an improved sample complexity for solving the resulting minimization problem in our outer-loop procedure using a careful decomposition of the estimation error caused by policy gradient estimation. We further improve the sample complexity to $\widetilde{\mathcal{O}}(\varepsilon^{-2})$ using a two-point ZO estimator under a stronger sampling assumption. (d) We provide the last-iterate convergence results for both deterministic and stochastic settings, which to the best of our knowledge are the first convergence results using the last-iterate measure for zero-sum LQ games.

Thesis organization. The rest of this thesis is structured as follows. In Chapter 1.1, we discuss related work. In Chapter 1.2, we introduce the stochastic zero-sum LQ games problem together with useful background. We present our model-free nested natural policy gradient algorithm to solve the problem in Chapter 2 and Chapter 3 presents our main results along with a proof sketch to highlight the key steps leading to sample complexity improvement and the last-iterate convergence. We conclude this thesis with possible future directions. The proofs of our results and the detailed version of some results are deferred to Appendix A.

[*] This formulation is under the continuous-time setting.

[†] Notice that the total sample complexity was not provided in Zhang et al. [2021b] but can be easily derived from their results, see Remark 3.6 for more details.

1.1 Related work

Policy optimization for LQ problems. Compared to zero-sum LQ games, policy optimization for single-agent LQ problems is a well-understood topic. Theoretical guarantees for model-based and model-free algorithms searching for the optimal policy were established in Fazel et al. [2018] for the discrete-time infinite-horizon setting. Several subsequent works improved over the polynomial sample complexity in Fazel et al. [2018] using single and two-point ZO estimation Malik et al. [2019], Mohammadi et al. [2020]. Additionally, the LQ model has been studied under different settings including finite-horizon Hambly et al. [2021] and continuous-time Fatkhullin and Polyak [2021], Giegrich et al. [2022], Mohammadi et al. [2021]. First-order methods have also been recently investigated for solving LQR Ju et al. [2023], Yang et al. [2019]. In Bu and Mesbahi [2020], they provided convergence analysis for possibly indefinite infinite-horizon LQR problems. In Guo and Hu [2022], they designed Goldstein subdifferential algorithms to solve the nonsmooth $\mathcal{H}_\infty$ control problem and left sample complexity analysis in the model-free setting as an important future direction. Other related problems include Markovian jump systems (Sun and Fazel [2021]), output control design (Fatkhullin and Polyak [2021], Furieri et al. [2020], Zhao et al. [2022]), decentralized control (Feng and Lavaei [2019], Li et al. [2020]), receding-horizon policy gradient methods (Zhang and Başar [2023]), and nonlinear dynamics (Han et al. [2022]). Interested readers are referred to the thorough review paper Hu et al. [2022] on policy optimization methods for learning control policies.

Zero-sum LQ games and beyond. Recent research efforts have been devoted to studying the more challenging zero-sum LQ games problem Bu et al. [2019], Zhang et al. [2019, 2021a,b]. In Zhang et al. [2019], they proposed projected nested gradient-based algorithms in which the projection step is difficult to implement in practice. Later, Bu et al. [2019] removed the projection step, but their analysis requires access to the exact solution of the inner maximization problem and cannot be easily extended to the model-free case. Meanwhile, Zhang et al. [2021a] introduced a nested natural gradient-based algorithm that demonstrates the IR property for the infinite-horizon $\mathcal{H}_2/\mathcal{H}_\infty$ control problem in the model-based case, where they utilize the equivalent zero-sum game formulation and design model-free algorithms without sample complexity analysis. In the model-free setting, Al-Tamimi et al. [2007] proposed a Q-learning-based method to solve zero-sum LQ games without providing a sample complexity analysis. In the context of mean-field games, counterparts of LQR and zero-sum LQ games were developed in Carmona et al. [2020, 2019], where the formulation of mean-field zero-sum LQ games reduces to two zero-sum LQ game problems. Recently, a N-player general-sum game formulation of LQR was studied in Hambly et al. [2022], Mazumdar et al. [2019], Yang [2022]. However, such a problem in the 2-player case is different from our zero-sum formulation. More generally, a tabular setting of two-player zero-sum games is considered in Chen et al. [2023] and the first finite-sample guarantees are provided for independent-learning algorithms.

Mixed $\mathcal{H}_2/\mathcal{H}_\infty$ & Risk-sensitive LQ control. It is well-known that mix $\mathcal{H}_2/\mathcal{H}_\infty$ problems can be formulated as risk-sensitive control problems or zero-sum dynamical games Glover and Doyle [1988], and the solutions of these two classes of problems oftentimes inspire each other Zhang et al. [2021a]. A nice presentation on the history of the connection among them can be found in Başar and Bernhard [1995], Zhang et al. [2021a]. Here we focus more on recent developments. Before Zhang et al. [2021a,b] provided the first results on the implicit regularization property and convergence of policy optimization methods, policy optimization methods had been widely applied to solve mixed $\mathcal{H}_2/\mathcal{H}_\infty$ control design problems with great empirical successes. Borrowing ideas from robust control theory, Zhang et al. [2020] identified the stability issue of the robust adversarial reinforcement learning problem on LQ systems and proposed a double-loop algorithm using proper initialization as the solution. Besides the nested natural gradient algorithm, Cui and Jiang [2022] designed a dual-loop algorithm instead where the outer loop approximately solves the generalized algebraic Riccati equation iteratively. Their algorithm also enjoys a last-iterate linear convergence in the deterministic case, which is similar to our result and can be extended to a model-free version. The continuous-time counterpart is studied in Cui and Molu [2022], Molu [2023]. In this thesis, we focus on the nested natural gradient algorithm and reveal more insights into its convergence and sample complexity properties.

1.2 Preliminaries

Notations. For any matrix $M \in \mathbb{R}^{n\times n}$, we denote by $\|M\|$ and $\|M\|_F$ its operator and Frobenius norms respectively. The spectral radius of a matrix M is denoted by $\rho(M)$ and a matrix is said to be (Schur) stable if $\rho(M) < 1$, i.e., all the absolute values of the eigenvalues of the matrix M are (strictly) smaller than 1. The smallest eigenvalue of a symmetric matrix M is denoted by $\lambda_{\min}(M)$. For N diagonal matrices X_i for $i \in \{0, \cdots N-1\}$ for some integer $N \geq 1$, the block-diagonal matrix with diagonal entries $X_0, \cdots, X_{N-1}$ is denoted by $\mathrm{diag}(X_{0-(N-1)})$. The uniform distribution over a set S is denoted as $\mathrm{Unif}(S)$.

Stochastic Zero-Sum Linear Quadratic Dynamic Games. We consider the zero-sum LQ games problem (following the exposition in Zhang et al. [2021b]) where the system state evolves as follows:

$$x_{h+1} = A_h x_h + B_h u_h + D_h w_h + \xi_h,\ h \in \{0, \cdots, N-1\}, \tag{1.1}$$

where N is a finite nonzero horizon, $x_0 \in \mathbb{R}^m$ is an initial random state and where for any stage $h \in \{0, \cdots, N-1\}$, $x_h \in \mathbb{R}^m$ is the system state, $u_h \in \mathbb{R}^d$ and $w_h \in \mathbb{R}^n$ are the control inputs of the min and max players respectively[‡] and ξ_h is a random variable describing noisy perturbations to the system while A_h, B_h, D_h are (possibly) time-dependent system matrices with appropriate dimensions.

Assumption 1.1. *The initial state x_0 and the noise ξ_h for $h \in \{0, \cdots, N-1\}$ are independent random variables following a distribution with zero-mean and positive-definite covariance. Moreover, there exists a positive scalar ϑ such that for all $h \in \{0, \cdots, N-1\}$, $\|x_0\| \leq \vartheta$ and $\|\xi_h\| \leq \vartheta$ almost surely.*[§]

Our objective is to solve the following zero-sum game:

$$\inf_{(u_h)} \sup_{(w_h)} \mathbb{E}_{\boldsymbol{\xi}} \left[\sum_{h=0}^{N-1} (x_h^\top Q_h x_h + u_h^\top R_h^u u_h - w_h^\top R_h^w w_h) + c_N \right] \tag{1.2}$$

where $c_N := x_N^\top Q_N x_N$ and $\boldsymbol{\xi} := \left[x_0^\top, \xi_0^\top, \cdots, \xi_{N-1}^\top\right]^\top$ and the system states follow the linear time-varying system dynamics described in (1.1) and for every $h \in \{0, \cdots, N-1\}$, $Q_h \succeq 0, R_h^u, R_h^w \succ 0$ are symmetric matrices defining the quadratic objective. In view of our robust control motivation, the two players can be seen as a min controller and a max

‡These controls depend on the history of state-control pairs at each time step h for now, stationary control policies will be sufficient as will be mentioned later on.

§The almost sure boundedness can be relaxed to consider sub-Gaussian distributions as noticed in prior work Furieri and Kamgarpour [2020], Malik et al. [2019].

disturbance. Under standard assumptions which we do not mention here for brevity[¶], the saddle-point control policies solving (1.2) are unique and have the linear state-feedback form. Thus, we can restrict our search to gain matrices $K_h \in \mathbb{R}^{d \times m}$ and $L_h \in \mathbb{R}^{n \times m}$ such that the controls are given by $u_h = -K_h x_h, w_h = -L_h x_h$ for $h \in \{0, \cdots, N-1\}$. Therefore, we will mainly focus on solving the following min-max policy optimization problem resulting from (1.2):

$$\min_{(K_h)} \max_{(L_h)} \mathbb{E}_{\xi} \left[\sum_{h=0}^{N-1} x_h^\top M_h x_h + c_N \right], \tag{1.3}$$

where $M_h := Q_h + K_h^\top R_h^u K_h - L_h^\top R_h^w L_h$ and the system state follows the dynamics $x_{h+1} = (A_h - B_h K_h - D_h L_h) x_h + \xi_h$ for $h \in \{0, \cdots, N-1\}$.

Compact reformulation To simplify the exposition and our analysis, we rewrite problem (1.3) under a more compact form following the reformulation proposed in Zhang et al. [2021b]. Consider the following notations:

$$\begin{aligned}
\boldsymbol{x} &:= [x_0^\top, \cdots, x_N^\top]^\top, \boldsymbol{u} := [u_0^\top, \cdots, u_{N-1}^\top]^\top, \\
\boldsymbol{w} &:= [w_0^\top, \cdots, w_{N-1}^\top]^\top, \boldsymbol{\xi} = [x_0^\top, \xi_0^\top, \cdots, \xi_{N-1}^\top]^\top, \\
\boldsymbol{A} &:= \begin{bmatrix} \mathbf{0}_{m \times mN} & \mathbf{0}_{m \times m} \\ \mathrm{diag}(A_{0-(N-1)}) & \mathbf{0}_{mN \times m} \end{bmatrix}, \boldsymbol{Q} := \mathrm{diag}(Q_{0-N}), \\
\boldsymbol{D} &:= \begin{bmatrix} \mathbf{0}_{m \times nN} \\ \mathrm{diag}(D_{0-(N-1)}) \end{bmatrix}, \boldsymbol{B} := \begin{bmatrix} \mathbf{0}_{m \times dN} \\ \mathrm{diag}(B_{0-(N-1)}) \end{bmatrix}, \\
\boldsymbol{R}^u &:= \mathrm{diag}(R^u_{0-(N-1)}), \boldsymbol{R}^w := \mathrm{diag}(R^w_{0-(N-1)}), \\
\boldsymbol{K} &:= \begin{bmatrix} \mathrm{diag}(K_{0-(N-1)}) & \mathbf{0}_{dN \times m} \end{bmatrix}, \qquad (1.4) \\
\boldsymbol{L} &:= \begin{bmatrix} \mathrm{diag}(L_{0-(N-1)}) & \mathbf{0}_{nN \times m} \end{bmatrix}. \qquad (1.5)
\end{aligned}$$

We denote by $\mathcal{S}_1 \subset \mathbb{R}^{dN \times m(N+1)}$ and $\mathcal{S}_2 \subset \mathbb{R}^{nN \times m(N+1)}$ the matrix subspaces induced by the sparsity patterns described in (1.4), (1.5) for the gain matrices $\boldsymbol{K}$ and $\boldsymbol{L}$ respectively. The subspaces $\mathcal{S}_1$, $\mathcal{S}_2$ where we search for the NE solution $(\boldsymbol{K}^*, \boldsymbol{L}^*)$, are of dimensions $d_{\boldsymbol{K}} := dmN$ and $d_{\boldsymbol{L}} := nmN$ respectively. Then, problem (1.3) can be rewritten as:

$$\min_{\boldsymbol{K} \in \mathcal{S}_1} \max_{\boldsymbol{L} \in \mathcal{S}_2} \mathcal{G}(\boldsymbol{K}, \boldsymbol{L}) := \mathbb{E}_{\boldsymbol{\xi}}[\boldsymbol{x}^\top (\boldsymbol{Q} + \boldsymbol{K}^\top \boldsymbol{R}^u \boldsymbol{K} - \boldsymbol{L}^\top \boldsymbol{R}^w \boldsymbol{L}) \boldsymbol{x}], \tag{1.6}$$

where the transition dynamics are described by $\boldsymbol{x} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{B}\boldsymbol{u} + \boldsymbol{D}\boldsymbol{w} + \boldsymbol{\xi} = (\boldsymbol{A} - \boldsymbol{B}\boldsymbol{K} - \boldsymbol{D}\boldsymbol{L})\boldsymbol{x} + \boldsymbol{\xi}$. Notice that our search for gain matrices $\boldsymbol{K}, \boldsymbol{L}$ is restricted to the matrices of the form described in (1.4), (1.5) as this set of sparse matrices is sufficient to find the NE we are looking for. For any gain matrices $\boldsymbol{K}$ and $\boldsymbol{L}$, we can rewrite the objective function value $\mathcal{G}(\boldsymbol{K}, \boldsymbol{L})$ as follows:

$$\mathcal{G}(\boldsymbol{K}, \boldsymbol{L}) = \mathbb{E}_{\boldsymbol{\xi}}[\mathcal{G}_{\boldsymbol{\xi}}(\boldsymbol{K}, \boldsymbol{L})] = \mathrm{Tr}(\boldsymbol{P}_{\boldsymbol{K},\boldsymbol{L}} \boldsymbol{\Sigma}_0) = \mathrm{Tr}\big((\boldsymbol{Q} + \boldsymbol{K}^\top \boldsymbol{R}^u \boldsymbol{K} - \boldsymbol{L}^\top \boldsymbol{R}^w \boldsymbol{L}) \boldsymbol{\Sigma}_{\boldsymbol{K},\boldsymbol{L}}\big),$$

where $\mathcal{G}_{\boldsymbol{\xi}}(\boldsymbol{K}, \boldsymbol{L}) := \boldsymbol{\xi}^\top \boldsymbol{P}_{\boldsymbol{K},\boldsymbol{L}} \boldsymbol{\xi}$, $\boldsymbol{\Sigma}_0 := \mathbb{E}_{\boldsymbol{\xi}}[\boldsymbol{\xi}\boldsymbol{\xi}^\top] \succ 0$ (see Assumption 1.1) and the matrices $\boldsymbol{P}_{\boldsymbol{K},\boldsymbol{L}}$, $\boldsymbol{\Sigma}_{\boldsymbol{K},\boldsymbol{L}} := \mathbb{E}_{\boldsymbol{\xi}}[\mathrm{diag}(x_0 x_0^\top, \cdots, x_N x_N^\top)]$ are the unique solutions to the recursive Lyapunov equations

$$\boldsymbol{P}_{\boldsymbol{K},\boldsymbol{L}} = \boldsymbol{A}_{\boldsymbol{K},\boldsymbol{L}}^\top \boldsymbol{P}_{\boldsymbol{K},\boldsymbol{L}} \boldsymbol{A}_{\boldsymbol{K},\boldsymbol{L}} + \boldsymbol{Q} + \boldsymbol{K}^\top \boldsymbol{R}^u \boldsymbol{K} - \boldsymbol{L}^\top \boldsymbol{R}^w \boldsymbol{L}, \tag{1.7}$$

$$\boldsymbol{\Sigma}_{\boldsymbol{K},\boldsymbol{L}} = \boldsymbol{A}_{\boldsymbol{K},\boldsymbol{L}} \boldsymbol{\Sigma}_{\boldsymbol{K},\boldsymbol{L}} \boldsymbol{A}_{\boldsymbol{K},\boldsymbol{L}}^\top + \boldsymbol{\Sigma}_0, \tag{1.8}$$

[¶]See Assumption 2.4 in Zhang et al. [2021b] for instance and the explanations in Remark 2.5 therein for further details, see also Başar and Bernhard [1995].

where $\boldsymbol{A}_{\boldsymbol{K},\boldsymbol{L}} := \boldsymbol{A} - \boldsymbol{B}\boldsymbol{K} - \boldsymbol{D}\boldsymbol{L}$. The objective $\mathcal{G}(\boldsymbol{K}, \boldsymbol{L})$ is nonconvex-nonconcave in general (see Lemma 3.1 in Zhang et al. [2021b]). From the above compact formulation, we observe that the finite-horizon case can be seen as a special case of infinite-horizon zero-sum LQ games with special constraints on sparsity patterns of matrices defined in (1.4), (1.5). Using this perspective, the time-varying case where model parameters such as A_h, B_h vary over $h \in \{0, \cdots, N-1\}$ is included in the compact formulation as shown in Zhang et al. [2021b].

Policy Gradients. The gradients of $\mathcal{G}$ w.r.t. $\boldsymbol{K}, \boldsymbol{L}$ (see Zhang et al. [2021b]) are given by the following expressions:

$$\nabla_{\boldsymbol{K}}\mathcal{G}(\boldsymbol{K}, \boldsymbol{L}) = 2\boldsymbol{F}_{\boldsymbol{K},\boldsymbol{L}}\boldsymbol{\Sigma}_{\boldsymbol{K},\boldsymbol{L}}\,, \quad \boldsymbol{F}_{\boldsymbol{K},\boldsymbol{L}} := (\boldsymbol{R}^u + \boldsymbol{B}^\top \boldsymbol{P}_{\boldsymbol{K},\boldsymbol{L}}\boldsymbol{B})\boldsymbol{K} - \boldsymbol{B}^\top \boldsymbol{P}_{\boldsymbol{K},\boldsymbol{L}}(\boldsymbol{A} - \boldsymbol{D}\boldsymbol{L})\,, \tag{1.9}$$

$$\nabla_{\boldsymbol{L}}\mathcal{G}(\boldsymbol{K}, \boldsymbol{L}) = 2\boldsymbol{E}_{\boldsymbol{K},\boldsymbol{L}}\boldsymbol{\Sigma}_{\boldsymbol{K},\boldsymbol{L}}\,, \quad \boldsymbol{E}_{\boldsymbol{K},\boldsymbol{L}} := (-\boldsymbol{R}^w + \boldsymbol{D}^\top \boldsymbol{P}_{\boldsymbol{K},\boldsymbol{L}}\boldsymbol{D})\boldsymbol{L} - \boldsymbol{D}^\top \boldsymbol{P}_{\boldsymbol{K},\boldsymbol{L}}(\boldsymbol{A} - \boldsymbol{B}\boldsymbol{K})\,. \tag{1.10}$$

If $\boldsymbol{P}_{\boldsymbol{K},\boldsymbol{L}} \succeq 0$ and $\boldsymbol{R}^w - \boldsymbol{D}^\top \boldsymbol{P}_{\boldsymbol{K},\boldsymbol{L}}\boldsymbol{D} \succ 0$ for a stationary point $(\boldsymbol{K}, \boldsymbol{L})$ of $\mathcal{G}$, then this stationary point is the unique NE of the game (see Lemma 3.2 in Zhang et al. [2021b]).

Remark 1.2. *In our finite-horizon scenario, $\rho(\boldsymbol{A}_{\boldsymbol{K},\boldsymbol{L}}) = 0$ since $\boldsymbol{A}_{\boldsymbol{K},\boldsymbol{L}}^{N+1} = 0$, see Lemma A.22. This means that the pair $(\boldsymbol{K}, \boldsymbol{L})$ defined in (1.4)-(1.5) is always stable. This property leads to the existence and uniqueness of the solution of the Lyapunov equation, see Lemma A.22.*

Chapter 2

Nested Derivative-Free Natural Policy Gradient (NPG) Algorithm

In this chapter, we present our model-free and derivative-free nested NPG algorithm inspired by the recent work Zhang et al. [2021b]. We start with the deterministic exact version of the algorithm assuming access to exact natural policy gradients.

2.1 Exact Nested NPG Algorithm

To prepare the stage for the model-free setting, we briefly introduce the nested NPG algorithm in the deterministic setting, i.e., when we have access to the policy gradients w.r.t. both control variables $\boldsymbol{K}, \boldsymbol{L}$ as reported in (1.9). This algorithm was considered for example in Zhang et al. [2021b] and we follow a similar exposition in this subchapter. We first solve the inner maximization problem in (1.6) for any fixed control gain matrix $\boldsymbol{K}$ to obtain a solution $\boldsymbol{L}(\boldsymbol{K})$ before solving the outer-loop minimization problem with the resulting objective $\mathcal{G}(\boldsymbol{K}, \boldsymbol{L}(\boldsymbol{K}))$. The following proposition that we report here from Lemma 3.3 in Zhang et al. [2021b] guarantees that there exists a unique solution $\boldsymbol{L}(\boldsymbol{K})$ to the inner maximization problem whenever the control gain matrix $\boldsymbol{K}$ lies in a set which is known to contain the optimal control gain matrix solving the min-max problem.

Lemma 2.1. *(Inner-loop well-definedness condition Zhang et al. [2021b]) Consider the Riccati equation*

$$\boldsymbol{P}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})} = \boldsymbol{Q} + \boldsymbol{K}^\top \boldsymbol{R}^u \boldsymbol{K} + \boldsymbol{A}_{\boldsymbol{K}}^\top \widetilde{\boldsymbol{P}}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})} \boldsymbol{A}_{\boldsymbol{K}}, \tag{2.1}$$

where $\boldsymbol{A}_{\boldsymbol{K}} := \boldsymbol{A} - \boldsymbol{B}\boldsymbol{K}$ *and* $\widetilde{\boldsymbol{P}}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})} := \boldsymbol{P}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})} + \boldsymbol{P}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})} \boldsymbol{D}(\boldsymbol{R}^w - \boldsymbol{D}^\top \boldsymbol{P}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})} \boldsymbol{D})^{-1} \boldsymbol{D}^\top \boldsymbol{P}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})}$ *and define the set*

$$\mathcal{K} := \left\{ \boldsymbol{K} \in \mathcal{S}_1 \;|(2.1) \textit{ admits a solution } \boldsymbol{P}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})} \succeq 0, \textit{and } \boldsymbol{R}^w - \boldsymbol{D}^\top \boldsymbol{P}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})} \boldsymbol{D} \succ 0 \right\}^* . \tag{2.2}$$

Then, for any $\boldsymbol{K} \in \mathcal{K}$*, there exists a unique solution* $\boldsymbol{L}(\boldsymbol{K})$ *to the inner maximization problem in 1.6 given by*

$$\boldsymbol{L}(\boldsymbol{K}) = (-\boldsymbol{R}^w + \boldsymbol{D}^\top \boldsymbol{P}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})} \boldsymbol{D})^{-1} \boldsymbol{D}^\top \boldsymbol{P}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})} (\boldsymbol{A} - \boldsymbol{B}\boldsymbol{K}).$$

Moreover, for any $\boldsymbol{K} \in \mathcal{K}$ *and any* $\boldsymbol{L} \in \mathcal{S}_2$, $\boldsymbol{P}_{\boldsymbol{K},\boldsymbol{L}} \preceq \boldsymbol{P}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})}$. *The proof is deferred to Appendix A.7.1.*

We are now ready to introduce the nested NPG algorithm which can be written as follows using positive step-sizes τ_1, τ_2 and indices $k \geq 0, t \geq 0$ for the inner and outer loops respectively:

$$\text{Inner loop: } \boldsymbol{L}_{k+1} = \boldsymbol{L}_k + \tau_1 \boldsymbol{E}_{\boldsymbol{K}_t,\boldsymbol{L}_k},\ k = 0, 1, \ldots \tag{2.3}$$

$$\text{Outer loop: } \boldsymbol{K}_{t+1} = \boldsymbol{K}_t - \tau_2 \boldsymbol{F}_{\boldsymbol{K}_t,\boldsymbol{L}(\boldsymbol{K}_t)},\ t = 0, 1, \ldots \tag{2.4}$$

The use of natural policy gradients and the nested structure of the algorithm have an important IR effect: They guarantee that the iterates remain in the feasible set defining admissible stable controls without any explicit regularization of the problem, as shown in Zhang et al. [2021b]. Maintaining the feasibility of the iterates during learning is important since it translates to preserving the robustness of the controllers in the face of adversarial perturbations. More formally, it was shown in Theorem 3.7 in Zhang et al. [2021b] that (a) the sequence $(\boldsymbol{P}_{\boldsymbol{K}_t,\boldsymbol{L}(\boldsymbol{K}_t)})_t$ is well-defined, satisfies the conditions in (2.2) for every $t \geq 0$ and is (most importantly) non-increasing and bounded below in the sense of positive definiteness; and as a consequence (b) for every $t \geq 0, \boldsymbol{K}_t \in \mathcal{K}$ when $\boldsymbol{K}_0 \in \mathcal{K}$.

2.2 Derivative-Free Nested NPG Algorithm

In this subchapter, we describe our algorithm to solve problem (1.6) in the model-free setting where we do not have access to exact gradients. In this setting for which system parameters are unknown, namely $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{D}, \boldsymbol{Q}, \boldsymbol{R}^u, \boldsymbol{R}^w$, we can simulate system trajectories, $(x_h)_{h=0,\cdots,N}$, using a pair of control gain matrices $(\boldsymbol{K}, \boldsymbol{L})$ and we have access to ZO information consisting of the (stochastic) cost $\mathcal{G}_{\boldsymbol{\xi}}(\boldsymbol{K}, \boldsymbol{L})$ incurred by this pair of controllers. In Algorithms 1 and 2, we denote by (1P) and (2P) the single-point and two-point ZO estimation procedures respectively.

Inner loop ZO-NPG algorithm (see Algorithm 1). In the light of the update rule (2.3) in the deterministic exact setting, for any fixed matrix $\boldsymbol{K}$ and any time index k, we replace the gradient $\nabla_{\boldsymbol{L}}\mathcal{G}(\boldsymbol{K}, \boldsymbol{L}_k)$ and the covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{K},\boldsymbol{L}_k}$ by ZO estimates denoted as $\widetilde{\nabla}_{\boldsymbol{L}}\mathcal{G}(\boldsymbol{K}, \boldsymbol{L}_k)$ and $\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{K},\boldsymbol{L}_k}$ respectively. By sampling two independent trajectories at each sample step, we firstly obtain an unbiased estimate of the gradient w.r.t. $\boldsymbol{L}$ of the smoothed objective $\mathcal{G}_{r_1}(\boldsymbol{K}, \boldsymbol{L}_k)$ in the sense that: $\mathbb{E}[\widetilde{\nabla}_{\boldsymbol{L}}\mathcal{G}(\boldsymbol{K}, \boldsymbol{L})] = \nabla_{\boldsymbol{L}}\mathcal{G}_{r_1}(\boldsymbol{K}, \boldsymbol{L}_k)$, $\mathcal{G}_{r_1}(\boldsymbol{K}, \boldsymbol{L}_k) := \mathbb{E}[\mathcal{G}(\boldsymbol{K}, \boldsymbol{L} + r_1\boldsymbol{U})]$, where $\boldsymbol{U}$ is uniformly sampled on a unit ball in $\mathcal{S}_1$. Secondly, we obtain an unbiased estimate of the covariance matrix, i.e., $\mathbb{E}[\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{K},\boldsymbol{L}_k}] = \boldsymbol{\Sigma}_{\boldsymbol{K},\boldsymbol{L}_k}$. For any given $\boldsymbol{K} \in \mathcal{K}$ and other proper choices of parameters, Algorithm 1 outputs $\boldsymbol{L}_{T_{in}}$ that satisfies the accuracy requirement. The detailed sampling and computation procedures can be found in Algorithm 1 of Zhang et al. [2021b] and here we repeat the algorithm as Algorithm 1 for completeness.

Remark 2.2. *In Algorithm 1 of Zhang et al. [2021b], two-point zeroth-order estimation is not covered but can be easily adapted. Here we include both single-point and two-point estimations in Algorithm 1.*

Outer loop ZO-NPG (see Algorithm 2). Similarly to the inner loop procedure, we now replace the unknown quantities $\nabla_{\boldsymbol{K}}\mathcal{G}(\boldsymbol{K}_t, \boldsymbol{L}(\boldsymbol{K}_t))$ and $\boldsymbol{\Sigma}_{\boldsymbol{K}_t,\boldsymbol{L}(\boldsymbol{K}_t)}$ in (2.4) by ZO estimates. As for the exact solution $\boldsymbol{L}(\boldsymbol{K}_t)$ to the inner maximization problem, we use the output of the inner loop ZO-NPG algorithm instead. Notice that the zeroth-order single-point estimate $\widetilde{\nabla}_{\boldsymbol{K}}\mathcal{G}(\boldsymbol{K}, \boldsymbol{L})$ as defined in Algorithm 2 is an unbiased estimate of the gradient w.r.t. $\boldsymbol{K}$ of the smoothed objective $\mathcal{G}_{r_2}(\boldsymbol{K}, \boldsymbol{L})$ in the sense that: $\mathbb{E}[\widetilde{\nabla}_{\boldsymbol{K}}\mathcal{G}(\boldsymbol{K}, \boldsymbol{L})] = \nabla_{\boldsymbol{K}}\mathcal{G}_{r_2}(\boldsymbol{K}, \boldsymbol{L})$, $\mathcal{G}_{r_2}(\boldsymbol{K}, \boldsymbol{L}) := \mathbb{E}[\mathcal{G}(\boldsymbol{K} + r_2\boldsymbol{V}, \boldsymbol{L})]$, where $\boldsymbol{V}$ is uniformly sampled on a unit ball in $\mathcal{S}_2$.

Comparison to the derivative free NPG algorithm in Zhang et al. [2021b]. We would like to point out here an important difference between our proposed algorithm and the zeroth order NPG algorithm in Zhang et al. [2021b] which inspired this work. This

Algorithm 1 (Algorithm 1 in Zhang et al. [2021b]) Inner-loop Zeroth-Order Maximization Oracle

Input: $\boldsymbol{K} \in \mathcal{K}$, $\boldsymbol{L}_0$, number of iterations T_{in}, sample size M_1, perturbation radius r_1, stepsize τ_1, horizon N, dimension $d_{\boldsymbol{L}} = nmN$.

Output: $\boldsymbol{L}_{\text{out}} = \boldsymbol{L}_{T_{in}}$.

for $k = 0, 1, \cdots, T_{in} - 1$ **do**

  Call Algorithm 1 to obtain $\boldsymbol{L}_t$.

  for $i = 0, 1, \cdots, M_1 - 1$ **do**

    Sample policies

- (1P): Sample $\boldsymbol{L}_k^i = \boldsymbol{L}_k + r_1\boldsymbol{U}_i$ where $\boldsymbol{U}_i$ is uniformly drawn from $\mathcal{S}_2$ with $\|\boldsymbol{U}_i\|_F = 1$.
- (2P): Sample $\boldsymbol{L}_k^{1,i} = \boldsymbol{L}_k + r_1\boldsymbol{U}_i$, $\boldsymbol{L}_k^{2,i} = \boldsymbol{L}_k - r_1\boldsymbol{U}_i$ where $\boldsymbol{U}_i$ is uniformly drawn from $\mathcal{S}_2$ with $\|\boldsymbol{U}_i\|_F = 1$.

    Simulate trajectories

- (1P): Simulate a first trajectory using control $(\boldsymbol{K}, \boldsymbol{L}_k^i)$ for horizon N under one realization of noises $\boldsymbol{\xi}_i$ and collect the cost $\mathcal{G}_{\boldsymbol{\xi}_i}(\boldsymbol{K}, \boldsymbol{L}_k^i)$.
- (2P): Simulate two trajectories using controls $(\boldsymbol{K}, \boldsymbol{L}_k^{1,i})$ and $(\boldsymbol{K}, \boldsymbol{L}_k^{2,i})$ for horizon N under the same realization of noises $\boldsymbol{\xi}_i$ and collect $\mathcal{G}_{\boldsymbol{\xi}_i}(\boldsymbol{K}, \boldsymbol{L}_k^{1,i})$, $\mathcal{G}_{\boldsymbol{\xi}_i}(\boldsymbol{K}, \boldsymbol{L}_k^{2,i})$.

    Simulate another independent trajectory using control $(\boldsymbol{K}, \boldsymbol{L}_k)$ for horizon N starting from $x_{0,i}$ and compute

$$\widetilde{\boldsymbol{\Sigma}}^i_{\boldsymbol{K},\boldsymbol{L}_k} = \operatorname{diag}\big(x_{0,i}x_{0,i}^\top, \cdots, x_{N,i}x_{N,i}^\top\big).$$

  end for

  Update $\boldsymbol{L}_{k+1} = \boldsymbol{L}_k + \tau_1 \widetilde{\nabla}_{\boldsymbol{L}}\mathcal{G}(\boldsymbol{K}, \boldsymbol{L}_k)\widetilde{\boldsymbol{\Sigma}}^{-1}_{\boldsymbol{K},\boldsymbol{L}_k}$ where $\widetilde{\nabla}_{\boldsymbol{L}}\mathcal{G}(\boldsymbol{K}, \boldsymbol{L}_k)$ equals

$$\text{(1P):}\ \frac{1}{M_1}\sum_{i=0}^{M_1-1}\frac{d_{\boldsymbol{L}}}{r_1}\mathcal{G}_{\boldsymbol{\xi}_i}(\boldsymbol{K}, \boldsymbol{L}_k^i)\boldsymbol{U}_i,$$

$$\text{(2P):}\ \frac{1}{M_1}\sum_{i=0}^{M_1-1}\frac{d_{\boldsymbol{L}}}{2r_1}\big(\mathcal{G}_{\boldsymbol{\xi}_i}(\boldsymbol{K}, \boldsymbol{L}_k^{1,i}) - \mathcal{G}_{\boldsymbol{\xi}_i}(\boldsymbol{K}, \boldsymbol{L}_k^{2,i})\big)\boldsymbol{U}_i,$$

  and $\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{K},\boldsymbol{L}_k} = \frac{1}{M_1}\sum_{i=0}^{M_1-1}\widetilde{\boldsymbol{\Sigma}}^i_{\boldsymbol{K},\boldsymbol{L}_k}$.

end for

difference lies in the outer loops of the algorithms: namely comparing Algorithm 2 and Algorithm 2 in Zhang et al. [2021b]. In their work, at each time step t of the outer loop, Algorithm 1 (which provides an approximate solution of the maximization problem) is called for each perturbation $\boldsymbol{K}_t^m$ (for $m = 0, \cdots, M_2 - 1$) of the control gain matrix $\boldsymbol{K}_t$ (see step 6: in their Algorithm 2) in order to control the gradient estimation error. In contrast to their work, observe that we only call Algorithm 1 once at each outer loop iteration t in Algorithm 2 and use the approximate maximizer $\boldsymbol{L}_t$ to compute our zeroth order estimates for updating the control gain matrix sequence $(\boldsymbol{K}_t)$. This observation is crucial for our sample complexity improvement as will be discussed in the next chapter.

Remark 2.3. *The single-point estimation Flaxman et al. [2004] might suffer high variance for a small smoothing radius r. We can reduce the variance and hence the sample complexity by using two-point estimation.*

Algorithm 2 Outer-loop Nested Natural Policy Gradient

Input: $\boldsymbol{K}_0 \in \mathcal{K}$, number of iterations T, sample size M_2, perturbation radius r_2, stepsize τ_2, horizon N, dimension $d_{\boldsymbol{K}} = dmN$.

Output: $\boldsymbol{K}_{\text{out}} = \boldsymbol{K}_i$ where $i \sim \text{Unif}(\{0, \cdots, T-1\})$.

1: **for** $t = 0, 1, \cdots, T$ **do**

2: Call Algorithm 1 to obtain $\boldsymbol{L}_t$.

3: **for** $m = 0, 1, \cdots, M_2 - 1$ **do**

4: Sample policies

- (1P): Sample $\boldsymbol{K}_t^m = \boldsymbol{K}_t + r_2 \boldsymbol{V}_m$ where $\boldsymbol{V}_m$ is uniformly drawn from $\mathcal{S}_1$ with $\|\boldsymbol{V}_m\|_F = 1$.
- (2P): Sample $\boldsymbol{K}_t^{1,m} = \boldsymbol{K}_t + r_2 \boldsymbol{V}_m$, $\boldsymbol{K}_t^{2,m} = \boldsymbol{K}_t - r_2 \boldsymbol{V}_m$ where $\boldsymbol{V}_m$ is uniformly drawn from $\mathcal{S}_1$ with $\|\boldsymbol{V}_m\|_F = 1$.

5: Simulate trajectories

- (1P): Simulate a first trajectory using control $(\boldsymbol{K}_t^m, \boldsymbol{L}_t)$ for horizon N under one realization of noises $\boldsymbol{\xi}_m$ and collect the cost $\mathcal{G}_{\boldsymbol{\xi}_m}(\boldsymbol{K}_t^m, \boldsymbol{L}_t)$.
- (2P): Simulate two trajectories using controls $(\boldsymbol{K}_t^{1,m}, \boldsymbol{L}_t)$ and $(\boldsymbol{K}_t^{2,m}, \boldsymbol{L}_t)$ for horizon N under the same realization of noises $\boldsymbol{\xi}_m$ and collect $\mathcal{G}_{\boldsymbol{\xi}_m}(\boldsymbol{K}_t^{1,m}, \boldsymbol{L}_t)$, $\mathcal{G}_{\boldsymbol{\xi}_m}(\boldsymbol{K}_t^{2,m}, \boldsymbol{L}_t)$.

6: Simulate another independent trajectory using control $(\boldsymbol{K}_t, \boldsymbol{L}_t)$ for horizon N starting from $x_{0,m}$ and compute

$$\widetilde{\boldsymbol{\Sigma}}^m_{\boldsymbol{K}_t,\boldsymbol{L}_t} = \text{diag}\big(x_{0,m} x_{0,m}^\top, \cdots, x_{N,m} x_{N,m}^\top\big).$$

7: **end for**

8: Update $\boldsymbol{K}_{t+1} = \boldsymbol{K}_t - \tau_2 \widetilde{\nabla}_{\boldsymbol{K}} \mathcal{G}(\boldsymbol{K}_t, \boldsymbol{L}_t) \widetilde{\boldsymbol{\Sigma}}^{-1}_{\boldsymbol{K}_t,\boldsymbol{L}_t}$ where $\widetilde{\nabla}_{\boldsymbol{K}} \mathcal{G}(\boldsymbol{K}_t, \boldsymbol{L}_t)$ equals

$$\text{(1P)}: \frac{1}{M_2} \sum_{m=0}^{M_2-1} \frac{d_{\boldsymbol{K}}}{r_2} \mathcal{G}_{\boldsymbol{\xi}_m}(\boldsymbol{K}_t^m, \boldsymbol{L}_t) \boldsymbol{V}_m,$$

$$\text{(2P)}: \frac{1}{M_2} \sum_{m=0}^{M_2-1} \frac{d_{\boldsymbol{K}}}{2r_2} \big(\mathcal{G}_{\boldsymbol{\xi}_m}(\boldsymbol{K}_t^{1,m}, \boldsymbol{L}_t) - \mathcal{G}_{\boldsymbol{\xi}_m}(\boldsymbol{K}_t^{2,m}, \boldsymbol{L}_t)\big) \boldsymbol{V}_m,$$

and $\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{K}_t,\boldsymbol{L}_t} = \frac{1}{M_2} \sum_{m=0}^{M_2-1} \widetilde{\boldsymbol{\Sigma}}^m_{\boldsymbol{K}_t,\boldsymbol{L}_t}$.

9: **end for**

Chapter 3

Sample Complexity and Convergence Analysis

In this chapter, (i) we analyze the sample complexity of the algorithm introduced in Chapter 2, i.e., the number of samples of system trajectories required to reach an ε-neighborhood of the NE. (ii) In addition to the average norm of natural gradients, we will present last-iterate convergence results in terms of cost values.

When using estimated natural gradients, the monotonicity of the sequence $(\boldsymbol{P}_{\boldsymbol{K}_t,\boldsymbol{L}(\boldsymbol{K}_t)})_{t\geq 0}$ is violated and the iterates $(\boldsymbol{K}_t)$ are no longer guaranteed to lie in the set $\mathcal{K}$ as we previously described in Chapter 2.1 for the deterministic counterpart of the algorithm. In the following, we consider a subset $\hat{\mathcal{K}}$ of $\mathcal{K}$ for which we prove that IR holds (with high probability) similarly to the result we reported in Lemma 2.1 for good enough ZO estimates as we shall precisely state later in this chapter. Consider an initial point $\boldsymbol{K}_0 \in \mathcal{K}$ and define the set

$$\hat{\mathcal{K}} := \Big\{ \boldsymbol{K} \in \mathcal{S}_1 \mid (2.1) \text{ admits a solution } \boldsymbol{P}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})} \succeq 0 \\ \text{and } \boldsymbol{P}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})} \preceq \boldsymbol{P}_{\boldsymbol{K}_0,\boldsymbol{L}(\boldsymbol{K}_0)} + \frac{\lambda_{\min}(\boldsymbol{H}_{\boldsymbol{K}_0,\boldsymbol{L}(\boldsymbol{K}_0)})}{2\|\boldsymbol{D}\|^2} \cdot \boldsymbol{I} \Big\}, \tag{3.1}$$

where $\boldsymbol{H}_{\boldsymbol{K},\boldsymbol{L}} := \boldsymbol{R}^w - \boldsymbol{D}^\top \boldsymbol{P}_{\boldsymbol{K},\boldsymbol{L}} \boldsymbol{D}$. Notice that $\hat{\mathcal{K}} \subset \mathcal{K}$ since

$$\boldsymbol{R}^w - \boldsymbol{D}^\top \boldsymbol{P}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})} \boldsymbol{D} \succeq \boldsymbol{R}^w - \boldsymbol{D}^\top (\boldsymbol{P}_{\boldsymbol{K}_0,\boldsymbol{L}(\boldsymbol{K}_0)} + \frac{\lambda_{\min}(\boldsymbol{H}_{\boldsymbol{K}_0,\boldsymbol{L}(\boldsymbol{K}_0)})}{2\|\boldsymbol{D}\|^2} \cdot \boldsymbol{I}) \boldsymbol{D} \\ \succeq \frac{\lambda_{\min}(\boldsymbol{H}_{\boldsymbol{K}_0,\boldsymbol{L}(\boldsymbol{K}_0)})}{2} \cdot \boldsymbol{I} \succ 0. \tag{3.2}$$

As can be observed from (3.1), we need to control the error induced by the inner loop solver which provides an approximation of $\boldsymbol{L}(\boldsymbol{K})$ in order to show the recurrence of the iterates $\boldsymbol{K}_t$ in the set $\hat{\mathcal{K}}$ with high probability. This inner maximization problem which takes the form of an LQR problem has been previously addressed in the literature in several works using for example a gradient ascent or a natural gradient ascent algorithm in both model-based and model-free settings Fazel et al. [2018], Malik et al. [2019], Zhang et al. [2021b]. We report in the next result an informal version of Theorem 4.1 in Zhang et al. [2021b] for the inner maximization problem in view of deriving the total sample complexity of our nested algorithm.

Lemma 3.1. *(Inner-loop sample complexity Zhang et al. [2021b]) Let $\delta_1, \varepsilon_1 \in (0,1)$ and let $\boldsymbol{K} \in \mathcal{K}$. Using $\tilde{\mathcal{O}}(\varepsilon_1^{-2} \log \delta_1^{-1})$ samples, Algorithm 1 outputs with probability at least $1-\delta_1$ a control gain matrix $\boldsymbol{L}$ satisfying: $\mathcal{G}(\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})) - \mathcal{G}(\boldsymbol{K},\boldsymbol{L}) \leq \varepsilon_1$, $\|\boldsymbol{L}(\boldsymbol{K}) - \boldsymbol{L}\|_F \leq \sqrt{\lambda_{\min}^{-1}(\boldsymbol{H}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})}) \cdot \varepsilon_1}$.*

Remark 3.2. *This $\tilde{\mathcal{O}}(\varepsilon_1^{-2})$ sample complexity reported in Lemma 3.1 can be further improved to $\widetilde{\mathcal{O}}(\varepsilon_1^{-1})$ using ZO two-point estimation Agarwal and Dekel [2010].*

It follows from Lemma 2 that any control gain matrix $\boldsymbol{L}$ produced by Algorithm 1 lies in the following bounded set:

$$\hat{\mathcal{L}} := \left\{ \boldsymbol{L} \in \mathcal{S}_2 \mid \|\boldsymbol{L}(\boldsymbol{K}) - \boldsymbol{L}\|_F \leq H, \boldsymbol{K} \in \hat{\mathcal{K}} \right\}, \quad H := \sup_{\boldsymbol{K} \in \hat{\mathcal{K}}} \lambda_{\min}^{-1}(\boldsymbol{H}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})}) \leq 2\lambda_{\min}^{-1}(\boldsymbol{H}_{\boldsymbol{K}_0,\boldsymbol{L}(\boldsymbol{K}_0)}). \tag{3.3}$$

3.1 Implicit Regularization

Using the sets $\hat{\mathcal{K}}$ and $\hat{\mathcal{L}}$ respectively defined in (3.1) and (3.3), we are now ready to state the IR of our model-free nested natural gradient algorithm w.r.t. both control gain matrices $\boldsymbol{K}$ and $\boldsymbol{L}$. More specifically, we will prove that the pair of iterates $(\boldsymbol{K}_t, \boldsymbol{L}_t)$ generated by Algorithms 1 and 2 will be maintained in the bounded set $\hat{\mathcal{K}} \times \hat{\mathcal{L}}$ with high probability for every t if we properly choose the batch sample size M_2, the smoothing radius r and the inner-loop accuracy ε_1. Before stating the IR result, we state some nice Lipschitzness properties over the set $\hat{\mathcal{K}} \times \hat{\mathcal{L}}$ that will contribute to our analysis.

Proposition 3.3. *Let $\boldsymbol{K}_0 \in \mathcal{K}$ and consider the corresponding set $\hat{\mathcal{K}}$. For any $(\boldsymbol{K}, \boldsymbol{L}) \in \hat{\mathcal{K}} \times \hat{\mathcal{L}}$, $\boldsymbol{K}' \in \mathcal{K}, \boldsymbol{L}'$, there exist positive constants D_1, D_2, l_1, l_2 such that if we let $\|\boldsymbol{K}' - \boldsymbol{K}\| \leq D_1$, $\|\boldsymbol{L}' - \boldsymbol{L}\| \leq D_2$ where $D_1, D_2 > 0$ are defined in Lemma A.6, then there exist positive constants l_1, l_2 such that $\|\mathbf{F}_{\boldsymbol{K}',\boldsymbol{L}} - \mathbf{F}_{\boldsymbol{K},\boldsymbol{L}}\| \leq l_1 \|\boldsymbol{K}' - \boldsymbol{K}\|$, and $\|\mathbf{F}_{\boldsymbol{K},\boldsymbol{L}'} - \mathbf{F}_{\boldsymbol{K},\boldsymbol{L}}\| \leq l_2 \|\boldsymbol{L}' - \boldsymbol{L}\|$. Similar results also hold when replacing $\mathbf{F}_{\boldsymbol{K},\boldsymbol{L}}$ by $\mathbf{E}_{\boldsymbol{K},\boldsymbol{L}}, \boldsymbol{\Sigma}_{\boldsymbol{K},\boldsymbol{L}}$, and $\mathbf{P}_{\boldsymbol{K},\boldsymbol{L}}$, see Lemma A.9, A.10 and A.11 for the proofs.*

The smoothness and continuity over the set $\hat{\mathcal{K}} \times \hat{\mathcal{L}}$ naturally motivate us to borrow the ideas from stochastic optimization. In particular, it is tempting to follow the analysis of stochastic nested algorithms for global Lipschitz smooth functions, see for instance Lin et al. [2020]. Unfortunately, such analysis is not directly applicable since the properties stated in Proposition 3.3, only hold locally within the set $\hat{\mathcal{K}} \times \hat{\mathcal{L}}$, therefore one needs to ensure that the iterates of Algorithm 2 remain in this set. This can be achieved by controlling the value matrix $\boldsymbol{P}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})}$ along the iterations. When the exact (natural) gradients are available, Zhang et al. [2021b] utilize this idea to show that the sequence $(\boldsymbol{P}_{\boldsymbol{K}_t,\boldsymbol{L}(\boldsymbol{K}_t)})$ is monotone along the trajectory in the positive semi-definite sense and refer to this property as *implicit regularization*. However, in the case when the estimated gradients (from ZO estimation) are used, the situation is more challenging. Such sequence is no longer monotone and the deviation from monotonicity must be controlled.

3.2 Sample Complexity Improvement

In this subchapter, we state one of our key technical results, which ensures that the iterates will remain in the set $\hat{\mathcal{K}} \times \hat{\mathcal{L}}$ with high probability. The key technical improvement over the similar result in Theorem 4.2 of Zhang et al. [2021b] is that we require a much smaller number of samples for achieving this. This improvement is crucial for achieving our better total sample complexity stated in Theorems 3.5 and 3.7.

Proposition 3.4. *(Implicit regularization using single-point estimation) Let Assumption 1.1 hold. Let $\boldsymbol{K}_0 \in \mathcal{K}$ and consider the corresponding $\hat{\mathcal{K}}$ set defined in (3.1). For any $\delta_1 \in (0,1), \varepsilon_1 > 0$ and for any $\boldsymbol{K} \in \mathcal{K}$, Algorithm 1 with single-point estimation outputs $\boldsymbol{L}$ such that $\mathcal{G}(\boldsymbol{K}, \boldsymbol{L}(\boldsymbol{K})) - \mathcal{G}(\boldsymbol{K}, \boldsymbol{L}) \leq \varepsilon_1$ with probability at least $1 - \delta_1$ using $T_{in} M_1 = \widetilde{\mathcal{O}}(\varepsilon_1^{-2})$ samples.*

Moreover for any $\delta_2 \in (0,1)$ and any integer $T \geq 1$, if the estimation parameters in Algorithm 2 satisfy $M_2 = \widetilde{\mathcal{O}}(T^2)$, $\tau_2 = \mathcal{O}(1)$, $r_2 = \mathcal{O}(T^{-1/2})$, $\varepsilon_1 = \mathcal{O}(T^{-1})$, $\delta_1 = \mathcal{O}(\delta_2/T)$, then, it holds with probability at least $1-\delta_2$ that $\boldsymbol{K}_t \in \hat{\mathcal{K}}$ for all $t = 1, \cdots, T$.

A detailed version of this proposition and its formal proof can be found in Appendix A.2. Here we provide a brief proof sketch, outlining the key steps of the proof.

Proof. The key step in the proof is a descent-like inequality for the value matrix sequence $(\boldsymbol{P}_{\boldsymbol{K}_t,\boldsymbol{L}(\boldsymbol{K}_t)})$ (in the positive semi-definite sense) which holds with high probability (see Lemma A.20 for more details):

$$\boldsymbol{P}_{\boldsymbol{K}_{t+1},\boldsymbol{L}(\boldsymbol{K}_{t+1})} - \boldsymbol{P}_{\boldsymbol{K}_t,\boldsymbol{L}(\boldsymbol{K}_t)} \preceq \tau_2(c_1 \cdot r_2^2 + c_2 \cdot \varepsilon_1 + c_3 \cdot \|V(\widetilde{\boldsymbol{F}}_{\boldsymbol{K}_t,\boldsymbol{L}_t})\|) \cdot I - \frac{\tau_2}{4}\boldsymbol{F}_{\boldsymbol{K}_t,\boldsymbol{L}(\boldsymbol{K}_t)}^\top \boldsymbol{F}_{\boldsymbol{K}_t,\boldsymbol{L}(\boldsymbol{K}_t)} \tag{3.4}$$

$$\preceq \tau_2(c_1 \cdot r_2^2 + c_2 \cdot \varepsilon_1 + c_3 \cdot \|V(\widetilde{\boldsymbol{F}}_{\boldsymbol{K}_t,\boldsymbol{L}_t})\|) \cdot I = \mathcal{O}\left(\frac{1}{T}\right) \cdot I, \tag{3.5}$$

where c_1, c_2, c_3 are positive constants and $V(\widetilde{\boldsymbol{F}}_{\boldsymbol{K}_t,\boldsymbol{L}_t}) := (\widetilde{\boldsymbol{F}}_{\boldsymbol{K}_t,\boldsymbol{L}_t} - \mathbb{E}[\widetilde{\boldsymbol{F}}_{\boldsymbol{K}_t,\boldsymbol{L}_t}])^\top(\widetilde{\boldsymbol{F}}_{\boldsymbol{K}_t,\boldsymbol{L}_t} - \mathbb{E}[\widetilde{\boldsymbol{F}}_{\boldsymbol{K}_t,\boldsymbol{L}_t}])$. From (3.5), we can observe that the deviation can be upperbounded by three sources of estimation errors: a $\mathcal{O}(r_2^2)$ bias term induced by the ZO estimate, the inner-loop error ε_1, and a variance-like term induced by the ZO estimation procedure. Hence, the deviation can be controlled by choosing $\varepsilon_1 = \mathcal{O}(1/T)$, $r_2 = \mathcal{O}(T^{-1/2})$ and a large enough M_2 such that $V(\widetilde{\boldsymbol{F}}_{\boldsymbol{K}_t,\boldsymbol{L}_t}) = \mathcal{O}(1/T)$. This control allows to show that $\boldsymbol{K}_{t+1}$ can be kept in $\hat{\mathcal{K}}$ for $t = 0, \cdots, T-1$. Inequality (3.4) follows from the Lipschitzness properties in Proposition 3.3 and borrows ideas from the analysis of stochastic double-loop algorithms for functions with similar curvature properties such as Lipschitz smoothness and continuity (see supplementary material of Lin et al. [2020], for example). □

Theorem 3.5. *Under the setting of Proposition 3.4, for every integer $T \geq 1$, it holds with probability at least $1-\delta_2$ that*

$$\frac{1}{T}\sum_{t=0}^{T-1} \|\boldsymbol{F}_{\boldsymbol{K}_t,\boldsymbol{L}(\boldsymbol{K}_t)}\|_F^2 = \mathcal{O}\left(\frac{1}{T}\right).$$

In other words, Algorithm 2 reaches with high probability an ε-stationary point (i.e., $\|\boldsymbol{F}_{\boldsymbol{K}_{out},\boldsymbol{L}(\boldsymbol{K}_{out})}\|_F^2 \leq \varepsilon$) and hence an ε-neighborhood of the NE[*] with a total sample complexity given by $T(T_{in}M_1 + M_2) = \widetilde{\mathcal{O}}(\varepsilon^{-3})$. A detailed version of this theorem can be found in Appendix A.3.

Proof. The convergence rate result follows from multiplying (3.4) by $\boldsymbol{\Sigma}_0$, taking the trace and summing up the resulting inequality to obtain with high probability:

$$\frac{1}{T}\sum_{t=0}^{T-1} \|\boldsymbol{F}_{\boldsymbol{K}_t,\boldsymbol{L}(\boldsymbol{K}_t)}\|_F^2 = \frac{1}{T}\sum_{t=0}^{T-1} \mathrm{Tr}\left(\boldsymbol{F}_{\boldsymbol{K}_t,\boldsymbol{L}(\boldsymbol{K}_t)}^\top \boldsymbol{F}_{\boldsymbol{K}_t,\boldsymbol{L}(\boldsymbol{K}_t)}\right) = \mathcal{O}\left(\frac{1}{T}\right),$$

We refer the reader to Appendix A.3 for the full proof. □

Remark 3.6. *Our $\widetilde{\mathcal{O}}(\varepsilon^{-3})$ total sample complexity result improves over the $\widetilde{\mathcal{O}}(\varepsilon^{-9})$ sample complexity shown in Zhang et al. [2021b]. The improvement of our algorithms comes from three elements: (a) we have a looser requirement for the inner-loop problem accuracy $\varepsilon_1 = \mathcal{O}(T^{-1})$ while in Zhang et al. [2021b] $\varepsilon_1 = \mathcal{O}(T^{-2})$; (b) we achieve a better sample complexity for the*

[*] Here the correspondence between stationary point and NE can be found in Lemma 3.2 of Zhang et al. [2021b].

outer-loop problem using a more careful decomposition of the estimation error caused by the estimated natural gradients: we only require $r_2 = \mathcal{O}(T^{-1/2})$ while Zhang et al. [2021b] chose $r_2 = \mathcal{O}(T^{-1})$ and (c) we reduce the number of inner-loop algorithm calls with a more natural version of the model-free nested algorithm (see the comparison at the end of Chapter 2). Hence the outer-loop sample complexity is improved from $\mathcal{O}(\varepsilon^{-5})$ to $TM_2 = \widetilde{\mathcal{O}}(\varepsilon^{-3})$. Combining all of these three elements, we improve the total sample complexity provided in Zhang et al. [2021b] which is given by: $T(T_{in}M_1M_2 + T_{in}M_1) = \mathcal{O}(\varepsilon^{-9})$[†].

In the following theorem, we utilize the two-point zeroth order estimation method which enjoys smaller variance and hence leads to improved sample complexity.

Theorem 3.7. *(Sample complexity using two-point estimation) Let Assumption 1.1 hold. Let $\boldsymbol{K}_0 \in \mathcal{K}$ and consider the corresponding set $\hat{\mathcal{K}}$ defined in (3.1). For any $\delta_1 \in (0,1), \varepsilon_1 > 0$ and for any $\boldsymbol{K} \in \mathcal{K}$, Algorithm 1 with two-point estimation outputs $\boldsymbol{L}$ such that $\mathcal{G}(\boldsymbol{K}, \boldsymbol{L}(\boldsymbol{K})) - \mathcal{G}(\boldsymbol{K}, \boldsymbol{L}) \leq \varepsilon_1$ with probability at least $1 - \delta_1$ using $T_{in}M_1 = \widetilde{\mathcal{O}}(\varepsilon_1^{-1})$ samples. Moreover for any $\delta_2 \in (0,1)$ and any integer $T \geq 1$, if the estimation parameters in Algorithm 2 satisfy $M_2 = \widetilde{\mathcal{O}}(T)$, $\tau_2 = \mathcal{O}(1)$, $r_2 = \mathcal{O}(T^{-1/2})$, $\varepsilon_1 = \mathcal{O}(T^{-1})$, $\delta_1 = \mathcal{O}(\delta_2/T)$. Then, it holds with probability at least $1 - \delta_2$ that $\boldsymbol{K}_t \in \hat{\mathcal{K}}$ for all $t = 1, \cdots, T$ and $\frac{1}{T}\sum_{t=0}^{T-1} \|\boldsymbol{F}_{\boldsymbol{K}_t, \boldsymbol{L}(\boldsymbol{K}_t)}\|_F^2 = \mathcal{O}\left(\frac{1}{T}\right)$. In other words, Algorithm 2 returns an ε-stationary point (i.e., $\|\boldsymbol{F}_{\boldsymbol{K}_{out}, \boldsymbol{L}(\boldsymbol{K}_{out})}\|_F^2 \leq \varepsilon$) after $\mathcal{O}(\varepsilon^{-1})$ iterations. The total sample complexity is given by $T\,(T_{in}\,M_1 + M_2) = \widetilde{\mathcal{O}}(\varepsilon^{-2})$. A detailed version of this theorem can be found in Appendix A.4.*

Remark 3.8. *(Two-point estimation) In order to obtain Theorem 3.7, we assume to have access to cost values at two different controllers $\boldsymbol{K}_t^1$ and $\boldsymbol{K}_t^2$ under the same realization of noise $\boldsymbol{\xi}_m$. This assumption can be limiting since it implies that $\boldsymbol{\xi}_m$ is generated in advance. Recently developed techniques of first-order estimation for single agent LQR (instead of ZO) Ju et al. [2023] might help to avoid this assumption in the future.*

3.3 Last-iterate Convergence

By now, we adopt the same convergence measure as Zhang et al. [2021b] and improve upon it. In the following results, we show new last-iterate convergence results using cost function values. Before we present the result, we prepare readers with the gradient domination proposition, which plays a crucial role in the proof.

Proposition 3.9. *(Gradient domination) Suppose $\boldsymbol{K} \in \hat{\mathcal{K}}$, then we have the following inequality*

$$\mathcal{G}(\boldsymbol{K}^*, \boldsymbol{L}^*) - \mathcal{G}(\boldsymbol{K}, \boldsymbol{L}(\boldsymbol{K})) \geq -s_2 Tr(\boldsymbol{F}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})}^\top \boldsymbol{F}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})}),$$

where $s_2 := \sigma_{\min}^{-1}(\boldsymbol{R}^u)s_4$, $s_4 := \sup_{\boldsymbol{K}\in\hat{\mathcal{K}}} \|\boldsymbol{\Sigma}_{\boldsymbol{K}^*, \widetilde{\boldsymbol{L}}_{\boldsymbol{K},\boldsymbol{K}^*}}\|$, and

$$\widetilde{\boldsymbol{L}}_{\boldsymbol{K},\boldsymbol{K}'} := \boldsymbol{L}(\boldsymbol{K}) - (-\boldsymbol{R}^w + \boldsymbol{D}^\top \boldsymbol{P}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})}\boldsymbol{D})^{-1}\boldsymbol{D}^\top \boldsymbol{P}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})}\boldsymbol{B}(\boldsymbol{K}' - \boldsymbol{K}).$$

Proof. We start from the more general matrix difference result, Lemma A.24. Then for

[†]Notice that the total sample complexity for inner and outer loops together was not explicitely stated in Zhang et al. [2021b], but can be inferred from their intermediate results.

$\boldsymbol{P}_{\boldsymbol{K}',\boldsymbol{L}(\boldsymbol{K}')}$ and $\boldsymbol{P}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})}$ where $\boldsymbol{K}'$ and $\boldsymbol{K}$ are arbitrary policies in $\mathcal{K}$, we have

$$\begin{aligned}\boldsymbol{P}_{\boldsymbol{K}',\boldsymbol{L}'} - \boldsymbol{P}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})} = \boldsymbol{A}_{\boldsymbol{K}',\boldsymbol{L}'}^\top(\boldsymbol{P}_{\boldsymbol{K}',\boldsymbol{L}'} - \boldsymbol{P}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})})\boldsymbol{A}_{\boldsymbol{K}',\boldsymbol{L}'} + (\boldsymbol{K}'-\boldsymbol{K})^\top \boldsymbol{F}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})} + \boldsymbol{F}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})}^\top(\boldsymbol{K}'-\boldsymbol{K})\\ + (\boldsymbol{K}'-\boldsymbol{K})^\top(\boldsymbol{R}^w + \boldsymbol{B}^\top \boldsymbol{P}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})}\boldsymbol{B})(\boldsymbol{K}'-\boldsymbol{K})\\ + (\boldsymbol{L}'-\boldsymbol{L}(\boldsymbol{K}))^\top \boldsymbol{D}^\top \boldsymbol{P}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})}\boldsymbol{B}(\boldsymbol{K}'-\boldsymbol{K})\\ + (\boldsymbol{K}'-\boldsymbol{K})^\top \boldsymbol{B}^\top \boldsymbol{P}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})}\boldsymbol{D}(\boldsymbol{L}'-\boldsymbol{L}(\boldsymbol{K}))\\ + (\boldsymbol{L}'-\boldsymbol{L})^\top(-\boldsymbol{R}^w + \boldsymbol{D}^\top \boldsymbol{P}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})}\boldsymbol{D})(\boldsymbol{L}'-\boldsymbol{L}).\end{aligned}$$

Again, multiply $\boldsymbol{\Sigma}_0$ at both sides at the same time and take the trace, we have

$$\begin{aligned}\mathcal{G}(\boldsymbol{K}',\boldsymbol{L}') - \mathcal{G}(\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})) &= \operatorname{Tr}((\boldsymbol{P}_{\boldsymbol{K}',\boldsymbol{L}'} - \boldsymbol{P}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})})\boldsymbol{\Sigma}_0)\\ &= \operatorname{Tr}((2(\boldsymbol{K}'-\boldsymbol{K})^\top \boldsymbol{F}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})} + (\boldsymbol{K}'-\boldsymbol{K})^\top(\boldsymbol{R}^u + \boldsymbol{B}^\top \boldsymbol{P}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})}\boldsymbol{B})(\boldsymbol{K}'-\boldsymbol{K})\\ &\quad + (\boldsymbol{L}'-\boldsymbol{L}(\boldsymbol{K}))^\top(-\boldsymbol{R}^w + \boldsymbol{D}^\top \boldsymbol{P}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})}\boldsymbol{D})(\boldsymbol{L}'-\boldsymbol{L}(\boldsymbol{K}))\\ &\quad + 2(\boldsymbol{L}'-\boldsymbol{L}(\boldsymbol{K}))^\top \boldsymbol{D}^\top \boldsymbol{P}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})}\boldsymbol{B}(\boldsymbol{K}'-\boldsymbol{K}))\boldsymbol{\Sigma}_{\boldsymbol{K}',\boldsymbol{L}'}).\end{aligned}$$

In the second inequality, we apply the dual Lyapunov equation lemma (Lemma A.23) with

$$\begin{aligned}&A = \boldsymbol{A}_{\boldsymbol{K}',\boldsymbol{L}'}, \quad X = \boldsymbol{P}_{\boldsymbol{K}',\boldsymbol{L}'} - \boldsymbol{P}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})}, \quad V = \boldsymbol{\Sigma}_0, \quad W = \boldsymbol{\Sigma}_{\boldsymbol{K}',\boldsymbol{L}'},\\ &Y = (\boldsymbol{K}'-\boldsymbol{K})^\top \boldsymbol{F}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})} + \boldsymbol{F}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})}^\top(\boldsymbol{K}'-\boldsymbol{K}) + (\boldsymbol{K}'-\boldsymbol{K})^\top(\boldsymbol{R}^w + \boldsymbol{B}^\top \boldsymbol{P}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})}\boldsymbol{B})(\boldsymbol{K}'-\boldsymbol{K})\\ &\quad + (\boldsymbol{L}'-\boldsymbol{L}(\boldsymbol{K}))^\top \boldsymbol{D}^\top \boldsymbol{P}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})}\boldsymbol{B}(\boldsymbol{K}'-\boldsymbol{K}) + (\boldsymbol{K}'-\boldsymbol{K})^\top \boldsymbol{B}^\top \boldsymbol{P}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})}\boldsymbol{D}(\boldsymbol{L}'-\boldsymbol{L}(\boldsymbol{K}))\\ &\quad + (\boldsymbol{L}'-\boldsymbol{L})^\top(-\boldsymbol{R}^w + \boldsymbol{D}^\top \boldsymbol{P}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})}\boldsymbol{D})(\boldsymbol{L}'-\boldsymbol{L}).\end{aligned}$$

If we choose $\boldsymbol{L}' = \widetilde{\boldsymbol{L}}_{\boldsymbol{K},\boldsymbol{K}'} := \boldsymbol{L}(\boldsymbol{K}) - (-\boldsymbol{R}^w + \boldsymbol{D}^\top \boldsymbol{P}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})}\boldsymbol{D})^{-1}\boldsymbol{D}^\top \boldsymbol{P}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})}\boldsymbol{B}(\boldsymbol{K}'-\boldsymbol{K})$, the maximum of the RHS is achieved since $-\boldsymbol{R}^w + \boldsymbol{D}^\top \boldsymbol{P}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})}\boldsymbol{D} \prec 0$. Then for the objective function values,

$$\begin{aligned}&\mathcal{G}(\boldsymbol{K}',\widetilde{\boldsymbol{L}}_{\boldsymbol{K},\boldsymbol{K}'}) - \mathcal{G}(\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K}))\\ &= \operatorname{Tr}((2(\boldsymbol{K}'-\boldsymbol{K})^\top \boldsymbol{F}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})} + (\boldsymbol{K}'-\boldsymbol{K})^\top(\boldsymbol{R}^u + \boldsymbol{B}^\top \boldsymbol{P}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})}\boldsymbol{B})(\boldsymbol{K}'-\boldsymbol{K})\\ &\quad - (\boldsymbol{K}'-\boldsymbol{K})^\top \boldsymbol{B}^\top \boldsymbol{P}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})}\boldsymbol{D}(-\boldsymbol{R}^w + \boldsymbol{D}^\top \boldsymbol{P}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})}\boldsymbol{D})^{-1}\boldsymbol{D}^\top \boldsymbol{P}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})}\boldsymbol{B}(\boldsymbol{K}'-\boldsymbol{K}))\boldsymbol{\Sigma}_{\boldsymbol{K}',\boldsymbol{L}'}).\end{aligned}$$

Moreover, let $\boldsymbol{K}' = \boldsymbol{K}^* \in \hat{\mathcal{K}}$, we have

$$\begin{aligned}&\mathcal{G}(\boldsymbol{K}^*,\boldsymbol{L}^*) - \mathcal{G}(\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})) \geq \mathcal{G}(\boldsymbol{K}^*,\widetilde{\boldsymbol{L}}_{\boldsymbol{K},\boldsymbol{K}^*}) - \mathcal{G}(\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})) \qquad (3.6)\\ &= \operatorname{Tr}((2(\boldsymbol{K}^*-\boldsymbol{K})^\top \boldsymbol{F}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})} + (\boldsymbol{K}^*-\boldsymbol{K})^\top(\boldsymbol{R}^u + \boldsymbol{B}^\top \boldsymbol{P}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})}\boldsymbol{B}\\ &\quad + \boldsymbol{B}^\top \boldsymbol{P}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})}\boldsymbol{D}(\boldsymbol{R}^w - \boldsymbol{D}^\top \boldsymbol{P}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})}\boldsymbol{D})^{-1}\boldsymbol{D}^\top \boldsymbol{P}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})}\boldsymbol{B})(\boldsymbol{K}^*-\boldsymbol{K}))\boldsymbol{\Sigma}_{\boldsymbol{K}^*,\widetilde{\boldsymbol{L}}_{\boldsymbol{K},\boldsymbol{K}^*}})\\ &\overset{(a)}{\geq} -\operatorname{Tr}(\boldsymbol{F}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})}^\top(\boldsymbol{R}^u + \boldsymbol{B}^\top \boldsymbol{P}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})}\boldsymbol{B}\\ &\quad + \boldsymbol{B}^\top \boldsymbol{P}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})}\boldsymbol{D}(\boldsymbol{R}^w - \boldsymbol{D}^\top \boldsymbol{P}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})}\boldsymbol{D})^{-1}\boldsymbol{D}^\top \boldsymbol{P}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})}\boldsymbol{B})^{-1}\boldsymbol{F}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})}\boldsymbol{\Sigma}_{\boldsymbol{K}^*,\widetilde{\boldsymbol{L}}_{\boldsymbol{K},\boldsymbol{K}^*}})\\ &\geq -\operatorname{Tr}(\boldsymbol{F}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})}^\top(\boldsymbol{R}^u)^{-1}\boldsymbol{F}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})}\boldsymbol{\Sigma}_{\boldsymbol{K}^*,\widetilde{\boldsymbol{L}}_{\boldsymbol{K},\boldsymbol{K}^*}})\\ &\overset{(b)}{\geq} -\sigma_{\min}^{-1}(\boldsymbol{R}^u)\|\boldsymbol{\Sigma}_{\boldsymbol{K}^*,\widetilde{\boldsymbol{L}}_{\boldsymbol{K},\boldsymbol{K}^*}}\|\operatorname{Tr}(\boldsymbol{F}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})}^\top \boldsymbol{F}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})}),\end{aligned}$$

where (a) holds since

$$(\boldsymbol{K}^* - \boldsymbol{K} - M^{-1}\boldsymbol{F}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})})^\top M(\boldsymbol{K}^* - \boldsymbol{K} - M^{-1}\boldsymbol{F}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})}) \succeq 0,$$

where $M := \mathbf{R}^u + \mathbf{B}^\top \mathbf{P}_{\mathbf{K},\mathbf{L}(\mathbf{K})}\mathbf{B} + \mathbf{B}^\top \mathbf{P}_{\mathbf{K},\mathbf{L}(\mathbf{K})}\mathbf{D}(\mathbf{R}^w - \mathbf{D}^\top \mathbf{P}_{\mathbf{K},\mathbf{L}(\mathbf{K})}\mathbf{D})^{-1}\mathbf{D}^\top \mathbf{P}_{\mathbf{K},\mathbf{L}(\mathbf{K})}\mathbf{B}$. And in (b), we apply $\mathbf{R}^w - \mathbf{D}^\top \mathbf{P}_{\mathbf{K},\mathbf{L}(\mathbf{K})}\mathbf{D} \succ 0$ since $\mathbf{K} \in \mathcal{K}$. Here $\|\mathbf{\Sigma}_{\mathbf{K}^*,\widetilde{\mathbf{L}}_{\mathbf{K},\mathbf{K}^*}}\|$ is bounded since

$$\begin{aligned}
\|\widetilde{\mathbf{L}}_{\mathbf{K},\mathbf{K}^*}\| &\leq \|\mathbf{L}^*\| + \|\mathbf{L}(\mathbf{K})\| + \sigma_{\min}^{-1}(\mathbf{R}^w - \mathbf{D}^\top \mathbf{P}_{\mathbf{K},\mathbf{L}(\mathbf{K})}\mathbf{D})\|\mathbf{D}\|\|\mathbf{P}_{\mathbf{K},\mathbf{L}(\mathbf{K})}\|\|\mathbf{B}\|\|\mathbf{K}^* - \mathbf{K}\| \\
&\leq \|\mathbf{L}^*\| + \sigma_{\min}^{-1}(\mathbf{R}^w - \mathbf{D}^\top \mathbf{P}_{\mathbf{K}_0,\mathbf{L}(\mathbf{K}_0)}\mathbf{D})\|\mathbf{D}\|\|\mathbf{P}_{\mathbf{K}_0,\mathbf{L}(\mathbf{K}_0)}\|\|\mathbf{A} - \mathbf{B}\mathbf{K}\| \\
&\quad + \sigma_{\min}^{-1}(\mathbf{R}^w - \mathbf{D}^\top \mathbf{P}_{\mathbf{K}_0,\mathbf{L}(\mathbf{K}_0)}\mathbf{D})\|\mathbf{D}\|\|\mathbf{P}_{\mathbf{K}_0,\mathbf{L}(\mathbf{K}_0)}\|\|\mathbf{B}\|\|\mathbf{K}^* - \mathbf{K}\|.
\end{aligned}$$

Hence when $\mathbf{K} \in \hat{\mathcal{K}}$, hence there exists a positive constant s_4 such that $\|\mathbf{\Sigma}_{\mathbf{K}^*,\widetilde{\mathbf{L}}_{\mathbf{K},\mathbf{K}^*}}\| \leq s_4$ where $s_4 := \sup_{\mathbf{K}\in\hat{\mathcal{K}}} \|\mathbf{\Sigma}_{\mathbf{K}^*,\widetilde{\mathbf{L}}_{\mathbf{K},\mathbf{K}^*}}\|$ holds for any $\mathbf{K} \in \hat{\mathcal{K}}$. Hence let $s_2 := \sigma_{\min}^{-1}(\mathbf{R}^u)s_4$, we have

$$\mathcal{G}(\mathbf{K}^*,\mathbf{L}^*) - \mathcal{G}(\mathbf{K},\mathbf{L}(\mathbf{K})) \geq -s_2 Tr(\mathbf{F}_{\mathbf{K},\mathbf{L}(\mathbf{K})}^\top \mathbf{F}_{\mathbf{K},\mathbf{L}(\mathbf{K})}). \tag{3.7}$$

□

Theorem 3.10. *(Last-iterate linear convergence in deterministic setting) Suppose $\mathbf{K}_0 \in \mathcal{K}$ and consider the nested natural gradient algorithm in the deterministic case: $\mathbf{K}_{t+1} = \mathbf{K}_t - \tau_2 \mathbf{F}_{\mathbf{K}_t,\mathbf{L}(\mathbf{K}_t)}$, let the stepsize τ_2 be a small enough constant. Then the iterates converge linearly in the sense that*

$$\mathcal{G}(\mathbf{K}_{t+1},\mathbf{L}(\mathbf{K}_{t+1})) - \mathcal{G}(\mathbf{K}^*,\mathbf{L}^*) \leq q(\mathcal{G}(\mathbf{K}_t,\mathbf{L}(\mathbf{K}_t)) - \mathcal{G}(\mathbf{K}^*,\mathbf{L}^*))$$

where the constant $q \in [0,1)$[‡] is the contractive coefficient. A detailed version of this theorem can be found in Appendix A.5.

Proof. Here we provide a proof sketch for this theorem. Consider one-step update of the algorithm: $\mathbf{K}' = \mathbf{K} - \tau_2 \mathbf{F}_{\mathbf{K},\mathbf{L}(\mathbf{K})}$. The key steps are to show the sufficient decrease and the gradient domination properties of the cost function, which are standard in proving linear convergence in optimization literature. More exactly, by choosing a small enough constant stepsize τ_2 we show that

$$\begin{aligned}
&\textbf{Sufficient decrease: } \mathcal{G}(\mathbf{K}',\mathbf{L}(\mathbf{K}')) - \mathcal{G}(\mathbf{K},\mathbf{L}(\mathbf{K})) \leq -s_1\tau_2 \operatorname{Tr}(\mathbf{F}_{\mathbf{K},\mathbf{L}(\mathbf{K})}^\top \mathbf{F}_{\mathbf{K},\mathbf{L}(\mathbf{K})}) \\
&\textbf{Gradient domination: } \mathcal{G}(\mathbf{K}^*,\mathbf{L}^*) - \mathcal{G}(\mathbf{K},\mathbf{L}(\mathbf{K})) \geq -s_2 \operatorname{Tr}(\mathbf{F}_{\mathbf{K},\mathbf{L}(\mathbf{K})}^\top \mathbf{F}_{\mathbf{K},\mathbf{L}(\mathbf{K})}),
\end{aligned}$$

where s_1, s_2 are positive constants. Combine these two inequalities, we obtain

$$\mathcal{G}(\mathbf{K}',\mathbf{L}(\mathbf{K}')) - \mathcal{G}(\mathbf{K}^*,\mathbf{L}^*) \leq (1 - \frac{s_1\tau_2}{s_2})(\mathcal{G}(\mathbf{K},\mathbf{L}(\mathbf{K})) - \mathcal{G}(\mathbf{K}^*,\mathbf{L}^*)),$$

where we require $\tau_2 \leq \frac{s_2}{s_1}$ in addition to the upperbound of τ_2 that ensures the sufficient decrease and gradient domination inequalities hold. We refer the reader to Appendix A.5 for the full proof. □

Theorem 3.11. *(Last-iterate convergence in stochastic setting) Let Assumption 1.1 hold. Let $\mathbf{K}_0 \in \mathcal{K}$ and consider the corresponding set $\hat{\mathcal{K}}$ defined in (3.1). For any $\delta_1 \in (0,1)$, $\varepsilon_1 > 0$ and for any $\mathbf{K} \in \mathcal{K}$, Algorithm 1 with single-point estimation outputs $\mathbf{L}$ such that $\mathcal{G}(\mathbf{K},\mathbf{L}(\mathbf{K})) - \mathcal{G}(\mathbf{K},\mathbf{L}) \leq \varepsilon_1$ with probability at least $1-\delta_1$ using $T_{in}M_1 = \widetilde{\mathcal{O}}(\varepsilon_1^{-2})$ samples. Moreover, for any $\delta_2 \in (0,1)$ and any accuracy requirement $\varepsilon \geq 0$, if the estimation parameters in Algorithm 2 satisfy $T = \mathcal{O}(\log(\varepsilon^{-1}))$, $M_2 = \widetilde{\mathcal{O}}(\varepsilon^{-2})$, $\tau_2 = \mathcal{O}(1)$, $r_2 = \mathcal{O}(\varepsilon^{-1/2})$, $\varepsilon_1 = \mathcal{O}(\varepsilon)$, $\delta_1 = \mathcal{O}(\delta_2/T)$. Then it holds with probability at least $1-\delta_2$ that $\mathbf{K}_t \in \hat{\mathcal{K}}$ for all $t = 1,\cdots,T$ and $\mathcal{G}(\mathbf{K}_T,\mathbf{L}(\mathbf{K}_T)) - \mathcal{G}(\mathbf{K}^*,\mathbf{L}^*) \leq \varepsilon$. The total sample complexity is given by $\mathcal{O}(T(T_{in}M_1 + T_{out}M_2)) = \widetilde{\mathcal{O}}(\varepsilon^{-2})$. A detailed version of this theorem is deferred to Appendix A.6.*

[‡]Note here choosing τ_2 such that q is arbitrarily close to 0 might not be possible since some upperbounds of τ_2 are required for the above contractive inequality to hold.

Proof. The proof of Theorem 3.11 is a generalization of the proof of Theorem 3.10. We apply Proposition 3.4 to guarantee the implicit regularization property of iterates. The key steps are also divided into two parts, sufficient decrease inequality (Inequality (3.4)) and gradient domination inequality (Proposition 3.9) are guaranteed with proper choice of constant stepsize τ_2:

$$\textbf{Sufficient decrease: } \mathcal{G}(\boldsymbol{K}', \boldsymbol{L}(\boldsymbol{K}')) - \mathcal{G}(\boldsymbol{K}, \boldsymbol{L}(\boldsymbol{K})) \leq -c_1\tau_2 \operatorname{Tr}(\boldsymbol{F}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})}^\top \boldsymbol{F}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})}) + \mathit{error}$$
$$\textbf{Gradient domination: } \mathcal{G}(\boldsymbol{K}^*, \boldsymbol{L}^*) - \mathcal{G}(\boldsymbol{K}, \boldsymbol{L}(\boldsymbol{K})) \geq -c_2 \operatorname{Tr}(\boldsymbol{F}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})}^\top \boldsymbol{F}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})}),$$

where c_1, c_2 are positive constants. Here an extra error term is introduced because of the estimated natural gradients. Combine these two inequalities and we obtain

$$\mathcal{G}(\boldsymbol{K}', \boldsymbol{L}(\boldsymbol{K}')) - \mathcal{G}(\boldsymbol{K}^*, \boldsymbol{L}^*) \leq q(\mathcal{G}(\boldsymbol{K}, \boldsymbol{L}(\boldsymbol{K})) - \mathcal{G}(\boldsymbol{K}^*, \boldsymbol{L}^*)) + \mathit{error}.$$

where $q \in [0, 1)$ is the contractive coefficient. Then this recursive inequality leads to the final convergence result. The complete proof is deferred to Appendix A.6. ☐

Remark 3.12. *Note here we don't use the two-point estimation assumption in Remark 3.8 to obtain the same sample complexity as Theorem 3.7 with a different convergence measure. More importantly, to the best of our knowledge, this is the first global last-iterate convergence result for policy optimization algorithms in search of the NE of zero-sum LQ games.*

Chapter 4

Simulations

In this chapter, we present simulation results[*] to further validate our contribution. We mainly present simulation results to show (i) convergence of Algorithm 2 in Zhang et al. [2021b] (benchmark algorithm, see Appendix A.10 for the complete algorithm) and Algorithm 2 when solving the same zero-sum LQ game using the same set of algorithm parameters; (ii) Algorithm 2 is more sample-efficient compared to the benchmark algorithm; (iii) the nested natural gradient algorithm has global linear last-iterate convergence in the deterministic case and Algorithm 2 demonstrates global last-iterate convergence.

Simulation setup. All the experiments are executed with Python 3.8.5 on a high-performance computing cluster where the reserved memory for executing experiments is 2000 MB. For the sake of comparison, we adopt the same set of model parameters as Zhang et al. [2021b]. Here we repeat the setting for completeness. The horizon length H is set to 5 and $A_t = A$, $B_t = B$, $D_t = D$, $Q_t = Q$, $R_t^u = R^u$, and $R_t^w = R_w$, where

$$A = \begin{bmatrix} 1 & 0 & -5 \\ -1 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad B = \begin{bmatrix} 1 & -10 & 0 \\ 0 & 3 & 1 \\ -1 & 0 & 2 \end{bmatrix}, \quad D = \begin{bmatrix} 0.5 & 0 & 0 \\ 0 & 0.2 & 0 \\ 0 & 0 & 0.2 \end{bmatrix},$$

$$Q = \begin{bmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{bmatrix}, \quad R^u = \begin{bmatrix} 4 & -1 & 0 \\ -1 & 4 & -2 \\ 0 & -2 & 3 \end{bmatrix}, \quad R^w = 5 \cdot I.$$

Applying the Nash equilibrium solution $(\boldsymbol{K}^*, \boldsymbol{L}^*)$ of the above game, $\mathcal{G}(\boldsymbol{K}^*, \boldsymbol{L}^*) \approx 3.2330$ and $\lambda_{\min}(\boldsymbol{H}_{\boldsymbol{K}^*,\boldsymbol{L}^*}) \approx 4.2860$. For the purpose of comparison, we choose the same set of parameters for both Algorithm 2 in Zhang et al. [2021b] and Algorithm 2 in this thesis. We choose $\boldsymbol{\Sigma}_0 = 0.05 \cdot I$ and the other parameters as follows

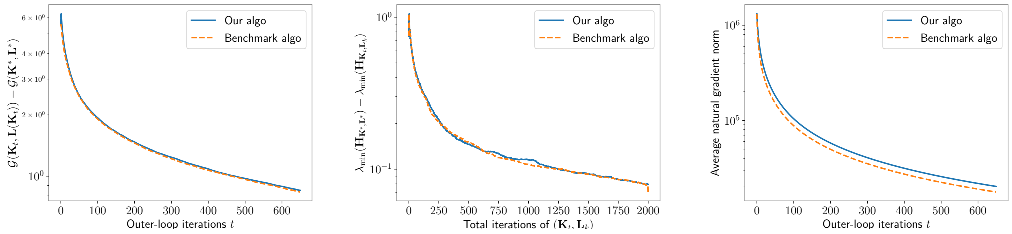
$$\boldsymbol{K}_0 = \begin{bmatrix} \mathrm{diag}(K, K, K, K, K) & \mathbf{0}_{15\times 3} \end{bmatrix}, \quad K := \begin{bmatrix} -0.08 & 0.35 & 0.62 \\ -0.21 & 0.19 & 0.32 \\ -0.06 & 0.10 & 0.41 \end{bmatrix}, \quad \boldsymbol{L}_0 = \mathbf{0}_{15\times 18},$$

$$r_2 = 0.08, \quad M_2 = 5 \times 10^5, \quad \epsilon_1 = 10^{-4}, \quad \tau_1 = 0.1, \quad \tau_2 = 4.67 \times 10^{-4}.$$

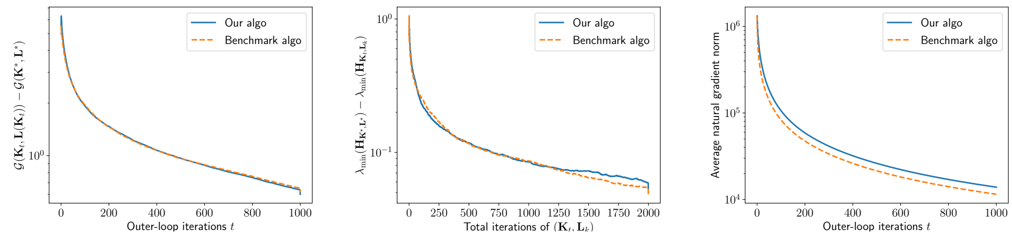
The experiments are executed using the above set of parameters and the single-point estimation without additional explanation.

Sample complexity improvement. As in Zhang et al. [2021b], we adopt the following two realizations: the inner-loop problem is solved using (i) the exact solutions and

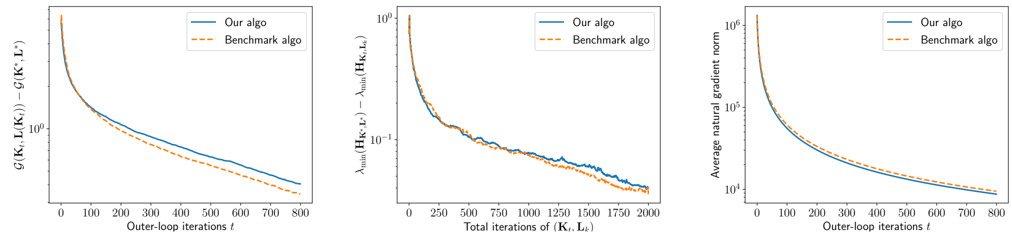
[*]The codes can be found at https://drive.google.com/drive/folders/1SVPAwxLAiC7K6EPhSKQXQZ_iyXXaZfoc?usp=sharing



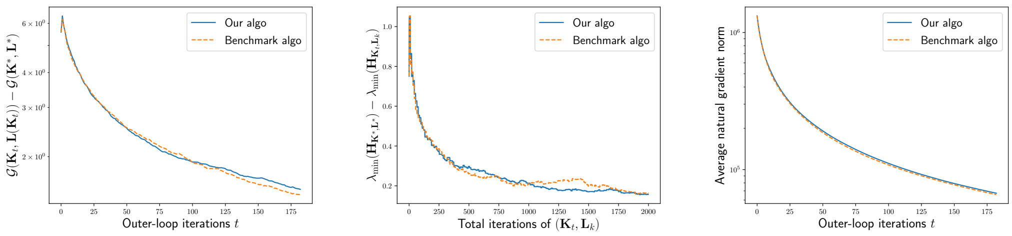
(a) Comparison of convergence between Algorithm 2 and the benchmark algorithm when using exact inner-loop natural gradient and estimated outer-loop natural gradients.



(b) Comparison of convergence between Algorithm 2 and the benchmark algorithm when using exact inner-loop solutions and estimated outer-loop natural gradients.



(c) Comparison of convergence between Algorithm 2 and the benchmark algorithm when using exact inner-loop natural gradient and estimated outer-loop natural gradients with $M_2 = 2.5 \times 10^5$ and $\tau_2 = 1 \times 10^{-3}$.

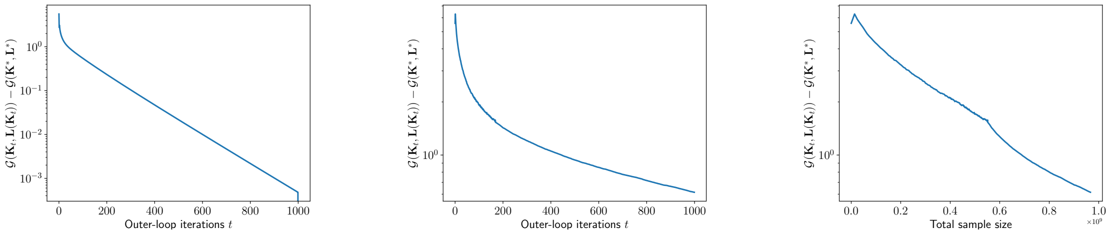


(d) Comparison of convergence between Algorithm 2 and the benchmark algorithm when using a fixed number of inner-loop iterations, exact inner-loop natural gradient, and estimated outer-loop natural gradients with $M_2 = 5 \times 10^5$, $\tau_1 = 0.1$, $\tau_2 = 2 \times 10^{-3}$, $r_2 = 0.02$, and $T_{in} = 10$.

Figure 4.1: Comparisons of Algorithm 2 and the benchmark algorithm under various settings. In the left, middle, and right figure, we show the convergence in terms of $\mathcal{G}(\boldsymbol{K}, \boldsymbol{L})$, $\lambda_{\min}(\boldsymbol{H}_{\boldsymbol{K},\boldsymbol{L}})$, and $T^{-1}\sum_{t=0}^{T-1} \|\boldsymbol{F}_{\boldsymbol{K}_t,\boldsymbol{L}(\boldsymbol{K}_t)}\|_F^2$ respectively.

(ii) exact natural gradients for efficiency[†]. From Figure 4.1a and 4.1b, we can see our algorithm shows a comparable convergence rate compared to the benchmark algorithm using the same set of parameters. In Figure 4.1c, we show that with smaller sample sizes $M_2 = 2.5 \times 10^5$ and larger stepsize $\tau_2 = 1 \times 10^{-3}$, Algorithm 2 also demonstrates convergence to $\mathcal{G}(\boldsymbol{K}^*, \boldsymbol{L}^*)$ and $\lambda_{\min}(\boldsymbol{H}_{\boldsymbol{K}^*,\boldsymbol{L}^*})$ with a comparable convergence rate compared with the benchmark algorithm. These results indicate Algorithm 2 is more sample-efficient than the benchmark algorithm. As in the benchmark algorithm (see Algorithm 3), an inner-loop problem still needs to be solved using samples at each sample step $m = 0, \cdots, M_2 - 1$ when the exact inner-loop solutions are not accessible.

[†]In the codes, we assume exact access to the solution of the inner-loop problem given each perturbed $\boldsymbol{K}_t^m$, i.e., $\boldsymbol{L}(\boldsymbol{K}_t^m)$ $m = 0, \cdots, M_2 - 1$. This practice is for the efficiency of the simulations and has an instrumental effect on the performance of the benchmark algorithm.



(a) Last-iterate convergence when using exact solutions of the inner-loop problem and the exact outer-loop natural gradients.

(b) Last-iterate convergence when using estimated inner-loop & outer-loop natural gradients.

(c) Sample size-last-iterate convergence when using estimated inner-loop & outer-loop natural gradients.

Figure 4.2: Last-iterate convergence results of Algorithm 2. We choose $r_1 = 0.5$, $M_1 = 10^6$, $\varepsilon_1 = 0.1$, and $\tau_1 = 0.04$.

Last-iterate convergence. To validate the convergence results in Theorem 3.10 and Theorem A.5, we conduct experiments using two sets of settings (i) exact solutions to the inner-loop problem and the exact outer-loop natural gradients and (ii) estimated inner-loop and estimated outer-loop natural gradients. Figure 4.2a displays the linear convergence rate of the cost difference and the linear curves support our last-iterate linear convergence result in the deterministic case. Figure 4.2b shows the last-iterate convergence of Algorithm 2 in the stochastic case where we run simulations using estimated inner-loop & outer-loop natural gradients. Moreover, in Figure 4.2c, we see the convergence of Algorithm 2 while the benchmark algorithm cannot complete one outer-loop iteration using the same number of sample sizes.

Fixed inner-loop iteration number. Besides using ε_1 to determine when to terminate the inner-loop iterations, we use a constant number of inner-loop iterations, T_{in}. This setting is closer to the practical scenario. We choose $T_{in} = 10$, exact inner-loop natural gradients, and estimated outer-loop natural gradients. In Figure 4.1d, we again observe similar convergence rates of Algorithm 2 and the benchmark algorithm. This observation supports the sample efficiency improvement of Algorithm 2 with a more realistic implementation.

Chapter 5

Conclusion

In this work, we showed a $\tilde{\mathcal{O}}(\varepsilon^{-3})$ sample complexity for a derivative-free nested natural policy gradient algorithm for solving the stochastic zero-sum linear quadratic dynamic game problem, improving over prior work. We further improved this sample complexity to $\tilde{\mathcal{O}}(\varepsilon^{-2})$ using zeroth order two-point estimation. Moreover, we provide the global last-iterate convergence result of nested algorithms for zero-sum LQ games in both deterministic and stochastic cases, which were not provided for zero-sum LQ games. Possible future research directions include (a) extending our analysis to continuous-time and infinite-horizon settings beyond our finite-horizon setting using techniques such as sensitivity analysis for stable continuous-time Lyapunov equations Hewer and Kenney [1988], (b) improving the dependence on problem dimensions and considering more general noise distributions since the boundedness of noises is not required by the stability constraint under the finite-horizon setting, and (c) establishing lower bounds for solving this problem. Designing theoretically grounded single-loop algorithms for zero-sum LQ games and considering more involved dynamics such as certain nonlinear dynamics Han et al. [2022, 2023] offer avenues of future research that merit further investigation.

Bibliography

Alekh Agarwal and Ofer Dekel. Optimal algorithms for online convex optimization with multi-point bandit feedback. In *Colt*, pages 28–40. Citeseer, 2010.

Asma Al-Tamimi, Frank L. Lewis, and Murad Abu-Khalaf. Model-free Q-learning designs for discrete-time zero-sum games with application to H-infinity control. In *2007 European Control Conference (ECC)*, page 1668–1675, Jul 2007.

Venkataramanan Balakrishnan and Lieven Vandenberghe. Semidefinite programming duality and linear time-invariant systems. *IEEE Transactions on Automatic Control*, 48(1): 30–41, 2003.

Tamer Başar and Pierre Bernhard. $\mathcal{H}_\infty$-optimal control and related minimax design problems. *Springer Book Archive-Mathematics*, 1995.

Tamer Başar and Geert Jan Olsder. *Dynamic Noncooperative Game Theory, 2nd Edition*. Society for Industrial and Applied Mathematics, 1998. doi: 10.1137/1.9781611971132. URL https://epubs.siam.org/doi/abs/10.1137/1.9781611971132.

Shankar P Bhattacharyya and Lee H Keel. Robust control: the parametric approach. In *Advances in control education 1994*, pages 49–52. Elsevier, 1995.

Jingjing Bu and Mehran Mesbahi. Global convergence of policy gradient algorithms for indefinite least squares stationary optimal control. *IEEE Control Systems Letters*, 4(3): 638–643, 2020.

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

Marco C Campi and Matthew R James. Nonlinear discrete-time risk-sensitive optimal control. *International Journal of Robust and Nonlinear Control*, 6(1):1–19, 1996.

René Carmona, Kenza Hamidouche, Mathieu Laurière, and Zongjun Tan. Linear-quadratic zero-sum mean-field type games: Optimality conditions and policy optimization. *arXiv preprint arXiv:2009.00578*, 2020.

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

Zaiwei Chen, Kaiqing Zhang, Eric Mazumdar, Asuman Ozdaglar, and Adam Wierman. A finite-sample analysis of payoff-based independent learning in zero-sum stochastic games. *arXiv preprint arXiv:2303.03100*, 2023.

Leilei Cui and Zhong-Ping Jiang. A reinforcement learning look at risk-sensitive linear quadratic gaussian control. *arXiv preprint arXiv:2212.02072*, 2022.

Leilei Cui and Lekan Molu. Mixed $\mathcal{H}_2/\mathcal{H}_\infty$ LQ games for robust policy optimization under unknown dynamics. *arXiv preprint arXiv:2209.04477*, 2022.

Ilyas Fatkhullin and Boris Polyak. Optimizing static linear feedback: Gradient method. *SIAM Journal on Control and Optimization*, 59(5):3887–3911, 2021.

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.

Han Feng and Javad Lavaei. On the exponential number of connected components for the feasible set of optimal decentralized control problems. In *2019 American Control Conference (ACC)*, pages 1430–1437. IEEE, 2019.

Claude-Nicolas Fiechter. PAC adaptive control of linear systems. In *Proc. 14th International Conference on Machine Learning*, pages 116–124. Morgan Kaufmann, 1997.

Abraham D Flaxman, Adam Tauman Kalai, and H Brendan McMahan. Online convex optimization in the bandit setting: gradient descent without a gradient. *arXiv preprint cs/0408007*, 2004.

Luca Furieri and Maryam Kamgarpour. First order methods for globally optimal distributed controllers beyond quadratic invariance. In *American Control Conference (ACC)*, page 4588–4593, Jul 2020.

Luca Furieri, Yang Zheng, and Maryam Kamgarpour. Learning the globally optimal distributed LQ regulator. In *Learning for Dynamics and Control*, pages 287–297. PMLR, 2020.

P.M. Gahinet, A.J. Laub, C.S. Kenney, and G.A. Hewer. Sensitivity of the stable discrete-time Lyapunov equation. *IEEE Transactions on Automatic Control*, 35(11):1209–1217, 1990. doi: 10.1109/9.59806.

Michael Giegrich, Christoph Reisinger, and Yufei Zhang. Convergence of policy gradient methods for finite-horizon stochastic linear-quadratic control problems. *arXiv preprint arXiv:2211.00617*, 2022.

Keith Glover and John C Doyle. State-space formulae for all stabilizing controllers that satisfy an $\mathcal{H}_\infty$-norm bound and relations to relations to risk sensitivity. *Systems & control letters*, 11(3):167–172, 1988.

Xingang Guo and Bin Hu. Global convergence of direct policy search for state-feedback $\mathcal{H}_\infty$ robust control: A revisit of nonsmooth synthesis with goldstein subdifferential. In *Thirty-Sixth Conference on Neural Information Processing Systems*, 2022.

Ben Hambly, Renyuan Xu, and Huining Yang. Policy gradient methods for the noisy linear quadratic regulator over a finite horizon. *arXiv preprint arXiv:2011.10300*, Jun 2021.

Ben Hambly, Renyuan Xu, and Huining Yang. Policy gradient methods find the nash equilibrium in n-player general-sum linear-quadratic games. *arXiv preprint arXiv:2107.13090*, Aug 2022.

Yinbin Han, Meisam Razaviyayn, and Renyuan Xu. Policy gradient finds global optimum of nearly linear-quadratic control systems. In *OPT 2022: Optimization for Machine Learning (NeurIPS 2022 Workshop)*, 2022.

Yinbin Han, Meisam Razaviyayn, and Renyuan Xu. Policy gradient converges to the globally optimal policy for nearly linear-quadratic regulators. *arXiv preprint arXiv:2303.08431*, 2023.

G. Hewer. An iterative technique for the computation of the steady state gains for the discrete optimal regulator. *IEEE Transactions on Automatic Control*, 16(4):382–384, 1971. doi: 10.1109/TAC.1971.1099755.

Gary Hewer and Charles Kenney. The sensitivity of the stable Lyapunov equation. *SIAM journal on control and optimization*, 26(2):321–344, 1988.

Roger A Horn and Charles R Johnson. *Matrix analysis*. Cambridge university press, 2012.

Bin Hu, Kaiqing Zhang, Na Li, Mehran Mesbahi, Maryam Fazel, and Tamer Başar. Towards a theoretical foundation of policy optimization for learning control policies. *Annual Review of Control, Robotics, and Autonomous Systems*, 2022.

Chi Jin, Praneeth Netrapalli, Rong Ge, Sham M Kakade, and Michael I Jordan. A short note on concentration inequalities for random vectors with subgaussian norm. *arXiv preprint arXiv:1902.03736*, 2019.

Caleb Ju, Georgios Kotsalis, and Guanghui Lan. A model-free first-order method for linear quadratic regulator with $\tilde{O}(1/\varepsilon)$ sampling complexity. *arXiv preprint arXiv:2212.00084*, Feb 2023.

Peter Lancaster and Leiba Rodman. *Algebraic Riccati equations*. Clarendon press, 1995.

Yingying Li, Yujie Tang, Runyu Zhang, and Na Li. Distributed reinforcement learning for decentralized linear quadratic control: A derivative-free policy optimization approach. *arXiv preprint arXiv:1912.09135*, Oct 2020.

Tianyi Lin, Chi Jin, and Michael Jordan. On gradient descent ascent for nonconvex-concave minimax problems. In *International Conference on Machine Learning*, pages 6083–6093. PMLR, 2020.

Lennart Ljung. *System Identification*. Birkhäuser Boston, 1998.

E.F. Mageirou and Y.C. Ho. Decentralized stabilization via game theoretic methods. *Automatica*, 13(4):393–399, 1977.

Perttim Makila and Hannut Toivonen. Computational methods for parametric LQ problems–a survey. *IEEE Transactions on Automatic Control*, 32(8):658–671, 1987.

Dhruv Malik, Ashwin Pananjady, Kush Bhatia, Koulik Khamaru, Peter Bartlett, and Martin Wainwright. Derivative-free methods for policy optimization: Guarantees for linear quadratic systems. In *The 22nd international conference on artificial intelligence and statistics*, pages 2916–2925. PMLR, 2019.

Eric Mazumdar, Lillian J Ratliff, Michael I Jordan, and S Shankar Sastry. Policy-gradient algorithms have no guarantees of convergence in linear quadratic games. *arXiv preprint arXiv:1907.03712*, 2019.

Hesameddin Mohammadi, Mahdi Soltanolkotabi, and Mihailo R Jovanović. On the linear convergence of random search for discrete-time LQR. *IEEE Control Systems Letters*, 5(3): 989–994, 2020.

Hesameddin Mohammadi, Armin Zare, Mahdi Soltanolkotabi, and Mihailo R Jovanović. Convergence and sample complexity of gradient methods for the model-free linear–quadratic regulator problem. *IEEE Transactions on Automatic Control*, 67(5):2435–2450, 2021.

Lekan Molu. Mixed $\mathcal{H}_2/\mathcal{H}_\infty$-policy learning synthesis, 2023.

Boris Polyak. Gradient methods for the minimisation of functionals. *USSR Computational Mathematics and Mathematical Physics*, page 864–878, 1963.

Yue Sun and Maryam Fazel. Learning optimal controllers by policy gradient: Global optimality via convex parameterization. In *2021 60th IEEE Conference on Decision and Control (CDC)*, page 4576–4581. IEEE, Dec 2021.

Huining Yang. *Policy gradient methods for linear quadratic problems*. PhD thesis, University of Oxford, 2022.

Zhuoran Yang, Yongxin Chen, Mingyi Hong, and Zhaoran Wang. On the global convergence of actor-critic: A case for linear quadratic regulator with ergodic cost. *arXiv preprint arXiv:1907.06246*, Jul 2019.

Kaiqing Zhang, Zhuoran Yang, and Tamer Basar. Policy optimization provably converges to nash equilibria in zero-sum linear quadratic games. *Advances in Neural Information Processing Systems*, 32, 2019.

Kaiqing Zhang, Bin Hu, and Tamer Basar. On the stability and convergence of robust adversarial reinforcement learning: A case study on linear quadratic systems. *Advances in Neural Information Processing Systems*, 33:22056–22068, 2020.

Kaiqing Zhang, Bin Hu, and Tamer Başar. Policy optimization for $\mathcal{H}_2$ linear control with $\mathcal{H}_\infty$ robustness guarantee: Implicit regularization and global convergence. *SIAM Journal on Control and Optimization*, 59(6):4081–4109, 2021a. doi: 10.1137/20M1347942. URL `https://doi.org/10.1137/20M1347942`.

Kaiqing Zhang, Xiangyuan Zhang, Bin Hu, and Tamer Basar. 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.

Xiangyuan Zhang and Tamer Başar. Revisiting lqr control from the perspective of receding-horizon policy gradient. *IEEE Control Systems Letters*, pages 1–1, 2023. doi: 10.1109/LCSYS.2023.3271594.

Feiran Zhao, Xingyun Fu, and Keyou You. Global convergence of policy gradient methods for output feedback linear quadratic control. *arXiv preprint arXiv:2211.04051*, Nov 2022.

Appendix A

Proofs and Auxiliary Results

A.1 Summary of Notations

Let $\boldsymbol{K}_0 \in \mathcal{K}$ and consider the following set

$$\begin{aligned}\hat{\mathcal{K}} := \Big\{ \boldsymbol{K} \mid \text{(2.1) admits a solution } \boldsymbol{P}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})} \succeq 0, \\ \text{and } \boldsymbol{P}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})} \preceq \boldsymbol{P}_{\boldsymbol{K}_0,\boldsymbol{L}(\boldsymbol{K}_0)} + \frac{\lambda_{\min}(\boldsymbol{H}_{\boldsymbol{K}_0,\boldsymbol{L}(\boldsymbol{K}_0)})}{2\|\boldsymbol{D}\|} \cdot \boldsymbol{I} \Big\}.\end{aligned}$$

Below is a summary of the intensely used notations for convenient lookup.

$$\begin{aligned}
\varphi &:= \lambda_{\min}(\boldsymbol{\Sigma}_0) \\
\boldsymbol{H}_{\boldsymbol{K},\boldsymbol{L}} &:= \boldsymbol{R}^w - \boldsymbol{D}^\top \boldsymbol{P}_{\boldsymbol{K},\boldsymbol{L}} \boldsymbol{D} \\
\boldsymbol{E}_{\boldsymbol{K},\boldsymbol{L}} &:= (-\boldsymbol{R}^w + \boldsymbol{D}^\top \boldsymbol{P}_{\boldsymbol{K},\boldsymbol{L}} \boldsymbol{D})\boldsymbol{L} - \boldsymbol{D}^\top \boldsymbol{P}_{\boldsymbol{K},\boldsymbol{L}}(\boldsymbol{A} - \boldsymbol{B}\boldsymbol{K}) \\
\boldsymbol{F}_{\boldsymbol{K},\boldsymbol{L}} &:= (\boldsymbol{R}^u + \boldsymbol{B}^\top \boldsymbol{P}_{\boldsymbol{K},\boldsymbol{L}} \boldsymbol{B})\boldsymbol{K} - \boldsymbol{B}^\top \boldsymbol{P}_{\boldsymbol{K},\boldsymbol{L}}(\boldsymbol{A} - \boldsymbol{D}\boldsymbol{L}) \\
\boldsymbol{G}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})} &:= \boldsymbol{R}^u + \boldsymbol{B}^\top \widetilde{\boldsymbol{P}}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})} \boldsymbol{B} \\
G &:= \sup_{\boldsymbol{K} \in \hat{\mathcal{K}}} \|\boldsymbol{G}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})}\| \\
\widetilde{\boldsymbol{P}}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})} &:= \boldsymbol{P}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})} + \boldsymbol{P}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})} \boldsymbol{D}(\boldsymbol{R}^w - \boldsymbol{D}^\top \boldsymbol{P}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})} \boldsymbol{D})^{-1} \boldsymbol{D}^\top \boldsymbol{P}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})} \\
\boldsymbol{A}_{\boldsymbol{K},\boldsymbol{L}} &:= \boldsymbol{A} - \boldsymbol{B}\boldsymbol{K} - \boldsymbol{D}\boldsymbol{L} \\
\boldsymbol{L}(\boldsymbol{K}) &:= \arg\max_{\boldsymbol{L}} \mathcal{G}(\boldsymbol{K},\boldsymbol{L}) = (-\boldsymbol{R}^w + \boldsymbol{D}^\top \boldsymbol{P}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})} \boldsymbol{D})^{-1} \boldsymbol{D}^\top \boldsymbol{P}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})}(\boldsymbol{A} - \boldsymbol{B}\boldsymbol{K}) \\
\Phi(\boldsymbol{K}) &:= \mathcal{G}(\boldsymbol{K}, \boldsymbol{L}(\boldsymbol{K})) \\
\widetilde{\boldsymbol{F}}_{\boldsymbol{K},\boldsymbol{L}} &:= \frac{1}{2} \widetilde{\nabla}_{\boldsymbol{K}} \mathcal{G}(\boldsymbol{K},\boldsymbol{L}) \widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{K},\boldsymbol{L}}^{-1} \\
\nabla_{\boldsymbol{K}} \mathcal{G}_{r_2}(\boldsymbol{K},\boldsymbol{L}) &:= \mathbb{E}[\widetilde{\nabla}_{\boldsymbol{K}} \mathcal{G}(\boldsymbol{K},\boldsymbol{L})] = \nabla_{\boldsymbol{K}} \mathbb{E}_{\boldsymbol{V}}[\mathcal{G}(\boldsymbol{K} + r_2 \boldsymbol{V}, \boldsymbol{L})], \\
\boldsymbol{F}^r_{\boldsymbol{K},\boldsymbol{L}} &:= \mathbb{E}[\widetilde{\boldsymbol{F}}_{\boldsymbol{K},\boldsymbol{L}}] \\
V(\widetilde{\boldsymbol{F}}_{\boldsymbol{K},\boldsymbol{L}}) &:= (\widetilde{\boldsymbol{F}}_{\boldsymbol{K},\boldsymbol{L}} - \boldsymbol{F}^r_{\boldsymbol{K},\boldsymbol{L}})^\top (\widetilde{\boldsymbol{F}}_{\boldsymbol{K},\boldsymbol{L}} - \boldsymbol{F}^r_{\boldsymbol{K},\boldsymbol{L}}) \\
d_{\boldsymbol{\Sigma}} &:= m^2(N+1), \quad d_{\boldsymbol{K}} := d\, m\, N, \quad d_{\boldsymbol{L}} := n\, m\, N, \quad d_{\boldsymbol{P}} := m^2(N+1) \\
H &:= \sup_{\boldsymbol{K} \in \hat{\mathcal{K}}} \lambda_{\min}^{-1}(\boldsymbol{H}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})}) \\
\varepsilon_l &:= \sup_{\boldsymbol{K} \in \hat{\mathcal{K}}} \Big\{ \sqrt{\lambda_{\min}^{-1}(\boldsymbol{H}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})}) \cdot \varepsilon_1} \Big\}
\end{aligned}$$

The following notations are defined for $h = 0, \cdots, N-1$.

$$
\begin{aligned}
A_{K_h,L_h} &:= A_h - B_hK_h - D_hL_h \\
\mathcal{R}_{K_h,K'_h} &:= (K'_h - K_h)^\top F_{K_h,L(K_h)} + F^\top_{K_h,L(K_h)}(K'_h - K_h) \\
&\quad + (K'_h - K_h)^\top (R^u_h + B_h^\top \widetilde{P}_{K_{h+1},L(K_{h+1})}B_h)(K'_h - K_h) \\
G_h &:= R^u_h + B_h^\top \widetilde{P}_{K_{h+1},L(K_{h+1})}B_h \\
F^r_{K_h,L_h} &:= \mathbb{E}[\widetilde{F}_{K_h,L_h}] \\
V(\widetilde{F}_{K_h,L_h}) &:= (\widetilde{F}_{K_h,L_h} - F^r_{K_h,L_h})^\top (\widetilde{F}_{K_h,L_h} - F^r_{K_h,L_h}) \\
\Xi_{K_h,K'_h} &:= -(R^w_h - D_h^\top P_{K_{h+1},L(K_{h+1})}D_h)L(K'_h) - D_h^\top P_{K_{h+1},L(K_{h+1})}(A_h - B_hK'_h) \\
E_{K_h,L_h} &:= (-R^w_h + D_h^\top P_{K_{h+1},L_{h+1}}D_h)L_h - D_h^\top P_{K_{h+1},L_{h+1}}(A_h - B_hK_h) \\
F_{K_h,L_h} &:= (R^u_h + B_h^\top P_{K_{h+1},L_{h+1}}B_h)K_h - B_h^\top P_{K_{h+1},L_{h+1}}(A_h - D_hL_h) \\
F_{K_h,L(K_h)} &:= (R^u_h + B_h^\top \widetilde{P}_{K_{h+1},L(K_{h+1})}B_h)K_h - B_h^\top \widetilde{P}_{K_{h+1},L(K_{h+1})}A_h \\
\widetilde{P}_{K_{h+1},L(K_{h+1})} &:= P_{K_{h+1},L(K_{h+1})} + P_{K_{h+1},L(K_{h+1})}D_h(R^w_h - D_h^\top P_{K_{h+1},L(K_{h+1})}D_h)^{-1}D_h^\top P_{K_{h+1},L(K_{h+1})}
\end{aligned}
$$

Assumption 1.1 ensures that $\mathbf{\Sigma}_0 \succ 0$ and hence $\varphi > 0$.

A.2 Proof of Implicit Regularization

Proposition A.1. *(Detailed version of Proposition 3.4) Let Assumption 1.1 hold. Let $\boldsymbol{K}_0 \in \mathcal{K}$ and consider the corresponding $\hat{\mathcal{K}}$ set defined in (3.1). For any $\delta_1 \in (0,1), \epsilon_1 > 0$ and for any $\boldsymbol{K} \in \mathcal{K}$, Algorithm 1 with single-point estimation outputs $\mathbf{L}$ such that $\mathcal{G}(\boldsymbol{K}, \mathbf{L}(\boldsymbol{K})) - \mathcal{G}(\boldsymbol{K}, \mathbf{L}) \leq \epsilon_1$ with probability at least $1-\delta_1$ using $M_1 = \widetilde{\mathcal{O}}(\epsilon_1^{-2})$ samples. Moreover for any $\delta_2 \in (0,1)$ and any integer $T \geq 1$, if the estimation parameters in Algorithm 2 satisfy*

$$
\begin{aligned}
&\tau_2 \leq \min\left\{\frac{\lambda_{\min}(\boldsymbol{H}_{\boldsymbol{K}_0,\boldsymbol{L}(\boldsymbol{K}_0)})}{6\|\boldsymbol{D}\|^2}, 1/(8G), B_2/(\sqrt{m(N+1)}B_4), B_1/(\sqrt{m(N+1)}B_4), 1\right\}, \\
&r_2 \leq \min\left\{D_1, \sqrt{1/(Tc_1)}\right\}, \quad \epsilon_1 \leq \min\left\{D_3, \frac{1}{Tc_2}\right\}, \quad \delta_1 \leq \delta_2/(2T), \\
&M_2 \geq \max\left\{M_{\mathbf{\Sigma}}(\varphi/2, \delta_2/(4T)), M_{\mathbf{\Sigma}}\Big(\frac{\varphi^2}{4O_1}\cdot\sqrt{\frac{2}{c_3T}}, \delta_2/(4T)\Big), M_V\Big(\frac{\varphi}{4}\cdot\sqrt{\frac{2}{c_3T}}, \delta_2/(4T)\Big)\right\} \\
&\quad = \widetilde{\mathcal{O}}(T^2),
\end{aligned}
$$

where $M_{\mathbf{\Sigma}}(\epsilon,\delta), M_V(\epsilon,\delta)$ are defined in Lemma A.19 and A.18 respectively. G, B_1, B_2, B_4, are defined in Lemma A.6, A.8, A.7, A.17. D_1, D_3 are defined in Lemma A.6. Then, it holds with probability at least $1-\delta_2$ that $\boldsymbol{K}_t \in \hat{\mathcal{K}}$ for all $t = 1, \cdots, T$.

Proof. The results for solving the inner-loop problem have been discussed in Theorem 3.8 of Zhang et al. [2021b] and hence omitted here. For the outer-loop algorithm, we firstly consider one step update from $\boldsymbol{K}_0 \in \mathcal{K}$ to $\boldsymbol{K}_1$ with $\boldsymbol{K}_1 = \boldsymbol{K}_0 - \tau_2\widetilde{\boldsymbol{F}}_{\boldsymbol{K}_0,\boldsymbol{L}_0}$ where $\boldsymbol{L}_0$ is the output of the max-oracle given $\boldsymbol{K}_0$. We already know that using exact outer-loop natural gradients, $2\boldsymbol{F}_{\boldsymbol{K}_t,\boldsymbol{L}(\boldsymbol{K}_t)}$, we can ensure the non-increasing monotonicity of $\boldsymbol{P}_{\boldsymbol{K}_t,\boldsymbol{L}(\boldsymbol{K}_t)}$ Zhang et al. [2021b]. While in the model-free setting, the estimated gradients will lead to deviations, we will prove shortly that the deviation can be well-controlled (within set $\hat{\mathcal{K}}$) using good approximations.

We try to control the deviation of $\boldsymbol{P}_{\boldsymbol{K}_t,\boldsymbol{L}(\boldsymbol{K}_t)}$ from non-increasing monotonity. We start from the upperbound we developed in Lemma A.15 for difference between matrices

$\boldsymbol{P}_{\boldsymbol{K}_1,\boldsymbol{L}(\boldsymbol{K}_1)}$ and $\boldsymbol{P}_{\boldsymbol{K}_0,\boldsymbol{L}(\boldsymbol{K}_0)}$:

$$\boldsymbol{P}_{\boldsymbol{K}',\boldsymbol{L}(\boldsymbol{K}')} - \boldsymbol{P}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})} \preceq \sum_{i=0}^{N} (\boldsymbol{A}_{\boldsymbol{K}',\boldsymbol{L}(\boldsymbol{K}')}^\top)^i (\boldsymbol{e}_{1,\boldsymbol{K},\boldsymbol{K}'} + \boldsymbol{e}_{2,\boldsymbol{K},\boldsymbol{K}'} + \boldsymbol{e}_{3,\boldsymbol{K},\boldsymbol{K}'})(\boldsymbol{A}_{\boldsymbol{K}',\boldsymbol{L}(\boldsymbol{K}')})^i - \frac{\tau_2}{4}\boldsymbol{F}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})}^\top \boldsymbol{F}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})},$$

where $\boldsymbol{e}_{1,\boldsymbol{K},\boldsymbol{K}'}, \boldsymbol{e}_{2,\boldsymbol{K},\boldsymbol{K}'}, \boldsymbol{e}_{3,\boldsymbol{K},\boldsymbol{K}'}$ are errors terms that we try to upperbound. Recall

$$\begin{aligned}
\boldsymbol{e}_{1,\boldsymbol{K},\boldsymbol{K}'} &:= (4\tau_2 + 4\tau_2^2\|\boldsymbol{G}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})}\|)(\boldsymbol{F}_{\boldsymbol{K},\boldsymbol{L}}^r - \boldsymbol{F}_{\boldsymbol{K},\boldsymbol{L}})^\top (\boldsymbol{F}_{\boldsymbol{K},\boldsymbol{L}}^r - \boldsymbol{F}_{\boldsymbol{K},\boldsymbol{L}}) \\
\boldsymbol{e}_{2,\boldsymbol{K},\boldsymbol{K}'} &:= \tau_2(\boldsymbol{F}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})} - \boldsymbol{F}_{\boldsymbol{K},\boldsymbol{L}})^\top (\boldsymbol{F}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})} - \boldsymbol{F}_{\boldsymbol{K},\boldsymbol{L}}) \\
\boldsymbol{e}_{3,\boldsymbol{K},\boldsymbol{K}'} &:= (2\tau_2 + \tau_2^2\|\boldsymbol{G}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})}\|) V(\widetilde{\boldsymbol{F}}_{\boldsymbol{K},\boldsymbol{L}}),
\end{aligned}$$

where $\boldsymbol{F}_{\boldsymbol{K},\boldsymbol{L}}^r := \mathbb{E}[\widetilde{\boldsymbol{F}}_{\boldsymbol{K},\boldsymbol{L}}]$, $V(\widetilde{\boldsymbol{F}}_{\boldsymbol{K},\boldsymbol{L}}) := (\widetilde{\boldsymbol{F}}_{\boldsymbol{K},\boldsymbol{L}} - \mathbb{E}[\widetilde{\boldsymbol{F}}_{\boldsymbol{K},\boldsymbol{L}}])^\top (\widetilde{\boldsymbol{F}}_{\boldsymbol{K},\boldsymbol{L}} - \mathbb{E}[\widetilde{\boldsymbol{F}}_{\boldsymbol{K},\boldsymbol{L}}])$. We can observe from the above definitions that $\boldsymbol{e}_{1,\boldsymbol{K},\boldsymbol{K}'}$ is caused by biased estimation of the natural gradients, $\boldsymbol{e}_{2,\boldsymbol{K},\boldsymbol{K}'}$ is from the errors in solving the inner-loop problem, and $\boldsymbol{e}_{3,\boldsymbol{K},\boldsymbol{K}'}$ is the variance-like term that can be controlled by choosing proper sample sizes M_2. The first two error terms can be controlled by choosing proper parameters r_2 and ϵ_1. Apply Lemma A.20, we know that by choosing

$$\begin{aligned}
&\tau_2 \le \min\Big\{1/(8G), B_2/(\sqrt{m(N+1)}B_4), B_1/(\sqrt{m(N+1)}B_4), 1\Big\}, \quad r_2 \le D_1, \\
&\epsilon_1 \le D_3, \quad M_2 \ge \max\{M_{\boldsymbol{\Sigma}}(\varphi/2, \delta/2), M_V(1, \delta/2)\}.
\end{aligned}$$

Then with probability at least $(1-\delta_1)(1-\delta)$, we have $\boldsymbol{K}' \in \mathcal{K}$ and

$$\boldsymbol{P}_{\boldsymbol{K}_1,\boldsymbol{L}(\boldsymbol{K}_1)} - \boldsymbol{P}_{\boldsymbol{K}_0,\boldsymbol{L}(\boldsymbol{K}_0)} \preceq \tau_2 \cdot (c_1 \cdot r_2^2 + c_2 \cdot \epsilon_1 + c_3 \cdot \|V(\widetilde{\boldsymbol{F}}_{\boldsymbol{K}_0,\boldsymbol{L}_0})\|) \cdot I. \tag{A.1}$$

To further constrain $\boldsymbol{K}_1$ within $\hat{\mathcal{K}}$ set, we choose small r_2, small τ_2, and large enough M_2 such that $V(\widetilde{\boldsymbol{F}}_{\boldsymbol{K}_0,\boldsymbol{L}_{\boldsymbol{K}_0}})$ is small: we require

$$c_1 \cdot r_2^2 \le 1, \quad c_2 \cdot \epsilon_1 \le 1, \quad c_3 \cdot \|V(\widetilde{\boldsymbol{F}}_{\boldsymbol{K}_0,\boldsymbol{L}_0})\| \le 1.$$

Additionally, let $\tau_2 \le \frac{\lambda_{\min}(\boldsymbol{H}_{\boldsymbol{K}_0,\boldsymbol{L}(\boldsymbol{K}_0)})}{6\|\boldsymbol{D}\|^2}$, then we have

$$\boldsymbol{P}_{\boldsymbol{K}_1,\boldsymbol{L}(\boldsymbol{K}_1)} \preceq \boldsymbol{P}_{\boldsymbol{K}_0,\boldsymbol{L}(\boldsymbol{K}_0)} + \frac{\lambda_{\min}(\boldsymbol{H}_{\boldsymbol{K}_0,\boldsymbol{L}(\boldsymbol{K}_0)})}{2\|\boldsymbol{D}\|^2} \cdot I.$$

Hence $\boldsymbol{K}_1 \in \hat{\mathcal{K}}$ is guaranteed. Now by applying the above reasoning recursively: by choosing

$$\begin{aligned}
&\tau_2 \le \min\left\{\frac{\lambda_{\min}(\boldsymbol{H}_{\boldsymbol{K}_0,\boldsymbol{L}(\boldsymbol{K}_0)})}{6\|\boldsymbol{D}\|^2}, 1/(8G), B_2/(\sqrt{m(N+1)}B_4), B_1/(\sqrt{m(N+1)}B_4), 1\right\}, \\
&r_2 \le \min\Big\{D_1, \sqrt{1/(Tc_1)}\Big\}, \quad \epsilon_1 \le \min\left\{D_3, \frac{1}{Tc_2}\right\}, \quad \delta_1 \le \delta_2/(2T), \\
&M_2 \ge \max\left\{M_{\boldsymbol{\Sigma}}(\varphi/2, \delta_2/(4T)), M_{\boldsymbol{\Sigma}}(\frac{\varphi^2}{4O_1} \cdot \sqrt{\frac{2}{c_3 T}}, \delta_2/(4T)), M_V(\frac{\varphi}{4} \cdot \sqrt{\frac{2}{c_3 T}}, \delta_2/(4T))\right\} \\
&\quad = \widetilde{\mathcal{O}}(T^2),
\end{aligned}$$

where we apply Lemma A.18 for the choice of M_2 to control $\|V(\widetilde{\boldsymbol{F}}_{\boldsymbol{K},\boldsymbol{L}_0})\|$, we obtain that

$$c_1 \cdot r_2^2 \cdot T \le 1, \quad c_2 \cdot \epsilon_1 \cdot T \le 1, \quad c_3 \cdot \sum_{t=0}^{T-1} \|V(\widetilde{\boldsymbol{F}}_{\boldsymbol{K}_t,\boldsymbol{L}_t})\| \le 1,$$

hold with probability at least $1-\delta_2$. Then we can compute the telescoping sum of (A.1)

$$\begin{aligned} \boldsymbol{P}_{\boldsymbol{K}_t,\boldsymbol{L}(\boldsymbol{K}_t)} &\preceq \boldsymbol{P}_{\boldsymbol{K}_0,\boldsymbol{L}(\boldsymbol{K}_0)} + \tau_2 \cdot \sum_{t=0}^{T-1} (c_1 r_2^2 + c_2\varepsilon_1 + c_3\|V(\widetilde{\boldsymbol{F}}_{\boldsymbol{K}_t,\boldsymbol{L}_t})\|) \cdot I \\ &\preceq \boldsymbol{P}_{\boldsymbol{K}_0,\boldsymbol{L}(\boldsymbol{K}_0)} + \frac{\lambda_{\min}(\boldsymbol{H}_{\boldsymbol{K}_0,\boldsymbol{L}(\boldsymbol{K}_0)})}{2\|\mathbf{D}\|^2} \cdot I, \end{aligned}$$

hold for $t = 0,\cdots,T$ with probability at least $1-\delta_2$, i.e., $\boldsymbol{K}_0,\cdots,\boldsymbol{K}_T$ stay in $\hat{\mathcal{K}}$ with probability at least $1-\delta_2$. $\square$

A.3 Proof of Sample Complexity Improvement with 1-Point Estimation

Theorem A.2. *(Detailed version of Theorem 3.5: Improved Sample Complexity for Outer-loop Algorithm in Zhang et al. [2021b]) Let Assumption 1.1 hold. Let $\boldsymbol{K}_0 \in \mathcal{K}$ and consider the corresponding $\hat{\mathcal{K}}$ set defined in (3.1). For any $\delta_1 \in (0,1), \varepsilon_1 > 0$ and for any $\boldsymbol{K} \in \mathcal{K}$, Algorithm 1 with single-point estimation outputs $\boldsymbol{L}$ such that $\mathcal{G}(\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})) - \mathcal{G}(\boldsymbol{K},\boldsymbol{L}) \le \varepsilon_1$ with probability at least $1-\delta_1$ using $T_{in}M_1 = \widetilde{\mathcal{O}}(\varepsilon_1^{-2})$ samples. Moreover for any $\delta_2 \in (0,1)$ and any integer $T \ge 1$, if the estimation parameters in Algorithm 2 satisfy*

$$\begin{aligned} &\tau_2 \le \min\left\{\frac{\lambda_{\min}(\boldsymbol{H}_{\boldsymbol{K}_0,\boldsymbol{L}(\boldsymbol{K}_0)})}{6\|\boldsymbol{D}\|^2}, 1/(8G), B_2/(\sqrt{m(N+1)}B_4), B_1/(\sqrt{m(N+1)}B_4), 1\right\}, \\ &r_2 \le \min\left\{D_1, \sqrt{1/(Tc_1)}\right\}, \quad \varepsilon_1 \le \min\left\{D_3, \frac{1}{Tc_2}\right\}, \quad \delta_1 \le \delta/(2T), \\ &M_2 \ge \max\left\{M_{\boldsymbol{\Sigma}}(\varphi/2, \delta/(4T)), M_{\boldsymbol{\Sigma}}(\frac{\varphi^2}{4O_1}\cdot\sqrt{\frac{1}{c_3T}}, \delta/(4T)), M_V(\frac{\varphi}{4}\cdot\sqrt{\frac{1}{c_3T}}, \delta/(4T))\right\} \\ &\quad = \widetilde{\mathcal{O}}(T^2), \end{aligned}$$

where $M_{\boldsymbol{\Sigma}}(\varepsilon,\delta), M_V(\varepsilon,\delta)$ are defined in Lemma A.19 and A.18 respectively. Here G, B_1, B_2, B_4 are defined in Lemma A.6, A.8, A.7, A.17. And D_1, D_3 are defined in Lemma A.6. Constants c_1, c_2, c_3 are defined in Lemma A.20. Then, it holds with probability at least $1-\delta_2$ that $\boldsymbol{K}_t \in \hat{\mathcal{K}}$ for all $t = 1,\cdots,T$. Moreover, if we require

$$\varepsilon = \frac{\lambda_{\min}(\boldsymbol{H}_{\boldsymbol{K}_0,\boldsymbol{L}(\boldsymbol{K}_0)})}{\tau_2\varphi T\|\boldsymbol{D}\|}\mathrm{Tr}(\boldsymbol{\Sigma}_0),$$

Algorithm 2 returns an ε-stationary point of $\varphi(\boldsymbol{K})$ in the sense that

$$\frac{1}{T}\sum_{t=0}^{T-1}\|\boldsymbol{F}_{\boldsymbol{K}_t,\boldsymbol{L}(\boldsymbol{K}_t)}\|_F^2 \le \varepsilon.$$

And hence the sample complexity is of order $\widetilde{\mathcal{O}}(\varepsilon^{-3})$.

Proof. The first part of the theorem is proved in Appendix A.2. As for the convergence rate, multiply the descent inequality with $\boldsymbol{\Sigma}_0$ and take the trace:

$$\begin{aligned} \boldsymbol{P}_{\boldsymbol{K}_{t+1},\boldsymbol{L}(\boldsymbol{K}_{t+1})} - \boldsymbol{P}_{\boldsymbol{K}_t,\boldsymbol{L}(\boldsymbol{K}_t)} &\preceq \sum_{i=0}^{N}(\boldsymbol{A}_{\boldsymbol{K}_{t+1},\boldsymbol{L}(\boldsymbol{K}_{t+1})}^\top)^i \boldsymbol{e}_{\boldsymbol{K}_t,\boldsymbol{K}_{t+1}}(\boldsymbol{A}_{\boldsymbol{K}_{t+1},\boldsymbol{L}(\boldsymbol{K}_{t+1})})^i - \tau_2\boldsymbol{F}_{\boldsymbol{K}_t,\boldsymbol{L}(\boldsymbol{K}_t)}^\top\boldsymbol{F}_{\boldsymbol{K}_t,\boldsymbol{L}(\boldsymbol{K}_t)}, \\ \boldsymbol{e}_{\boldsymbol{K}_t,\boldsymbol{K}_{t+1}} &:= \boldsymbol{e}_{1,\boldsymbol{K}_t,\boldsymbol{K}_{t+1}} + \boldsymbol{e}_{2,\boldsymbol{K}_t,\boldsymbol{K}_{t+1}} + \boldsymbol{e}_{3,\boldsymbol{K}_t,\boldsymbol{K}_{t+1}} \end{aligned}$$

By computing the telescoping sum, we can see that

$$
\begin{aligned}
\Phi(\boldsymbol{K}_T) - \Phi(\boldsymbol{K}_0) &\leq \sum_{t=0}^{T-1} \mathrm{Tr}\Big(\sum_{i=0}^{N} (\boldsymbol{A}_{\boldsymbol{K}_{t+1},\boldsymbol{L}(\boldsymbol{K}_{t+1})}^{\top})^i \boldsymbol{e}_{\boldsymbol{K}_t,\boldsymbol{K}_{t+1}} (\boldsymbol{A}_{\boldsymbol{K}_{t+1},\boldsymbol{L}(\boldsymbol{K}_{t+1})})^i \boldsymbol{\Sigma}_0\Big) \\
&\quad - \tau_2 \,\mathrm{Tr}(\boldsymbol{F}_{\boldsymbol{K}_t,\boldsymbol{L}(\boldsymbol{K}_t)}^{\top} \boldsymbol{F}_{\boldsymbol{K}_t,\boldsymbol{L}(\boldsymbol{K}_t)} \boldsymbol{\Sigma}_0), \\
\frac{1}{T}\sum_{t=0}^{T-1} \|\boldsymbol{F}_{\boldsymbol{K}_t,\boldsymbol{L}(\boldsymbol{K}_t)}\|_F^2 &\leq \frac{1}{T}\sum_{t=0}^{T-1} \mathrm{Tr}(\boldsymbol{F}_{\boldsymbol{K}_t,\boldsymbol{L}(\boldsymbol{K}_t)}^{\top} \boldsymbol{F}_{\boldsymbol{K}_t,\boldsymbol{L}(\boldsymbol{K}_t)}) \\
&\leq \frac{1}{\tau_2 \varphi T}\Bigg(\Phi(\boldsymbol{K}_0) - \Phi(\boldsymbol{K}_T) \\
&\quad + \sum_{t=0}^{T-1} \mathrm{Tr}\Big(\sum_{i=0}^{N} (\boldsymbol{A}_{\boldsymbol{K}_{t+1},\boldsymbol{L}(\boldsymbol{K}_{t+1})}^{\top})^i \boldsymbol{e}_{\boldsymbol{K}_t,\boldsymbol{K}_{t+1}} (\boldsymbol{A}_{\boldsymbol{K}_{t+1},\boldsymbol{L}(\boldsymbol{K}_{t+1})})^i \boldsymbol{\Sigma}_0\Big)\Bigg) \\
&\leq \frac{1}{\tau_2 \varphi T}\Bigg(\Bigg(\frac{\lambda_{\min}(\boldsymbol{H}_{\boldsymbol{K}_0,\boldsymbol{L}(\boldsymbol{K}_0)})}{2\|\boldsymbol{D}\|^2} \\
&\quad + \sum_{t=0}^{T-1} \Big\|\sum_{i=0}^{N} (\boldsymbol{A}_{\boldsymbol{K}_{t+1},\boldsymbol{L}(\boldsymbol{K}_{t+1})}^{\top})^i \boldsymbol{e}_{\boldsymbol{K}_t,\boldsymbol{K}_{t+1}} (\boldsymbol{A}_{\boldsymbol{K}_{t+1},\boldsymbol{L}(\boldsymbol{K}_{t+1})})^i\Big\|\Bigg) \mathrm{Tr}(\boldsymbol{\Sigma}_0)\Bigg) \\
&\leq \frac{1}{\tau_2 \varphi T}\left(\frac{\lambda_{\min}(\boldsymbol{H}_{\boldsymbol{K}_0,\boldsymbol{L}(\boldsymbol{K}_0)})}{\|\boldsymbol{D}\|^2} \mathrm{Tr}(\boldsymbol{\Sigma}_0)\right).
\end{aligned}
$$

where in the fifth inequality, we apply the same reasoning in Appendix A.2: by the same choice of parameters, we can control the deviation:

$$
\sum_{t=0}^{T-1} \Big\|\sum_{i=0}^{N} (\boldsymbol{A}_{\boldsymbol{K}_{t+1},\boldsymbol{L}(\boldsymbol{K}_{t+1})}^{\top})^i \boldsymbol{e}_{\boldsymbol{K}_t,\boldsymbol{K}_{t+1}} (\boldsymbol{A}_{\boldsymbol{K}_{t+1},\boldsymbol{L}(\boldsymbol{K}_{t+1})})^i\Big\| \leq \frac{\lambda_{\min}(\boldsymbol{H}_{\boldsymbol{K}_0,\boldsymbol{L}(\boldsymbol{K}_0)})}{2\|\boldsymbol{D}\|^2}
$$

with high probability. Conclusively, we obtain high probability sublinear convergence in the measure of

$$
\frac{1}{T}\sum_{t=0}^{T-1} \|\boldsymbol{F}_{\boldsymbol{K}_t,\boldsymbol{L}(\boldsymbol{K}_t)}\|_F^2 \leq \frac{1}{T}\cdot\frac{1}{\tau_2\varphi}\left(\frac{\lambda_{\min}(\boldsymbol{H}_{\boldsymbol{K}_0,\boldsymbol{L}(\boldsymbol{K}_0)})}{\|\boldsymbol{D}\|^2} \mathrm{Tr}(\boldsymbol{\Sigma}_0)\right).
$$

If we choose

$$
\varepsilon = \frac{\lambda_{\min}(\boldsymbol{H}_{\boldsymbol{K}_0,\boldsymbol{L}(\boldsymbol{K}_0)})}{\tau_2\varphi T\|\boldsymbol{D}\|} \mathrm{Tr}(\boldsymbol{\Sigma}_0),
$$

then we have

$$
\frac{1}{T}\sum_{t=0}^{T-1} \|\boldsymbol{F}_{\boldsymbol{K}_t,\boldsymbol{L}(\boldsymbol{K}_t)}\|_F^2 \leq \varepsilon.
$$

Let $T_{in} \cdot M_1$ be the sample complexity of the inner problem, M_2 be the number of samples required for each iteration of the outer loop. From Zhang et al. [2021b], we already know the sample complexity of the inner-loop is $\tilde{\mathcal{O}}(\varepsilon_1^{-2}) \cdot T$. From the choice of ε_1, ε, we know that $\varepsilon_1 = \mathcal{O}(\varepsilon) = \mathcal{O}(T^{-1})$. Then the total complexity can be computed as

$$
T \cdot (T_{in} \cdot M_1 + M_2) = \tilde{\mathcal{O}}(\varepsilon^{-1}(\varepsilon_1^{-2} + M_2)) = \tilde{\mathcal{O}}(\varepsilon^{-3}).
$$

$\square$

A.4 Proof of Sample Complexity Improvement with 2-Point Estimation

Theorem A.3. *(Detailed version of Theorem 3.7) Let Assumption 1.1 hold. Let $\boldsymbol{K}_0 \in \mathcal{K}$ and consider the corresponding $\hat{\mathcal{K}}$ set defined in (3.1). For any $\delta_1 \in (0,1), \varepsilon_1 > 0$ and for any $\boldsymbol{K} \in \mathcal{K}$, Algorithm 1 with two-point estimation outputs $\mathbf{L}$ such that $\mathcal{G}(\boldsymbol{K}, \boldsymbol{L}(\boldsymbol{K})) - \mathcal{G}(\boldsymbol{K}, \boldsymbol{L}) \leq \varepsilon_1$ with probability at least $1 - \delta_1$ using $T_{in}M_1 = \widetilde{O}(\varepsilon_1^{-1})$ samples. Moreover for any $\delta_2 \in (0,1)$ and any integer $T \geq 1$, if the estimation parameters in Algorithm 2 satisfy*

$$
\begin{aligned}
&\tau_2 \leq \min\left\{\frac{\lambda_{\min}(\boldsymbol{H}_{\boldsymbol{K}_0, \boldsymbol{L}(\boldsymbol{K}_0)})}{6\|\boldsymbol{D}\|^2}, 1/(8G), B_2/(\sqrt{m(N+1)}B_4), B_1/(\sqrt{m(N+1)}B_4), 1\right\},\\
&r_2 \leq \min\left\{D_1, \sqrt{1/(Tc_1)}\right\}, \quad \varepsilon_1 \leq \min\left\{D_3, \frac{1}{Tc_2}\right\}, \quad \delta_1 \leq \delta/(2T),\\
&M_2 \geq \max\left\{M_{\boldsymbol{\Sigma}}(\varphi/2, \delta/(4T)), M_{\boldsymbol{\Sigma}}(\frac{\varphi^2}{4O_1}\cdot\sqrt{\frac{1}{c_3T}}, \delta/(4T)), M_V'(\frac{\varphi}{4}\cdot\sqrt{\frac{1}{c_3T}}, \delta/(4T))\right\}\\
&\quad = \widetilde{\mathcal{O}}(T),\\
&M_V'(\varepsilon,\delta) := (\frac{O_2'}{\varepsilon})^2 \cdot \log(\frac{2d_{\boldsymbol{K}}}{\delta}), \quad O_2' := d_{\boldsymbol{K}}(N+1)\vartheta^2 l_5 + O_1,
\end{aligned}
$$

where positive constant O_1 is defined in Lemma A.18, $M_{\boldsymbol{\Sigma}}(\varepsilon,\delta), M_V(\varepsilon,\delta)$ are defined in Lemma A.19 and A.18 respectively. Here G, B_1, B_2, B_4 are defined in Lemma A.6, A.8, A.7, A.17. And D_1, D_3 are defined in Lemma A.6. Constants c_1, c_2, c_3 are defined in Lemma A.20. Then, it holds with probability at least $1-\delta_2$ that $\boldsymbol{K}_t \in \hat{\mathcal{K}}$ for all $t = 1, \cdots, T$. Moreover, if we require

$$
\varepsilon = \frac{\lambda_{\min}(\boldsymbol{H}_{\boldsymbol{K}_0, \boldsymbol{L}(\boldsymbol{K}_0)})}{\tau_2 \varphi T \|\boldsymbol{D}\|} \operatorname{Tr}(\boldsymbol{\Sigma}_0),
$$

Algorithm 2 returns an ε-stationary point of $\varphi(\boldsymbol{K})$ in the sense that

$$
\frac{1}{T}\sum_{t=0}^{T-1} \|\boldsymbol{F}_{\boldsymbol{K}_t, \boldsymbol{L}(\boldsymbol{K}_t)}^\top\|_F^2 \leq \varepsilon.
$$

And the sample complexity is of order $\widetilde{\mathcal{O}}(\varepsilon^{-2})$.

Proof. Here the main steps are the same as Proposition 3.4 and Theorem 3.5. The difference now is the relationship between the variance of gradient estimation and the sample size, in other words, Lemma A.18 needs adaptations for the two-point estimation method. By using two-point estimation, we do not have variance $\propto r_2^{-2}$. As in Lemma A.18, we consider random variable $\frac{d_{\boldsymbol{K}}}{2r_2}(\mathcal{G}_{\boldsymbol{\xi}}(\boldsymbol{K} + r_2\boldsymbol{V}, \boldsymbol{L}) - \mathcal{G}_{\boldsymbol{\xi}}(\boldsymbol{K} - r_2\boldsymbol{V}, \boldsymbol{L}))\boldsymbol{V}$ where $\boldsymbol{V}$ is sampled uniformly randomly from the unit sphere.

$$
\begin{aligned}
&\left\|\frac{d_{\boldsymbol{K}}}{2r_2}\left(\mathcal{G}_{\boldsymbol{\xi}}(\boldsymbol{K} + r_2\boldsymbol{V}, \boldsymbol{L})\boldsymbol{V} - \mathcal{G}_{\boldsymbol{\xi}}(\boldsymbol{K} - r_2\boldsymbol{V}, \boldsymbol{L})\boldsymbol{V}\right)\right\|_F\\
&= \left|\frac{d_{\boldsymbol{K}}}{2r_2}\boldsymbol{\xi}^\top(\boldsymbol{P}_{\boldsymbol{K}+r_2\boldsymbol{V},\boldsymbol{L}} - \boldsymbol{P}_{\boldsymbol{K}-r_2\boldsymbol{V},\boldsymbol{L}})\boldsymbol{\xi}\right| \leq \frac{d_{\boldsymbol{K}}}{2r_2}\|\boldsymbol{\xi}\|^2\left(\|\boldsymbol{P}_{\boldsymbol{K}+r_2\boldsymbol{V},\boldsymbol{L}} - \boldsymbol{P}_{\boldsymbol{K},\boldsymbol{L}}\| + \|\boldsymbol{P}_{\boldsymbol{K},\boldsymbol{L}} - \boldsymbol{P}_{\boldsymbol{K}-r_2\boldsymbol{V},\boldsymbol{L}}\|\right)\\
&\leq \frac{d_{\boldsymbol{K}}}{2r_2}(N+1)\vartheta^2 l_5 \cdot r_2 \cdot 2 = d_{\boldsymbol{K}}(N+1)\vartheta^2 l_5,
\end{aligned}
$$

with probability at least $1-\delta_1$. In the second inequality, we apply Lemma A.9 and choose

$$
r_2 \leq D_1, \quad \varepsilon_1 \leq D_3.
$$

Hence $\frac{d_{\boldsymbol{K}}}{r_2}\mathcal{G}(\boldsymbol{K}+r_2\boldsymbol{V},\boldsymbol{L})\boldsymbol{V} - \nabla_{\boldsymbol{K}}\mathcal{G}_{r_2}(\boldsymbol{K},\boldsymbol{L})$ is bounded for any $\boldsymbol{V}$ with probability at least $1-\delta_1$, and hence norm-subGaussian with probability at least $1-\delta_1$. Then we apply Corollary 7 in Jin et al. [2019], with probability at least $(1-\delta_1)(1-\delta)$, we have

$$\begin{aligned}
&\|\widetilde{\nabla}_{\boldsymbol{K}}\mathcal{G}(\boldsymbol{K},\boldsymbol{L}) - \nabla_{\boldsymbol{K}}\mathcal{G}_{r_2}(\boldsymbol{K},\boldsymbol{L})\|_F \\
&= \left\| \frac{1}{M_2}\sum_{m=0}^{M_2-1}\frac{d_{\boldsymbol{K}}}{2r_2}(\mathcal{G}_{\boldsymbol{\xi}_m}(\boldsymbol{K}+r_2\boldsymbol{V}_m,\boldsymbol{L})\boldsymbol{V}_m - \mathcal{G}_{\boldsymbol{\xi}_m}(\boldsymbol{K}-r_2\boldsymbol{V}_m,\boldsymbol{L})\boldsymbol{V}_m) - \nabla_{\boldsymbol{K}}\mathcal{G}_{r_2}(\boldsymbol{K},\boldsymbol{L})\right\|_F \\
&\le \frac{1}{M_2}\cdot\sqrt{M_2}\cdot(d_{\boldsymbol{K}}(N+1)\vartheta^2 l_5 + O_1)\cdot\sqrt{\log(\frac{2d_{\boldsymbol{K}}}{\delta})},
\end{aligned}$$

where positive constant O_1 is defined in Lemma A.18. Hence when we sample

$$\begin{aligned}
M_2 &\ge \max\{M_V'(\frac{\sqrt{2\varepsilon}\cdot\varphi}{4},\frac{\delta}{2}), M_{\boldsymbol{\Sigma}}(\varphi/2,\delta/2)\}, \\
M_V'(\varepsilon,\delta) &:= (\frac{O_2'}{\varepsilon})^2\cdot\log(\frac{2d_{\boldsymbol{K}}}{\delta}) = \mathcal{O}(\frac{1}{\varepsilon^2}\cdot\log(\frac{1}{\delta})), \quad O_2' := d_{\boldsymbol{K}}(N+1)\vartheta^2 l_5 + O_1,
\end{aligned}$$

we have

$$\|\widetilde{\nabla}_{\boldsymbol{K}}\mathcal{G}(\boldsymbol{K},\boldsymbol{L}) - \nabla_{\boldsymbol{K}}\mathcal{G}_{r_2}(\boldsymbol{K},\boldsymbol{L})\|^2\cdot\|\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{K},\boldsymbol{L}}^{-1}\|^2 \le \|\widetilde{\nabla}_{\boldsymbol{K}}\mathcal{G}(\boldsymbol{K},\boldsymbol{L}) - \nabla_{\boldsymbol{K}}\mathcal{G}_{r_2}(\boldsymbol{K},\boldsymbol{L})\|_F^2\cdot\|\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{K},\boldsymbol{L}}^{-1}\|^2 \le \frac{\varepsilon}{2},$$

with probability at least $(1-\delta_1)(1-\delta)$. Moreover, apply Lemma A.19 and choose

$$M_2 \ge \max\left\{M_{\boldsymbol{\Sigma}}(\varphi/2,\delta/2), M_{\boldsymbol{\Sigma}}(\frac{\varphi^2\sqrt{2\varepsilon}}{4O_1},\delta/2)\right\}$$

We can bound

$$\begin{aligned}
\|\nabla_{\boldsymbol{K}}\mathcal{G}_{r_2}(\boldsymbol{K},\boldsymbol{L})\|^2\cdot\|\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{K},\boldsymbol{L}}^{-1} - \boldsymbol{\Sigma}_{\boldsymbol{K},\boldsymbol{L}}^{-1}\|^2 &\le \|\nabla_{\boldsymbol{K}}\mathcal{G}_{r_2}(\boldsymbol{K},\boldsymbol{L})\|^2\cdot\|\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{K},\boldsymbol{L}}^{-1}\|^2\|\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{K},\boldsymbol{L}} - \boldsymbol{\Sigma}_{\boldsymbol{K},\boldsymbol{L}}\|^2\|\boldsymbol{\Sigma}_{\boldsymbol{K},\boldsymbol{L}}^{-1}\|^2 \\
&\le \frac{4}{\varphi^4}(O_1)^2\|\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{K},\boldsymbol{L}} - \boldsymbol{\Sigma}_{\boldsymbol{K},\boldsymbol{L}}\|^2 \le \frac{\varepsilon}{2}
\end{aligned}$$

with probability at least $(1-\delta_1)(1-\delta/2)$. In conclusion, by sampling

$$\begin{aligned}
M_2 &\ge \max\left\{M_{\boldsymbol{\Sigma}}(\varphi/2,\delta/2), M_{\boldsymbol{\Sigma}}(\frac{\varphi^2\sqrt{2\varepsilon}}{4O_1},\delta/2), M_V'(\frac{\sqrt{2\varepsilon}\varphi}{4},\delta/2)\right\} \\
&= \mathcal{O}(\frac{1}{\varepsilon}\cdot\log(\frac{1}{\delta}) + \varepsilon^{-1}\cdot\log(\frac{1}{\delta})),
\end{aligned}$$

we have

$$\begin{aligned}
\|V(\widetilde{\boldsymbol{F}}_{\boldsymbol{K},\boldsymbol{L}})\| &\le \|\widetilde{\nabla}_{\boldsymbol{K}}\mathcal{G}(\boldsymbol{K},\boldsymbol{L}) - \nabla_{\boldsymbol{K}}\mathcal{G}_{r_2}(\boldsymbol{K},\boldsymbol{L})\|^2\cdot\|\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{K},\boldsymbol{L}}^{-1}\|^2 + \|\nabla_{\boldsymbol{K}}\mathcal{G}_{r_2}(\boldsymbol{K},\boldsymbol{L})\|^2\cdot\|\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{K},\boldsymbol{L}}^{-1} - \boldsymbol{\Sigma}_{\boldsymbol{K},\boldsymbol{L}}^{-1}\|^2 \\
&\le \varepsilon/2 + \varepsilon/2 = \varepsilon,
\end{aligned}$$

with probability at least $(1-\delta_1)(1-\delta)$. Here to adapt the proof of Proposition 3.4 and Theorem 3.5, substitute the relationship between the sample size M_2 and $\|V(\widetilde{\boldsymbol{F}}_{\boldsymbol{K},\boldsymbol{L}})\|_F$ in Lemma A.18 with the proof above, the other parts still apply and hence omitted here.

From Malik et al. [2019], we know the inner-loop problem be solved by sample size $M_1 = \mathcal{O}(\varepsilon_1^{-1})$ using two-point estimation. Hence let $T_{in}\cdot M_1$ be the sample complexity of the inner problem, M_2 be the number of samples required for each iteration of the outer loop. From the choice of ε_1, ε, we know that $\varepsilon_1 = \mathcal{O}(\varepsilon) = \mathcal{O}(T^{-1})$. Then the total complexity can be computed as

$$T\cdot(T_{in}\cdot M_1 + M_2) = \tilde{\mathcal{O}}(\varepsilon^{-1}(\varepsilon_1^{-1}+\varepsilon^{-1})) = \tilde{\mathcal{O}}(\varepsilon^{-2}).$$

□

A.5 Proof of Last-iterate Convergence (Deterministic)

Theorem A.4. *(Detailed version of Theorem 3.10) Let $\boldsymbol{K}_0 \in \mathcal{K}$. For any $\boldsymbol{K} \in \mathcal{K}$, we assume access to the exact solution of the inner-loop problem, $\boldsymbol{L}(\boldsymbol{K})$. Consider the nested natural gradient algorithm using the exact natural gradients: $\boldsymbol{K}_{t+1} = \boldsymbol{K}_t - \tau_2 \boldsymbol{F}_{\boldsymbol{K}_t,\boldsymbol{L}(\boldsymbol{K}_t)}$. Let stepsize τ_2 satisfy*

$$\tau_2 \leq \min\left\{\frac{1}{\|\boldsymbol{G}_{\boldsymbol{K}_0,\boldsymbol{L}(\boldsymbol{K}_0)}\|}, \frac{s_2}{\varphi}\right\},$$

where

$$s_2 := \sigma_{\min}^{-1}(\boldsymbol{R}^u) s_4, \quad s_4 := \sup_{\boldsymbol{K}\in\hat{\mathcal{K}}} \|\boldsymbol{\Sigma}_{\boldsymbol{K}^*,\widetilde{\boldsymbol{L}}_{\boldsymbol{K},\boldsymbol{K}^*}}\|,$$
$$\widetilde{\boldsymbol{L}}_{\boldsymbol{K},\boldsymbol{K}'} := \boldsymbol{L}(\boldsymbol{K}) - (-\boldsymbol{R}^w + \boldsymbol{D}^\top \boldsymbol{P}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})}\boldsymbol{D})^{-1}\boldsymbol{D}^\top \boldsymbol{P}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})}\boldsymbol{B}(\boldsymbol{K}' - \boldsymbol{K}).$$

We have linear convergence as follows

$$\mathcal{G}(\boldsymbol{K}_{t+1}, \boldsymbol{L}(\boldsymbol{K}_{t+1})) - \mathcal{G}(\boldsymbol{K}^*, \boldsymbol{L}^*) \leq (1 - \frac{\varphi\tau_2}{s_2})(\mathcal{G}(\boldsymbol{K}_t, \boldsymbol{L}(\boldsymbol{K}_t)) - \mathcal{G}(\boldsymbol{K}^*, \boldsymbol{L}^*)).$$

Proof. The proof is divided into two main parts: (i) proving sufficient decrease and (ii) gradient domination respectively. Then we conclude the linear convergence rate using these two properties, which is standard in optimization.

Sufficient decrease. We firstly consider one step update $\boldsymbol{K}' = \boldsymbol{K} - \tau_2 \boldsymbol{F}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})}$ where $\boldsymbol{K} \in \mathcal{K}$. We start from the matrix difference lemma for $\boldsymbol{P}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})}$ (see Lemma B.1 in Zhang et al. [2021b] for example)

$$\begin{aligned}\boldsymbol{P}_{\boldsymbol{K}',\boldsymbol{L}(\boldsymbol{K}')} - \boldsymbol{P}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})} = \boldsymbol{A}_{\boldsymbol{K}',\boldsymbol{L}(\boldsymbol{K}')}^\top (\boldsymbol{P}_{\boldsymbol{K}',\boldsymbol{L}(\boldsymbol{K}')} - \boldsymbol{P}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})})\boldsymbol{A}_{\boldsymbol{K}',\boldsymbol{L}(\boldsymbol{K}')} + \boldsymbol{\mathcal{R}}_{\boldsymbol{K},\boldsymbol{K}'} \\ - \boldsymbol{\Xi}_{\boldsymbol{K},\boldsymbol{K}'}^\top (\boldsymbol{R}^w - \boldsymbol{D}^\top \boldsymbol{P}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})}\boldsymbol{D})^{-1}\boldsymbol{\Xi}_{\boldsymbol{K},\boldsymbol{K}'}\end{aligned}$$

where

$$\boldsymbol{\mathcal{R}}_{\boldsymbol{K},\boldsymbol{K}'} := \mathrm{diag}(\mathcal{R}_{K_0,K_0'}, \cdots, \mathcal{R}_{K_{N-1},K_{N-1}'}, \boldsymbol{0}_{m\times m}), \quad \boldsymbol{\Xi}_{\boldsymbol{K},\boldsymbol{K}'} := \begin{bmatrix} \boldsymbol{0}_{m\times nN} \\ \mathrm{diag}(\Xi_{K_0,K_0'}, \cdots, \Xi_{K_{N-1},K_{N-1}'}) \end{bmatrix}.$$

Then from Lemma A.23 and

$$\boldsymbol{\mathcal{R}}_{\boldsymbol{K},\boldsymbol{K}'} - \boldsymbol{\Xi}_{\boldsymbol{K},\boldsymbol{K}'}^\top (\boldsymbol{R}^w - \boldsymbol{D}^\top \boldsymbol{P}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})}\boldsymbol{D})^{-1}\boldsymbol{\Xi}_{\boldsymbol{K},\boldsymbol{K}'} \preceq \boldsymbol{\mathcal{R}}_{\boldsymbol{K},\boldsymbol{K}'},$$

we know $\boldsymbol{P}_{\boldsymbol{K}',\boldsymbol{L}(\boldsymbol{K}')} - \boldsymbol{P}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})}$ is upperbounded by the solution of the Lyapunov equation below

$$\boldsymbol{P}_{\boldsymbol{K}',\boldsymbol{L}(\boldsymbol{K}')} - \boldsymbol{P}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})} = \boldsymbol{A}_{\boldsymbol{K}',\boldsymbol{L}(\boldsymbol{K}')}^\top (\boldsymbol{P}_{\boldsymbol{K}',\boldsymbol{L}(\boldsymbol{K}')} - \boldsymbol{P}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})})\boldsymbol{A}_{\boldsymbol{K}',\boldsymbol{L}(\boldsymbol{K}')} + \boldsymbol{\mathcal{R}}_{\boldsymbol{K},\boldsymbol{K}'}.$$

If we compute the product of $\boldsymbol{P}_{\boldsymbol{K}',\boldsymbol{L}(\boldsymbol{K}')} - P_{K,L(K)}$ and $\boldsymbol{\Sigma}_0$, and take the trace, we have

$$\begin{aligned}\mathcal{G}(\boldsymbol{K}', \boldsymbol{L}(\boldsymbol{K}')) - \mathcal{G}(\boldsymbol{K}, \boldsymbol{L}(\boldsymbol{K})) &\leq \mathrm{Tr}(\boldsymbol{\mathcal{R}}_{\boldsymbol{K},\boldsymbol{K}'}\boldsymbol{\Sigma}_{\boldsymbol{K}',\boldsymbol{L}(\boldsymbol{K}')}) \\ &= \mathrm{Tr}((-2\tau_2 \boldsymbol{F}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})}^\top \boldsymbol{F}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})} \\ &\quad + \tau_2^2 \boldsymbol{F}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})}^\top (\boldsymbol{R}^u + \boldsymbol{B}^\top \widetilde{\boldsymbol{P}}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})}\boldsymbol{B})\boldsymbol{F}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})})\boldsymbol{\Sigma}_{\boldsymbol{K}',\boldsymbol{L}(\boldsymbol{K}')}).\end{aligned}$$

From the implicit regularization property of the nested natural gradient method (see Theorem 3.7 in Zhang et al. [2021b]), by choosing $\tau_2 \leq 1/\|\boldsymbol{G}_{\boldsymbol{K}_0,\boldsymbol{L}(\boldsymbol{K}_0)}\|$, we obtain the monotonically non-increasing sequence $(\boldsymbol{P}_{\boldsymbol{K}_t,\boldsymbol{L}(\boldsymbol{K}_t)})_t$ and lowerbounded by $\mathbf{0}$. Then since $\boldsymbol{K}_0 \in \mathcal{K}$, we can bound $\widetilde{\boldsymbol{P}}_{\boldsymbol{K}_t,\boldsymbol{L}(\boldsymbol{K}_t)}$ for any $t \geq 0$

$$\|\widetilde{\boldsymbol{P}}_{\boldsymbol{K}_t,\boldsymbol{L}(\boldsymbol{K}_t)}\| \leq \|\boldsymbol{P}_{\boldsymbol{K}_0,\boldsymbol{L}(\boldsymbol{K}_0)}\| + \|\boldsymbol{P}_{\boldsymbol{K}_0,\boldsymbol{L}(\boldsymbol{K}_0)}\|^2 \|\boldsymbol{D}\|^2 \sigma_{\min}^{-1}(\boldsymbol{R}^w - \boldsymbol{D}^\top \boldsymbol{P}_{\boldsymbol{K}_0,\boldsymbol{L}(\boldsymbol{K}_0)}\boldsymbol{D}) =: f_3.$$

Hence by choosing $\tau_2 \leq \frac{2}{\|\boldsymbol{R}^u\|+s_3\|\boldsymbol{B}\|^2}$, we have

$$\begin{aligned}
& -2\tau_2 \boldsymbol{F}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})}^\top \boldsymbol{F}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})} + \tau_2^2 \boldsymbol{F}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})}^\top (\boldsymbol{R}^u + \boldsymbol{B}^\top \widetilde{\boldsymbol{P}}_{\boldsymbol{K}',\boldsymbol{L}(\boldsymbol{K}')}\boldsymbol{B}) \boldsymbol{F}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})} \\
& \preceq -2\tau_2 \boldsymbol{F}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})}^\top \boldsymbol{F}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})} + \tau_2^2 (\|\boldsymbol{R}^u\| + s_3 \|\boldsymbol{B}\|^2) \boldsymbol{F}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})}^\top \boldsymbol{F}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})} \preceq 0.
\end{aligned}$$

Hence

$$\begin{aligned}
& \operatorname{Tr}((-2\tau_2 \boldsymbol{F}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})}^\top \boldsymbol{F}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})} + \tau_2^2 \boldsymbol{F}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})}^\top (\boldsymbol{R}^u + \boldsymbol{B}^\top \widetilde{\boldsymbol{P}}_{\boldsymbol{K}',\boldsymbol{L}(\boldsymbol{K}')}\boldsymbol{B}) \boldsymbol{F}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})}) \boldsymbol{\Sigma}_{\boldsymbol{K}',\boldsymbol{L}(\boldsymbol{K}')}) \\
& \leq \operatorname{Tr}((-2\tau_2 \boldsymbol{F}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})}^\top \boldsymbol{F}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})} + \tau_2^2 \boldsymbol{F}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})}^\top (\boldsymbol{R}^u + \boldsymbol{B}^\top \widetilde{\boldsymbol{P}}_{\boldsymbol{K}',\boldsymbol{L}(\boldsymbol{K}')}\boldsymbol{B}) \boldsymbol{F}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})}) \boldsymbol{\Sigma}_0) \\
& \leq \varphi \operatorname{Tr}(-2\tau_2 \boldsymbol{F}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})}^\top \boldsymbol{F}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})} + \tau_2^2 \boldsymbol{F}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})}^\top (\boldsymbol{R}^u + \boldsymbol{B}^\top \widetilde{\boldsymbol{P}}_{\boldsymbol{K}',\boldsymbol{L}(\boldsymbol{K}')}\boldsymbol{B}) \boldsymbol{F}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})}) \\
& \leq -\varphi\tau_2 \operatorname{Tr}(\boldsymbol{F}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})}^\top \boldsymbol{F}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})}),
\end{aligned}$$

where in the first inequality, we apply Lemma A.27, and in the second inequality, we apply Lemma A.29. In summary:

$$\mathcal{G}(\boldsymbol{K}',\boldsymbol{L}(\boldsymbol{K}')) - \mathcal{G}(\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})) \leq -\varphi\tau_2 \operatorname{Tr}(\boldsymbol{F}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})}^\top \boldsymbol{F}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})}). \tag{A.2}$$

Gradient domination. Since we know that the sequence $(\boldsymbol{P}_{\boldsymbol{K}_t,\boldsymbol{L}(\boldsymbol{K}_t)})_t$ is non-increasing in deterministic case, $\boldsymbol{K}_t \in \hat{\mathcal{K}}$ is ensured. Apply Proposition 3.9, we have

$$\mathcal{G}(\boldsymbol{K}^*,\boldsymbol{L}^*) - \mathcal{G}(\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})) \geq -s_2 Tr(\boldsymbol{F}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})}^\top \boldsymbol{F}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})}).$$

Convergence rate. Combining (A.2) and (3.7), we have

$$\begin{aligned}
& \frac{1}{s_2}(-\mathcal{G}(\boldsymbol{K}^*,\boldsymbol{L}^*) + \mathcal{G}(\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K}))) \leq \operatorname{Tr}(\boldsymbol{F}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})}^\top \boldsymbol{F}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})}) \leq \frac{1}{\varphi\tau_2}(-\mathcal{G}(\boldsymbol{K}',\boldsymbol{L}(\boldsymbol{K}')) + \mathcal{G}(\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K}))) \\
& \Leftrightarrow \mathcal{G}(\boldsymbol{K}',\boldsymbol{L}(\boldsymbol{K}')) - \mathcal{G}(\boldsymbol{K}^*,\boldsymbol{L}^*) \leq (1 - \frac{\varphi\tau_2}{s_2})(\mathcal{G}(\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})) - \mathcal{G}(\boldsymbol{K}^*,\boldsymbol{L}^*)).
\end{aligned}$$

Here to make the above inequality meaningful, we further choose $\tau_2 < \frac{s_2}{\varphi}$. Then we conclude the linear convergence rate. □

A.6 Proof of Last-iterate Convergence (Stochastic)

Theorem A.5. *(Detailed version of Theorem 3.11) Let Assumption 1.1 holds. Let $\boldsymbol{K}_0 \in \mathcal{K}$ and consider the corresponding $\hat{\mathcal{K}}$ set defined in (3.1). For any $\delta_1 \in (0,1)$, $\varepsilon_1 > 0$ and for any $\boldsymbol{K} \in \mathcal{K}$, Algorithm 1 with single-point estimation outputs $\mathbf{L}$ such that $\mathcal{G}(\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})) - \mathcal{G}(\boldsymbol{K},\boldsymbol{L}) \leq \varepsilon_1$ with probability at least $1-\delta_1$ using $T_{in}M_1 = \widetilde{\mathcal{O}}(\varepsilon_1^{-2})$ samples. Moreover for any $\delta_2 \in (0,1)$ and* *any accuracy requirement $\varepsilon \geq 0$, if the estimation parameters in Algorithm 2 satisfy*

$$\tau_2 \leq \min\left\{\frac{\lambda_{\min}(\boldsymbol{H}_{\boldsymbol{K}_0,\boldsymbol{L}(\boldsymbol{K}_0)})}{6\|\boldsymbol{D}\|^2}, 1/(8G), B_2/(\sqrt{m(N+1)}B_4), B_1/(\sqrt{m(N+1)}B_4), 1\right\},$$
$$T \geq \frac{\log(\varepsilon/2(\mathcal{G}(\boldsymbol{K}_0,\boldsymbol{L}(\boldsymbol{K}_0)) - \mathcal{G}(\boldsymbol{K}^*,\boldsymbol{L}^*)))}{\log(1-\varphi\tau_2/4s_2)} = \mathcal{O}(\log\varepsilon^{-1})$$
$$r_2 \leq \min\left\{D_1, \sqrt{\varepsilon/(2c_1)}\right\}, \quad \varepsilon_1 \leq \min\left\{D_3, \frac{\varepsilon}{2c_2}\right\}, \quad \delta_1 \leq \delta/(2T),$$
$$M_2 \geq \max\left\{M_{\boldsymbol{\Sigma}}(\varphi/2, \delta/(4T)), M_{\boldsymbol{\Sigma}}(\frac{\varphi^2}{4O_1}\cdot\sqrt{\frac{\varepsilon}{2c_3}}, \delta/(4T)), M'_V(\frac{\varphi}{4}\cdot\sqrt{\frac{\varepsilon}{2c_3}}, \delta/(4T))\right\}$$
$$= \widetilde{\mathcal{O}}(\varepsilon^{-1}),$$
$$M'_V(\varepsilon,\delta) := (\frac{O'_2}{\varepsilon})^2\cdot\log(\frac{2d_{\boldsymbol{K}}}{\delta}), \quad O'_2 := d_{\boldsymbol{K}}(N+1)\vartheta^2 l_5 + O_1,$$

where positive constant O_1 is defined in Lemma A.18, $M_{\boldsymbol{\Sigma}}(\varepsilon,\delta), M_V(\varepsilon,\delta)$ are defined in Lemma A.19 and A.18 respectively. Here G, B_1, B_2, B_4 are defined in Lemma A.6, A.8, A.7, A.17. And D_1, D_3 are defined in Lemma A.6. Constants c_1, c_2, c_3 are defined in Lemma A.20. Then, it holds with probability at least $1-\delta_2$ that $\boldsymbol{K}_t \in \hat{\mathcal{K}}$ for all $t = 1,\cdots,T$ and we have

$$\mathcal{G}(\boldsymbol{K}_T, \boldsymbol{L}(\boldsymbol{K}_T)) - \mathcal{G}(\boldsymbol{K}^*,\boldsymbol{L}^*) \leq \varepsilon,$$

with probability at least $1-\delta_2$ using a total sample complexity $\mathcal{O}(T(T_{in}M_1 + T_{out}M_2)) = \mathcal{O}(T\cdot\varepsilon^{-2}) = \widetilde{\mathcal{O}}(\varepsilon^{-2})$.

Proof. The proof is a generalization of Theorem A.4. Here we try to develop almost sufficient decrease and gradient domination properties of the objective function under the stochastic setting.

Gradient domination. In this part we can follow the same proof as Theorem A.4. We have

$$\mathcal{G}(\boldsymbol{K}^*,\boldsymbol{L}^*) - \mathcal{G}(\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})) \geq -s_2 Tr(\boldsymbol{F}^\top_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})}\boldsymbol{F}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})}) \tag{A.3}$$

where the positive constant s_2 is defined in Theorem A.4.

Sufficient decrease. We firstly consider one step update $\boldsymbol{K}' = \boldsymbol{K} - \tau_2\boldsymbol{F}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})}$ where $\boldsymbol{K} \in \mathcal{K}$. We start from the following inequality which is already proved in Theorem A.4

$$\mathcal{G}(\boldsymbol{K}',\boldsymbol{L}(\boldsymbol{K}')) - \mathcal{G}(\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})) \leq \mathrm{Tr}(\boldsymbol{\mathcal{R}}_{\boldsymbol{K},\boldsymbol{K}'}\boldsymbol{\Sigma}_{\boldsymbol{K}',\boldsymbol{L}(\boldsymbol{K}')}).$$

If we choose $\tau_2 \leq \min\{1/(8G), 1\}$ where G is defined in Lemma A.6, we have

$$
\begin{aligned}
\mathcal{G}(\boldsymbol{K}', \boldsymbol{L}(\boldsymbol{K}')) - \mathcal{G}(\boldsymbol{K}, \boldsymbol{L}(\boldsymbol{K})) &\leq \operatorname{Tr}(\boldsymbol{\mathcal{R}}_{\boldsymbol{K},\boldsymbol{K}'} \boldsymbol{\Sigma}_{\boldsymbol{K}',\boldsymbol{L}(\boldsymbol{K}')}) \\
&\leq \operatorname{Tr}(((4\tau_2 + 4\tau_2^2 \|\boldsymbol{G}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})}\|)(\boldsymbol{F}^r_{\boldsymbol{K},\boldsymbol{L}} - \boldsymbol{F}_{\boldsymbol{K},\boldsymbol{L}})^\top (\boldsymbol{F}^r_{\boldsymbol{K},\boldsymbol{L}} - \boldsymbol{F}_{\boldsymbol{K},\boldsymbol{L}}) \\
&\quad + \tau_2 (\boldsymbol{F}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})} - \boldsymbol{F}_{\boldsymbol{K},\boldsymbol{L}})^\top (\boldsymbol{F}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})} - \boldsymbol{F}_{\boldsymbol{K},\boldsymbol{L}}) \\
&\quad + (2\tau_2 + \tau_2^2 \|\boldsymbol{G}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})}\|) V(\widetilde{\boldsymbol{F}}_{\boldsymbol{K},\boldsymbol{L}}) - \frac{\tau_2}{4} \boldsymbol{F}^\top_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})} \boldsymbol{F}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})}) \boldsymbol{\Sigma}_{\boldsymbol{K}',\boldsymbol{L}(\boldsymbol{K}')}) \\
&= \operatorname{Tr}(((4\tau_2 + 4\tau_2^2 \|\boldsymbol{G}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})}\|)(\boldsymbol{F}^r_{\boldsymbol{K},\boldsymbol{L}} - \boldsymbol{F}_{\boldsymbol{K},\boldsymbol{L}})^\top (\boldsymbol{F}^r_{\boldsymbol{K},\boldsymbol{L}} - \boldsymbol{F}_{\boldsymbol{K},\boldsymbol{L}}) \\
&\quad + \tau_2 (\boldsymbol{F}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})} - \boldsymbol{F}_{\boldsymbol{K},\boldsymbol{L}})^\top (\boldsymbol{F}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})} - \boldsymbol{F}_{\boldsymbol{K},\boldsymbol{L}}) \\
&\quad + (2\tau_2 + \tau_2^2 \|\boldsymbol{G}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})}\|) V(\widetilde{\boldsymbol{F}}_{\boldsymbol{K},\boldsymbol{L}})) \boldsymbol{\Sigma}_{\boldsymbol{K}',\boldsymbol{L}(\boldsymbol{K}')}) \\
&\quad - \frac{\tau_2}{4} \operatorname{Tr}(\boldsymbol{F}^\top_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})} \boldsymbol{F}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})} \boldsymbol{\Sigma}_{\boldsymbol{K}',\boldsymbol{L}(\boldsymbol{K}')}) \\
&\leq \operatorname{Tr}(((4\tau_2 + 4\tau_2^2 \|\boldsymbol{G}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})}\|)(\boldsymbol{F}^r_{\boldsymbol{K},\boldsymbol{L}} - \boldsymbol{F}_{\boldsymbol{K},\boldsymbol{L}})^\top (\boldsymbol{F}^r_{\boldsymbol{K},\boldsymbol{L}} - \boldsymbol{F}_{\boldsymbol{K},\boldsymbol{L}}) \\
&\quad + \tau_2 (\boldsymbol{F}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})} - \boldsymbol{F}_{\boldsymbol{K},\boldsymbol{L}})^\top (\boldsymbol{F}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})} - \boldsymbol{F}_{\boldsymbol{K},\boldsymbol{L}}) \\
&\quad + (2\tau_2 + \tau_2^2 \|\boldsymbol{G}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})}\|) V(\widetilde{\boldsymbol{F}}_{\boldsymbol{K},\boldsymbol{L}})) \boldsymbol{\Sigma}_{\boldsymbol{K}',\boldsymbol{L}(\boldsymbol{K}')}) \\
&\quad - \frac{\tau_2 \varphi}{4} \operatorname{Tr}(\boldsymbol{F}^\top_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})} \boldsymbol{F}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})}),
\end{aligned}
$$

where in the second inequality, we apply the upperbound for $\boldsymbol{\mathcal{R}}_{\boldsymbol{K},\boldsymbol{K}'}$ in the proof of Lemma A.15 and in the third inequality, we apply Lemma A.29. Then we obtain the following inequality, which is a counterpart of (A.2) in stochastic setting

$$
\begin{aligned}
\mathcal{G}(\boldsymbol{K}', \boldsymbol{L}(\boldsymbol{K}')) - \mathcal{G}(\boldsymbol{K}, \boldsymbol{L}(\boldsymbol{K})) \leq &-\frac{\varphi \tau_2}{4} \operatorname{Tr}(\boldsymbol{F}^\top_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})} \boldsymbol{F}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})}) \\
&+ \operatorname{Tr}(((4\tau_2 + 4\tau_2^2 \|\boldsymbol{G}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})}\|)(\boldsymbol{F}^r_{\boldsymbol{K},\boldsymbol{L}} - \boldsymbol{F}_{\boldsymbol{K},\boldsymbol{L}})^\top (\boldsymbol{F}^r_{\boldsymbol{K},\boldsymbol{L}} - \boldsymbol{F}_{\boldsymbol{K},\boldsymbol{L}}) \\
&\quad + \tau_2 (\boldsymbol{F}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})} - \boldsymbol{F}_{\boldsymbol{K},\boldsymbol{L}})^\top (\boldsymbol{F}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})} - \boldsymbol{F}_{\boldsymbol{K},\boldsymbol{L}}) \\
&\quad + (2\tau_2 + \tau_2^2 \|\boldsymbol{G}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})}\|) V(\widetilde{\boldsymbol{F}}_{\boldsymbol{K},\boldsymbol{L}})) \boldsymbol{\Sigma}_{\boldsymbol{K}',\boldsymbol{L}(\boldsymbol{K}')}). \qquad \text{(A.4)}
\end{aligned}
$$

Convergence rate and sample complexity. Combine (A.3) and (A.4), we obtain

$$
\begin{aligned}
&\frac{1}{f_2}(-\mathcal{G}(\boldsymbol{K}^*, \boldsymbol{L}^*) + \mathcal{G}(\boldsymbol{K}, \boldsymbol{L}(\boldsymbol{K}))) \\
&\leq \operatorname{Tr}(\boldsymbol{F}^\top_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})} \boldsymbol{F}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})}) \\
&\leq \frac{4}{\varphi \tau_2}(-\mathcal{G}(\boldsymbol{K}', \boldsymbol{L}(\boldsymbol{K}')) + \mathcal{G}(\boldsymbol{K}, \boldsymbol{L}(\boldsymbol{K}))) \\
&\quad + \frac{4}{\varphi \tau_2}\Big(\operatorname{Tr}(((4\tau_2 + 4\tau_2^2 \|\boldsymbol{G}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})}\|)(\boldsymbol{F}^r_{\boldsymbol{K},\boldsymbol{L}} - \boldsymbol{F}_{\boldsymbol{K},\boldsymbol{L}})^\top (\boldsymbol{F}^r_{\boldsymbol{K},\boldsymbol{L}} - \boldsymbol{F}_{\boldsymbol{K},\boldsymbol{L}}) \\
&\quad + \tau_2 (\boldsymbol{F}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})} - \boldsymbol{F}_{\boldsymbol{K},\boldsymbol{L}})^\top (\boldsymbol{F}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})} - \boldsymbol{F}_{\boldsymbol{K},\boldsymbol{L}}) \\
&\quad + (2\tau_2 + \tau_2^2 \|\boldsymbol{G}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})}\|) V(\widetilde{\boldsymbol{F}}_{\boldsymbol{K},\boldsymbol{L}})) \boldsymbol{\Sigma}_{\boldsymbol{K}',\boldsymbol{L}(\boldsymbol{K}')})\Big) \\
\Leftrightarrow &\mathcal{G}(\boldsymbol{K}', \boldsymbol{L}(\boldsymbol{K}')) - \mathcal{G}(\boldsymbol{K}^*, \boldsymbol{L}^*) \\
&\leq (1 - \frac{\varphi \tau_2}{4 f_2})(\mathcal{G}(\boldsymbol{K}, \boldsymbol{L}(\boldsymbol{K})) - \mathcal{G}(\boldsymbol{K}^*, \boldsymbol{L}^*)) \\
&\quad + \operatorname{Tr}(((4\tau_2 + 4\tau_2^2 \|\boldsymbol{G}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})}\|)(\boldsymbol{F}^r_{\boldsymbol{K},\boldsymbol{L}} - \boldsymbol{F}_{\boldsymbol{K},\boldsymbol{L}})^\top (\boldsymbol{F}^r_{\boldsymbol{K},\boldsymbol{L}} - \boldsymbol{F}_{\boldsymbol{K},\boldsymbol{L}}) \\
&\quad + \tau_2 (\boldsymbol{F}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})} - \boldsymbol{F}_{\boldsymbol{K},\boldsymbol{L}})^\top (\boldsymbol{F}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})} - \boldsymbol{F}_{\boldsymbol{K},\boldsymbol{L}}) \\
&\quad + (2\tau_2 + \tau_2^2 \|\boldsymbol{G}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})}\|) V(\widetilde{\boldsymbol{F}}_{\boldsymbol{K},\boldsymbol{L}})) \boldsymbol{\Sigma}_{\boldsymbol{K}',\boldsymbol{L}(\boldsymbol{K}')}).
\end{aligned}
$$

To control the deviation term in the above inequality, we apply the implicit regularization property, Proposition 3.4. Then we know that by choosing $M_2 = \widetilde{\mathcal{O}}(T^2)$, $\tau_2 = \mathcal{O}(1)$, $r_2 = \mathcal{O}(T^{-1/2})$, $\varepsilon_1 = \mathcal{O}(T^{-1})$, $\delta_1 = \mathcal{O}(\delta_2/T)$, then, it holds with probability at least $1-\delta_2$ that $\boldsymbol{K}_t \in \hat{\mathcal{K}}$ for all $t = 1, \cdots, T$. Hence by controlling $\tau_2 < 4s_2/\varphi$, we have the following convergence result

$$
\begin{aligned}
\mathcal{G}(\boldsymbol{K}_T, \boldsymbol{L}(\boldsymbol{K}_T)) - \mathcal{G}(\boldsymbol{K}^*, \boldsymbol{L}^*) \leq (1 - \frac{\varphi\tau_2}{4s_2})^T (\mathcal{G}(\boldsymbol{K}_0, \boldsymbol{L}(\boldsymbol{K}_0)) - \mathcal{G}(\boldsymbol{K}^*, \boldsymbol{L}^*)) \\
+ \sum_{t=0}^{T-1} (1 - \frac{\varphi\tau_2}{4s_2})^t \tau_2 (c_1 \cdot r_2^2 + c_2 \cdot \varepsilon_1 + c_3 V) \\
\leq (1 - \frac{\varphi\tau_2}{4s_2})^T (\mathcal{G}(\boldsymbol{K}_0, \boldsymbol{L}(\boldsymbol{K}_0)) - \mathcal{G}(\boldsymbol{K}^*, \boldsymbol{L}^*)) \\
+ \frac{4s_2}{\varphi} (c_1 r_2^2 + c_2 \varepsilon_1 + c_3 V),
\end{aligned}
$$

where V denotes the variance-like term. Hence to achieve an accuracy of $\mathcal{G}(\boldsymbol{K}_T, \boldsymbol{L}(\boldsymbol{K}_T)) - \mathcal{G}(\boldsymbol{K}^*, \boldsymbol{L}^*) \leq \varepsilon$, firstly we choose

$$
T \geq \frac{\log(\varepsilon/2(\mathcal{G}(\boldsymbol{K}_0, \boldsymbol{L}(\boldsymbol{K}_0)) - \mathcal{G}(\boldsymbol{K}^*, \boldsymbol{L}^*)))}{\log(1 - \varphi\tau_2/4s_2)}
$$

such that

$$
(1 - \frac{\varphi\tau_2}{4s_2})^T (\mathcal{G}(\boldsymbol{K}_0, \boldsymbol{L}(\boldsymbol{K}_0)) - \mathcal{G}(\boldsymbol{K}^*, \boldsymbol{L}^*)) \leq \varepsilon/2
$$

Moreover, apply Lemma A.18 and choose

$$
\begin{aligned}
&\tau_2 \leq \min\left\{ \frac{\lambda_{\min}(\boldsymbol{H}_{\boldsymbol{K}_0, \boldsymbol{L}(\boldsymbol{K}_0)})}{6\|\boldsymbol{D}\|^2}, 1/(8G), B_2/(\sqrt{m(N+1)}B_4), B_1/(\sqrt{m(N+1)}B_4), 1 \right\}, \\
&r_2 \leq \min\left\{ D_1, \sqrt{\varphi\varepsilon/(8s_2c_1)} \right\}, \quad \varepsilon_1 \leq \min\left\{ D_3, \frac{\varphi\varepsilon}{8s_2c_2} \right\}, \quad \delta_1 \leq \delta/(2T), \\
&M_2 \geq \max\left\{ M_{\boldsymbol{\Sigma}}(\varphi/2, \delta/(4T)), M_{\boldsymbol{\Sigma}}(\frac{\varphi^2}{4O_1} \cdot \sqrt{\frac{\varphi\varepsilon}{8s_2c_3}}, \delta/(4T)), M_V'(\frac{\varphi}{4} \cdot \sqrt{\frac{\varphi\varepsilon}{8s_2c_3}}, \delta/(4T)) \right\} \\
&\quad = \widetilde{\mathcal{O}}(\varepsilon^{-1}),
\end{aligned}
$$

we further have

$$
\frac{4s_2}{\varphi} (c_1 r_2^2 + c_2 \varepsilon_1 + c_3 V) \leq \varepsilon/2,
$$

with probability at least $1-\delta_2$. Hence the total sample complexity is $\mathcal{O}(T(T_{in}M_1 + T_{out}M_2)) = \widetilde{\mathcal{O}}(\varepsilon^{-2})$. □

A.7 Structural Properties of Zero-sum LQ Games

This section summarizes basic results for LQ games, some of which are similar to results in Fazel et al. [2018], Zhang et al. [2019], and Zhang et al. [2021b]. Before we introduce the lemmas for zero-sum LQ games, we define the following bounds:

Lemma A.6. *(Uniform bounds over $\hat{\mathcal{K}} \times \hat{\mathcal{L}}$) Let $\boldsymbol{K}_0 \in \mathcal{K}$ and consider the following set*

$$
\begin{aligned}
\hat{\mathcal{K}} := \Big\{ \boldsymbol{K} \mid (2.1) \text{ admits a solution } \boldsymbol{P}_{\boldsymbol{K}, \boldsymbol{L}(\boldsymbol{K})} \succeq 0, \\
\text{and } \boldsymbol{P}_{\boldsymbol{K}, \boldsymbol{L}(\boldsymbol{K})} \preceq \boldsymbol{P}_{\boldsymbol{K}_0, \boldsymbol{L}(\boldsymbol{K}_0)} + \frac{\lambda_{\min}(\boldsymbol{H}_{\boldsymbol{K}_0, \boldsymbol{L}(\boldsymbol{K}_0)})}{2\|\boldsymbol{D}\|} \cdot \boldsymbol{I} \Big\}.
\end{aligned}
$$

For any $(\boldsymbol{K}, \boldsymbol{L}) \in \hat{\mathcal{K}} \times \hat{\mathcal{L}}$, *the following amounts are positive and well-defined.*

$$
\begin{aligned}
D_1 &:= \inf_{(\boldsymbol{K},\boldsymbol{L})\in\hat{\mathcal{K}}\times\hat{\mathcal{L}}} \min\left\{1, \frac{\varphi}{2\operatorname{Tr}(\boldsymbol{\Sigma}_{\boldsymbol{K},\boldsymbol{L}})\|\boldsymbol{B}\|(2\|\boldsymbol{A}_{\boldsymbol{K},\boldsymbol{L}}\| + \|\boldsymbol{B}\|)}\right\}, \\
D_2 &:= \inf_{(\boldsymbol{K},\boldsymbol{L})\in\hat{\mathcal{K}}\times\hat{\mathcal{L}}} \min\left\{1, \frac{\varphi}{2\operatorname{Tr}(\boldsymbol{\Sigma}_{\boldsymbol{K},\boldsymbol{L}})\|\boldsymbol{D}\|(2\|\boldsymbol{A}_{\boldsymbol{K},\boldsymbol{L}}\| + \|\boldsymbol{D}\|)}\right\}, \\
D_3 &:= \inf_{(\boldsymbol{K},\boldsymbol{L})\in\hat{\mathcal{K}}\times\hat{\mathcal{L}}} \min\left\{1, H^{-1}, H^{-1}\left(\frac{\varphi}{2\operatorname{Tr}(\boldsymbol{\Sigma}_{\boldsymbol{K},\boldsymbol{L}})\|\boldsymbol{D}\|(2\|\boldsymbol{A}_{\boldsymbol{K},\boldsymbol{L}}\| + \|\boldsymbol{D}\|)}\right)^2\right\}, \\
G &:= \sup_{\boldsymbol{K}\in\hat{\mathcal{K}}} \|\boldsymbol{G}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})}\|,
\end{aligned}
$$

where $\boldsymbol{G}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})}$ *and* H *are defined in Appendix A.1. These uniform bounds will contribute to the proof of IR property.*

Proof. The proof is immediate from the boundedness of $\hat{\mathcal{K}}$, $\hat{\mathcal{L}}$ as well as Lemma A.16 since

$$
\begin{aligned}
&\operatorname{Tr}(\boldsymbol{\Sigma}_{\boldsymbol{K},\boldsymbol{L}}) \le m(N+1)\|\boldsymbol{\Sigma}_{\boldsymbol{K},\boldsymbol{L}}\| \le m(N+1)\sum_{i=0}^{N}(\|\boldsymbol{A}_{\boldsymbol{K},\boldsymbol{L}}\|)^{2i}\|\boldsymbol{\Sigma}_0\|, \\
&\|\boldsymbol{A}_{\boldsymbol{K},\boldsymbol{L}}\| \le \|\boldsymbol{A}\| + \|\boldsymbol{B}\|\|\boldsymbol{K}\| + \|\boldsymbol{D}\|\|\boldsymbol{L}\|.
\end{aligned}
$$

□

Lemma A.7. *(Lemma B.9 in Zhang et al. [2021b]) For any* $\boldsymbol{K} \in \mathcal{K}$, *there exists some* $\mathcal{B}_{2,\boldsymbol{K}} > 0$ *such that all* $\boldsymbol{K}'$ *satisfying* $\|\boldsymbol{K}' - \boldsymbol{K}\|_F \le \mathcal{B}_{2,\boldsymbol{K}}$ *satisfy* $\boldsymbol{K}' \in \mathcal{K}$. *Then for any* $\boldsymbol{K} \in \hat{\mathcal{K}}$ *with* $\boldsymbol{K}_0 \in \mathcal{K}$, *there exists a positive constant* B_2 *such that all* $\boldsymbol{K}'$ *satisfying* $\|\boldsymbol{K}' - \boldsymbol{K}\|_F \le B_2$ *satisfy* $\boldsymbol{K}' \in \mathcal{K}$.

Lemma A.8. *(Lemma B.7 in Zhang et al. [2021b]: Local Lipschitz continuity of* $\boldsymbol{\Sigma}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})}$, $\boldsymbol{L}(\boldsymbol{K})$, $\boldsymbol{P}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})}$*) For any* $\boldsymbol{K}, \boldsymbol{K}' \in \mathcal{K}$, *there exist some* $\mathcal{B}_{1,\boldsymbol{K}}$, $\mathcal{B}_{\boldsymbol{P},\boldsymbol{K}}$, $\mathcal{B}_{\boldsymbol{L}(\boldsymbol{K}),\boldsymbol{K}}$, $\mathcal{B}_{\boldsymbol{\Sigma},\boldsymbol{K}} > 0$ *that are continuous functions of* $\boldsymbol{K}$ *such that all* $\boldsymbol{K}'$ *satisfying* $\|\boldsymbol{K}' - \boldsymbol{K}\|_F \le \mathcal{B}_{1,\boldsymbol{K}}$ *satisfy*

$$
\begin{aligned}
\|\boldsymbol{P}_{\boldsymbol{K}',\boldsymbol{L}(\boldsymbol{K}')} - \boldsymbol{P}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})}\|_F &\le \mathcal{B}_{\boldsymbol{P},\boldsymbol{K}} \cdot \|\boldsymbol{K}' - \boldsymbol{K}\|_F, \\
\|\boldsymbol{\Sigma}_{\boldsymbol{K}',\boldsymbol{L}(\boldsymbol{K}')} - \boldsymbol{\Sigma}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})}\|_F &\le \mathcal{B}_{\boldsymbol{\Sigma},\boldsymbol{K}} \cdot \|\boldsymbol{K}' - \boldsymbol{K}\|_F, \\
\|\boldsymbol{L}(\boldsymbol{K}') - \boldsymbol{L}(\boldsymbol{K})\|_F &\le \mathcal{B}_{\boldsymbol{L}(\boldsymbol{K}),\boldsymbol{K}} \cdot \|\boldsymbol{K}' - \boldsymbol{K}\|_F.
\end{aligned}
$$

Especially, we define the following positive constants

$$
B_1 := \inf_{\boldsymbol{K}\in\hat{\mathcal{K}}} \mathcal{B}_{1,\boldsymbol{K}}, \quad B_{\boldsymbol{P}} := \sup_{\boldsymbol{K}\in\hat{\mathcal{K}}} \mathcal{B}_{\boldsymbol{P},\boldsymbol{K}}, \quad B_{\boldsymbol{\Sigma}} := \sup_{\boldsymbol{K}\in\hat{\mathcal{K}}} \mathcal{B}_{\boldsymbol{\Sigma},\boldsymbol{K}}, \quad B_{\boldsymbol{L}(\boldsymbol{K})} := \sup_{\boldsymbol{K}\in\hat{\mathcal{K}}} \mathcal{B}_{\boldsymbol{L}(\boldsymbol{K}),\boldsymbol{K}}.
$$

Even though four constants are defined above, only $B_1, B_{\boldsymbol{\Sigma}}$ *are used in our proof.*

Lemma A.9. *(Local Lipschitz continuity of* $P_{K,L}$*) Let* $\boldsymbol{K}_0 \in \mathcal{K}$ *and consider the following set*

$$
\begin{aligned}
\hat{\mathcal{K}} := \Bigg\{\boldsymbol{K} \mid (2.1) \text{ admits a solution } \boldsymbol{P}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})} \succeq 0, \\
\text{and } \boldsymbol{P}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})} \preceq \boldsymbol{P}_{\boldsymbol{K}_0,\boldsymbol{L}(\boldsymbol{K}_0)} + \frac{\lambda_{\min}(\boldsymbol{H}_{\boldsymbol{K}_0,\boldsymbol{L}(\boldsymbol{K}_0)})}{2\|\boldsymbol{D}\|} \cdot \boldsymbol{I}\Bigg\}.
\end{aligned}
$$

For any $(\boldsymbol{K}, \boldsymbol{L}) \in \hat{\mathcal{K}} \times \hat{\mathcal{L}}$ *with structures defined in (1.4) and* $(\boldsymbol{K}' \in \mathcal{K}, \boldsymbol{L}')$ *that satisfy*

$$
\|\boldsymbol{K}' - \boldsymbol{K}\| \le D_1, \quad \|\boldsymbol{L}' - \boldsymbol{L}\| \le D_2,
$$

where D_1, D_2 *are defined in Lemma A.6. Then there exist positive constants* l_5, l_6 *such that*

$$
\|\boldsymbol{P}_{\boldsymbol{K}',\boldsymbol{L}} - \boldsymbol{P}_{\boldsymbol{K},\boldsymbol{L}}\| \le l_5\|\boldsymbol{K}' - \boldsymbol{K}\|, \quad \|\boldsymbol{P}_{\boldsymbol{K},\boldsymbol{L}'} - \boldsymbol{P}_{\boldsymbol{K},\boldsymbol{L}}\| \le l_6\|\boldsymbol{L}' - \boldsymbol{L}\|.
$$

Proof. For the simplicity of the proof, we use $\Delta_K := K' - K$ and $\Delta_L := L' - L$ hereafter. We apply Lemma A.22 and sensitivity analysis in Lemma A.26 with $F = A_{K',L}$, $M = Q + (K')^\top R^u (K') - L^\top R^v L$, H is the solution of the following Lyapunov equation

$$I + A_{K,L}^\top H A_{K,L} = H.$$

and $X = P_{K',L}$. Then if let $\|H\|\|B\Delta_K\|(2\|A_{K,L}\| + \|B\Delta_K\|) < 1$, we have

$$\begin{aligned}\|P_{K',L} - P_{K,L}\| \leq\ & (1 - \|H\|\|B\Delta_K\|(2\|A_{K,L}\| + \|B\Delta_K\|))^{-1}\|H\| \\ & \cdot (\|(K')^\top R^u \Delta_K + \Delta_K^\top R^u K\| + \|B\Delta_K\|(2\|A_{K,L}\| + \|B\Delta_K\|)\|P_{K,L}\|) \\ \leq\ & (1 - \|H\|\|B\Delta_K\|(2\|A_{K,L}\| + \|B\Delta_K\|))^{-1}\|H\| \cdot ((\|K\| + \|\Delta_K\|)\|R^u\| \\ & + \|R^u\|\|K\| + \|B\|(2\|A_{K,L}\| + \|B\|\|\Delta_K\|)\|P_{K,L}\|) \cdot \|\Delta_K\|.\end{aligned}$$

When we apply Lemma A.25 to bound $\|H\|$ and choose K' such that $\|\Delta_K\| \leq D_1$, then

$$\|H\|\|B\Delta_K\|(2\|A_{K,L}\| + \|B\Delta_K\|) \leq \|H\|\|B\|\|\Delta_K\|(2\|A_{K,L}\| + \|B\|\|\Delta_K\|) < 1.$$

Then we can ensure that

$$\begin{aligned}\|P_{K',L} - P_{K,L}\| \leq\ & 2\,\mathrm{Tr}(\Sigma_{K,L})/\varphi \cdot ((\|K\| + 1)\|R^u\| \\ & + \|R^u\|\|K\| + \|B\|(2\|A_{K,L}\| + \|B\|)\|P_{K,L}\|) \cdot \|\Delta_K\|.\end{aligned}$$

For L' and L, the proof is similar and hence omitted

$$\begin{aligned}\|P_{K,L'} - P_{K,L}\| \leq\ & 2\,\mathrm{Tr}(\Sigma_{K,L})/\varphi \cdot ((\|L\| + 1)\|R^v\| \\ & + \|R^v\|\|L\| + \|D\|(2\|A_{K,L}\| + \|D\|)\|P_{K,L}\|) \cdot \|\Delta_L\|,\end{aligned}$$

when $\|\Delta_L\| \leq D_2$. Since $\|K\|, \|L\|, \|P_{K,L}\|, \|\Sigma_{K,L}\|$ are bounded over $\hat{\mathcal{K}} \times \hat{\mathcal{L}}$, we can define positive constants l_5 and l_6 as follows

$$\begin{aligned} l_5 &:= \sup_{(K,L)\in\hat{\mathcal{K}}\times\hat{\mathcal{L}}} 2\,\mathrm{Tr}(\Sigma_{K,L})/\varphi \cdot ((\|K\| + 1)\|R^u\| + \|R^u\|\|K\| + \|B\|(2\|A_{K,L}\| + \|B\|)\|P_{K,L}\|), \\ l_6 &:= \sup_{(K,L)\in\hat{\mathcal{K}}\times\hat{\mathcal{L}}} 2\,\mathrm{Tr}(\Sigma_{K,L})/\varphi \cdot ((\|L\| + 1)\|R^v\| + \|R^v\|\|L\| + \|D\|(2\|A_{K,L}\| + \|D\|)\|P_{K,L}\|). \end{aligned}$$

$\square$

Lemma A.10. *(Lipschitz continuity of $E_{K,L}, F_{K,L}$) Let $K_0 \in \mathcal{K}$ and consider the following set*

$$\hat{\mathcal{K}} := \Big\{ K \mid (2.1) \textit{ admits a solution } P_{K,L(K)} \succeq 0, \\ \textit{and } P_{K,L(K)} \preceq P_{K_0,L(K_0)} + \frac{\lambda_{\min}(H_{K_0,L(K_0)})}{2\|D\|} \cdot I \Big\}.$$

For any $(K, L) \in \hat{\mathcal{K}} \times \hat{\mathcal{L}}$ and $(K' \in \mathcal{K}, L')$ with structures defined in (1.4), (1.5) that satisfy

$$\|K' - K\| \leq D_1, \quad \|L' - L\| \leq D_2,$$

where D_1, D_2 are positive constants defined in Lemma A.6. There exist positive constants l_1, l_2, l_3, and l_4 such that

$$\begin{aligned} \|F_{K',L} - F_{K,L}\| &\leq l_1\|K' - K\|, \quad \|F_{K,L'} - F_{K,L}\| \leq l_2\|L' - L\|, \\ \|E_{K',L} - E_{K,L}\| &\leq l_3\|K' - K\|, \quad \|E_{K,L'} - E_{K,L}\| \leq l_4\|L' - L\|. \end{aligned}$$

Proof. For the simplicity of the proof, we use $\Delta_K := K' - K$ and $\Delta_L := L' - L$. We start from the explicit expression of $E_{K,L}$ and $F_{K,L}$

$$\begin{aligned}\|F_{K',L} - F_{K,L}\| &= \|B^\top(P_{K',L} - P_{K,L})(BK' + DL - A) + (R^u + B^\top P_{K,L}B)(K' - K)\| \\ &\leq \left(\|B\| l_5(\|A_{K,L}\| + \|B\|) + \|R^u\| + \|B\|^2\|P_{K,L}\|\right) \cdot \|K' - K\| \\ &= l_1\|K' - K\|.\end{aligned}$$

For the first inequality, we apply Lemma A.9. Then we require $\|\Delta_K\| \leq D_1$, and apply Lemma A.6. Following the same spirit of proving, we have

$$\begin{aligned}\|F_{K,L'} - F_{K,L}\| &= \| - B^\top(P_{K,L'} - P_{K,L})A_{K,L'} + B^\top P_{K,L}D(L' - L)\| \\ &\leq (l_6\|B\|(\|A_{K,L}\| + \|D\|) + \|B\|\|P_{K,L}\|\|D\|) \cdot \|L' - L\|, \\ \|E_{K',L} - E_{K,L}\| &= -D^\top(P_{K',L} - P_{K,L})A_{K',L} + D^\top P_{K,L}B(K' - K) \\ &\leq (l_5\|D\|(\|A_{K,L}\| + \|B\|) + \|D\|\|P_{K,L}\|\|B\|) \cdot \|K' - K\|, \\ \|E_{K,L'} - E_{K,L}\| &= \| - D^\top(P_{K,L'} - P_{K,L})A_{K,L'} + (-R^w + D^\top P_{K,L}D)(L' - L)\| \\ &\leq \left(l_6\|D\|(\|A_{K,L}\| + \|D\|) + \|R^w\| + \|D\|^2\|P_{K,L}\|\right) \cdot \|L' - L\|,\end{aligned}$$

when we also require $\|\Delta_L\| \leq D_2$. The constant coefficients are defined as

$$\begin{aligned}l_1 &:= \sup_{(K,L)\in\hat{\mathcal{K}}\times\hat{\mathcal{L}}} l_5\|B\|(\|A_{K,L}\| + \|B\|) + \|R^u\| + \|B\|^2\|P_{K,L}\| \\ l_2 &:= \sup_{(K,L)\in\hat{\mathcal{K}}\times\hat{\mathcal{L}}} l_6\|B\|(\|A_{K,L}\| + \|D\|) + \|B\|\|P_{K,L}\|\|D\| \\ l_3 &:= \sup_{(K,L)\in\hat{\mathcal{K}}\times\hat{\mathcal{L}}} l_5\|D\|(\|A_{K,L}\| + \|B\|) + \|D\|\|P_{K,L}\|\|B\| \\ l_4 &:= \sup_{(K,L)\in\hat{\mathcal{K}}\times\hat{\mathcal{L}}} l_6\|D\|(\|A_{K,L}\| + \|D\|) + \|R^w\| + \|D\|^2\|P_{K,L}\|\end{aligned}$$

Bounds are well-defined since $\|K\|, \|L\|, \|P_{K,L}\|, \|\Sigma_{K,L}\|$ are bounded over $\hat{\mathcal{K}} \times \hat{\mathcal{L}}$. $\square$

Lemma A.11. *(Local Lipschitz continuity of $\Sigma_{K,L}$) Let $K_0 \in \mathcal{K}$ and consider the following set*

$$\hat{\mathcal{K}} := \Big\{ K \mid (2.1) \text{ admits a solution } P_{K,L(K)} \succeq 0, \\ \text{and } P_{K,L(K)} \preceq P_{K_0,L(K_0)} + \frac{\lambda_{\min}(H_{K_0,L(K_0)})}{2\|D\|} \cdot I \Big\}.$$

For any $(K, L) \in \hat{\mathcal{K}} \times \hat{\mathcal{L}}$ and $(K' \in \mathcal{K}, L')$ with structures defined in (1.4), (1.5) that satisfy

$$\|K' - K\| \leq D_1, \quad \|L' - L\| \leq D_2,$$

where D_1, D_2 are defined in Lemma A.6. Then there exist positive constants l_7, l_8 such that

$$\|\Sigma_{K',L} - \Sigma_{K,L}\| \leq l_7\|K' - K\|, \quad \|\Sigma_{K,L'} - \Sigma_{K,L}\| \leq l_8\|L' - L\|.$$

Proof. For the simplicity of notations, we use $\Delta_K := K' - K$ and $\Delta_L := L' - L$. Here we apply Lemma A.22 and Lemma A.26 with $F = A_{K,L'}^\top$, $M = \Sigma_0$, $X = \Sigma_{K,L}$, and H is the solution of the Lyapunov equation below

$$I + A_{K,L}HA_{K,L}^\top = H.$$

Then if let $\|\boldsymbol{H}\|\|\boldsymbol{B}\Delta_{\boldsymbol{K}}\|(2\|\boldsymbol{A}_{\boldsymbol{K},\boldsymbol{L}}\|+2\|\boldsymbol{B}\Delta_{\boldsymbol{K}}\|)<1$, we have

$$
\begin{aligned}
\|\boldsymbol{\Sigma}_{\boldsymbol{K}',\boldsymbol{L}}-\boldsymbol{\Sigma}_{\boldsymbol{K},\boldsymbol{L}}\| \leq (1-\|\boldsymbol{H}\|\|\boldsymbol{B}\Delta_{\boldsymbol{K}}\|(2\|\boldsymbol{\Sigma}_{\boldsymbol{K},\boldsymbol{L}}\|+\|\boldsymbol{B}\Delta_{\boldsymbol{K}}\|))^{-1} \\
\cdot \|\boldsymbol{H}\|\|\boldsymbol{B}\Delta_{\boldsymbol{K}}\|(2\|\boldsymbol{\Sigma}_{\boldsymbol{K},\boldsymbol{L}}\|+\|\boldsymbol{B}\Delta_{\boldsymbol{K}}\|)\|\boldsymbol{\Sigma}_{\boldsymbol{K},\boldsymbol{L}}\|.
\end{aligned}
$$

Hence when we choose $\boldsymbol{K}'$ such that $\|\Delta_{\boldsymbol{K}}\|\leq D_1$, then

$$
\|\boldsymbol{H}\|\|\boldsymbol{B}\Delta_{\boldsymbol{K}}\|(2\|\boldsymbol{\Sigma}_{\boldsymbol{K},\boldsymbol{L}}\|+2\|\boldsymbol{B}\Delta_{\boldsymbol{K}}\|)\leq\|\boldsymbol{H}\|\|\boldsymbol{B}\|\|\Delta_{\boldsymbol{K}}\|(2\|\boldsymbol{\Sigma}_{\boldsymbol{K},\boldsymbol{L}}\|+2\|\boldsymbol{B}\|\|\Delta_{\boldsymbol{K}}\|)<1.
$$

Again, we apply Lemma A.25 to guarantee the above requirement for $\Delta_{\boldsymbol{K}}$. We have

$$
\|\boldsymbol{\Sigma}_{\boldsymbol{K}',\boldsymbol{L}}-\boldsymbol{\Sigma}_{\boldsymbol{K},\boldsymbol{L}}\|\leq 2\operatorname{Tr}(\boldsymbol{\Sigma}_{\boldsymbol{K},\boldsymbol{L}})/\varphi\cdot\|\boldsymbol{B}\|(2\|\boldsymbol{\Sigma}_{\boldsymbol{K},\boldsymbol{L}}\|+\|\boldsymbol{B}\|)\|\boldsymbol{\Sigma}_{\boldsymbol{K},\boldsymbol{L}}\|\|\Delta_{\boldsymbol{K}}\|.
$$

Similarly, we can prove that by choosing $\|\Delta_{\boldsymbol{L}}\|\leq D_2$,, we have

$$
\|\boldsymbol{\Sigma}_{\boldsymbol{K},\boldsymbol{L}'}-\boldsymbol{\Sigma}_{\boldsymbol{K},\boldsymbol{L}}\|\leq 2\operatorname{Tr}(\boldsymbol{\Sigma}_{\boldsymbol{K},\boldsymbol{L}})/\varphi\cdot\|\boldsymbol{D}\|(2\|\boldsymbol{\Sigma}_{\boldsymbol{K},\boldsymbol{L}}\|+\|\boldsymbol{D}\|)\|\boldsymbol{\Sigma}_{\boldsymbol{K},\boldsymbol{L}}\|\|\Delta_{\boldsymbol{L}}\|.
$$

The positive constant coefficients are defined as

$$
\begin{aligned}
l_7 &:= \sup_{(\boldsymbol{K},\boldsymbol{L})\in\hat{\mathcal{K}}\times\hat{\mathcal{L}}} 2\operatorname{Tr}(\boldsymbol{\Sigma}_{\boldsymbol{K},\boldsymbol{L}})/\varphi\cdot\|\boldsymbol{B}\|(2\|\boldsymbol{\Sigma}_{\boldsymbol{K},\boldsymbol{L}}\|+\|\boldsymbol{B}\|)\|\boldsymbol{\Sigma}_{\boldsymbol{K},\boldsymbol{L}}\|,\\
l_8 &:= \sup_{(\boldsymbol{K},\boldsymbol{L})\in\hat{\mathcal{K}}\times\hat{\mathcal{L}}} 2\operatorname{Tr}(\boldsymbol{\Sigma}_{\boldsymbol{K},\boldsymbol{L}})/\varphi\cdot\|\boldsymbol{D}\|(2\|\boldsymbol{\Sigma}_{\boldsymbol{K},\boldsymbol{L}}\|+\|\boldsymbol{D}\|)\|\boldsymbol{\Sigma}_{\boldsymbol{K},\boldsymbol{L}}\|.
\end{aligned}
$$

Coefficients are well-defined since $\|\boldsymbol{K}\|,\|\boldsymbol{L}\|,\|\boldsymbol{P}_{\boldsymbol{K},\boldsymbol{L}}\|,\|\boldsymbol{\Sigma}_{\boldsymbol{K},\boldsymbol{L}}\|$ are bounded over $\hat{\mathcal{K}}\times\hat{\mathcal{L}}$. $\square$

Lemma A.12. *(Local Lipschitz continuity of $\nabla_{\boldsymbol{K}}\mathcal{G}(\boldsymbol{K},\boldsymbol{L}),\nabla_{\boldsymbol{L}}\mathcal{G}(\boldsymbol{K},\boldsymbol{L})$) Let $\boldsymbol{K}_0\in\mathcal{K}$ and consider the following set*

$$
\begin{aligned}
\hat{\mathcal{K}} := \Big\{\boldsymbol{K}\mid (2.1)\ \textit{admits a solution}\ \boldsymbol{P}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})}\succeq 0,\\
\textit{and}\ \boldsymbol{P}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})}\preceq\boldsymbol{P}_{\boldsymbol{K}_0,\boldsymbol{L}(\boldsymbol{K}_0)}+\frac{\lambda_{\min}(\boldsymbol{H}_{\boldsymbol{K}_0,\boldsymbol{L}(\boldsymbol{K}_0)})}{2\|\boldsymbol{D}\|}\cdot\boldsymbol{I}\Big\}.
\end{aligned}
$$

For any $(\boldsymbol{K},\boldsymbol{L})\in\hat{\mathcal{K}}\times\hat{\mathcal{L}}$ and $(\boldsymbol{K}'\in\mathcal{K},\boldsymbol{L}')$ with structures defined in (1.4), (1.5) that satisfy

$$
\|\boldsymbol{K}'-\boldsymbol{K}\|\leq D_1,\quad \|\boldsymbol{L}'-\boldsymbol{L}\|\leq D_2,
$$

where D_1,D_2 are defined in Lemma A.6. Then the following inequalities hold

$$
\begin{aligned}
\|\nabla_{\boldsymbol{K}}\mathcal{G}(\boldsymbol{K}',\boldsymbol{L})-\nabla_{\boldsymbol{K}}\mathcal{G}(\boldsymbol{K},\boldsymbol{L})\| &\leq(2l_1\|\boldsymbol{\Sigma}_{\boldsymbol{K},\boldsymbol{L}}\|+2l_1\cdot l_7+2l_8\|\boldsymbol{F}_{\boldsymbol{K},\boldsymbol{L}}\|)\|\boldsymbol{K}'-\boldsymbol{K}\|\\
\|\nabla_{\boldsymbol{K}}\mathcal{G}(\boldsymbol{K},\boldsymbol{L}')-\nabla_{\boldsymbol{K}}\mathcal{G}(\boldsymbol{K},\boldsymbol{L})\| &\leq(2l_2\|\boldsymbol{\Sigma}_{\boldsymbol{K},\boldsymbol{L}}\|+2l_2\cdot l_8+2l_8\|\boldsymbol{F}_{\boldsymbol{K},\boldsymbol{L}}\|)\|\boldsymbol{L}'-\boldsymbol{L}\|\\
\|\nabla_{\boldsymbol{L}}\mathcal{G}(\boldsymbol{K}',\boldsymbol{L})-\nabla_{\boldsymbol{L}}\mathcal{G}(\boldsymbol{K},\boldsymbol{L})\| &\leq(2l_3\|\boldsymbol{\Sigma}_{\boldsymbol{K},\boldsymbol{L}}\|+2l_3\cdot l_7+2l_7\|\boldsymbol{E}_{\boldsymbol{K},\boldsymbol{L}}\|)\|\boldsymbol{K}'-\boldsymbol{K}\|\\
\|\nabla_{\boldsymbol{L}}\mathcal{G}(\boldsymbol{K},\boldsymbol{L}')-\nabla_{\boldsymbol{L}}\mathcal{G}(\boldsymbol{K},\boldsymbol{L}')\| &\leq(2l_4\|\boldsymbol{\Sigma}_{\boldsymbol{K},\boldsymbol{L}}\|+2l_4\cdot l_8+2l_8\|\boldsymbol{E}_{\boldsymbol{K},\boldsymbol{L}}\|)\|\boldsymbol{L}'-\boldsymbol{L}\|,
\end{aligned}
$$

where l_7,l_8 are defined in Lemma A.11. And l_1, l_2, l_3, and l_4 are defined in Lemma A.10. This lemma is important for controlling the estimation bias caused by using ZO estimation.

Proof. We use the explicit expressions of $\nabla_{\boldsymbol{K}}\mathcal{G}(\boldsymbol{K},\boldsymbol{L})$ and $\nabla_{\boldsymbol{L}}\mathcal{G}(\boldsymbol{K},\boldsymbol{L})$

$$
\begin{aligned}
\|\nabla_{\boldsymbol{K}}\mathcal{G}(\boldsymbol{K}',\boldsymbol{L})-\nabla_{\boldsymbol{K}}\mathcal{G}(\boldsymbol{K},\boldsymbol{L})\| &= \|2\boldsymbol{F}_{\boldsymbol{K}',\boldsymbol{L}}\boldsymbol{\Sigma}_{\boldsymbol{K}',\boldsymbol{L}}-2\boldsymbol{F}_{\boldsymbol{K},\boldsymbol{L}}\boldsymbol{\Sigma}_{\boldsymbol{K},\boldsymbol{L}}\|\\
&= \|2\boldsymbol{F}_{\boldsymbol{K}',\boldsymbol{L}}\boldsymbol{\Sigma}_{\boldsymbol{K}',\boldsymbol{L}}-2\boldsymbol{F}_{\boldsymbol{K},\boldsymbol{L}}\boldsymbol{\Sigma}_{\boldsymbol{K}',\boldsymbol{L}}+2\boldsymbol{F}_{\boldsymbol{K},\boldsymbol{L}}\boldsymbol{\Sigma}_{\boldsymbol{K}',\boldsymbol{L}}-2\boldsymbol{F}_{\boldsymbol{K},\boldsymbol{L}}\boldsymbol{\Sigma}_{\boldsymbol{K},\boldsymbol{L}}\|\\
&\leq 2\|\boldsymbol{F}_{\boldsymbol{K}',\boldsymbol{L}}-\boldsymbol{F}_{\boldsymbol{K},\boldsymbol{L}}\|\|\boldsymbol{\Sigma}_{\boldsymbol{K}',\boldsymbol{L}}\|+2\|\boldsymbol{F}_{\boldsymbol{K},\boldsymbol{L}}\|\|\boldsymbol{\Sigma}_{\boldsymbol{K}',\boldsymbol{L}}-\boldsymbol{\Sigma}_{\boldsymbol{K},\boldsymbol{L}}\|\\
&\leq (2l_1\|\boldsymbol{\Sigma}_{\boldsymbol{K}',\boldsymbol{L}}\|+2l_7\|\boldsymbol{F}_{\boldsymbol{K},\boldsymbol{L}}\|)\|\boldsymbol{K}'-\boldsymbol{K}\|\\
&\leq (2l_1\|\boldsymbol{\Sigma}_{\boldsymbol{K},\boldsymbol{L}}\|+2l_1\cdot l_7+2l_7\|\boldsymbol{F}_{\boldsymbol{K},\boldsymbol{L}}\|)\cdot\|\boldsymbol{K}'-\boldsymbol{K}\|,
\end{aligned}
$$

where in the last inequality we use the fact that $D_1 \leq 1$. In the second inequality, we apply Lemma A.10, A.11 and require

$$\|\boldsymbol{K}' - \boldsymbol{K}\| \leq D_1, \quad \|\boldsymbol{L}' - \boldsymbol{L}\| \leq D_2,$$

where D_1, D_2 are defined in Lemma A.6. The rest inequalities can be obtained similarly and hence omitted. $\square$

Lemma A.13. *(Bound for Natural Gradients) Let $\boldsymbol{K}_0 \in \mathcal{K}$ and consider the following set*

$$\hat{\mathcal{K}} := \Big\{ \boldsymbol{K} \mid (2.1) \text{ admits a solution } \boldsymbol{P}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})} \succeq 0, \\ \text{and } \boldsymbol{P}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})} \preceq \boldsymbol{P}_{\boldsymbol{K}_0,\boldsymbol{L}(\boldsymbol{K}_0)} + \frac{\lambda_{\min}(\boldsymbol{H}_{\boldsymbol{K}_0,\boldsymbol{L}(\boldsymbol{K}_0)})}{2\|\boldsymbol{D}\|} \cdot \boldsymbol{I} \Big\}.$$

For $\boldsymbol{K} \in \hat{\mathcal{K}}$ and $\boldsymbol{L}$ be the output of Algorithm 1 given $\boldsymbol{K}$ with structures defined in (1.4). If we choose $\varepsilon_1 \leq D_3$, there exists a positive constant B_3 such that $\|\boldsymbol{F}_{\boldsymbol{K},\boldsymbol{L}}\| \leq B_3$ with probability at least $1 - \delta_1$.

Proof. Apply Lemma A.16

$$\begin{aligned}
\|\boldsymbol{F}_{\boldsymbol{K},\boldsymbol{L}}\| &= \|(\boldsymbol{R}^u + \boldsymbol{B}^\top \boldsymbol{P}_{\boldsymbol{K},\boldsymbol{L}} \boldsymbol{B})\boldsymbol{K} - \boldsymbol{B}^\top \boldsymbol{P}_{\boldsymbol{K},\boldsymbol{L}}(\boldsymbol{A} - \boldsymbol{D}\boldsymbol{L})\| \\
&\leq (\|\boldsymbol{R}^u\| + \|\boldsymbol{B}\|^2(\|\boldsymbol{P}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})}\| + 1/\varphi))\|\boldsymbol{K}\| + \|\boldsymbol{B}\|(\|\boldsymbol{P}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})}\| + 1/\varphi) \\
&\quad \cdot (\|\boldsymbol{A}\| + \|\boldsymbol{D}\|\|\boldsymbol{L}(\boldsymbol{K})\| + \|\boldsymbol{D}\|), \\
B_3 &:= \sup_{\boldsymbol{K} \in \hat{\mathcal{K}}} (\|\boldsymbol{R}^u\| + \|\boldsymbol{B}\|^2(\|\boldsymbol{P}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})}\| + 1/\varphi))\|\boldsymbol{K}\| + \|\boldsymbol{B}\|(\|\boldsymbol{P}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})}\| + 1/\varphi) \\
&\quad \cdot (\|\boldsymbol{A}\| + \|\boldsymbol{D}\|\|\boldsymbol{L}(\boldsymbol{K})\| + \|\boldsymbol{D}\|).
\end{aligned}$$

The bound B_3 over $\hat{\mathcal{K}}$ is well-defined because of the compactness of $\hat{\mathcal{K}}$. $\square$

In the next lemma, we essentially discuss a different way to upperbound $\boldsymbol{P}_{\boldsymbol{K}',\boldsymbol{L}(\boldsymbol{K}')} - \boldsymbol{P}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})}$ from Zhang et al. [2021b] by developing a careful upperbound of $\mathcal{R}_{K_h,K_h'}$. Here $\mathcal{R}_{K_h,K_h'}$ is a term in the difference between $\boldsymbol{P}_{\boldsymbol{K}',\boldsymbol{L}(\boldsymbol{K}')} - \boldsymbol{P}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})}$, which is crucial to upperbound the difference and hence keep $\boldsymbol{K}'$ in $\hat{\mathcal{K}}$.

Lemma A.14. *(An upperbound for $\mathcal{R}_{K_h,K_h'}$) Let $\boldsymbol{K}_0 \in \mathcal{K}$. Assume $\boldsymbol{K} \in \hat{\mathcal{K}}$, $\boldsymbol{K}' \in \mathcal{K}$ with $\boldsymbol{K}' = \boldsymbol{K} - \tau_2 \widetilde{\boldsymbol{F}}_{\boldsymbol{K},\boldsymbol{L}}$, $\widetilde{\boldsymbol{F}}_{\boldsymbol{K},\boldsymbol{L}} := \frac{1}{2}\widetilde{\nabla}_{\boldsymbol{K}} \mathcal{G}(\boldsymbol{K},\boldsymbol{L})\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{K},\boldsymbol{L}}^{-1}$. If we choose $\tau_2 \leq 1/(8G)$ where G is defined in Lemma A.6, then for*

$$\begin{aligned}
\mathcal{R}_{K_h,K_h'} &:= (K_h' - K_h)^T F_{K_h,L(K_h)} + F_{K_h,L(K_h)}^T (K_h' - K_h) \\
&\quad + (K_h' - K_h)^T (R_h^u + B_h^T \widetilde{P}_{K_{h+1},L(K_{h+1})} B_h)(K_h' - K_h). \\
\widetilde{P}_{K_{h+1},L(K_{h+1})} &:= P_{K_{h+1},L(K_{h+1})} + P_{K_{h+1},L(K_{h+1})} D_h (R_h^w - D_h^T P_{K_{h+1},L(K_{h+1})} D_h)^{-1} D_h^T P_{K_{h+1},L(K_{h+1})},
\end{aligned}$$

we have

$$\begin{aligned}
\mathcal{R}_{K_h,K_h'} &\preceq W_h - \frac{\tau_2}{4} F_{K_h,L(K_h)}^\top F_{K_h,L(K_h)}, \\
W_h &:= (4\tau_2 + 4\tau_2^2\|G_h\|)(F_{K_h,L_h}^r - F_{K_h,L_h})^\top (F_{K_h,L_h}^r - F_{K_h,L_h}) \\
&\quad + \tau_2 (F_{K_h,L(K_h)} - F_{K_h,L_h})^\top (F_{K_h,L(K_h)} - F_{K_h,L_h}) \\
&\quad + (2\tau_2 + \tau_2^2\|G_h\|) V(\widetilde{F}_{K_h,L_h}).
\end{aligned} \tag{A.5}$$

holds for $h = 0, \cdots, N-1$ where

$$F_{K_h,L_h}^r := \mathbb{E}[\widetilde{F}_{K_h,L_h}], \quad V(\widetilde{F}_{K_h,L_h}) := (\widetilde{F}_{K_h,L_h} - F_{K_h,L_h}^r)^\top (\widetilde{F}_{K_h,L_h} - F_{K_h,L_h}^r).$$

Proof. Consider the recursive inequality of $P_{K_h,L(K_h)}$ for $h = 0, \cdots, N-1$ from Lemma B.1 of Zhang et al. [2021b], we know that for any $\boldsymbol{K}, \boldsymbol{K}' \in \mathcal{K}$

$$P_{K'_h,L(K'_h)} - P_{K_h,L(K_h)} = A^\top_{K'_h,L(K'_h)}(P_{K'_{h+1},L(K'_{h+1})} - P_{K_{h+1},L(K_{h+1})})A_{K'_h,L(K'_h)} + \mathcal{R}_{K_h,K'_h} - \Xi^\top_{K_h,K'_h}(R^w_h - D^\top_h P_{K_{h+1},L(K_{h+1})}D_h)^{-1}\Xi_{K_h,K'_h},$$

holds for $h = 0, \cdots, N-1$. Here $K'_h = K_h - \tau_2 \widetilde{F}_{K_h,L_h}$ and $\boldsymbol{L}$ is the output of Algorithm 1. We denote $R^u_h + B^\top_h \widetilde{P}_{K_{h+1},L(K_{h+1})}B_h$ as G_h.

$$\begin{aligned}
\mathcal{R}_{K_h,K'_h} &= (K'_h - K_h)^\top F_{K_h,L(K_h)} + F^\top_{K_h,L(K_h)}(K'_h - K_h) + (K'_h - K_h)^\top G_h(K'_h - K_h) \\
&= -\tau_2 \widetilde{F}^T_{K_h,L_h}F_{K_h,L(K_h)} - \tau_2 F^\top_{K_h,L(K_h)}\widetilde{F}_{K_h,L_h} + \tau_2^2 \widetilde{F}^\top_{K_h,L_h}G_h\widetilde{F}_{K_h,L_h} \\
&= -\tau_2(\widetilde{F}_{K_h,L_h} - F^r_{K_h,L_h} + F^r_{K_h,L_h})^\top F_{K_h,L(K_h)} \\
&\quad - \tau_2 F^\top_{K_h,L(K_h)}(\widetilde{F}_{K_h,L_h} - F^r_{K_h,L_h} + F^r_{K_h,L_h}) \\
&\quad + \tau_2^2(\widetilde{F}_{K_h,L_h} - F^r_{K_h,L_h} + F^r_{K_h,L_h})^\top G_h(\widetilde{F}_{K_h,L_h} - F^r_{K_h,L_h} + F^r_{K_h,L_h}).
\end{aligned}$$

We denote the expectation $\mathbb{E}[\widetilde{F}_{K_h,L_h}]$ as $F^r_{K_h,L_h}$ where expectation is taken w.r.t. the randomness when estimating the natural gradient $2F_{K_h,L_h}$.

$$\begin{aligned}
\mathcal{R}_{K_h,K'_h} &\preceq -\tau_2(\widetilde{F}_{K_h,L_h} - F^r_{K_h,L_h})^\top F_{K_h,L(K_h)} - \tau_2 F^\top_{K_h,L(K_h)}(\widetilde{F}_{K_h,L_h} - F^r_{K_h,L_h}) \\
&\quad - \tau_2(F^r_{K_h,L_h})^\top F_{K_h,L(K_h)} - \tau_2 F^\top_{K_h,L(K_h)}F^r_{K_h,L_h} \\
&\quad + \tau_2^2\|G_h\|(\widetilde{F}_{K_h,L_h} - F^r_{K_h,L_h})^\top(\widetilde{F}_{K_h,L_h} - F^r_{K_h,L_h}) \\
&\quad + \tau_2^2\|G_h\|(\widetilde{F}_{K_h,L_h} - F^r_{K_h,L_h})^\top F^r_{K_h,L_h} + \tau_2^2\|G_h\|(F^r_{K_h,L_h})^\top(\widetilde{F}_{K_h,L_h} - F^r_{K_h,L_h}) \\
&\quad + \tau_2^2\|G_h\|(F^r_{K_h,L_h})^\top F^r_{K_h,L_h} \\
&\preceq \underbrace{(2\tau_2 + \tau_2^2\|G_h\|)(\widetilde{F}_{K_h,L_h} - F^r_{K_h,L_h})^\top(\widetilde{F}_{K_h,L_h} - F^r_{K_h,L_h})}_{(1)} \\
&\quad \underbrace{-\tau_2(F^r_{K_h,L_h})^\top F_{K_h,L(K_h)} - \tau_2 F^\top_{K_h,L(K_h)}F^r_{K_h,L_h}}_{(2)} \\
&\quad \underbrace{+\frac{\tau_2}{2}F^\top_{K_h,L(K_h)}F_{K_h,L(K_h)}}_{(3)} + \underbrace{2\tau_2^2\|G_h\|(F^r_{K_h,L_h})^\top F^r_{K_h,L_h}}_{(4)}.
\end{aligned}$$

In the first inequality, we apply Lemma A.27. In the second inequality, we apply Lemma A.31 and use

$$\begin{aligned}
(\widetilde{F}_{K_h,L_h} - F^r_{K_h,L_h})^\top F^r_{K_h,L_h} + (F^r_{K_h,L_h})^\top(\widetilde{F}_{K_h,L_h} - F^r_{K_h,L_h}) &\preceq (\widetilde{F}_{K_h,L_h} - F^r_{K_h,L_h})^\top(\widetilde{F}_{K_h,L_h} - F^r_{K_h,L_h}) \\
&\quad + (F^r_{K_h,L_h})^\top F^r_{K_h,L_h}, \\
-(\widetilde{F}_{K_h,L_h} - F^r_{K_h,L_h})^\top F_{K_h,L(K_h)} - F^\top_{K_h,L(K_h)}(\widetilde{F}_{K_h,L_h} - F^r_{K_h,L_h}) &\preceq 2(\widetilde{F}_{K_h,L_h} - F^r_{K_h,L_h})^\top(\widetilde{F}_{K_h,L_h} - F^r_{K_h,L_h}) \\
&\quad + \frac{1}{2}F^\top_{K_h,L(K_h)}F_{K_h,L(K_h)}.
\end{aligned}$$

We use the following notation for term (1)

$$V(\widetilde{F}_{K_h,L_h}) := (\widetilde{F}_{K_h,L_h} - F^r_{K_h,L_h})^\top(\widetilde{F}_{K_h,L_h} - F^r_{K_h,L_h}).$$

To control (2), we apply Lemma A.31 and $M^\top N + N^\top M = M^\top M + N^\top N - (M-N)^\top(M-N)$.

$$
\begin{aligned}
(2) &= -\tau_2(F^r_{K_h,L_h} - F_{K_h,L_h})^\top F_{K_h,L(K_h)} - \tau_2 F^\top_{K_h,L(K_h)}(F^r_{K_h,L_h} - F_{K_h,L_h}) \\
&\quad - \tau_2 F^\top_{K_h,L_h} F_{K_h,L(K_h)} - \tau_2 F^\top_{K_h,L(K_h)} F_{K_h,L_h} \\
&\preceq \tau_2 \left(4(F^r_{K_h,L_h} - F_{K_h,L_h})^\top(F^r_{K_h,L_h} - F_{K_h,L_h}) + \frac{1}{4}F^\top_{K_h,L(K_h)}F_{K_h,L(K_h)}\right) \\
&\quad - \tau_2\left(F^\top_{K_h,L(K_h)}F_{K_h,L(K_h)} + F^\top_{K_h,L_h}F_{K_h,L_h} - (F_{K_h,L(K_h)} - F_{K_h,L_h})^\top(F_{K_h,L(K_h)} - F_{K_h,L_h})\right) \\
&= 4\tau_2(F^r_{K_h,L_h} - F_{K_h,L_h})^\top(F^r_{K_h,L_h} - F_{K_h,L_h}) - \frac{3\tau_2}{4}F^\top_{K_h,L(K_h)}F_{K_h,L(K_h)} \\
&\quad - \tau_2 F^\top_{K_h,L_h}F_{K_h,L_h} + \tau_2(F_{K_h,L(K_h)} - F_{K_h,L_h})^\top(F_{K_h,L(K_h)} - F_{K_h,L_h}).
\end{aligned}
$$

As for (4), we use $(M+N)^\top(M+N) \preceq 2M^\top M + 2N^\top N$ and obtain

$$
\begin{aligned}
(4) = 2\tau_2^2\|G_h\|(F^r_{K_h,L_h} - F_{K_h,L_h} + F_{K_h,L_h})^\top(F^r_{K_h,L_h} - F_{K_h,L_h} + F_{K_h,L_h}) \preceq \\
4\tau_2^2\|G_h\|[(F^r_{K_h,L_h} - F_{K_h,L_h})^\top(F^r_{K_h,L_h} - F_{K_h,L_h}) + F^\top_{K_h,L_h}F_{K_h,L_h}].
\end{aligned}
$$

Now we organize terms and obtain

$$
\begin{aligned}
\mathcal{R}_{K_h,K'_h} &\preceq (4\tau_2 + 4\tau_2^2\|G_h\|)(F^r_{K_h,L_h} - F_{K_h,L_h})^\top(F^r_{K_h,L_h} - F_{K_h,L_h}) \\
&\quad - \frac{\tau_2}{4}F^\top_{K_h,L(K_h)}F_{K_h,L(K_h)} + (4\tau_2^2\|G_h\| - \frac{\tau_2}{2})F^\top_{K_h,L_h}F_{K_h,L_h} \\
&\quad + \tau_2(F_{K_h,L(K_h)} - F_{K_h,L_h})^\top(F_{K_h,L(K_h)} - F_{K_h,L_h}) \\
&\quad + (2\tau_2 + \tau_2^2\|G_h\|)V(\widetilde{F}_{K_h,L_h}).
\end{aligned}
$$

By choosing $\tau_2 \leq 1/(8G) \leq 1/(8\|\boldsymbol{G}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})}\|) \leq 1/(8\|G_h\|)$, we have for $h = 0, \cdots, N-1$

$$
\begin{aligned}
\mathcal{R}_{K_h,K'_h} &\preceq W_h - \frac{\tau_2}{4}F^\top_{K_h,L(K_h)}F_{K_h,L(K_h)}, \\
W_h &:= (4\tau_2 + 4\tau_2^2\|G_h\|)(F^r_{K_h,L_h} - F_{K_h,L_h})^\top(F^r_{K_h,L_h} - F_{K_h,L_h}) \\
&\quad + \tau_2(F_{K_h,L(K_h)} - F_{K_h,L_h})^\top(F_{K_h,L(K_h)} - F_{K_h,L_h}) \\
&\quad + (2\tau_2 + \tau_2^2\|G_h\|)V(\widetilde{F}_{K_h,L_h}).
\end{aligned}
$$

□

Lemma A.15. *(Descent-like inequality) Let $\boldsymbol{K}_0 \in \mathcal{K}$ and consider the following set*

$$
\hat{\mathcal{K}} := \left\{\boldsymbol{K} \mid (2.1) \text{ admits a solution } \boldsymbol{P}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})} \succeq 0, \right.
$$
$$
\left. \text{and } \boldsymbol{P}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})} \preceq \boldsymbol{P}_{\boldsymbol{K}_0,\boldsymbol{L}(\boldsymbol{K}_0)} + \frac{\lambda_{\min}(\boldsymbol{H}_{\boldsymbol{K}_0,\boldsymbol{L}(\boldsymbol{K}_0)})}{2\|\boldsymbol{D}\|}\cdot \boldsymbol{I}\right\}.
$$

For $\boldsymbol{K} \in \hat{\mathcal{K}}$, $\boldsymbol{K}' \in \mathcal{K}$ defined in with $\boldsymbol{K}' = \boldsymbol{K} - \tau_2\widetilde{\boldsymbol{F}}_{\boldsymbol{K},\boldsymbol{L}}$ and $\widetilde{\boldsymbol{F}}_{\boldsymbol{K}_t,\boldsymbol{L}_t} := \frac{1}{2}\widetilde{\nabla}_{\boldsymbol{K}}\mathcal{G}(\boldsymbol{K}_t,\boldsymbol{L}_t)\widetilde{\boldsymbol{\Sigma}}^{-1}_{\boldsymbol{K}_t,\boldsymbol{L}_t}$. If we choose

$$
\tau_2 \leq \min\{1/(8G), 1\},
$$

where G is defined in Lemma A.6. Then we have the following inequality

$$
\begin{aligned}
\boldsymbol{P}_{\boldsymbol{K}',\boldsymbol{L}(\boldsymbol{K}')} - \boldsymbol{P}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})} &\preceq \sum_{i=0}^{N}(\boldsymbol{A}^\top_{\boldsymbol{K}',\boldsymbol{L}(\boldsymbol{K}')})^i(\boldsymbol{e}_{1,\boldsymbol{K},\boldsymbol{K}'} + \boldsymbol{e}_{2,\boldsymbol{K},\boldsymbol{K}'} + \boldsymbol{e}_{3,\boldsymbol{K},\boldsymbol{K}'})(\boldsymbol{A}_{\boldsymbol{K}',\boldsymbol{L}(\boldsymbol{K}')})^i - \frac{\tau_2}{4}\boldsymbol{F}^\top_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})}\boldsymbol{F}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})}, \\
\boldsymbol{e}_{1,\boldsymbol{K},\boldsymbol{K}'} &:= (4\tau_2 + 4\tau_2^2\|\boldsymbol{G}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})}\|)(\boldsymbol{F}^r_{\boldsymbol{K},\boldsymbol{L}} - \boldsymbol{F}_{\boldsymbol{K},\boldsymbol{L}})^\top(\boldsymbol{F}^r_{\boldsymbol{K},\boldsymbol{L}} - \boldsymbol{F}_{\boldsymbol{K},\boldsymbol{L}}) \\
\boldsymbol{e}_{2,\boldsymbol{K},\boldsymbol{K}'} &:= \tau_2(\boldsymbol{F}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})} - \boldsymbol{F}_{\boldsymbol{K},\boldsymbol{L}})^\top(\boldsymbol{F}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})} - \boldsymbol{F}_{\boldsymbol{K},\boldsymbol{L}}) \\
\boldsymbol{e}_{3,\boldsymbol{K},\boldsymbol{K}'} &:= (2\tau_2 + \tau_2^2\|\boldsymbol{G}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})}\|)V(\widetilde{\boldsymbol{F}}_{\boldsymbol{K},\boldsymbol{L}}),
\end{aligned}
$$

where

$$F^r_{K,L} := \mathbb{E}[\widetilde{F}_{K,L}], \quad V(\widetilde{F}_{K,L}) := (\widetilde{F}_{K,L} - \mathbb{E}[\widetilde{F}_{K,L}])^\top (\widetilde{F}_{K,L} - \mathbb{E}[\widetilde{F}_{K,L}]).$$

This descent inequality is similar to the smooth inequality if we only look at $e_{K,K'}$ on the RHS.

Proof. We try to develop an upperbound for $P_{K',L(K')} - P_{K,L(K)}$. We start by discussing the difference between $P_{K_h,L(K_h)}$ and $P_{K'_h,L(K'_h)}$, which can be obtained by computing the difference between two Lyapunov equations and reorganizing the terms (see Lemma B.1 in Zhang et al. [2021b] for example).

$$\begin{aligned} P_{K'_h,L(K'_h)} - P_{K_h,L(K_h)} &= A^T_{K'_h,L(K'_h)}(P_{K'_{h+1},L(K'_{h+1})} - P_{K_{h+1},L(K_{h+1})})A_{K'_h,L(K'_h)} \\ &\quad + \mathcal{R}_{K_h,K'_h} - \Xi^T_{K_h,K'_h}(R^w_h - D^T_h P_{K_{h+1},L(K_{h+1})} D_h)^{-1}\Xi_{K_h,K'_h},\ h = 0,\cdots,N-1, \end{aligned}$$

and $P_{K_N,L(K_N)} = P_{K'_N,L(K'_N)} = Q_N$. Here Ξ_{K_h,K'_h} and $\mathcal{R}_{K_h,K'_h}$ are defined in Appendix A.1. For the simplicity of notations and proof, we write the above equation in a compact matrix form

$$\begin{aligned} P_{K',L(K')} - P_{K,L(K)} &= A^\top_{K',L(K')}(P_{K',L(K')} - P_{K,L(K)})A_{K',L(K')} + \mathcal{R}_{K,K'} \\ &\quad - \Xi^\top_{K,K'}(R^w - D^\top P_{K,L(K)} D)^{-1}\Xi_{K,K'} \end{aligned}$$

where

$$\mathcal{R}_{K,K'} := \mathrm{diag}(\mathcal{R}_{K_0,K'_0},\cdots,\mathcal{R}_{K_{N-1},K'_{N-1}},\mathbf{0}_{m\times m}), \quad \Xi_{K,K'} := \begin{bmatrix} \mathbf{0}_{m\times nN} \\ \mathrm{diag}(\Xi_{K_0,K'_0},\cdots,\Xi_{K_{N-1},K'_{N-1}}) \end{bmatrix}.$$

Then apply Lemma A.23, we observe that in order to upperbound $P_{K',L(K')} - P_{K,L(K)}$, we only need to upperbound $\mathcal{R}_{K,K'} - \Xi^\top_{K,K'}(R^w - D^\top P_{K,L(K)} D)^{-1}\Xi_{K,K'}$. Here we choose the step size $\tau_2 \le \min\{1/(8G), 1\}$ where G is defined in Appendix A.6. Then from the upperbound we developed for $\mathcal{R}_{K,K'}$ in Lemma A.14 and the fact that $R^w - D^\top P_{K,L(K)} D \succ 0$, we know

$$\begin{aligned} \mathcal{R}_{K,K'} - \Xi^\top_{K,K'}(R^w - D^\top P_{K,L(K)} D)^{-1}\Xi_{K,K'} \preceq \mathcal{R}_{K,K'} &\preceq \mathrm{diag}(W_0,\cdots,W_{N-1},\mathbf{0}_{m\times m}) \\ &\quad - \frac{\tau_2}{4} F^\top_{K,L(K)} F_{K,L(K)}. \end{aligned}$$

Hence we can upperbound $P_{K',L(K')} - P_{K,L(K)}$ with the solution to the Lyapunov equation below

$$\begin{aligned} P_{K',L(K')} - P_{K,L(K)} &= A^\top_{K',L(K')}(P_{K',L(K')} - P_{K,L(K)})A_{K',L(K')} + \mathrm{diag}(W_0,\cdots,W_{N-1},\mathbf{0}_{m\times m}) \\ &\quad - \frac{\tau_2}{4} F^\top_{K,L(K)} F_{K,L(K)}. \end{aligned}$$

Then apply Lemma A.22, we have

$$\begin{aligned} P_{K',L(K')} - P_{K,L(K)} \preceq \sum_{i=0}^{\infty} (A^\top_{K',L(K')})^i \big(&\mathrm{diag}(W_0,\cdots,W_{N-1},\mathbf{0}_{m\times m}) \\ &- \frac{\tau_2}{4} F^\top_{K,L(K)} F_{K,L(K)}\big)(A_{K',L(K')})^i. \end{aligned}$$

We also easily observe from the definition of W_h, (A.5), that

$$\begin{aligned} \mathrm{diag}(W_0,W_1,\cdots,W_{N-1},\mathbf{0}_{m\times m}) &\preceq (4\tau_2 + 4\tau_2^2\|G_{K,L(K)}\|)(F^r_{K,L} - F_{K,L})^\top (F^r_{K,L} - F_{K,L}) \\ &\quad + \tau_2 (F_{K,L(K)} - F_{K,L})^\top (F_{K,L(K)} - F_{K,L}) \\ &\quad + (2\tau_2 + \tau_2^2\|G_{K,L(K)}\|)V(\widetilde{F}_{K,L}). \end{aligned}$$

Hence we conclude that

$$
\begin{aligned}
P_{K',L(K')} - P_{K,L(K)} &\preceq \sum_{i=0}^{\infty} (A_{K',L(K')}^\top)^i (e_{1,K,K'} + e_{2,K,K'} + e_{3,K,K'})(A_{K',L(K')})^i - \frac{\tau_2}{4} F_{K,L(K)}^\top F_{K,L(K)} \\
&= \sum_{i=0}^{N} (A_{K',L(K')}^\top)^i (e_{1,K,K'} + e_{2,K,K'} + e_{3,K,K'})(A_{K',L(K')})^i - \frac{\tau_2}{4} F_{K,L(K)}^\top F_{K,L(K)}, \\
e_{1,K,K'} &:= (4\tau_2 + 4\tau_2^2 \|G_{K,L(K)}\|)(F_{K,L}^r - F_{K,L})^\top (F_{K,L}^r - F_{K,L}) \\
e_{2,K,K'} &:= \tau_2 (F_{K,L(K)} - F_{K,L})^\top (F_{K,L(K)} - F_{K,L}) \\
e_{3,K,K'} &:= (2\tau_2 + \tau_2^2 \|G_{K,L(K)}\|) V(\widetilde{F}_{K,L}).
\end{aligned}
$$

where in the first inequality, we apply $\sum_{i=1}^{\infty}(A_{K',L(K')}^\top)^i F_{K,L(K)}^\top F_{K,L(K)} (A_{K',L(K')})^i \succeq 0$ and in the first equation, we apply the fact that $(A_{K',L(K')})^{N+1} = \mathbf{0}$, see Lemma A.22. $\square$

The following lemma shows that: the output L of the inner-loop algorithm that satisfies the accuracy requirement not only implies the boundedness of $\mathcal{G}(K, L)$ and L but also the boundedness of $P_{K,L}$ and $\Sigma_{K,L}$.

Lemma A.16. *(Bounded output of the inner-loop algorithm) Let $K_0 \in \mathcal{K}$ and consider the following set*

$$
\begin{aligned}
\hat{\mathcal{K}} := \Big\{ K \mid (2.1) \text{ admits a solution } P_{K,L(K)} \succeq 0, \\
\text{and } P_{K,L(K)} \preceq P_{K_0,L(K_0)} + \frac{\lambda_{\min}(H_{K_0,L(K_0)})}{2\|D\|} \cdot I \Big\}.
\end{aligned}
$$

Let $K \in \hat{\mathcal{K}}$ and L be the output of Algorithm 1 given K with structures defined in (1.4), which satisfies

$$
\mathcal{G}(K, L(K)) - \mathcal{G}(K, L) \leq \varepsilon_1,
$$

with probability at least $1 - \delta_1$. Let ε_1 satisfy $\varepsilon_1 \leq D_3$ where D_3 is defined in Lemma A.6. Then we have

$$
\begin{aligned}
\|P_{K,L}\| &\leq \|P_{K,L(K)}\| + \varepsilon_1/\varphi \leq \|P_{K,L(K)}\| + 1/\varphi, \\
\|\Sigma_{K,L}\| &\leq \|\Sigma_{K,L(K)}\| + l_8 \sqrt{\lambda_{\min}^{-1}(H_{K,L(K)}) \cdot \varepsilon_1} \leq \|\Sigma_{K,L(K)}\| + l_8,
\end{aligned}
$$

with probability at least $1 - \delta_1$.

Proof. Consider

$$
\begin{aligned}
\|P_{K,L}\| &\leq \|P_{K,L(K)}\| + \|P_{K,L(K)} - P_{K,L}\| \leq \|P_{K,L(K)}\| + \mathrm{Tr}(P_{K,L(K)} - P_{K,L}) \\
&\leq \|P_{K,L(K)}\| + \varphi^{-1} \mathrm{Tr}((P_{K,L(K)} - P_{K,L})\Sigma_0) \\
&\leq \|P_{K,L(K)}\| + \varphi^{-1} (\mathcal{G}(K, L(K)) - \mathcal{G}(K, L)) \\
&\leq \|P_{K,L(K)}\| + \varepsilon_1/\varphi \leq \|P_{K,L(K)}\| + 1/\varphi
\end{aligned}
$$

In the second inequality, we utilize the optimality of $P_{K,L(K)}$ (see Lemma 2.1) and Lemma A.32. In the third inequality, we apply Lemma A.29. For the second result, since output L of the max-oracle satisfies

$$
\|L(K) - L\|_F \leq \sqrt{\lambda_{\min}^{-1}(H_{K,L(K)}) \cdot \varepsilon_1}
$$

with probability at least $1-\delta_1$. Then apply Lemma A.11 and choose $\varepsilon_1 \le D_3$ where D_3 is defined in Lemma A.6. We have

$$\|\boldsymbol{\Sigma}_{\boldsymbol{K},\boldsymbol{L}} - \boldsymbol{\Sigma}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})}\| \le l_8 \sqrt{\lambda_{\min}^{-1}(\boldsymbol{H}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})}) \cdot \varepsilon_1}$$
$$\|\boldsymbol{\Sigma}_{\boldsymbol{K},\boldsymbol{L}}\| \le \|\boldsymbol{\Sigma}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})}\| + l_8 \sqrt{\lambda_{\min}^{-1}(\boldsymbol{H}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})}) \cdot \varepsilon_1} \le \|\boldsymbol{\Sigma}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})}\| + l_8$$

holds with probability at least $1-\delta_1$. □

A.7.1 Proof of Optimality

In this subsection, we provide the proof of Lemma 2.1 indicating the optimality of $\boldsymbol{P}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})}$.

Proof. The proof is an immediate result by applying Lemma A.24. Since $\boldsymbol{K}$ is fixed and $\boldsymbol{E}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})} = 0$, the difference can be simplified below

$$\begin{aligned}
\boldsymbol{P}_{\boldsymbol{K},\boldsymbol{L}} - \boldsymbol{P}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})} &= \boldsymbol{A}_{\boldsymbol{K},\boldsymbol{L}}^\top (\boldsymbol{P}_{\boldsymbol{K},\boldsymbol{L}} - \boldsymbol{P}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})}) \boldsymbol{A}_{\boldsymbol{K},\boldsymbol{L}} \\
&\quad + (\boldsymbol{L} - \boldsymbol{L}(\boldsymbol{K}))^\top (-\boldsymbol{R}^w + \boldsymbol{D}^\top \boldsymbol{P}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})} \boldsymbol{D})(\boldsymbol{L} - \boldsymbol{L}(\boldsymbol{K})) \\
\Rightarrow \boldsymbol{P}_{\boldsymbol{K},\boldsymbol{L}} - \boldsymbol{P}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})} &\preceq 0 \\
\boldsymbol{P}_{\boldsymbol{K},\boldsymbol{L}} &\preceq \boldsymbol{P}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})}.
\end{aligned}$$

In the first inequality, we use the fact that $-\boldsymbol{R}^w + \boldsymbol{D}^\top \boldsymbol{P}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})} \boldsymbol{D} \prec 0$ and apply Lemma A.23. □

A.8 Minibatch Approximation

Lemma A.17. *(Bounded gradient estimates) Let $\boldsymbol{K}_0 \in \mathcal{K}$ and consider the following set*

$$\hat{\mathcal{K}} := \Big\{ \boldsymbol{K} \mid (2.1) \text{ admits a solution } \boldsymbol{P}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})} \succeq 0, \\ \text{and } \boldsymbol{P}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})} \preceq \boldsymbol{P}_{\boldsymbol{K}_0,\boldsymbol{L}(\boldsymbol{K}_0)} + \frac{\lambda_{\min}(\boldsymbol{H}_{\boldsymbol{K}_0,\boldsymbol{L}(\boldsymbol{K}_0)})}{2\|\boldsymbol{D}\|} \cdot \boldsymbol{I} \Big\}.$$

Let $\boldsymbol{L}$ be the output of Algorithm 1 given $\boldsymbol{K} \in \hat{\mathcal{K}}$, by choosing

$$r_2 \le D_1, \quad \varepsilon_1 \le D_3,$$
$$M_2 \ge \max\{M_{\boldsymbol{\Sigma}}(\varphi/2, \delta/2), M_V(1, \delta/2)\} = \mathcal{O}(\frac{1}{r_2^2} \cdot \log(\frac{1}{\delta}) + \log(\frac{1}{\delta})),$$

where D_1, D_3 are defined in Lemma A.6, and $M_{\boldsymbol{\Sigma}}(\cdot,\cdot)$, $M_V(\cdot,\cdot)$ are defined in Lemma A.19, A.18 respectively. Then we have

$$\|\widetilde{\boldsymbol{F}}_{\boldsymbol{K},\boldsymbol{L}}\|_F \le B_4,$$
$$B_4 := \sup_{(\boldsymbol{K},\boldsymbol{L}) \in \hat{\mathcal{K}} \times \hat{\mathcal{L}}} \Big(1 + (2l_1 \cdot \|\boldsymbol{\Sigma}_{\boldsymbol{K},\boldsymbol{L}}\| + 2l_1 \cdot l_7 + 2l_7 \|\boldsymbol{F}_{\boldsymbol{K},\boldsymbol{L}}\|) \cdot m(N+1) + \|\nabla_{\boldsymbol{K}} \mathcal{G}(\boldsymbol{K},\boldsymbol{L})\|_F \Big) \cdot \frac{1}{\varphi},$$

hold with probability at least $(1-\delta_1)(1-\delta)$. Hence by choosing the proper stepsize, we can maintain the iterates within the desired set such as $\mathcal{K}$ with high probability.

Proof. Since $\widetilde{\boldsymbol{F}}_{\boldsymbol{K},\boldsymbol{L}} = \frac{1}{2}\widetilde{\nabla}_{\boldsymbol{K}}\mathcal{G}(\boldsymbol{K},\boldsymbol{L})\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{K},\boldsymbol{L}}^{-1}$, we discuss the bound for $\|\widetilde{\nabla}_{\boldsymbol{K}}\mathcal{G}(\boldsymbol{K},\boldsymbol{L})\|_F$ first. Consider

$$\begin{aligned}\|\widetilde{\nabla}_{\boldsymbol{K}}\mathcal{G}(\boldsymbol{K},\boldsymbol{L})\|_F &= \|\widetilde{\nabla}_{\boldsymbol{K}}\mathcal{G}(\boldsymbol{K},\boldsymbol{L}) - \nabla_{\boldsymbol{K}}\mathcal{G}_{r_2}(\boldsymbol{K},\boldsymbol{L}) + \nabla_{\boldsymbol{K}}\mathcal{G}_{r_2}(\boldsymbol{K},\boldsymbol{L}) - \nabla_{\boldsymbol{K}}\mathcal{G}(\boldsymbol{K},\boldsymbol{L}) + \nabla_{\boldsymbol{K}}\mathcal{G}(\boldsymbol{K},\boldsymbol{L})\|_F \\ &\leq \|\widetilde{\nabla}_{\boldsymbol{K}}\mathcal{G}(\boldsymbol{K},\boldsymbol{L}) - \nabla_{\boldsymbol{K}}\mathcal{G}_{r_2}(\boldsymbol{K},\boldsymbol{L})\|_F + \|\nabla_{\boldsymbol{K}}\mathcal{G}_{r_2}(\boldsymbol{K},\boldsymbol{L}) - \nabla_{\boldsymbol{K}}\mathcal{G}(\boldsymbol{K},\boldsymbol{L})\|_F \\ &\quad + \|\nabla_{\boldsymbol{K}}\mathcal{G}(\boldsymbol{K},\boldsymbol{L})\|_F,\end{aligned}$$

where $\nabla_{\boldsymbol{K}}\mathcal{G}_{r_2}(\boldsymbol{K},\boldsymbol{L}) := \mathbb{E}[\widetilde{\nabla}_{\boldsymbol{K}}\mathcal{G}(\boldsymbol{K},\boldsymbol{L})]$, the expectation is taken w.r.t. all the randomness in estimating the gradients. Apply Lemma A.12 and (A.6)-(A.8) in Lemma A.18. If we choose

$$r_2 \leq D_1, \quad M_2 \geq M_V(1,\delta/2),$$

we have

$$\|\widetilde{\nabla}_{\boldsymbol{K}}\mathcal{G}(\boldsymbol{K},\boldsymbol{L})\|_F \leq 1 + (2l_1\cdot\|\boldsymbol{\Sigma}_{\boldsymbol{K},\boldsymbol{L}}\| + 2l_1\cdot l_7 + 2l_7\|\boldsymbol{F}_{\boldsymbol{K},\boldsymbol{L}}\|)\cdot m(N+1) + \|\nabla_{\boldsymbol{K}}\mathcal{G}(\boldsymbol{K},\boldsymbol{L})\|_F,$$

holds with probability at least $(1-\delta_1)(1-\delta/2)$. Then we apply Lemma A.19 and conclude that when

$$\varepsilon_1 \leq D_3, \quad M_2 \geq \max\big\{M_{\boldsymbol{\Sigma}}(\varphi/2,\delta/2), M_V(1,\delta/2)\big\},$$

we obtain

$$\begin{aligned}\|\widetilde{\boldsymbol{F}}_{\boldsymbol{K},\boldsymbol{L}}\|_F &\leq \Big(1 + (2l_1\cdot\|\boldsymbol{\Sigma}_{\boldsymbol{K},\boldsymbol{L}}\| + 2l_1\cdot l_7 + 2l_7\|\boldsymbol{F}_{\boldsymbol{K},\boldsymbol{L}}\|)\cdot m(N+1) + \|\nabla_{\boldsymbol{K}}\mathcal{G}(\boldsymbol{K},\boldsymbol{L})\|_F\Big)\cdot\frac{1}{\varphi} \\ &\leq \sup_{(\boldsymbol{K},\boldsymbol{L})\in\hat{\mathcal{K}}\times\hat{\mathcal{L}}}\Big(1 + (2l_1\cdot\|\boldsymbol{\Sigma}_{\boldsymbol{K},\boldsymbol{L}}\| + 2l_1\cdot l_7 + 2l_7\|\boldsymbol{F}_{\boldsymbol{K},\boldsymbol{L}}\|)\cdot m(N+1) \\ &\quad + \|\nabla_{\boldsymbol{K}}\mathcal{G}(\boldsymbol{K},\boldsymbol{L})\|_F\Big)\cdot\frac{1}{\varphi} =: B_4,\end{aligned}$$

holds with probability at least $(1-\delta_1)(1-\delta)$. □

Lemma A.18. *(Natural gradient estimation variance and sample size) Let $\boldsymbol{K}_0 \in \mathcal{K}$ and consider the following set*

$$\begin{aligned}\hat{\mathcal{K}} := \Big\{\boldsymbol{K} \mid (2.1) \text{ admits a solution } \boldsymbol{P}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})} \succeq 0, \\ \text{and } \boldsymbol{P}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})} \preceq \boldsymbol{P}_{\boldsymbol{K}_0,\boldsymbol{L}(\boldsymbol{K}_0)} + \frac{\lambda_{\min}(\boldsymbol{H}_{\boldsymbol{K}_0,\boldsymbol{L}(\boldsymbol{K}_0)})}{2\|\boldsymbol{D}\|}\cdot\boldsymbol{I}\Big\}.\end{aligned}$$

In Algorithm 2, input $\boldsymbol{K} \in \hat{\mathcal{K}}$, and $\boldsymbol{L}$ is the output of Algorithm 1 given $\boldsymbol{K}$. For $\delta_1 \in (0,1), \varepsilon_1 > 0$, $\boldsymbol{L}$ satisfies

$$\mathcal{G}(\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})) - \mathcal{G}(\boldsymbol{K},\boldsymbol{L}) \leq \varepsilon_1, \quad \|\boldsymbol{L}(\boldsymbol{K}) - \boldsymbol{L}\|_F \leq \sqrt{\lambda_{\min}^{-1}(\boldsymbol{H}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})})\cdot\varepsilon_1},$$

with probability at least $1-\delta_1$. If we choose

$$\begin{aligned}&r_2 \leq D_1, \quad \varepsilon_1 \leq D_3, \\ &M_2 \geq \max\Big\{M_{\boldsymbol{\Sigma}}(\varphi/2,\delta/2), M_{\boldsymbol{\Sigma}}(\frac{\varphi^2\sqrt{2\varepsilon}}{4O_1},\delta/2), M_V(\frac{\sqrt{2\varepsilon}\varphi}{4},\delta/2)\Big\} \\ &\quad = \mathcal{O}(r_2^{-2}\varepsilon^{-1}\cdot\log(\frac{1}{\delta}) + \varepsilon^{-1}\log(\frac{1}{\delta})),\end{aligned}$$

where D_1, D_3 are defined in Lemma A.6 and

$$M_V(\varepsilon,\delta) := \varepsilon^{-2}(\frac{O_2}{r_2} + O_1)^2 \cdot \log(\frac{2d_K}{\delta}) = \mathcal{O}(\frac{1}{r_2^2\varepsilon^2} \cdot \log(\frac{2}{\delta})),$$

$$O_1 := \sup_{(\boldsymbol{K},\boldsymbol{L})\in\hat{\mathcal{K}}\times\hat{\mathcal{L}}} (N+1)(d+m)(2l_1\|\boldsymbol{\Sigma}_{\boldsymbol{K},\boldsymbol{L}}\| + 2l_1 \cdot l_7 + 2l_7\|\boldsymbol{F}_{\boldsymbol{K},\boldsymbol{L}}\|) + 2\|\boldsymbol{F}_{\boldsymbol{K},\boldsymbol{L}}\boldsymbol{\Sigma}_{\boldsymbol{K},\boldsymbol{L}}\|_F,$$

$$O_2 := \sup_{\boldsymbol{K}\in\hat{\mathcal{K}}} d_{\boldsymbol{K}}(l_5 + \|\boldsymbol{P}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})}\| + 1/\varphi)\vartheta^2(N+1),$$

and $M_{\boldsymbol{\Sigma}}(\cdot,\cdot)$ is defined in Lemma A.19. Positive constants l_1, l_5, l_7 are defined in Lemma A.9, A.10, A.11. Then We have

$$\|V(\widetilde{\boldsymbol{F}}_{\boldsymbol{K},\boldsymbol{L}})\| \leq \varepsilon,$$

holds with probability at least $(1-\delta_1)(1-\delta)$. This lemma describe the relationship between the sample size M_2 and the algorithm parameters r_2, ε_1, and will be important for determining the total sample complexity.

Proof. According to Algorithm 2,

$$\begin{aligned}
V(\widetilde{\boldsymbol{F}}_{\boldsymbol{K},\boldsymbol{L}}) &= \frac{1}{2}(\widetilde{\nabla}_{\boldsymbol{K}}\mathcal{G}(\boldsymbol{K},\boldsymbol{L})\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{K},\boldsymbol{L}}^{-1} - \nabla_{\boldsymbol{K}}\mathcal{G}_{r_2}(\boldsymbol{K},\boldsymbol{L})\boldsymbol{\Sigma}_{\boldsymbol{K},\boldsymbol{L}}^{-1})^\top \cdot (\widetilde{\nabla}_{\boldsymbol{K}}\mathcal{G}(\boldsymbol{K},\boldsymbol{L})\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{K},\boldsymbol{L}}^{-1} - \nabla_{\boldsymbol{K}}\mathcal{G}_{r_2}(\boldsymbol{K},\boldsymbol{L})\boldsymbol{\Sigma}_{\boldsymbol{K},\boldsymbol{L}}^{-1}) \\
&= \frac{1}{2}(\widetilde{\nabla}_{\boldsymbol{K}}\mathcal{G}(\boldsymbol{K},\boldsymbol{L})\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{K},\boldsymbol{L}}^{-1} - \nabla_{\boldsymbol{K}}\mathcal{G}_{r_2}(\boldsymbol{K},\boldsymbol{L})\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{K},\boldsymbol{L}}^{-1} + \nabla_{\boldsymbol{K}}\mathcal{G}_{r_2}(\boldsymbol{K},\boldsymbol{L})\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{K},\boldsymbol{L}}^{-1} - \nabla_{\boldsymbol{K}}\mathcal{G}_{r_2}(\boldsymbol{K},\boldsymbol{L})\boldsymbol{\Sigma}_{\boldsymbol{K},\boldsymbol{L}}^{-1})^\top \\
&\quad \cdot (\widetilde{\nabla}_{\boldsymbol{K}}\mathcal{G}(\boldsymbol{K},\boldsymbol{L})\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{K},\boldsymbol{L}}^{-1} - \nabla_{\boldsymbol{K}}\mathcal{G}_{r_2}(\boldsymbol{K},\boldsymbol{L})\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{K},\boldsymbol{L}}^{-1} + \nabla_{\boldsymbol{K}}\mathcal{G}_{r_2}(\boldsymbol{K},\boldsymbol{L})\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{K},\boldsymbol{L}}^{-1} - \nabla_{\boldsymbol{K}}\mathcal{G}_{r_2}(\boldsymbol{K},\boldsymbol{L})\boldsymbol{\Sigma}_{\boldsymbol{K},\boldsymbol{L}}^{-1}) \\
&\preceq (\widetilde{\nabla}_{\boldsymbol{K}}\mathcal{G}(\boldsymbol{K},\boldsymbol{L})\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{K},\boldsymbol{L}}^{-1} - \nabla_{\boldsymbol{K}}\mathcal{G}_{r_2}(\boldsymbol{K},\boldsymbol{L})\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{K},\boldsymbol{L}}^{-1})^\top(\widetilde{\nabla}_{\boldsymbol{K}}\mathcal{G}(\boldsymbol{K},\boldsymbol{L})\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{K},\boldsymbol{L}}^{-1} - \nabla_{\boldsymbol{K}}\mathcal{G}_{r_2}(\boldsymbol{K},\boldsymbol{L})\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{K},\boldsymbol{L}}^{-1}) \\
&\quad + (\nabla_{\boldsymbol{K}}\mathcal{G}_{r_2}(\boldsymbol{K},\boldsymbol{L})\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{K},\boldsymbol{L}}^{-1} - \nabla_{\boldsymbol{K}}\mathcal{G}_{r_2}(\boldsymbol{K},\boldsymbol{L})\boldsymbol{\Sigma}_{\boldsymbol{K},\boldsymbol{L}}^{-1})^\top(\nabla_{\boldsymbol{K}}\mathcal{G}_{r_2}(\boldsymbol{K},\boldsymbol{L})\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{K},\boldsymbol{L}}^{-1} - \nabla_{\boldsymbol{K}}\mathcal{G}_{r_2}(\boldsymbol{K},\boldsymbol{L})\boldsymbol{\Sigma}_{\boldsymbol{K},\boldsymbol{L}}^{-1}).
\end{aligned}$$

For the first inequality, we apply Lemma A.31. Furthermore, since $V(\widetilde{\boldsymbol{F}}_{\boldsymbol{K},\boldsymbol{L}}) \succeq 0$

$$\begin{aligned}
\|V(\widetilde{\boldsymbol{F}}_{\boldsymbol{K},\boldsymbol{L}})\| &\leq \|\widetilde{\nabla}_{\boldsymbol{K}}\mathcal{G}(\boldsymbol{K},\boldsymbol{L})\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{K},\boldsymbol{L}}^{-1} - \nabla_{\boldsymbol{K}}\mathcal{G}_{r_2}(\boldsymbol{K},\boldsymbol{L})\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{K},\boldsymbol{L}}^{-1}\|^2 \\
&\quad + \|\nabla_{\boldsymbol{K}}\mathcal{G}_{r_2}(\boldsymbol{K},\boldsymbol{L})\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{K},\boldsymbol{L}}^{-1} - \nabla_{\boldsymbol{K}}\mathcal{G}_{r_2}(\boldsymbol{K},\boldsymbol{L})\boldsymbol{\Sigma}_{\boldsymbol{K},\boldsymbol{L}}^{-1}\|^2 \\
&\leq \|\widetilde{\nabla}_{\boldsymbol{K}}\mathcal{G}(\boldsymbol{K},\boldsymbol{L}) - \nabla_{\boldsymbol{K}}\mathcal{G}_{r_2}(\boldsymbol{K},\boldsymbol{L})\|^2 \cdot \|\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{K},\boldsymbol{L}}^{-1}\|^2 + \|\nabla_{\boldsymbol{K}}\mathcal{G}_{r_2}(\boldsymbol{K},\boldsymbol{L})\|^2 \cdot \|\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{K},\boldsymbol{L}}^{-1} - \boldsymbol{\Sigma}_{\boldsymbol{K},\boldsymbol{L}}^{-1}\|^2.
\end{aligned}$$

Consider random variable $\frac{d_K}{r_2}\mathcal{G}_{\boldsymbol{\xi}}(\boldsymbol{K}+r_2\boldsymbol{V},\boldsymbol{L})\boldsymbol{V} - \nabla_{\boldsymbol{K}}\mathcal{G}_{r_2}(\boldsymbol{K},\boldsymbol{L})$ where $\boldsymbol{V}$ is sampled uniformly from the unit sphere and $\boldsymbol{\xi}$ is sampled following distribution $\mathcal{D}$.

$$\left\|\frac{d_K}{r_2}\mathcal{G}_{\boldsymbol{\xi}}(\boldsymbol{K}+r_2\boldsymbol{V},\boldsymbol{L})\boldsymbol{V} - \nabla_{\boldsymbol{K}}\mathcal{G}_{r_2}(\boldsymbol{K},\boldsymbol{L})\right\|_F \leq \underbrace{\frac{d_K}{r_2}\|\mathcal{G}_{\boldsymbol{\xi}}(\boldsymbol{K}+r_2\boldsymbol{V},\boldsymbol{L})\boldsymbol{V}\|_F}_{(1)} + \underbrace{\|\nabla_{\boldsymbol{K}}\mathcal{G}_{r_2}(\boldsymbol{K},\boldsymbol{L})\|_F}_{(2)}.$$

For term (1), we have

$$
\begin{aligned}
(1) = \frac{d_{\boldsymbol{K}}}{r_2}\|\mathcal{G}_{\boldsymbol{\xi}}(\boldsymbol{K}+r_2\boldsymbol{V},\boldsymbol{L})\boldsymbol{V}\|_F &= \frac{d_{\boldsymbol{K}}}{r_2}|\mathcal{G}_{\boldsymbol{\xi}}(\boldsymbol{K}+r_2\boldsymbol{V},\boldsymbol{L})| \\
&\le \frac{d_{\boldsymbol{K}}}{r_2}\|\boldsymbol{P}_{\boldsymbol{K}+r_2\boldsymbol{V},\boldsymbol{L}}\|\|\boldsymbol{\xi}\boldsymbol{\xi}^\top\| \\
&\le \frac{d_{\boldsymbol{K}}}{r_2}(\|\boldsymbol{P}_{\boldsymbol{K}+r_2\boldsymbol{V},\boldsymbol{L}}-\boldsymbol{P}_{\boldsymbol{K},\boldsymbol{L}}\|+\|\boldsymbol{P}_{\boldsymbol{K},\boldsymbol{L}}\|)\|\boldsymbol{\xi}\boldsymbol{\xi}^\top\|_F \\
&\le \frac{d_{\boldsymbol{K}}}{r_2}(l_5\cdot r_2+\|\boldsymbol{P}_{\boldsymbol{K},\boldsymbol{L}}\|)\|\boldsymbol{\xi}\boldsymbol{\xi}^\top\|_F \\
&\le \frac{d_{\boldsymbol{K}}}{r_2}(l_5\cdot r_2+\|\boldsymbol{P}_{\boldsymbol{K},\boldsymbol{L}}\|)\vartheta^2(N+1) \\
&\le \frac{d_{\boldsymbol{K}}}{r_2}(l_5+\|\boldsymbol{P}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})}\|+1/\varphi)\vartheta^2(N+1)
\end{aligned}
$$

holds with probability at least $1-\delta_1$. In the fifth inequality, we apply Lemma A.16. In the third inequality, we apply Lemma A.9 and require

$$r_2 \le D_1, \quad \varepsilon_1 \le D_3$$

where D_1, D_3 are defined in Lemma A.6. For term (2), we have

$$
\begin{aligned}
(2) &= \|\nabla_{\boldsymbol{K}}\mathcal{G}_{r_2}(\boldsymbol{K},\boldsymbol{L})-\nabla_{\boldsymbol{K}}\mathcal{G}(\boldsymbol{K},\boldsymbol{L})+\nabla_{\boldsymbol{K}}\mathcal{G}(\boldsymbol{K},\boldsymbol{L})\|_F \\
&\le \|\nabla_{\boldsymbol{K}}\mathcal{G}_{r_2}(\boldsymbol{K},\boldsymbol{L})-\nabla_{\boldsymbol{K}}\mathcal{G}(\boldsymbol{K},\boldsymbol{L})\|_F+\|\nabla_{\boldsymbol{K}}\mathcal{G}(\boldsymbol{K},\boldsymbol{L})\|_F \\
&\le (N+1)(d+m)(2l_1\|\boldsymbol{\Sigma}_{\boldsymbol{K},\boldsymbol{L}}\|+2l_1\cdot l_7+2l_7\|\boldsymbol{F}_{\boldsymbol{K},\boldsymbol{L}}\|)\cdot r_2+2\|\boldsymbol{F}_{\boldsymbol{K},\boldsymbol{L}}\boldsymbol{\Sigma}_{\boldsymbol{K},\boldsymbol{L}}\|_F \\
&\le (N+1)(d+m)(2l_1\|\boldsymbol{\Sigma}_{\boldsymbol{K},\boldsymbol{L}}\|+2l_1\cdot l_7+2l_7\|\boldsymbol{F}_{\boldsymbol{K},\boldsymbol{L}}\|)+2\|\boldsymbol{F}_{\boldsymbol{K},\boldsymbol{L}}\boldsymbol{\Sigma}_{\boldsymbol{K},\boldsymbol{L}}\|_F \\
&\le \sup_{(\boldsymbol{K},\boldsymbol{L})\in\hat{\mathcal{K}}\times\hat{\mathcal{L}}}(N+1)(d+m)(2l_1\|\boldsymbol{\Sigma}_{\boldsymbol{K},\boldsymbol{L}}\|+2l_1\cdot l_7+2l_7\|\boldsymbol{F}_{\boldsymbol{K},\boldsymbol{L}}\|)+2\|\boldsymbol{F}_{\boldsymbol{K},\boldsymbol{L}}\boldsymbol{\Sigma}_{\boldsymbol{K},\boldsymbol{L}}\|_F =: O_1
\end{aligned}
$$

holds with probability at least $1-\delta_1$. In the second inequality, we apply Lemma A.12. Summarizing the above inequalities we have

$$
\begin{aligned}
\left\|\frac{d_{\boldsymbol{K}}}{r_2}\mathcal{G}(\boldsymbol{K}+r_2\boldsymbol{V},\boldsymbol{L})-\nabla_{\boldsymbol{K}}\mathcal{G}_{r_2}(\boldsymbol{K},\boldsymbol{L})\right\|_F &\le \sup_{\boldsymbol{K}\in\hat{\mathcal{K}}}\frac{d_{\boldsymbol{K}}}{r_2}(l_5+\|\boldsymbol{P}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})}\|+1/\varphi)\vartheta^2(N+1)+O_1 \\
&= \frac{O_2}{r_2}+O_1, \\
O_2 &=: \sup_{\boldsymbol{K}\in\hat{\mathcal{K}}} d_{\boldsymbol{K}}(l_5+\|\boldsymbol{P}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})}\|+1/\varphi)\vartheta^2(N+1),
\end{aligned}
$$

holds with probability at least $1-\delta_1$. In the first inequality, we apply Lemma A.16. When the output of Algorith 1, $\boldsymbol{L}$, satisfies the accuracy requirement, term $\frac{d_{\boldsymbol{K}}}{r_2}\mathcal{G}(\boldsymbol{K}+r_2\boldsymbol{V},\boldsymbol{L})\boldsymbol{V}-\nabla_{\boldsymbol{K}}\mathcal{G}_{r_2}(\boldsymbol{K},\boldsymbol{L})$ is bounded, and hence norm-subGaussian w.r.t. random variable $\boldsymbol{V}$. Then we apply Corollary 7 in Jin et al. [2019], with probability at least $(1-\delta_1)(1-\delta)$, we have

$$
\begin{aligned}
\|\widetilde{\nabla}_{\boldsymbol{K}}\mathcal{G}(\boldsymbol{K},\boldsymbol{L})-\nabla_{\boldsymbol{K}}\mathcal{G}_{r_2}(\boldsymbol{K},\boldsymbol{L})\|_F &= \left\|\frac{1}{M_2}\sum_{m=0}^{M_2-1}(\frac{d_{\boldsymbol{K}}}{r_2}\mathcal{G}(\boldsymbol{K}+r_2\boldsymbol{V}_m,L)\boldsymbol{V}_m-\nabla_{\boldsymbol{K}}\mathcal{G}_{r_2}(\boldsymbol{K},\boldsymbol{L}))\right\|_F \\
&\le \frac{1}{M_2}\cdot\sqrt{M_2}\cdot(\frac{O_2}{r_2}+O_1)\cdot\sqrt{\log(\frac{2d_{\boldsymbol{K}}}{\delta})}
\end{aligned}
$$

Hence when we sample

$$M_2 \ge \max\left\{M_{\boldsymbol{\Sigma}}(\varphi/2,\delta/2), M_V(\frac{\sqrt{2\varepsilon\varphi}}{4},\delta/2)\right\} = \mathcal{O}(\frac{1}{r_2^2\varepsilon}\cdot\log(\frac{1}{\delta})), \tag{A.6}$$

$$M_V(\varepsilon,\delta) := \varepsilon^{-2}(\frac{O_2}{r_2}+O_1)^2\cdot\log(\frac{2}{\delta}), \tag{A.7}$$

we have

$$\begin{aligned}\|\widetilde{\nabla}_{K}\mathcal{G}(K,L)-\nabla_{K}\mathcal{G}_{r_2}(K,L)\|^2\cdot\|\widetilde{\Sigma}_{K,L}^{-1}\|^2 &\leq \|\widetilde{\nabla}_{K}\mathcal{G}(K,L)-\nabla_{K}\mathcal{G}_{r_2}(K,L)\|_F^2\cdot\|\widetilde{\Sigma}_{K,L}^{-1}\|^2\\ &\leq (\frac{\sqrt{2\varepsilon}\varphi}{4})^2\cdot(\frac{2}{\varphi})^2\leq\frac{\varepsilon}{2},\end{aligned}\tag{A.8}$$

with probability at least $(1-\delta_1)(1-\delta)$. Moreover, apply Lemma A.19, by sampling

$$M_2\geq\max\left\{M_{\Sigma}(\varphi/2,\delta/2),M_{\Sigma}(\frac{\varphi^2\sqrt{2\varepsilon}}{4O_1},\delta/2)\right\}=\mathcal{O}(\varepsilon^{-1}\cdot\log(\frac{1}{\delta})),$$

we can bound

$$\begin{aligned}\|\nabla_{K}\mathcal{G}_{r_2}(K,L)\|^2\cdot\|\widetilde{\Sigma}_{K,L}^{-1}-\Sigma_{K,L}^{-1}\|^2 &\leq\|\nabla_{K}\mathcal{G}_{r_2}(K,L)\|^2\cdot\|\widetilde{\Sigma}_{K,L}^{-1}\|^2\cdot\|\widetilde{\Sigma}_{K,L}-\Sigma_{K,L}\|^2\cdot\|\Sigma_{K,L}^{-1}\|^2\\ &\leq\frac{4}{\varphi^4}(O_1)^2\|\widetilde{\Sigma}_{K,L}-\Sigma_{K,L}\|^2\leq\frac{\varepsilon}{2},\end{aligned}$$

with probability at least $(1-\delta_1)(1-\delta/2)$. In conclusion, by sampling

$$\begin{aligned}M_2 &\geq\max\left\{M_{\Sigma}(\varphi/2,\delta/2),M_{\Sigma}(\frac{\varphi^2\sqrt{2\varepsilon}}{4O_1},\delta/2),M_V(\frac{\sqrt{2\varepsilon}\varphi}{4},\delta/2)\right\}\\ &=\mathcal{O}(\frac{1}{r_2^2\varepsilon}\cdot\log(\frac{1}{\delta})+\varepsilon^{-1}\log(\frac{1}{\delta})),\end{aligned}$$

we have

$$\begin{aligned}\|V(\widetilde{F}_{K,L})\|_F &\leq\|\widetilde{\nabla}_{K}\mathcal{G}(K,L)-\nabla_{K}\mathcal{G}_{r_2}(K,L)\|_F^2\cdot\|\widetilde{\Sigma}_{K,L}^{-1}\|_F^2+\|\nabla_{K}\mathcal{G}_{r_2}(K,L)\|_F^2\cdot\|\widetilde{\Sigma}_{K,L}^{-1}-\Sigma_{K,L}^{-1}\|_F^2\\ &\leq\varepsilon/2+\varepsilon/2=\varepsilon,\end{aligned}$$

holds with probability at least $(1-\delta_1)(1-\delta)$.

□

Lemma A.19. *(Bounded estimated covariance matrix) For any sampled trajectory following policies K, L defined in (1.4), (1.5), we have*

$$\|\widetilde{\Sigma}_{K,L,\xi}\|\leq\operatorname{Tr}(\Sigma_{K,L})\cdot m(N+1)^2\vartheta/\varphi,\quad a.s.$$

holds for any initial condition ξ that satisfies Assumption 1.1. Especially, consider $K\in\hat{\mathcal{K}}$ and L is the output of the max-oracle given K. Moreover, if we require $\varepsilon_1\leq D_3$ where D_3 is a positive constant defined in Lemma A.6. Then by sampling

$$\begin{aligned}M_2 &\geq M_{\Sigma}(\varepsilon,\delta)=\mathcal{O}(\varepsilon^{-2}\cdot\log(\frac{1}{\delta})),\\ M_{\Sigma}(\varepsilon,\delta) &:=\sup_{K\in\hat{\mathcal{K}}}\varepsilon^{-2}(m^2(N+1)^2\vartheta^2/\varphi+\sqrt{m(N+1)})^2(\|\Sigma_{K,L(K)}\|+l_8)^2\log(\frac{2d_{\Sigma}}{\delta}),\end{aligned}$$

independent trajectories in Algorithm 2, we can guarantee

$$\|\widetilde{\Sigma}_{K,L}-\Sigma_{K,L}\|\leq\|\widetilde{\Sigma}_{K,L}-\Sigma_{K,L}\|_F\leq\varepsilon,$$

with probability at least $(1-\delta_1)(1-\delta)$. And by choosing $\varepsilon\leq\varphi/2$, we have

$$\lambda_{\min}(\widetilde{\Sigma}_{K,L})\geq\varphi/2\Rightarrow\|\widetilde{\Sigma}_{K,L}^{-1}\|\leq\frac{2}{\varphi}.$$

Here positive constant l_8 is defined in Lemma A.11.

Proof. As for the upperbound, since we assume that $\|x_0\|, \|\xi_h\| \leq \vartheta$ almost surely for $h = 0, 1, \cdots, N-1$ in Assumption 1.1. Then for any sampled trajectory, as long as $\|\boldsymbol{K}\|, \|\boldsymbol{L}\|$ are bounded almost surely, $\|\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{K},\boldsymbol{L}}\|$ is also bounded almost surely. For any sampled $\boldsymbol{\xi}$, apply Lemma A.22 and we have

$$\begin{aligned}
\boldsymbol{\Sigma}_{\boldsymbol{K},\boldsymbol{L},\boldsymbol{\xi}} &= \boldsymbol{A}_{\boldsymbol{K},\boldsymbol{L}}\boldsymbol{\Sigma}_{\boldsymbol{K},\boldsymbol{L},\boldsymbol{\xi}} + \boldsymbol{A}_{\boldsymbol{K},\boldsymbol{L}}^\top + \boldsymbol{\xi}\boldsymbol{\xi}^\top, \\
\boldsymbol{\Sigma}_{\boldsymbol{K},\boldsymbol{L},\boldsymbol{\xi}} &:= \sum_{h=0}^{N} \boldsymbol{A}_{\boldsymbol{K},\boldsymbol{L}}^h \boldsymbol{\xi}\boldsymbol{\xi}^\top (\boldsymbol{A}_{\boldsymbol{K},\boldsymbol{L}}^\top)^h \preceq \|\boldsymbol{\xi}\boldsymbol{\xi}^\top\| \cdot \sum_{h=0}^{N} \boldsymbol{A}_{\boldsymbol{K},\boldsymbol{L}}^h (\boldsymbol{A}_{\boldsymbol{K},\boldsymbol{L}}^\top)^h, \\
\|\boldsymbol{\Sigma}_{\boldsymbol{K},\boldsymbol{L},\boldsymbol{\xi}}\|_F &\leq \|\boldsymbol{\xi}\boldsymbol{\xi}^\top\| \cdot \left\| \sum_{h=0}^{N} \boldsymbol{A}_{\boldsymbol{K},\boldsymbol{L}}^h (\boldsymbol{A}_{\boldsymbol{K},\boldsymbol{L}}^\top)^h \right\|_F \leq \mathrm{Tr}(\boldsymbol{\xi}\boldsymbol{\xi}^\top) \cdot \mathrm{Tr}(\boldsymbol{\Sigma}_{\boldsymbol{K},\boldsymbol{L}})/\varphi \\
&= \mathrm{Tr}(\boldsymbol{\xi}^\top\boldsymbol{\xi}) \cdot \mathrm{Tr}(\boldsymbol{\Sigma}_{\boldsymbol{K},\boldsymbol{L}})/\varphi \leq \mathrm{Tr}(\boldsymbol{\Sigma}_{\boldsymbol{K},\boldsymbol{L}}) \cdot m(N+1)\vartheta^2/\varphi \quad a.s..
\end{aligned}$$

In the third inequality, we apply Lemma A.25. As an a.s. bounded random variable, we know $\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{K},\boldsymbol{L}}$ is norm-subGaussian Jin et al. [2019]. Hence we know that with probability at least $1-\delta$

$$\begin{aligned}
\|\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{K},\boldsymbol{L}} - \boldsymbol{\Sigma}_{\boldsymbol{K},\boldsymbol{L}}\|_F &= \left\| \frac{1}{M_2} \sum_{m=0}^{M_2-1} \boldsymbol{\Sigma}_{\boldsymbol{K},\boldsymbol{L},\boldsymbol{\xi}_m} - \boldsymbol{\Sigma}_{\boldsymbol{K},\boldsymbol{L}} \right\|_F = \frac{1}{M_2} \left\| \sum_{m=0}^{M_2-1} (\boldsymbol{\Sigma}_{\boldsymbol{K},\boldsymbol{L},\boldsymbol{\xi}_m} - \boldsymbol{\Sigma}_{\boldsymbol{K},\boldsymbol{L}}) \right\|_F \\
&\leq \frac{1}{M_2} \cdot \sqrt{M_2} \cdot (\mathrm{Tr}(\boldsymbol{\Sigma}_{\boldsymbol{K},\boldsymbol{L}}) \cdot m(N+1)\vartheta^2/\varphi + \|\boldsymbol{\Sigma}_{\boldsymbol{K},\boldsymbol{L}}\|_F) \cdot \sqrt{\log(\frac{2d_{\boldsymbol{\Sigma}}}{\delta})}.
\end{aligned}$$

When $\boldsymbol{L}$ is the output of Algorithm 1 given $\boldsymbol{K}$, apply Lemma A.16 and require $\epsilon_1 \leq D_3$, then we have

$$\|\boldsymbol{L} - \boldsymbol{L}(\boldsymbol{K})\| \leq \|\boldsymbol{L} - \boldsymbol{L}(\boldsymbol{K})\|_F \leq \sqrt{\lambda_{\min}^{-1}(\boldsymbol{H}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})}) \cdot \epsilon_1} \leq D_2, \ w.p. \geq 1-\delta_1,$$

where D_3, D_2 are defined in Lemma A.6. We choose

$$\begin{aligned}
M_2 &\geq \sup_{\boldsymbol{K}\in\hat{\mathcal{K}}} \epsilon^{-2}(m^2(N+1)^2\vartheta^2/\varphi + \sqrt{m(N+1)})^2(\|\boldsymbol{\Sigma}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})}\| + l_8)^2 \log(\frac{2d_{\boldsymbol{\Sigma}}}{\delta}) \\
&=: M_{\Sigma}(\epsilon,\delta) = \mathcal{O}(\frac{1}{\epsilon^2} \cdot \log(\frac{2}{\delta})) \\
&\geq \epsilon^{-2}(\mathrm{Tr}(\boldsymbol{\Sigma}_{\boldsymbol{K},\boldsymbol{L}}) \cdot m(N+1)\vartheta^2/\varphi + \|\boldsymbol{\Sigma}_{\boldsymbol{K},\boldsymbol{L}}\|_F)^2 \log(\frac{2d_{\boldsymbol{\Sigma}}}{\delta}).
\end{aligned}$$

In the first inequality, we apply Lemma A.16 to bound $\|\boldsymbol{\Sigma}_{\boldsymbol{K},\boldsymbol{L}}\|_F$ and $\mathrm{Tr}(M) \leq \sqrt{n}\|M\|_F$ where $M \in \mathbb{R}^{n\times n}$. Conclusively we have

$$\|\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{K},\boldsymbol{L}} - \boldsymbol{\Sigma}_{\boldsymbol{K},\boldsymbol{L}}\|_F \leq \epsilon,$$

with probability at least $(1-\delta_1)(1-\delta)$. Specially, by choosing $\epsilon = \varphi/2$, we have

$$\|\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{K},\boldsymbol{L}} - \boldsymbol{\Sigma}_{\boldsymbol{K},\boldsymbol{L}}\| \leq \|\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{K},\boldsymbol{L}} - \boldsymbol{\Sigma}_{\boldsymbol{K},\boldsymbol{L}}\|_F \leq \frac{\varphi}{2}.$$

Then

$$\begin{aligned}
\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{K},\boldsymbol{L}} &= \boldsymbol{\Sigma}_{\boldsymbol{K},\boldsymbol{L}} - (\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{K},\boldsymbol{L}} - \boldsymbol{\Sigma}_{\boldsymbol{K},\boldsymbol{L}}) \geq \boldsymbol{\Sigma}_{\boldsymbol{K},\boldsymbol{L}} - \|\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{K},\boldsymbol{L}} - \boldsymbol{\Sigma}_{\boldsymbol{K},\boldsymbol{L}}\| \cdot I \geq \frac{\varphi}{2} \cdot I \\
\Rightarrow \lambda_{\min}(\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{K},\boldsymbol{L}}) &\geq \frac{\varphi}{2} \Rightarrow \|\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{K},\boldsymbol{L}}^{-1}\| \leq \frac{2}{\varphi},
\end{aligned}$$

where in the second inequality, we use the fact that $\boldsymbol{\Sigma}_{\boldsymbol{K},\boldsymbol{L}} \succeq \boldsymbol{\Sigma}_0$. ☐

Lemma A.20. *(Descent-like inequality) For $\boldsymbol{K} \in \hat{\mathcal{K}}$ with $\boldsymbol{K}_0 \in \mathcal{K}$, $\boldsymbol{K}'$ with $\boldsymbol{K}' = \boldsymbol{K} - \tau_2 \widetilde{\boldsymbol{F}}_{\boldsymbol{K},\boldsymbol{L}}$ where $\boldsymbol{L}$ is the output of the max oracle given $\boldsymbol{K}$. If we require*

$$\tau_2 \leq \min\left\{1/(8G), B_2/(\sqrt{m(N+1)}B_4), B_1/(\sqrt{m(N+1)}B_4), 1\right\}, \quad r_2 \leq D_1,$$
$$\varepsilon_1 \leq D_3, \quad M_2 \geq \max\{M_{\boldsymbol{\Sigma}}(\varphi/2, \delta/2), M_V(1, \delta/2)\},$$

where G is defined in Lemma A.6, D_1 and D_3 are defined in Lemma A.6, B_1, B_2, B_4 are defined in Lemma A.8, A.7, A.17. Then with probability at least $(1-\delta_1)(1-\delta)$, we have $\boldsymbol{K}' \in \mathcal{K}$ and positive constants c_1, c_2, c_3 such that

$$\begin{aligned}
\boldsymbol{P}_{\boldsymbol{K}',\boldsymbol{L}(\boldsymbol{K}')} - \boldsymbol{P}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})} &\preceq \tau_2 \cdot (c_1 \cdot r_2^2 + c_2 \cdot \varepsilon_1 + c_3 \cdot \|V(\widetilde{\boldsymbol{F}}_{\boldsymbol{K},\boldsymbol{L}})\|), \\
c_1 &:= \sup_{(\boldsymbol{K},\boldsymbol{L}) \in \hat{\mathcal{K}} \times \hat{\mathcal{L}}} \quad (4+4G)/\varphi^3 \cdot (l_1 \|\boldsymbol{\Sigma}_{\boldsymbol{K},\boldsymbol{L}}\| + l_1 \cdot l_7 + l_7 \|\boldsymbol{F}_{\boldsymbol{K},\boldsymbol{L}}\|)^2 \\
&\quad \cdot (\|\boldsymbol{\Sigma}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})}\|_F + \mathcal{B}_{\boldsymbol{\Sigma}}), \\
c_2 &:= \sup_{\boldsymbol{K} \in \hat{\mathcal{K}}} (l_2)^2 \cdot H(\|\boldsymbol{\Sigma}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})}\|_F + \mathcal{B}_{\boldsymbol{\Sigma}})/\varphi, \\
c_3 &:= \sup_{\boldsymbol{K} \in \hat{\mathcal{K}}} (2+G) \cdot (\|\boldsymbol{\Sigma}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})}\|_F + \mathcal{B}_{\boldsymbol{\Sigma}})/\varphi.
\end{aligned}$$

In other words, in this Lemma, we decompose the deviation from monotonicity into three sources: (a) biased estimation term with r_2^2; (b) estimation error caused by using approximate solution of the inner-loop problem; (c) variance-like term that can be controlled via large enough sample size.

Proof. Apply Lemma A.17, we know that by sampling

$$M_2 \geq \max\{M_{\boldsymbol{\Sigma}}(\varphi/2, \delta/2), M_V(1, \delta/2)\},$$

we can ensure $\|\widetilde{\boldsymbol{F}}_{\boldsymbol{K},\boldsymbol{L}}\| \leq B_4$ with probability at least $(1-\delta_1)(1-\delta)$. We can choose

$$\tau_2 \leq \min\{1, B_{2,\boldsymbol{K}}, \mathcal{B}_{1,\boldsymbol{K}}\}/(\sqrt{m(N+1)}B_4) \Rightarrow \|\boldsymbol{K}' - \boldsymbol{K}\|_F \leq \min\{1, B_2, B_1\}.$$

This ensures $\boldsymbol{K}' \in \mathcal{K}$ with probability at least $(1-\delta_1)(1-\delta)$ by applying Lemma A.7. Recall from Lemma A.15,

$$\boldsymbol{e}_{1,\boldsymbol{K},\boldsymbol{K}'} \succeq 0, \quad \boldsymbol{e}_{2,\boldsymbol{K},\boldsymbol{K}'} \succeq 0, \quad \boldsymbol{e}_{3,\boldsymbol{K},\boldsymbol{K}'} \succeq 0.$$

Then via Lemma A.30, we have

$$\begin{aligned}
\boldsymbol{e}_{1,\boldsymbol{K},\boldsymbol{K}'} &\preceq (4\tau_2 + 4\tau_2^2 \|\boldsymbol{G}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})}\|) \|\boldsymbol{F}^r_{\boldsymbol{K},\boldsymbol{L}} - \boldsymbol{F}_{\boldsymbol{K},\boldsymbol{L}}\|^2 \cdot I, \\
\boldsymbol{e}_{2,\boldsymbol{K},\boldsymbol{K}'} &\preceq \tau_2 \|\boldsymbol{F}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})} - \boldsymbol{F}_{\boldsymbol{K},\boldsymbol{L}}\|^2 \cdot I, \\
\boldsymbol{e}_{3,\boldsymbol{K},\boldsymbol{K}'} &\preceq (2\tau_2 + \tau_2^2 \|\boldsymbol{G}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})}\|) \|V(\widetilde{\boldsymbol{F}}_{\boldsymbol{K},\boldsymbol{L}})\| \cdot I.
\end{aligned}$$

To bound these errors, we apply Lemma A.12 and choose r, ε_1 such that

$$r_2 \leq D_1, \quad \varepsilon_1 \leq D_3,$$

where D_1, D_3 are defined in Lemma A.6. Recall that $\mathbb{E}[\widetilde{\nabla}_{\boldsymbol{K}} \mathcal{G}(\boldsymbol{K},\boldsymbol{L})] = \nabla_{\boldsymbol{K}} \mathcal{G}_{r_2}(\boldsymbol{K},\boldsymbol{L}) = \nabla_{\boldsymbol{K}} \mathbb{E}_{\boldsymbol{V}}[\mathcal{G}(\boldsymbol{K} + r_2 \boldsymbol{V}, \boldsymbol{L})]$ and $\boldsymbol{F}^r_{\boldsymbol{K},\boldsymbol{L}} = \nabla_{\boldsymbol{K}} \mathcal{G}_{r_2}(\boldsymbol{K},\boldsymbol{L}) \boldsymbol{\Sigma}^{-1}_{\boldsymbol{K},\boldsymbol{L}}$. Then we have

$$\begin{aligned}
\boldsymbol{e}_{1,\boldsymbol{K},\boldsymbol{K}'} &\preceq (2\tau_2 + 2\tau_2^2 \|\boldsymbol{G}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})}\|) \|\nabla_{\boldsymbol{K}} \mathcal{G}_{r_2}(\boldsymbol{K},\boldsymbol{L}) \boldsymbol{\Sigma}^{-1}_{\boldsymbol{K},\boldsymbol{L}} - \nabla_{\boldsymbol{K}} \mathcal{G}(\boldsymbol{K},\boldsymbol{L}) \boldsymbol{\Sigma}^{-1}_{\boldsymbol{K},\boldsymbol{L}}\|^2 \cdot I \\
&\preceq (4\tau_2 + 4\tau_2^2 \|\boldsymbol{G}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})}\|)/\varphi^2 \cdot (l_1 \|\boldsymbol{\Sigma}_{\boldsymbol{K},\boldsymbol{L}}\| + l_1 \cdot l_7 + l_7 \|\boldsymbol{F}_{\boldsymbol{K},\boldsymbol{L}}\|)^2 r^2 \cdot I \\
\boldsymbol{e}_{2,\boldsymbol{K},\boldsymbol{K}'} &\preceq \tau_2 \|\boldsymbol{F}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})} - 2\boldsymbol{F}_{\boldsymbol{K},\boldsymbol{L}}\|^2 \cdot I \preceq \tau_2 (l_2)^2 \cdot H\varepsilon_1 \cdot I \\
\Rightarrow \|\boldsymbol{e}_{\boldsymbol{K},\boldsymbol{K}'}\| &\leq (4\tau_2 + 4\tau_2^2 \|\boldsymbol{G}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})}\|)/\varphi^2 \cdot (l_1 \|\boldsymbol{\Sigma}_{\boldsymbol{K},\boldsymbol{L}}\| + l_1 \cdot l_7 + l_7 \|\boldsymbol{F}_{\boldsymbol{K},\boldsymbol{L}}\|)^2 r^2 \\
&\quad + \tau_2 (l_2)^2 \cdot H\varepsilon_1 + (2\tau_2 + \tau_2^2 \|\boldsymbol{G}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})}\|) \|V(\widetilde{\boldsymbol{F}}_{\boldsymbol{K},\boldsymbol{L}})\|,
\end{aligned}$$

holds with probability at least $(1-\delta_1)(1-\delta)$ where $\boldsymbol{e}_{\boldsymbol{K},\boldsymbol{K}'} = \boldsymbol{e}_{1,\boldsymbol{K},\boldsymbol{K}'} + \boldsymbol{e}_{2,\boldsymbol{K},\boldsymbol{K}'} + \boldsymbol{e}_{3,\boldsymbol{K},\boldsymbol{K}'}$. Moreover

$$
\begin{aligned}
\sum_{t=0}^{N} (\boldsymbol{A}_{\boldsymbol{K}',\boldsymbol{L}(\boldsymbol{K}')}^\top)^t \boldsymbol{e}_{\boldsymbol{K},\boldsymbol{K}'} (\boldsymbol{A}_{\boldsymbol{K}',\boldsymbol{L}(\boldsymbol{K}')})^t &\preceq \|\boldsymbol{e}_{\boldsymbol{K},\boldsymbol{K}'}\| \sum_{t=0}^{N} (\boldsymbol{A}_{\boldsymbol{K}',\boldsymbol{L}(\boldsymbol{K}')}^\top)^t (\boldsymbol{A}_{\boldsymbol{K}',\boldsymbol{L}(\boldsymbol{K}')})^t \\
&\preceq \frac{\|\boldsymbol{e}_{\boldsymbol{K},\boldsymbol{K}'}\|}{\varphi} \sum_{t=0}^{N} (\boldsymbol{A}_{\boldsymbol{K}',\boldsymbol{L}(\boldsymbol{K}')}^\top)^t \boldsymbol{\Sigma}_0 (\boldsymbol{A}_{\boldsymbol{K}',\boldsymbol{L}(\boldsymbol{K}')})^t \\
&\preceq \frac{\|\boldsymbol{e}_{\boldsymbol{K},\boldsymbol{K}'}\|}{\varphi} \left\| \sum_{t=0}^{N} (\boldsymbol{A}_{\boldsymbol{K}',\boldsymbol{L}(\boldsymbol{K}')}^\top)^t \boldsymbol{\Sigma}_0 (\boldsymbol{A}_{\boldsymbol{K}',\boldsymbol{L}(\boldsymbol{K}')})^t \right\| \cdot I \\
&= \frac{\|\boldsymbol{e}_{\boldsymbol{K},\boldsymbol{K}'}\|}{\varphi} \left\| \sum_{t=0}^{N} (\boldsymbol{A}_{\boldsymbol{K}',\boldsymbol{L}(\boldsymbol{K}')})^t \boldsymbol{\Sigma}_0 (\boldsymbol{A}_{\boldsymbol{K}',\boldsymbol{L}(\boldsymbol{K}')}^\top)^t \right\| \cdot I \\
&\preceq \|\boldsymbol{e}_{\boldsymbol{K},\boldsymbol{K}'}\| (\|\boldsymbol{\Sigma}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})}\|_F + B_{\boldsymbol{\Sigma}} \|\boldsymbol{K}' - \boldsymbol{K}\|_F) / \varphi \cdot I.
\end{aligned}
$$

In the first and second inequality, we apply Lemma A.27 and Lemma A.30. For the third inequality, we apply Lemma A.25. In the first equation, we apply the fact that $\|AA^\top\| = \|A^\top A\|$. In the fourth inequality, we apply Lemma A.8 and require $\|\boldsymbol{K}' - \boldsymbol{K}\|_F \leq B_1$. Hence, we obtain the following inequality holds with probability at least $(1-\delta_1)(1-\delta)$

$$
\begin{aligned}
\boldsymbol{P}_{\boldsymbol{K}',\boldsymbol{L}(\boldsymbol{K}')} - \boldsymbol{P}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})} &\preceq \sum_{t=0}^{N} (\boldsymbol{A}_{\boldsymbol{K}',\boldsymbol{L}(\boldsymbol{K}')}^\top)^t \boldsymbol{e}_{\boldsymbol{K},\boldsymbol{K}'} (\boldsymbol{A}_{\boldsymbol{K}',\boldsymbol{L}(\boldsymbol{K}')})^t - \frac{\tau_2}{4} \boldsymbol{F}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})}^\top \boldsymbol{F}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})} \\
&\preceq \Big((4\tau_2 + 4\tau_2^2 \|\boldsymbol{G}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})}\|) / \varphi^2 \cdot (l_1 \|\boldsymbol{\Sigma}_{\boldsymbol{K},\boldsymbol{L}}\| + l_1 \cdot l_7 + l_7 \|\boldsymbol{F}_{\boldsymbol{K},\boldsymbol{L}}\|)^2 r^2 \\
&\quad + \tau_2 (l_2)^2 \cdot H\varepsilon_1 + (2\tau_2 + \tau_2^2 \|\boldsymbol{G}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})}\|) \|V(\widetilde{\boldsymbol{F}}_{\boldsymbol{K},\boldsymbol{L}})\| \Big) \\
&\quad \cdot (\|\boldsymbol{\Sigma}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})}\|_F + \mathcal{B}_{\boldsymbol{\Sigma}}) / \varphi \cdot I - \frac{\tau_2}{4} \boldsymbol{F}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})}^\top \boldsymbol{F}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})} \\
&\preceq \tau_2 \cdot (c_1 \cdot r_2^2 + c_2 \cdot \varepsilon_1 + c_3 \cdot \|V(\widetilde{\boldsymbol{F}}_{\boldsymbol{K},\boldsymbol{L}})\|) - \frac{\tau_2}{4} \boldsymbol{F}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})}^\top \boldsymbol{F}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})}.
\end{aligned}
$$

where the first inequality is the descent-like inequality in Lemma A.15. The positive constants are defined as

$$
\begin{aligned}
c_1 &:= \sup_{(\boldsymbol{K},\boldsymbol{L}) \in \hat{\mathcal{K}} \times \hat{\mathcal{L}}} (4 + 4G) / \varphi^3 \cdot (l_1 \|\boldsymbol{\Sigma}_{\boldsymbol{K},\boldsymbol{L}}\| + l_1 \cdot l_7 + l_7 \|\boldsymbol{F}_{\boldsymbol{K},\boldsymbol{L}}\|)^2 \cdot (\|\boldsymbol{\Sigma}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})}\|_F + \mathcal{B}_{\boldsymbol{\Sigma}}), \\
c_2 &:= \sup_{\boldsymbol{K} \in \hat{\mathcal{K}}} (l_2)^2 \cdot H(\|\boldsymbol{\Sigma}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})}\|_F + \mathcal{B}_{\boldsymbol{\Sigma}}) / \varphi, \quad c_3 := \sup_{\boldsymbol{K} \in \hat{\mathcal{K}}} (2 + G) \cdot (\|\boldsymbol{\Sigma}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})}\|_F + \mathcal{B}_{\boldsymbol{\Sigma}}) / \varphi.
\end{aligned}
$$

where G, H, φ are defined in Appendix A.1. □

A.9 Useful Technical Lemma

Basic Results of LQ Problems

Lemma A.21. *(Dual Lyapunov equations) Let matrix A be Schur stable, and X be the solution to the Lyapunov equation*

$$A^\top X A + W = X.$$

Let Y be the solution to the dual Lyapunov equation

$$A Y A^\top + V = Y.$$

Then $\mathrm{Tr}(XV) = \mathrm{Tr}(YW)$.

Proof. The solutions to these two Lyapunov equations satisfy

$$X = \sum_{i=0}^{\infty}(A^\top)^i W A^i, \quad Y = \sum_{i=0}^{\infty} A^i W (A^\top)^i,$$
$$\mathrm{Tr}(XV) = \mathrm{Tr}(\sum_{i=0}^{\infty}(A^\top)^i W A^i V) = \mathrm{Tr}(\sum_{i=0}^{\infty} W A^i V (A^\top)^i) = \mathrm{Tr}(WY) = \mathrm{Tr}(YW).$$

□

Lemma A.22. *The solution to the Lyapunov equation*

$$\boldsymbol{X} = \boldsymbol{A}_{\boldsymbol{K},\boldsymbol{L}}^\top \boldsymbol{X} \boldsymbol{A}_{\boldsymbol{K},\boldsymbol{L}} + \boldsymbol{Z},$$

is unique and has the explicit expression

$$\boldsymbol{X} = \sum_{i=0}^{N}(\boldsymbol{A}_{\boldsymbol{K},\boldsymbol{L}}^\top)^i \boldsymbol{Z} (\boldsymbol{A}_{\boldsymbol{K},\boldsymbol{L}})^i,$$

for any $\boldsymbol{K}, \boldsymbol{L}$ defined in (1.4), (1.5), *$\boldsymbol{Z}$ is an arbitrary real matrix with proper dimensions.*

Proof. First of all, we can easily verify that the explicit expression is a solution to the Lyapunov equation. Then assume we have two different solutions $\boldsymbol{X}_1, \boldsymbol{X}_2$

$$\boldsymbol{X}_1 = \boldsymbol{A}_{\boldsymbol{K},\boldsymbol{L}}^\top \boldsymbol{X}_1 \boldsymbol{A}_{\boldsymbol{K},\boldsymbol{L}} + \boldsymbol{Z}, \quad \boldsymbol{X}_2 = \boldsymbol{A}_{\boldsymbol{K},\boldsymbol{L}}^\top \boldsymbol{X}_2 \boldsymbol{A}_{\boldsymbol{K},\boldsymbol{L}} + \boldsymbol{Z}.$$

Then we know

$$\boldsymbol{X}_1 - \boldsymbol{X}_2 = \boldsymbol{A}_{\boldsymbol{K},\boldsymbol{L}}^\top (\boldsymbol{X}_1 - \boldsymbol{X}_2) \boldsymbol{A}_{\boldsymbol{K},\boldsymbol{L}}.$$

Then, iteratively

$$\boldsymbol{X}_1 - \boldsymbol{X}_2 = (\boldsymbol{A}_{\boldsymbol{K},\boldsymbol{L}}^\top)^{N+2} (\boldsymbol{X}_1 - \boldsymbol{X}_2) (\boldsymbol{A}_{\boldsymbol{K},\boldsymbol{L}})^{N+2} = 0.$$

Here we observe that for any $\boldsymbol{K}$, $\boldsymbol{L}$ with the structure defined in (1.4), (1.5), we have

$$\boldsymbol{A}_{\boldsymbol{K},\boldsymbol{L}}^{N+1} = \boldsymbol{0}.$$

To see this, we can easily compute $\boldsymbol{A}_{\boldsymbol{K},\boldsymbol{L}}$ and observe that

$$\begin{aligned}\boldsymbol{A}_{\boldsymbol{K},\boldsymbol{L}} &= \begin{bmatrix} 0_{m\times mN} & 0_{m\times m} \\ \mathrm{diag}(A_{0-(N-1)}) & 0_{mN\times m}\end{bmatrix} + \begin{bmatrix} 0_{m\times dN} \\ \mathrm{diag}(B_0 - (N-1))\end{bmatrix} \begin{bmatrix}\mathrm{diag}(K_{0-(N-1)}) & 0_{dN\times m}\end{bmatrix} \\ &\quad - \begin{bmatrix} 0_{m\times nN} \\ \mathrm{diag}(D_{0-(N-1)})\end{bmatrix} \begin{bmatrix}\mathrm{diag}(L_{0-(N-1)}) & 0_{nN\times m}\end{bmatrix} \\ &= \begin{bmatrix} 0_{m\times mN} & 0_{m\times m} \\ \mathrm{diag}(A_{K_0,L_0} - A_{K_{N-1},L_{N-1}}) & 0_{mN\times m}\end{bmatrix}.\end{aligned}$$

Note $A_{K_h,L_h} \in \mathbb{R}^{m\times m}$ and hence diagonal entries of $\boldsymbol{A}_{\boldsymbol{K},\boldsymbol{L}}$ are all zeros. Then we conclude that $\boldsymbol{A}_{\boldsymbol{K},\boldsymbol{L}}$ is nilpotent.

Therefore, we have $\boldsymbol{X}_1 = \boldsymbol{X}_2$ which contradicts our assumption. Hence the solution is unique and has the explicit expression above. We can easily see the same result holds for $\boldsymbol{A}_{\boldsymbol{K},\boldsymbol{L}}^\top$ by replacing $\boldsymbol{A}_{\boldsymbol{K},\boldsymbol{L}}$ with $\boldsymbol{A}_{\boldsymbol{K},\boldsymbol{L}}^\top$ in the above proof. □

Lemma A.23. *Let $Q_1 \succ Q_2$ and X_1, X_2 be solutions to the solutions to Lyapunov equations:*

$$X_1 = A^\top X_1 A + Q_1, \quad X_2 = A^\top X_2 A + Q_2.$$

where A is stable. Then $X_1 \succ X_2$.

Lemma A.24. *(Value matrix difference Zhang et al. [2019]) For any $\boldsymbol{K}, \boldsymbol{K}' \in \mathcal{K}$, $\boldsymbol{L}, \boldsymbol{L}'$ defined in (1.4), (1.5), we have the following equation*

$$\begin{aligned}\boldsymbol{P}_{\boldsymbol{K}',\boldsymbol{L}'} - \boldsymbol{P}_{\boldsymbol{K},\boldsymbol{L}} &= (\boldsymbol{A} - \boldsymbol{B}\boldsymbol{K}' - \boldsymbol{D}\boldsymbol{L}')^\top (\boldsymbol{P}_{\boldsymbol{K}',\boldsymbol{L}'} - \boldsymbol{P}_{\boldsymbol{K},\boldsymbol{L}})(\boldsymbol{A} - \boldsymbol{B}\boldsymbol{K}' - \boldsymbol{D}\boldsymbol{L}') \\ &+ (\boldsymbol{K}' - \boldsymbol{K})^\top \boldsymbol{F}_{\boldsymbol{K},\boldsymbol{L}} + \boldsymbol{F}_{\boldsymbol{K},\boldsymbol{L}}^\top (\boldsymbol{K}' - \boldsymbol{K}) + (\boldsymbol{K}' - \boldsymbol{K})^\top (\boldsymbol{R}^u + \boldsymbol{B}^\top \boldsymbol{P}_{\boldsymbol{K},\boldsymbol{L}} \boldsymbol{B})(\boldsymbol{K}' - \boldsymbol{K}) \\ &+ (\boldsymbol{L}' - \boldsymbol{L})^\top \boldsymbol{E}_{\boldsymbol{K},\boldsymbol{L}} + \boldsymbol{E}_{\boldsymbol{K},\boldsymbol{L}}^\top (\boldsymbol{L}' - \boldsymbol{L}) + (\boldsymbol{L}' - \boldsymbol{L})^\top (-\boldsymbol{R}^w + \boldsymbol{D}^\top \boldsymbol{P}_{\boldsymbol{K},\boldsymbol{L}} \boldsymbol{D})(\boldsymbol{L}' - \boldsymbol{L}) \\ &+ (\boldsymbol{L}' - \boldsymbol{L})^\top \boldsymbol{D}^\top \boldsymbol{P}_{\boldsymbol{K},\boldsymbol{L}} \boldsymbol{B}(\boldsymbol{K}' - \boldsymbol{K}) + (\boldsymbol{K}' - \boldsymbol{K})^\top \boldsymbol{B}^\top \boldsymbol{P}_{\boldsymbol{K},\boldsymbol{L}} \boldsymbol{D}(\boldsymbol{L}' - \boldsymbol{L}).\end{aligned}$$

Proof. The proof can be easily obtained by subtracting two Lyapunov equations and basic algebraic computations.

$$\begin{aligned}\boldsymbol{P}_{\boldsymbol{K}',\boldsymbol{L}'} &= (\boldsymbol{A} - \boldsymbol{B}\boldsymbol{K}' - \boldsymbol{D}\boldsymbol{L}')^\top \boldsymbol{P}_{\boldsymbol{K}',\boldsymbol{L}'} (\boldsymbol{A} - \boldsymbol{B}\boldsymbol{K}' - \boldsymbol{D}\boldsymbol{L}') + \boldsymbol{Q} + (\boldsymbol{K}')^\top \boldsymbol{R}^u \boldsymbol{K}' - (\boldsymbol{L}')^\top \boldsymbol{R}^w \boldsymbol{L}', \\ \boldsymbol{P}_{\boldsymbol{K},\boldsymbol{L}} &= (\boldsymbol{A} - \boldsymbol{B}\boldsymbol{K} - \boldsymbol{D}\boldsymbol{L})^\top \boldsymbol{P}_{\boldsymbol{K},\boldsymbol{L}} (\boldsymbol{A} - \boldsymbol{B}\boldsymbol{K} - \boldsymbol{D}\boldsymbol{L}) + \boldsymbol{Q} + \boldsymbol{K}^\top \boldsymbol{R}^u \boldsymbol{K} - \boldsymbol{L}^\top \boldsymbol{R}^w \boldsymbol{L}.\end{aligned}$$

□

Lemma A.25. *(Preliminary Lemma for Bounded Perturbation) For any control pair $(\boldsymbol{K}, \boldsymbol{L})$ defined in (1.4), (1.5), let $\boldsymbol{H}, \boldsymbol{H}'$ be the solution of the following Lyapunov equation**

$$I + \boldsymbol{A}_{\boldsymbol{K},\boldsymbol{L}}^\top \boldsymbol{H} \boldsymbol{A}_{\boldsymbol{K},\boldsymbol{L}} = \boldsymbol{H}, \tag{A.9}$$
$$I + \boldsymbol{A}_{\boldsymbol{K},\boldsymbol{L}} \boldsymbol{H}' \boldsymbol{A}_{\boldsymbol{K},\boldsymbol{L}}^\top = \boldsymbol{H}'. \tag{A.10}$$

we know

$$\|\boldsymbol{H}\| = \|\boldsymbol{H}'\| \leq \mathrm{Tr}(\boldsymbol{\Sigma}_{\boldsymbol{K},\boldsymbol{L}})/\varphi.$$

Proof. Then apply Lemma A.21 and consider the following dual Lyapunov equation of (A.9) with solution $\boldsymbol{\Sigma}_{\boldsymbol{K},\boldsymbol{L}}$

$$\boldsymbol{\Sigma}_{\boldsymbol{K},\boldsymbol{L}} = \boldsymbol{A}_{\boldsymbol{K},\boldsymbol{L}} \boldsymbol{\Sigma}_{\boldsymbol{K},\boldsymbol{L}} \boldsymbol{A}_{\boldsymbol{K},\boldsymbol{L}}^\top + \boldsymbol{\Sigma}_0.$$

Apply Lemma A.21, we have

$$\mathrm{Tr}(\boldsymbol{\Sigma}_{\boldsymbol{K},\boldsymbol{L}}) = \mathrm{Tr}(\boldsymbol{H}\boldsymbol{\Sigma}_0) \geq \lambda_{\min}(\boldsymbol{\Sigma}_0) \cdot \mathrm{Tr}(\boldsymbol{H}) \geq \lambda_{\min}(\boldsymbol{\Sigma}_0) \cdot \|\boldsymbol{H}\|_F \geq \lambda_{\min}(\boldsymbol{\Sigma}_0) \cdot \|\boldsymbol{H}\|.$$

In the first inequality, we apply Lemma A.29. Since $\|AA^\top\| = \|A^\top A\|$, the proof for H' is immediate. □

Sensitivity Analysis for Stable Discrete-time Lyapunov Equations With the stability of Lyapunov equations discussed in Remark 1.2, we can always apply the result of the sensitivity analysis in Gahinet et al. [1990] to our case for the local Lipschitz continuity of our objective function. Here we include this result for the sake of completeness. For any stable F and discrete-time Lyapunov equation,

$$M = X - F^\top X F, \tag{A.11}$$

we have the following sensitivity analysis result for Lyapunov equation A.11. We respectively define the norms of an arbitrary linear operator $\Theta : \mathbb{R}^{n\times n} \to \mathbb{R}^{n\times n}$ as

$$\|\Theta\|_F = \max_{M\in\mathbb{R}^{n\times n}, \|M\|_F=1} \|\Theta(M)\|_F, \quad \|\Theta\| = \max_{M\in\mathbb{R}^{n\times n}, \|M\|=1} \|\Theta(M)\|.$$

*The solutions of the above Lyapunov equations uniquely exist by applying Lemma A.22. Hence $\boldsymbol{H}, \boldsymbol{H}'$ are well-defined.

Lemma A.26. *Consider two stable Lyapunov equations that admit unique solutions*

$$\begin{aligned} M &= X - F^\top X F, \\ M + \Delta M &= X + \Delta X - (F + \Delta F)^\top (X + \Delta X)(F + \Delta F). \end{aligned}$$

If $\|U\|\|\Delta F\|(2\|F\| + \|\Delta F\|) < 1$, *then we have*

$$\|\Delta X\| \leq \left(1 - \|U\|\|\Delta F\|(2\|F\| + \|\Delta F\|)\right)^{-1} \cdot \left(\|U\|\|\Delta M\| + \|U\|\|\Delta F\|(2\|F\| + \|\Delta F\|)\|X\|\right),$$

where U *is the unique solution of* $X - F^\top U F = I$.

This result will become useful when we try to control the error caused by the estimated output of the inner-loop oracle.

Proof. When F is stable, we know that the discrete-time Lyapunov operator $\Omega_F(X) := X - F^\top X F$ is a nonsingular linear operator Gahinet et al. [1990]. And the Lyapunov equation has a unique solution

$$\Omega_F(X) = M, \quad X = \sum_{k=0}^{+\infty} (F^\top)^k M F^k,$$

where M is an arbitrary matrix with proper dimensions. When perturbed F, i.e., $F + \Delta F$ is also stable, assume $X + \Delta X$ is the solution of the Lyapunov equation below

$$\Omega_{F+\Delta F}(X + \Delta X) = M + \Delta M.$$

We apply Lemma 2.3 in Gahinet et al. [1990] and obtain

$$\|\Delta X\| \leq \|\Omega_F^{-1}\|(\|\Delta M\| + \|\Delta \Omega\|\|X + \Delta X\|),$$

where $\Delta\Omega = \Omega_{F+\Delta F} - \Omega_F$. Then we apply Lemma 2.4 in Gahinet et al. [1990] to bound $\|\Delta\Omega\|$ with

$$\|\Delta\Omega\| \leq \|\Delta F\|(2\|F\| + \|\Delta F\|).$$

Note that the above result holds both for the Frobenius norm and the spectral norm. Finally, we apply Theorem 4.1 in Gahinet et al. [1990] which says $\|\Omega^{-1}\| = \|U\|$ and obtain

$$\begin{aligned} \|\Delta X\| &\leq \|U\|\left(\|\Delta M\| + \|\Delta F\|(2\|F\| + \|\Delta F\|)\|X + \Delta X\|\right) \\ &\leq \|U\|\left(\|\Delta M\| + \|\Delta F\|(2\|F\| + \|\Delta F\|)(\|X\| + \|\Delta X\|)\right). \end{aligned}$$

Hence when $\|U\|\|\Delta F\|(2\|F\| + \|\Delta F\|) < 1$, we have

$$\|\Delta X\| \leq \left(1 - \|U\|\|\Delta F\|(2\|F\| + \|\Delta F\|)\right)^{-1} \cdot \|U\|\left(\|\Delta M\| + \|\Delta F\|(2\|F\| + \|\Delta F\|)\|X\|\right).$$

Our proof slightly adapts the proof of Theorem 2.6 in Gahinet et al. [1990] and removes their assumption that $M + \Delta M \neq 0$. ☐

Matrix inequalities This subsection summarizes some basic matrix inequalities used in our proofs. Some proofs of well-known results are omitted and can be found in Horn and Johnson [2012].

Lemma A.27. *For any matrix M, N such that $M \succeq N$, for any real matrix A with proper dimensions, we have*

$$A^T M A \succeq A^T N A.$$

Lemma A.28. *For any real matrix $M \in \mathbb{R}^{d_1 \times d_2}$, if we know $\|M\| \leq c$. Then we have*

$$M^T M \preceq c^2 \cdot I_{d_2}, \quad MM^T \preceq c^2 \cdot I_{d_1}.$$

where I_{d_2} denotes the identity matrix of dimension d_2 and I_{d_1} denotes the d_1 dimension identity matrix.

Proof. Since both $M^T M$ and MM^T are positive semi-definite matrices, and for any positive semi-definite matrix A we know that $A \preceq \|A\| \cdot I$, and hence

$$\begin{aligned} M^T M &\preceq \|M^T M\| \cdot I \preceq \|M\|^2 \cdot I \preceq c^2 \cdot I \\ MM^T &\preceq \|MM^T\| \cdot I \preceq \|M\|^2 \cdot I \preceq c^2 \cdot I \end{aligned}$$

□

Lemma A.29. *For any positive semi-definite matrices M, N with proper dimensions, we have*

$$\lambda_{\min}(N) \cdot \mathrm{Tr}(M) \leq \mathrm{Tr}(MN) \leq \lambda_{max}(N) \cdot \mathrm{Tr}(M)$$

Lemma A.30. *For any real and symmetric matrix M, if we know $\|M\| \leq c$ where c is a positive constant, then we have*

$$-c \cdot I \preceq M \preceq c \cdot I$$

Proof. We know that any real symmetric matrix is similar to a diagonal matrix D with diagonal elements being the eigenvalues of M. Moreover, we know that $\max_i |\lambda_i(M)| \leq c$. Then

$$\begin{aligned} M + c \cdot I &= PDP^T + c \cdot PP^T = P(D + c \cdot I)P^T \succeq 0 \\ M - c \cdot I &= PDP^T - c \cdot PP^T = P(D - c \cdot I)P^T \preceq 0 \end{aligned}$$

where P is an orthogonal matrix. □

Lemma A.31. *For any matrix M, N with proper dimensions, we have*

$$M^T N + N^T M \preceq \gamma M^T M + \gamma^{-1} N^T N$$

where $\gamma > 0$ is an arbitrary constant

Lemma A.32. *(Trace and norms) For any matrix $M \succeq 0$, we have*

$$\mathrm{Tr}(M) \geq \|M\|_F \geq \|M\|$$

A.10 Benchmark Algorithm

In this section, we repeat Algorithm 2 of Zhang et al. [2021b] for completeness. Different parts are colored in red for easier comparison.

Algorithm 3 (Algorithm 2 of Zhang et al. [2021b]) Benchmark Outer-loop Nested Natural Policy Gradient Algorithm

Input: $\boldsymbol{K}_0 \in \mathcal{K}$, number of iterations T, sample size M_2, perturbation radius r_2, stepsize τ_2, horizon N, dimension $d_{\boldsymbol{K}} = dmN$.

Output: $\boldsymbol{K}_{\text{out}} = \boldsymbol{K}_i$ where $i \sim \text{Unif}(\{0, \cdots, T-1\})$.

1: **for** $t = 0, 1, \cdots, T$ **do**

2: Call Algorithm 1 to obtain $\boldsymbol{L}_t$.

3: **for** $m = 0, 1, \cdots, M_2 - 1$ **do**

4: Sample $\boldsymbol{K}_t^m = \boldsymbol{K}_t + r_2 \boldsymbol{V}_m$ where $\boldsymbol{V}_m$ is uniformly drawn from $\mathcal{S}_1$ with $\|\boldsymbol{V}_m\|_F = 1$.

5: Call Algorithm 1 to obtain $\widetilde{\boldsymbol{L}}(\boldsymbol{K}_t^m)$ such that $\mathcal{G}(\boldsymbol{K}_t^m, \widetilde{\boldsymbol{L}}(\boldsymbol{K}_t^m)) \geq \mathcal{G}(\boldsymbol{K}_t^m, \boldsymbol{L}(\boldsymbol{K}_t^m)) - \epsilon_1$.

6: Simulate a first trajectory using control $(\boldsymbol{K}_t^m, \widetilde{\boldsymbol{L}}(\boldsymbol{K}_t^m))$ for horizon N under one realization of noises $\boldsymbol{\xi}_m$ and collect the cost $\mathcal{G}_{\boldsymbol{\xi}_m}(\boldsymbol{K}_t^m, \widetilde{\boldsymbol{L}}(\boldsymbol{K}_t^m))$.

7: Simulate another independent trajectory using control $(\boldsymbol{K}_t, \boldsymbol{L}_t)$ for horizon N starting from $x_{0,m}$ and compute

$$\widetilde{\boldsymbol{\Sigma}}^m_{\boldsymbol{K}_t,\boldsymbol{L}_t} = \text{diag}\big(x_{0,m}x_{0,m}^\top, \cdots, x_{N,m}x_{N,m}^\top\big).$$

8: **end for**

9: Update $\boldsymbol{K}_{t+1} = \boldsymbol{K}_t - \tau_2 \widetilde{\nabla}_{\boldsymbol{K}} \mathcal{G}(\boldsymbol{K}_t, \boldsymbol{L}_t) \widetilde{\boldsymbol{\Sigma}}^{-1}_{\boldsymbol{K}_t,\boldsymbol{L}_t}$ where $\widetilde{\nabla}_{\boldsymbol{K}} \mathcal{G}(\boldsymbol{K}_t, \boldsymbol{L}_t)$ equals

$$\frac{1}{M_2} \sum_{m=0}^{M_2-1} \frac{d_{\boldsymbol{K}}}{r_2} \mathcal{G}_{\boldsymbol{\xi}_m}(\boldsymbol{K}_t^m, \widetilde{\boldsymbol{L}}(\boldsymbol{K}_t^m)) \boldsymbol{V}_m,$$

and $\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{K}_t,\boldsymbol{L}_t} = \frac{1}{M_2} \sum_{m=0}^{M_2-1} \widetilde{\boldsymbol{\Sigma}}^m_{\boldsymbol{K}_t,\boldsymbol{L}_t}$.

10: **end for**

Acknowledgements

I would like to express my heartfelt gratitude to the following individuals and groups who have played significant roles in the completion of this work and throughout my master's studies.

First and foremost, my most profound gratitude goes to my supervisors Prof. Dr. Niao He, Ilyas Fatkhullin, and Dr. Anas Barakat for their continuous guidance and invaluable insights. Without them, I wouldn't have been able to manage my thesis as well as my Ph.D. applications at the same time. Weekly hours-long discussions with them inspired the topic of my thesis and gradually shape it. I have been nourished so much by their supervision not only in terms of technical capacities but also in visions of research in general. It's been a wonderful journey to conduct my thesis and submit my first scientific paper to CDC under their mentorship. I'm also grateful to Dr. Siqi Zhang for mentoring my semester project within the Optimization and Decision Intelligence (ODI) Group which led to my thesis, and all the group members of for their constructive feedback and suggestions, which helped me refine and improve my work.

I would also like to express endless appreciation to Prof. Dr. Dr. Frank Schweitzer, Dr. Giacomo Vaccario, and Dr. Giona Casiraghi for their tremendous help during my semester project within the Chair of Systems Design and Ph.D. applications. Their understanding, patience, and guidance make the stressful and challenging process enjoyable. I also want to give big thanks to Prof. Dr. Dafeng Zuo for his kindness and support during my applications for doctoral positions.

My family and friends have been my pillars of strength, and I cannot thank my parents and family enough for their unconditional love and support. Tons of love to Fan Wang for your constant encouragement and comfort, how lucky I am to meet you at Zurich airport. Numerous love goes to Jingqi Li for inspiring me with your ambition and being there for me whenever I need it. Deep thanks and love go to my roommate Malvika Srivastava, who has been an incredibly caring elder sister and sincere friend to me during the past three years. Warmest gratitude and love to Siyu Bian for being my sunshine and rainbow. Immense love to Yabei Xia for always being there for me, I can always regain energy from your persistence against obstacles. Extraordinary love to Yingxue Yu for being my life mentor and supporting me with your insightful and wise opinions. My biggest thanks to all of my friends who helped and spent time with me. The year 2020 might not be the best year to join ETH, but your company, whether physical or mental, has made coming to Zurich one of the best decisions I ever made.