Derivative-Free Policy Optimization for Risk-Sensitive and Robust Control Design: Implicit Regularization and Sample Complexity

Anonymous Author(s)
Affiliation
Address
email

Abstract

Policy-based model-free reinforcement learning (RL) methods have shown great promise for continuous control applications. However, their performances on risk-sensitive/robust control tasks have not been fully understood, which has been generally considered to be one important open problem in the seminal work (Fazel et al., 2018). We make a step toward addressing this open problem, by providing the first sample complexity results for policy gradient (PG) methods in two fundamental risk-sensitive/robust control settings: the linear exponential quadratic Gaussian, and the linear-quadratic (LQ) disturbance attenuation problems. The optimization landscapes for these problems are by nature more challenging than that of the LQ regulator problem, due to lack of coercivity of their objective functions. To overcome this challenge, we obtain the first *implicit regularization* results for model-free PG methods, certifying that the controller *remains robust* during the learning process, which further lead to the sample complexity guarantees. As a by-product, our results also provide the first sample complexity of PG methods in two-player zero-sum LQ dynamic games, a baseline in multi-agent RL.

1 Introduction

Recent years have witnessed the rapid development of reinforcement learning (RL) methods in handling continuous control tasks [1, 2, 3]. Central to the success of RL are policy optimization (PO) methods, including policy gradient (PG) [4, 5, 6], actor-critic [7, 8], and other variants [9, 10]. Progress reported in the literature has clearly shown an increasing interest in understanding theoretical properties of PO methods for relatively simple baselines such as various linear control problems [11, 12, 13, 14, 15, 16, 17, 18, 19, 20]. However, the theory of model-free PO methods on *risk-sensitive/robust* control remains underdeveloped in the literature. Since risk-sensitivity and robustness are important issues for designing safety-critical systems, it is natural to bring up the questions of whether and how model-free PO methods would converge for these continuous control tasks.

Our work in this paper is motivated by the above concern, and studies the sample complexity of model-free PG methods on two important baseline problems in risk-sensitive/robust control, namely the linear exponential quadratic Gaussian (LEQG), and the linear quadratic (LQ) disturbance attenuation problems. The former covers a fundamental setting in risk-sensitive control, and the latter is an important baseline for robust control. Based on the well-known equivalence between these problems and LQ dynamic games [21, 22, 23, 24, 25], we develop a *unified* PO perspective for both. A common feature for the above two problems is that their optimization landscapes are by nature more challenging than that of the linear quadratic regulator (LQR) problem, and existing proof techniques for model-free PG methods [11, 16, 17, 18] are no longer effective due to lack

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

of coercivity of the objective functions. Specifically, when applying PG methods to LQR, the feasible set for the resultant constrained optimization problem is the set of all linear state-feedback controllers that *stabilize* the closed-loop dynamics. The objective function of LQR is coercive on this feasible set and serves as a barrier function itself [13], guaranteeing for the PG iterates to stay in the feasible set and converge to the globally optimal controller. For the LEQG and LQ disturbance attenuation problems, the risk-sensitivity/robustness conditions have reshaped the feasible set in a way that the objective function becomes non-coercive, i.e. the objective value can remain finite while approaching the boundary. The objective is no longer a barrier function, and new proof techniques are needed to show that model-free PG iterates will stay in the feasible set. This is significant for safety-critical control systems with model uncertainty, as the iterates' feasibility here is equivalent to *risk-sensitivity/robustness* of the controller (cf. Remark A.8), and the failure to preserve robustness during learning can cause catastrophic effects, e.g., destabilizing the system in face of disturbances.

The most relevant result was developed in [15], which proposes implicit regularization (IR) arguments to show the convergence of several PG methods on the mixed $\mathcal{H}_2/\mathcal{H}_\infty$ control design problem (which can be viewed as the infinite-horizon variant of the LQ disturbance attenuation problem studied in this paper). The main finding there is that two specific PG search directions (with perfect model information) are automatically biased towards the interior of the robustness-related feasible set. In [15], it is emphasized that IR is a feature of both the *problem* and the *algorithm*, contrasting to that the stability-preserving nature of PG methods for LQR problems is based on the barrier function property of the objective and hence is *algorithm-agnostic*. Although the idea of IR is relevant to PO problems with non-coercive objective functions, the arguments in [15] only apply to the setting with a *known* model, since they rely on a specialized perturbation technique which may potentially generate arbitrarily small "margins". However, in the model-free setting, a uniform margin is required for provable tolerance of statistical errors. See a detailed comparison of the literature in §A.1.

In this paper, for the LEQG and LQ disturbance attenuation problems, we overcome the above margin issue and obtain the first IR result in the model-free setting. This enables the first model-free PG method that provably solves these control problems with polynomial number of samples. We highlight our contributions as follows.

Contributions. We provide the first sample complexity results for model-free PG methods for solving linear control problems with risk-sensitivity/robustness concerns (the LEQG and LQ disturbance attenuation problems), which were viewed as important open problems in the seminal work [11]. From the robust control perspective, one feature of our algorithms is that, during the learning process, a certain level of *robustness/risk-sensitivity* of the controller is proved to be preserved. This has generalized the results in [15] with a known model, and has thus enabled the first finite-sample convergence guarantees of PO methods for risk-sensitive/robust control design. Our algorithms and sample complexity results also address two-player zero-sum LQ dynamic games in the finite-horizon time-varying setting, which are among the first sample complexity results for the global convergence of policy-based methods for competitive multi-agent RL. Second, in the context of minimax optimization, our results address a class of nonconvex-nonconcave minimax constrained optimization problems, using *zeroth-order* multi-step gradient descent-ascent methods. Finally, part of our results provide the first sample complexity analysis for PG methods that solve the finite-horizon time-varying LQR problem with system noises and a possibly *indefinite* state-weighting matrix.

Limitations and Broader Impacts. We believe that our results and insights for these fundamental LQ robust control settings help pave the way for future investigations into whether and how *model-free* PO methods converge for *general safety-critical* controller designs. We note that as an initial attempt, our complexity bounds are not optimized, and accelerated algorithms with sharper rates could be developed based on the intricate analysis we established in §4.

More importantly, our results bridge the gap between classical robust control concepts and modern policy-based RL theories, pushing forward the development of RL for *safety-critical* controller designs in the broader community of learning for control. Based on this robust control perspective, researchers could potentially improve the robustness of existing policy-based RL algorithms, and also develop new *model-free* RL algorithms with strong *safe on-the-fly* robustness guarantees.

Lastly, we do not believe our work will cause any potential negative societal impacts.

2 Background

We first introduce two classic settings in risk-sensitive and robust control, namely LEQG, and LQ disturbance attenuation, and their equivalence to zero-sum LQ dynamic games.

2.1 Linear Exponential Quadratic Gaussian

Consider the finite-horizon LEQG problem [21, 26, 27], with time-varying systems dynamics described by $x_{t+1} = A_t x_t + B_t u_t + w_t$, $t \in \{0, \cdots, N-1\}$, where $x_t \in \mathbb{R}^m$ represents the system state; $u_t \in \mathbb{R}^d$ is the control input; $w_t \in \mathbb{R}^m$ is an independent (across time) Gaussian random noise with $w_t \sim \mathcal{N}(\mathbf{0}, W)$ for some $W > 0$; the initial state $x_0 \sim \mathcal{N}(\mathbf{0}, X_0)$ for some $X_0 > 0$, which is independent of $\{w_t\}$; and A_t, B_t are time-varying system matrices. The objective function is given by

$$\mathcal{J}\big(\{u_t\}\big) := \frac{2}{\beta} \log \mathbb{E} \exp\Big[\frac{\beta}{2}\Big(\sum_{t=0}^{N-1} (x_t^\top Q_t x_t + u_t^\top R_t u_t) + x_N^\top Q_N x_N\Big)\Big], \tag{2.1}$$

where $Q_t, Q_N \geq 0$ and $R_t > 0$ are symmetric weighting matrices; and $\beta > 0$ is a parameter capturing the degree of risk-sensitivity, which is upper-bounded by some $\beta^* > 0$ [22, 27, 28].

The goal in the LEQG problem is to find the $\mathcal{J}$-minimizing optimal control policy $\mu_t^* : (\mathbb{R}^m \times \mathbb{R}^d)^t \times \mathbb{R}^m \to \mathbb{R}^d$ that maps, at time t, the history of state-control pairs up to time t and the current state x_t to the control u_t. It has been shown in [21] that μ_t^* has a linear state-feedback form $\mu_t^*(x_t) = -K_t^* x_t$, where $K_t^* \in \mathbb{R}^{d \times m}$, for all t. Therefore, it suffices to search K_t^* in the matrix space $\mathbb{R}^{d \times m}$ for all $t \in \{0, \cdots, N-1\}$, without losing any optimality. The resulting PO problem as well as its closed-form objective function and PG are provided in §A.3.1.

2.2 LQ Disturbance Attenuation

Second, consider the LQ disturbance attenuation problem, with time-varying dynamical systems described by $x_{t+1} = A_t x_t + B_t u_t + D_t w_t$ and $z_t = C_t x_t + E_t u_t$, where $x_t \in \mathbb{R}^m$ is the system state; $u_t \in \mathbb{R}^d$ is the control input; $w_t \in \mathbb{R}^n$ is the (unknown) disturbance input; $z_t \in \mathbb{R}^l$ is the controlled output; A_t, B_t, C_t, D_t, E_t are system matrices with appropriate dimensions; and $x_0 \in \mathbb{R}^m$ is unknown. In addition, we assume that $E_t^\top [C_t \; E_t] = [\mathbf{0} \; R_t]$ for some $R_t > 0$ for normalization, with no loss of generality (see §3.5.1 of [24] for a simple procedure that can transform the general problem into a form that satisfies this "normalization" assumption). Subsequently, we denote the ℓ^2-norms of the vectors $\omega := \big[x_0^\top C_0^\top, w_0^\top, \cdots, w_{N-1}^\top\big]^\top$ and $z := \big[z_0^\top, \cdots, z_{N-1}^\top, x_N^\top Q_N^{1/2}\big]^\top$ as $\|\omega\|$ and $\|z\| = \Big\{\mathcal{C}\big(\{u_t\}, \{w_t\}\big)\Big\}^{1/2}$, where $Q_N \geq 0$. Then, the robustness of a designed controller can be guaranteed by a constraint on the ratio between $\|z\|$ and $\|\omega\|$. Specifically, the goal of disturbance attenuation is the following: Given a $\gamma > \gamma^*$, where $\gamma^* = \sqrt{(\beta^*)^{-1}} > 0$ is the *optimal* (*minimax*) level of disturbance attenuation at the output and recall that β^* is the upper bound for the risk-sensitivity parameter in LEQG [22, 24], find a control policy $\mu_t = -K_t x_t$, $t \in \{0, \cdots, N-1\}$, that solves

$$\min_{\{\mu_t\}} \overline{\mathcal{J}}\big(\{\mu_t(x_t)\}\big) \qquad \text{s.t.} \quad \sup_{x_0, \{w_t\}} \frac{\Big\{\mathcal{C}\big(\{\mu_t(x_t)\}, \{w_t\}\big)\Big\}^{1/2}}{\|\omega\|} < \gamma, \tag{2.2}$$

where $\overline{\mathcal{J}}(\{\mu_t(x_t)\})$ is an upper bound of the LQG cost [29]. We defer the precise formulations of the LQ disturbance attenuation problem to §A.3.2. Also, see Remark A.3 for the challenges when addressing LEQG and LQ disturbance attenuation directly using derivative-free PG methods.

2.3 An Equivalent Dynamic Game Formulation

To overcome the challenges spelled out in Remark A.3, we now introduce an equivalent *dynamic game* formulation to the LEQG and the LQ disturbance attenuation problems due to that the natural PGs of the objective function of the game can be directly sampled (as we will show in §4). Under certain conditions to be introduced shortly, the saddle-point gain matrix of the minimizing player in the dynamic game (if exists) also addresses these two robust/risk-sensitive control problems. We aim to address these three classes of problems in a unified fashion.

Now, consider a two-player zero-sum LQ stochastic dynamic game (henceforth, game) with closed-loop perfect-state information pattern, characterized by $x_{t+1} = A_t x_t + B_t u_t + D_t w_t + \xi_t$, where

$x_t \in \mathbb{R}^m$ is the system state, and $u_t := -K_t x_t \in \mathbb{R}^d$ (resp., $w_t := -L_t x_t \in \mathbb{R}^n$) is the linear state-feedback controller of the minimizing (resp., maximizing) player[1], which suffices to attain the saddle point whenever exists [25]. The initial state x_0 and the additive process noises ξ_t are drawn independently from a distribution $\mathcal{D}$ and A_t, B_t, D_t are system matrices. Further, we assume that $\mathcal{D}$ is zero-mean, has a positive-definite covariance, and satisfies almost surely that $\|x_0\|, \|\xi_t\| \leq \vartheta$, for all $t \in \{0, \cdots, N-1\}$ and some constant ϑ. The goal of the minimizing (resp. maximizing) player is to minimize (resp. maximize) a quadratic objective function, namely to solve the game

$$\inf_{\{K_t\}} \sup_{\{L_t\}} \mathbb{E}_{x_0, \xi_0, \cdots, \xi_{N-1}} \left[\sum_{t=0}^{N-1} x_t^\top (Q_t + K_t^\top R_t^u K_t - L_t^\top R_t^w L_t) x_t + x_N^\top Q_N x_N \right], \tag{2.3}$$

where $Q_t, Q_N \geq 0$, $R_t^u, R_t^w > 0$ are symmetric weighting matrices. Whenever the solution to (2.3) exists such that the inf and sup can be interchanged, the value (2.3) is the *value* of the game, and the corresponding policies are the saddle-point/Nash equilibrium policies. To characterize the solution to (2.3), we first introduce the following time-varying Riccati difference equation (RDE):

$$P_t^* = Q_t + A_t^\top P_{t+1}^* \Lambda_t^{-1} A_t, \quad t \in \{0, \cdots, N-1\}, \tag{2.4}$$

where $\Lambda_t := \boldsymbol{I} + \left(B_t (R_t^u)^{-1} B_t^\top - D_t (R_t^w)^{-1} D_t^\top \right) P_{t+1}^*$ and $P_N^* = Q_N$. From [25], whenever a saddle point exists, the saddle-point control policies are linear state-feedback (i.e., $\mu_{u,t}^*(h_{u,t}) = -K_t^* x_t$ and $\mu_{w,t}^*(h_{w,t}) = -L_t^* x_t$), and the gain matrices $K_t^* = (R_t^u)^{-1} B_t^\top P_{t+1}^* \Lambda_t^{-1} A_t$ and $L_t^* = -(R_t^w)^{-1} D_t^\top P_{t+1}^* \Lambda_t^{-1} A_t$ are unique, where $P_{t+1}^* \geq 0$ is generated by (2.4). Then, we introduce a standard assumption that ensures the existence of the value of the game [24].

Assumption 2.1 $R_t^w - D_t^\top P_{t+1}^* D_t > 0$, *for all* $t \in \{0, \cdots, N-1\}$, *where* $P_{t+1}^* \geq 0$ *is generated by* (2.4).

Under Assumption 2.1, the value in (2.3) is attained by the sequence $\left(\{K_t^*\}, \{L_t^*\} \right)$, where $K_t^* \in \mathbb{R}^{d \times m}$ and $L_t^* \in \mathbb{R}^{n \times m}$ are gain matrices at the saddle-point/Nash equilibrium. Thus, we can replace the inf and sup in (2.3) with min and max, respectively. Some further notes on Assumption 2.1 are provided in Remark A.5. Lastly, we state the equivalences between three classes of problems, with its proof provided in §B.3.

Lemma 2.2 *(Connections) For any fixed* $\gamma > \gamma^*$ *in the LQ disturbance attenuation problem, we can introduce an equivalent LEQG problem and an equivalent zero-sum LQ dynamic game. Specifically, if we set* $\beta^{-1}\boldsymbol{I}, R_t, C_t^\top C_t, W$ *in LEQG,* $\gamma^2 \boldsymbol{I}, R_t, C_t^\top C_t, D_t D_t^\top$ *in the LQ disturbance attenuation problem, and* $R_t^w, R_t^u, Q_t, D_t D_t^\top$ *in the game to be the same, for all* $t \in \{0, \cdots, N-1\}$, *then the optimal gain matrices in LEQG, the gain matrix in the LQ disturbance attenuation problem, and the Nash equilibrium gain matrix for the minimizing player in the game are the same.*

By Lemma 2.2 and Remark A.3, we will hereafter focus on solving the game formulation in §2.3 using PG methods. As a result, the minimizing controller we obtain solves three classes of problems introduced above altogether.

3 Policy Gradient Methods

For ease of analysis, we define

$$\begin{aligned}
&\boldsymbol{x} = \left[x_0^\top, \cdots, x_N^\top\right]^\top, \ \boldsymbol{u} = \left[u_0^\top, \cdots, u_{N-1}^\top\right]^\top, \ \boldsymbol{w} = \left[w_0^\top, \cdots, w_{N-1}^\top\right]^\top, \ \boldsymbol{\xi} = \left[x_0^\top, \xi_0^\top, \cdots, \xi_{N-1}^\top\right]^\top, \ \boldsymbol{Q} = diag(Q_{0-N}), \\
&\boldsymbol{A} = \begin{bmatrix} \boldsymbol{0}_{m \times mN} & \boldsymbol{0}_{m \times m} \\ diag(A_{0-(N-1)}) & \boldsymbol{0}_{mN \times m} \end{bmatrix}, \ \boldsymbol{B} = \begin{bmatrix} \boldsymbol{0}_{m \times dN} \\ diag(B_{0-(N-1)}) \end{bmatrix}, \ \boldsymbol{D} = \begin{bmatrix} \boldsymbol{0}_{m \times nN} \\ diag(D_{0-(N-1)}) \end{bmatrix}, \ \boldsymbol{R}^u = diag(R_{0-(N-1)}^u), \\
&\boldsymbol{R}^w = diag(R_{0-(N-1)}^w), \quad \boldsymbol{K} = \begin{bmatrix} diag(K_{0-(N-1)}) & \boldsymbol{0}_{dN \times m} \end{bmatrix}, \quad \boldsymbol{L} = \begin{bmatrix} diag(L_{0-(N-1)}) & \boldsymbol{0}_{nN \times m} \end{bmatrix}.
\end{aligned} \tag{3.1}$$

where $diag(X_{0-N})$ denotes the block-diagonal matrix with $X_0, \cdots, X_N$ on the diagonal block entries. Some other notations are introduced in §A.2. Now, using the compact notations, we develop

[1] Hereafter we will use player and agent interchangeably. Also, restrictions to instantaneous linear state-feedback policies do not lead any loss of generality, as the results of §2.3 hold for general square-integrable policies, allowed to use the entire state history.

PG methods with exact PG accesses that provably converge to the Nash equilibrium of the game, $(\boldsymbol{K}^*, \boldsymbol{L}^*) \in \mathcal{S}(d,m,N) \times \mathcal{S}(n,m,N)$, where $\mathcal{S}(d,m,N)$ and $\mathcal{S}(n,m,N)$ are the subspaces that we confine our searches of $\boldsymbol{K}$ and $\boldsymbol{L}$ to, respectively. In particular, we only search over $\boldsymbol{K} \in \mathbb{R}^{Nd\times(N+1)m}$ and $\boldsymbol{L} \in \mathbb{R}^{Nn\times(N+1)m}$, and $\boldsymbol{K}, \boldsymbol{L}$ satisfy the sparsity patterns in (3.1), without losing any optimality. Then, for any $(\boldsymbol{K}, \boldsymbol{L})$, the objective function $\mathcal{G}(\boldsymbol{K}, \boldsymbol{L})$ is given by

$$\mathcal{G}(\boldsymbol{K}, \boldsymbol{L}) = \mathbb{E}_\xi\Big[\xi^\top \boldsymbol{P}_{\boldsymbol{K},\boldsymbol{L}} \xi\Big] = \mathrm{Tr}\Big(\boldsymbol{P}_{\boldsymbol{K},\boldsymbol{L}} \Sigma_0\Big) = \mathrm{Tr}\Big((\boldsymbol{Q} + \boldsymbol{K}^\top \boldsymbol{R}^u \boldsymbol{K} - \boldsymbol{L}^\top \boldsymbol{R}^w \boldsymbol{L})\Sigma_{\boldsymbol{K},\boldsymbol{L}}\Big), \tag{3.2}$$

where $\boldsymbol{P}_{\boldsymbol{K},\boldsymbol{L}} := diag(P_{K_0,L_0}, \cdots, P_{K_n,L_n})$ and P_{K_t,L_t}, $t \in \{0, \cdots, N-1\}$, solves the Lyapunov equation

$$P_{K_t,L_t} = (A_t - B_t K_t - D_t L_t)^\top P_{K_{t+1},L_{t+1}} (A_t - B_t K_t - D_t L_t) + Q_t + K_t^\top R_t^u K_t - L_t^\top R_t^w L_t, \tag{3.3}$$

with $P_{K_N,L_N} := Q_N$. Moreover, $\boldsymbol{P}_{\boldsymbol{K},\boldsymbol{L}}$, $\Sigma_{\boldsymbol{K},\boldsymbol{L}}$ are the solutions to the Lyapunov equations

$$\boldsymbol{P}_{\boldsymbol{K},\boldsymbol{L}} = (\boldsymbol{A} - \boldsymbol{B}\boldsymbol{K} - \boldsymbol{D}\boldsymbol{L})^\top \boldsymbol{P}_{\boldsymbol{K},\boldsymbol{L}} (\boldsymbol{A} - \boldsymbol{B}\boldsymbol{K} - \boldsymbol{D}\boldsymbol{L}) + \boldsymbol{Q} + \boldsymbol{K}^\top \boldsymbol{R}^u \boldsymbol{K} - \boldsymbol{L}^\top \boldsymbol{R}^w \boldsymbol{L}, \tag{3.4}$$

$$\Sigma_{\boldsymbol{K},\boldsymbol{L}} = (\boldsymbol{A} - \boldsymbol{B}\boldsymbol{K} - \boldsymbol{D}\boldsymbol{L})\Sigma_{\boldsymbol{K},\boldsymbol{L}} (\boldsymbol{A} - \boldsymbol{B}\boldsymbol{K} - \boldsymbol{D}\boldsymbol{L})^\top + \Sigma_0,\ \Sigma_0 := \mathbb{E}\Big[diag(x_0 x_0^\top, \xi_0 \xi_0^\top, \cdots, \xi_{N-1}\xi_{N-1}^\top)\Big] > 0, \tag{3.5}$$

where $\Sigma_0 > 0$ is full-rank because $x_0, \xi_0, \cdots, \xi_{N-1}$ are drawn independently from $\mathcal{D}$ which has a positive-definite covariance matrix. The solutions to (3.4) and (3.5) always exist and are unique because (3.4) and (3.5) are recursive formulas in blocks. Since $\|x_0\|, \|\xi_t\| \le \vartheta$ almost surely, for all t, $\Big\|diag(x_0 x_0^\top, \xi_0\xi_0^\top, \cdots, \xi_{N-1}\xi_{N-1}^\top)\Big\|_F \le (N+1)\vartheta^2$almost surely. Next, define the following notations:

$$c_0 := (N+1)\vartheta^2, \qquad d_\Sigma := m^2(N+1), \qquad \phi := \lambda_{\min}(\Sigma_0) > 0, \qquad \boldsymbol{H}_{\boldsymbol{K},\boldsymbol{L}} := \boldsymbol{R}^w - \boldsymbol{D}^\top \boldsymbol{P}_{\boldsymbol{K},\boldsymbol{L}} \boldsymbol{D},$$

$$\boldsymbol{E}_{\boldsymbol{K},\boldsymbol{L}} := (-\boldsymbol{R}^w + \boldsymbol{D}^\top \boldsymbol{P}_{\boldsymbol{K},\boldsymbol{L}} \boldsymbol{D})\boldsymbol{L} - \boldsymbol{D}^\top \boldsymbol{P}_{\boldsymbol{K},\boldsymbol{L}}(\boldsymbol{A} - \boldsymbol{B}\boldsymbol{K}), \quad \boldsymbol{F}_{\boldsymbol{K},\boldsymbol{L}} := (\boldsymbol{R}^u + \boldsymbol{B}^\top \boldsymbol{P}_{\boldsymbol{K},\boldsymbol{L}} \boldsymbol{B})\boldsymbol{K} - \boldsymbol{B}^\top \boldsymbol{P}_{\boldsymbol{K},\boldsymbol{L}}(\boldsymbol{A} - \boldsymbol{D}\boldsymbol{L}), \tag{3.6}$$

$$\boldsymbol{G}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})} := \boldsymbol{R}^u + \boldsymbol{B}^\top \widetilde{\boldsymbol{P}}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})} \boldsymbol{B}, \quad \widetilde{\boldsymbol{P}}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})} := \boldsymbol{P}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})} + \boldsymbol{P}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})} \boldsymbol{D}(\boldsymbol{R}^w - \boldsymbol{D}^\top \boldsymbol{P}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})} \boldsymbol{D})^{-1} \boldsymbol{D}^\top \boldsymbol{P}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})}. \tag{3.7}$$

Our goal is to solve the minimax optimization problem $\min_{\boldsymbol{K}} \max_{\boldsymbol{L}} \mathcal{G}(\boldsymbol{K}, \boldsymbol{L})$ such that $\mathcal{G}(\boldsymbol{K}^*, \boldsymbol{L}) \le \mathcal{G}(\boldsymbol{K}^*, \boldsymbol{L}^*) \le \mathcal{G}(\boldsymbol{K}, \boldsymbol{L}^*)$ for any $\boldsymbol{K}$ and $\boldsymbol{L}$. Some properties of $\mathcal{G}(\boldsymbol{K}, \boldsymbol{L})$ are presented in the following lemmas, with their proofs being deferred to §B.4 and §B.5, respectively.

Lemma 3.1 *(Nonconvexity-Nonconcavity) There exist zero-sum LQ dynamic games such that the objective function $\mathcal{G}(\boldsymbol{K}, \boldsymbol{L})$ is nonconcave in $\boldsymbol{L}$ for a fixed $\boldsymbol{K}$, and nonconvex in $\boldsymbol{K}$ for a fixed $\boldsymbol{L}$.*

Lemma 3.2 *(PG & No Spurious Local Minimum) The PGs of $\mathcal{G}(\boldsymbol{K}, \boldsymbol{L})$ can be computed as $\nabla_{\boldsymbol{K}} \mathcal{G}(\boldsymbol{K}, \boldsymbol{L}) = 2\boldsymbol{F}_{\boldsymbol{K},\boldsymbol{L}} \Sigma_{\boldsymbol{K},\boldsymbol{L}}$ and $\nabla_{\boldsymbol{L}} \mathcal{G}(\boldsymbol{K}, \boldsymbol{L}) = 2\boldsymbol{E}_{\boldsymbol{K},\boldsymbol{L}} \Sigma_{\boldsymbol{K},\boldsymbol{L}}$. Also, if at some stationary point $(\boldsymbol{K}, \boldsymbol{L})$ of $\mathcal{G}(\boldsymbol{K}, \boldsymbol{L})$ (i.e., where $\nabla_{\boldsymbol{K}} \mathcal{G}(\boldsymbol{K}, \boldsymbol{L}) = \boldsymbol{0}$ and $\nabla_{\boldsymbol{L}} \mathcal{G}(\boldsymbol{K}, \boldsymbol{L}) = \boldsymbol{0}$), it holds that $\boldsymbol{P}_{\boldsymbol{K},\boldsymbol{L}} \ge 0$ and $\boldsymbol{R}^w - \boldsymbol{D}^\top \boldsymbol{P}_{\boldsymbol{K},\boldsymbol{L}} \boldsymbol{D} > 0$, then $(\boldsymbol{K}, \boldsymbol{L})$ is the unique Nash equilibrium.*

3.1 Double-Loop Scheme

We introduce a *double-loop* update scheme. Specifically, we first fix $\boldsymbol{K}$, and solve for the optimal $\boldsymbol{L}$, denoted as $\boldsymbol{L}(\boldsymbol{K})$, whenever exists, by maximizing $\mathcal{G}(\boldsymbol{K}, \cdot)$ over $\boldsymbol{L} \in \mathcal{S}(n,m,N)$. Then, we update $\boldsymbol{K}$ to minimize $\mathcal{G}(\boldsymbol{K}, \boldsymbol{L}(\boldsymbol{K}))$ over $\boldsymbol{K} \in \mathcal{S}(d,m,N)$. For each fixed $\boldsymbol{K}$, the inner-loop is an *indefinite* LQR problem where $-\boldsymbol{Q} - \boldsymbol{K}^\top \boldsymbol{R}^u \boldsymbol{K}$ is not positive semi-definite (p.s.d.). Our goal is to find, via PG methods, the maximizing gain matrix $\boldsymbol{L}(\boldsymbol{K})$ whenever the objective function $\mathcal{G}(\boldsymbol{K}, \cdot)$ admits a finite upper bound value. We formally state the precise condition for the solution to the inner-loop maximization problem to be *well-defined* as follows, with its proof being deferred to §B.6.

Lemma 3.3 *(Inner-Loop Well-Definedness Condition) For the Riccati equation*

$$\boldsymbol{P}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})} = \boldsymbol{Q} + \boldsymbol{K}^\top \boldsymbol{R}^u \boldsymbol{K} + (\boldsymbol{A} - \boldsymbol{B}\boldsymbol{K})^\top \widetilde{\boldsymbol{P}}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})} (\boldsymbol{A} - \boldsymbol{B}\boldsymbol{K}), \tag{3.8}$$

define the following set of outer-loop gain matrices $\boldsymbol{K}$:

$$\mathcal{K} := \Big\{\boldsymbol{K} \in \mathcal{S}(d,m,N) \mid (3.8) \text{ admits a solution } \boldsymbol{P}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})} \ge 0, \text{ and } \boldsymbol{R}^w - \boldsymbol{D}^\top \boldsymbol{P}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})} \boldsymbol{D} > 0\Big\}. \tag{3.9}$$

Then, $\boldsymbol{K} \in \mathcal{K}$ is sufficient *for the inner-loop solution $\boldsymbol{L}(\boldsymbol{K})$ to be well defined. Also, $\boldsymbol{L}(\boldsymbol{K})$ is unique, and takes the form of:*

$$\boldsymbol{L}(\boldsymbol{K}) = (-\boldsymbol{R}^w + \boldsymbol{D}^\top \boldsymbol{P}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})} \boldsymbol{D})^{-1} \boldsymbol{D}^\top \boldsymbol{P}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})} (\boldsymbol{A} - \boldsymbol{B}\boldsymbol{K}). \tag{3.10}$$

Moreover, $\boldsymbol{K} \in \mathcal{K}$ is also almost necessary, *in that if $\boldsymbol{K} \notin \overline{\mathcal{K}}$, where*

$$\overline{\mathcal{K}} := \Big\{\boldsymbol{K} \in \mathcal{S}(d,m,N) \mid (3.8) \text{ admits a solution } \boldsymbol{P}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})} \ge 0, \text{ and } \boldsymbol{R}^w - \boldsymbol{D}^\top \boldsymbol{P}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})} \boldsymbol{D} \ge 0\Big\}, \tag{3.11}$$

then the solution to the inner loop is not well defined, i.e., the objective $\mathcal{G}(\boldsymbol{K}, \cdot)$ could be driven to arbitrarily large values. Lastly, the solution to (3.8) satisfies $\boldsymbol{P}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})} \ge \boldsymbol{P}_{\boldsymbol{K},\boldsymbol{L}}, \forall \boldsymbol{L}$ in the p.s.d. sense.

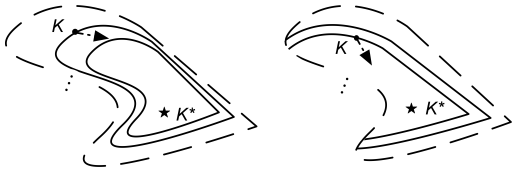

Figure 1: *Left:* Optimization landscape of LQR, where the dashed line represents the boundary of the stabilizing controller set. *Right:* Optimization landscape of the outer loop, with the dashed line representing the boundary of $\mathcal{K}$. The solid lines represent the contour lines of the objective function, K denotes the control gain of one iterate, and ★ is the global minimizer.

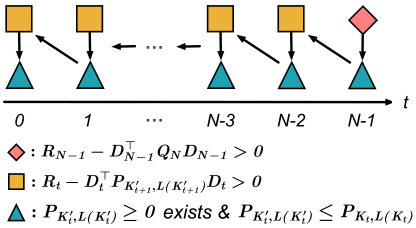

Figure 2: Proof idea for Theorem 3.7: For any $\boldsymbol{K} \in \mathcal{K}$, denote the gain matrix after one-step PG update as $\boldsymbol{K}'$. We construct an iterative argument to find a constant stepsize such that $P_{K'_t,L(K'_t)} \geq 0$ exists and satisfies $P_{K'_t,L(K'_t)} \leq P_{K_t,L(K_t)}$ for all t. Specifically, any K'_t satisfying ■ also satisfies ▲. Moreover, ♦ is enforced by Assumption 2.1. Combined, $\boldsymbol{K}' \in \mathcal{K}$ is ensured.

Note that $\mathcal{K}$ is nonempty due to Assumption 2.1, but it might be *unbounded*. Also, every $\boldsymbol{K} \in \mathcal{K}$ is equivalent to a control gain matrix that attains a γ-level of disturbance in the original disturbance attenuation problem (cf. Lemma 2.2). Then, by Lemmas 3.1 and 3.3, the inner-loop maximization problem for a fixed $\boldsymbol{K} \in \mathcal{K}$ is nonconcave in $\boldsymbol{L}$ and always admits a unique maximizing solution. Thus, applying a variant of the Danskin's theorem [30] yields that $\max_{\boldsymbol{L}} \mathcal{G}(\boldsymbol{K},\boldsymbol{L})$ is differentiable with respect to $\boldsymbol{K}$ and $\nabla_{\boldsymbol{K}}\{\max_{\boldsymbol{L}} \mathcal{G}(\boldsymbol{K},\boldsymbol{L})\} = \nabla_{\boldsymbol{K}} \mathcal{G}(\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K}))$. The outer loop is then a nonconvex (as we will show in Lemma 3.5) constrained optimization problem in $\boldsymbol{K}$, expressed as $\min_{\boldsymbol{K}\in\mathcal{K}} \mathcal{G}(\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K}))$. Then, the PG of the outer loop is computed as $\nabla_{\boldsymbol{K}} \mathcal{G}(\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})) = 2\boldsymbol{F}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})}\Sigma_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})}$. Lastly, we comment on the double-loop scheme we study in Remark A.7.

3.2 Optimization Landscape

For a fixed $\boldsymbol{K}$, we present the optimization landscape of the inner-loop and defer the proof to §B.7.

Lemma 3.4 *(Inner-Loop Landscape) There exists $\boldsymbol{K} \in \mathcal{K}$ such that $\mathcal{G}(\boldsymbol{K},\boldsymbol{L})$ is nonconcave in $\boldsymbol{L}$. For a fixed $\boldsymbol{K} \in \mathcal{K}$, $\mathcal{G}(\boldsymbol{K},\boldsymbol{L})$ is coercive (i.e., $\mathcal{G}(\boldsymbol{K},\boldsymbol{L}) \to -\infty$ as $\|\boldsymbol{L}\|_F \to \infty$), and the superlevel set*

$$\mathcal{L}_{\boldsymbol{K}}(a) := \left\{\boldsymbol{L} \in \mathcal{S}(n,m,N) \mid \mathcal{G}(\boldsymbol{K},\boldsymbol{L}) \geq a\right\} \tag{3.12}$$

is compact for any a such that $\mathcal{L}_{\boldsymbol{K}}(a) \neq \emptyset$. Moreover, there exist some $l_{\boldsymbol{K},\boldsymbol{L}}, \psi_{\boldsymbol{K},\boldsymbol{L}}, \rho_{\boldsymbol{K},\boldsymbol{L}} > 0$ such that $\mathcal{G}(\boldsymbol{K},\boldsymbol{L})$ is $(l_{\boldsymbol{K},\boldsymbol{L}}, \rho_{\boldsymbol{K},\boldsymbol{L}})$ locally Lipschitz and $(\psi_{\boldsymbol{K},\boldsymbol{L}}, \rho_{\boldsymbol{K},\boldsymbol{L}})$ locally smooth at $(\boldsymbol{K},\boldsymbol{L})$. Further, there exist some $l_{\boldsymbol{K},a}, \psi_{\boldsymbol{K},a} > 0$ such that $\mathcal{G}(\boldsymbol{K},\boldsymbol{L})$ is $l_{\boldsymbol{K},a}$-globally Lipschitz and $\psi_{\boldsymbol{K},a}$-globally smooth over $\mathcal{L}_{\boldsymbol{K}}(a)$. Lastly, $\mathcal{G}(\boldsymbol{K},\boldsymbol{L})$ is $\mu_{\boldsymbol{K}}$-PL for some $\mu_{\boldsymbol{K}} > 0$ which depends only on $\boldsymbol{K}$ and the problem parameters.

The optimization landscape of the inner loop is similar to that of the LQR [11]. In our setting, independent process noises (together with the random initial state) with a positive-definite covariance matrix guarantee the full-rankness of $\Sigma_{\boldsymbol{K},\boldsymbol{L}}$, which leads to the stationary point of the inner-loop objective function $\mathcal{G}(\boldsymbol{K},\boldsymbol{L})$ also being the unique optimum (cf. Lemma 3.2). Further, the full-rankness of $\Sigma_{\boldsymbol{K},\boldsymbol{L}}$ also ensures that $\mathcal{G}(\boldsymbol{K},\boldsymbol{L})$ is PL (cf. Lemma 3.4). These two properties together are essential for establishing the global convergence of PG methods in our nonconvex-nonconcave setting.

Subsequently, we analyze the optimization landscape of the outer loop (proved in §B.8), subject to $\mathcal{K}$ defined by (3.9). Note that $\mathcal{K}$ is critical, as by Lemma 3.3, it is a sufficient and almost necessary condition to ensure that the solution to the associated inner-loop subproblem is well defined. More importantly, from a robust control perspective, such a set $\mathcal{K}$ represents the set of control gains that enjoy a certain level of *robustness*, which share the same vein as the celebrated $\mathcal{H}_\infty$-norm constraint. Indeed, they both enforce the gain matrix to *attenuate* a prescribed level of disturbance. This level of robustness also corresponds to the level of *risk-sensitivity* of the controllers in LEQG problems.

Lemma 3.5 *(Outer-Loop Landscape) There exist zero-sum LQ dynamic games such that $\mathcal{G}(\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K}))$ is nonconvex and noncoercive on $\mathcal{K}$. Specifically, as $\boldsymbol{K}$ approaches the boundary of $\mathcal{K}$, $\mathcal{G}(\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K}))$ does not necessarily approach $+\infty$. Moreover, the stationary point of $\mathcal{G}(\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K}))$ in $\mathcal{K}$, denoted as $(\boldsymbol{K}^*,\boldsymbol{L}(\boldsymbol{K}^*))$, is unique and constitutes the unique Nash equilibrium of the game.*

Lack of the coercivity brings up challenges for convergence analysis, as a decrease in the value of the objective function cannot ensure feasibility of the updated gain matrix, in sharp contrast to the

LQR problem [11, 13]. We illustrate the difficult landscape of the outer loop in Figure 1. To address this challenge, we will show next that the natural PG (NPG) and Gauss-Newton (GN) updates, with some stepsize choices, can automatically preserve the feasibility of the iterates on-the-fly, termed the *implicit regularization* (IR) property.

3.3 Update Rules and Global Convergence

We use $l, k \geq 0$ to represent the iteration indices of the inner- and outer-loop updates, respectively, and introduce three PG-based update rules as follows:

$$\text{PG:} \quad \boldsymbol{L}_{l+1} = \boldsymbol{L}_l + \eta \nabla_{\boldsymbol{L}} \mathcal{G}(\boldsymbol{K}_k, \boldsymbol{L}_l), \tag{3.13} \qquad \boldsymbol{K}_{k+1} = \boldsymbol{K}_k - \alpha \nabla_{\boldsymbol{K}} \mathcal{G}(\boldsymbol{K}_k, \boldsymbol{L}(\boldsymbol{K}_k)), \tag{3.16}$$

$$\text{NPG:} \quad \boldsymbol{L}_{l+1} = \boldsymbol{L}_l + \eta \boldsymbol{E}_{\boldsymbol{K}_k, \boldsymbol{L}_l}, \tag{3.14} \qquad \boldsymbol{K}_{k+1} = \boldsymbol{K}_k - \alpha \boldsymbol{F}_{\boldsymbol{K}_k, \boldsymbol{L}(\boldsymbol{K}_k)}, \tag{3.17}$$

$$\text{GN:} \quad \boldsymbol{L}_{l+1} = \boldsymbol{L}_l + \eta \boldsymbol{H}^{-1}_{\boldsymbol{K}_k, \boldsymbol{L}_l} \boldsymbol{E}_{\boldsymbol{K}_k, \boldsymbol{L}_l}, \tag{3.15} \qquad \boldsymbol{K}_{k+1} = \boldsymbol{K}_k - \alpha \boldsymbol{G}^{-1}_{\boldsymbol{K}_k, \boldsymbol{L}(\boldsymbol{K}_k)} \boldsymbol{F}_{\boldsymbol{K}_k, \boldsymbol{L}(\boldsymbol{K}_k)}, \tag{3.18}$$

where $\eta, \alpha > 0$ are constant stepsizes for the inner loop and the outer loop, respectively. We first present the convergence results for three inner-loop PG updates.

Theorem 3.6 *(Inner-Loop Global Convergence) For a fixed $\boldsymbol{K} \in \mathcal{K}$ and an arbitrary $\boldsymbol{L}_0$ that induces a finite $\mathcal{G}(\boldsymbol{K}, \boldsymbol{L}_0)$, we define a superlevel set $\mathcal{L}_{\boldsymbol{K}}(a)$ as in (3.12), where $a < \mathcal{G}(\boldsymbol{K}, \boldsymbol{L}_0)$ is arbitrary. Then, iterates $\boldsymbol{L}_l$ of the updates (3.13), (3.14), and (3.15) with stepsizes satisfying*

$$\text{PG: } \eta \leq 1/\psi_{\boldsymbol{K},a}, \quad \text{NPG: } \eta \leq 1/(2\|\boldsymbol{H}_{\boldsymbol{K},\boldsymbol{L}_0}\|), \quad \text{GN: } \eta \leq 1/2,$$

converge to $\boldsymbol{L}(\boldsymbol{K})$ with globally linear rates, where $\psi_{\boldsymbol{K},a}$ is the smoothness constant of $\mathcal{G}(\boldsymbol{K}, \boldsymbol{L})$ over $\mathcal{L}_{\boldsymbol{K}}(a)$. Moreover, with $\eta = 1/2$, GN (3.15) converges to $\boldsymbol{L}(\boldsymbol{K})$ with a locally Q-quadratic rate.

In contrast to the standard convergence proof for first-order methods in nonconvex optimization problems where the objective function is PL, the proof of Theorem 3.6 is more involved since we do not have the *global* smoothness property in our control setting. A careful analysis that addresses this issue has been carried out in §B.9. For the outer loop, we require the iterates of $\boldsymbol{K}$ to stay within $\mathcal{K}$ in order for the solution to the associated inner-loop subproblem to be well defined. To meet this requirement, we introduce the IR property for the NPG (3.17) and GN (3.18) updates in Theorem 3.7, with its proof being provided in §B.10.

Theorem 3.7 *(IR) Let $\boldsymbol{K}_0 \in \mathcal{K}$ and let the stepsizes satisfy*

$$\text{NPG: } \alpha \leq 1/\|\boldsymbol{G}_{\boldsymbol{K}_0, \boldsymbol{L}(\boldsymbol{K}_0)}\|, \qquad \text{GN: } \alpha \leq 1.$$

Then, the iterates $\boldsymbol{K}_k \in \mathcal{K}$ for all $k \geq 0$. In other words, the sequence of solutions to (3.8), $\{\boldsymbol{P}_{\boldsymbol{K}_k, \boldsymbol{L}(\boldsymbol{K}_k)}\}$, exists, and for all $k \geq 0$, $\boldsymbol{P}_{\boldsymbol{K}_k, \boldsymbol{L}(\boldsymbol{K}_k)}$ always satisfies the conditions in (3.9). Furthermore, the sequence $\{\boldsymbol{P}_{\boldsymbol{K}_k, \boldsymbol{L}(\boldsymbol{K}_k)}\}$ is monotonically non-increasing and bounded below by $\boldsymbol{P}_{\boldsymbol{K}^*, \boldsymbol{L}(\boldsymbol{K}^*)}$, in the p.s.d. sense.

A key step of the proof for Theorem 3.7 is to ensure the *existence* of a solution to (3.8) along the iterations, by carefully controlling the stepsizes of *certain* descent directions. We provide an illustration of the proof idea in Figure 2. In particular, the IR property holds for NPG and GN directions because they can ensure *matrix-wise* decrease of $\boldsymbol{P}_{\boldsymbol{K}_k, \boldsymbol{L}(\boldsymbol{K}_k)}$, while other descent directions (e.g., vanilla PG) can only decrease $\mathcal{G}(\boldsymbol{K}, \boldsymbol{L}(\boldsymbol{K}))$, which is a *scalar*. We highlight the importance of IR in Remark A.8, and establish the convergence result for the outer loop in the following theorem.

Theorem 3.8 *(Outer-Loop Global Convergence) Suppose $\boldsymbol{K}_0 \in \mathcal{K}$ and let the stepsizes satisfy*

$$\text{NPG: } \alpha \leq 1/(2\|\boldsymbol{G}_{\boldsymbol{K}_0, \boldsymbol{L}(\boldsymbol{K}_0)}\|), \qquad \text{GN: } \alpha \leq 1/2.$$

Then, the sequence of average natural gradient norm squares $\{k^{-1} \sum_{\kappa=0}^{k-1} \|\boldsymbol{F}_{\boldsymbol{K}_\kappa, \boldsymbol{L}(\boldsymbol{K}_\kappa)}\|_F^2\}$, $k \geq 1$, converges to $\boldsymbol{0}$ with $\mathcal{O}(1/k)$ rate. Moreover, this convergence is towards the unique Nash equilibrium. Lastly, the NPG (3.17) and GN (3.18) updates enjoy locally linear and Q-quadratic rates, respectively, around the Nash equilibrium.

The proof of Theorem 3.8 is deferred to §B.11. In the derivative-free setting where PGs are estimated through samples of system trajectories, if we can uniformly control the estimation bias using a fixed number of samples per iterate, then Theorems 3.6 and 3.8 together imply that the global convergence to the Nash equilibrium also holds. We will substantiate this in the next section.

4 Derivative-Free Policy Gradient Methods

We prove the sample complexity of our double-loop algorithm, when the exact PG is not accessible, and can only be estimated through samples of system trajectories. In particular, we propose a zeroth-order NPG (ZO-NPG) algorithm with a (zeroth-order) maximization oracle that approximately solves the inner-loop subproblem. Details of the algorithms are deferred to Algorithms 1 and 2 in §D.

4.1 Inner-Loop Maximization Oracle

Sample complexities of zeroth-order PG algorithms for solving standard infinite-horizon LQR have been investigated in both discrete-time [11, 18, 31] and continuous-time [16] settings. Our inner-loop maximization oracle extends the sample complexity result to finite-horizon time-varying LQR with system noises and a possibly *indefinite* state-weighting matrix. In particular, we show that zeroth-order PG and NPG with a one-point minibatch estimation scheme enjoy $\widetilde{\mathcal{O}}(\epsilon_1^{-2})$ sample complexities, where ϵ_1 is the desired accuracy level in terms of inner-loop objective values, i.e., $\mathcal{G}(\boldsymbol{K},\overline{\boldsymbol{L}}(\boldsymbol{K})) \geq \mathcal{G}(\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})) - \epsilon_1$ for the $\overline{\boldsymbol{L}}(\boldsymbol{K})$ returned by the algorithm. The two specific zeroth-order PG updates are introduced as follows:

$$\textit{ZO-PG: } \boldsymbol{L}_{l+1} = \boldsymbol{L}_l + \eta \overline{\nabla}_{\boldsymbol{L}} \mathcal{G}(\boldsymbol{K},\boldsymbol{L}_l), \qquad (4.1) \qquad\qquad \textit{ZO-NPG: } \boldsymbol{L}_{l+1} = \boldsymbol{L}_l + \eta \overline{\nabla}_{\boldsymbol{L}} \mathcal{G}(\boldsymbol{K},\boldsymbol{L}_l)\overline{\Sigma}_{\boldsymbol{K},\boldsymbol{L}_l}^{-1}, \qquad (4.2)$$

where $\eta > 0$ is the stepsize to be chosen, $l \geq 0$ is the iteration index, $\overline{\nabla}_{\boldsymbol{L}} \mathcal{G}(\boldsymbol{K},\boldsymbol{L})$ and $\overline{\Sigma}_{\boldsymbol{K},\boldsymbol{L}}$ are the noisy estimates of $\nabla_{\boldsymbol{L}} \mathcal{G}(\boldsymbol{K},\boldsymbol{L})$ and $\Sigma_{\boldsymbol{K},\boldsymbol{L}}$, respectively, obtained through zeroth-order oracles. The procedure of our inner-loop maximization oracle is presented in Algorithm 1 in §D and we establish its sample complexity in the following theorem.

Theorem 4.1 *(Inner-Loop Sample Complexity) Let the desired accuracy level be $\epsilon_1, \delta_1 \in (0,1)$, and parameters of the minibatch zeroth-order oracle (Algorithm 1 in §D) satisfy*

$$\textit{Minibatch Size: } M_1 \sim \widetilde{\Omega}(\epsilon_1^{-2}), \quad \textit{Stepsize: } \eta \sim \mathcal{O}(1), \quad \textit{Smoothing Radius: } r_1 \sim \Theta(\sqrt{\epsilon_1}).$$

Then, with probability at least $1-\delta_1$ and $L \sim \mathcal{O}(\log(\epsilon_1^{-1}))$ steps, updates (4.1) and (4.2) output some $\overline{\boldsymbol{L}}(\boldsymbol{K}) := \boldsymbol{L}_L$ satisfying $\mathcal{G}(\boldsymbol{K},\overline{\boldsymbol{L}}(\boldsymbol{K})) \geq \mathcal{G}(\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})) - \epsilon_1$ and $\|\boldsymbol{L}(\boldsymbol{K}) - \overline{\boldsymbol{L}}(\boldsymbol{K})\|_F \leq \sqrt{\lambda_{\min}^{-1}(\boldsymbol{H}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})}) \cdot \epsilon_1}$.

Complete versions of Theorem 4.1 for PG and NPG updates (4.1) and (4.2) are given respectively in §B.12 and §B.13, along with their proofs. The total sample complexity of both (4.1) and (4.2) scales as $M_1 \cdot L \sim \widetilde{\mathcal{O}}(\epsilon_1^{-2} \cdot \log(\delta_1^{-1}))$, where the logarithmic dependence on ϵ_1 is suppressed. In contrast to [11], our Algorithm 1 uses an *unperturbed* pair of gain matrices to generate the state sequence for estimating $\Sigma_{\boldsymbol{K},\boldsymbol{L}}$. This modification avoids the estimation bias induced by the perturbations on the gain matrix, while only adding a constant factor of 2 to the total sample complexity.

4.2 Outer-Loop ZO-NPG

With the approximate inner-loop solution $\overline{\boldsymbol{L}}(\boldsymbol{K})$ obtained from Algorithm 1, the outer-loop ZO-NPG algorithm approximately solves $\min_{\boldsymbol{K}\in\mathcal{K}} \mathcal{G}(\boldsymbol{K},\overline{\boldsymbol{L}}(\boldsymbol{K}))$, with the following update rule:

$$\textit{ZO-NPG:} \quad \boldsymbol{K}_{k+1} = \boldsymbol{K}_k - \alpha \overline{\nabla}_{\boldsymbol{K}} \mathcal{G}(\boldsymbol{K}_k,\overline{\boldsymbol{L}}(\boldsymbol{K}_k))\overline{\Sigma}_{\boldsymbol{K}_k,\overline{\boldsymbol{L}}(\boldsymbol{K}_k)}^{-1}, \qquad (4.3)$$

where $\alpha > 0$ is the stepsize, $k \geq 0$ is the iteration index, and $\overline{\nabla}_{\boldsymbol{K}} \mathcal{G}(\boldsymbol{K},\overline{\boldsymbol{L}}(\boldsymbol{K}))$ and $\overline{\Sigma}_{\boldsymbol{K},\overline{\boldsymbol{L}}(\boldsymbol{K})}$ are the estimated PG and state correlation matrices, obtained from Algorithm 2 in §D. Similar to the modification in §4.1, we estimate the correlation matrix $\overline{\Sigma}(\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K}))$ using the state sequence generated by the *unperturbed* gain matrices, $(\boldsymbol{K},\overline{\boldsymbol{L}}(\boldsymbol{K}))$, to avoid the estimation bias induced by the perturbations on the gain matrices. In the following theorem (Theorem 4.2), with its complete version and proof being deferred to §B.14, we have that the IR property in Theorem 3.7 also holds in the derivative-free setting, with high probability. We provide here a sketch of the proof of Theorem 4.2.

Theorem 4.2 *(IR: Derivative-Free Setting) For any $\boldsymbol{K}_0 \in \mathcal{K}$, define the following strict subset of $\mathcal{K}$:*

$$\widehat{\mathcal{K}} := \left\{ \boldsymbol{K} \mid (3.8) \textit{ admits a solution } \boldsymbol{P}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})} \geq 0, \textit{ and } \boldsymbol{P}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})} \leq \boldsymbol{P}_{\boldsymbol{K}_0,\boldsymbol{L}(\boldsymbol{K}_0)} + \frac{\lambda_{\min}(\boldsymbol{H}_{\boldsymbol{K}_0,\boldsymbol{L}(\boldsymbol{K}_0)})}{2\|\boldsymbol{D}\|^2} \cdot \boldsymbol{I} \right\}.$$

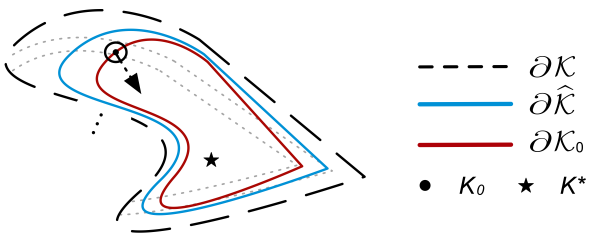

Figure 3: Illustrating the proof idea for Theorem 4.2. For any $\boldsymbol{K}_0 \in \mathcal{K}$, we construct two compact sets $\widehat{\mathcal{K}}$ and $\mathcal{K}_0$ shown in blue and red, respectively, that are independent of the contour lines of the objective function. Our analysis proves that the iterates following (4.3) stay within $\widehat{\mathcal{K}}$ with high probability, thus uniformly separated from $\partial\mathcal{K}$, with high probability.

Let $\delta_2 \in (0,1)$ and parameters of the minibatch zeroth-order algorithm (Algorithm 2 in §D) satisfy

$$\text{Minibatch Size: } M_2 \sim \widetilde{\Omega}(K^4), \quad \text{Stepsize: } \alpha \sim \mathcal{O}(1), \quad \text{Smoothing Radius: } r_2 \sim \mathcal{O}(K^{-1}),$$
$$\text{Inner-Loop Accuracy: } \epsilon_1 \sim \mathcal{O}(K^{-2}), \qquad \delta_1 \sim \mathcal{O}(\delta_2 M_2^{-1} K^{-1}).$$

Then, it holds with probability at least $1-\delta_2$ that $\boldsymbol{K}_k \in \widehat{\mathcal{K}} \subset \mathcal{K}$ for all $k \in \{1, \cdots, K\}$.

Proof Sketch. Due to lack of coercivity, the objective function of the outer loop cannot act as a barrier function to guarantee that the iterates of PG updates stay in $\mathcal{K}$ (cf. Figure 1), and thus new candidates are needed. To this end, we construct two compact sets $\widehat{\mathcal{K}}$ and $\mathcal{K}_0$ for a given $\boldsymbol{K}_0 \in \mathcal{K}$, as shown in blue and red in Figure 3, respectively, where $\mathcal{K}_0 := \left\{\boldsymbol{K} \mid \boldsymbol{K} \in \mathcal{K}, \text{ and } \boldsymbol{P}_{\boldsymbol{K},\boldsymbol{L}(\boldsymbol{K})} \leq \boldsymbol{P}_{\boldsymbol{K}_0,\boldsymbol{L}(\boldsymbol{K}_0)}\right\}$. Clearly, $\boldsymbol{K}_0 \in \partial\mathcal{K}_0$ and a sequence of iterates $\{\boldsymbol{K}_k\}$ that stays in $\widehat{\mathcal{K}}$ or $\mathcal{K}_0$ is also uniformly separated from $\partial\mathcal{K}$ (dashed lines in Figure 3). Denote the gain matrix after one step of the exact NPG update (3.17) as $\widetilde{\boldsymbol{K}}_1$. Then, under an appropriate stepsize, Theorem 3.7 demonstrates that $\boldsymbol{P}_{\widetilde{\boldsymbol{K}}_1,\boldsymbol{L}(\widetilde{\boldsymbol{K}}_1)} \geq 0$ exists and satisfies $\boldsymbol{P}_{\widetilde{\boldsymbol{K}}_1,\boldsymbol{L}(\widetilde{\boldsymbol{K}}_1)} \leq \boldsymbol{P}_{\boldsymbol{K}_0,\boldsymbol{L}(\boldsymbol{K}_0)}$ almost surely, implying $\widetilde{\boldsymbol{K}_1} \in \mathcal{K}_0$ almost surely. In contrast, when the model is not known, one step of the ZO-NPG update (4.3) using estimated gradients sampled through system trajectories could drive the gain matrix outside of $\mathcal{K}_0$ (even worse, outside of $\mathcal{K}$) due to the induced statistical errors. Moreover, these errors accumulate over all the iterations, raising significant challenges to find a uniform "margin" that is needed to safely select the parameters of Algorithm 2.

To overcome this challenge, we establish some arguments stronger than that in Theorem 3.7, i.e., $\boldsymbol{K}_k$ stays in $\widehat{\mathcal{K}}$, for all k, with high probability. We first show that with polynomial samples, the estimated NPG could be accurate enough such that $\boldsymbol{K}_1$, the iterate after applying one step of (4.3), is close to $\widetilde{\boldsymbol{K}}_1$, and thus stays in $\widehat{\mathcal{K}}$, with high probability. The same arguments could be iteratively applied to all future iterations, because starting from any $\boldsymbol{K}_k \in \mathcal{K}$ and choosing an appropriate stepsize, Theorem 3.7 guarantees that the iterates following the exact NPG direction are monotonically moving toward the interior of $\mathcal{K}$. Also, there exist parameters of Algorithm 2 such that the NPG estimates could be arbitrarily close to the exact ones. These two properties together imply that we can control, with high probability, the rate under which the iterates following (4.3) is moving "outward", i.e. toward $\partial\mathcal{K}$. Therefore, we manage to demonstrate that even in the worse case, the iterates of the ZO-NPG update will not travel beyond $\partial\widehat{\mathcal{K}}$ (the blue line), with high probability. Since $\widehat{\mathcal{K}}$ is compact, we can then safely choose the parameters of Algorithm 2 when analyzing the convergence rate of (4.3). ■

Theorem 4.2 appears to be the first IR result of PO in robust control in the derivative-free setting, with previous work in the literature [15] focusing only on the case with exact PG accesses. With Theorem 4.2 at hand, we now present the sample complexity of the outer-loop ZO-NPG algorithm (4.3), deferring its proof to §B.15.

Theorem 4.3 *(Outer-Loop Sample Complexity) For any $\boldsymbol{K}_0 \in \mathcal{K}$, let $\epsilon_2 \leq \phi/2$, $\delta_2, \epsilon_1, \delta_1, M_2, r_2, \alpha$ in Algorithm 2 have the same order as the requirements introduced in Theorem 4.2. Then, it holds with probability at least $1-\delta_2$ that the sequence $\{\boldsymbol{K}_k\}$, $k \in \{0, \cdots, K\}$, converges with $\mathcal{O}(1/K)$ rate in the sense that $K^{-1}\sum_{k=0}^{K-1} \|\boldsymbol{F}_{\boldsymbol{K}_k,\boldsymbol{L}(\boldsymbol{K}_k)}\|_F^2 \leq \epsilon_2$ with $K = \mathrm{Tr}(\boldsymbol{P}_{\boldsymbol{K}_0,\boldsymbol{L}(\boldsymbol{K}_0)} - \boldsymbol{P}_{\boldsymbol{K}^*,\boldsymbol{L}(\boldsymbol{K}^*)})/[\alpha\phi\epsilon_2]$.*

Hence, ZO-NPG with maximization oracle converges to the ϵ_2-neighborhood of the Nash equilibrium with probability at least $1-\delta_2$, after using $M_2 \cdot K \sim \widetilde{\mathcal{O}}(\epsilon_2^{-5} \cdot \log(\delta_2^{-1}))$ samples. Moreover, the convergent gain matrix $\boldsymbol{K}_K$ also solves the LEQG and the LQ disturbance attenuation problems, due to their equivalences in Lemma 2.2.

References

[1] John Schulman, Philipp Moritz, Sergey Levine, Michael Jordan, and Pieter Abbeel. High-dimensional continuous control using generalized advantage estimation. *arXiv preprint arXiv:1506.02438*, 2015.

[2] Timothy P Lillicrap, Jonathan J Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971*, 2015.

[3] Benjamin Recht. A tour of reinforcement learning: The view from continuous control. *Annual Review of Control, Robotics, and Autonomous Systems*, 2:253–279, 2019.

[4] Richard S Sutton, David A McAllester, Satinder P Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. In *Advances in Neural Information Processing Systems*, pages 1057–1063, 2000.

[5] Sham M Kakade. A natural policy gradient. In *Advances in Neural Information Processing Systems*, pages 1531–1538, 2002.

[6] Kaiqing Zhang, Alec Koppel, Hao Zhu, and Tamer Başar. Global convergence of policy gradient methods to (almost) locally optimal policies. *SIAM Journal on Control and Optimization*, 2020.

[7] Vijay R Konda and John N Tsitsiklis. Actor-critic algorithms. In *Advances in Neural Information Processing Systems*, pages 1008–1014, 2000.

[8] Shalabh Bhatnagar, Richard S Sutton, Mohammad Ghavamzadeh, and Mark Lee. Natural actor-critic algorithms. *Automatica*, 45(11):2471–2482, 2009.

[9] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

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

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

[12] Stephen Tu and Benjamin Recht. The gap between model-based and model-free methods on the linear quadratic regulator: An asymptotic viewpoint. In *Conference on Learning Theory*, pages 3036–3083, 2019.

[13] Jingjing Bu, Afshin Mesbahi, Maryam Fazel, and Mehran Mesbahi. LQR through the lens of first order methods: Discrete-time case. *arXiv preprint arXiv:1907.08921*, 2019.

[14] Kaiqing Zhang, Zhuoran Yang, and Tamer Başar. Policy optimization provably converges to Nash equilibria in zero-sum linear quadratic games. In *Advances in Neural Information Processing Systems*, pages 11602–11614, 2019.

[15] 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. *arXiv preprint arXiv:1910.09496*, 2019.

[16] 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. *arXiv preprint arXiv:1912.11899*, 2019.

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

[18] Dhruv Malik, Ashwin Pananjady, Kush Bhatia, Koulik Khamaru, Peter L Bartlett, and Martin J Wainwright. Derivative-free methods for policy optimization: Guarantees for linear quadratic systems. *Journal of Machine Learning Research*, 21(21):1–51, 2020.

[19] Paulo Jansch-Porto, Bin Hu, and Geir Dullerud. Convergence guarantees of policy optimization methods for Markovian jump linear systems. In *American Control Conference*, pages 2882–2887, 2020.

[20] Ilyas Fatkhullin and Boris Polyak. Optimizing static linear feedback: Gradient method. *arXiv preprint arXiv:2004.09875*, 2020.

[21] David Jacobson. Optimal stochastic linear systems with exponential performance criteria and their relation to deterministic differential games. *IEEE Transactions on Automatic Control*, 18(2):124–131, 1973.

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

[23] Tamer Başar. A dynamic games approach to controller design: Disturbance rejection in discrete-time. *IEEE Transactions on Automatic Control*, 36(8):936–952, 1991.

[24] Tamer Başar and Pierre Bernhard. *H-Infinity Optimal Control and Related Minimax Design Problems: A Dynamic Game Approach*. Springer Science & Business Media, 2008.

[25] Tamer Başar and Geert Jan Olsder. *Dynamic Noncooperative Game Theory*. SIAM, 1998.

[26] Peter Whittle. Risk-sensitive linear/quadratic/gaussian control. *Advances in Applied Probability*, pages 764–777, 1981.

[27] Peter Whittle. *Risk-Sensitive Optimal Control*, volume 2. Wiley, 1990.

[28] Haruki Nishimura, Negar Mehr, Adrien Gaidon, and Mac Schwager. RAT ILQR: A risk auto-tuning controller to optimally account for stochastic model mismatch. *IEEE Robotics and Automation Letters*, 6(2):763–770, 2021.

[29] Denis Mustafa and Dennis S Bernstein. LQG cost bounds in discrete-time $\mathcal{H}_2/\mathcal{H}_\infty$ control. *Transactions of the Institute of Measurement and Control*, 13(5):269–275, 1991.

[30] Pierre Bernhard and Alain Rapaport. On a theorem of Danskin with an application to a theorem of Von Neumann-Sion. *Nonlinear Analysis: Theory, Methods & Applications*, 24(8):1163–1181, 1995.

[31] 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.

[32] George Zames. Feedback and optimal sensitivity: Model reference transformations, multiplicative seminorms, and approximate inverses. *IEEE Transactions on Automatic Control*, 26(2):301–320, 1981.

[33] J Doyle, K Glover, P Khargonekar, and B Francis. State-space solutions to standard $\mathcal{H}_2$ and $\mathcal{H}_\infty$ control problems. *IEEE Transactions on Automatic Control*, 34:831–847, 1989.

[34] Tamer Başar. Minimax disturbance attenuation in LTV plants in discrete time. In *American Control Conference*, pages 3112–3113, 1990.

[35] Tankred Rautert and Ekkehard W Sachs. Computational design of optimal output feedback controllers. *SIAM Journal on Optimization*, 7(3):837–852, 1997.

[36] Karl Mårtensson and Anders Rantzer. Gradient methods for iterative distributed control synthesis. In *IEEE Conference on Decision and Control held jointly with Chinese Control Conference*, pages 549–554, 2009.

[37] 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.

[38] Benjamin Gravell, Peyman Mohajerin Esfahani, and Tyler Summers. Learning robust controllers for linear quadratic systems with multiplicative noise via policy gradient. *arXiv preprint arXiv:1907.03680*, 2019.

[39] 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.

[40] Vincent Roulet, Maryam Fazel, Siddhartha Srinivasa, and Zaid Harchaoui. On the convergence of the iterative linear exponential quadratic gaussian algorithm to stationary points. In *American Control Conference*, pages 132–137. IEEE, 2020.

[41] Ben M Hambly, Renyuan Xu, and Huining Yang. Policy gradient methods for the noisy linear quadratic regulator over a finite horizon. *arXiv preprint arXiv:2011.10300*, 2020.

[42] 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*, 2019.

[43] Joao Paulo Jansch-Porto, Bin Hu, and Geir Dullerud. Policy learning of MDPs with mixed continuous/discrete variables: A case study on model-free control of Markovian jump systems. In *Learning for Dynamics and Control*, pages 947–957, 2020.

[44] Constantinos Daskalakis and Ioannis Panageas. The limit points of (optimistic) gradient descent in min-max optimization. In *Advances in Neural Information Processing Systems*, pages 9236–9246, 2018.

[45] Ya-Ping Hsieh, Chen Liu, and Volkan Cevher. Finding mixed Nash equilibria of generative adversarial networks. In *International Conference on Machine Learning*, pages 2810–2819, 2019.

[46] Chi Jin, Praneeth Netrapalli, and Michael I Jordan. What is local optimality in nonconvex-nonconcave minimax optimization? In *International Conference on Machine Learning*, 2020.

[47] Emmanouil-Vasileios Vlatakis-Gkaragkounis, Lampros Flokas, and Georgios Piliouras. Poincaré recurrence, cycles and spurious equilibria in gradient-descent-ascent for non-convex non-concave zero-sum games. In *Advances in Neural Information Processing Systems*, pages 10450–10461, 2019.

[48] Junchi Yang, Negar Kiyavash, and Niao He. Global convergence and variance-reduced optimization for a class of nonconvex-nonconcave minimax problems. In *Advances in Neural Information Processing Systems*, 2020.

[49] Jelena Diakonikolas, Constantinos Daskalakis, and Michael I Jordan. Efficient methods for structured nonconvex-nonconcave min-max optimization. *arXiv preprint arXiv:2011.00364*, 2020.

[50] Constantinos Daskalakis, Stratis Skoulakis, and Manolis Zampetakis. The complexity of constrained min-max optimization. *arXiv preprint arXiv:2009.09623*, 2020.

[51] Kaiqing Zhang, Zhuoran Yang, and Tamer Başar. Multi-agent reinforcement learning: A selective overview of theories and algorithms. *arXiv preprint arXiv:1911.10635*, 2019.

[52] Constantinos Daskalakis, Dylan J Foster, and Noah Golowich. Independent policy gradient methods for competitive reinforcement learning. In *Advances in Neural Information Processing Systems*, 2020.

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

[54] Lerrel Pinto, James Davidson, Rahul Sukthankar, and Abhinav Gupta. Robust adversarial reinforcement learning. In *International Conference on Machine Learning*, pages 2817–2826, 2017.

[55] Kaiqing Zhang, Bin Hu, and Tamer Başar. On the stability and convergence of robust adversarial reinforcement learning: A case study on linear quadratic systems. In *Advances in Neural Information Processing Systems*, 2020.

[56] Arnab Nilim and Laurent El Ghaoui. Robust control of Markov decision processes with uncertain transition matrices. *Operations Research*, 53(5):780–798, 2005.

[57] Garud N Iyengar. Robust dynamic programming. *Mathematics of Operations Research*, 30(2):257–280, 2005.

[58] Shiau Hong Lim, Huan Xu, and Shie Mannor. Reinforcement learning in robust Markov decision processes. In *Advances in Neural Information Processing Systems*, pages 701–709, 2013.

[59] Shiau Hong Lim and Arnaud Autef. Kernel-based reinforcement learning in robust markov decision processes. In *International Conference on Machine Learning*, pages 3973–3981, 2019.

[60] Daniel J Mankowitz, Nir Levine, Rae Jeong, Abbas Abdolmaleki, Jost Tobias Springenberg, Timothy Mann, Todd Hester, and Martin Riedmiller. Robust reinforcement learning for continuous control with model misspecification. *arXiv preprint arXiv:1906.07516*, 2019.

[61] Yingjie Fei, Zhuoran Yang, Yudong Chen, Zhaoran Wang, and Qiaomin Xie. Risk-sensitive reinforcement learning: Near-optimal risk-sample tradeoff in regret. *arXiv preprint arXiv:2006.13827*, 2020.

[62] Xuezhou Zhang, Yiding Chen, Xiaojin Zhu, and Wen Sun. Robust policy gradient against strong data corruption. *arXiv preprint arXiv:2102.05800*, 2021.

[63] Isaac Kaminer, Pramod P Khargonekar, and Mario A Rotea. Mixed $\mathcal{H}_2/\mathcal{H}_\infty$ control for discrete-time systems via convex optimization. *Automatica*, 29(1):57–70, 1993.

[64] Lloyd S Shapley. Stochastic games. *Proceedings of the National Academy of Sciences*, 39(10):1095–1100, 1953.

[65] Michael L Littman. Markov games as a framework for multi-agent reinforcement learning. In *Machine Learning Proceedings*, pages 157–163. Elsevier, 1994.

[66] Kaiqing Zhang, Sham M Kakade, Tamer Başar, and Lin F Yang. Model-based multi-agent RL in zero-sum markov games with near-optimal sample complexity. *arXiv preprint arXiv:2007.07461*, 2020.

[67] Maher Nouiehed, Maziar Sanjabi, Tianjian Huang, Jason D Lee, and Meisam Razaviyayn. Solving a class of non-convex min-max games using iterative first order methods. In *Advances in Neural Information Processing Systems*, pages 14934–14942, 2019.

[68] Tianyi Lin, Chi Jin, and Michael I Jordan. On gradient descent ascent for nonconvex-concave minimax problems. In *International Conference on Machine Learning*, 2020.

[69] Tanner Fiez and Lillian Ratliff. Gradient descent-ascent provably converges to strict local minmax equilibria with a finite timescale separation. *arXiv preprint arXiv:2009.14820*, 2020.

[70] Christian Berg. *Kompleks Funktionsteori*. Universitetsbogladen, 2001.

[71] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[72] Abraham D Flaxman, Adam Tauman Kalai, and H Brendan McMahan. Online convex optimization in the bandit setting: gradient descent without a gradient. In *ACM-SIAM Symposium on Discrete Algorithms*, pages 385–394, 2005.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See the "Limitations and Broader Impacts" paragraph right before §2.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] See the "Limitations and Broader Impacts" paragraph right before §2.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] The only assumption that we have made is Assumption 2.1, which is standard in the literature.
   (b) Did you include complete proofs of all theoretical results? [Yes] The proofs are deferred to the supplementary material, see the first page of our appendix for an outline.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] All code and data are included in the supplementary material.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See §D of our appendix.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [No] For the exact PG updates (3.13)-(3.18), randomness enters only when we are randomly searching for an initial $\boldsymbol{K}_0 \in \mathcal{K}$ in $\mathcal{S}(d, m, N)$, where $\mathcal{S}(d, m, N)$ is defined in §3. For a given $\boldsymbol{K}_0 \in \mathcal{K}$ fixed, exact PG updates are deterministic and thus not affected by varying random seed. For the zeroth-order updates (4.1), (4.2), and (4.3), randomness enters into the optimization process in addition to initialization. However, we utilize mini-batch gradient estimators with large batchsizes to sample PGs at a high accuracy, thus the effects of random seeds are minimal.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See §D of our appendix.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]