Online Policy Gradient for Model Free Learning of Linear Quadratic Regulators with $\sqrt{T}$ Regret

Asaf Cassel [1] **Tomer Koren** [1 2]

Abstract

We consider the task of learning to control a linear dynamical system under fixed quadratic costs, known as the Linear Quadratic Regulator (LQR) problem. While model-free approaches are often favorable in practice, thus far only model-based methods, which rely on costly system identification, have been shown to achieve regret that scales with the optimal dependence on the time horizon T. We present the first model-free algorithm that achieves similar regret guarantees. Our method relies on an efficient policy gradient scheme, and a novel and tighter analysis of the cost of exploration in policy space in this setting.

1. Introduction

Model-free, policy gradient algorithms have become a staple of Reinforcement Learning (RL) with both practical successes (Lillicrap et al., 2015; Haarnoja et al., 2018), and strong theoretical guarantees in several settings (Sutton et al., 1999; Silver et al., 2014). In this work we study the design and analysis of such algorithms for the adaptive control of Linear Quadratic Regulator (LQR) systems, as seen through the lens of regret minimization (Abbasi-Yadkori & Szepesvári, 2011; Cohen et al., 2019; Mania et al., 2019). In this continuous state and action reinforcement learning setting, an agent chooses control actions u_t and the system state x_t evolves according to the noisy linear dynamics

$$x_{t+1} = A_\star x_t + B_\star u_t + w_t,$$

where $A_\star$ and $B_\star$ are transition matrices and w_t are i.i.d zero-mean noise terms. The cost is a quadratic function of the current state and action, and the regret is measured with respect to the class of linear policies, which are known to be optimal for this setting.

Model-based methods, which perform planning based on a system identification procedure that estimates the transition matrices, have been studied extensively in recent years. This started with Abbasi-Yadkori & Szepesvári (2011), which established an $O(\sqrt{T})$ regret guarantee albeit with a computationally intractable method. More recently, Cohen et al. (2019); Mania et al. (2019) complemented this result with computationally efficient methods, and Simchowitz & Foster (2020); Plevrakis & Hazan (2020) respectively give model-based algorithms with similar guarantees for the settings of strongly-convex adversarial costs, and general fixed convex costs, which both encompass the quadratic setting. Cassel et al. (2020); Simchowitz & Foster (2020) provided lower bounds, showing that this rate is generally unavoidable, regardless of whether the algorithm is model free or not, with Simchowitz & Foster (2020) also establishing a near-optimal dependence on the dimension parameters. In comparison, the best existing model-free algorithms are policy iteration procedures by Krauth et al. (2019) and Abbasi-Yadkori et al. (2019) that respectively achieve $\widetilde{O}(T^{2/3})$ and $\widetilde{O}(T^{2/3+\epsilon})$ regret for $\epsilon = \Theta(1/\log T)$.

Our main result is an efficient (in fact, linear time per step) policy gradient algorithm that achieves $\widetilde{O}(\sqrt{T})$ regret, thus closing the (theoretical) gap between model based and free methods for the LQR model. An interesting feature of our approach is that while the policies output by the algorithm are clearly state dependent, the tuning of their parameters requires no such access. Instead, we only rely on observations of the incurred cost, similar to bandit models (e.g., Cassel & Koren, 2020). We note that our results focus strictly on the time horizon parameter, and it remains open whether model-free methods could also achieve dimension optimal regret guarantees.

One of the main challenges of regret minimization in LQRs (and more generally, in reinforcement learning) is that it is generally infeasible to change policies as often as one likes. Roughly, this is due to a burn-in period following a policy change, during which the system converges to a new steady distribution, and typically incurs an additional cost proportional to the change in steady states, which is in turn proportional to the distance between policies. There are several ways to overcome this impediment. The simplest is

[1]Blavatnik School of Computer Science, Tel Aviv University [2]Google Research, Tel Aviv. Correspondence to: Asaf Cassel <acassel@mail.tau.ac.il>, Tomer Koren <tkoren@tauex.tau.ac.il>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

to restrict the number of policy updates and explore directly in the action space via artificial noise (see e.g., Simchowitz & Foster, 2020). Another approach by Cohen et al. (2019) considers a notion of slowly changing policies, however, these can be very prohibitive for exploration in policy space. Other works (e.g., Agarwal et al., 2019) consider a policy parameterization that converts the problem into online optimization with memory, which also relies on slowly changing policies. This last method is also inherently model-based and thus not adequate for our purpose.

A key technical contribution that we make is to overcome this challenge by exploring directly in policy space. While the idea itself is not new, we provide a novel and tighter analysis that allows us to use larger perturbations, thus reducing the variance of the resulting gradient estimates. We achieve this by showing that the additional cost depends only quadratically on the exploration radius, which is a crucial ingredient for overcoming the $O(T^{2/3})$ barrier of previous model-free methods.

The final ingredient of the analysis involves a sensitivity analysis of the gradient descent procedure that uses the estimated gradients. While similar analyses of gradient methods exist (see, e.g., Theorem 21 in Malik et al., 2020), they do not directly account for (small) adversarial corruptions to the gradient, which is crucial for our application. We provide a general result that gives appropriate conditions for which the optimization error depends only quadratically on the error in the gradients, regardless of whether its source is stochastic or adversarial.

Related work. Policy gradient methods in the context of LQR have seen significant interest in recent years. Notably, Fazel et al. (2018) establish its global convergence in the perfect information setting, and give complexity bounds for sample based methods. Subsequently, Malik et al. (2020) refined their analysis to obtain the correct $1/\varepsilon^2$ scaling of the sample complexity, however, they do not address the cost of exploration. Hambly et al. (2020) also improve the sample efficiency, but in a finite horizon setting. Mohammadi et al. (2020) give sample complexity bounds for the continuous-time variant of LQR. Finally, Tu & Recht (2019) show that a model based method can potentially outperform the sample complexity of policy gradient by factors of the input and output dimensions. While we observe similar performance gaps in our regret bounds, these were not our main focus and may potentially be improved by a more refined analysis. Moving away from policy gradients, Yang et al. (2019); Jin et al. (2020); Yaghmaie & Gustafsson (2019) analyze the convergence and sample complexity of other model free methods such as policy iteration and temporal difference (TD) learning, but they do not include any regret guarantees.

2. Preliminaries

2.1. Setup: Learning in LQR

We consider the problem of regret minimization in the LQR model. At each time step t, a state $x_t \in \mathbb{R}^{d_x}$ is observed and action $u_t \in \mathbb{R}^{d_u}$ is chosen. The system evolves according to

$$x_{t+1} = A_\star x_t + B_\star u_t + w_t, \quad (x_0 = 0 \text{ w.l.o.g.}),$$

where the state-state $A_\star \in \mathbb{R}^{d_x \times d_x}$ and state-action $B_\star \in \mathbb{R}^{d_x \times d_u}$ matrices form the transition model and the w_t are bounded, zero mean, i.i.d. noise terms with a positive definite covariance matrix $\Sigma_w \succ 0$. Formally, there exist $\sigma, W > 0$ such that

$$\mathbb{E}w_t = 0 \quad , \|w_t\| \le W \quad , \Sigma_w = \mathbb{E}w_t w_t^\mathsf{T} \succ \sigma^2 I.$$

The bounded noise assumption is made for simplicity of the analysis, and in the full version of the paper (Cassel & Koren, 2021) we show how to accommodate Gaussian noise via a simple reduction to this setting. At time t, the instantaneous cost is

$$c_t = x_t^\mathsf{T} Q x_t + u_t^\mathsf{T} R u_t,$$

where $0 \prec Q, R \preceq I$ are positive definite. We note that the upper bound is without loss of generality since multiplying Q and R by a constant factor only re-scales the regret.

A policy of the learner is a potentially time dependent mapping from past history to an action $u \in \mathbb{R}^{d_u}$ to be taken at the current time step. Classic results in linear control establish that, given the system parameters $A_\star, B_\star, Q$ and R, a linear transformation of the current state is an optimal policy for the infinite horizon setting. We thus consider policies of the form $u_t = Kx_t$ and define their infinite horizon expected cost,

$$J(K) = \lim_{T\to\infty} \frac{1}{T} \mathbb{E}\left[\sum_{t=1}^{T} x_t^\mathsf{T}\left(Q + K^\mathsf{T} R K\right) x_t\right],$$

where the expectation is taken with respect to the random noise variables w_t. Let $K_\star = \arg\min_K J(K)$ be a (unique) optimal policy and $J_\star = J(K_\star)$ denote the optimal infinite horizon expected cost, which are both well defined under mild assumptions.[1] We are interested in minimizing the *regret* over T decision rounds, defined as

$$R_T = \sum_{t=1}^{T} \left(x_t^\mathsf{T} Q x_t + u_t^\mathsf{T} R u_t - J_\star\right).$$

[1] These are valid under standard, very mild stabilizability assumptions (see Bertsekas, 1995) that hold in our setting (see strong-stability).

We focus on the setting where the learner does not have a full a-priori description of the transition parameters $A_\star$ and $B_\star$, and has to learn the optimal control strategy while minimizing the regret.

Throughout, we assume that the learner has knowledge of constants $\alpha_0 > 0$ and $\psi \geq 1$ such that

$$\|Q^{-1}\|, \|R^{-1}\| \leq 1/\alpha_0, \text{ and } \|B_\star\| \leq \psi.$$

We also assume that there is a known stable (not necessarily optimal) policy K_0 and $\nu > 0$ such that $J(K_0) \leq \frac{1}{4}\nu$. We note that all of the aforementioned parameters could be easily estimated at the cost of an additive constant regret term by means of a warm-up period. However, recovering the initial control K_0 gives an additive constant that depends exponentially on the problem parameters as shown by Chen & Hazan (2020); Mania et al. (2019); Cohen et al. (2019).

Finally, denote the set of all "admissable" controllers

$$\mathcal{K} = \{K \mid J(K) \leq \nu\}.$$

By definition, $K_0 \in \mathcal{K}$. As discussed below, over the set $\mathcal{K}$ the LQR cost function J has certain regularity properties that we will use throughout.

2.2. Smooth Optimization

Fazel et al. (2018) show that while the objective $J(\cdot)$ is non-convex, it has properties that make it amenable to standard gradient based optimization schemes. We summarize these here as they are used in our analysis.

Definition 1 (PL-condition)**.** A function $f : \mathcal{X} \to \mathbb{R}$ with global minimum f^* is said to be μ-PL if it satisfies the Polyak-Lojasiewicz (PL) inequality with constant $\mu > 0$, given by

$$\mu(f(x) - f^*) \leq \|\nabla f(x)\|^2 \quad , \forall x \in \mathcal{X}.$$

Definition 2 (Smoothness)**.** A function $f : \mathbb{R}^d \to \mathbb{R}$ is locally β, D_0-*smooth* over $\mathcal{X} \subseteq \mathbb{R}^d$ if for any $x \in \mathcal{X}$ and $y \in \mathbb{R}^d$ with $\|y - x\| \leq D_0$

$$\|\nabla f(x) - \nabla f(y)\| \leq \beta \|x - y\|.$$

Definition 3 (Lipschitz)**.** A function $f : \mathbb{R}^d \to \mathbb{R}$ is locally G, D_0-*Lipschitz* over $\mathcal{X} \subseteq \mathbb{R}^d$ if for any $x \in \mathcal{X}$ and $y \in \mathbb{R}^d$ with $\|y - x\| \leq D_0$

$$|f(x) - f(y)| \leq G\|x - y\|.$$

It is well-known that for functions satisfying the above conditions and for sufficiently small step size η, the gradient descent update rule

$$x_{t+1} = x_t - \eta \nabla f(x_t)$$

converges exponentially fast, i.e., there exists $0 \leq \rho < 1$ such that $f(x_t) - f^* \leq \rho^t (f(x_0) - f^*)$ (e.g., Nesterov, 2003). This setting has also been investigated in the absence of a perfect gradient oracle. Here we provide a clean result that shows that the exponential convergence continue until reaching a noise-floor that depends only on the *squared error* of any gradient estimate.

Finally, we require the notion of a one point gradient estimate (Flaxman et al., 2005). Let $f : \mathcal{X} \to \mathbb{R}$ and define its smoothed version with parameter $r > 0$ as

$$f^r(x) = \mathbb{E}_B f(x + rB), \tag{1}$$

where $B \in \mathcal{B}^d$ is a uniform random vector over the Euclidean unit ball. The following lemma is standard (we include a proof in the full version of the paper (Cassel & Koren, 2021) for completeness).

Lemma 1. *If f is (D_0, β)-*locally smooth *and $r \leq D_0$, then:*

(i) $\nabla f^r(x) = \frac{d}{r}\mathbb{E}_U[f(x + rU)U]$, *where $U \in \mathcal{S}^d$ is a uniform random vector of the unit sphere;*

(ii) $\|\nabla f^r(x) - \nabla f(x)\| \leq \beta r, \ \forall x \in \mathcal{X}$.

2.3. Background on LQR

It is well-known for the LQR problem that

$$J(K) = \mathrm{Tr}(P_K \Sigma_w) = \mathrm{Tr}\big((Q + K^\mathsf{T} R K)\Sigma_K\big),$$

where P_K, Σ_K are the positive definite solutions to

$$P_K = Q + K^\mathsf{T} R K + (A_\star + B_\star K)^\mathsf{T} P_K (A_\star + B_\star K), \tag{2}$$

$$\Sigma_K = \Sigma_w + (A_\star + B_\star K)\Sigma_K (A_\star + B_\star K)^\mathsf{T}. \tag{3}$$

Another important notion is that of strong stability (Cohen et al., 2018). This is essentially a quantitative version of classic stability notions in linear control.

Definition 4 (strong stability)**.** A matrix M is (κ, γ)-strongly stable (for $\kappa \geq 1$ and $0 < \gamma \leq 1$) if there exist matrices $H \succ 0$ and L such that $M = HLH^{-1}$ with $\|L\| \leq 1 - \gamma$ and $\|H\|\|H^{-1}\| \leq \kappa$. A controller K for is $(\kappa, \gamma)-$strongly stable if $\|K\| \leq \kappa$ and the matrix $A_\star + B_\star K$ is (κ, γ)-strongly stable.

The following lemma, due to Cohen et al. (2019), relates the infinite horizon cost of a controller to its strong stability parameters.

Lemma 2 (Cohen et al., 2019, Lemma 18)**.** *Suppose that $K \in \mathcal{K}$ then K is $(\kappa, \gamma)-$strongly stable with $\kappa = \sqrt{\nu/\alpha_0 \sigma^2}$ and $\gamma = 1/2\kappa^2$.*

The following two lemmas, due to Cohen et al. (2018); Cassel et al. (2020), show that the state covariance converges exponentially fast, and that the state is bounded as long as controllers are allowed to mix.

Lemma 3 (Cohen et al., 2018, Lemma 3.2)**.** *Suppose we play some fixed $K \in \mathcal{K}$ starting from some $x_0 \in \mathbb{R}^{d_x}$, then*

$$\|\mathbb{E}[x_t x_t^\mathsf{T}] - \Sigma_K\| \le \kappa^2 e^{-2\gamma t}\|x_0 x_0^\mathsf{T} - \Sigma_K\|,$$
$$|\mathbb{E}[c_t] - J(K)| \le \frac{\nu\kappa^2}{\sigma^2} e^{-2\gamma t}\|x_0 x_0^\mathsf{T} - \Sigma_K\|.$$

Lemma 4 (Cassel et al., 2020, Lemma 39)**.** *Suppose we have $K_1, K_2, \ldots \in \mathcal{K}$. If we play each controller K_i for at least $\tau \ge 2\kappa^2 \log 2\kappa$ rounds before switching to K_{i+1} then for all $t \ge 1$ we have that $\|x_t\| \le 6\kappa^4 W$ and $c_t \le 36\nu\kappa^8 W^2/\sigma^2$.*

The following is a summary of results from Fazel et al. (2018) that describe the main properties of $\Sigma_K, P_K, J(K)$. See the full version of the paper (Cassel & Koren, 2021) for the complete details.

Lemma 5 (Fazel et al., 2018, Lemmas 11, 13, 16, 27 and 28)**.** *Let $K \in \mathcal{K}$ and $K' \in \mathbb{R}^{d_u \times d_x}$ with*

$$\|K - K'\| \le \frac{1}{8\psi\kappa^3} = D_0,$$

then we have that

(i) $\mathrm{Tr}(P_K) \le J(K)/\sigma^2$; $\mathrm{Tr}(\Sigma_K) \le J(K)/\alpha_0$;

(ii) $\|\Sigma_K - \Sigma_{K'}\| \le (8\psi\nu\kappa^3/\alpha_0)\|K - K'\|$;

(iii) $\|P_K - P_{K'}\| \le 16\psi\kappa^7\|K - K'\|$;

(iv) J satisfies the local Lipschitz condition (Definition 3) over $\mathcal{K}$ with D_0 and $G = 4\psi\nu\kappa^7/\alpha_0$;

(v) J satisfies the local smoothness condition (Definition 2) over $\mathcal{K}$ with D_0 and $\beta = 112\sqrt{d_x}\nu\psi^2\kappa^8/\alpha_0$;

(vi) J satisfies the PL condition (Definition 1) with $\mu = 4\nu/\kappa^4$.

3. Algorithm and Overview of Analysis

We are now ready to present our main algorithm for model free regret minimization in LQR. The algorithm, given in Algorithm 1, optimizes an underlying controller K_j over epochs of exponentially increasing duration. Each epoch consists of sub-epochs, during which a perturbed controller $K_{j,i}$ centered at K_j is drawn and played for τ rounds. At the end of each epoch, the algorithm uses $c_{j,i,\tau}$, which is the cost incurred during the final round of playing the controller $K_{j,i}$, to construct a gradient estimate which in turn is used to calculate the next underlying controller K_{j+1}. Interestingly, we do not make any explicit use of the state observation x_t which is only used implicitly to calculate the control signal, via $u_t = K_t x_t$. Furthermore, the algorithm makes only $O(d_u d_x)$ computations per time step.

Algorithm 1 LQR Online Policy Gradient

input: initial controller $K_0 \in \mathcal{K}$, step size η, mixing length τ, parameters μ, r_0, m_0
for epoch $j = 0, 1, 2, \ldots$ **do**
  set $r_j = r_0(1-\mu\eta/3)^{j/2}$, $m_j = m_0(1-\mu\eta/3)^{-2j}$
  for $i = 1, \ldots, m_j$ **do**
    draw $\widetilde{U}_{j,i} \in \mathbb{R}^{d_u \times d_x}$ with i.i.d. $\mathcal{N}(0,1)$ entries
    set $U_{j,i} = \widetilde{U}_{j,i}/\|\widetilde{U}_{j,i}\|_F$
    play $K_{j,i} = K_j + r_j U_{j,i}$ for τ rounds
    observe the cost of the final round $c_{j,i,\tau}$
  calculate $\hat{g}_j = \frac{d_x d_u}{m_j r_j}\sum_{i=1}^{m_j} c_{j,i,\tau} U_{j,i}$
  update $K_{j+1} = K_j - \eta\hat{g}_j$

Our main result regarding Algorithm 1 is stated in the following theorem: a high-probability $O(\sqrt{T})$ regret guarantee with a polynomial dependence on the problem parameters.

Theorem 1. *Let $\kappa = \sqrt{\nu/\alpha_0\sigma^2}$ and suppose we run Algorithm 1 with parameters*

$$\eta = \frac{\alpha_0}{128\nu\psi^2\kappa^{10}},\ \tau = 2\kappa^2\log(7\kappa T),$$
$$\mu = \frac{4\nu}{\kappa^4},\ r_0 = \frac{\alpha_0}{448\sqrt{d_x}\psi^2\kappa^{10}},$$
$$\sqrt{m_0} = \frac{2^{17} d_u d_x^{3/2}\psi^2\kappa^{20}W^2}{\alpha_0\sigma^2}\sqrt{\log\frac{240T^4}{\delta}},$$

then with probability at least $1-\delta$,

$$R_T = O\left(\frac{d_u d_x^{3/2}\psi^4\kappa^{36}W^2}{\alpha_0}\sqrt{T\tau\log\frac{T}{\delta}}\right).$$

Here we give an overview of the main steps in proving Theorem 1, deferring the details of each step to later sections. Our first step is analyzing the utility of the policies K_j computed at the end of each epoch. We show that the regret of each K_j (over epoch j) in terms of its long-term (steady state) cost compared to that of the optimal $K_\star$, is controlled by the inverse square-root of the epoch length m_j. While, a similar result was proven in Malik et al., 2020, it pertains to the setting where we have access to unbiased estimates of the infinite horizon cost, making it non-applicable for our single trajectory setting.

Lemma 6 (exploitation)**.** *Under the parameter choices of Theorem 1, for any $j \ge 0$ we have that with probability at least $1 - \delta/8T^2$,*

$$J(K_j) - J_\star = O\left(\nu\sqrt{\frac{m_0}{m_j}}\right)$$

$$= O\left(d_u d_x^{3/2}\psi^2\kappa^{22}W^2\sqrt{\frac{1}{m_j}\log\frac{T}{\delta}}\right),$$

and further that $J(K_j) \le \nu/2$.

The proof of the lemma is based on a careful analysis of gradient descent with inexact gradients and crucially exploits the PL and local-smoothness properties of the loss $J(\cdot)$. More details can be found in Section 4.

The more interesting (and challenging) part of our analysis pertains to controlling the costs associated with exploration, namely, the penalties introduced by the perturbations of the controllers K_j. The direct cost of exploration is clear: instead of playing the K_j intended for exploitation, the algorithm actually follows the perturbed controllers $K_{j,i}$ and thus incurs the differences in long-term costs $J(K_{j,i}) - J(K_j)$. Our following lemma bounds the accumulation of these penalties over an epoch j; importantly, it shows that while the bound scales linearly with the length of the epoch m_j, it has a *quadratic* dependence on the exploration radius r_j.

Lemma 7 (direct exploration cost)**.** *Under the parameter choices of Theorem 1, for any* $j \ge 0$ *we have that with probability at least* $1 - \delta/4T$,

$$\sum_{i=1}^{m_j} J(K_{j,i}) - J(K_j) = O\left(\frac{\sqrt{d_x}\nu\psi^2\kappa^8}{\alpha_0} r_j^2 m_j + \nu\sqrt{m_j\log\frac{T}{\delta}}\right).$$

There are additional, indirect costs associated with exploration however: within each epoch the algorithm switches frequently between different policies, thereby suffering the indirect costs that stem from their "burn-in" period. This is precisely what gives rise to the differences between the realized cost $c_{j,i,s}$ and the long-term cost $J(K_{j,i})$ of the policy $K_{j,i}$. The cumulative effect of these is bounded in the next lemma, which is the technical crux of our results. Here again, note the quadratic dependence on the exploration radius r_j which is essential for obtaining our $\sqrt{T}$-regret result.

Lemma 8 (indirect exploration cost)**.** *Under the parameter choices of Theorem 1, for any* $j \ge 0$ *we have that with probability at least* $1 - \delta/4T$,

$$\sum_{i=1}^{m_j}\sum_{s=1}^{\tau}\left(c_{j,i,s} - J(K_{j,i})\right) = O\left(\frac{\nu\kappa^8 W^2}{\sigma^2}\tau\sqrt{m_j\log\frac{T}{\delta}} + \frac{d_x\nu\psi^2\kappa^{10}}{\alpha_0} m_j r_j^2\right).$$

The technical details for Lemmas 7 and 8 are discussed in Section 5. We now have all the main pieces required for proving our main result.

Proof of Theorem 1. Taking a union bound, we conclude that Lemmas 6 to 8 hold for all $j \ge 0$ with probability at least $1 - \delta$. Now, notice that our choice of parameters is such that

$$r_j^2 m_j = r_0^2\sqrt{m_0 m_j} = O\left(\frac{\sqrt{d_x}d_u\alpha_0 W^2}{\psi^2\sigma^2}\sqrt{m_j\log\frac{T}{\delta}}\right).$$

Plugging this back into Lemmas 7 and 8 we get that for all j,

$$\sum_{i=1}^{m_j}\sum_{s=1}^{\tau}\left(c_{j,i,s} - J(K_{j,i})\right) = O\left(\frac{d_u d_x^{\frac{3}{2}}\nu\kappa^{10}W^2}{\sigma^2}\tau\sqrt{m_j\log\frac{T}{\delta}}\right),$$

$$\tau\sum_{i=1}^{m_j} J(K_{j,i}) - J(K_j) = O\left(\frac{d_u d_x\nu\kappa^8 W^2}{\sigma^2}\tau\sqrt{m_j\log\frac{T}{\delta}}\right).$$

We conclude that the regret during epoch j is bounded as

$$\begin{aligned}\sum_{i=1}^{m_j}\sum_{s=1}^{\tau}\left(c_{j,i,s} - J_\star\right) &= \left[\sum_{i=1}^{m_j}\sum_{s=1}^{\tau}\left(c_{j,i,s} - J(K_{j,i})\right)\right] \\ &+ \left[\tau\sum_{i=1}^{m_j} J(K_{j,i}) - J(K_j)\right] + \left[\tau m_j(J(K_j) - J_\star)\right] \\ &= O\left(d_u d_x^{3/2}\psi^2\kappa^{22}W^2\tau\sqrt{m_j\log\frac{T}{\delta}}\right),\end{aligned}$$

where the second step also used the fact that $\nu/\sigma^2 \le \kappa^2$. Finally, a simple calculation (see Lemma 12) shows that

$$\sum_{j=0}^{n-1}\sqrt{m_j} = O\left(\frac{1}{\mu\eta}\sqrt{T/\tau}\right) = O\left(\frac{\psi^2\kappa^{14}}{\alpha_0}\sqrt{T/\tau}\right),$$

and thus summing over the regret accumulated in each epoch concludes the proof. ■

4. Optimization Analysis

At its core, Algorithm 1 is a policy gradient method with K_j being the prediction after j gradient steps. In this section we analyze the sub-optimality gap of the underlying controllers K_j culminating in the proof of Lemma 6. To achieve this, we first consider a general optimization problem with a corrupted gradient oracle, and show that the optimization rate is limited only by the square of the corruption magnitude. We follow this with an analysis of the LQR gradient estimation from which the overall optimization cost follows readily.

4.1. Inexact First-Order Optimization

Let $f : \mathbb{R}^d \to \mathbb{R}$ be a function with global minimum $f_* > -\infty$. Suppose there exists $\bar{f} \in \mathbb{R}$ such that f is μ-*PL*, (D_0, β)-*locally smooth*, and (D_0, G)-*locally Lipschitz* over the sub-level set $\mathcal{X} = \{x \mid f(x) \le \bar{f}\}$. We consider the update rule

$$x_{t+1} = x_t - \eta \hat{g}_t, \tag{4}$$

where $f(x_0) \le \bar{f}$, and $\hat{g}_t \in \mathbb{R}^d$ is a corrupted gradient oracle that satisfies

$$\|\hat{g}_t - \nabla f(x_t)\| \le \varepsilon_t, \tag{5}$$

where $\varepsilon_t \le \min\{G, \sqrt{(\bar{f} - f_*)\mu}/2\}$ is the magnitude of the corruption at step t. Define the effective corruption up to round t as

$$\bar{\varepsilon}_t^2 = \max_{s \le t} \left\{ \varepsilon_s^2 [1 - (\mu\eta/3)]^{t-s} \right\},$$

and notice that if $\varepsilon_s[1 - (\mu\eta/3)] \le \varepsilon_{s+1}$ then $\bar{\varepsilon}_t = \varepsilon_t$.

The following result shows that this update rule achieves a linear convergence rate up to an accuracy that depends quadratically on the corruptions. The proof follows similar ideas to those in Theorem 21 of Malik et al., 2020, but crucially, distills the dependence on the gradient errors regardless of their estimation method. See proof in the full version of the paper (Cassel & Koren, 2021).

Theorem 2 (corrupted gradient descent)**.** *Suppose that* $\eta \le \min\{1/\beta, 4/\mu, D_0/2G.\}$ *Then for all* $t \ge 0$*,*

$$f(x_t) - f_* \le \max\left\{ \frac{4\bar{\varepsilon}_{t-1}^2}{\mu}, \left[1 - \frac{\mu\eta}{3}\right]^t (f(x_0) - f_*) \right\},$$

and consequently $x_t \in \mathcal{X}$*.*

4.2. Gradient Estimation

The gradient estimate $\hat{g}_j$ is a batched version of the typical one-point gradient estimator. We bound it in the next lemma using the following inductive idea: if $J(K_j) \le \nu/2$, then $K_{j,i} \in \mathcal{K}$ and standard concentration arguments imply that the estimation error is small with high probability and thus Theorem 2 implies that $J(K_{j+1}) \le \nu/2$.

Lemma 9 (Gradient estimation error)**.** *Under the parameter choices of Theorem 1, for any* $j \ge 0$ *we have that with probability at least* $1 - (\delta/8T^3)$*,*

$$\|\hat{g}_j - \nabla J(K_j)\|_F \le \frac{\sqrt{\mu\nu}}{4}\left(1 - \frac{\mu\eta}{3}\right)^{j/2}.$$

Proof of Lemma 9. Assume that conditioned on the event $J(K_j') \le \nu/2$ for all $j' \le j$, the claim holds with probability at least $1 - \delta/8T^4$. We show by induction that we can peel-off the conditioning by summing the failure probability of each epoch. Concretely, we show by induction that the claim holds for all $j' \le j$ with probability at least $1 - j\delta/8T^4$. Since the number of epochs is less than T (in fact logarithmic in T), this will conclude the proof.

The induction base follows immediately by our conditional assumption and the fact that $J(K_0) \le \nu/4$. Now, assume the hypothesis holds up to $j-1$. We show that the conditions of Theorem 2 are satisfied with $\bar{f} = \nu/2$ up to round j, and thus $J(K_{j'}) \le \nu/2$ for all $j' \le j$. We can then invoke our conditional assumption and a union bound to conclude the induction step.

We verify the conditions of Theorem 2. First, the Lipschitz, smoothness, and PL conditions hold by Lemma 5. Next, notice that by definition $J_\star \le J(K_0) \le \nu/4$, and so by the induction hypothesis $\|\hat{g}_{j'} - \nabla J(K_{j'})\|_F \le \sqrt{\nu\mu}/4 \le \sqrt{(\bar{f} - f_*)\mu}/2 \le G$, for all $j' < j$. Finally, noticing that $\kappa^2 > d_x$ it is easy to verify the condition on η.

It remains to show the conditional claim holds. The event $J(K_{j'}) \le \nu/2$ for all $j' \le j$ essentially implies that the policy gradient scheme did not diverge up to the start of epoch j. Importantly, this event is independent of any randomization during epoch j and thus will not break any i.i.d. assumptions within the epoch. Moreover, by Lemma 5 and since $r_0 \le \nu/2G$, this implies that $J(K_{j',i}) \le J(K_j) + Gr_j \le \nu$, i.e., $K_{j',i} \in \mathcal{K}$ for all i and $j' \le j$. For the remainder of the proof, we implicitly assume that this holds, allowing us to invoke Lemmas 3 to 5. For ease of notation, we will not specify this explicitly.

Now, let J^r be the smoothed version of J as in Eq. (1). Since $r_j \le D_0$ we can use Lemma 1 to get that

$$\begin{aligned}
&\|\hat{g}_j - \nabla J(K_j)\|_F \\
&\le \|\hat{g}_j - \nabla J^{r_j}(K_j)\|_F + \|\nabla J^{r_j}(K_j) - \nabla J(K_j)\|_F \\
&\le \beta r_j + \|\hat{g}_j - \nabla J^{r_j}(K_j)\|_F,
\end{aligned}$$

Next, we decompose the remaining term using the triangle inequality to get that

$$\begin{aligned}
&\|\hat{g}_j - \nabla J^{r_j}(K_j)\|_F \\
&= \left\| \frac{1}{m_j} \sum_{i=1}^{m_j} \left(\frac{d_x d_u}{r_j} c_{j,i,\tau} U_{j,i} - \nabla J^{r_j}(K_j) \right) \right\|_F \\
&\le \left\| \frac{1}{m_j} \sum_{i=1}^{m_j} \left(\frac{d_x d_u}{r_j} J(K_{j,i}) U_{j,i} - \nabla J^{r_j}(K_j) \right) \right\|_F \\
&\quad + \left\| \frac{1}{m_j} \sum_{i=1}^{m_j} \left(\frac{d_x d_u}{r_j} (c_{j,i,\tau} - J(K_{j,i})) U_{j,i} \right) \right\|_F.
\end{aligned}$$

By Lemma 1, we notice that, conditioned on K_j, the first term is a sum of zero-mean i.i.d random vectors with norm bounded by $2d_u d_x \nu / r_j$. We thus invoke Lemma 13 (Vector

Azuma) to get that with probability at least $1-\delta/16T^4$

$$\left\|\frac{1}{m_j}\sum_{i=1}^{m_j}\frac{d_x d_u}{r_j}J(K_{j,i})U_{j,i}-\nabla J^{r_j}(K_j)\right\|_F \le \frac{d_u d_x \nu}{r_j}\sqrt{\frac{8}{m_j}\log\frac{240T^4}{\delta}}.$$

Next, denote $Z_i = \frac{d_x d_u}{r_j}(c_{j,i,\tau}-J(K_{j,i}))U_{j,i}$, and notice that the remaining term is exactly $\|\frac{1}{m_j}\sum_{i=1}^{m_j}Z_i\|_F$. Let $x_{j,i,\tau}$ be the state during the final round of playing controller $K_{j,i}$, and $\mathcal{F}_i$ be the filtration defined by $K_j, x_{j,1,\tau},\dots,x_{j,i,\tau},U_{j,1},\dots,U_{j,i}$. We use Jensen's inequality and Lemma 3 to get that

$$\begin{aligned}
&\|\mathbb{E}[Z_i\mid\mathcal{F}_{i-1}]\|_F\\
&\le \mathbb{E}[\|\mathbb{E}[Z_i\mid\mathcal{F}_{i-1},K_{j,i}]\|_F\mid\mathcal{F}_{i-1}]\\
&\le \frac{d_x d_u}{r_j}\mathbb{E}\left[|\mathbb{E}[c_{j,i,\tau}\mid\mathcal{F}_{i-1},K_{j,i}]-J(K_{j,i})|\;\middle|\;\mathcal{F}_{i-1}\right]\\
&\le \frac{d_x d_u \nu\kappa^2}{r_j\sigma^2}e^{-2\gamma\tau}\mathbb{E}\left[\|x_{j,i,1}x_{j,i,1}^\mathsf{T}-\Sigma_{K_{j,i}}\|\;\middle|\;\mathcal{F}_{i-1}\right]\\
&\le \frac{37 d_x d_u \nu\kappa^{10}W^2}{r_j\sigma^2}e^{-2\gamma\tau}\\
&\le \frac{d_x d_u \nu\kappa^8 W^2}{r_j\sigma^2 T^2},
\end{aligned}$$

where the last step plugged in the value of τ and the one before that used Lemmas 4 and 5 to bound $\|\Sigma_{K_{j,i}}\|\le\nu/\alpha_0=\kappa^2\sigma^2$ and $\|x_{j,i,1}\|\le 6\kappa^4 W$. Further using Lemma 4 to bound $c_{j,i,\tau}$, we also get that

$$\begin{aligned}
&\|Z_i-\mathbb{E}[Z_i\mid\mathcal{F}_{i-1}]\|_F\\
&\le \|Z_i\|_F+\|\mathbb{E}[Z_i\mid\mathcal{F}_{i-1}]\|_F\\
&\le \frac{d_x d_u c_{j,i,\tau}}{r_j}+\|\mathbb{E}[Z_i\mid\mathcal{F}_{i-1}]\|_F\\
&\le \frac{37 d_x d_u \nu\kappa^8 W^2}{r_j\sigma^2}.
\end{aligned}$$

Since Z_i is $\mathcal{F}_i-$measurable we can invoke Lemma 13 (Vector Azuma) to get that with probability at least $1-\frac{\delta}{16T^4}$,

$$\begin{aligned}
&\left\|\frac{1}{m_j}\sum_{i=1}^{m_j}Z_i\right\|_F \le \frac{1}{m_j}\left\|\sum_{i=1}^{m_j}Z_i-\mathbb{E}[Z_i\mid\mathcal{F}_{i-1}]\right\|_F\\
&+\frac{1}{m_j}\sum_{i=1}^{m_j}\|\mathbb{E}[Z_i\mid\mathcal{F}_{i-1}]\|_F\\
&\le \frac{d_x d_u \nu\kappa^8 W^2}{r_j\sigma^2}\left[37\sqrt{\frac{2}{m_j}\log\frac{240T^4}{\delta}}+\frac{1}{T^2}\right]\\
&\le \frac{54 d_x d_u \nu\kappa^8 W^2}{r_j\sigma^2}\sqrt{\frac{1}{m_j}\log\frac{240T^4}{\delta}}.
\end{aligned}$$

Using a union bound and putting everything together, we conclude that with probability at least $1-(\delta/8T^4)$,

$$\begin{aligned}
&\|\hat{g}_j-\nabla J(K_j)\|_F\\
&\le \beta r_j+\frac{54 d_x d_u \nu\kappa^8 W^2}{r_j\sigma^2}\sqrt{\frac{1}{m_j}\log\frac{240T^4}{\delta}}\\
&= \left[\beta r_0+\frac{54 d_x d_u \nu\kappa^8 W^2}{\sigma^2 r_0 m_0^{1/2}}\sqrt{\log\frac{240T^4}{\delta}}\right]\left(1-\frac{\mu\eta}{3}\right)^{j/2}\\
&\le 2\beta r_0\left(1-\frac{\mu\eta}{3}\right)^{j/2}\\
&\le \frac{\sqrt{\mu\nu}}{4}\left(1-\frac{\mu\eta}{3}\right)^{j/2},
\end{aligned}$$

where the last steps plugged in the values of μ,β,r_0, and m_0. ■

4.3. Proof of Lemma 6

Lemma 6 is a straightforward consequence of the previous results.

Proof. For $j=0$ the claim holds trivially by our assumption that $J(K_0)\le\nu/4$. Now, for $j\ge 1$, we use a union bound on Lemma 9 to get that with probability at least $1-\delta/8T^2$

$$\|\hat{g}_j-\nabla J(K_j)\|\le\frac{\sqrt{\mu\nu}}{4}\left(1-\frac{\mu\eta}{3}\right)^{j/2},\qquad\forall j\ge 0.$$

Then by Theorem 2 we have that

$$\begin{aligned}
J(K_j)&\le J_\star+\frac{\nu}{4}\left(1-\frac{\mu\eta}{3}\right)^{j-1}\\
&\le \min\left\{\frac{\nu}{2}, J_\star+\frac{\nu}{2}\left(1-\frac{\mu\eta}{3}\right)^{j}\right\},
\end{aligned}$$

where the last step used the facts that $J_\star\le J(K_0)\le\nu/4$ and $1-\mu\eta/3\ge 1/2$. ■

5. Exploration Cost Analysis

In this section we demonstrate that exploring near a given initial policy does not incur linear regret in the exploration radius (as more straightforward arguments would give), and use this crucial observation for proving Lemmas 7 and 8.

We begin with Lemma 8. The main difficulty in the proof is captured by the following basic result, which roughly shows that the expected cost for transitioning between two i.i.d. copies of a given random policy scales with the *variance* of the latter. This would in turn give the quadratic dependence on the exploration radius we need.

Lemma 10. *Let $K\in\mathcal{K}$ be fixed. Suppose K_1,K_2 are i.i.d. random variables such that $\mathbb{E}K_i=K$, and $\|K_i-K\|_F\le r\le D_0$. If $x_\tau(K_1)$ is the result of playing K_1 for $\tau\ge 1$ rounds starting at $x_0\in\mathbb{R}^{d_x}$, then*

$$\mathbb{E}[x_\tau(K_1)^\mathsf{T}(P_{K_2}-P_{K_1})x_\tau(K_1)]$$

$$\leq \frac{256 d_x \nu \psi^2 \kappa^{10}}{\alpha_0} r^2 + 32 d_x \psi \kappa^9 (\|x_0\|^2 + \kappa^2 \sigma^2) r e^{-2\gamma\tau}.$$

Proof. Notice that the expectation is with respect to both controllers and the τ noise terms, all of which are jointly independent. We begin by using Lemmas 3 and 5 to get that

$$\begin{aligned}
&\mathrm{Tr}\big((P_{K_2} - P_{K_1})(\mathbb{E}[x_\tau(K_1) x_\tau(K_1)^\mathsf{T} \mid K_1] - \Sigma_{K_1})\big) \\
&\leq 32 d_x \psi \kappa^7 r \|\mathbb{E}[x_\tau(K_1) x_\tau(K_1)^\mathsf{T} \mid K_1] - \Sigma_{K_1})\| \\
&\leq 32 d_x \psi \kappa^9 r e^{-2\gamma\tau} \|x_0 x_0^\mathsf{T} - \Sigma_{K_1}\| \\
&\leq 32 d_x \psi \kappa^9 (\|x_0\|^2 + \kappa^2 \sigma^2) r e^{-2\gamma\tau},
\end{aligned}$$

where the last step also used the fact that $\kappa^2 \sigma^2 = \nu / \alpha_0$. Now, since P_{K_1}, P_{K_2} do not depend on the noise, we can use the law of total expectation to get that

$$\begin{aligned}
&\mathbb{E}[x_\tau(K_1)^\mathsf{T} (P_{K_2} - P_{K_1}) x_\tau(K_1)] \\
&= \mathbb{E}[\mathrm{Tr}\big((P_{K_2} - P_{K_1}) \mathbb{E}[x_\tau(K_1) x_\tau(K_1)^\mathsf{T} \mid K_1]\big)] \\
&\leq \mathbb{E}\mathrm{Tr}((P_{K_2} - P_{K_1}) \Sigma_{K_1}) \\
&\quad + 4 d_x \alpha_0 \kappa^2 (\|x_0\|^2 + \kappa^2 \sigma^2) e^{-2\gamma\tau}.
\end{aligned}$$

To bound the remaining term, notice that since K_1, K_2 are i.i.d, we may change their roles without changing the expectation, i.e.,

$$\mathbb{E}[\mathrm{Tr}((P_{K_2} - P_{K_1})\Sigma_{K_1})] = \mathbb{E}[\mathrm{Tr}((P_{K_1} - P_{K_2})\Sigma_{K_2})],$$

we conclude that

$$\begin{aligned}
&\mathbb{E}[\mathrm{Tr}((P_{K_2} - P_{K_1})\Sigma_{K_1})] \\
&= \frac{1}{2}\mathbb{E}[\mathrm{Tr}((P_{K_2} - P_{K_1})(\Sigma_{K_1} - \Sigma_{K_2}))] \\
&\leq \frac{d_x}{2} \|P_{K_2} - P_{K_1}\| \|\Sigma_{K_2} - \Sigma_{K_1}\| \\
&\leq \frac{256 d_x \nu \psi^2 \kappa^{10}}{\alpha_0} r^2,
\end{aligned}$$

where the last step also used Lemma 5. ■

5.1. Proof of Lemma 8

Before proving Lemma 8 we introduce a few simplifying notations. Since the lemma pertains to a single epoch, we omit its notation j wherever it is clear from context. For example, $K_{j,i}$ will be shortened to K_i and $x_{j,i,s}$ to $x_{i,s}$. In any case, we reserve the index j for epochs and i for sub-epochs. In this context, we also denote the gap between realized and idealized costs during sub-epoch i by

$$\Delta C_i = \sum_{s=1}^{\tau} (c_{i,s} - J(K_i)),$$

and the filtration $\mathcal{H}_i = \sigma(w_{1,1}, \ldots, w_{i,\tau-1}, K_1, \ldots, K_i)$. We note that K_i and ΔC_i are $\mathcal{H}_i-$measurable. The following lemma uses Eq. (2) to decompose the cost gap at the various time resolutions. See proof in the full version of the paper (Cassel & Koren, 2021).

Lemma 11. *If the epoch initial controller satisfies $J(K_j) \leq \nu/2$ then (recall that P_K is the positive definite solution to Eq. (2)):*

(i) $c_{i,s} - J(K_i)$
$= x_{i,s}^\mathsf{T} P_{K_i} x_{i,s} - \mathbb{E}_{w_{i,s}}[x_{i,s+1}^\mathsf{T} P_{K_i} x_{i,s+1}]$;

(ii) $\mathbb{E}[\Delta C_i \mid \mathcal{H}_{i-1}]$
$= \mathbb{E}\big[x_{i,1}^\mathsf{T} P_{K_i} x_{i,1} - x_{i+1,1}^\mathsf{T} P_{K_i} x_{i+1,1} \mid \mathcal{H}_{i-1}\big]$;

(iii) $\sum_{i=1}^{m_j} \mathbb{E}[\Delta C_i \mid \mathcal{H}_{i-1}] \leq \mathbb{E}[x_{1,1}^\mathsf{T} P_{K_1} x_{1,1}]$
$+ \sum_{i=2}^{m_j} \big(\mathbb{E}[x_{i,1}^\mathsf{T} P_{K_i} x_{i,1} \mid \mathcal{H}_{i-1}]$
$- \mathbb{E}[x_{i,1}^\mathsf{T} P_{K_{i-1}} x_{i,1} \mid \mathcal{H}_{i-2}]\big)$.

We are now ready to prove the main lemma of this section.

Proof of Lemma 8. First, by Lemma 6, the event $J(K_{j'}) \leq \nu/2$ for all $j' \leq j$ holds with probability at least $1 - \delta/8T$. As in the proof of Lemma 9, we will implicitly assume that this event holds, which will not break any i.i.d assumptions during epoch j and implies that $K_i \in \mathcal{K}$ for all $1 \leq i \leq m_j$. We also use this to invoke Lemmas 4 and 5 to get that for any $1 \leq i, i' \leq m_j$ and $1 \leq s \leq \tau$ we have $x_{i,s}^\mathsf{T} P_{K_{i'}} x_{i,s} \leq 36 \nu \kappa^8 W^2 / \sigma^2 = \nu_0$.

Now, recall that ΔC_i is $\mathcal{H}_i$-measurable and thus $\Delta C_i - \mathbb{E}[\Delta C_i \mid \mathcal{H}_{i-1}]$ is a martingale difference sequence. Using the first part of Lemma 11 we also conclude that each term bounded by $\tau \nu_0$. Applying Azuma's inequality we get that with probability at least $1 - (\delta/16T)$

$$\begin{aligned}
\sum_{i=1}^{m_j} \Delta C_i &= \sum_{i=1}^{m_j} \Delta C_i - \mathbb{E}[\Delta C_i \mid \mathcal{H}_{i-1}] + \mathbb{E}[\Delta C_i \mid \mathcal{H}_{i-1}] \\
&\leq \sqrt{2 m_j \tau^2 \nu_0^2 \log \frac{16T}{\delta}} + \sum_{i=1}^{m_j} \mathbb{E}[\Delta C_i \mid \mathcal{H}_{i-1}].
\end{aligned}$$

Now, recall from Lemma 11 that

$$\begin{aligned}
&\sum_{i=1}^{m_j} \mathbb{E}[\Delta C_i \mid \mathcal{H}_{i-1}] \\
&\leq \mathbb{E}[x_{1,1}^\mathsf{T} P_{K_1} x_{1,1}] \\
&+ \sum_{i=2}^{m_j} \mathbb{E}[x_{i,1}^\mathsf{T} P_{K_i} x_{i,1} \mid \mathcal{H}_{i-1}] - \mathbb{E}[x_{i,1}^\mathsf{T} P_{K_{i-1}} x_{i,1} \mid \mathcal{H}_{i-2}] \\
&= \mathbb{E}[x_{1,1}^\mathsf{T} P_{K_1} x_{1,1}] \\
&+ \sum_{i=2}^{m_j} \mathbb{E}[x_{i,1}^\mathsf{T} P_{K_i} x_{i,1} \mid \mathcal{H}_{i-1}] - \mathbb{E}[x_{i,1}^\mathsf{T} P_{K_i} x_{i,1} \mid \mathcal{H}_{i-2}] \\
&+ \mathbb{E}[x_{i,1}^\mathsf{T} (P_{K_i} - P_{K_{i-1}}) x_{i,1} \mid \mathcal{H}_{i-2}].
\end{aligned}$$

The first two terms in the sum form a martingale difference sequence with each term being bound by ν_0. We thus have that with probability at least $1 - \delta/16T$,

$$\sum_{i=1}^{m_j} \mathbb{E}[\Delta C_i \mid \mathcal{H}_{i-1}] \leq \nu_0 + \sqrt{2 m_j \nu_0^2 \log \frac{16T}{\delta}}$$

$$+\sum_{i=2}^{m_j}\mathbb{E}[x_{i,1}^\mathsf{T}(P_{K_i}-P_{K_{i-1}})x_{i,1}\mid\mathcal{H}_{i-2}].$$

Notice that the summands in the remaining term fit the setting of Lemma 10 and thus

$$\begin{aligned}
&\sum_{i=2}^{m_j}\mathbb{E}[x_{i,1}^\mathsf{T}(P_{K_i}-P_{K_{i-1}})x_{i,1}\mid\mathcal{H}_{i-2}]\\
&\le\frac{256d_x\nu\psi^2\kappa^{10}}{\alpha_0}r_j^2m_j\\
&+\sum_{i=1}^{m_j}32d_x\psi\kappa^9(\|x_{i,1}\|^2+\kappa^2\sigma^2)r_je^{-2\gamma\tau}\\
&\le\frac{256d_x\nu\psi^2\kappa^{10}}{\alpha_0}r_j^2m_j+\frac{25d_x\psi\kappa^{15}W^2r_jm_j}{T^2}\\
&\le\frac{257d_x\nu\psi^2\kappa^{10}}{\alpha_0}r_j^2m_j,
\end{aligned}$$

where the second transition plugged in τ and used Lemma 4 to bound $\|x_{i,1}\|$, and the third transition used the fact that $T^{-2}\le m_j^{-2}\le r_j/m_0$. Plugging in the value of ν_0 and using a union bound, we conclude that with probability at least $1-\delta/4T$,

$$\begin{aligned}
\sum_{i=1}^{m_j}\Delta C_i&\le\frac{144\nu\kappa^8W^2}{\sigma^2}\tau\sqrt{m_j\log\frac{16T}{\delta}}\\
&\quad+\frac{257d_x\nu\psi^2\kappa^{10}}{\alpha_0}r_j^2m_j,
\end{aligned}$$

as desired. ■

5.2. Proof of Lemma 7

Proof of Lemma 7. By Lemma 6, the event $J(K_j)\le\nu/2$ occurs with probability at least $1-\delta/8T^2$. Similarly to Lemmas 8 and 9, we implicitly assume that this event holds, which does not break i.i.d assumptions inside the epoch and implies that $K_{j,i}\in\mathcal{K}$ for all $1\le i\le m_j$. Now, notice that $\mathbb{E}[K_{j,i}\mid K_j]=K_j$. Since $K_j\in\mathcal{K}$ and $r_j\le D_0$, we can invoke the local smoothness of $J(\cdot)$ (see Lemma 5) to get that

$$\begin{aligned}
\mathbb{E}[J(K_{j,i})\mid K_j]&\le J(K_j)+\nabla J(K_j)^\mathsf{T}\mathbb{E}[K_{j,i}-K_j\mid K_j]\\
&\quad+\frac{1}{2}\beta\mathbb{E}[\|K_{j,i}-K_j\|^2\mid K_j]\\
&=J(K_j)+\frac{1}{2}\beta r_j^2.
\end{aligned}$$

We thus have that

$$\begin{aligned}
&\sum_{i=1}^{m_j}J(K_{j,i})-J(K_j)\\
&\le\frac{1}{2}\beta r_j^2m_j+\sum_{i=1}^{m_j}J(K_{j,i})-\mathbb{E}[J(K_{j,i})\mid K_j].
\end{aligned}$$

The remaining term is a sum of zero-mean i.i.d. random variables that are bounded by ν. We use Hoeffding's inequality and a union bound to get that with probability at least $1-\delta/4T$

$$\sum_{i=1}^{m_j}J(K_{j,i})-J(K_j)\le\frac{1}{2}\beta r_j^2m_j+\nu\sqrt{\frac{1}{2}m_j\log\frac{8T}{\delta}},$$

and plugging in the value of β from Lemma 5 concludes the proof. ■

Acknowledgements

We thank Nadav Merlis for numerous helpful discussions. This work was partially supported by the Israeli Science Foundation (ISF) grant 2549/19, by the Len Blavatnik and the Blavatnik Family foundation, and by the Yandex Initiative in Machine Learning.

References

Abbasi-Yadkori, Y. and Szepesvári, C. Regret bounds for the adaptive control of linear quadratic systems. In *Proceedings of the 24th Annual Conference on Learning Theory*, pp. 1–26, 2011.

Abbasi-Yadkori, Y., Lazic, N., and Szepesvári, C. Model-free linear quadratic control via reduction to expert prediction. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 3108–3117. PMLR, 2019.

Agarwal, N., Hazan, E., and Singh, K. Logarithmic regret for online control. In *Advances in Neural Information Processing Systems*, pp. 10175–10184, 2019.

Bertsekas, D. P. *Dynamic programming and optimal control*, volume 1. Athena scientific Belmont, MA, 1995.

Cassel, A. and Koren, T. Bandit linear control. *Advances in Neural Information Processing Systems*, 33, 2020.

Cassel, A. and Koren, T. Online policy gradient for model free learning of linear quadratic regulators with $\sqrt{T}$ regret. *arXiv preprint arXiv:2102.12608*, 2021.

Cassel, A., Cohen, A., and Koren, T. Logarithmic regret for learning linear quadratic regulators efficiently. In *International Conference on Machine Learning*, pp. 1328–1337. PMLR, 2020.

Chen, X. and Hazan, E. Black-box control for linear dynamical systems. *arXiv preprint arXiv:2007.06650*, 2020.

Cohen, A., Hasidim, A., Koren, T., Lazic, N., Mansour, Y., and Talwar, K. Online linear quadratic control. In *International Conference on Machine Learning*, pp. 1029–1038, 2018.

Cohen, A., Koren, T., and Mansour, Y. Learning linear-quadratic regulators efficiently with only $\sqrt{T}$ regret. In *International Conference on Machine Learning*, pp. 1300–1309, 2019.

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

Flaxman, A. D., Kalai, A. T., and McMahan, H. B. Online convex optimization in the bandit setting: gradient descent without a gradient. In *Proceedings of the sixteenth annual ACM-SIAM symposium on Discrete algorithms*, pp. 385–394, 2005.

Haarnoja, T., Zhou, A., Abbeel, P., and Levine, S. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International Conference on Machine Learning*, pp. 1861–1870. PMLR, 2018.

Hambly, B. M., Xu, R., and Yang, H. Policy gradient methods for the noisy linear quadratic regulator over a finite horizon. *Available at SSRN*, 2020.

Hanson, D. L. and Wright, F. T. A bound on tail probabilities for quadratic forms in independent random variables. *The Annals of Mathematical Statistics*, 42(3):1079–1083, 1971.

Hayes, T. P. A large-deviation inequality for vector-valued martingales. *Combinatorics, Probability and Computing*, 2005.

Hsu, D., Kakade, S., Zhang, T., et al. A tail inequality for quadratic forms of subgaussian random vectors. *Electronic Communications in Probability*, 17, 2012.

Jin, Z., Schmitt, J. M., and Wen, Z. On the analysis of model-free methods for the linear quadratic regulator. *arXiv preprint arXiv:2007.03861*, 2020.

Krauth, K., Tu, S., and Recht, B. Finite-time analysis of approximate policy iteration for the linear quadratic regulator. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Lillicrap, T. P., Hunt, J. J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., Silver, D., and Wierstra, D. Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971*, 2015.

Malik, D., Pananjady, A., Bhatia, K., Khamaru, K., Bartlett, P. L., and Wainwright, M. J. Derivative-free methods for policy optimization: Guarantees for linear quadratic systems. *Journal of Machine Learning Research*, 21(21): 1–51, 2020.

Mania, H., Tu, S., and Recht, B. Certainty equivalence is efficient for linear quadratic control. In *Advances in Neural Information Processing Systems*, volume 32, pp. 10154–10164, 2019.

Mohammadi, H., Jovanovic, M. R., and Soltanolkotabi, M. Learning the model-free linear quadratic regulator via random search. In *Learning for Dynamics and Control*, pp. 531–539. PMLR, 2020.

Nesterov, Y. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2003.

Plevrakis, O. and Hazan, E. Geometric exploration for online control. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 7637–7647. Curran Associates, Inc., 2020.

Silver, D., Lever, G., Heess, N., Degris, T., Wierstra, D., and Riedmiller, M. Deterministic policy gradient algorithms. In *International conference on machine learning*, pp. 387–395. PMLR, 2014.

Simchowitz, M. and Foster, D. Naive exploration is optimal for online lqr. In *International Conference on Machine Learning*, pp. 8937–8948. PMLR, 2020.

Sutton, R. S., McAllester, D. A., Singh, S. P., Mansour, Y., et al. Policy gradient methods for reinforcement learning with function approximation. In *NIPs*, volume 99, pp. 1057–1063. Citeseer, 1999.

Tu, S. and Recht, B. The gap between model-based and model-free methods on the linear quadratic regulator: An asymptotic viewpoint. In *Conference on Learning Theory*, pp. 3036–3083. PMLR, 2019.

Wright, F. T. A bound on tail probabilities for quadratic forms in independent random variables whose distributions are not necessarily symmetric. *The Annals of Probability*, pp. 1068–1070, 1973.

Yaghmaie, F. A. and Gustafsson, F. Using reinforcement learning for model-free linear quadratic control with process and measurement noises. In *2019 IEEE 58th Conference on Decision and Control (CDC)*, pp. 6510–6517. IEEE, 2019.

Yang, Z., Chen, Y., Hong, M., and Wang, Z. Provably global convergence of actor-critic: A case for linear quadratic regulator with ergodic cost. In *Advances in Neural Information Processing Systems*, volume 32, 2019.