2024 American Control Conference (ACC)
July 8-12, 2024. Toronto, Canada

Oracle Complexity Reduction for Model-free LQR: A Stochastic Variance-Reduced Policy Gradient Approach

Leonardo F. Toso, Han Wang, and James Anderson

***Abstract*— We investigate the problem of learning an ϵ-approximate solution for the discrete-time Linear Quadratic Regulator (LQR) problem via a Stochastic Variance-Reduced Policy Gradient (SVRPG) approach. Whilst policy gradient methods have proven to converge linearly to the optimal solution of the model-free LQR problem, the substantial requirement for two-point cost queries in gradient estimations may be intractable, particularly in applications where obtaining cost function evaluations at two distinct control input configurations is exceptionally costly. To this end, we propose an oracle-efficient approach. Our method combines both one-point and two-point estimations in a dual-loop variance-reduced algorithm. It achieves an approximate optimal solution with only $\mathcal{O}\left(\log\left(1/\epsilon\right)^{\beta}\right)$ two-point cost information for $\beta \in (0,1)$.**

I. Introduction

Policy gradient (PG) methods have attracted significant attention in model-free reinforcement learning (RL), in large part due to their simplicity of implementation. Within the context of control, and the LQR problem specifically (where analytic solutions are known), a lot of recent work has focused on connecting system theoretic properties such as controllability, with learning theoretic measures such as sample complexity [1]. As first shown in [2] and further analyzed in [3]–[5], PG methods converge to the global optimal solutions despite the lack of convexity in the LQR problem. This significant result, combined with the adaptability of PG in the model-free setting, has opened up a line of research that addresses classical control problems using PG-based approaches [6], [7].

In the model-free LQR setting, policy gradient descent relies on a finite-sample estimate of the true gradient, often acquired through derivative-free (otherwise known as zeroth-order) methods. We refer the reader to [4] for specific application of zeroth-order methods to LQR control and [8] for general background. Zeroth-order gradient estimation approaches are particularly valuable for applications where the computational resources needed for exact gradient evaluations may be impractical, or when cost-query information is *only* accessible through a black-box procedure.

Despite providing flexibility by avoiding the explicit computation of gradients, zeroth-order gradient estimations with one-point (ZO1P) or two-point (ZO2P) queries frequently produce biased estimations accompanied by large variances [4]. In order to counteract this, large sample sizes are required to accurately estimate the gradients.

Whilst ZO2P provides a reduced variance relative to ZO1P, it necessitates querying the cost function at two distinct control input configurations, which can be prohibitively impractical for certain applications (e.g., robot path planning [9]). Addressing this limitation is crucial for developing efficient approaches applicable to real-world scenarios.

Motivated by these challenges, one line of work focuses on leveraging data from multiple similar systems to mitigate variance and thereby reduce the sample complexity of policy gradient methods [10], [11]. However, for the single-agent setting it remains unclear how we can devise a more computationally efficient approach without resorting to second-order techniques.

On the other hand, in supervised learning and RL, SVRPG approaches have demonstrated their effectiveness in significantly reducing variance and enhancing sample efficiency for PG methods [12], [13]. Such methods leverage the well-known control variate analysis, which incorporates both current and past gradient information to form a descent direction that reduces the estimation's variance. This concept motivates the following question addressed in this work:

Can we design an oracle-efficient solution for addressing the model-free LQR problem by building upon the success of stochastic variance-reduced approaches?.

Our Contributions: Toward this end, our main contributions are summarized as follows:

- This is the first work to propose a stochastic variance-reduced policy gradient algorithm featuring a mixed zeroth-order gradient estimation scheme for tackling the model-free and discrete-time LQR problem.
- Theoretical guarantees demonstrate the convergence (Theorem 2) of our approach, while ensuring stability of the system under the iterated policy (Theorem 1).
- We establish conditions on the problem parameters under which our approach achieves an ϵ-approximate solution with $\mathcal{O}\left(\log\left(1/\epsilon\right)^{3-2\beta}\right)$ queries, while utilizing only $\mathcal{O}\left(\log\left(1/\epsilon\right)^{\beta}\right)$ two-point query information for $\beta \in (0,1)$. This oracle complexity improves upon the best known result $\mathcal{O}\left(\log\left(1/\epsilon\right)\right)$ by a factor of $\mathcal{O}\left(\log\left(1/\epsilon\right)^{1-\beta}\right)$ (Corollary 2).

Main result overview: The SVRPG method we propose requires a slightly larger number of queries, specifically we require $\mathcal{O}\left(\log\left(1/\epsilon\right)^{3-2\beta}\right)$, (this includes one *and* two-point

This material is based upon work supported in part by NSF awards 2144634 & 2231350. Leonardo F. Toso is funded by the Columbia Presidential Fellowship. The authors are with the Department of Electrical Engineering, Columbia University in the City of New York, New York, NY, 10027, USA. Email: {lt2879, hw2786, james.anderson}@columbia.edu.

queries) in comparison to $\mathcal{O}\left(\log(1/\epsilon)\right)$ required by the standard ZO2P approach, in order to achieve an ϵ-approximate solution – the difference is only a logarithmic factor, for *large* β. However, our approach requires considerably fewer two-point queries, specifically a factor of $\mathcal{O}\left(\log\left(1/\epsilon\right)^{1-\beta}\right)$ fewer, for *small* β. This underscores the benefit of our technique, particularly in applications where conducting two-point function evaluations is prohibitively costly.

A. Related Work

Model-free LQR via Policy Gradient Methods: PG methods have been extensively explored as a solution to solve the model-free LQR problem in both discrete [1]–[6] and continuous-time settings [14]–[16]. Despite of the non-convexity of the LQR landscape under the policy search, Fazel et al. [2] proved theoretical guarantees for the global convergence of PG methods for both model-based and model-free settings. Table I summarizes the sample complexity of the aforementioned work.

Although there has been an evident sample complexity reduction from $\mathcal{O}(\frac{1}{\epsilon}\log\left(1/\epsilon\right))$ [4] to $\mathcal{O}\left(\log\left(1/\epsilon\right)\right)$ [5], this is primarily a result of a more refined analysis rather than algorithmic development.[1] In this work, we propose a SVRPG algorithm to reduce the number of two-point queries required to obtain an ϵ-approximate solution for the LQR problem.

Stochastic Variance-Reduced Policy Gradient: Stochastic variance-reduced gradient descent (SVRG) have emerged as a sample-efficient solution technique for non-convex finite-sum optimization problems. Whilst SVRG methods have long been established for non-convex optimization problems (e.g., SVRG [12], SAG [17], and SAGA [18]), their extension to online RL settings is a relatively recent development (e.g., SVRPG [13], [19], [20]). This extension has presented unique challenges, primarily stemming from policy non-stationarity and approximations in the computation of the gradient. Furthermore, SVRPG approaches generally rely on the assumption of unbiased gradient estimation, a condition that rarely holds for derivative-free techniques. This has been addressed in [21], [22] for finite-sum, non-convex problems.

We emphasize that our work does not revolve around a simple extension of the results in [13], [19], [20] (online RL setting) or [21], [22] (non-convex finite-sum problem). In contrast to the latter, our LQR setting encompasses an online optimization problem with a single cost function. As a result, the sampling variance reduction benefit of using zeroth-order variance-reduced methods cannot be simply extended to our setting. On the other hand, in our setting we have the stabilizing policy requirement which is commonly taken for granted in the Markov Decision Process (MDP) case [13], [19], [20] with irreducibly and aperiodicity assumptions on the policy. Moreover, the zeroth-order gradient estimation produces biased estimations. This necessitates further derivations to control this bias as we will discuss later.

[1]We use big-O notation $\mathcal{O}(\cdot)$ to omit constant factors in the argument.

II. Preliminaries

We summarize key policy gradient results for the LQR problem as well as derivative-free optimization techniques.

A. Discrete-time Linear Quadratic Regulator

Consider the discrete-time LTI system

$$x_{\tau+1} = Ax_\tau + Bu_\tau, \quad x_0 \overset{\text{i.i.d.}}{\sim} \mathcal{X}_0, \tag{1}$$

where $x_\tau \in \mathbb{R}^{n_x}$, $u_\tau \in \mathbb{R}^{n_u}$, and x_0 denote the state and input at time τ, and the initial condition. The optimal LQR policy associated with (1) is $u_\tau = -K^* x_\tau$ where K^* solves

$$\underset{K\in\mathcal{K}}{\operatorname{argmin}} \left\{ C(K) := \mathbb{E}_{x_0\sim\mathcal{X}_0}\left[\sum_{\tau=0}^{\infty} x_\tau^\top Q x_\tau + u_\tau^\top R u_\tau\right]\right\}, \quad \text{subject to (1)} \tag{2}$$

where $Q \in \mathbb{S}^{n_x}_{\succ 0}$, $R \in \mathbb{S}^{n_u}_{\succ 0}$, and $\mathcal{K} := \{K | \rho(A-BK) < 1\}$ denotes all stabilizing controllers $K \in \mathbb{R}^{n_u \times n_x}$. The optimal cost is assumed to be finite. This is satisfied when (A, B) is controllable.

In the model-based setting the optimal controller is given by: $K^* := -\left(R + B^\top P B\right)^{-1} B^\top P A$, where $P \in \mathbb{S}^{n_x}_{\succ 0}$ is the solution of the Algebraic Riccati Equation (ARE) [23]. In the absence of the system model (A, B), there is no way to implement an ARE-derived controller. Notably, motivated by the fact that traditional RL techniques aim to find optimal policies for unknown MDPs through direct exploration of the policy space, the line of work led by Fazel et al. [2] and followed by [3]–[5], [14], [16], [24] have proved guarantees for the global convergence of PG methods for both model-based and model-free LQR. This is achieved by leveraging fundamental properties of the LQR cost function. Next, we revisit the updating rule of the model-free LQR problem through policy gradient, as well as its important properties.

Suppose that instead of having the true gradient $\nabla C(K_l)$ at the l-th iteration, we posses a finite-sample estimate $\widehat{\nabla} C(K_l)$. The policy gradient method's update rule for the LQR problem can be expressed as follows:

$$K_{l+1} = K_l - \eta \widehat{\nabla} C(K_l), \quad l = 0, 1, \ldots, L-1 \tag{3}$$

where η represents a positive scalar step-size. We require the following standard assumption [2]–[5].

Assumption 1: We have access to an initial stabilizing controller K_0 such that $\rho(A - BK_0) < 1$.

Remark 1: Note that if the initial controller K_0 fails to stabilize system (1), the PG in (3) cannot iteratively converge to a stabilizing policy since $\widehat{\nabla} C(K_0)$ becomes undefined.

Definition 1: The sublevel set of stabilizing feedback controllers $\mathcal{G} \subseteq \mathcal{K}$ is defined as follows

$$\mathcal{G} := \{K \mid C(K) - C(K^*) \leq \xi \Delta_0\},$$

where $\Delta_0 = C(K_0) - C(K^*)$ and ξ is any positive constant.

Lemma 1: Given two stabilizing policies $K', K \in \mathcal{G}$ such that $\|K' - K\|_F \leq h_\Delta(K) < \infty$, it holds that

$$|C(K') - C(K)| \leq h_{\text{cost}}(K) C(K) \|K' - K\|_F,$$
$$\|\nabla C(K') - \nabla C(K)\|_F \leq h_{\text{grad}}(K) \|K' - K\|_F.$$

TABLE I: Comparison on the sample complexity ($\mathbb{S}_c$), and two-point oracle complexity ($\mathcal{N}_{\text{ZO2P}}$) required to achieve $\mathbb{E}\left(C(K_{\text{out}}) - C(K^*)\right) \leq \epsilon$. Here $\beta \in (0,1)$.

Methods	**Sample Complexity** ($\mathbb{S}_c$)	**Two-point Oracle Complexity** ($\mathcal{N}_{\text{ZO2P}}$)
PG - ZO1P (Fazel et al (2018), [2])	$\mathcal{O}(1/\epsilon^4 \cdot \log(1/\epsilon))$	-
PG - ZO1P (Gravell et al (2019), [3])	$\mathcal{O}(1/\epsilon^4 \cdot \log(1/\epsilon))$	-
PG - ZO1P (Malik et al. (2019), [4])	$\mathcal{O}(1/\epsilon^2 \cdot \log(1/\epsilon))$	-
PG - ZO2P (Malik et al. (2019), [4])	$\mathcal{O}(1/\epsilon \cdot \log(1/\epsilon))$	$\mathcal{O}(1/\epsilon \cdot \log(1/\epsilon))$
PG - ZO2P (Mohammadi et al. (2020), [5])	$\mathcal{O}(\log(1/\epsilon))$	$\mathcal{O}(\log(1/\epsilon))$
SVRPG - Algorithm 2 (This paper)	$\mathcal{O}\left(\log(1/\epsilon)^{3-2\beta}\right)$	$\mathcal{O}\left(\log(1/\epsilon)^{\beta}\right)$

Lemma 2: Let $K^* \in \mathcal{G}$ be the optimal policy that solves (2). Thus, it holds that

$$C(K) - C(K^*) \leq \frac{1}{\lambda}\|\nabla C(K)\|_F^2,$$

for any stabilizing controller $K \in \mathcal{G}$.

A detailed proof of the above lemmas, along with the explicit expressions for $h_\Delta(K)$, $h_{\text{cost}}(K)$, $h_{\text{grad}}(K)$, and λ, can be found in [3]. We direct the reader to [25, Appendix A] for the definition of $\bar{h}_{\text{grad}}$, $\bar{h}_{\text{cost}}$, and $\underline{h}_\Delta$ that are positive coefficients we use further in our derivations.

B. Zeroth-Order Gradient Estimation

Given a positive scalar smoothing radius, denoted as r, and randomly sampled matrices $U_1, \ldots, U_m$ drawn i.i.d. from the uniform distribution $\mathcal{S}_r$ of matrices with $\|U\|_F = r$, and considering a given stabilizing policy $K \in \mathcal{G}$, we define the one-point and two-point zeroth-order gradient estimations of the true gradient $\nabla C(K)$ as follows:

$$\textbf{ZO1P}: \overline{\nabla} C(K) := \sum_{i=1}^{m} \frac{dC(K+U_i)U_i}{mr^2},$$

$$\textbf{ZO2P}: \widetilde{\nabla} C(K) := \sum_{i=1}^{m} \frac{d\left(C(K+U_i) - C(K-U_i)\right)U_i}{2mr^2},$$

where $d = n_x n_u$ and $C(\cdot)$ denotes the true cost value provided by an oracle.

We emphasize that, in practice, we have a finite number of samples denoted by m to compute ZO1P and ZO2P. Consequently, both ZO1P and ZO2P gradient estimation schemes exhibit an inherent bias. In addition, for simplicity we assume access to the true cost, as provided by an oracle [4]. In reality, practical limitations prevent us from simulating our system over an infinite horizon. However, as in [3, Appendix B] the finite horizon approximation for the cost is upper-bounded by the true cost, with the approximation error controllable by the horizon length. Our work can thus be readily extended to this finite-horizon approximated cost setting.

Moreover, the expressions of ZO1P and ZO2P shed light on the fact that whilst ZO2P requires more computational resources due to the need for two cost-query information for each sampled matrix $U \overset{\text{i.i.d.}}{\sim} \mathcal{S}_r$, it offers a lower-variance estimation, which results in a more efficient sample complexity, compared to ZO1P [4]. This makes ZO2P a more favorable choice over ZO1P gradient estimation. Next, we present the PG algorithm with ZO2P gradient estimations for solving the model-free LQR.

Algorithm 1 PG with ZO2P Gradient Estimation.

Input: L, η, n_1, r, K_0
for $l = 0, \ldots, L-1$ **do**
  Compute $\widetilde{\nabla} C(K_l)$ with r via ZO2P
  $K_{l+1} = K_l - \eta \widetilde{\nabla} C(K_l)$
end for
Output $K_{\text{out}} := K_L$

It is well-established [5] that under certain conditions on the quality of the estimated gradient, i.e., with n_1 large and r small, Algorithm 1 converges linearly to the optimal solution of (2) while ensuring $K_l \in \mathcal{G}$ at each iteration. However, due to the still high variance of the gradient estimation step, the required number of two-point queries to achieve an ϵ-approximate solution may become prohibitively large.

III. SVRPG Algorithm for model-free LQR

With the purpose of reducing the number of two-cost query information to achieve an ϵ-approximate solution we propose a SVRPG approach featuring a mixed gradient estimation scheme. The idea is to use a ZO2P gradient estimate in the outer-loop and a ZO1P estimate in the inner-loop so as to lower the computational complexity associated with two-point cost queries compared to Algorithm 1. The need for two-point cost query information arises only periodically instead of at each iteration.

Algorithm 2 LQR via SVRPG

Input: N, T, η, n_1, n_2, r_{out}, r_{in}, $K_T^0 := \widetilde{K}^0 := K_0$.
for $n = 0, \ldots, N-1$ **do**
  $K_0^{n+1} := \widetilde{K}^n := K_T^n$
  Compute $\tilde{\mu} = \widetilde{\nabla} C(\widetilde{K}^n)$ with r_{out} ▷ ZO2P
  for $t = 0, \ldots, T-1$ **do**
    Compute $\overline{\nabla} C(K_t^{n+1})$, $\overline{\nabla} C(\tilde{K}^n)$ with r_{in} ▷ ZO1P
    $v_t^{n+1} = \tilde{\mu} + \overline{\nabla} C(K_t^{n+1}) - \overline{\nabla} C(\tilde{K}^n)$
    $K_{t+1}^{n+1} = K_t^{n+1} - \eta v_t^{n+1}$
  end for
end for
Output $K_{\text{out}} := K_T^N$.

In contrast to Algorithm 1 our SVRPG algorithm divides the total number of iterations into N epochs, each of length T. For each epoch (outer-loop), we estimate gradients using n_1 samples with smoothing radius r_{out}, whereas inside each epoch (inner-loop) we use n_2 samples with smoothing radius r_{in}. In line 3, we fix the current policy $\tilde{K}^n$ and compute $\widetilde{\nabla} C(\tilde{K}^n)$ via ZO2P. Throughout the inner-loop iterations, we estimate $\overline{\nabla} C(K_t^{n+1})$ and $\overline{\nabla} C(\tilde{K}^n)$ with the same set of samples via ZO1P. Finally, in line 8 we perform a gradient descent step, using the stochastic variance-reduced gradient computed in line 7.

To close this section, we briefly discuss the idea behind SVRG-based methods. Consider a fixed stabilizing policy $\tilde{K} \in \mathcal{G}$ and estimate $\widetilde{\nabla} C(\tilde{K})$ using n_1 samples. Then perform $K \leftarrow K - \eta v$, with

$$v = \widetilde{\nabla} C(\tilde{K}) + \overline{\nabla} C(K) - \overline{\nabla} C(\tilde{K}),$$

where $\overline{\nabla} C(K)$ and $\overline{\nabla} C(\tilde{K})$ are estimated by using the *same* set of n_2 samples. Note that $\mathbb{E}\widetilde{\nabla} C(\tilde{K}) = \mathbb{E}\overline{\nabla} C(\tilde{K})$ (see the extended version of our work [25, Appendix D]). Therefore, since $\overline{\nabla} C(K)$, and $\overline{\nabla} C(\tilde{K})$ are correlated through their samples, the variance of the stochastic gradient v might be reduced by controlling the covariance across the gradient estimations. That is, $\mathbf{var}(v) = \mathbf{var}(X - Y) = \mathbf{var}(X) + \mathbf{var}(Y) - 2\,\mathbf{cov}(X, Y)$, with $X = \overline{\nabla} C(K)$, $Y = \widetilde{\nabla} C(\tilde{K}) - \overline{\nabla} C(\tilde{K})$, and $\mathbf{cov}(\cdot, \cdot)$ denotes the covariance operator.

IV. Theoretical Guarantees

Without loss of generality and for the purpose of the theoretical analysis only, set $r_{\text{out}} = r_{\text{in}} = r$ in Algorithm 2. In Proposition 1 we first establish the convergence rate of Algorithm 1. This allows for a fair comparison on the sample and oracle complexities of Algorithm 2, detailed in Corollaries 1 and 2. Moreover, we outline the conditions under which Algorithm 2 converge to the optimal solution (Theorem 2), all while staying within the stabilizing sub-level set (Theorem 1) throughout the algorithm's iterations.

Proposition 1: (Convergence of Algorithm 1) Suppose the smoothing radius, number of samples, and number of iterations are in the order of $n_1 = \mathcal{O}(1)$, $r = \mathcal{O}(\sqrt{\epsilon})$ and $L = \mathcal{O}\left(\log(1/\epsilon)\right)$, respectively. Then, Algorithm 1 achieves and ϵ-approximate solution with $\mathcal{O}\left(\log(1/\epsilon)\right)$ samples.

Remark 2: We stress that linear convergence with ZO2P was first established in [5] for this problem and extended to continuous-time in [16], [24]. However, in [25, Appendix B] we present an alternative and straightforward proof, one that relies simply on the upper bound of the expectation[2] of the estimated gradient, i.e., $\mathbb{E}\|\widetilde{\nabla}(K)\|_F^2$ (Lemma 4) and does not involve proving that $\langle \widetilde{\nabla} C(K), \nabla C(K) \rangle \geq \mu_1 \|\nabla C(K)\|_F^2$, and $\|\widetilde{\nabla} C(K)\|_F^2 \leq \mu_2 \|\nabla C(K)\|_F^2$ are satisfied with high probability, for $\mu_1, \mu_2 \in \mathbb{R}_+$ [5, Section V].

Assumption 2: Let $\overline{g}(K) = \frac{d}{r^2} C(K + U) U$ be a single sample ZO1P gradient estimation with $U \overset{\text{i.i.d.}}{\sim} \mathcal{S}_r$. Then, for any two stabilizing policies K, $K' \in \mathcal{G}$, we assume that

$$\mathbb{E}\|\overline{g}(K) - \overline{g}(K')\|_F \leq C_g \mathbb{E}\|K - K'\|_F.$$

[2]Expectation is taken with respect to $U \overset{\text{i.i.d.}}{\sim} \mathcal{S}_r$ and $x_0 \overset{\text{i.i.d.}}{\sim} \mathcal{X}_0$.

for some positive constant C_g.

Remark 3: Note that this assumption on the local smoothness of the estimated gradient is a standard requirement for variance-reduced algorithms, as established in [26], [27]. In the context of the LQR problem, this assumption has the same flavor as the local Lipschitz condition on the empirical cost function in [4, Section 2].

Next, we present two auxiliary results that are instrumental in proving our main results. First, we control the bias in the zeroth-order gradient estimation (Lemma 3) and establish a uniform bound for ZO2P estimated gradient (Lemma 4).

Lemma 3: (Controlling the bias) Let $\widehat{\nabla} C(K)$ be the ZO1P or ZO2P gradient estimations evaluated at the stabilizing policy $K \in \mathcal{G}$. Then,

$$\mathbb{E}\|\nabla C(K) - \mathbb{E}\widehat{\nabla} C(K)\|_F^2 \leq \mathcal{B}(r) := \left(\bar{h}_{\text{grad}} r\right)^2.$$

Proof: See [25, Appendix D]. ∎

Lemma 4: Let $\widetilde{\nabla}(K)$ be the ZO2P gradient estimation. For any stabilizing policy $K \in \mathcal{G}$, it holds that

$$\mathbb{E}\|\widetilde{\nabla}(K)\|_F^2 \leq 8d^2 \mathcal{B}(r) + 2d^2 \mathbb{E}\|\nabla C(K)\|_F^2.$$

Proof: See [25, Appendix C]. ∎

A. Stability Analysis

We now introduce the conditions on the number of samples $\{n_1, n_2\}$, step-size η and smoothing radius r to ensure that Algorithm 2 produces a stabilizing policy K_{t+1}^{n+1} at each epoch $n \in \{0, \ldots, N-1\}$ and each $t \in \{0, \ldots, T-1\}$.

Theorem 1: (Per-iteration Stability) Given $K_0 \in \mathcal{G}$, suppose we set the number of outer and inner-loop samples such that satisfies $\{n_1, n_2\} \gtrsim \bar{h}_s\left(\frac{\psi}{6}, \delta\right)$, the step-size $\eta \lesssim \frac{r^2 \Delta_0}{\bar{h}_{\text{grad}} d^2}$, and the smoothing radius

$$r \leq \underline{h}_r\left(\frac{\psi}{6}\right) := \min\left\{\underline{h}_\Delta, \frac{1}{\bar{h}_{\text{cost}}}, \frac{\psi}{6\bar{h}_{\text{grad}}}\right\},$$

with $\delta \in (0, 1)$, $\psi := \sqrt{\frac{\lambda \Delta_0}{4}}$. Then, with probability $1 - \delta$, it holds that $K_{t+1}^{n+1} \in \mathcal{G}$, for all n and t.

Proof: A detailed proof with the explicitly expression of $\bar{h}_s\left(\frac{\psi}{6}, \delta\right)$ is provided [25, Appendix E]. ∎

Discussion: We emphasize that, unlike the RL setting in [13], [19], in the LQR optimal control problem, it is imperative to ensure the closed-loop stability of (1) under K_{t+1}^{n+1} for all $n \in \{0, \ldots, N-1\}$ and $t \in \{0, \ldots, T-1\}$. However, despite its dual-loop structure, demonstrating that $K_{t+1}^{n+1} \in \mathcal{G}$ throughout the iterations of Algorithm 2 can be achieved by following a similar approach as outlined in previous works without variance reduction [2]–[5].

To this end, we first set the first iteration as the base case and demonstrate that as long as $K_0 \in \mathcal{G}$ (Assumption 1), then $C(K_1^1) - C(K^*) \leq C(K_0) - C(K^*)$ holds true, indicating that $K_1^1 \in \mathcal{G}$. To establish this, we use the Lipschitz property of the cost function (Lemma 1), along with the gradient domination condition (Lemma 2), and the matrix Bernstein inequality [28, Section 6]. The latter provides the necessary conditions on n_1, n_2 and r to upper

bound $\|\nabla C(K) - v_0^1\|_F \leq \psi$. The stability analysis is then completed by applying an induction step to this base case.

B. Convergence Analysis

We now proceed with our analysis to provide the necessary conditions on the number of samples $\{n_1, n_2\}$, smoothing radius r, step-size η, and total number of iterations NT to ensure the global convergence of Algorithm 2.

Theorem 2: (Convergence Analysis) Suppose we select $n_2 \geq \max\left\{96d^2, \frac{\left(3C_g^2+12\bar{h}_{\text{grad}}^2 d^2\right)T^2}{\bar{h}_{\text{grad}}^2}\right\}$, and $\eta \leq \frac{1}{4\bar{h}_{\text{grad}}}$. Then, the policy K_{out} returned by Algorithm 2 after NT iterations enjoys the following property:

$$\mathbb{E}\left(C(K_{\text{out}}) - C(K^*)\right) \leq \Delta_0 \times \left(1 - \frac{\eta\lambda}{16}\right)^{NT} + \frac{\mathcal{B}(r)\phi}{\lambda n_2}.$$

with $\phi = 120 + 192d^2$.

Proof: Below we provide the proof strategy for this theorem. A detailed proof is presented in [25, Appendix F]. ∎

Proof Sketch: Theorem 2 is proved as follows:
1) With the fact that $K_{t+1}^{n+1} \in \mathcal{G}$ for all $n \in \{0, \ldots, N-1\}$ and $t \in \{0, \ldots, T-1\}$ (Theorem 1), along with Lemma 1 and Young's inequality we can write

$$\mathbb{E}\left(C(K_{t+1}^{n+1}) - C(K_t^{n+1})\right) \leq \frac{3\eta}{4}\mathbb{E}\|\nabla C(K_t^{n+1}) - v_t^{n+1}\|_F^2 - \frac{\eta}{8}\mathbb{E}\|\nabla C(K_t^{n+1})\|_F^2 - \frac{\bar{h}_{\text{grad}}}{2}\mathbb{E}\|K_{t+1}^{n+1} - K_t^{n+1}\|_F^2, \quad (4)$$

2) We control $\mathbb{E}\|\nabla C(K_t^{n+1}) - v_t^{n+1}\|_F^2$ in the above expression by decomposing it into bias and variance terms. In particular, we have: biases from the inner and outer-loop estimations + variance of the ZO2P outer-loop estimation + ZO1P gradient estimation difference at K_t^{n+1} and $\tilde{K}^n$. Both ZO1P and ZO2P biases are controlled in Lemma 3. For the variance of the ZO2P gradient estimation we use Lemma 4 and for the ZO1P gradient difference term we assume local smoothness (Assumption 2). Thus, with $n_2 \geq 96d^2$, we have

$$\mathbb{E}\|\nabla C(K_t^{n+1}) - v_t^{n+1}\|_F^2 \leq \frac{\phi\eta\mathcal{B}(r)}{16n_2} + \tilde{\phi}\mathbb{E}\|K_t^{n+1} - \tilde{K}^n\|_F^2 + \frac{1}{16}\mathbb{E}\|\nabla C(K_t^{n+1})\|_F^2, \text{ with } \tilde{\phi} = \frac{4}{3n_2}\left(\frac{3C_g^2}{2} + 6\bar{h}_{\text{grad}}^2 d^2\right).$$

3) The proof is completed by using the PL condition (Lemma 2) and telescoping (4) over outer and inner-loop iterations, with $n_2 \geq \frac{\left(3C_g^2+12\bar{h}_{\text{grad}}^2 d^2\right)T^2}{\bar{h}_{\text{grad}}^2}$, and $\eta \leq \frac{1}{4\bar{h}_{\text{grad}}}$.

Corollary 1: (Sample Complexity) Under the conditions of Theorem 2, and suppose we select the total number of iterations and smoothing radius according to

$$NT \geq \frac{16\log\left(2\Delta_0/\epsilon\right)}{\eta\lambda}, \quad r \leq \sqrt{\frac{n_2\lambda\epsilon}{2\phi\bar{h}_{\text{grad}}^2}},$$

then Algorithm 2 achieves $\mathbb{E}\left(C(K_{\text{out}}) - C(K^*)\right) \leq \epsilon$ with $\mathcal{O}\left(\log\left(1/\epsilon\right)^{3-2\beta}\right)$ cost queries.

Proof: The total number of cost queries required in Algorithm 2 is given by $\mathbb{S}_c := NTn_2 + Nn_1$. Therefore, since $n_1 = \mathcal{O}(1)$, the sample complexity of Algorithm 2 is dominated by the order of NTn_2. As a result, by setting $N = \mathcal{O}\left(\log\left(1/\epsilon\right)\right)^{\beta}$ and $T = \mathcal{O}\left(\log\left(1/\epsilon\right)\right)^{1-\beta}$, with $\beta \in (0,1)$, Algorithm 2 returns an ϵ-approximate solution with $\mathcal{O}\left(\log\left(1/\epsilon\right)^{3-2\beta}\right)$ total number of cost queries. ∎

Corollary 2: (Oracle Complexity Reduction) Under the conditions of Theorem 2 and Corollary 1, it holds that Algorithm 2 achieves an ϵ-approximate solution with a reduction of $\mathcal{O}\left(\log\left(1/\epsilon\right)\right)^{1-\beta}$ in the two-point cost queries when compared to Algorithm 1, where $\beta \in (0,1)$.

Discussion: Similar to Corollary 1, we select $N = \mathcal{O}\left(\log\left(1/\epsilon\right)\right)^{\beta}$ and $T = \mathcal{O}\left(\log\left(1/\epsilon\right)\right)^{1-\beta}$. Then, we observe that Algorithm 2, with number of outer-loop samples $n_1 = \mathcal{O}(1)$, demands only $\mathcal{O}\left(\log\left(1/\epsilon\right)\right)^{\beta}$ two-point queries (i.e., the more resource-intensive cost queries to obtain) to achieve an ϵ-approximate solution. This improves upon the two-point oracle complexity of Algorithm 1 by a factor of $\mathcal{O}\left(\log\left(1/\epsilon\right)\right)^{1-\beta}$. To verify this we simply note that our algorithm necessitates $\mathcal{N}_{\text{ZO2P}} = Nn_1 = \mathcal{O}\left(\log\left(1/\epsilon\right)\right)^{\beta}$, whereas Algorithm 1 requires $\mathcal{N}_{\text{ZO2P}} = \mathcal{O}\left(\log\left(1/\epsilon\right)\right)$ two-point cost queries to attain $\mathbb{E}\left(C(K_{\text{out}}) - C(K^*)\right) \leq \epsilon$.

V. Numerical Experiments

Numerical experiments [3] are now conducted to illustrate and evaluate the effectiveness of Algorithm 2. To ensure a fair comparison on the performance of the algorithms we set $x_0^\top = [1,1,1]$ for computing the normalized cost gap between the current and optimal cost, namely, $\frac{C(K_l)-C(K^*)}{C(K_0)-C(K^*)}$, and $\mathcal{X}_0 \overset{d}{=} \mathcal{N}(0, I_{n_x})$ for the cost oracle generation.

Consider a unstable system with $n_x = 3$ states and $n_u = 1$ input, where the system and cost matrices are detailed in [25, Appendix G]. We set the initialization parameters of Algorithms 1 and 2 as follows: 1) $r = 1 \times 10^{-4}$, $n_1 = 50$, $\eta = 1 \times 10^{-4}$. 2) $r_{\text{in}} = 5 \times 10^{-2}$, $r_{\text{out}} = 1 \times 10^{-4}$, $n_1 = 50$, $n_2 = 25$, $N = 125$, $T = 4$, $\eta = 1 \times 10^{-4}$.

Figure 1 demonstrates the convergence of Algorithms 1 and 2. It also includes the result for the policy gradient descent under the model-based setting. The latter highlights the limit of how well the PG algorithms discussing in this work can do without knowing the system model.

The figure shows that both Algorithms 1 and 2 achieve an equivalent convergence performance for the specified parameters. We emphasize that Algorithms 1 and 2 use $\mathcal{S}_c = 50000$ and $\mathcal{S}_c = 37500$ cost queries, respectively, to attain $\epsilon = 3 \times 10^{-2}$. Moreover, in terms of two-point queries, Algorithm 2 necessitates only $\mathcal{N}_{\text{ZO2P}} = Nn_1 = 6250$, whereas Algorithm 1 is entirely reliant on two-point queries, requiring 25000 to achieve the same accuracy as shown in the figure. The figure also shows that the performance of Algorithm 1 degrades when the number of two-point queries decreases to 6500. This demonstrates that with our SVRPG approach we are able to effectively reduce the two-point oracle complexity for solving the model-free LQR problem.

[3] Code for exact reproduction of the proposed experiments can be downloaded from `https://github.com/jd-anderson/LQR_SVRPG`

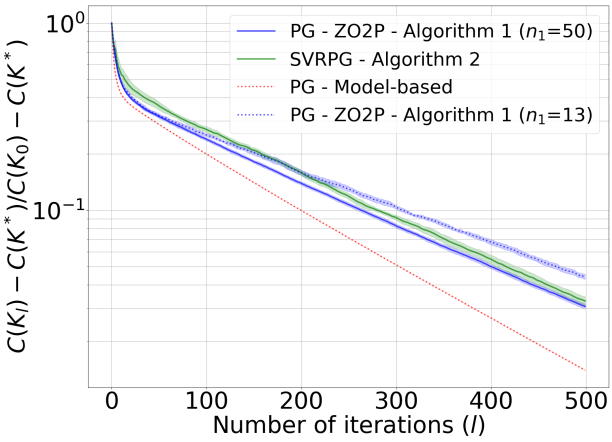

Fig. 1: Normalized gap between the current and optimal cost with respect to the iteration count.

VI. Conclusions and Future Work

We proposed an oracle efficient algorithm to solve the model-free LQR problem. Our approach combines a SVRPG-based approach with a mixed zeroth-order gradient estimation scheme. This mixed gradient estimation yields a reduction in the number of two-point cost queries required to achieve an ϵ-approximate solution since the more resource-expensive queries are now required less frequently. We proved that our approach improves by a factor of $\mathcal{O}\left(\log\left(1/\epsilon\right)\right)^{1-\beta}$ two-point query information upon the standard ZO2P gradient estimation method. Future work will involve exploring loop-less variants and recursive momentum-based approaches to further reduce the two-point oracle complexity required to solve the model-free LQR problem.

References

[1] I. Ziemann, A. Tsiamis, H. Sandberg, and N. Matni, "How are policy gradient methods affected by the limits of control?" in *2022 IEEE 61st Conference on Decision and Control (CDC)*. IEEE, 2022, pp. 5992–5999.

[2] M. Fazel, R. Ge, S. Kakade, and M. Mesbahi, "Global convergence of policy gradient methods for the linear quadratic regulator," in *International conference on machine learning*. PMLR, 2018, pp. 1467–1476.

[3] B. Gravell, P. M. Esfahani, and T. Summers, "Learning optimal controllers for linear systems with multiplicative noise via policy gradient," *IEEE Transactions on Automatic Control*, vol. 66, no. 11, pp. 5283–5298, 2020.

[4] D. Malik, A. Pananjady, K. Bhatia, K. Khamaru, P. Bartlett, and M. Wainwright, "Derivative-free methods for policy optimization: Guarantees for linear quadratic systems," in *The 22nd international conference on artificial intelligence and statistics*. PMLR, 2019, pp. 2916–2925.

[5] H. Mohammadi, M. Soltanolkotabi, and M. R. Jovanović, "On the linear convergence of random search for discrete-time lqr," *IEEE Control Systems Letters*, vol. 5, no. 3, pp. 989–994, 2020.

[6] B. Hu, K. Zhang, N. Li, M. Mesbahi, M. Fazel, and T. Başar, "Toward a theoretical foundation of policy optimization for learning control policies," *Annual Review of Control, Robotics, and Autonomous Systems*, vol. 6, pp. 123–158, 2023.

[7] J. Perdomo, J. Umenberger, and M. Simchowitz, "Stabilizing dynamical systems via policy gradient methods," *Advances in neural information processing systems*, vol. 34, pp. 29 274–29 286, 2021.

[8] J. C. Spall, *Introduction to stochastic search and optimization: estimation, simulation, and control*. John Wiley & Sons, 2005.

[9] J. Peters and S. Schaal, "Reinforcement learning of motor skills with policy gradients," *Neural networks*, vol. 21, no. 4, pp. 682–697, 2008.

[10] H. Wang, L. F. Toso, A. Mitra, and J. Anderson, "Model-free Learning with Heterogeneous Dynamical Systems: A Federated LQR Approach," *arXiv preprint arXiv:2308.11743*, 2023.

[11] Y. Tang, Z. Ren, and N. Li, "Zeroth-order feedback optimization for cooperative multi-agent systems," *Automatica*, vol. 148, p. 110741, 2023.

[12] R. Johnson and T. Zhang, "Accelerating stochastic gradient descent using predictive variance reduction," *Advances in neural information processing systems*, vol. 26, 2013.

[13] M. Papini, D. Binaghi, G. Canonaco, M. Pirotta, and M. Restelli, "Stochastic variance-reduced policy gradient," in *International conference on machine learning*. PMLR, 2018, pp. 4026–4035.

[14] H. Mohammadi, A. Zare, M. Soltanolkotabi, and M. R. Jovanović, "Global exponential convergence of gradient methods over the nonconvex landscape of the linear quadratic regulator," in *2019 IEEE 58th Conference on Decision and Control (CDC)*. IEEE, 2019, pp. 7474–7479.

[15] H. Mohammadi, M. Soltanolkotabi, and M. R. Jovanovic, "Random search for learning the linear quadratic regulator," in *2020 American Control Conference (ACC)*. IEEE, 2020, pp. 4798–4803.

[16] H. Mohammadi, A. Zare, M. Soltanolkotabi, and M. R. Jovanović, "Convergence and sample complexity of gradient methods for the model-free linear–quadratic regulator problem," *IEEE Transactions on Automatic Control*, vol. 67, no. 5, pp. 2435–2450, 2021.

[17] N. Roux, M. Schmidt, and F. Bach, "A stochastic gradient method with an exponential convergence _rate for finite training sets," *Advances in neural information processing systems*, vol. 25, 2012.

[18] A. Defazio, F. Bach, and S. Lacoste-Julien, "SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives," *Advances in neural information processing systems*, vol. 27, 2014.

[19] P. Xu, F. Gao, and Q. Gu, "An improved convergence analysis of stochastic variance-reduced policy gradient," in *Uncertainty in Artificial Intelligence*. PMLR, 2020, pp. 541–551.

[20] Y. Liu, K. Zhang, T. Basar, and W. Yin, "An improved analysis of (variance-reduced) policy gradient and natural policy gradient methods," *Advances in Neural Information Processing Systems*, vol. 33, pp. 7624–7636, 2020.

[21] K. Ji, Z. Wang, Y. Zhou, and Y. Liang, "Improved zeroth-order variance reduced algorithms and analysis for nonconvex optimization," in *International conference on machine learning*. PMLR, 2019, pp. 3100–3109.

[22] S. Liu, B. Kailkhura, P.-Y. Chen, P. Ting, S. Chang, and L. Amini, "Zeroth-order stochastic variance reduction for nonconvex optimization," *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[23] G. Hewer, "An iterative technique for the computation of the steady state gains for the discrete optimal regulator," *IEEE Transactions on Automatic Control*, vol. 16, no. 4, pp. 382–384, 1971.

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

[25] L. F. Toso, H. Wang, and J. Anderson, "Oracle Complexity Reduction for Model-free LQR: A Stochastic Variance-Reduced Policy Gradient Approach," *arXiv preprint arXiv:2309.10679*, 2023.

[26] P. Khanduri, P. Sharma, H. Yang, M. Hong, J. Liu, K. Rajawat, and P. Varshney, "Stem: A stochastic two-sided momentum algorithm achieving near-optimal sample and communication complexities for federated learning," *Advances in Neural Information Processing Systems*, vol. 34, pp. 6050–6061, 2021.

[27] C. Fang, C. J. Li, Z. Lin, and T. Zhang, "Spider: Near-optimal nonconvex optimization via stochastic path-integrated differential estimator," *Advances in neural information processing systems*, vol. 31, 2018.

[28] J. A. Tropp, "User-friendly tail bounds for sums of random matrices," *Foundations of computational mathematics*, vol. 12, pp. 389–434, 2012.