Convergence and Sample Complexity of Policy Gradient Methods for Stabilizing Linear Systems

Feiran Zhao, Xingyun Fu, and Keyou You, Senior Member, IEEE

Abstract—System stabilization via policy gradient (PG) methods has drawn increasing attention in both control and machine learning communities. In this article, we study their convergence and sample complexity for stabilizing linear time-invariant systems in terms of the number of system rollouts. Our analysis is built upon a discounted linear quadratic regulator (LQR) method which alternatively updates the policy and the discount factor of the LQR problem. First, we propose an explicit rule to adaptively adjust the discount factor by exploring the stability margin of a linear control policy. Then, we establish the sample complexity of PG methods for stabilization, which only adds a coefficient logarithmic in the spectral radius of the state matrix to that for solving the LQR problem with a prior stabilizing policy. Finally, we perform simulations to validate our theoretical findings and demonstrate the effectiveness of our method on a class of nonlinear systems.

Index Terms—Policy gradient (PG), sample complexity, stabilization of linear systems, the discounted linear quadratic regulator (LQR).

I. INTRODUCTION

R
cent years have witnessed tremendous success in applications of reinforcement learning (RL) in continuous control [1], [2], [3], and sequential decision-making problems [4], [5]. Most of these successes hinge on the developments of the policy gradient (PG) method, an essential approach in modern RL, which directly searches over the policy space to minimize a cost function. Such an end-to-end approach only requires samples of the cost from system rollouts, and thus is conceptually simple to implement. In fact, the PG method has been directly applied to a wide variety of tasks for complex dynamics [6].

The success of RL has inspired an increasing interest in studying the convergence and sample complexity of PG methods on classical control problems, which serve as ideal benchmarks to understand their efficiency and complexities [7], [8], [9], [10], [11], [12], [13], [14], [15], [16], [17], [18], [19], [20], [21], [22], [23], [24]. The seminal work of [7] first showed that PG methods have global convergence guarantees for the celebrated linear quadratic regulator (LQR) problem. Then, sample complexities of PG methods in terms of the number of system rollouts are established for both discrete-time [9], [10] and continuous-time LQR [11], and they have also been applied to many variants of LQR problems, such as risk-sensitive/constrained LQ control [12], [13], adaptive LQR [24], linear quadratic Gaussian (LQG) [15], [16] and decentralized control [17], [18]. Though these advances lead to fruitful and profound results for PG methods, they all require a common assumption, i.e., a stabilizing policy must be known a prior. This is essential to PG methods [7], [12], [14], [15], since their iterative local search makes sense only if the cost is finite. In fact, learning a stabilizing policy from samples of the cost is a nontrivial task and even of the same importance as solving the LQR problem [25], [26].

Very recently, discount methods have been introduced to address the stabilization problem for linear systems [27], [28] and nonlinear systems [29]. The discount method is referred to as a class of iterative methods that solve a sequence of discounted LQR problems with increasing discount factors, and is originally proposed for escaping local optimal policies in multiagent control systems [30], [31]. However, [27] and [28] did not provide finite-time convergence guarantees. Using a simulator to collect cost samples, [29] proposed a PG method with finite-time convergence to learn a stabilizing policy. However, it does not show the sample complexity in terms of the number of system rollouts even for linear systems.

The sample complexity cannot be explicitly provided in [29] since: 1) their discount method requires to completely solve the discounted LQR problem via PG methods each iteration; and 2) the update of the discount factor relies on a binary or random search, which involves an unpredictable number of iterations. In contrast, we propose a new discount method where: 1) we only need to reduce the cost to a uniform level each iteration via PG methods; and 2) we design an explicit update rule to recursively adjust the discount factor. Our key is to establish a connection between the stability margin of a linear control policy and its discounted LQR cost by using Lyapunov theory, which can be viewed as a natural generalization of the stability margin of the LQR [32]. Using a simulator returning samples of the cost, we characterize the sample complexity for PG methods each iteration and the estimate of the update rate of the discount factor. Then, we provide finite-time convergence guarantees and overall sample complexity for our discount method. Finally, we
validate our theoretical results via simulations and verify that our methods can also efficiently stabilize nonlinear dynamical systems around the equilibrium point.

Our results are the first to reveal that the sample complexity of PG methods for stabilizing unstable linear systems only adds a coefficient logarithmic in the spectral radius of the state matrix to that for solving the LQR problem with a prior stabilizing policy [11]. Moreover, our explicit update rule can be easily used to study sample complexity for other RL-based discount methods, e.g., [27] and [28]. Since only samples of the cost are used, the proposed approach can be generalized to stabilize more complex systems where only a simulator is accessible.

To clarify our contributions in the literature, we further compare with recent methods for system stabilization, which can be broadly categorized as model-based approach using state space data, behavioral-based direct data-driven control, and RL approach using a simulator.

**Model-based approach using state space data:** It learns a descriptive model via system identification (SysID) using state space data [25], [33], [34], [35], [36], [37], [38], [39], [40]. In this context, the sample complexity is quantified in terms of the number of state-input pairs for SysID. A necessary condition for SysID is the persistent excitation (PE) condition, which requires at least \( n + m \) samples with \( n \) and \( m \) being the state and input dimension, respectively. Consequently, their sample complexity typically scales linearly in \( n \). While a recent work [25] improved the complexity to be linear in the number of unstable eigenvalues of the state matrix, they only focused on linear systems with diagonalizable state matrix. In this work, our goal is to strengthen the theoretical foundation of PG methods. Since we only have access to a scalar cost per system rollout, our sample complexity is quadratic in \( n \) in terms of the number of system rollouts. Thus, it is comparable with model-based approach where each sample contains a state vector with dimension \( n \).

**Behavioral-based direct data-driven control:** There is another line of research that directly synthesizes controllers from a given set of system trajectories. The seminal work in behavior system theory [41] establishes Willems’ fundamental lemma to show that all the trajectories of a linear system can be represented by a linear combination of historical trajectories under the PE condition. Then, a data-based semidefinite program (SDP) is designed in [42] and [43] to find a stabilizing controller when additive noise of the system dynamics satisfies a quadratic matrix inequality. Without the PE condition, [44] provided a necessary and sufficient condition for stabilizing noiseless linear systems with state feedback. For the stochastic case, a so-called matrix S-lemma was proposed in [45] and [46] to find a common stabilizing controller of a set of admissible systems consistent with data samples. While this approach provides a novel perspective on the stabilization problem, it relies on the solution to an SDP, the existence of which is sensitive to the data quality [45]. Different from this work, the above references cannot provide any sample complexity in finding a stabilizing controller. In fact, a recent work [26] showed that the sample complexity of stabilization via direct data-driven approaches is almost identical to that of SysID, which is \( n + m \) in terms of the number of state-input pairs. We refer the readers to [47] for a comprehensive review of this direction.

**RL approach using a simulator:** Instead of identifying an explicit dynamical model or solving an SDP, the RL approach directly updates the policy using a simulator [27], [29], [48]. Lamperski [27] was the first to combine the discount method and Q-learning to stabilize deterministic linear systems, while they require to estimate the Q-function using the states and control inputs. In contrast, our PG method does not need to identify any explicit Q-function and only requires the cost of a system rollout from the simulator. Moreover, using our explicit update rule for the discount factor, we can provide finite-time convergence guarantees for their method. A recent work [48] considered the continuous-time linear systems and learned a stabilizing controller by a so-called homotopy policy iteration scheme. This approach gradually pushes the unstable poles to the left-half complex plane with a lower bounded distance each iteration, instead of multiplying a damping factor. Again, they fail to prove finite-time convergence guarantees as well as the sample complexity.

The rest of this article is organized as follows. In Section II, we describe the stabilization problem of linear systems with randomized initial states. In Section III, we propose our explicit discount method algorithm where the update rule for the discount factor is designed by studying the stability margin of a linear control policy. Section IV shows the convergence and sample complexity of our discount method for stabilization in terms of the number of rollouts. Section V performs simulations on both linear and nonlinear systems to verify the effectiveness of the proposed method. Finally, Section VI concludes this article.

**Notations:** We use \( \rho(\cdot) \) to denote the spectral radius of a matrix. We use \( \| \cdot \|_F \) and \( \| \cdot \|_2 \) to denote the Frobenius norm and 2-norm of a matrix, respectively. Let \( \sigma(\cdot) \) be the minimal singular value of a matrix. \( \text{Tr}(\cdot) \) denotes the trace function. Let \( I_n \) be the \( n \)-by-\( n \) identity matrix. Let \( \mathbb{B}^d \) be the unit sphere of dimension \( d - 1 \). We use \( O(\epsilon) \) to denote some constant proportional to \( \epsilon \). We use \( \tilde{O}(\cdot) \) to neglect the logarithmic term in the dimension \( n \) of \( \tilde{O}(\cdot) \).

## II. PROBLEM FORMULATION

Consider the following discrete-time linear time-invariant system:

\[
x_{t+1} = Ax_t + Bu_t, \quad x_0 \sim \mathcal{D}
\]

where \( x_t \in \mathbb{R}^n \) is the state vector and \( u_t \in \mathbb{R}^m \) is the control input vector. The initial state \( x_0 \) is sampled from a distribution \( \mathcal{D} \), on which we make the following assumption.

**Assumption 1:** The distribution \( \mathcal{D} \) has zero mean and covariance \( \mathbb{E}[x_0 x_0^T] = I \) with a bounded support, i.e., \( \|x_0\| \leq d \) for some constant \( d > 0 \).

**Remark 1:** Assumption 1 is made only for simplicity of theoretical analysis, and our results can be extended to the unbounded distributions with strong tail decay, e.g., Gaussian distributions, by using the high-probability bounds and standard truncation
arguments in [9]. In fact, we have validated our main results over the Gaussian distribution in the simulation.

The PG method for the design of (1) first parameterizes the control policy with $K$ and then, updates it in an iterative form

$$K_{j+1} = K_j - \eta \cdot \nabla J(K_j), j \in \mathbb{N}$$

where $\eta > 0$ is a stepsize, $\nabla J(\cdot)$ is the PG of some cost function $J(\cdot)$ to evaluate the performance of the closed-loop system (1), and can be either computed using the knowledge of $(A, B)$ or estimated via samples of $J(K)$. While the PG method has become an essential approach in RL, how to theoretically quantify the sample complexity for convergence certificates remains unclear. Recently, many classical control problems have been adopted as performance benchmarks of the PG method, e.g., LQR [7], risk-constrained LQR [13], and linear quadratic Gaussian [15]. Unfortunately, a stabilizing initial policy $K_0$, i.e., $\rho(A + BK_0) < \infty$, seems indispensable in those works. This is nontrivial to achieve if $(A, B)$ is unknown, and has been acknowledged as an important problem [7]. Though the PG method has been further used to find a stabilizing policy [29], it is unable to quantify its sample complexity, rendering the efficiency of PG methods still unclear. Without loss of generality, we resolve it under the following assumption.

Assumption 2: The system $(A, B)$ is controllable and the open-loop system is unstable, i.e., $\rho(A) \geq 1$. Moreover, we have an a priori knowledge on the upper bound of $\rho(A)$.

Instead of using the exact values of $(A, B)$, we follow the literature [7], [8], [9], [10], [11], [29] that a simulator is available to return a random sample of the cost function, e.g., a sample of $J(K)$ under a randomly selected initial state vector $x_0$. Note that our setting is also consistent with the standard RL approach [3].

III. Stabilization via Discount Methods

In this section, we first introduce the discount method, and then propose an explicit rule to update its discount factor to find a stabilizing gain of (1) via a simulator.

A. Discount Method

Consider the discounted LQR problem

$$\begin{align*}
\text{minimize} & \quad J_\gamma(K) := \mathbb{E}_{x_0} \sum_{t=0}^{\infty} \gamma^t (x_t^T Q x_t + u_t^T R u_t) \\
\text{subject to} & \quad (1) \text{ and } u_t = -K x_t
\end{align*}$$

(2)

where $\gamma \in (0, 1)$ is a discount factor. It is well-known that the discounted cost in (2) is equivalent to the standard LQR cost of a damped dynamical system [9], [29]

$$x_{t+1} = \sqrt{\gamma}(A - BK)x_t, \quad x_0 \sim \mathcal{D}$$

(3)

That is, it holds that

$$J_\gamma(K) = \mathbb{E}_{x_0} \sum_{t=0}^{\infty} (x_t^T Q x_t + u_t^T R u_t)$$

subject to (3) and $u_t = -K x_t$.

(4)

Clearly, the condition for a finite cost $J_\gamma(K)$ is characterized by the stability region of the damped system (3)

$$\mathcal{S}_\gamma = \{K | \sqrt{\gamma} \rho(A - BK) < 1\}.$$ 

Lemma 1 (See [7]): Consider the discounted LQR problem (2). Then, $J_\gamma(K) < \infty$ if and only if $K \in \mathcal{S}_\gamma$. Moreover, the finite cost has a closed-form expression

$$J_\gamma(K) = \text{Tr}(P(\gamma)) = \text{Tr}((Q + K^T R K) \Sigma_K(\gamma))$$

where $P(\gamma)$ and $\Sigma_K(\gamma)$ are positive semidefinite solutions to the Lyapunov equations

$$P(\gamma) = Q + K^T R K + \gamma (A - BK)^T P(\gamma) (A - BK),$$

$$\Sigma_K(\gamma) = I + \gamma (A - BK) \Sigma_K(\gamma) (A - BK)^T,$$

respectively.

(5)

(6)

For simplicity of notation, we use the shorthand $\Sigma := \Sigma_K(\gamma)$ when $K$ and $\gamma$ are clear from the context. To stabilize the system, the discount methods in [27] and [29] alternate between the following two procedures:

(a) Solve $K^{k+1} \in \arg \min_K J_{\gamma^k}(K)$,

(7a)

(b) Search $\gamma^{k+1} > \gamma^k$ such that $\sqrt{\gamma^k} \rho(A - BK^{k+1}) < 1$.

(7b)

Given a fixed $\gamma^k$, the step (7a) finds an optimal gain $K^{k+1}$ of the discounted LQR problem. By the well known stability margin of the LQR [32], the optimal gain $K^*$ is bounded away from the boundary of $\mathcal{S}_{\gamma^k}$, i.e., there exists $\gamma^{k+1} > \gamma^k$ such that $\sqrt{\gamma^k} \rho(A - BK^{k+1}) < 1$. Then, a binary or random search method is used to find such a discount factor $\gamma^{k+1}$ in (7b). For example, a binary search was proposed in [29] to find $\gamma^{k+1} \in (\gamma^k, 1]$ such that

$$2.5 \cdot J_{\gamma^k}(K^{k+1}) \leq J_{\gamma^{k+1}}(K^{k+1}) \leq 8 \cdot J_{\gamma^k}(K^{k+1}).$$

(7b)

Since their discount factor in (7b) is implicitly computed, we refer to (7) as the implicit discount method. In general, the implicit discount method (7) is challenging to implement in practice. Specifically, the minimization of the discounted LQR cost in (7a) requires to evaluate $J_{\gamma^k}(K^{k+1}) - J_{\gamma^k}(K) \leq \epsilon$, which is difficult as $J_{\gamma^k} := \min_K J_{\gamma^k}(K)$ is unknown. Moreover, the sample complexity analysis of (7) in intractable, since the binary or random search in (7b) may result in an unpredictable number of iterations.

In this article, we propose an explicit discount method alternating as

(a) Find $K^{k+1}$ such that $J_{\gamma^k}(K^{k+1}) < D$

(8a)

(b) Update $\gamma^{k+1} = (1 + \xi \alpha^k) \gamma^k$

(8b)

where $D$ and $\xi \in (0, 1)$ are two constants to be specified later, and $\alpha^k > 0$ is the update rate of the discount factor that can be explicitly computed using the value of $J_{\gamma^k}(K^{k+1})$. Instead of minimizing the cost in (7a), we only need to reduce it to a uniform level $D$ in (8a). This ensures a uniform lower bound of the update rate $\alpha^k$ in the explicit rule of the discount factor (8b). By alternating between (8a) and (8b), the explicit discount
method provably returns a stabilizing policy in $S_1$. We provide an illustration of our method in Fig. 1.

In the sequel, we propose a zeroth-order PG method to achieve the cost reduction of step (8a) and design an explicit update rule of the discount factor for the policy in (8b).

**B. Cost Reduction in (8a) via Zeroth-Order PG Methods**

In this section, we propose a zeroth-order PG method to reduce the discounted LQR cost $J_{k_0}(K)$ to a uniform level $D$, i.e., (8a) is satisfied. In fact, the parameter $D$ should be selected such that $D > J_{k_0}^*$ for all $k \in \mathbb{N}$. Thanks to Lemma 2, it suffices to set $D > J_1^*$.

**Lemma 2:** For $0 < \gamma_1 < \gamma_2 \leq 1$, it holds that $J_{\gamma_1}^* < J_{\gamma_2}^*$.

The proof of Lemma 2 follows directly from the definition of discounted LQR cost in (2). Under a fixed discount factor $\gamma$, the zeroth-order PG method for the cost reduction (8a) is proposed as

$$K_{j+1} = K_j - \eta \nabla J_{\gamma}(K_j)$$

where $\eta > 0$ is a constant stepsize\(^1\) and $K_0 \in S_\gamma$. The gradient estimate $\nabla J_{\gamma}(K)$ is obtained via zeroth-order methods, such as minibatching [7] and one-point or two-point methods [9], [11]. In view of [9], we adopt the two-point method to obtain $\nabla J_{\gamma}(K_j)$ in Algorithm 1. In addition, we consider to use a simulator to evaluate the following value function $V_{\gamma,j}^*(K, x_0)$:

$$V_{\gamma,j}^*(K, x_0) = \sum_{i=0}^{\tau-1} (x_i^T Q x_i + u_i^T R u_i)$$

subject to (3) and $u_i = -K x_i$ \hspace{1cm} (10)

where $x_0$ is an initial state vector drawn from the distribution $D$, and $\tau$ is the length of time horizon.

Next, we show that the number of iterations of (9) required to achieve (8a) is finite. In fact, the PG update (9) meets global convergence under a proper stepsize, which is established by exploiting the gradient dominance property and local smoothness of the discounted LQR cost [11]. Specifically, for all $K$ and $K'$ such that the line segment connecting them lies in the sublevel set $\{K | J_{\gamma}(K) \leq \alpha\}$, it holds that

$$J_{\gamma}(K) - J_{\gamma}^* \leq \frac{1}{2\mu} ||\nabla J_{\gamma}(K)||^2,$$

and

$$J_{\gamma}(K') - J_{\gamma}(K) \leq \langle \nabla J_{\gamma}(K), K' - K \rangle + \frac{L(a)}{2} ||K' - K||^2,$$

where $a$ is a positive constant, $\mu = \sigma(R)/(2||\Sigma(\gamma)||)$ is the gradient dominance constant with $\Sigma(\gamma)$ being the solution of (6) under the optimal policy, and $L(a)$ is the smoothness parameter that polynomially depends on the LQR parameters $(a, ||A||, ||B||, ||Q||, ||R||, \gamma)$.

**Lemma 3:** Consider the zeroth-order PG method in (9) to solve the discounted LQR problem (2) with an initial policy $K_0 \in S_\gamma$, let $\nabla J_{\gamma}(K)$ be computed by Algorithm 1, where $\tau \geq \theta_1 \log(1/(D - J_{\gamma}^*))$

$$N_c \geq c(1 + \beta^4 \phi^2 \theta_2^2 \log^6 n)n$$

and $r < \theta_3 (D - J_{\gamma}^*)^{1/2}, \eta \leq 1/(\omega L)$. If the number of iteration $j$ satisfies

$$j \geq \frac{\log((J_{\gamma}(K_0) - J_{\gamma}^*)/(D - J_{\gamma}^*)))}{\log(n/\mu/8 - 1)}$$

Then $J_{\gamma}(K_j) \leq D$ with probability not smaller than

$$1 - c_j(n^{-\beta} + N_c e^{-\beta\theta_2^2} + e^{-c_j N_c}).$$

Here, $\omega = \phi(\sqrt{m} + \beta \phi^2 \theta_2 \sqrt{\log n})$, $\phi$ is a function of $d$, $\mu$ is the gradient dominance constant, $L$ is the smoothness parameter over the sublevel set $\{K \in \mathbb{R}^{n \times n} | J_{\gamma}(K) \leq J_{\gamma}(K_0)\}$, $\beta, c, c', \phi'$ are positive constants, and $L, \theta_1, \theta_2, \theta_3$ are polynomials in the parameters of the discounted LQR problem ($J_{\gamma}(K_0), ||A||, ||B||, ||Q||, ||R||, \varphi(Q), \varphi(R), \gamma$).

The proof essentially follows from [10, Theorem 1] with the focus on our discounted LQR problem and is provided in

---

\(^1\)We use the subscript $j$ to denote the iteration of the policy under a fixed discount factor and use the superscript $k$ to denote the iteration of the discount factor.

---

**Algorithm 1: Two-Point Gradient Estimate of the Cost.**

**Input:** A discount factor $\gamma$, a policy $K \in S_\gamma$, a smoothing radius $r$, a time horizon $\tau$, and a number of system rollouts $N_c$.

1: for $i = 1, \ldots, N_c$ do
2: Sample a perturbation matrix $U_i$ uniformly from the unit sphere $U^{m \times 1}$.
3: Set $K_{i,1} = K + r \sqrt{m} U_i$ and $K_{i,2} = K - r \sqrt{m} U_i$.
4: Sample an initial state $x_{i,0}$ using the distribution $D$.
5: Simulate system (3) to obtain $V_{\gamma}^*(K_{i,1}, x_{i,0}^*)$ and $V_{\gamma}^*(K_{i,2}, x_{i,0}^*)$.
6: end for

**Output:** Gradient estimate

$$\nabla J_{\gamma}(K) = \frac{1}{2r N_c} \sum_{i=1}^{N_c} \left( V_{\gamma}^*(K_{i,1}, x_{i,0}^*) - V_{\gamma}^*(K_{i,2}, x_{i,0}^*) \right) U_i.$$
Appendix A. It is worthy of mentioning that $\tilde{O}(\log(1/\epsilon))$ is the state-of-the-art sample complexity to achieve an optimality gap $|J_{\gamma}(K) - J_{\gamma}^*| \leq \epsilon$ of (9) [10]. For a fixed discount factor $\gamma^k$, Lemma 3 provides the total sample complexity to achieve (8a). In view of (11) and (12), such a sample complexity depends crucially on the initial cost $J_{\gamma}(K_0)$ and the parameters $\gamma, \mu, L, \omega$, $\theta_i, i \in \{1, 2, 3\}$, which vary with the iteration index $k$. Then, we cannot simply take a sum over sample complexities of (9) for each discount factor, and the design of the update rule in (8b) is essential. For example, the one in (7b) of the existing works [27], [29] cannot be used to quantify the overall sample complexity for finding a stabilizing gain. In the sequel, we specify the update rule for the discount factor in (8b) to address it.

C. Explicit Update Rule for the Discount Factor in (8b)

To specify the update rule in (8b), we exploit the relation between the stability margin and the cost of a linear control policy. Different from the classical stability margin of the LQR [32], we study the following problem regarding the stability margin of an arbitrary policy $K$:

**Given a policy $K \in S_\gamma$, explicitly find a larger discount factor $\gamma' > \gamma$ such that $K \in S_{\gamma'}$ without the knowledge of $(A, B)$.**

We first provide a sufficient condition on $\gamma'$ by using the Lyapunov stability theory.

**Lemma 4:** If $K \in S_\gamma$, and $\gamma'$ satisfies

\[
(1 - \gamma/\gamma') P < Q + K^\top R K
\]

where $P$ is a positive semidefinite solution of the Lyapunov equation in (5), then $K \in S_{\gamma'}$.

**Proof:** Consider the following damped system:

\[
x_{t+1} = \sqrt{\gamma'}(A - BK)x_t.
\]

Define $V(x) = x^\top P x$. Clearly, $V(x)$ satisfies that $V(0) = 0$ and $V(x) > 0 \forall x \neq 0$. Moreover, we obtain

\[
V(x_{t+1}) - V(x_t) = \gamma' x_t^\top(A - BK)\top P(A - BK)x_t - x_t^\top P x_t
\]

\[
= x_t^\top \left( \frac{\gamma'}{\gamma}(P - Q - K^\top R K) - P \right) x_t
\]

where the last equality follows from (5).

Jointly with (14), it follows that $V(x_{t+1}) - V(x_t) < 0$. Thus, $V(x)$ is a Lyapunov function for (15) and the damped system (15) is asymptotically stable.

However, it is impossible to directly obtain $\gamma'$ by Lemma 4 as the computation of $P$ involves the unknown $(A, B)$. Next, we use Lemma 1 to show that (14) can be achieved by using the discounted cost $J_{\gamma}(K)$.

**Theorem 1:** If $K \in S_\gamma$ and $\gamma' \leq (1 + \alpha)\gamma$ with

\[
\alpha = \frac{\sigma(Q + K^\top R K)}{J_{\gamma}(K) - \sigma(Q + K^\top R K)},
\]

Then, $K \in S_{\gamma'}$.

**Proof:** A sufficient condition for (14) is that

\[
1 - \gamma/\gamma' < \sigma(Q + K^\top R K)/\text{Tr}(P).
\]

The proof follows from Lemma 1 that $J_{\gamma}(K) = \text{Tr}(P)$.

Theorem 1 reveals how much the system matrices $A$ and $B$ can be scaled to maintain stability of $\sqrt{\gamma}(A - BK)$ for a given policy $K$. Noting that a sufficient condition for (16) is $\alpha \geq \sigma(Q)/(J_{\gamma}(K) - \sigma(Q))$, the scaling is inversely proportional to the cost $J_{\gamma}(K)$. This is consistent with the intuition that a lower cost usually corresponds to a smaller spectral radius of the closed-loop system.

**Remark 2:** Herein we only consider that $A$ and $B$ are scaled in our discount method, and use the discount factor to quantify the stability margin. How to rigorously evaluate its connection to the classical notion of stability margin [32] is beyond the scope of this article.

Theorem 1 suggests an explicit update rule for the discount factor in the form of (8b). Specifically, we can set the new discount factor as $\gamma' = (1 + \alpha)\gamma$, where $\alpha$ is the update rate given by (16). The computation of $\alpha$ is trivial as $\sigma(Q + K^\top R K)$ is known, and the cost function $J_{\gamma}(K)$ can be simply estimated by random sampling, i.e.,

\[
\hat{J}_{\gamma}^N(K) = \frac{1}{N} \sum_{i=1}^{N} V_{\gamma}(K, x_0)
\]

where $N \in \mathbb{N}_+$ and $x_0^i, i \in \{1, 2, \ldots, N\}$ are initial states independently drawn from the distribution $D$. Then, we need to show that the estimate error $|\hat{J}_{\gamma}^N(K) - J_{\gamma}(K)|$ induced by Monte Carlo sampling and truncation can be well controlled with high probability. However, since the existing results (e.g., [7, Lemma 30]) make the number of system rollouts depend on the accuracy and the discount factor, they cannot be directly adopted.

In the following lemma, we show that the estimate error can be bounded by using only a constant number of system rollouts $N$.

**Lemma 5:** For $K \in S_\gamma$ and a constant $0 < \delta < 1$, let the horizon $\tau$ and the number of rollouts $N$ satisfy

\[
\tau \geq \frac{2 J_{\gamma}(K)}{\sigma(Q)} \log \left( \frac{J_{\gamma}(K)\delta^2}{\sigma(Q)} \right), \quad N \geq 8d^4 \log \left( \frac{2}{\delta} \right).
\]

Then, with probability not smaller than $1 - \delta$, it holds that

\[
|J_{\gamma}(K) - \hat{J}_{\gamma}^N(K)| \leq \frac{1}{2} J_{\gamma}(K).
\]

To prove it, we first show that an upper bound of $|V_{\infty}(K, x_0) - V_{\gamma}(K, x_0)|$ decays exponentially with the horizon $\tau$ in Lemma 9, which is a discrete-time counterpart of [11, Lemma 11]. Then, we select the horizon $\tau$ and the right-hand side of (19) such that $N$ is independent of the discount factor. Finally, we use concentration bounds for the bounded random variable $x_0$ to complete the proof. We provide the details in Appendix B.

By Lemma 5, the cost $J_{\gamma}(K)$ can be upper bounded by the estimate $2\hat{J}_{\gamma}^N(K)$ with high probability. Thus, we can design the update rate by (16) as

\[
\alpha = \frac{\sigma(Q + K^\top R K)}{2 \hat{J}_{\gamma}^N(K) - \sigma(Q + K^\top R K)}.
\]
Algorithm 2: The Explicit Discount Method.

Input: An initial policy $K^0 = 0$, a discount factor $\gamma^0 < 1/\rho^2(A)$, a decay factor $0 < \xi < 1$, a number of PG iterations $M$, a stepsize $\eta$, a bound of the initial state $d$, and a probability $\delta$.

1: for $k = 0, 1, \cdots$ do
2:   Cost reduction: Let $K_0 = K^k$. Iterate
3:   
4:   Update of the discount factor: $\gamma^{k+1} = (1 + \xi \alpha^k) \gamma^k$ with
5:   
6: end for

Output: A stabilizing policy.

Moreover, Lemma 5 shows that the sample complexity for estimating $\alpha^k$ is a constant during the iteration. In comparison, it is impossible for [29] to derive such an explicit bound via binary or random search.

In our discount method (8b), we additionally multiply a decay factor $\xi \in (0, 1)$ to the update rate. The motivation is that, while Theorem 1 ensures the stability of the damped closed-loop system with the new discount factor, it does not provide any explicit bound for the spectral radius. By introducing the decay factor, we ensure that $\sqrt{\gamma^{k+1}} \rho(A - BK^{k+1})$ is “strongly” stable for $k \in \mathbb{N}$, i.e., the spectral radius of the damped closed-loop matrix is bounded strictly away from one. We will elaborate on this point in the next section.

Our explicit discount method in the form of (8) is summarized in Algorithm 2. By using the prior knowledge of an upper bound of $\rho(A)$, we can set $\gamma^0 \leq 1/\rho^2(A)$ such that the zero policy $K^0 = 0$ is able to stabilize the damped system with $\gamma^0$.

In the next section, we provide convergence and sample complexity guarantees for Algorithm 2.

IV. CONVERGENCE AND SAMPLE COMPLEXITY OF THE PG-BASED DISCOUNT METHOD FOR STABILIZATION

In this section, we first show finite-iteration convergence of Algorithm 2 and the sample complexity in terms of the number of system rollouts. Then, we discuss our sample complexity result, compare it with the model-based approach, and discuss the extensions of the explicit discount method.

A. Convergence and Sample Complexity of Algorithm 2

Without loss of generality, we assume that $\rho(A) \geq 1$, which leads to $\gamma^0 \leq 1/\rho^2(A) \leq 1$. We now show the convergence and sample complexity of Algorithm 2.

Theorem 2: Let the parameters in Algorithms 1 and 2 satisfy

$$ N \geq 8d^4 \log\left(\frac{2}{\delta}\right), \quad D = \max\{2J^*, J_\gamma(K^0)\}, $$

$$ \tau \geq \max\left\{2D \frac{\|Q\|}{\|\sigma(Q)\|} \log\left(\frac{1}{J^*}\right) \right\}, $$

$$ N_c \geq c(1 + \beta^4 \delta^2 \log^6 n)n, \quad r \geq \theta \sqrt{J^*_1}, \quad \eta \leq \frac{1}{wL}, $$

$$ M \geq \frac{2\tau J^*_1 \eta^2}{\xi^2 \|Q\|^2} \log\left(\frac{3D^2}{\|\sigma(Q)\|^{1.8}}\right). $$

(22)

If

$$ k \geq \frac{\log(1/\delta)}{\log\left(1 + \xi \frac{\|Q\|}{3D - \|\sigma(Q)\|}\right)} $$

Then, with probability not smaller than

$$ 1 - \delta = c'M(n^{-\beta} + N_{e^{-\beta}} + N_c e^{-\beta} + e^{-cN_c}) $$

(23)

Algorithm 2 returns a stabilizing policy $K^k$ such that

$$ \rho(A - BK^k) < \sqrt{1 + \frac{(1 - \xi)\|\sigma(Q)\|}{3D}}. $$

Here, $\bar{w} = c'\left(\sqrt{\bar{m}} + \bar{\phi} \bar{\theta} \sqrt{\bar{m} \log n}\right)^{\frac{1}{2}}, \bar{\theta}_1, \bar{\theta}_2, \bar{\theta}_3$, $\bar{\mathcal{L}}$ are positive polynomials in $(\|A\|, \|B\|, \|Q\|, \|R\|, \|\sigma(Q)\|, \|\sigma(R)\|)$, $0 < \delta < 1$ is a user-specified constant, and $\phi, \beta, c, c'$ are the constants in Lemma 3.

As many variants of LQR problems in [12], [13], [14], [15], [16], [17], [18], [19], [20], [21], and [22], we also adopt the convergence analysis approaches in [7], [8], [9], [10], and [11] for the proof. Since each LQ problem has its unique features, using those approaches to a specific LQ problem is nontrivial. Specifically, here we are dealing with a sequence of discounted LQR problems where both the discount factor and the initial cost vary with iterations and require to provide uniform bounds. Moreover, algorithm parameters need to be carefully designed to quantify the sample complexity. Thus, the existing references cannot be used to solve these technical challenges.

To prove Theorem 2, we show that under the parameter selection in (22), the update rate of the discount factor $\alpha^k$ satisfies Theorem 1. Hence, we are able to provide uniform bounds for the initial cost per discount factor by deriving novel bounds of Lyapunov equations. Under the parameter selection in (22), we show that the requirement in (8a) can be satisfied with high probability. Furthermore, the update rate $\alpha^k$ in (8b) has a uniform lower bound, which leads to finite-iteration convergence guarantees of Algorithm 2.

B. Proof of Theorem 2

We first specify a condition for Algorithm 2 to return a stabilizing policy in finite iterations.

Lemma 6: Suppose that for all $k$, it holds that

$$ J_\gamma^*(K^{k+1}) \leq D, \quad J_\gamma^*(K^{k+1}) - J_\gamma^*(K^{k+1}) \leq \frac{J_\gamma^*(K^{k+1})}{2}, $$

(25a)

(25b)
Then, for $k \geq (3D - \sigma(Q)) \log(1/\gamma^0)/(\xi \sigma(Q))$, Algorithm 2 returns a stabilizing policy $K^k$ with
\[
\rho(A - BK^k) < \sqrt{1 - \frac{(1 - \xi)\sigma(Q)}{3D}}.
\] (26)

**Proof:** By the hypothesis (25), the update rate $\alpha^k$ in (21) satisfies
\[
\alpha^k \geq \frac{\sigma(Q)}{2J_{\gamma,N}^k(K^{k+1}) - \sigma(Q)} \geq \frac{\sigma(Q)}{3D - \sigma(Q)}.
\] (27)
Since $\alpha^k$ has uniform lower bounds, Algorithm 2 returns a discount factor larger than one in
\[
\frac{\log(1/\gamma^0)}{\log(1 + \xi \sigma(Q)/(3D - \sigma(Q)))}
\] iterations.

Next, we show that the spectral radius of the damped system is upper bounded, i.e.,
\[
\sqrt{\gamma^k} \rho(A - BK^k) \leq \sqrt{1 - \frac{(1 - \xi)\sigma(Q)}{3D}} \quad \forall k \in \mathbb{N}_+.
\] (28)

By hypothesis (25a), at the $k$th iteration we have $J_{\gamma,N}(K^{k+1}) \leq D$ and $\sqrt{\gamma^k} \rho(A - BK^{k+1}) < 1$. Thus, the lower bound of $\alpha^k$ in (27) satisfies
\[
\alpha^k \geq \frac{\sigma(Q)}{3D - \sigma(Q)} \geq \frac{\sigma(Q)}{J_{\gamma,N}(K^{k+1}) - \sigma(Q)}.
\]
By Theorem 1, we have $\sqrt{(1 + \alpha^k)^k} \rho(A - BK^{k+1}) < 1$. Then, the new discount factor $\gamma^{k+1} = (1 + \xi \alpha^k)^k$ yields that
\[
\sqrt{\gamma^{k+1}} \rho(A - BK^{k+1})
\]
\[
= \sqrt{(1 + \xi \alpha^k)^k} \rho(A - BK^{k+1})
\]
\[
= \sqrt{(1 + \xi \alpha^k)^k} \cdot (1 + \alpha^k)^k \rho(A - BK^{k+1})
\]
\[
< \sqrt{(1 + \xi \alpha^k)^k}
\]
\[
= \sqrt{1 - \frac{1 - \xi}{(\alpha^k)^{-1} + 1}}
\]
By the bounds on $\alpha^k$ in (27), it further leads to (28). Thus, Algorithm 2 returns a stabilizing policy with the spectral radius of the closed-loop system being upper bounded by $\sqrt{1 - (1 - \xi)\sigma(Q)/(3D)}$.

**Remark 3:** Lemma 6 explicitly shows the dependence of the number of iterations and performance of Algorithm 2 on the user-specified parameters $\gamma^0$, $D$, and $\xi$. In particular, the upper bound of $\rho(A - BK^k)$ significantly relies on the coefficient $\xi$. If $\xi = 1$, then the resulted closed-loop matrix is only asymptotically stable and it is impossible to find a uniform bound for $\rho(A - BK^k)$.

We use induction to prove Theorem 2, i.e., if at the current iteration (25) holds, then with high probability it still holds in the next iteration. Then, Lemma 6 can be applied to prove the convergence. The proof relies on the following technical lemma.

**Lemma 7:** At the $k$th iteration, if the condition (25) holds, then it follows that:
\[
\alpha^k \geq \frac{\sigma(Q)}{3D - \sigma(Q)},
\] (29)
\[
J_{\gamma+1}(K^{k+1}) \leq \frac{3D^2}{(1 - \xi)\sigma(Q)}.
\] (30)

**Proof:** By the derivation in (27), (29) holds. Next, we show that (30) also holds.

It follows from the hypothesis $J_{\gamma,N}(K^{k+1}) \leq D$ in (25a) and the trace inequality that
\[
D \geq J_{\gamma,N}(K^{k+1})
\]
\[
= \text{Tr}((Q + (K^{k+1})^{T} R K^{k+1})\Sigma_{K^{k+1}}(\gamma^k))
\]
\[
\geq \text{Tr}((Q + (K^{k+1})^{T} R K^{k+1})\Sigma(\Sigma_{K^{k+1}}(\gamma^k)))
\]
\[
\geq \text{Tr}(Q + (K^{k+1})^{T} R K^{k+1})
\]
where the last inequality follows from $\Sigma_{K^{k+1}}(\gamma^k) \geq I_n$. By Lemma 1, it holds that
\[
J_{\gamma,N}(K^{k+1}) = \text{Tr}((Q + (K^{k+1})^{T} R K^{k+1})\Sigma_{K^{k+1}}(\gamma^k))
\]
\[
\leq \text{Tr}(Q + (K^{k+1})^{T} R K^{k+1})||\Sigma_{K^{k+1}}(\gamma^k)||
\]
\[
\leq D||\Sigma_{K^{k+1}}(\gamma^k)||.
\] (31)
For brevity, let $\Sigma^k := \Sigma_{K^{k}}(\gamma^k)$. By definition, it holds that $\Sigma^k = \sum_{t=0}^{\infty}(\gamma^k)^t(A - BK^t)^t(A - BK^t)^t$. Then, it follows from (28) that:
\[
||\Sigma^k|| \leq \sum_{t=0}^{\infty} \left(1 - \frac{1 - \xi}{3D} \sigma(Q)\right)^t = \frac{3D}{(1 - \xi)\sigma(Q)}.
\]
Inserting the above into (31) completes the proof.

**Lemma 8:** Under the parameter selection in (22), if at iteration $k - 1$ the condition (25) holds, then at iteration $k$ it still holds with probability not smaller than
\[
1 - \left(\delta + \epsilon N e^{-\beta} + N e^{-\gamma} + e^{c N e}\right) \quad \forall k \in \mathbb{N}.
\]

**Proof:** Suppose that at iteration $k - 1$ the condition (25) holds. We derive the failure probability of (25a) and (25b) at iteration $k$, respectively, and use union bound to prove this lemma.

First, we derive the failure probability of (25a). To this end, we show that the parameters $\tau, N, r, \beta, M$ in (22) satisfy the conditions in Lemma 3.

By Lemma 7, we have
\[
J_{\gamma,N}(K^{k+1}) \leq \frac{3D^2}{(1 - \xi)\sigma(Q)}.
\]

2The parameters $\gamma, \mu, L, \omega, \theta_i, i \in \{1, 2, 3\}$ in Lemma 3 depend on the iteration index $k$. In this subsection, we add a superscript to these notations, i.e., $\gamma^k, \mu^k, L^k, \omega^k, \theta^k_i, i \in \{1, 2, 3\}$.
Moreover, by definition it holds that $J_{\gamma}^* \leq J_{\gamma}^*(K^k)$. Then, since the parameters $L^k, \theta_1^k, \theta_2^k, \theta_3^k$ in Lemma 3 are polynomials in $(||A||, ||B||, ||R||, \sigma(Q), \varrho(R), \gamma, J_{\gamma}^*(K^k))$, there exist uniform bounds $\bar{\theta}_1, \bar{\theta}_2, \bar{\theta}_3$ such that $L^k \leq \bar{\theta}_1, \theta_1^k \leq \bar{\theta}_2, \theta_2^k \geq \bar{\theta}_3$. Combining $D - J_{\gamma}^* > 2J_1 - J_1^* = J_1^*$, it follows that $\tau, N_e, \tau, \eta$ in (22) are a sufficient condition for that in Lemma 3.

To show that $M$ satisfy the conditions in Lemma 3, we provide an upper bound for the right-hand side of (12), which at the $k$th iteration is

$$\frac{-\log((J_{\gamma}^*(K^k) - J_{\gamma}^k)/(D - J_{\gamma}^k)))}{\log(1 - \mu^k \eta^k/\gamma)}$$

where $\mu^k = \varrho(R) / (2||\Sigma^*(\gamma)||)$ is the gradient dominance constant, and $\eta^k$ is the stepsize. Using the inequality $\log(1 + x) \leq x$ by

$$8 \log((J_{\gamma}^*(K^k) - J_{\gamma}^k)/(D - J_{\gamma}^k))) < \log \left(\frac{3D^2}{(1 - \xi)(\varrho(Q)J_1^*)}\right).$$

The denominator of (33) can be upper bounded by

$$8 \log \left(\frac{J_{\gamma}^*(K^k) - J_{\gamma}^k}{D - J_{\gamma}^k}) < \log \left(\frac{3D^2}{(1 - \xi)(\varrho(Q)J_1^*)}\right).$$

For the denominator of (33), we note that the gradient dominance constant has a uniform lower bound $\mu^k \geq \varrho(R) / (2J_1^*)$ for $k \in \mathbb{N}$, which follows from $J_1^* \geq J_{\gamma}^k \geq ||\Sigma^*(\gamma)||$. The stepsize $\eta^k$ is upper bounded by $1/(\sqrt{\bar{\theta}_1 L})$ as indicated in (22). Then, it suffices to set the number of PG updates $M$ as

$$M \geq \frac{LJ_1^*}{\varrho(R)} \log \left(\frac{3D^2}{(1 - \xi)(\varrho(Q)J_1^*)}\right).$$

Thus, by Lemma 3, the PG updates under the condition in (22) fail to return $J^k(K^{k+1}) \leq D$ with probability no larger than $c^k M(n^{-\beta} + N_\alpha^{-\beta} + N_e^{-\beta} + e^{-c N_e})$.

Next, we derive the failure probability of (25b), assuming that $J^k(K^{k+1}) \leq D$. It is straightforward to check that the parameters in (22) satisfy the assumptions in Lemma 5. Clearly, $\tau$ and $N$ in (22) are sufficient conditions for (18) since $J_{\gamma}^*(K^{k+1}) \leq D$. Thus, at the $k$th iteration, $|J_{\gamma}^*(K^k) - J_{\gamma}^k| \leq 2J_1^*/2$ has the failure probability not larger than $\delta$. Using union bound, the proof is completed.

Following the proof of Lemma 8, it is straightforward to show that at $k = 0$, the condition (25) also holds with probability not smaller than

$$1 - \left(\delta + c^k M(n^{-\beta} + N_\alpha^{-\beta} + N_e^{-\beta} + e^{-c N_e})\right).$$

Thus, by union bound the condition (25) holds for $k \in \mathbb{N}$ with probability not smaller than

$$1 - k \left(\delta + c^k M(n^{-\beta} + N_\alpha^{-\beta} + N_e^{-\beta} + e^{-c N_e})\right), \forall k \in \mathbb{N}.$$

Applying Lemma 6 completes the proof of Theorem 2.

C. Discussion on the Convergence, Sample Complexity, and Extensions of Our Discount Method

Theorem 2 shows that with a large probability, our PG-based discount method returns a stabilizing policy in finite iterations. Using the inequality $\log(1 + x) \approx x$, the number of iterations is simplified as $(3D - \sigma(Q))/(1/\gamma)$. While (29) also provided finite-iteration convergence, their number of iterations is a fourth order polynomial of ours. Theorem 2 also provides performance guarantees for the stabilizing policy in (24). Specifically, the spectral radius of the closed-loop system is strictly smaller than one, and depends significantly on the decay factor $0 < \xi < 1$.

By Theorem 2, the total number of system rollouts is

$$k(N + N_e) \approx 3D - \sigma(Q)$$

$$\log \left(\frac{1}{\gamma}\right) \left(8d^4 \log \left(\frac{2}{\delta}\right) + 1 + \beta^4 \varrho^2 \phi^4 \log^6 n\right)$$

$$\times \frac{CJ_1^*}{\varrho(R)} \left(1 + \beta \varrho \phi^2 \sqrt{n} \log n\right)^2 mn \log \left(\frac{3D^2}{(1 - \xi)(\varrho(Q)J_1^*)}\right)$$

where $C$ is a constant. Clearly, the sample complexity has a slight dependence on the user-specified parameters $\xi$ and $D$. To elucidate the dependence on important system parameters $n, m, \rho(A)$, we let $\gamma^* = 1/(2\rho^2(A))$ and write the sample complexity more compactly as

$$\log(\rho(A)) \cdot \tilde{O}(n^2 m) \cdot \log(||A||, ||B||, ||Q||, ||R||, J^*).$$

(34)

This reveals that the sample complexity of PG methods for stabilizing linear systems only adds a coefficient logarithmic in the spectral radius of the state matrix to that for solving the LQR problem with a prior stabilizing policy, which is $\tilde{O}(n^2 m)$ [11]. That is, the difficulty of stabilizing an unstable linear system grows with the maximal unstable eigenvalue of the system. While the sample complexity of model-based system stabilization has a dimension dependence linear in $n$ [9], it is comparable to our sample complexity quadratic in $n$. This is because each observation in the model-based approach is a state vector with dimension $n$, while herein a sample of the cost is a scalar.

Finally, we remark some extensions of our explicit discount method. First, we can leverage the explicit update rule to provide finite-time convergence guarantees and sample complexity for other RL-based discount methods, e.g., [27] and [28]. Then, our PG-based discount method is straightforward to extend to linear systems with stochastic noises by leveraging the equivalence between LQR formulations with noises and random initial conditions established in [9, Lemma 10]. Finally, it can be used to stabilize nonlinear systems around the equilibrium point, which we shall confirm in the simulation.

V. SIMULATION

In this section, we first validate the theoretical results of our PG-based discount method in linear systems. Then, we perform it over a classical nonlinear system to show the effectiveness of the proposed method. We observe that our method results in a
larger Region of Attraction (ROA) of initial states than that of using the LQR of the system linearized at the equilibrium point.

### A. Linear Systems

We first perform simulations over a 2-D example for illustration. Consider the following unstable dynamical system with single control input:

\[
A = \begin{bmatrix} 4 & 3 \\ 3 & 1.5 \end{bmatrix}, \quad B = \begin{bmatrix} 2 \\ 2 \end{bmatrix}, \quad Q = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad R = 2.
\]

Clearly, \((A, B)\) is controllable. Let the initial state distribution \(D\) be the standard normal distribution. The selections of the parameters in Algorithm 2 are as follows. The number of rollouts in function evaluation is set to \(N = 20\), and the horizon is set to \(\tau = 100\). We select an initial policy \(K^0 = 0\), the coefficient \(\xi = 0.9\), and discount factor \(\gamma^0 = 10^{-3} < 1/\rho^2(A) = 1/36\). We apply one-step gradient descent each iteration in Algorithm 2, i.e., the policy is updated only once for each \(k\). The stepsize of gradient descent is set to a constant \(\eta = 10^{-3}\), and the smooth radius and the number of rollouts for gradient estimation in Algorithm 1 are set to \(r = 2 \times 10^{-3}\) and \(N_e = 20\), respectively. Fig. 1 illustrates the policy trajectory. In less than 100 iterations, Algorithm 2 returns a stabilizing policy.

Next, we show the convergence of the discount factor in Fig. 2, where we report the results of 20 independent trials. From \(\gamma^0 = 10^{-3}\), the adaptive discount factor (yellow solid line) grows almost exponentially to 1 within 100 iterations. The sample complexity can be calculated by \(100 \times (N + N_e)\), i.e., total number of \(4 \times 10^3\) rollouts with horizon \(\tau = 100\). For comparison, we also plot the maximal discount factor under \(K^\tau\) each iteration, given as \(\gamma_{opt} = 1/\rho^2(A - BK^\tau)\), denoted by the “optimal” discount factor (blue dashed line). We observe that the update rate \(\alpha^\tau\) in (21) well approximates the upper bound in the first 50 iterations, and the gap increases reasonably due to the approximation of condition (14). Moreover, the variance induced by independent trials is competitively small, considering our low sample complexity. We also display a model-based implementation of Algorithm 2, where we assume that \((A, B)\) is known and compute the update rate by (16) in Theorem 1. It is shown that using the exact \(J_s(K)\) and the gradient, the required number of iterations can be reduced to less than 50.

To validate the sample complexity result in Theorem 2, we study the total number of system rollouts required by Algorithm 2 as a function of the state dimension \(n\) and the input dimension \(m\). We first validate the dependence on \(n\) in (34), where we fix the system parameters \(m = 8, Q\) and \(R\) as identity matrices, and randomly sample \(A = 2(A + A^\top)/\|A + A^\top\|\) with each element of \(A\) subject to the standard normal distribution and \(B = B/\|B\|\) with \(B\) also being Gaussian such that \(\|A\| = 2, \|B\| = 1\) and \((A, B)\) is stabilizable with probability one. For algorithm parameters, we set the initial discount factor to \(0.9/\rho^2(A) = 0.225\), \(N_e = 20 \times n\) as suggested by Lemma 3, and other parameters the same as before. For each \(n\), we repeat 20 independent trials and the mean of the sample complexity is recorded. The result is plotted in Fig. 3. It can be observed that the sample complexity is approximately quadratic in the state dimension \(n\). Then, we validate the dependence on \(m\) in (34), where \(n = 2\) is fixed and other parameters are set as before. The result is plotted in Fig. 4. As indicated by (34), the sample complexity is approximately linear in the input dimension \(m\).
B. Nonlinear Systems

In this section, we extend our explicit discount method over the classical cart-pole system with the following continuous dynamics [29]:

\[
\begin{bmatrix}
    m_p + m_c & -m_p l \cos(\theta) \\
    -m_p l \cos(\theta) & m_pl^2
\end{bmatrix}
\begin{bmatrix}
    \ddot{x} \\
    \ddot{\theta}
\end{bmatrix}
= \begin{bmatrix}
    u - m_p l \sin(\theta) \dot{\theta}^2 \\
    m_p gl \sin(\theta)
\end{bmatrix}
\]

where \(z\) and \(\dot{z}\) denote the position and velocity of the cart, \(\theta\) and \(\dot{\theta}\) denote the angular position and velocity of the pole, and the control input \(u\) is the horizontal force to the cart. The state is defined as \(x = (z, \theta, \dot{z}, \dot{\theta})\). The goal is to regulate the system state to the upright equilibrium with \(\theta = 0\). The system parameters include the mass of the pole \(m_p\) and the cart \(m_c\), the length of the pendulum \(l\), and the gravity acceleration \(g\). In the simulation, we set them to \(m_p = m_c = l = g = 1\).

For implementation, we use the forward Euler rule to discretize the continuous dynamics with sampling time \(dt = 0.02\) s. We limit the policy to be linear feedback \(u = -Kx\) and implement Algorithm 2 to search a stabilizing gain over a neighbourhood around the equilibrium point. We set the penalty matrices to \(Q = 2 \times I_{4 \times 4}\) and \(R = 1\). The initial discount factor is \(\gamma = 0.01\). Other parameters are set to \(\tau = 10^3\), \(N_c = N = 20\), \(r = 0.01\), \(n = 10^{-3}\), \(\xi = 1\). In the simulator, we obtain samples of the following cost:

\[
J_{\text{test}}(K) = \mathbb{E}_{x_0} \sum_{t=0}^{\infty} (x_t^\top Q x_t + u_t^\top R u_t)
\]

subject to \(x_{t+1} = \sqrt{\gamma} f(x_t, u_t), x_0 \sim D_{\text{ini}}\) and \(u_t = -K x_t\)

where \(x_{t+1} = f(x_t, u_t)\) denotes the discrete-time cart-pole system and the distribution \(D_{\text{ini}}\) of the initial state \(x_0\) is uniform over the cube \([-r_{\text{ini}}, r_{\text{ini}}]^4\). Since the covariance of \(D_{\text{ini}}\) is not unitary, the update rate at the \(k\)th iteration is accordingly modified as

\[
\alpha_k = \frac{g(Q + (K^{k+1})^\top RK^{k+1})}{3J_{\text{test}}(K^{k+1})/r_{\text{ini}} - g(Q + (K^{k+1})^\top RK^{k+1})}
\]

where \(J_{\text{test}}(K^{k+1})\) is estimated from the simulator.

To evaluate the performance of the resulted policy, we introduce the notion of ROA which describes the region of the initial states that can asymptotically converge to the equilibrium point. Overall, it reflects the ability of a controller to locally stabilize the system. In this simulation, we use the largest ball of the initial states that asymptotically converge to the origin to estimate the ROA based on Monte Carlo sampling. To this end, we randomly select an initial state from the sphere \(r_{\text{test}} \times I^4\) and simulate the closed-loop nonlinear dynamics for \(2 \times 10^3\) steps. We repeat it for \(10^3\) times to see whether \(\|x_{2 \times 10^3}\| \leq r_{\text{test}}/5\) (the state converges) in all cases. If the answer is positive, then we conclude that \(r_{\text{test}} > r_{\text{ini}}\). We then increase \(r_{\text{test}}\) and repeat the test until for a radius \(r_{\text{test}}\) the state norm fails to converge to within \(r_{\text{test}}/5\). Hence, we can estimate the ROA as the ball with radius \(r_{\text{test}} = r_{\text{test}}\).

We perform our algorithm with initial state radius \(r_{\text{ini}} = 0.1, 0.3, 0.5\), each with three independent trials, and estimate the average ROA radius of the resulting controllers. We also compare the results with the optimal LQR controller of the linearized dynamical model at the origin. Our result is summarized in Table I. Interestingly, though the LQR controller has the well known stability margin, our explicit discount method algorithm returns a stabilizing policy with a larger ROA radius when \(r_{\text{ini}} = 0.3, 0.5\). We also observe that the ROA radius becomes slightly larger as \(r_{\text{ini}}\) increases. This is due to that a large \(r_{\text{ini}}\) can facilitate the exploration of the unknown nonlinear system, which helps learn a better controller.

| Region of Attraction Radius \(r_{\text{ROA}}\) | \(r_{\text{ini}}\) | \(0.1\) | \(0.3\) | \(0.5\) |
|----------------|------------|------------|------------|------------|
| LQR            | \(r_{\text{ini}}\) | \(0.6\) | \(0.7\) | \(0.79\) | \(0.66\) |
| \(r_{\text{ROA}}\) | \(0.61\) | \(0.70\) | \(0.79\) | \(0.66\) |

VI. Conclusion

In this article, we have proposed a PG method to learn a stabilizing policy for linear systems by designing an update rule for the discount factor. Moreover, we have shown that the algorithm converges in finite iterations and proved explicit sample complexity in terms of the total number of system rollouts. Simulations have validated the theoretical results and shown the effectiveness of the proposed method on nonlinear systems. As the first to provide a formal sample complexity guarantee for the stabilization problem, we hope our work helps pave the way for understanding the performance of PG methods for general systems.

We now discuss some future directions. The prior work [29] has designed a binary search procedure of the noisy cost to update the discount factor, which is applicable to smooth nonlinear dynamics. Since our update rule for the discount factor is derived based on the structure of linear systems, it would be valuable to examine whether it can also be extended to nonlinear systems from a theoretical perspective. It would also be interesting to further exploit the connections between the notion in [32] and ours of the stability margin, as noted by Remark 2. The cost reduction step in Algorithm 2 can be replaced with other RL methods (e.g., Q-learning and actor-critic) to study their performance and sample complexities for stabilization. Another direction is to study the PG methods for finite-horizon LQR problems with a single feedback gain \(K\). Similar to the discount methods, the idea is to gradually increase the horizon to find a stabilizing gain. The major challenge is that the optimal policy for finite-horizon LQR may not be unique and the optimization landscape is unclear yet. Despite solving the stabilization problem, we believe it is valuable to further investigate the role of the discount factor in the PG methods of LQR problems. Existing literature in the RL field has shown that the discount factor works as a regularizer for generalization [49], and can even accelerate the convergence [50]. However, a systematic understanding of such phenomena is still lacking in the LQR setting, which is left as our future work.

APPENDIX A

PROOF OF LEMMA 3

Due to the equivalence between the discounted LQR in (2) and the standard LQR with damped systems in (4), we can leverage the results [10, Theorem 1] for standard LQR to our problem.

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
Note that the distribution of their initial state is assumed to have a bounded sub-Gaussian norm. This can be satisfied by the boundedness of the distribution \( D \), i.e., there exists \( \phi > 0 \) as a function of \( d \) such that \( \|x_0\|_{\Phi_d} \leq \phi \).

Since we require that \( J_{\gamma,k}(K_k) < D \), the optimality gap is no more than \( \epsilon = D - J_{\gamma,k} \). Then, noting that our PG update starts from \( K^k \) over the sublevel set \( \{ K | J_{\gamma}(K) \leq J_{\gamma}(K^k) \} \) and applying \cite[Theorem 1]{10} complete the proof.

**APPENDIX B**

**PROOF OF LEMMA 5**

We first present a technical lemma, which is a discrete-time counterpart of \cite[Lemma 11]{11}.

**Lemma 9:** Consider the system \( (3) \). Let \( K \in S_\gamma \) and \( P \) satisfy (5). Then, for \( t \in \mathbb{N} \), it holds that

\[
\|x_t\|^2 \leq \left(1 - \frac{\sigma(Q)}{\|P\|}\right)^t \text{Tr}(P) \|x_0\|^2.
\]

**Proof:** The function \( V(x) = x^\top Px \) is a Lyapunov function of the system \( x_{t+1} = \sqrt{\gamma}(A - BK)x_t \) since

\[
V(x_{t+1}) - V(x_t) \leq -cV(x_t)
\]

with \( c := \sigma(Q)/\|P\| \). Then, from an initial state \( x_0 \) we have

\[
\sigma(P)\|x_t\|^2 \leq V(x_t) \leq (1 - c)^t V(x_0) \leq (1 - c)^t \|x_0\|^2 \text{Tr}(P)
\]

which completes the proof.

**Proof of Lemma 5:** We first establish an upper bound of \( |V_\gamma^\infty(K, x_0) - V_{\gamma,N}^\infty(K, x_0)| \) induced by finite horizon, which has exponential dependence on \( \tau \). For a policy \( K \in S_\gamma \), it follows that:

\[
V_\gamma^\infty(K, x_0) - V_{\gamma,N}^\infty(K, x_0) = \sum_{i=0}^{\infty} (x_i^\top Q x_i + u_i^\top R u_i) - \left(\sum_{i=0}^{\tau-1} (x_i^\top Q x_i + u_i^\top R u_i) + \sum_{i=\tau}^{\infty} (x_i^\top Q x_i + u_i^\top R u_i)\right) = x_\tau^\top P x_\tau \leq \text{Tr}(P) \|x_\tau\|^2
\]

where \( P \) is the solution of (5) and \( x_\tau = (\sqrt{\gamma}(A - BK))^\tau x_0 \).

By Lemma 9, the above is further upper bounded by

\[
V_\gamma^\infty(K, x_0) - V_{\gamma,N}^\infty(K, x_0) \leq \left(1 - \frac{\sigma(Q)}{\|P\|}\right)^t \left(\text{Tr}(P)\right)^2 \|x_0\|^2 \leq \frac{d^2 J_{\gamma,k}(K)}{\sigma(Q)} \left(1 - \frac{\sigma(Q)}{J_{\gamma,k}(K)}\right)^t.
\]

Then, we use concentration inequalities to bound the total error \( |J_{\gamma,k}(K) - J_{\gamma,N,k}(K)| \).

Let \( x_i \in \{1, 2, \ldots, N\} \) be \( N \) random initial states sampled independently from \( D \). Since the support of \( D \) is bounded by \( \|x_0\| \leq d \), the random variable \( V_\gamma^\infty(K, x_0) \) is bounded by \( 0 \leq V_\gamma^\infty(K, x_0) = \text{Tr}(P x_0 x_0^\top) \leq J_{\gamma,k}(K) d^2 \forall \in \{1, 2, \ldots, N\} \). For a given constant \( \epsilon \), we let the horizon be

\[
\tau = \frac{-\log(2d) - \log(1 - \frac{\sigma(Q)}{\|P\|})}{\log(1 - \frac{\sigma(Q)}{J_{\gamma,k}(K)})} \approx \frac{J_{\gamma,k}(K) d^2}{\sigma(Q)} \log \left(\frac{2d^2}{\sigma(Q) \epsilon} \frac{\|P\|^2}{\|x_0\|^2}\right).
\]

such that \( V_\gamma^\infty(K, x_0) - V_{\gamma,N}^\infty(K, x_0) \leq \epsilon/2 \). This implies that

\[
\frac{1}{N} \sum_{i=0}^{N-1} V_\gamma^\infty(K, x_0^i) - \frac{1}{N} \sum_{i=0}^{N-1} V_{\gamma,N}^\infty(K, x_0^i) \leq \epsilon/2.
\]

Hence, the Hoeffding’s inequality for the random variable \( V_\gamma^\infty(K, x_0) \) yields that

\[
\Pr(|\hat{J}_{\gamma,N} - J_{\gamma}| \leq \epsilon) = \Pr\left(\frac{1}{N} \sum_{i=0}^{N-1} V_\gamma^\infty(K, x_0^i) - \mathbb{E}_{x_0} [V_{\gamma,N}^\infty(K, x_0)] \leq \epsilon\right) \geq \Pr\left(\frac{1}{N} \sum_{i=0}^{N-1} V_\gamma^\infty(K, x_0^i) - \mathbb{E}_{x_0} [V_{\gamma,N}^\infty(K, x_0)] \leq \frac{\epsilon}{2}\right) \geq 1 - 2 \exp\left(-\frac{N\epsilon^2}{2J_{\gamma,k}(K)^2 d^4}\right).
\]

where the first inequality follows from (35) and the triangle inequality. Let \( \delta = 2 \exp(-N\epsilon^2/(2J_{\gamma,k}(K)^2 d^4)) \). Then, we conclude that at least with probability \( 1 - \delta \), the estimate error is bounded by \( |\hat{J}_{\gamma,N} - J_{\gamma}| \leq \epsilon \). Letting \( \epsilon = J_{\gamma,k}/2 \) completes the proof.

**ACKNOWLEDGMENT**

The authors would like to sincerely thank the Associate Editor and anonymous reviewers for their constructive comments, which significantly improve the quality of this article.

**REFERENCES**

\[1\] J. Tobin, R. Fong, A. Ray, J. Schneider, W. Zaremba, and P. Abbeel, “Domain randomization for transferring deep neural networks from simulation to the real world,” in Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst., 2017, pp. 23–30.

\[2\] O. M. Andrychowicz et al., “Learning dexterous in-hand manipulation,” Int. J. Robot. Res., vol. 39, no. 1, pp. 3–20, 2020.

\[3\] B. Recht, “A tour of reinforcement learning: The view from continuous control,” Ann. Rev. Control, Robot., Auton. Syst., vol. 2, pp. 253–279, 2019.

\[4\] V. Mihelcic et al., “Human-level control through deep reinforcement learning,” Nature, vol. 518, no. 7540, pp. 529–533, 2015.

\[5\] D. Silver et al., “Mastering the game of go with deep neural networks and tree search,” Nature, vol. 529, no. 7587, pp. 484–489, 2016.

\[6\] S. Levine, C. Finn, T. Darrell, and P. Abbeel, “End-to-end training of deep visuomotor policies,” J. Mach. Learn. Res., vol. 17, no. 1, pp. 1334–1373, 2016.

\[7\] M. Fazel, R. Ge, S. Kakade, and M. Mesbahi, “Global convergence of policy gradient methods for the linear quadratic regulator,” in Proc. Int. Conf. Mach. Learn., 2018, pp. 1467–1476.

\[8\] J. Bu, A. Mesbahi, M. Fazel, and M. Mesbahi, “LQR through the lens of first order methods: Discrete-time case,” 2019, arXiv:1907.08921.

\[9\] 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 Proc. 22nd Int. Conf. Intell. Statist., 2019, pp. 2916–2925.

\[10\] H. Mohammadi, M. Soltanolkotabi, and M. R. Jovanović, “On the linear convergence of random search for discrete-time LQR,” IEEE Control Syst. Lett., vol. 5, no. 3, pp. 989–994, Jul. 2021.

\[11\] 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 Trans. Automat. Control, vol. 67, no. 5, pp. 2435–2450, May 2022.

\[12\] K. Zhang, B. Hu, and T. Basar, “Policy optimization for \( H_\infty \) linear control with \( H_\infty \) robustness guarantee: Implicit regularization and global convergence,” SIAM J. Control Optim., vol. 59, no. 6, pp. 4081–4109, 2021.
[13] F. Zhao, K. You, and T. Başar, “Global convergence of policy gradient primal-dual methods for risk-constrained LQRs,” IEEE Trans. Autom. Control, vol. 65, no. 5, pp. 2934–2949, May 2023.

[15] K. Zhang, Z. Yang, and T. Başar, “Policy optimization provably converges to Nash equilibrium in zero-sum linear quadratic games,” in Proc. Annu. Conf. Neural Inf. Process. Syst., 2019, pp. 11598–11608.

[18] Y. Tang, Y. Zheng, and N. Li, “Analysis of the optimization landscape of linear quadratic Gaussian (LQG) control,” Math. Program., vol. 202, no. 1, pp. 399–444, 2023.

[21] J. Duan, W. Cao, Y. Zheng, and L. Zhao, “On the optimization landscape of dynamic con-trollability feedback linear quadratic control,” IEEE Trans. Autom. Control, vol. 66, no. 2, pp. 920–935, Feb. 2024.

[24] F. Zhao, F. Dörfler, A. Chiuso, and K. You, “Data-enabled policy optimization approach,” IEEE Trans. Autom. Control, vol. 67, no. 12, pp. 6429–6444, Dec. 2022.

[26] S. Kang and K. You, “Minimum input design for direct data-driven property identification of unknown linear systems,” Automatica, vol. 156, 2023, Art. no. 111130.

[28] The author expresses his gratitude to the editors and reviewers for their valuable comments and suggestions. The paper is a revised version of a paper presented at the 59th IEEE Conference on Decision and Control, Miami, FL, Dec. 2020.

[30] Dr. You was the recipient of the Guan Zhaozhi award at the 29th Chinese Control Conference 2020. He also received the IEEE Control Systems Society Young Author Award in 2021.

Feiran Zhao received the B.S. degree in control science and engineering from the Harbin Institute of Technology, Harbin, China, in 2018, and the Ph.D. degree in control science and engineering from Tsinghua University, Beijing, China, in 2024.

He held a Visiting position with ETH Zürich, Zurich, Switzerland. His research interests include data-driven control, adaptive control, reinforcement learning, and their applications.

Xingyue Fu received the B.S. degree in control science and engineering, in 2020, from the Department of Automation, Tsinghua University, Beijing, China, where he is currently working toward the Ph.D. degree in control science and engineering.

His research interests include reinforcement learning and data-driven control.

Keyou You (Senior Member, IEEE) received the B.S. degree in statistical science from Sun Yat-sen University, Guangzhou, China, in 2007, and the Ph.D. degree in electrical and electronic engineering from Nanyang Technological University (NTU), Singapore, in 2012.

After briefly working as a Research Fellow with NTU, he joined Tsinghua University, Beijing, China, where he is currently a Full Professor with the Department of Automation. He held Visiting positions with Politecnico di Torino, Turin, Italy, Hong Kong University of Science and Technology, Hong Kong, University of Melbourne, Melbourne, Australia, etc. His research interests include intersections between control, optimization, and learning as well as their applications in autonomous systems.

Dr. You was the recipient of the Guan Zhaozhi award at the 29th Chinese Control Conference in 2010, the Asian Control Association (ACA) Young Researcher Education Award in 2019, the National Natural Science Funds for Excellent Young Scholars in 2017, and for Distinguished Young Scholars in 2023. He is currently an Associate Editor for Automatica, IEEE Transactions on Control of Network Systems, and IEEE Transactions on Cybernetics.

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.