Computing Stabilizing Feedback Gains via a Model-Free Policy Gradient Method

Ibrahim K. Ozaslan, *Graduate Student Member, IEEE*,
Hesameddin Mohammadi, *Graduate Student Member, IEEE*,
and Mihailo R. Jovanović, *Fellow, IEEE*

***Abstract*—In spite of the lack of convexity, convergence and sample complexity properties were recently established for the random search method applied to the linear quadratic regulator (LQR) problem. Since policy gradient approaches require an initial stabilizing controller, we propose a model-free algorithm that searches over the set of state-feedback gains and returns a stabilizing controller in a finite number of iterations. Our algorithm involves a sequence of relaxed LQR problems for which the associated domains converge to the set of stabilizing controllers for the original continuous-time linear time-invariant system. Starting from a stabilizing controller for the relaxed problem, the proposed approach alternates between updating the controller via policy gradient iterations and decreasing relaxation parameter in the LQR cost while preserving stability at all iterations. By properly tuning the relaxation parameter updates we ensure that the cost values do not exceed a uniform threshold and establish computable bounds on the total number of iterations.**

***Index Terms*—Data-driven control, linear quadratic regulator, model-free control, nonconvex optimization, random search method, reinforcement learning, sample complexity.**

I. Introduction

MODEL-FREE approaches to Reinforcement Learning (RL) have attracted significant attention in recent years because of their simple implementation. Regardless of the intricacies of the application domain, these approaches involve two simple steps: estimating values of a cost function without identifying a model and utilizing a gradient descent to obtain the policy that optimizes the cost function. Moreover, these approaches are shown to perform remarkably well in practice, e.g., in learning complex locomotion tasks via neural network dynamics [1] and in playing Atari games using deep-RL [2].

In model-free RL, a considerable effort has been devoted to policy gradient methods because of their simplicity and convenience for large-scale problems. For continuous control tasks such as the infinite-horizon Linear Quadratic Regulator (LQR), convergence and sample complexity of policy gradient methods have been established in [3]–[7]. Extensions to robustness analysis through implicit regularization [8], Markovian jump linear systems [9], distributed LQR [10], [11], and the output-feedback problem [12] have also been made.

A common requirement in many existing works is that the policy gradient method needs to be initialized with a stabilizing controller. However, in a model-free setting obtaining such a controller can be equally challenging even for linear time-invariant (LTI) systems. In the absence of a stabilizing controller, simulating the system can result in unbounded signals which hinders the application of data-driven techniques. Furthermore, the set of stabilizing feedback gains is not convex [13] and convergence of local search methods to feasible solutions is not well understood.

In this letter, we consider the problem of finding a stabilizing state-feedback controller for a *continuous-time* LTI system with unknown state-space parameters. Solving this problem for LTI systems is an important first step towards addressing similar challenge in more complicated settings including Markovian jump linear systems [8], static output-feedback design [14], and structured feedback synthesis [15], [16]. We propose a model-free algorithm based on a policy gradient method that can search over state-feedback gains without requiring the initial controller to be stabilizing. By introducing exponentially decaying weights to the state and control signals, we determine gradients with respect to a relaxation of the LQR objective function. The domain of the relaxed LQR problem contains the set of stabilizing feedback gains and we use it to guide our search for a stabilizing controller. Starting from a controller in the domain of the relaxed LQR problem, our algorithm alternates between updating the controller via policy gradient iterations and decreasing the relaxation parameter.

Our approach is inspired by the recent efforts to identify a stabilizing feedback gain matrix for *discrete-time* LQR problem using discounted LQR cost [17]–[19]. We examine the continuous-time problem and demonstrate how exponentially decaying weights in the LQR cost [20] can be used to perform policy gradient updates without requiring initialization with a stabilizing feedback gain. By properly tuning the

Manuscript received 21 March 2022; revised 30 May 2022; accepted 21 June 2022. Date of publication 4 July 2022; date of current version 19 July 2022. This work was supported in part by the National Science Foundation (NSF) under Award ECCS-1708906 and Award ECCS-1809833. Recommended by Senior Editor J. Daafouz. *(Corresponding author: Ibrahim K. Ozaslan.)*

The authors are with the Ming Hsieh Department of Electrical and Computer Engineering, University of Southern California Los Angeles, Los Angeles, CA 90089 USA (e-mail: ozaslan@usc.edu; hesamedm@usc.edu; mihailo@usc.edu).

Digital Object Identifier 10.1109/LCSYS.2022.3188180

updates of the relaxation parameter we ensure that the cost values do not exceed a threshold given by a constant factor of the optimal objective value for the original non-relaxed LQR problem. This allows us to establish finite-time convergence guarantees and bound the required relative accuracy of the policy gradient method by a constant factor, thereby keeping the total number of required policy gradient updates finite.

The rest of this letter is structured as follows. We formulate the problem in Section II and introduce the model-free algorithm for finding a stabilizing state-feedback gain matrix in Section III. We establish finite-time convergence guarantees in Section IV and provide an example to demonstrate the merits and the effectiveness of our approach in Section V. We offer concluding remarks in Section VI and provide proofs of technical results in appendices.

Notation: We use $\mathrm{Re}\,(\cdot)$ to denote the real part, $\|\cdot\|$ to denote the Euclidean norm, and $\|\cdot\|_W$ to denote the weighted norm for a positive definite matrix W, i.e., $\|x\|_W := \|W^{1/2}x\|$. For matrices, $\|\cdot\|$ is the spectral norm, $\|\cdot\|_F$ is the Frobenius norm, $\underline{\sigma}(\cdot)$ is the smallest singular value, and $\mathrm{tr}\,(\cdot)$ is the matrix trace.

II. Problem Formulation

We study continuous-time LTI systems

$$\dot{x} \;=\; A\,x \,+\, B\,u \tag{1}$$

where $x(t) \in \mathbb{R}^n$ is the state, $u(t) \in \mathbb{R}^m$ is the control input, and A and B are constant matrices. For a stabilizable pair (A, B), the closed-loop system associated with the control law $u(t) = -Kx(t)$ is stable if and only if the matrix K belongs to the set of stabilizing feedback gains,

$$\mathcal{S} \;:=\; \{K \,\in\, \mathbb{R}^{m\times n} \,|\, A \,-\, B\,K \text{ is Hurwitz}\}.$$

We are interested in finding a feedback gain $K \in \mathcal{S}$ when the model parameters A and B are unknown. Our approach utilizes the family of relaxations $\mathcal{S}_\alpha \supseteq \mathcal{S}$, parameterized by a scalar $\alpha \geq 0$ with $A_\alpha := A - \alpha I$, where

$$\mathcal{S}_\alpha \;:=\; \{K \in \mathbb{R}^{m\times n} \,|\, A_\alpha \,-\, B\,K \text{ is Hurwitz}\}.$$

It is straightforward to verify that for any $K \in \mathbb{R}^{m\times n}$ there exists an α such that $K \in \mathcal{S}_\alpha$; hence, solving the relaxed problem is trivial. In addition, the set $\mathcal{S}_\alpha$ characterizes the domain of the discounted LQR problem [20],

$$\underset{K}{\text{minimize}}\; J_\alpha(K) \;:=\; \begin{cases} \mathrm{tr}(P\,\Omega) & K \,\in\, S_\alpha \\ \infty & \text{otherwise} \end{cases} \tag{2a}$$

where P is the solution of the Lyapunov equation

$$(A_\alpha - BK)^T P + P(A_\alpha - BK) + Q + K^T R\,K = 0 \tag{2b}$$

and Q, R, and Ω are positive-definite matrices. For system (1) with $u(t) = -Kx(t)$ and a random initial condition $x(0) = x_0$ with the covariance matrix Ω, the objective function in (2) can also be written as,

$$J_\alpha(K) \;=\; \mathbb{E}\bigg[\int_0^\infty \mathrm{e}^{-2\alpha t}\big(x^T(t)Qx(t) + u^T(t)Ru(t)\big)\mathrm{d}t\bigg].$$

Assumption 1: The distribution $\mathcal{D}$ of the initial condition $x(0) = x_0$ has zero mean, covariance $\mathbb{E}\big[x_0 x_0^T\big] = \Omega \succ 0$, and bounded support, i.e., $\|x_0\| \leq \theta$ for $x_0 \sim \mathcal{D}$.

We next summarize basic properties of the set $\mathcal{S}_\alpha$ and its relation to the objective function $J_\alpha(K)$. These are utilized in the rest of this letter and the proof can be found in [13], [20].

Lemma 1: The set $\mathcal{S}_\alpha$ satisfies the following properties:

(i) For a stabilizable pair (A, B) and $\alpha \geq 0$, the set $\mathcal{S}_\alpha$ is open, unbounded, and increasing in α, i.e.,

$$\mathcal{S}_\alpha \,\subseteq\, \mathcal{S}_\beta \;\iff\; \alpha \,\leq\, \beta.$$

(ii) For any scalars a and α, the sublevel set

$$\mathcal{S}_\alpha(a) \;:=\; \big\{K \,\in\, \mathbb{R}^{m\times n} \,|\, J_\alpha(K) \,\leq\, a\big\}$$

is compact and increasing in both a and α, i.e.,

$$\begin{aligned} \mathcal{S}_\alpha(a) \subseteq \mathcal{S}_\alpha(b) &\iff a \leq b \\ \mathcal{S}_\alpha(a) \subseteq \mathcal{S}_\beta(a) &\iff \alpha \leq \beta. \end{aligned} \tag{3}$$

(iii) For any scalar α and a feedback gain matrix K, $K \in \mathcal{S}_\alpha$ if and only if $\alpha > \max_i \mathrm{Re}\,(\lambda_i)$, where λ_i is the ith eigenvalue of the matrix $A - BK$.

III. Algorithm

We now describe our approach to finding a stabilizing feedback gain $K \in \mathcal{S}_0$ for system (1) with unknown model parameters (A, B). We start by choosing an initial feedback gain $K \in \mathbb{R}^{m\times n}$ and a sufficiently large relaxation parameter α such that $K \in \mathcal{S}_\alpha$. Then, the algorithm alternates between decreasing the relaxation parameter α and updating K via policy gradient updates until we achieve $\alpha \leq 0$ while preserving the condition $K \in \mathcal{S}_\alpha$ in all iterations. At any iteration with K and α such that $K \in \mathcal{S}_\alpha$, as we decrease α the set $\mathcal{S}_\alpha$ shrinks and it may no longer contain K. Lemma 2 ensures that this situation does not arise provided that the decrease in α is sufficiently small.

Lemma 2: For any positive scalars α, a, and $c > 1$, there exists a scalar $\tilde{\alpha} > 0$ such that

$$\mathcal{S}_\beta(a) \;\subseteq\; \mathcal{S}_{\beta - \tilde{\beta}}(c\,a) \tag{4}$$

for all $\tilde{\beta} \leq \tilde{\alpha}$ and $\beta \leq \alpha$.

Proof: See Section IV-B. ∎

For any $c > 1$ and a cost upper bound a with $K \in \mathcal{S}_\alpha(a)$, Lemma 2 implies that the relaxation parameter can be decreased to $\alpha^+ = \alpha - \tilde{\alpha}$ by increasing the cost upper bound by a factor of c, i.e., $K \in \mathcal{S}_{\alpha^+}(ca)$. To avoid arbitrary small decrements $\tilde{\alpha}$ (which may prevent convergence of α to 0) our approach is to bring down the cost upper bound back to a. This is achieved using a sequence of policy gradient updates of the form [7]

$$K^{k+1} \;=\; K^k \,-\, \eta\,\overline{\nabla} J_{\alpha^+}(K^k) \tag{5}$$

initialized with $K^0 = K$ and applied to discounted LQR problem (2) associated with α^+ to update the feedback gain

$$K^+ \in \mathcal{S}_{\alpha^+}(a).$$

In (5), k is the iteration index, η is a constant step size and, $\overline{\nabla} J_{\alpha^+}(K)$ is an empirical approximation of the gradient $\nabla J_{\alpha^+}(K)$ computed via simulations of system (1) for randomly perturbed K.

Algorithm 1 Finding Stabilizing Controller

Input: Relaxation parameter α_0, feedback gain K_0, simulation time τ, number of samples N, the objective threshold b, state and control weight matrices Q and R, distribution $\mathcal{D}$ of the initial condition.

1: **Initialization:** Set $\alpha \leftarrow \alpha_0$ and $K \leftarrow K_0$.
2: **while** $\alpha > 0$ **do**
3: Compute $\bar{J}_\alpha(K)$ using (6).
4: Set $\tilde{\alpha} \leftarrow \dfrac{\underline{\sigma}^2(Q)\,\underline{\sigma}(\Omega)}{8\bar{J}_\alpha(K)\|Q + K^T R K\|_F}$.
5: Compute K^+ using policy gradient update (5) initialized at K so that $J_{\alpha^+}(K^+) \leq b$, where $\alpha^+ = \alpha - \tilde{\alpha}$.
6: Set $K \leftarrow K^+$ and $\alpha \leftarrow \alpha^+$.
7: **end while**

Output: Stabilizing controller K for system (1).

To update the relaxation parameter α, we need an estimate of the objective function. In the model-free setting, this estimate can be obtained via simulation of system (1),

$$\bar{J}_\alpha(K) \;=\; \frac{1}{N}\sum_{i=1}^{N} J^{\tau}_{\alpha,x_0^i}(K) \tag{6}$$

where the random vectors x_0^i represent N i.i.d. samples from the distribution $\mathcal{D}$ and the cost function

$$J^{\tau}_{\alpha,v}(K) \;:=\; \int_0^{\tau} \mathrm{e}^{-2\alpha t}\big(x^T(t)Qx(t) + u^T(t)Ru(t)\big)\mathrm{d}t \tag{7}$$

corresponds to system (1) with the initial condition v and the feedback law $u(t) = -Kx(t)$ simulated up to time τ. As we demonstrate in Section IV, the cost estimate in (6) can achieve any desired accuracy for sufficiently large sample size N and simulation time τ.

We outline the proposed strategy in Algorithm 1. Theoretical convergence guarantees along with guidelines for selecting the initialization parameters are provided in Section IV.

IV. Theoretical Guarantees

We next provide theoretical guarantees for Algorithm 1.

A. Main Result

We next establish conditions under which Algorithm 1 achieves finite-time convergence. Our main result is summarized in Theorem 1, where the optimal objective value of problem (2) is given by $J^\star_\alpha$ and, for $\alpha = 0$, we let $J^\star := J^\star_0$.

Theorem 1: Let the cost threshold in Algorithm 1 be given by $b = 2J^\star$ and let the initial relaxation parameter α_0, the initial feedback gain K_0, the simulation time τ, and the number of samples N satisfy

$$b \geq J_{\alpha_0}(K_0), \quad N \;\geq\; \frac{8\,\theta^4}{\underline{\sigma}^2(\Omega)}\log(1/\zeta)$$
$$\tau \geq \frac{2\,b}{\underline{\sigma}(\Omega)\,\underline{\sigma}(Q)}\log\left(\frac{4\,\theta^2}{\underline{\sigma}(\Omega)}\right) \tag{8}$$

for some positive scalar ζ, where θ is an upper bound on the norm of the initial conditions x_0^i in (6). Then, with probability at least $1 - \zeta\alpha_0/\tilde{\alpha}_{\min}$, Algorithm 1 terminates after at most $\alpha_0/\tilde{\alpha}_{\min}$ iterations, where $\tilde{\alpha}_{\min}$ is a lower bound on the decrement $\tilde{\alpha}$,

$$\tilde{\alpha}_{\min} \;:=\; \frac{\underline{\sigma}^2(Q)\,\underline{\sigma}(\Omega)\,\underline{\sigma}(R)\nu_{\alpha_0}}{16\,J^\star(\|Q\|_F\,\underline{\sigma}(R)\,\nu_{\alpha_0} \;+\; 16\,\|R\|(J^\star)^2)}$$

and the constant ν_{α_0} given by (14) only depends on the problem parameters. Moreover, the required relative accuracy

$$\epsilon \;:=\; (b \;-\; J^\star_{\alpha^+})/(J_{\alpha^+}(K) \;-\; J^\star_{\alpha^+})$$

for the policy gradient updates in Step 5 satisfies $\epsilon \geq 1/4$.

Proof: See Section IV-C. ∎

Theorem 1 establishes that the decrement $\tilde{\alpha}$ used in updating the relaxation parameter α is lower bounded by a constant and that we can find a stabilizing feedback gain for system (1) after a finite number of iterations. We note that the assumption $b = 2J^\star$ on the cost threshold in Theorem 1 is only made to simplify our presentation and that it can be relaxed to $b = cJ^\star$ for any $c > 1$.

Remark 1: Theorem 1 requires the initial relaxation parameter α_0 and the feedback gain K_0 to satisfy $b \geq J_{\alpha_0}(K_0)$. Below are two simple techniques to find such α_0 and K_0.

1) By Lemma 1, for any scalar α_0 and feedback gain K, we have $K \in \mathcal{S}_{\alpha_0}$ if $\alpha_0 > \max_i \mathrm{Re}\,(\lambda_i)$, where λ_i is the ith eigenvalue of $A - BK$. Under this condition, we can run the policy gradient method initialized with K applied to the α_0-relaxed LQR problem to obtain a feedback gain K_0 that satisfies $b \geq J_{\alpha_0}(K_0)$.
2) From $\lim_{\alpha\to\infty} J_\alpha(K_0) = 0$, it follows that for any K_0 we can select α_0 large enough to satisfy $b \geq J_{\alpha_0}(K_0)$. We note that because of potentially larger α_0 compared to the former approach, this technique may require more iterations (i.e., more α-updates) in Algorithm 1.

Remark 2: The complexity of model-free policy gradient updates in Step 5 depends on the gradient estimation scheme. In the one-point gradient estimation setting, the policy gradient method achieves relative accuracy $\epsilon \leq c/k$ after k updates [21], where c is a constant. Thus, the inequality $\epsilon > 1/4$ established by Theorem 1 ensures that Algorithm 1 requires at most $4c$ updates in Step 5. This requirement further reduces to $\log(4)/\log(1/\rho)$ updates in the two-point gradient estimation setting because of linear convergence with rate ρ, i.e., $\epsilon \leq \rho^k$ [7].

We next provide lemmas that we use to prove Theorem 1.

B. Technical Lemmas

We first address the α-update step of Algorithm 1.

Lemma 3: For any $K \in \mathcal{S}_\alpha$ and $c \geq 1$, if

$$\delta \;\leq\; \frac{\underline{\sigma}^2(Q + K^T R K)\,\underline{\sigma}(\Omega)}{2\,J_\alpha(K)\,\|Q + K^T R K\|_F}\,\frac{c \;-\; 1}{c} \tag{9a}$$

then

$$J_{\alpha-\delta}(K) \;\leq\; c\,J_\alpha(K). \tag{9b}$$

Proof: See the Appendix A. ∎

As we show in Section IV-C, for any scalar a the Frobenius norm of the feedback gain matrix $\|K\|_F$ is bounded over the sublevel set $\mathcal{S}_\alpha(a)$. This fact combined with Lemma 3 can be used to prove Lemma 2.

In Lemma 4, we show that the error in the τ-truncated cost estimate $J^\tau_{\alpha,x_0}(K)$ is an exponentially decaying function of τ. This allows us to establish a bound on the required simulation time τ in Theorem 1.

Lemma 4: For any $\tau \geq 0$, we have

$$J^\infty_{\alpha,x_0}(K) \;-\; J^\tau_{\alpha,x_0}(K) \;\leq\; c_1 \mathrm{e}^{-\,c_2\,\tau} \tag{10}$$

where $J^\tau_{\alpha,x_0}(K)$ is defined in (7) and

$$c_1 \;=\; \|x_0\|^2 J_\alpha(K)/\underline{\sigma}(\Omega),\;\; c_2 \;=\; \underline{\sigma}(\Omega)\,\underline{\sigma}(Q)/J_\alpha(K).$$

Proof: See the Appendix B. ∎

Using Lemma 4, we can show that, for any scalar γ, the estimation error bound

$$J^\infty_{\alpha,x_0}(K) \;-\; J^\tau_{\alpha,x_0}(K) \;\leq\; \gamma \tag{11a}$$

can be achieved by choosing a simulation time

$$\tau \;\geq\; \log\!\left(\frac{\|x_0\|^2 J_\alpha(K)}{\gamma\,\underline{\sigma}(\Omega)}\right)\frac{J_\alpha(K)}{\underline{\sigma}(\Omega)\,\underline{\sigma}(Q)}. \tag{11b}$$

Lemma 5 shows that the estimated cost remains within a factor of the actual cost with probability not smaller than $1-\zeta$ if the number of samples is proportional to $\log(1/\zeta)$.

Lemma 5: Let Assumption 1 hold and let the simulation time τ be such that

$$J^\infty_{\alpha,x_0}(K) \;-\; J^\tau_{\alpha,x_0}(K) \;\leq\; J_\alpha(K)/4. \tag{12}$$

Then, with probability not smaller than $1-\zeta$, we have

$$J_\alpha(K) \;\leq\; 2\,\bar{J}_\alpha(K) \;\leq\; 3J_\alpha(K)$$

where $\zeta = 2\exp(-N\,\underline{\sigma}^2(\Omega)/(8\theta^4))$.

Proof: See the Appendix C. ∎

Lemma 5 provides sufficient conditions to obtain an upper bound for the actual cost by using the cost estimate in (6).

We are now ready to prove our main result.

C. Proof of Theorem 1

We demonstrate that $J_\alpha(K)$ and $J_{\alpha^+}(K)$ never exceed $2b = 4J^\star$ throughout Algorithm 1. From the initial condition (8) and the cost threshold in Step 5 it follows that $J_\alpha(K) \leq b$ at all iterations. Thus, it remains to verify the condition $J_{\alpha^+}(K) \leq 2b$. Using the lower bound in (8) on the simulation time τ, we can combine (12) with the condition on the sample size N in (8) and apply Lemma 5 to obtain

$$J_\alpha(K) \;\leq\; 2\bar{J}_\alpha(K) \;\leq\; 3J_\alpha(K) \tag{13}$$

with probability not smaller than $1-\zeta$. In addition, for the decrement $\tilde{\alpha}$ in Step 4 of the algorithm, the condition in Lemma 3 holds for $c = 2$. Thus, we can combine the first inequality in (13) with Lemma 3 to obtain $J_{\alpha^+}(K) \leq 2\,J_\alpha(K) \leq 4J^\star$. To derive the uniform lower bound $\tilde{\alpha}_{\min}$ on $\tilde{\alpha}$, we can write

$$\tilde{\alpha} = \frac{\underline{\sigma}^2(Q)\,\underline{\sigma}(\Omega)}{8\bar{J}_\alpha(K)\|Q+K^TRK\|_F}$$

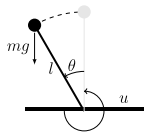

Fig. 1. An inverted pendulum of length l with a lumped mass m controlled with an input torque u.

$$\geq \frac{\underline{\sigma}^2(Q)\,\underline{\sigma}(\Omega)}{24\,b\,\|Q+K^TRK\|_F}$$
$$\geq \frac{\underline{\sigma}^2(Q)\,\underline{\sigma}(\Omega)}{24\,b\big(\|Q\|_F+\|R\|\|K\|_F^2\big)}$$
$$\geq \frac{\underline{\sigma}^2(Q)\,\underline{\sigma}(\Omega)\,\underline{\sigma}(R)\nu_{\alpha_0}}{24\,b\big(\|Q\|_F\;\underline{\sigma}(R)\;\nu_{\alpha_0}+4\|R\|b^2\big)}$$

where the first inequality follows from (13), and the second inequality is obtained by applying the triangle inequality. To show the last inequality, we utilize [7, Lemma 16], which shows that $\|K\|_F \leq a/\sqrt{\nu_\alpha\,\underline{\sigma}(R)}$ for any $K \in \mathcal{S}_\alpha(a)$, where

$$\nu_\alpha \;:=\; \frac{\underline{\sigma}^2(\Omega)}{4}\left(\frac{\|A\|+\alpha}{\sqrt{\underline{\sigma}(Q)}}+\frac{\|B\|}{\sqrt{\underline{\sigma}(R)}}\right)^{-2}. \tag{14}$$

This upper bound with $a = J_\alpha(K) \leq 4J^\star$ combined with the fact that ν_α is decreasing in α yields the last inequality. Accounting for failure in each iteration yields the success probability of the algorithm of at least $1-\zeta\alpha_0/\tilde{\alpha}_{\min}$.

Finally, to prove the lower bound on the required relative accuracy ϵ, we can write

$$\epsilon \;=\; (b-J^\star_\alpha)/(J_\alpha(K)-J^\star_\alpha) \;\geq\; J^\star/(4J^\star) \;=\; 1/4.$$

Here, we have used the facts that $0 < J^\star_\alpha < J^\star$ and $J_\alpha(K) < 4J^\star$ for all $\alpha > 0$.

V. Computational Experiments

To demonstrate the utility of our approach, we use a torque-controlled inverted pendulum with length $l = g$ and mass $m = 1/l^2$ where g is gravitational acceleration; see Fig. 1. Linearization around the upright equilibrium point yields system (1) with

$$A \;=\; \begin{bmatrix}0 & 1\\ 1 & 0\end{bmatrix},\quad B \;=\; \begin{bmatrix}0\\ 1\end{bmatrix}$$

where the state vector $x = \begin{bmatrix}\theta & \dot{\theta}\end{bmatrix}^T$ contains angle and angular velocity of the pendulum and the feedback gain matrix K is determined by $K := \begin{bmatrix}k_1 & k_2\end{bmatrix} \in \mathbb{R}^{1\times 2}$.

We let the distribution $\mathcal{D}$ be the standard normal and choose $Q = \mathrm{diag}\,(10, 1)$ and $R = 1$ in the LQR cost. This yields the optimal cost value $J^\star = 13.398$. Furthermore, we set the number of samples N to 10, the simulation time τ to 100, and initialize Algorithm 1 using the first technique proposed in Remark 1. In particular, we set $K_0 = 0$ and, to choose the initial relaxation parameter α_0 without using the sufficient condition $\alpha_0 \geq \max_i \mathrm{Re}\,(\lambda_i(A)) = 1$ (which requires the knowledge of model parameters), we start from $\alpha_0 = 0$ and gradually increase it until the cost estimate achieves the threshold $b \geq \bar{J}_{\alpha_0}(K_0)$, where we set $b = 30$. This approach

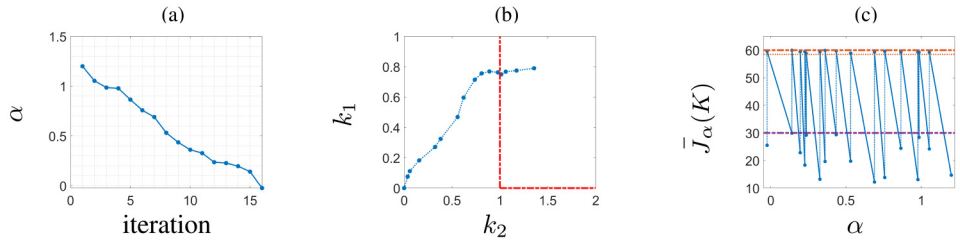

Fig. 2. (a) The dependence of the relaxation parameter α on the iteration count of Algorithm 1; (b) the feedback gains k_1 and k_2 obtained at each iteration (blue) starting with $k_1 = k_2 = 0$; the boundary of the set of stabilizing feedback gains $\{K | k_1 > 0, k_2 > 1\}$ for the actual system (red) is also shown; and (c) the dependence of the cost estimate $J_\alpha(K)$ on α along with the cost thresholds $b = 30$ (purple), $1.9b$, and $2b$ (brown). Starting from $\alpha_0 = 1.2$, the algorithm alternates between updating α (solid blue) and updating k_1 and k_2 via policy gradient (dotted blue) until $\alpha \leq 0$ is obtained.

yields $\alpha_0 = 1.2$. For the policy gradient update in Step 5, we use the two-point gradient estimation scheme proposed in [7] with step size 10^{-4} and a smoothing constant 10^{-5}.

The analytical decrement $\tilde{\alpha}$ given in Step 4 of Algorithm 1 is chosen small enough to guarantee that $J_{\alpha^+}(K)$ does not exceed the threshold $2b$. However, this decrement can be very conservative in practice. To address this issue, for the α-update step we use bisection to find the updated α^+ such that $1.9b \leq \bar{J}_{\alpha^+}(K) \leq 2b$.

Figure 2(a) shows the dependence of the relaxation parameter α on the iteration count of Algorithm 1 and Fig. 2(b) illustrates the corresponding feedback gains k_1 and k_2 obtained at each iteration along with the boundary of the set of stabilizing feedback gains $\{K \,|\, k_1 > 0,\ k_2 > 1\}$ for the actual system. We observe that starting from the initial values $K_0 = 0$ and $\alpha_0 = 1.2$, the algorithm achieves a stabilizing feedback gain and $\alpha \leq 0$ in only 16 steps.

Finally, Fig. 2(c) shows the dependence of the cost estimate $\bar{J}_\alpha(K)$ on the relaxation parameter α. Guided by the two thresholds b and $2b$, the algorithm alternates between decreasing the relaxation parameter and updating the feedback gain K via policy gradient until $\alpha \leq 0$ is achieved.

VI. Concluding Remarks

We have proposed a model-free algorithm based on a policy gradient method that does not require an initial stabilizing feedback gain for a continuous-time LTI system. By utilizing exponentially decaying weights on the state and control signals, our algorithm introduces a sequence of relaxations to the standard LQR objective function. Using policy gradient updates with respect to the relaxed LQR problems, the algorithm reduces the relaxation parameter by a constant factor at each iteration and converges in finite time. Our convergence guarantees are obtained by ensuring that the cost values do not exceed a threshold given by a constant factor of the optimal objective value for the original non-relaxed LQR problem. Future directions include extensions to model-free stabilization of systems with small nonlinear components and systems with partially-available model parameters.

Appendix A
Proof of Lemma 3

We first show that for any $K \in \mathcal{S}_\alpha$ and a scalar δ such that (9a) holds, $K \in \mathcal{S}_{\alpha'}$, where $\alpha' := \alpha - \delta$. Let P be the unique positive definite solution to Lyapunov equation (2b). For any scalar δ such that $2\delta P \prec Q_K := Q + K^T R K$, we can use the Lyapunov function candidate $V(x) = x^T P x$ to verify that the system $\dot{x} = (A_{\alpha'} - BK)x$ is stable, i.e., $K \in \mathcal{S}_{\alpha'}$. The condition $2\delta P \prec Q_K$ holds if we let

$$\delta \ \leq \ 0.5\, \underline{\sigma}(Q_K)\, \underline{\sigma}(\Omega)/J_\alpha(K). \tag{15}$$

This can be verified by noting that

$$J_\alpha(K) \ = \ \mathrm{tr}(P\,\Omega) \ \geq \ \mathrm{tr}(P)\, \underline{\sigma}(\Omega) \ \geq \ \|P\|\, \underline{\sigma}(\Omega).$$

The upper bound in (9a) ensures that (15) holds. Thus, we obtain $K \in \mathcal{S}_{\alpha'}$ and next show that (9b) holds.

Let $\widetilde{X} := X' - X$, where X and X' are the unique positive definite solutions to

$$(A_\alpha \ - \ B\,K)X \ + \ X(A_\alpha \ - \ B\,K)^T \ = -\Omega \tag{16a}$$
$$(A_{\alpha'} \ - \ B\,K)X' \ + \ X'(A_{\alpha'} \ - \ B\,K)^T \ = -\Omega. \tag{16b}$$

Subtracting (16a) from (16b) and rearranging terms yields

$$(A_\alpha - BK)\widetilde{X} + \widetilde{X}(A_\alpha - BK)^T + 2\delta X' \ = \ 0.$$

Equivalently, we can write $\widetilde{X} = \mathcal{F}(2\delta X')$, where

$$\mathcal{F}(W) \ := \ \int_0^\infty \mathrm{e}^{(A_\alpha - B\,K)t}\, W\, \mathrm{e}^{(A_\alpha - B\,K)^T t}\, \mathrm{d}t$$

is the inverse Lyapunov operator acting on a symmetric matrix W [7]. Hence, the cost difference satisfies

$$\begin{aligned} J_{\alpha'}(K) \ - \ J_\alpha(K) \ &= \ \mathrm{tr}(\widetilde{X} Q_K) \ = \ \mathrm{tr}(\mathcal{F}(2\delta X')Q_K) \\ &\leq 2\delta \|\mathcal{F}(X')\|_F \|Q_K\|_F \\ &\leq 2\delta \|\mathcal{F}\| \|X'\|_F \|Q_K\|_F \\ &\leq \ 2\delta \frac{\mathrm{tr}(X)}{\underline{\sigma}(\Omega)} \|X'\|_F \|Q_K\|_F \ \leq \ \beta J_{\alpha'}(K) \end{aligned}$$

where $\beta := 2\delta \|Q_K\|_F \mathrm{tr}(X)/(\underline{\sigma}(Q_K)\,\underline{\sigma}(\Omega))$. Here, the first inequality follows from the Cauchy-Schwartz inequality and the linearity of the inverse Lyapunov operator, the second inequality follows from definition of the operator norm $\|\mathcal{F}\| := \sup_M \|\mathcal{F}(M)\|_F/\|M\|_F$, the third inequality follows from [7, Lemma 24], and the last inequality follows from the upper bound $\|X'\|_F \leq J_{\alpha'}(K)/\underline{\sigma}(Q_K)$. Rearranging the terms and setting $(1-\beta)^{-1} \leq c$ yields that $J_{\alpha'}(K) \leq c J_\alpha(K)$ if

$$\delta \ \leq \ (1 - 1/c)\ \underline{\sigma}(Q_K)\ \underline{\sigma}(\Omega)/(2\mathrm{tr}(X)\|Q_K\|_F).$$

The proof is completed by replacing $\mathrm{tr}(X)$ with its upper bound $J_\alpha(K)/\underline{\sigma}(Q_K)$.

APPENDIX B
PROOF OF LEMMA 4

The cost $J^{\tau}_{\alpha,x_0}(K)$ in (7) can also be written as

$$J^{\tau}_{\alpha,x_0}(K) = \int_0^{\tau} x^T(t)\big(Q + K^T R K\big)x(t)\mathrm{d}t$$

where $x(t) = \exp((A_\alpha - BK)t)x_0$ is the state of system $\dot{x} = (A_\alpha - BK)x$ with the initial condition $x(0) = x_0$. Using this expression, it is easy to verify that

$$\begin{aligned} J^{\infty}_{\alpha,x_0}(K) - J^{\tau}_{\alpha,x_0}(K) &= J^{\infty}_{\alpha,x(\tau)}(K) \\ &= (x(\tau))^T P x(\tau) \le \|x_0\|^2 \, \|P^{1/2}\mathrm{e}^{G\tau}\|^2 \end{aligned} \tag{17}$$

where P is the unique positive definite solution to the Lyapunov equation (2b) and $G := A_\alpha - BK$.

Using the notion of logarithmic norm [22, Sec. II.8], we have [23, Th. 2.3],

$$\|\mathrm{e}^{G\tau}\|_P^2 \;:=\; \max_{z\neq 0} \|\mathrm{e}^{G\tau}z\|_P^2 \big/ \|z\|_P^2 \;\le\; \mathrm{e}^{\mu\tau} \tag{18}$$

where $\mu := -\|Q_K^{-1/2} P Q_K^{-1/2}\|^{-1} \le -\|P\|^{-1}\,\underline{\sigma}(Q)$ and $Q_K := Q + K^T R K$. Now, by combining (18) and the above upper bound on μ, we obtain

$$\begin{aligned} \|P^{1/2}\mathrm{e}^{G\tau}\|^2 &= \|\mathrm{e}^{\tilde{G}\tau}P^{1/2}\|^2 \\ &\le \|P\|\,\|\mathrm{e}^{\tilde{G}\tau}\|^2 = \|P\|\,\|\mathrm{e}^{G\tau}\|_P^2 \\ &\le \|P\|\,\mathrm{e}^{-(\underline{\sigma}(Q)/\|P\|)\tau} \\ &\le (J_\alpha(K)/\,\underline{\sigma}(\Omega))\,\mathrm{e}^{-(\underline{\sigma}(\Omega)\,\underline{\sigma}(Q)/J_\alpha(K))\tau} \end{aligned} \tag{19}$$

where $\tilde{G} := P^{1/2} G P^{-1/2}$. The last inequality comes from $\|P\| \le J_\alpha(K)/\,\underline{\sigma}(\Omega)$ and combining the inequalities in (17) and (19) completes the proof.

APPENDIX C
PROOF OF LEMMA 5

For any scalar γ with $J^{\infty}_{\alpha,x_0^i}(K) - J^{\tau}_{\alpha,x_0^i}(K) \le \gamma/2$, for $i = 1,\ldots,N$, we have

$$\left|\frac{1}{N}\sum_{i=1}^{N} J^{\infty}_{\alpha,x_0^i}(K) \;-\; \bar{J}_\alpha(K)\right| \le \frac{\gamma}{2} \tag{20}$$

where $\bar{J}_\alpha(K)$ is given by (6). From $J^{\infty}_{\alpha,x_0^i}(K) = (x_0^i)^T P x_0^i$ and the bound $\|x_0^i\| \le \theta$, it follows that $J^{\infty}_{\alpha,x_0^i}(K) \le \theta^2 \mathrm{tr}\,(P)$, where P is the solution to (2b). We can now write

$$\begin{aligned} &\mathbb{P}\big(\big|\bar{J}_\alpha(K) \;-\; J_\alpha(K)\big| \;\le\; \gamma\big) \\ &= \mathbb{P}\left(\left|\frac{1}{N}\sum_{i=1}^{N} J^{\tau}_{\alpha,x_0^i}(K) \;-\; \mathbb{E}_{x_0}\big[J^{\infty}_{\alpha,x_0}(K)\big]\right| \;\le\; \gamma\right) \\ &\ge \mathbb{P}\left(\left|\frac{1}{N}\sum_{i=1}^{N} J^{\infty}_{\alpha,x_0^i}(K) \;-\; \mathbb{E}_{x_0}\big[J^{\infty}_{\alpha,x_0}(K)\big]\right| \;\le\; \frac{\gamma}{2}\right) \\ &\ge 1 \;-\; 2\exp\Big(-N\gamma^2/\Big(2\mathrm{tr}(P)^2\theta^4\Big)\Big) \end{aligned}$$

where the first inequality follows from combining (20) and the triangle inequality, and the last line follows from Hoeffding's inequality [24, Sec. 2.2]. Replacing $\mathrm{tr}(P)$ with its upper bound $J_\alpha(K)/\,\underline{\sigma}(\Omega)$ and setting $\gamma = J_\alpha(K)/2$ completes the proof.

REFERENCES

[1] A. Nagabandi, G. Kahn, R. S. Fearing, and S. Levine, "Neural network dynamics for model-based deep reinforcement learning with model-free fine-tuning," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2018, pp. 7559–7566.

[2] V. Mnih *et al.*, "Playing Atari with deep reinforcement learning," 2013, *arXiv:1312.5602*.

[3] 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.

[4] S. Tu and B. Recht, "The gap between model-based and model-free methods on the linear quadratic regulator: An asymptotic viewpoint," in *Proc. Conf. Learn. Theory*, 2019, pp. 3036–3083.

[5] 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 *Proc. IEEE Conf. Decis. Control*, 2019, pp. 7474–7479.

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

[7] 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. Autom. Control*, vol. 67, no. 5, pp. 2435–2450, May 2022.

[8] K. Zhang, B. Hu, and T. Basar, "Policy optimization for $\mathcal{H}_2$ linear control with $\mathcal{H}_\infty$ robustness guarantee: Implicit regularization and global convergence," *SIAM J. Control Optim.*, vol. 59, no. 6, pp. 4081–4109, Jul. 2021.

[9] J. P. Jansch-Porto, B. Hu, and G. E. Dullerud, "Convergence guarantees of policy optimization methods for Markovian jump linear systems," in *Proc. Amer. Control Conf.*, 2020, pp. 2882–2887.

[10] L. Furieri, Y. Zheng, and M. Kamgarpour, "Learning the globally optimal distributed LQ regulator," in *Proc. Learn. Dyn. Control*, 2020, pp. 287–297.

[11] T.-J. Chang and S. Shahrampour, "Distributed online linear quadratic control for linear time-invariant systems," in *Proc. Amer. Control Conf.*, 2021, pp. 923–928.

[12] I. Fatkhullin and B. Polyak, "Optimizing static linear feedback: Gradient method," *SIAM J. Control Optim.*, vol. 59, no. 5, pp. 3887–3911, 2021.

[13] H. T. Toivonen, "A globally convergent algorithm for the optimal constant output feedback problem," *Int. J. Control*, vol. 41, no. 6, pp. 1589–1599, Jun. 1985.

[14] T. Rautert and E. W. Sachs, "Computational design of optimal output feedback controllers," *SIAM J. Optim.*, vol. 7, no. 3, pp. 837–852, Aug. 1997.

[15] F. Lin, M. Fardad, and M. R. Jovanović, "Design of optimal sparse feedback gains via the alternating direction method of multipliers," *IEEE Trans. Autom. Control*, vol. 58, no. 9, pp. 2426–2431, Sep. 2013.

[16] M. R. Jovanović and N. K. Dhingra, "Controller architectures: Tradeoffs between performance and structure," *Eur. J. Control*, vol. 30, pp. 76–91, Jul. 2016.

[17] J. C. Perdomo, J. Umenberger, and M. Simchowitz, "Stabilizing dynamical systems via policy gradient methods," in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 29274–29286.

[18] A. Lamperski, "Computing stabilizing linear controllers via policy iteration," in *Proc. IEEE Conf. Decis. Control*, 2020, pp. 1902–1907.

[19] F. Zhao, X. Fu, and K. You, "Learning stabilizing controllers of linear systems via discount policy gradient," 2021, *arXiv:2112.09294*.

[20] H. Feng and J. Lavaei, "Escaping locally optimal decentralized control polices via damping," in *Proc. Amer. Control Conf.*, 2020, pp. 50–57.

[21] D. Malik, A. Panajady, K. Bhatia, K. Khamaru, P. L. Bartlett, and M. J. Wainwright, "Derivative-free methods for policy optimization: Guarantees for linear quadratic systems," *J. Mach. Learn. Res.*, vol. 21, no. 21, pp. 1–51, Feb. 2020.

[22] C. A. Desoer and M. Vidyasagar, *Feedback Systems: Input-Output Properties*. Philadelphia, PA, USA: SIAM, 2009.

[23] G.-D. Hu and G.-D. Hu, "A relation between the weighted logarithmic norm of a matrix and the Lyapunov equation," *BIT Numer. Math.*, vol. 40, no. 3, pp. 606–610, Sep. 2000.

[24] R. Vershynin, *High-Dimensional Probability: An Introduction With Applications in Data Science*. Cambridge, U.K.: Cambridge Univ. Press, 2018.