POLICY GRADIENT CONVERGENCE IN DISCOUNTED LQR

ANDREA AGAZZI AND CRAIG CHEN

1. INTRODUCTION

The aim of this paper is to provide a convergence guarantee for policy gradient algorithms in the setting of the Linear-Quadratic Regulator (LQR). Specifically, we show that *any* random initialization of a linear (In the state space) policy will converge to the global optimum of the undiscounted LQR in finite time. In related existing works, it is assumed that the initial policy is stabilizing; however, our result shows that — at the cost of more training iterations — it is possible to do away with this assumption. Additionally, we provide an example to show that a policy that is linear in the *parameters* does not converge in this setting.

Notation. We use $\|\cdot\|$ as the operator norm of a matrix or the euclidean norm of a vector and $\varrho(\cdot)$ to denote the spectral radius. We also use $\lambda_{max}(\cdot)$ and $\sigma_{min}(\cdot)$ to refer to the largest eigenvalue and smallest singular value of a matrix, respectively.

1.1. Preliminaries and Background.

1.1.1. *Markov Decision Processes.*

We denote a Markov Decision Process (MDP) by the tuple $(X, U, \mathcal{P}, \mathcal{R}, \gamma)$ where X denotes the state-space, U the action-space, $\mathcal{P}$ the state transition function, $\mathcal{R}$ the immediate (real-valued) cost function, and $\gamma \in [0, 1]$ is a discount factor. Note that $\mathcal{P}$ and $\mathcal{R}$ may depend on the action $u \in U$ in addition to the state $x \in X$. MDPs provide a general setting that is useful for studying optimization problems. The Linear-Quadratic Regulator is such an MDP with $X, U = \mathbb{R}^n, \mathbb{R}^m$ and $\mathcal{P}, \mathcal{R}, \gamma$ as described below.

In our analysis, we do not assume that $\mathcal{R}$ is bounded on $X \times U$, instead, we guarantee that the cumulative, infinite-horizon cost $\sum_{t\geq 0} \gamma^t \mathcal{R}(x_t, u_t)$ is finite through discounting.

1.1.2. *The Linear-Quadratic Regulator.*

The Linear-Quadratic Regulator (LQR) is a classic optimal control problem. In general, optimal control problems are concerned with finding the control to given dynamics that minimizes a given cost function. In this paper, we are concerned with the special case where the dynamics are *linear*, time-invariant with no disturbance or added noise and the cost function is *quadratic* in the state and the control action. We consider the discounted infinite time-horizon problem,

$$\begin{aligned} \text{minimize} \quad & \mathbb{E}_{x_0\sim\mathcal{D}}\left[\sum_{t=0}^{\infty} \gamma^t \left(x_t^\intercal Q x_t + u_t^\intercal R u_t\right)\right] \\ \text{with} \quad & x_{t+1} = Ax_t + Bu_t \end{aligned}$$

where A, B are the system (or transition) matrices, Q, R are positive definite cost matrices, x_0 is randomly distributed with distribution $\mathcal{D}$, and $\gamma \in (0, 1]$ is the discount factor. In the entirety of this work, the pair (A, B) is assumed to be controllable.

Definition 1. Stability and Controllability

(1) We call a square matrix M stable (or stabilizing) if $\varrho(M) < 1$.

(2) A linear dynamical system with $x_t \in \mathbb{R}^n, u_t \in \mathbb{R}^m$ is called controllable if all states can be reached in n-steps (*i.e.*, if the system is controllable, we can always find an input sequence $(u_0, ..., u_{t-1})$ that moves the state from x_0 to x_{goal} in time = n steps).

Optimal control theory shows that the optimal control for the LQR problem is a linear function of the state.

$$u_t = K^* x_t$$

If A, B, Q, R, γ are known, we have an explicit form for the optimal control. Let P_γ denote the unique positive definite solution to the discounted discrete-time+ algebraic Riccati equation (DARE)

$$P_\gamma = \gamma A^\intercal P_\gamma A - \gamma^2 A^\intercal P_\gamma B \left(R + \gamma B^\intercal P_\gamma B\right)^{-1} B^\intercal P_\gamma A + Q$$

Later in this work, we omit the subscript γ and assume the dependencies to be clear from context. We can then write the optimal control as

$$K_\gamma^* = -\gamma \left(R + \gamma B^\intercal P_\gamma B\right)^{-1} B^\intercal P_\gamma A$$

It is important to note that, in general, the optimal policy for the discounted LQR is not always the same as what is optimal for the undiscounted LQR. For instance, the optimal policy w.r.t to the discounted LQR may not even be stabilizing [8]. In Section 3, we address this issue by prescribing an iterative method that converges to the undiscounted global optimum by iteratively updating the discount factor γ.

Throughout the remainder of the paper, we will use γ to denote the discount rate which we choose given an initial policy K_0. Specifically, we choose γ such that the infinite-horizon cost of K_0 will not diverge in the worst case. That is, we choose $\gamma < \min(\frac{1}{\varrho(A+BK_0)^2}, 1)$ such that

$$C(K) = \sum_{t=0}^{\infty} \gamma^t x_t^\intercal (Q + K^\intercal R K) x_t < \infty$$

where $\{x_t\}$ is the sequence generated by the dynamics $x_{t+1} = (A + BK)x_t$.

1.1.3. *Policy Gradient Methods.*

The advantage of directly learning a policy rather than first learning a value function is that policy-approximating methods can easily learn stochastic policies, whereas action-value based methods have no natural *and* flexible way to find a stochastic policy. For example, in situations with imperfect information, the best decision can often be to do different actions with specific probabilities (think Poker).

Although it is possible to generate stochastic policies with an action-value function by selecting actions from a softmax over the action-values of a state, this approach lacks the flexibility of its policy-based counterpart. Without modifying a temperature parameter, the softmax method can not approach a deterministic policy as the true action-values will always differ by a finite amount, which translates to specific probabilities that aren't 0 or 1. One could use a scheduled decay of the temperature parameter; however, in practice, this is often difficult to properly implement. On the other hand, with the policy-based approach, if the optimal policy is deterministic – parameterization permitting – the action preferences of the optimal actions will be pushed infinitely higher than those of the sub-optimal actions, resulting in convergence towards a deterministic policy.

Lastly, there is also an important theoretical result that gives policy-parameterization methods an advantage over action-value based methods. The policy gradient theorem [12] provides an explicit formula for the gradient of the performance in terms of the gradient of the policy w.r.t its

parameters – importantly, the gradient of the state distribution is *not* needed. Thus, in addition to their flexibility, the policy gradient theorem shows that these approaches are also practical.

Intuitively, the main idea behind policy gradient methods is to increase the probabilities of actions that lead to lower costs (higher returns) and decrease the probabilities of actions that lead to higher costs (lower rewards), until eventually arriving at the optimal policy. There are many different variations of policy gradients. Here, we review **REINFORCE** [15], a classic sampled-based Monte-Carlo policy gradient algorithm that immediately follows from the policy gradient theorem.

Let $\pi_\vartheta(u|x)$ be a policy parameterized by ϑ, where $u \sim \pi_\vartheta(\cdot|x)$. The policy gradient theorem [12] states that

$$\begin{aligned}\nabla C(\vartheta) &\propto \sum_x \mu(x) \sum_u Q_\pi(x,u) \nabla_\vartheta \pi_\vartheta(u|x) \\ &= \mathbb{E}_\pi \left[\sum_u Q_\pi(x_t,u) \nabla_\vartheta \pi_\vartheta(u|x_t) \right]\end{aligned}$$

where μ is the on-policy distribution of states following π and $Q_\pi(\cdot,\cdot)$ denotes the state-action values; the equality comes from the fact that the proportionality constant can be absorbed into the step-size. We can further simplify the above expression by handling the sum over actions in the same way that we replaced the sum over states. The expectation is with respect to the trajectory $\{x_t, u_t\}$ experienced under π.

$$\begin{aligned}\nabla C(\vartheta) &= \mathbb{E}_\pi \left[\sum_u \frac{\pi(u|x_t)}{\pi(u|x_t)} Q_\pi(x_t,u) \nabla_\vartheta \pi(u|x_t) \right] \\ &= \mathbb{E}_\pi \left[Q_\pi(x_t,u_t) \frac{\nabla_\vartheta \pi(u_t|x_t)}{\pi(u_t|x_t)} \right] \\ &= \mathbb{E}_\pi \left[G_t \frac{\nabla_\vartheta \pi(u_t|x_t)}{\pi(u_t|x_t)} \right]\end{aligned}$$

where $G_t = \sum_{k=0} \gamma^{t+k} r(x_{t+k}, u_{t+k})$ is the sample return (sum of discounted costs(rewards)) following time t. Thus, we have an expression that can be used in our stochastic gradient descent (ascent) algorithm.

$$\vartheta_{t+1} = \vartheta_t - \alpha G_t \frac{\nabla_\vartheta \pi(u_t|x_t)}{\pi(u_t|x_t)} = \vartheta_t - \alpha G_t \nabla_\vartheta \log \pi(u_t|x_t)$$

More recent advances in Reinforcement Learning theory have introduced the idea of learning with an entropy-regularized objective. The main motivation behind entropy-regularization is to encourage exploration by promoting stochastic policies [1]. Experiments have shown that maximum entropy – or so called, "soft" – approaches work rather well, making them an interesting object for theoretical study [4, 6]. Encouraging exploration in a natural way (as opposed to an ε-greedy approach) allows for an agent to choose actions that may yield larger long-run rewards without enforcing certain assumptions about the system that may or may not be verifiable [11].

In this framework, we regularize the reward (or cost) of every action as follows

$$r(x,u) \to r(x,u) - \tau \log \pi_\vartheta(u|x)$$

where τ is the temperature parameter. Then, the value function and action-value function read:

$$V^\pi(x_0) = \mathbb{E}_{x_0, u \sim \pi_\vartheta(\cdot|x)} \left[\sum_t \gamma^t \left(r(x_t,u_t) - \tau \log \pi_\vartheta(u_t|x_t) \right) \right]$$

$$
\begin{aligned}
Q^{\pi}(x_0, u_0) &= r(x_0, u_0) + \gamma \mathbb{E}_{x_0, u_0}\left[V^{\pi}(x_1)\right] \\
&= r(x_0, u_0) + \mathbb{E}_{x_0, u \sim \pi_{\vartheta}(\cdot|x)}\left[\sum_t \gamma^t \left(r(x_t, u_t) - \tau \log \pi_{\vartheta}(u_t|x_t)\right)\right]
\end{aligned}
$$

so that we have

$$
V^{\pi}(x_0) = \mathbb{E}_{u \sim \pi}\left[Q^{\pi}(x_0, u) - \tau \log \pi_{\vartheta}(u|x_0)\right]
$$

In later sections, we refer to these as the "soft" value or Q-functions and denote them $\tilde{V}$ and $\tilde{Q}$, respectively.

1.2. **Related Works.**

In light of the advantages provided by policy-based approaches to reinforcement learning, there has been a recent wave of research seeking to provide convergence guarantees for algorithms that seem to work in practice. Work in [5] has shown that in finite action spaces, our intuition is correct and algorithms like TRPO [9] and PPO [10] do indeed converge to the globally optimal policy. Concurrent work in [7] has demonstrated similar convergence results for softmax policy gradient methods with an entropy-regularized objective in the tabular setting. Interestingly, their work also demonstrated that the entropy-regularized objective yields a much faster convergence rate. Another perspective on the policy gradient methods with finite action spaces in [14] proves convergence for the policy class of over-parameterized two-layer neural networks by using a shared network architecture to ensure the compatibility condition for the actor and critic is met. However, in the aforementioned publications, the assumption of a finite action space is essential in obtaining the results; when considering continuous action spaces, we use the LQR as a proxy for more general environments since the LQR setting is well understood and allows for more straightforward analysis.

In the continuous action-space setting, [3] shows that policy gradient methods can converge in the LQR; however, they leverage existing knowledge about LQR optimality and consider only the class of linear policies. Similarly, [16] provides a convergence guarantee for Actor-Critic methods in the LQR setting by considering the class of linear-Gaussian policies with fixed variance. The work herein also primarily works with linear policies, but takes a slightly different approach in showing convergence. Where most works consider the evolution of the policy with respect to time/iterations, we also consider the evolution of the optimal policy with respect to the discount factor.

1.3. **Contributions.**

Our work continues along a line of work using the LQR as a proxy for more general Reinforcement Learning environments [2, 13, 3]. We show that the common assumption of a stable initial policy can be relaxed by using a homotopy-based approach to iteratively update the policy until reaching a stabilizing policy. We address the issue of an unstable initial policy by introducing a discount factor to force all costs to be finite. We then show that systematically updating the discount factor allows us to reach a stabilizing policy, where then we can run the policy gradient algorithm to converge to the global optimum.

Lastly, it's important to highlight that our work uses a model-based approach. To generalize our results to the model-free perspective, one could proceed in a similar fashion to [3] by showing that when the roll-out is sufficiently long, one can accurately approximate the cost function and covariance and that with enough samples, one can estimate the true gradient within a desired accuracy.

Summarizing, the main contributions of this work are as follows:

- (*Arbitrary Initialization*) With the class of linear (in the state) policies, we show that the standard assumption of a stabilizing initial policy is not necessary to guarantee convergence of policy gradient methods to the globally optimal solution.
- (*Generalizations*) We show that our arbitrary initialization result also applies to the class of linear-Gaussian policies. We entertain a simple non-linear policy in the one-dimensional LQR and show that policy gradient will converge in this simple setting. Additionally, we show that, we can not always expect policy gradient methods to converge even with somewhat well-behaved function approximators. We provide an example policy that is linear in the parameters, but will get stuck at a locally optimal policy.

In the remainder of this paper, we first provide the convergence proof for Policy Gradient methods in the discounted LQR (Section 2). Then, we show that the γ-iteration algorithm converges to the optimal policy of the undiscounted LQR (Section 3). Finally, we provide preliminary calculations for potential directions to generalize our work (Section 4). The discussion is contained in Section 5.

2. Convergence of Policy Gradient Methods in the Discounted LQR

In this section, we show the convergence result for a linearly parameterized policy in the states. This section extends the results in [3] to the discounted regime. The work in [3] proves the result for the undiscounted LQR. We modify their proofs for the undiscounted LQR to show that the result also holds when using a discounted cost function.

2.1. **Adapted Lemmas.**

Before continuing, we highlight the difference between eigenvalues and singular values. It is possible for a matrix to have all eigenvalues within the unit circle of the complex plane, but to have singular values outside of it. This is important because the criteria for stability of $A + BK$ is for its spectral radius to be less than 1. Geometrically, the spectral radius of a matrix is the largest factor we can stretch a vector in its original direction, whereas the "stretching" captured by singular values is not necessarily in the direction of the original vector.

We start with a few definitions for convenience of notation.

Definition 2. Define P_K^γ as the unique solution to:

$$P_K^\gamma = Q + K^\intercal R K + \gamma (A + BK)^\intercal P_K^\gamma (A + BK)$$

As a consequence we have: $V_K(x) = x^\intercal P_K^\gamma x$.

Definition 3. For notational simplicity:

$$\begin{aligned} E_K^\gamma &= (R + \gamma B^\intercal P_K^\gamma B) K + \gamma B^\intercal P_K^\gamma A \\ \Sigma_K^\gamma &= \mathbb{E}_{x_0}\left[\sum_{t=0}^{\infty} \gamma^t x_t x_t^\intercal\right] \\ \mu &= \sigma_{min} \mathbb{E}\left[x_0 x_0^\intercal\right] \end{aligned}$$

Lemma 1. (*Policy Gradient Expression*) *The policy gradient can be written:*

$$\nabla C(K) = 2 E_K^\gamma \Sigma_K^\gamma$$

Proof of Lemma 1. Given x_0, we have:

$$\begin{aligned} C(K)|_{x_0} &= x_0^\intercal P_K^\gamma x_0 \\ &= x_0^\intercal \big(Q + K^\intercal R K\big) x_0 + \gamma x_0^\intercal (A + BK)^\intercal P_K^\gamma (A + BK) x_0 \end{aligned}$$

$$= x_0^\intercal \left(Q + K^\intercal R K\right) x_0 + \gamma C(K)|_{(A+BK)x_0}$$

Now, taking the gradient with respect to K and using recursion – notice there are two dependencies on K in $\nabla C(K)|_{(A+BK)x_0}$,

$$\begin{aligned}
\nabla C(K)|_{x_0} &= 2RKx_0x_0^\intercal + 2\gamma B^\intercal P_K^\gamma (A+BK)x_0x_0^\intercal + \gamma \nabla C(K)|_{x_1} \\
&= \sum_{t=0}^{\infty} \gamma^t 2\left((R+\gamma B^\intercal P_K^\gamma B)K + \gamma B^\intercal P_K^\gamma A\right) x_t x_t^\intercal \\
&= 2\left((R+\gamma B^\intercal P_K^\gamma B)K + \gamma B^\intercal P_K^\gamma A\right) \Sigma_K^\gamma \\
&= 2E_K^\gamma \Sigma_K^\gamma
\end{aligned}$$

□

Lemma 2. (*Cost Difference*) *Let K and K' be two policies. Let $\{x_t'\}$ and $\{u_t'\}$ be the state and action sequences generated by K', and let the initial value for both policies be $x = x_0 = x_0'$. With discount factor $\gamma \in (0,1)$, we have:*

$$V_{K'}(x) - V_K(x) = \sum_{t=0}^{\infty} \gamma^t A_K(x_t', u_t')$$

Furthermore, we can write the advantage at any x as (notice the distinction between the behavior policy and the evaluator policy):

$$A_K(x, K'x) = 2x^\intercal (K'-K)^\intercal E_K^\gamma x + x^\intercal (K'-K)^\intercal (R+\gamma B^\intercal P_K^\gamma B)(K'-K)x$$

Proof of Lemma 2. Let $\{c_t'\}$ be the cost sequence generated by K'. We have that:

$$\begin{aligned}
V_{K'}(x) - V_K(x) &= \sum_{t=0}^{\infty} \gamma^t c_t' - V_K(x) \\
&= \sum_{t=0}^{\infty} \gamma^t \left(c_t' - V_K(x_t') + V_K(x_t')\right) - V_K(x)
\end{aligned}$$

Since $x_0' = x_0 = x$

$$\begin{aligned}
&= \sum_{t=0}^{\infty} \gamma^t (c_t' + \gamma V_K(x_{t+1}') - V_K(x_t')) \\
&= \sum_{t=0}^{\infty} \gamma^t A_K(x_t', u_t')
\end{aligned}$$

For the second claim, observe that:

$$V_K(x) = x^\intercal \big(Q + K^\intercal R K\big)x + \gamma x^\intercal (A+BK)^\intercal P_K^\gamma (A+BK)x$$

Thus, for $u' = K'x$:

$$\begin{aligned}
A_K(x, K'x) &= Q_K(x, K'x) - V_K(x) \\
&= x^\intercal \left(Q + (K')^\intercal R K'\right)x + \gamma V_K\big((A + BK')x\big) - V_K(x) \\
&= x^\intercal \left(Q + (K')^\intercal R K'\right)x + \gamma x^\intercal (A + BK')^\intercal P_K^\gamma (A + BK')x - V_K(x) \\
&= x^\intercal \left(Q + (\mathbf{K}' - \mathbf{K} + \mathbf{K})^\intercal R (\mathbf{K}' - \mathbf{K} + \mathbf{K})\right)x \\
&\quad + \gamma x^\intercal \left(A + B(\mathbf{K}' - \mathbf{K} + \mathbf{K})\right)^\intercal P_K^\gamma \left(A + B(\mathbf{K}' - \mathbf{K} + \mathbf{K})\right)x - V_K(x) \\
&= x^\intercal \left(Q + K^\intercal R K\right)x + \gamma x^\intercal (A + BK)^\intercal P_K^\gamma (A + BK)x - V_K(x) \\
&\quad + x^\intercal (K' - K)^\intercal R (K' - K)x + 2x^\intercal (K' - K) R K x \\
&\quad + \gamma x^\intercal (K' - K)^\intercal B^\intercal P_K^\gamma B (K' - K)x + \gamma 2 x^\intercal (K' - K)^\intercal B^\intercal P_K^\gamma (A + BK)x
\end{aligned}$$

By the observation above:

$$\begin{aligned}
&= x^\intercal (K' - K)^\intercal (R + \gamma B^\intercal P_K^\gamma B)(K' - K)x \\
&\quad + 2x^\intercal (K' - K)^\intercal \left(RK + \gamma B^\intercal P_K^\gamma (A + BK)\right)x \\
&= x^\intercal (K' - K)^\intercal (R + \gamma B^\intercal P_K^\gamma B)(K' - K)x \\
&\quad + 2x^\intercal (K' - K)^\intercal \left((R + \gamma B^\intercal P_K^\gamma B)K + \gamma B^\intercal P_K^\gamma A\right)x
\end{aligned}$$

which completes the second claim. □

Lemma 3. (*Gradient Domination*) *Let K^* denote the optimal policy. It holds that:*

$$C(K) - C(K^*) \leq \frac{\|\Sigma_{\mathbf{K}^*}^\gamma\|}{\sigma_{min}(R)\mu^2} \operatorname{Tr}\left(\nabla C(K)^\intercal \nabla C(K)\right)$$

where σ_{min} is the minimum singular value.
For a lower bound, it holds that:

$$C(K) - C(K^*) \geq \frac{\mu}{\|R + \gamma B^\intercal P_K^\gamma B\|} \operatorname{Tr}\left((E_K^\gamma)^\intercal E_K^\gamma\right)$$

Proof of Lemma 3. First, we provide a bound for the advantage. By Lemma 2:

$$\begin{aligned}
&A_K(x, K'x) \\
&= 2x^\intercal (K' - K)^\intercal E_K^\gamma x + x^\intercal (K' - K)^\intercal (R + \gamma B^\intercal P_K^\gamma B)(K' - K)x \\
&= 2 \operatorname{Tr}\left(xx^\intercal (K' - K)^\intercal E_K^\gamma\right) + \operatorname{Tr}\left(xx^\intercal (K' - K)^\intercal (R + \gamma B^\intercal P_K^\gamma B)(K' - K)\right)
\end{aligned}$$

Now, completing the square

$$\begin{aligned}
&= \operatorname{Tr}\left(xx^\intercal \left((K' - K) + (R + \gamma B^\intercal P_K^\gamma B)^{-1} E_K^\gamma\right)^\intercal (R + \gamma B^\intercal P_K^\gamma B)\left((K' - K) + (R + \gamma B^\intercal P_K^\gamma B)^{-1} E_K^\gamma\right)\right) \\
&\quad - \operatorname{Tr}\left(xx^\intercal (E_K^\gamma)^\intercal (R + \gamma B^\intercal P_K^\gamma B)^{-1} E_K^\gamma\right) \\
&\geq - \operatorname{Tr}\left(xx^\intercal (E_K^\gamma)^\intercal (R + \gamma B^\intercal P_K^\gamma B)^{-1} E_K^\gamma\right)
\end{aligned}$$

where equality holds when $K' = K - (R + \gamma B^\intercal P_K^\gamma B)^{-1} E_K^\gamma$.
For the upper bound. By Lemma 2:

$$\begin{aligned}
C(K) - C(K^*) &= -\mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t A_K(x_t^*, u_t^*)\right] \\
&\leq \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \operatorname{Tr}\left(x_t^* x_t^{*\intercal} (E_K^\gamma)^\intercal (R + \gamma B^\intercal P_K^\gamma B)^{-1} E_K^\gamma\right)\right] \\
&\leq \operatorname{Tr}\left(\Sigma_{K^*}^\gamma (E_K^\gamma)^\intercal (R + \gamma B^\intercal P_K^\gamma B)^{-1} E_K^\gamma\right)
\end{aligned}$$

$$
\begin{aligned}
&\leq \|\Sigma_{K^*}^\gamma\| \operatorname{Tr}\left((R+\gamma B^\intercal P_K^\gamma B)^{-1} E_K^\gamma (E_K^\gamma)^\intercal\right) \\
&\leq \|\Sigma_{K^*}^\gamma\| \|(R+\gamma B^\intercal P_K^\gamma B)^{-1}\| \operatorname{Tr}\left(E_K^\gamma (E_K^\gamma)^\intercal\right) \\
&\leq \frac{\|\Sigma_{K^*}^\gamma\|}{\sigma_{min}(R+\gamma B^\intercal P_K^\gamma B)} \operatorname{Tr}\left(E_K^\gamma (E_K^\gamma)^\intercal\right) \\
&= \frac{\|\Sigma_{K^*}^\gamma\|}{\sigma_{min}(R+\gamma B^\intercal P_K^\gamma B)} \operatorname{Tr}\left(\frac{1}{4}(\Sigma_K^\gamma)^{-1}(\Sigma_K^\gamma)^{-1}\nabla C(K)^\intercal \nabla C(K)\right) \\
&\leq \frac{\|\Sigma_{K^*}^\gamma\|}{\sigma_{min}(R)\sigma_{min}(\Sigma_K^\gamma)^2} \operatorname{Tr}\left(\nabla C(K)^\intercal \nabla C(K)\right)
\end{aligned}
$$

For the lower bound, consider $K' = K - (R+\gamma B^\intercal P_K^\gamma B)^{-1} E_K^\gamma$:

$$
\begin{aligned}
C(K) - C(K^*) &\geq C(K) - C(K') \\
&= -\mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t A_K(x_t', u_t')\right] \\
&= \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \operatorname{Tr}\left(x_t'(x_t')^\intercal (E_K^\gamma)^\intercal (R+\gamma B^\intercal P_K^\gamma B)^{-1} E_K^\gamma\right)\right] \\
&= \operatorname{Tr}\left(\Sigma_{K'}^\gamma (E_K^\gamma)^\intercal (R+\gamma B^\intercal P_K^\gamma B)^{-1} E_K^\gamma\right) \\
&\geq \frac{\mu}{\|R+\gamma B^\intercal P_K^\gamma B\|} \operatorname{Tr}\left((E_K^\gamma)^\intercal E_K^\gamma\right)
\end{aligned}
$$

□

Lemma 4. *This lemma provides useful upper bounds on P_K^γ and Σ_K^γ.*

$$
\|P_K^\gamma\| \leq \frac{C(K)}{\sigma_{min}\left(\mathbb{E}\left[x_0 x_0^\intercal\right]\right)} \qquad \|\Sigma_K^\gamma\| \leq \frac{C(K)}{\sigma_{min}(Q)}
$$

Proof of Lemma 4. For P_K^γ, we have:

$$
\begin{aligned}
C(K) = \mathbb{E}\left[x_0^\intercal P_K^\gamma x_0\right] &= \operatorname{Tr}\left(\mathbb{E}\left[x_0 x_0^\intercal\right] P_K^\gamma\right) \\
&\geq \|P_K^\gamma\| \sigma_{min}\left(\mathbb{E}\left[x_0 x_0^\intercal\right]\right)
\end{aligned}
$$

For the second claim:

$$
\begin{aligned}
C(K) &= \sum_{t=0}^{\infty} \gamma^t x_t^\intercal (Q + K^\intercal R K) x_t \\
&\geq \operatorname{Tr}(\Sigma_K^\gamma)\sigma_{min}(Q) \\
&\geq \|\Sigma_K^\gamma\| \sigma_{min}(Q)
\end{aligned}
$$

□

2.2. **Exact Gradient Descent Convergence.**

Theorem 1. *Gradient Descent Progress*

Let $K' = K - \alpha \nabla C(K)$ where the step-size

$$
\alpha \leq \min\left\{\frac{1}{16}\left(\frac{\sigma_{min}(Q)\mu}{C(K)}\right)^2 \frac{1}{\|B\|\|\nabla C(K)\|(1+\|A+BK\|)}, \frac{3}{8}\frac{\sigma_{min}(Q)}{C(K)\|R+\gamma B^T P_K^\gamma B\|}\right\}.
$$

Then, it holds that:

$$C(K') - C(K^*) \leq \left(1 - \alpha \frac{\sigma_{min}(R)\mu^2}{2\|\Sigma^\gamma_{K^*}\|}\right)(C(K) - C(K^*))$$

Proof of Theorem 1. By Lemma 2 (notice the switch of signs from $(K' - K)$ to $-(K - K')$):

$$\begin{aligned}
&C(K') - C(K) \\
&= \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \left(-2(x'_t)^\intercal (K - K')^\intercal E^\gamma_K x'_t + (x'_t)^\intercal (K - K')^\intercal (R + \gamma B^\intercal P^\gamma_K B)(K - K')x'_t\right)\right] \\
&= -2 \operatorname{Tr}\left(\Sigma^\gamma_{K'}(K - K')^\intercal E^\gamma_K\right) + \operatorname{Tr}\left(\Sigma^\gamma_{K'}(K - K')^\intercal (R + \gamma B^\intercal P^\gamma_K B)(K - K')\right) \\
&= -2\alpha \operatorname{Tr}\left(\Sigma^\gamma_{K'}(\nabla C(K))^\intercal E^\gamma_K\right) + \alpha^2 \operatorname{Tr}\left(\Sigma^\gamma_{K'}(\nabla C(K))^\intercal (R + \gamma B^\intercal P^\gamma_K B)\nabla C(K)\right)
\end{aligned}$$

Replacing $\Sigma^\gamma_{\mathbf{K}'}$ with $(\Sigma^\gamma_{\mathbf{K}'} - \Sigma^\gamma_K + \Sigma^\gamma_K)$ in the first term.

$$\begin{aligned}
&= -\alpha \operatorname{Tr}\left(\nabla C(K)^\intercal \nabla C(K)\right) + 2\alpha \operatorname{Tr}\left((\Sigma^\gamma_{K'} - \Sigma^\gamma_K)\nabla C(K)^\intercal E^\gamma_K\right) \\
&\quad + \alpha^2 \operatorname{Tr}\left(\Sigma^\gamma_{K'} \nabla C(K)^\intercal (R + \gamma B^\intercal P^\gamma_K B)\nabla C(K)\right) \\
&\leq -\alpha \operatorname{Tr}\left(\nabla C(K)^\intercal \nabla C(K)\right) + 2\alpha \|\Sigma^\gamma_{K'} - \Sigma^\gamma_K\| \operatorname{Tr}\left(\nabla C(K)^\intercal E^\gamma_K\right) \\
&\quad + \alpha^2 \|\Sigma^\gamma_{K'}\| \|R + \gamma B^\intercal P^\gamma_K B\| \operatorname{Tr}\left(\nabla C(K)^\intercal \nabla C(K)\right) \\
&\leq -\alpha \operatorname{Tr}\left(\nabla C(K)^\intercal \nabla C(K)\right) + \alpha \frac{\|\Sigma^\gamma_{K'} - \Sigma^\gamma_K\|}{\sigma_{min}(\Sigma^\gamma_K)} \operatorname{Tr}\left(\nabla C(K)^\intercal \nabla C(K)\right) \\
&\quad + \alpha^2 \|\Sigma^\gamma_{K'}\| \|R + \gamma B^\intercal P^\gamma_K B\| \operatorname{Tr}\left(\nabla C(K)^\intercal \nabla C(K)\right) \\
&= -\alpha \left(1 - \frac{\|\Sigma^\gamma_{K'} - \Sigma^\gamma_K\|}{\sigma_{min}(\Sigma^\gamma_K)} - \frac{\alpha}{2}\|\Sigma^\gamma_{K'}\| \|R + \gamma B^\intercal P^\gamma_K B\|\right) \operatorname{Tr}\left(\nabla C(K)^\intercal \nabla C(K)\right)
\end{aligned}$$

By Lemma 3

$$\leq \alpha \frac{\sigma_{min}(R)\mu^2}{\|\Sigma^\gamma_{K^*}\|} \left(1 - \frac{\|\Sigma^\gamma_{K'} - \Sigma^\gamma_K\|}{\mu} - \frac{\alpha}{2}\|\Sigma^\gamma_{K'}\| \|R + \gamma B^\intercal P^\gamma_K B\|\right)(C(K^*) - C(K))$$

To conclude the proof, we provide a lemma bounding the necessary terms,

Lemma 5. (Σ^γ_K *Perturbation and Bound*) *Using the given conditions of the theorem, it holds that:*

$$\|\Sigma^\gamma_{K'} - \Sigma^\gamma_K\| \leq \frac{\mu}{4}$$

and consequently:

$$\|\Sigma^\gamma_{K'}\| \leq \frac{4C(K)}{3\sigma_{min}(Q)}$$

Proof of Lemma 5. First, we prove the second claim (assuming the first claim to be true):

$$\|\Sigma^\gamma_{K'}\| \leq \|\Sigma^\gamma_{K'} - \Sigma^\gamma_K\| + \|\Sigma^\gamma_K\| \leq \frac{\mu}{4} + \frac{C(K)}{\sigma_{min}(Q)} \leq \frac{\|\Sigma^\gamma_K\|}{4} + \frac{C(K)}{\sigma_{min}(Q)}$$

where the middle inequality comes from Lemma 4 and the first claim.

For the first claim, we require some technical details. For these, we reference the reader to the appendix of [3]. Notice that their Lemmas 16 through 23 will also hold in the discounted regime due to our choice of discount factor γ, and that their restrictions on the spectral norm of the state transition matrix $A + BK$ are not needed the discounted setting. By Lemma 16 of [3], we have:

$$\frac{\|\Sigma^\gamma_{K'} - \Sigma^\gamma_K\|}{\mu} \leq 4\left(\frac{C(K)}{\sigma_{min}(Q)\mu}\right)^2 \|B\| \left(\|A + BK\| + 1\right) \|K - K'\|$$

$$\leq 4\alpha \left(\frac{C(K)}{\sigma_{min}(Q)\mu}\right)^2 \|B\| \left(\|A+BK\|+1\right))\|\nabla C(K)\| \leq \frac{1}{4}$$

by the condition on α. □

Temporarily, for convenience, let $\zeta = \left(1 - \frac{\|\Sigma^\gamma_{K'} - \Sigma^\gamma_K\|}{\mu} - \frac{\alpha}{2}\|\Sigma^\gamma_{K'}\|\|R+\gamma B^\intercal P^\gamma_K B\|\right)$. Now, returning to the proof of the theorem, we have:

$$C(K') - C(K) \leq \alpha\zeta \frac{\sigma_{min}(R)\mu^2}{\|\Sigma^\gamma_{K^*}\|} \left(C(K^*) - C(K)\right)$$
$$C(K') - C(K^*) \leq \left(1 - \alpha\zeta \frac{\sigma_{min}(R)\mu^2}{\|\Sigma^\gamma_{K^*}\|}\right) \left(C(K) - C(K^*)\right)$$

Since $\zeta \leq 1$, it remains to show that $\zeta > 0$. By the above Lemma 4 and Lemma 5:

$$\zeta = 1 - \frac{\|\Sigma^\gamma_{K'} - \Sigma^\gamma_K\|}{\mu} - \frac{\alpha}{2}\|\Sigma^\gamma_{K'}\|\|R+\gamma B^\intercal P^\gamma_K B\|$$
$$\geq 1 - \frac{1}{4} - \alpha\frac{2C(K)}{3\sigma_{min}(Q)}\|R+\gamma B^\intercal P^\gamma_K B\| \geq \frac{1}{2}$$

by the condition on α.

Finally, we conclude that

$$C(K') - C(K^*) \leq \left(1 - \alpha \frac{\sigma_{min}(R)\mu^2}{2\|\Sigma^\gamma_{K^*}\|}\right) \left(C(K) - C(K^*)\right).$$

□

Lemma 6. *This lemma provides two more useful bounds:*

$$\|\nabla C(K)\| \leq \frac{C(K)}{\sigma_{min}Q}\sqrt{\frac{\|R+\gamma B^\intercal P^\gamma_K B\|\left(C(K) - C(K^*)\right)}{\mu}}$$

and

$$\|K\| \leq \frac{1}{\sigma_{min}R}\left(\sqrt{\frac{\|R+\gamma B^\intercal P^\gamma_K B\|\left(C(K) - C(K^*)\right)}{\mu}} - \gamma\|B^\intercal P^\gamma_K A\|\right)$$

Proof of Lemma 6. For the first claim, using Lemma 1 and Lemma 4:

$$\|\nabla C(K)\|^2 \leq 4\,\mathrm{Tr}\left(\Sigma^\gamma_K (E^\gamma_K)^\intercal E^\gamma_K \Sigma^\gamma_K\right) \leq \left(\frac{C(K)}{\sigma_{min}Q}\right)^2 \mathrm{Tr}\left((E^\gamma_K)^\intercal E^\gamma_K\right)$$

Using Lemma 3 finishes the proof. For the second claim, also using Lemma 3:

$$\begin{aligned}
\|K\| &\leq \|(R+\gamma B^\intercal P^\gamma_K B)^{-1}\|\|(R+\gamma B^\intercal P^\gamma_K B)K\| \\
&\leq \frac{1}{\sigma_{min}R}\|(R+\gamma B^\intercal P^\gamma_K B)K\| \\
&\leq \frac{1}{\sigma_{min}R}\left(\|(R+\gamma B^\intercal P^\gamma_K B)K + \gamma B^\intercal P^\gamma_K A\| - \gamma\|B^\intercal P^\gamma_K A\|\right) \\
&= \frac{E^\gamma_K}{\sigma_{min}R} - \gamma\frac{\|B^\intercal P^\gamma_K A\|}{\sigma_{min}R} \\
&\leq \frac{\sqrt{\mathrm{Tr}\left((E^\gamma_K)^\intercal E^\gamma_K\right)}}{\sigma_{min}R} - \frac{\|B^\intercal P^\gamma_K A\|}{\sigma_{min}R}
\end{aligned}$$

$$\leq \frac{1}{\sigma_{min} R} \left(\sqrt{\frac{\|R + \gamma B^\intercal P_K^\gamma B\| \left(C(K) - C(K^*)\right)}{\mu}} \gamma \|B^\intercal P_K^\gamma A\| \right)$$

□

Theorem 2. *For an appropriate (constant) setting of the step-size α.*

$$\alpha = \text{poly}\left(\frac{\mu \sigma_{min}(Q)}{C(K_0)}, \frac{1}{\|A\|}, \frac{1}{\|B\|}, \frac{1}{\|R\|}, \sigma_{min}(R)\right)$$

and for

$$N \geq \frac{2\|\Sigma_{K^*}^\gamma\|}{\alpha \mu^2 \sigma_{min}(R)} \log \frac{C(K_0) - C(K^*)}{\varepsilon}$$

the gradient descent algorithm satisfies the following performance bound:

$$C(K_N) - C(K^*) \leq \varepsilon$$

Proof of Theorem 2. By choosing the appropriate step-size and using Lemma 6, we satisfy the condition in Theorem 1. Therefore, given K_0, we have:

$$C(K_1) - C(K^*) \leq \left(1 - \alpha \frac{\sigma_{min}(R)\mu^2}{2\|\Sigma_{K^*}^\gamma\|}\right) \left(C(K_0) - C(K^*)\right)$$

By induction, suppose at time $t > 1$, $C(K_t) \leq C(K_0)$, since the step-size is constant, the conditions remain satisfied and we can apply Theorem 1 again:

$$\begin{aligned} C(K_{t+1}) - C(K^*) &\leq \left(1 - \alpha \frac{\sigma_{min}(R)\mu^2}{2\|\Sigma_{K^*}^\gamma\|}\right) \left(C(K_t) - C(K^*)\right) \\ &\leq \left(1 - \alpha \frac{\sigma_{min}(R)\mu^2}{2\|\Sigma_{K^*}^\gamma\|}\right)^{t+1} \left(C(K_0) - C(K^*)\right) \end{aligned}$$

Thus, using the fact that $\log(1 - x) \leq -x$ for small, positive x, we conclude that

$$N \geq \frac{2\|\Sigma_{K^*}^\gamma\|}{\alpha \mu^2 \sigma_{min}(R)} \log \frac{C(K_0) - C(K^*)}{\varepsilon} \implies C(K_N) - C(K^*) \leq \varepsilon$$

□

3. Convergence to Undiscounted Optimal

Here, we provide our main result concerning the convergence of Algorithm 1. Again, we are working with a linearly parameterized policy K – for clarity, the policy is linear in the state space. We show that any random initial policy K_0 will converge to the global optimum of the undiscounted LQR in finite time. Lemma 7 is useful when producing a bound that is independent of γ and Lemma 8 shows that the optimal policy is more "stable" than the initial policy.

Now, we present the γ-iteration algorithm that will yield convergence to the optimal undiscounted policy with a random initial policy. In a more general sense, one could view this process as moving along a homotopy between the initial discounted optimal policy and the final undiscounted optimal policy, with γ functioning as the "slider" variable. We discuss this idea a bit more in the conclusion.

Lemma 7. *Let $1 \geq \gamma > \varrho > 0$ be two discount rates. Let the matrices Γ and P denote the unique PSD solutions to the discounted DAREs produced by discount rates γ and ϱ, respectively. If we assume that the initial distribution D has identity covariance, Then,*

$$\text{Tr}(\mathrm{P}) \leq \frac{\varrho}{\gamma} \text{Tr}(\Gamma)$$

Algorithm 1 Policy Gradient w/ Random Initialization

Input: K_0 random policy, System matrices A,B,Q,R, tolerance ε.
$\gamma \leftarrow \min\left(\frac{1}{\varrho(A+BK)^2}, 1\right)$
while $\gamma < 1$ **do**
  Run standard policy gradient algorithm until $C^\gamma(K) - C^\gamma(K^*) \leq \varepsilon$
  $\gamma \leftarrow \min\left(\frac{1}{\varrho(A+BK^*)^2}, 1\right)$
  $K_0 \leftarrow K^*$
end while
Run standard policy gradient until convergence to undiscounted optimal

Proof of Lemma 7. Let K_γ^* and K_ϱ^* be the optimal policies w.r.t. the discount factors. Then, we have:

$$C_\varrho(K_\varrho^*) \leq C_\varrho(K_\gamma^*) \leq C_\gamma(K_\gamma^*)$$

where the first inequality follows from the optimality of K_ϱ^* w.r.t to the ϱ-discounted LQR, and the second since $\varrho < \gamma$.

It follows that

$$\begin{aligned}
&C_\varrho(K_\varrho^*) - C_\gamma(K_\gamma^*) \\
&\leq C_\varrho(K_\gamma^*) - C_\gamma(K_\gamma^*) \\
&\leq \mathbb{E}_{x_0}\left[\sum_{t=0}^{\infty}\left((\varrho^t - \gamma^t)x_0^\intercal\left[(A+BK_\gamma^*)^\intercal\right]^t\left(Q + (K_\gamma^*)^\intercal R K_\gamma^*\right)\left[A+BK_\gamma^*\right]^t x_0\right)\right] \\
&= \mathbb{E}_{x_0}\left[\sum_{t=1}^{\infty}\left((\frac{\varrho^t}{\gamma^t} - 1)\gamma^t x_0^\intercal\left[(A+BK_\gamma^*)^\intercal\right]^t\left(Q + (K_\gamma^*)^\intercal R K_\gamma^*\right)\left[A+BK_\gamma^*\right]^t x_0\right)\right] \\
&\leq \left(\frac{\varrho}{\gamma} - 1\right) C_\gamma(K_\gamma^*)
\end{aligned}$$

which implies

$$C_\varrho(K_\varrho^*) \leq \frac{\varrho}{\gamma} C_\gamma(K_\gamma^*)$$

This yields:

$$\begin{aligned}
\mathbb{E}_{x_0}\left[x_0^\intercal \mathrm{P} x_0\right] &\leq \frac{\varrho}{\gamma}\mathbb{E}_{x_0}\left[x_0^\intercal \Gamma x_0\right] \\
\mathrm{Tr}(\mathrm{P}\Sigma_0) &\leq \frac{\varrho}{\gamma}\mathrm{Tr}(\Gamma\Sigma_0)
\end{aligned}$$

where the claim holds since $\Sigma_0 = I$. □

In the following Lemma, we assume that K_0 is *not* stabilizing; if it were, then there would be no need for a discount factor to which we defer to the work in [3].

Lemma 8. *Let K_0 be the unstable initial policy which yields discount rate γ. Let K^* denote the optimal policy with respect to the γ-discounted LQR. Then, we have the following improvement bound:*

$$\varrho\left(A + BK^*\right) \leq \sqrt{1 - \frac{\lambda_{min}(Q)}{\mathrm{Tr}(\mathcal{P})}}\,\varrho\left(A + BK_0\right)$$

where $\mathcal{P}$ is the unique stabilizing solution to the undiscounted discrete-time Algebraic Riccati equation.

Proof of Lemma 8. We know that $\gamma(A+BK^*)$ must be stable by optimality. Thus, we can write P as follows. For convenience, let $\mathcal{A} = A + BK^*$.

$$P = \sum_{t=0}^{\infty} \gamma^t \left(\mathcal{A}^\intercal\right)^t \left(Q + (K^*)^\intercal R K^*\right) \mathcal{A}^t$$

Now, let v be a unit vector in the eigenspace of the largest (modulus) eigenvalue of $\mathcal{A}$. If it is the case that v is a complex vector, then we multiply by the conjugate transpose on the left. Then,

$$\begin{aligned}
\lambda_{max}(P) \geq v^\intercal P v &= \sum_{t=0}^{\infty} \gamma^t v^\intercal \left(\mathcal{A}^\intercal\right)^t \left(Q + (K^*)^\intercal R K^*\right) \mathcal{A}^t v \\
&\geq \lambda_{min}\left(Q + (K^*)^\intercal R K^*\right) \sum_{t=0}^{\infty} \gamma^t v^\intercal \left(\mathcal{A}^\intercal\right)^t \mathcal{A}^t v \\
&= \lambda_{min}\left(Q + (K^*)^\intercal R K^*\right) \sum_{t=0}^{\infty} \gamma^t \|\lambda_{max}(\mathcal{A})^t v\|^2 \\
&= \frac{\lambda_{min}\left(Q + (K^*)^\intercal R K^*\right)}{1 - \gamma \varrho\left(\mathcal{A}\right)^2}
\end{aligned}$$

Lastly, since $Q, R \succ 0$ and by the definition of P,

$$0 < \frac{\lambda_{min}(Q)}{\mathrm{Tr}(\mathcal{P})} \leq \frac{\lambda_{min}(Q)}{\mathrm{Tr}(P)} \leq \frac{\lambda_{min}\left(Q + (K^*)^\intercal R K^*\right)}{\lambda_{max}(P)} < 1$$

where the second inequality follows from Lemma 7 □

Theorem 3. *For γ determined by random initialization, after*

$$N \geq \frac{\mathrm{Tr}(\mathcal{P})}{\lambda_{min}(Q)} \log \gamma^{-1}$$

iterations, Algorithm 1 converges to a stabilizing regime. Here, $\mathcal{P}$ represents the unique stabilizing solution to the undiscounted discrete-time Algebraic Riccati equation.

Proof of Theorem 3. Given K_0, if K_0 is not stabilizing, then by Lemma 8 we have

$$\varrho\left(A + BK_1\right) \leq \sqrt{1 - \frac{\lambda_{min}(Q)}{\mathrm{Tr}(\mathcal{P})}}\, \varrho\left(A + BK_0\right)$$

Since the step size is constant and independent of the initialization, after t iterations, we have that

$$\varrho\left(A + BK_t\right) \leq \left(1 - \frac{\lambda_{min}(Q)}{\mathrm{Tr}(\mathcal{P})}\right)^{t/2} \varrho\left(A + BK_0\right)$$

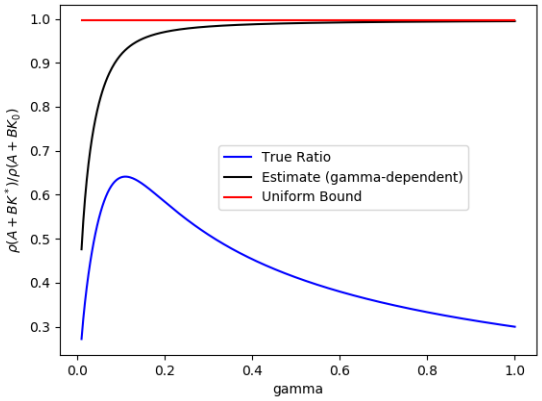
Thus, using the fact that $\log(1-x) \leq -x$ for small, positive x, we conclude that

$$N \geq \frac{2\,\mathrm{Tr}(\mathcal{P})}{\lambda_{min}(Q)} \log\left(\varrho\left(A + BK_0\right)\right) \implies \varrho\left(A + BK_N\right) < 1$$

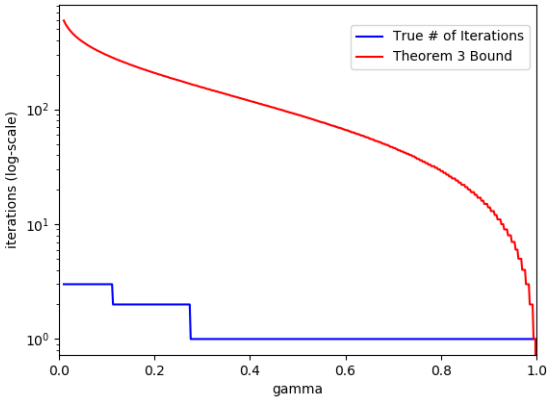
□

Remark 3.1. *If K_0 were stabilizing, then we would have $\gamma = 1$ and there would be no need for Algorithm 1.*

Remark 3.2. *This is a rather pessimist bound; however, this is an intentional choice. In the proof of Lemma 8 we can produce a tighter bound at the cost of adding an additional dependency on γ in $\lambda_{max}(P)$ - we choose to omit that dependency for a cleaner bound that only depends on the initialization through $\log\left(\varrho\left(A + BK_0\right)\right)$.*



(A) Improvement Ratio vs. Bound



(B) # Iterations vs. Bound

FIGURE 1. System matrices A, B both random 10×10 matrices reproduced below. The estimate bound is the bound that retains the secondary dependency on γ, and the uniform bound is the one used in our results. To generate these figures, we use $\varrho\left(A + BK_0\right) = 1/\sqrt{\gamma}$.

$$A = \begin{pmatrix} 0.87 & 0.82 & 2.87 & -0.89 & -0.25 & 0.91 & -2.17 & -0.46 & -0.26 & 0.53 \\ 1.23 & 2.15 & 0.67 & -0.93 & -1.11 & -0.26 & 0.88 & 1.73 & 0.54 & 0.83 \\ -0.59 & 0.13 & -0.26 & 1.57 & 0.14 & -0.90 & 0.40 & -0.43 & -0.94 & 0.35 \\ -0.60 & -2.06 & 0.68 & 1.41 & 0.71 & -1.03 & 0.12 & -1.23 & 0.76 & -0.99 \\ 0.69 & -0.91 & -0.59 & -0.93 & -0.10 & -2.67 & -1.13 & -0.15 & 1.99 & 0.08 \\ 1.89 & 0.34 & 1.49 & 0.39 & 0.44 & 0.24 & -0.03 & -0.10 & 1.54 & -1.00 \\ -0.65 & -0.12 & -0.82 & 1.30 & -0.42 & 0.03 & -0.84 & -1.27 & 0.36 & 0.18 \\ -0.17 & -0.21 & -0.72 & -0.62 & -1.15 & 0.64 & 0.79 & 1.23 & -2.04 & 0.15 \\ -0.68 & -0.46 & 0.65 & -0.28 & -1.78 & -0.07 & -0.24 & -1.35 & -0.79 & -0.33 \\ -1.90 & -0.48 & 0.22 & -1.34 & 1.00 & -1.60 & 1.33 & -0.09 & -0.26 & 0.57 \end{pmatrix}$$

$$B = \begin{pmatrix} -0.82 & -0.53 & -1.22 & 0.32 & -0.42 & -0.11 & 1.20 & 0.21 & -0.99 & -1.53 \\ -0.48 & 0.93 & 1.07 & -0.08 & 0.26 & -0.01 & 0.34 & -0.76 & -1.93 & -0.06 \\ 0.44 & 0.23 & 0.69 & -0.24 & 1.25 & 1.11 & -1.21 & -0.62 & 1.06 & 1.05 \\ -0.37 & -1.98 & 1.87 & -0.90 & 1.19 & -2.13 & -1.31 & 0.85 & 1.00 & 1.13 \\ 0.72 & -0.41 & -0.18 & 1.48 & -0.47 & 0.68 & -0.41 & 1.69 & 1.00 & 1.39 \\ -0.87 & 0.86 & 1.13 & 0.69 & -0.36 & 1.40 & -0.56 & -0.29 & 1.17 & 1.08 \\ -0.46 & -0.81 & 1.53 & 1.70 & 1.68 & 0.52 & -0.16 & -1.26 & 0.06 & -0.55 \\ -0.07 & -0.26 & 1.09 & 0.81 & -0.93 & 0.57 & 0.25 & 0.10 & -2.46 & 0.14 \\ -1.05 & -1.52 & -0.65 & -1.44 & 1.28 & -1.09 & 0.46 & 0.45 & -1.07 & 1.24 \\ -1.91 & 0.88 & 1.65 & -0.24 & -0.61 & 1.18 & -1.41 & -0.86 & 1.88 & 1.01 \end{pmatrix}$$

4. PRELIMINARY GENERALIZATIONS

In this section, we take a brief look into possible approaches to generalizing our main result. We first consider a natural generalization to stochastic policies with entropy regularization, and then we consider a simple non-linear case in the one-dimensional setting. Lastly, we provide an counterexample using a policy-approximator that is linear in parameters to show that we should not always expect policy gradient methods to converge to the global optimum.

4.1. **Linear-Gaussian Policies.**

Here, we show that Policy Gradient methods also converge for the class of linear-Gaussian policies (fixed σ).

$$\left\{\pi(\cdot|x) = N(Kx, \sigma^2 I_{k\times k})\right\} \implies u = Kx + \sigma\eta$$

where η is standard normal.

From optimal control theory, we know that the Q function can be written as follows

$$Q(x,u) = \begin{bmatrix} x \\ u \end{bmatrix}^\intercal \begin{pmatrix} Q + A^\intercal P_K A & A^\intercal P_K B \\ B^\intercal P_K A & R + B^\intercal P_K B \end{pmatrix} \begin{bmatrix} x \\ u \end{bmatrix}$$

using column concatenation of x and u. Thus, by policy gradient theorem and Stein's lemma

$$\begin{aligned}
\nabla_K C(K) &= \mathbb{E}_{x\sim\nu_K}\left[\nabla_K \log \pi_K(u|x) Q_K(x,u)\right] \\
&= \mathbb{E}\left[\nabla_K \frac{-(u-Kx)^2}{2\sigma^2} Q_K(x,u)\right] \\
&= \mathbb{E}\left[\frac{u-Kx}{\sigma^2} x^\intercal Q_K(x, Kx+\sigma\eta)\right] \\
&= \sigma^{-2}\mathbb{E}\left[\sigma\eta x^\intercal Q_K(x, Kx+\sigma\eta)\right] \\
\text{now, using Stein's lemma} &= \mathbb{E}\left[\nabla_u Q_K|_{(x,Kx+\sigma\eta)} x^\intercal\right] \\
&= 2\mathbb{E}\left[((R+B^\intercal P_K B)u + B^\intercal P_K A x)\,|_{(x,Kx+\sigma\eta)} x^\intercal\right] \\
&= 2\mathbb{E}\left[((R+B^\intercal P_K B)(Kx+\sigma\eta) + B^\intercal P_K A x)\, x^\intercal\right] \\
&= 2\mathbb{E}\left[((R+B^\intercal P_K B)K + B^\intercal P_K A)\, x x^\intercal\right] \\
&= 2E_K \Sigma_K
\end{aligned}$$

where $\nu_K = N(0, \Sigma_K)$ is the invariant distribution under the policy π_K.

Since the Policy Gradient takes the same form as in our previous analysis, the convergence results will also hold.

4.2. **Piecewise-Linear LQR in 1D.**

We consider a policy of the form

$$u = K_1 \max(x, 0) + K_2 \min(x, 0)$$

where $\vartheta = (K_1, K_2)$ are the parameters to be learned. This class of policies clearly contains the optimal policy (*i.e.*, $K_1 = K_2 = K^*$).

In this setting, there are three possibilities for the dynamics of the system.

(1) The dynamics produced by K_1 and K_2 are both forward invariant. Thus, if $x_0 > 0$, then $x_t > 0\ \forall t$, and similarly for $x_0 < 0$.
(2) After the first time step, the region corresponding to K_1 is immediately mapped to the region of K_2, and K_2 is forward invariant. For any x_0, $x_1 < 0$ and $x_t < 0\ \forall t \geq 1$ (this applies the other way around as well).
(3) The state bounces back and forth between the two regions.

To show that policy gradient methods converge in with this class of policies, we would like to show that the gradient takes the same form as that of Lemma 1, and that we can produce a similar cost-difference bound.

In the first setting, for any trajectory, the gradient of the policy is simply that of the "normal" LQR.

In the second setting, the gradient takes the form

$$\nabla C(\vartheta) = (2RK_1 x_0 x_0^\intercal, \gamma \nabla_{K_2} C(K_2))^\intercal$$

which shows that K_2 will move towards K^* and K_1 will move towards zero. As K_1 updates with this gradient, two things can happen: K_1 will update in a way that pushes the dynamics to either setting 1) or setting 3).

The third setting is the most interesting to consider.

Definition 4. Define P_1 and P_2 as the unique positive definite solutions to the following Lyapunov equations for positive and negative x_0, respectively.

$$\begin{aligned}
P_1 &= \left(Q + K_1^2 R\right) + \gamma (A + BK_1)\left(Q + K_2^2 R\right)(A + BK_1) \\
&\quad + \gamma^2 (A + BK_1)(A + BK_2) P_1 (A + BK_2)(A + BK_1) \\
P_2 &= \left(Q + K_2^2 R\right) + \gamma (A + BK_2)\left(Q + K_1^2 R\right)(A + BK_2) \\
&\quad + \gamma^2 (A + BK_2)(A + BK_1) P_2 (A + BK_1)(A + BK_2)
\end{aligned}$$

Also, we define the corresponding notation shortcuts $E_1, E_2, \Sigma_{2n}^K, \Sigma_{2n+1}^K$

$$\begin{aligned}
E_1 &= (R + \gamma B P_2 B) K_1 + B P_2 A \\
E_2 &= (R + \gamma B P_1 B) K_2 + B P_1 A \\
\Sigma_{2n}^K &= \sum_{t=0}^{\infty} \gamma^{2t} x_{2t}^2 \\
\Sigma_{2n+1}^K &= \sum_{t=0}^{\infty} \gamma^{2t+1} x_{2t+1}^2
\end{aligned}$$

Proposition 4. *P_1 and P_2 are related as follows*

$$\begin{aligned}
P_1 &= \left(Q + K_1^2 R\right) + \gamma (A + BK_1) P_2 (A + BK_1) \\
P_2 &= \left(Q + K_2^2 R\right) + \gamma (A + BK_2) P_1 (A + BK_2)
\end{aligned}$$

Then, we can write the cost as $C_\vartheta(x_0) = x_0 P_1 x_0$ or $C_\vartheta(x_0) = x_0 P_2 x_0$ depending on the value of x_0. Now, computing the gradient

$$\begin{aligned}
\nabla_{K_1} C_\vartheta(x_0) &= \left(2RK_1 + \gamma 2B(Q + K_2^2 R)(A + BK_1)\right) x_0^2 + \gamma^2 \nabla_{K_1} C_\vartheta(x_2)) \\
&= \left(2RK_1 + \gamma 2B(Q + K_2^2 R)(A + BK_1)\right) x_0^2 \\
&\quad + \gamma^2 2B(A + BK_2) P_1 (A + BK_2)(A + BK_1) x_0^2 + \gamma^2 \nabla_{K_1} C_\vartheta(x_2)|_{x_2 = (A+BK_2)(A+BK_1)x_0} \\
&= \sum_{t=0}^{\infty} \gamma^{2t} 2 \Bigg(\left(R + \gamma B \left((Q + K_2^2 R) + \gamma (A + BK_2) P_1 (A + BK_2)\right) B\right) K_1 \\
&\quad + \gamma B \left((Q + K_2^2 R) + \gamma (A + BK_2) P_1 (A + BK_2)\right) A \Bigg) x_{2t}^2 \\
&= \sum_{t=0}^{\infty} \gamma^{2t} 2 \Bigg(\left(R + \gamma B P_2 B\right) K_1 + \gamma B P_2 A \Bigg) x_{2t}^2 \\
&= 2E_1 \sum_{t=0}^{\infty} \gamma^{2t} x_{2t}^2
\end{aligned}$$

Using the same recursion trick for K_2

$$\begin{aligned}
\nabla_{K_2} C(\vartheta)|_{x_0} &= \gamma 2RK_2 \left((A + BK_1)x_0\right)^2 + \gamma^2 \nabla_{K_2} C_\vartheta(x_2)) \\
&= \gamma 2RK_2 \left((A + BK_1)x_0\right)^2 + \gamma^2 2BP_1(A + BK_2) \left((A + BK_1)x_0\right)^2
\end{aligned}$$

$$
\begin{aligned}
&+ \gamma^2 \nabla_{K_2} C_\vartheta(x_2)|_{x_2=(A+BK_2)(A+BK_1)x_0} \\
&= \sum_{t=0}^{\infty} \gamma^{2t+1} 2 \Big(\left(R + \gamma B P_1 B\right) K_2 + \gamma B P_1 A \Big) x_{2t+1}^2 \\
&= 2E_2 \sum_{t=0}^{\infty} \gamma^{2t+1} x_{2t+1}^2
\end{aligned}
$$

Definition 5. Here, we define the two-step versions of the Q and Advantage functions.

$$
\begin{aligned}
Q_{2,K}(x, u, u') &\coloneqq r(x, u) + \gamma r(x', u') + \gamma^2 V_K(x'') \\
A_{2,K}(x, u, u') &\coloneqq Q_{2,K}(x, u, u') - V_K(x)
\end{aligned}
$$

where V is the value function and $r(x, u)$ is the reward (cost) for a given state and action.

Lemma 9. *We can express the two-step advantage in terms of the one-step advantage*

$$
A_{2,K}(x, u, u') = A_K(x, u) + \gamma A_K(x', u')
$$

Additionally, we can write the cost-difference between two policies as follows

$$
V_{\hat{K}}(x) - V_K(x) = \sum_{t=0}^{\infty} \gamma^{2t} A_{2,K}(\hat{x}_{2t}, \hat{u}_{2t}, \hat{u}'_{2t+1})
$$

Proof of Lemma 9. The first claim follows from the defintion of the two-step advantage,

$$
\begin{aligned}
A_{2,K}(x, u, u') &= Q_{2,K}(x, u, u') - V_K(x) \\
&= r(x, u) + \gamma r(x', u') + \gamma^2 V_K(x'') - V_K(x) \\
&= r(x, u) + \gamma V_K(x') - V_K(x) + \gamma \left(r(x', u') + \gamma V_K(x'') - V_K(x')\right) \\
&= A_K(x, u) + \gamma A_K(x', u')
\end{aligned}
$$

where the next state x' is determined by x and u.

The next claim comes from Lemma 2

$$
\begin{aligned}
V_{\hat{K}}(x) - V_K(x) &= \sum_{t=0}^{\infty} \gamma^t A_K(\hat{x}_t, \hat{u}_t) \\
&= \sum_{t=0}^{\infty} \gamma^{2t} \left(A_K(\hat{x}_{2t}, \hat{u}_{2t}) + \gamma A_K(\hat{x}_{2t+1}, \hat{u}_{2t+1})\right) \\
&= \sum_{t=0}^{\infty} \gamma^{2t} A_{2,K}(\hat{x}_{2t}, \hat{u}_{2t}, \hat{u}'_{2t+1})
\end{aligned}
$$

□

Now, if K^* also produces "back-and-forth" dynamics (*i.e.*, $A + BK^* < 0$), then we can write the cost difference as follows

$$
\begin{aligned}
&C(\vartheta) - C(K^*) \\
&= -\mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^{2t} A_{2,\vartheta}(x^*_{2t}, u^*_{2t}, u^*_{2t+1})\right] \\
&= -\sum_{t=0}^{\infty} \gamma^{2t} \Bigg(2(K^* - K_1)\big((R + \gamma B P_2 B)K_1 + B P_2 A\big) x_{2t}^2 + (K^* - K_1)(R + \gamma B P_2 B)(K^* - K_1) x_{2t}^2 \\
&\quad + \gamma \left(2(K^* - K_2)\big((R + \gamma B P_1 B)K_2 + B P_1 A\big) x_{2t+1}^2 + (K^* - K_2)(R + \gamma B P_1 B)(K^* - K_2) x_{2t+1}^2\right) \Bigg)
\end{aligned}
$$

$$
\begin{aligned}
&= -\sum_{t=0}^{\infty} \gamma^{2t} \Big(2(K^* - K_1) E_1 x_{2t}^2 + (K^* - K_1)(R + \gamma B P_2 B)(K^* - K_1) x_{2t}^2 \\
&\quad + \gamma \left(2(K^* - K_2) E_2 x_{2t+1}^2 + (K^* - K_2)(R + \gamma B P_1 B)(K^* - K_2) x_{2t+1}^2 \right) \Big) \\
&\leq \mathbb{E} \left[\sum_{t=0}^{\infty} \gamma^{2t} E_1 (R + \gamma B P_2 B)^{-1} E_1 x_{2t}^2 + \sum_{t=0}^{\infty} \gamma^{2t+1} E_2 (R + \gamma B P_1 B)^{-1} E_2 x_{2t+1}^2 \right] \\
&= E_1 (R + \gamma B P_2 B)^{-1} E_1 \Sigma_{2n}^* + E_2 (R + \gamma B P_1 B)^{-1} E_2 \Sigma_{2n+1}^* \\
&\leq \frac{\Sigma_{2n}^{K^*}}{\|R\| \|\Sigma_{2n}^{K_1}\|^2} \left(\nabla_{K_1} C(\vartheta) \right)^2 + \frac{\Sigma_{2n+1}^{K^*}}{\|R\| \|\Sigma_{2n+1}^{K_2}\|^2} \left(\nabla_{K_2} C(\vartheta) \right)^2
\end{aligned}
$$

Now, if K^* produces the forward invariant dynamics described in (1), then, we can write the cost difference with the one-step advantage as in Lemma 2, using K_1 or K_2 depending on the sign of x_0. Supposing $x_0 > 0$, we would have

$$
\begin{aligned}
C(\vartheta) - C(K^*) &= -\mathbb{E} \left[\sum_{t=0}^{\infty} \gamma^t A_\vartheta(x_t^*, u_t^*) \right] \\
&\leq \frac{\Sigma_n^{K^*}}{\|R\| \|\Sigma_{2n}^{K_1}\|^2} \left(\nabla_{K_1} C(\vartheta) \right)^2
\end{aligned}
$$

and similarly for K_2 if $x_0 < 0$.

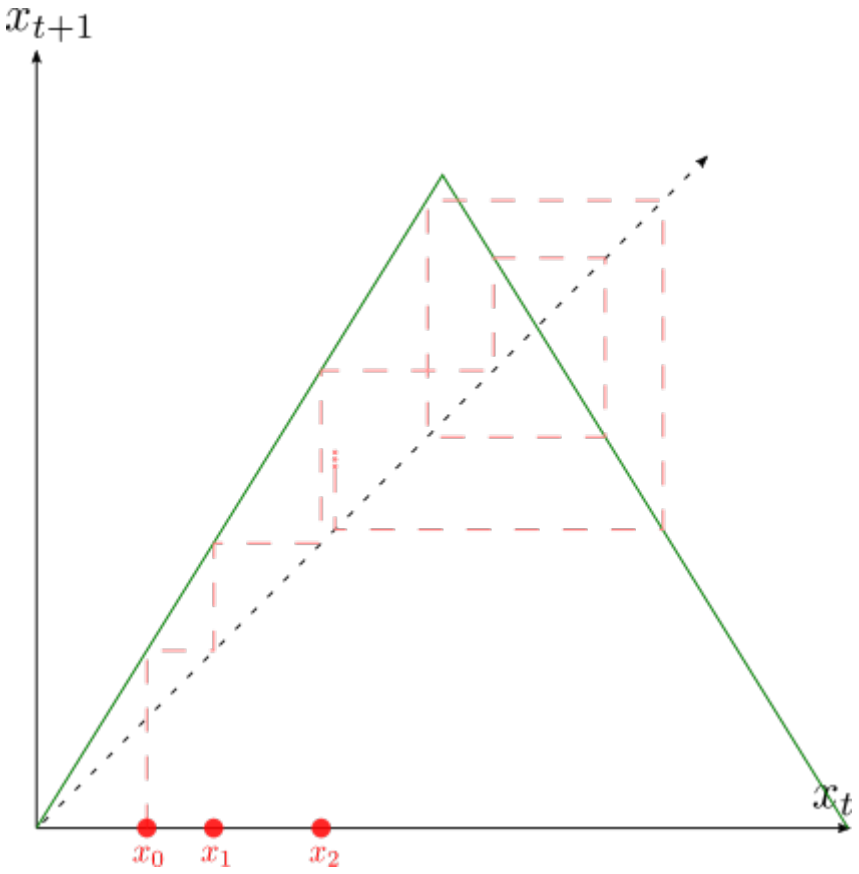
4.3. **Counterexample.**

In this section we show that there exists a policy that is linear in the parameters that does not converge to the global optimum of the LQR.

We first provide an example of a dynamics diagram that we will use later on in this section (Example 4.1). This type of diagram is commonly used to depict the behavior of dynamical systems produced by iterating the map of interest; in the case of Policy Gradient methods in the LQR, that map is determined by the system matrices (A, B) and the policy π.

Example 4.1. *Consider the state dynamics supported on* $(0, 2)$

$$
x_{t+1} = \begin{cases} 2x_t & 0 \leq x \leq 1 \\ -2(x_t - 2) & 1 \leq x \leq 2 \end{cases}
$$



The tent map dynamics are represented by the green line and the black dotted line represents $x_{t+1} = x_t$. The pink dotted line represents the "trajectory" from x_0. The first seven steps of the dynamics under the tent map are shown, with explicit labels for the first three.

Now continuing, we consider a simple setting in 1D with the dynamics:

$$x_{t+1} = Ax_t + Bu_t$$
$$u_{t+1} = F_\vartheta(x_t) = \vartheta_1 x_t + \vartheta_2 \Lambda(x_t)$$

where Λ is to be a function such that the state dynamics behave as depicted in Fig. 2. The optimal policy for the LQR belongs to this policy class and corresponds to the parameters $(\vartheta_1^*, 0)$.

To better understand this example, we consider the how changing ϑ_1 and changing ϑ_2 affect the dynamics and the cost of a trajectory given x_0. Specifically, let us consider $(A, B) = (0, 1)$ and $(Q, R) = (1, 1)$.

Lemma 10. *Given $(A, B) = (0, 1), (Q, R) = (1, 1)$, to update the parameters without dramatically increasing the cost, then we must update along the path determined by the differential equation below*

$$\mathrm{d}_t x_1 = \mathrm{d}_t A x_0 + \mathrm{d}_t B u_0 = 0 \implies \mathrm{d}_t \vartheta_1 = -\frac{\Lambda(x_0)\mathrm{d}_t \vartheta_2}{B x_0} \tag{4.1}$$

Proof. We first notice that independently changing ϑ_2 does not affect any of the behavior in the region Θ. Intuitively, increasing and decreasing ϑ_2 has the effect of rotating the blue section of Fig. 2 clockwise or anti-clockwise, respectively. Consequently, x_0 will no longer map to the single point that yields convergence of the state to zero, thus, severely increasing the cost of the trajectory. A similar fate is met if we increase or decrease ϑ_1. □

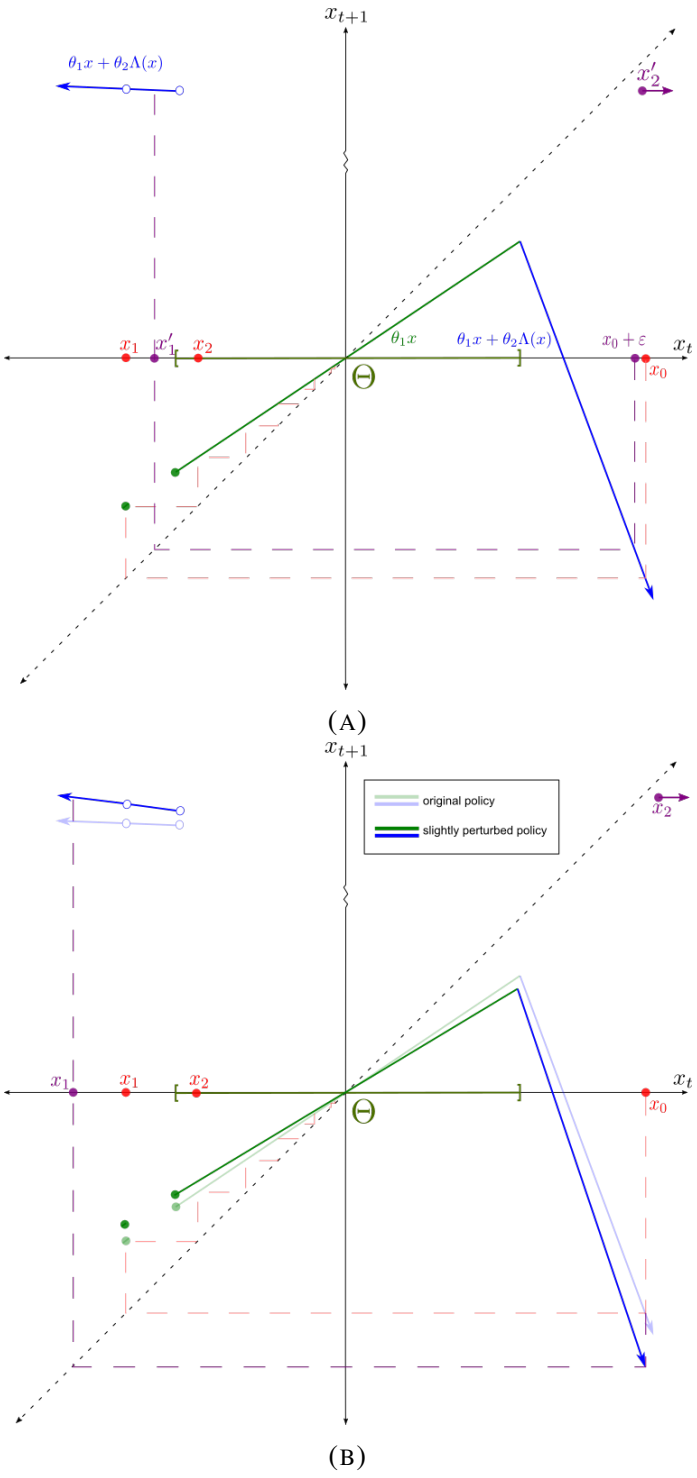


FIGURE 2. The policy is a linear function of the state in Θ ($\forall x \in \Theta, \Lambda(x) = 0$). Outside Θ, there is a single point x_1 such that $\Lambda(x_1) = 0$ as well.
(A) shows that any small change to x_0 will result in a large increase to the cost of the trajectory since x_t will not tend to zero unless x_0 is in Θ or precisely x_0.
(B) shows a similar example where the policy is slightly perturbed. Like above, any such change will result in a large increase to the cost function.

Theorem 5. *Let* $(A, B) = (0, 1), (Q, R) = (1, 1)$, *and* $(\vartheta_1, \vartheta_2) = (0.75, 1)$. *Additionally, let* $\Theta = [-2.25, 1]$, $x_0 = 4$ *and* $x_1 = -3$. *Referencing the idea of Fig. 2, let*

$$\Lambda = \begin{cases} -2x + 100, & x \in \mathbb{R}_{<0} \setminus \Theta \cup \{x_1\} \\ 0, & x \in \Theta \cup \{x_1\} \\ -2x + 2, & x \in \mathbb{R}_{>0} \setminus \Theta \cup \{x_1\} \end{cases}$$

Then, in this setting, there exists a locally optimal policy that is not the globally optimal policy for the LQR.

Proof. For these example parameters, the optimal policy $\vartheta^* = (0, 0)$. With all this in mind, we can continue forwards in determining the updates for the policy parameters. It's important to notice that in this setting, the cost function may not be $\mathscr{C}^1$; however, we can instead use subgradients to yield a similar update process.

Substituting values into equation 4.1, we see that $\mathrm{d}_t\vartheta_1 = \frac{3}{2}\mathrm{d}_t\vartheta_2$. Now, assuming we have chosen the step-size appropriately, as we proceed along this update process, for ϑ_1 to reach ϑ_1^*, the sum of all the updates will equal $\vartheta_1 - \vartheta_1^* = \sum_t \mathrm{d}_t\vartheta_1 = 0.75$. Therefore, we can also determine the net change for ϑ_2 since we must update along the path described by equation 4.1; thus, we see that the final value $\vartheta_2 = \vartheta_2(\vartheta_1) = -0.125 \neq \vartheta_2^*$. □

5. Discussion and Conclusion

This paper provides a convergence guarantee for model-based policy gradient methods in the setting of the LQR. Building on existing results, we use discounting to ensure finite costs, which allow us to show that the assumption of a stabilizing initial policy is not necessary for convergence to the globally optimal policy. Similar results can be obtained in the finite-horizon.

Future Work. Using a discount factor to guarantee finite costs can also be applied to solving issues involving a chaotic policy. In our work, we focused on divergent costs caused by divergent states; however, it is possible for the cost to diverge while the states remain in some compact set – this can be seen in some piece-wise linear policies (see Example 4.1). There are a few directions to extend the work from this paper.

Model-Free Case. The results in this paper are all use knowledge of system dynamics. A natural way forwards would be to demonstrate that our results are also true in the model-free scenario where simulated trajectories are used in a stochastic policy gradient method.

Nonlinear Extensions. In light of the piecewise-linear example in one dimension, it may be possible to extend this to higher dimensions. Some difficulties we encountered were how to suitably extend the idea of a two-parameter policy to $\mathbb{R}^n$. The end goal of this line of work would be to provide a convergence result for a simple ReLU neural network in the LQR, or to show that there is no such guarantee. As with all non-linear extensions, the friendly linear dynamics of the LQR will no longer be present, making the analysis much more challenging. One could also consider all of the above approaches from the entropy-regularized perspective.

Homotopy-based Iteration. The main result of this paper is to show that sequentially updating the discount factor is a viable way of moving from an unstable policy to a stable policy. We think it would be interesting to take a more general perspective on this method and apply it to more general reinforcement learning settings. It is unclear if a discounting approach would succeed in an environment where the optimal policies are not homotopic in the discount factor.

References

[1] Z. Ahmed, N. Le Roux, M. Norouzi, and D. Schuurmans. Understanding the impact of entropy on policy optimization. In: K. Chaudhuri and R. Salakhutdinov, eds., *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research* (PMLR, 2019).

[2] S. J. Bradtke. Reinforcement learning applied to linear quadratic regulation. In: *Advances in Neural Information Processing Systems 5* (Morgan-Kaufmann, 1993), pp. 295–302.

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

[4] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In: J. Dy and A. Krause, eds., *Proceedings of the 35th*

International Conference on Machine Learning, volume 80 of *Proceedings of Machine Learning Research* (PMLR, 2018).
[5] B. Liu, Q. Cai, Z. Yang, and Z. Wang. Neural trust region/proximal policy optimization attains globally optimal policy. In: *Advances in Neural Information Processing Systems 32* (Curran Associates, Inc., 2019), pp. 10565–10576.
[6] J. Liu, X. Gu, D. Zhang, and S. Liu. On-policy reinforcement learning with entropy regularization (2019).
[7] J. Mei, C. Xiao, C. Szepesvari, and D. Schuurmans. On the global convergence rates of softmax policy gradient methods. In: *Proceedings of Machine Learning and Systems* (PMLR, 2020), pp. 10170–10179.
[8] R. Postoyan, L. Buşoniu, D. Nešić, and J. Daafouz. Stability analysis of discrete-time infinite-horizon optimal control with discounted cost. *IEEE Transactions on Automatic Control* **62** (2017), 2736–2749.
[9] J. Schulman, S. Levine, P. Abbeel, M. Jordan, and P. Moritz. Trust region policy optimization. In: *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research* (PMLR, 2015).
[10] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov. Proximal policy optimization algorithms. *CoRR* **abs/1707.06347**.
[11] R. Sutton and A. Barto. *Reinforcement Learning: An Introduction*. Adaptive Computation and Machine Learning series (MIT Press, 2018).
[12] R. Sutton, D. Mcallester, S. Singh, and Y. Mansour. Policy gradient methods for reinforcement learning with function approximation. *Adv. Neural Inf. Process. Syst* **12** (2000), 1057–1063.
[13] S. Tu and B. Recht. Least-squares temporal difference learning for the linear quadratic regulator. In: *Proceedings of the 35th International Conference on Machine Learning*, volume 80 (PMLR, 2018).
[14] L. Wang, Q. Cai, Z. Yang, and Z. Wang. Neural policy gradient methods: Global optimality and rates of convergence. In: *International Conference on Learning Representations* (2020).
[15] R. J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning* **8** (1992), 229–256.
[16] Z. Yang, Y. Chen, M. Hong, and Z. Wang. Provably global convergence of actor-critic: A case for linear quadratic regulator with ergodic cost. In: *Advances in Neural Information Processing Systems 32* (Curran Associates, Inc., 2019), pp. 8353–8365.

APPENDIX A. SOME ESTIMATES FOR THE CONVERGENCE OF NONLINEAR LQR

In this section we consider the framework of entropy-regularized reinforcement learning. In the same vein as the above section, we look further into the prospect of developing a convergence result with a stochastic policy. This section is also motivated by the results from [7, 16, 6, 4].

Proposition 6. *For a given soft Q-function $\tilde{Q}$, the "optimal" policy $\pi^{\mathcal{B}}$ is the Boltzmann (or energy-based) policy given by*

$$\begin{aligned}\pi^{\mathcal{B}}(a|s) &= \arg\max_{\pi}\left(\mathbb{E}_{a\sim\pi}\left[\tilde{Q}^{\pi}(s_0,a)-\tau\log\pi_{\vartheta}(a_t|s_t)\right]\right)\\ &= \frac{\exp\left(\tau^{-1}\tilde{Q}(s,a)\right)}{\int_{\mathcal{A}}\pi(a'|s)\exp\left(\tau^{-1}\tilde{Q}(s,a')\right)\mathrm{d}a'}\end{aligned}$$

where τ is an adjustable hyperparameter.

Proof of Proposition 6. We first assume that $\tilde{Q}$ and $\int\exp\left(\tau^{-1}\tilde{Q}(s,a)\right)\mathrm{d}\pi$ are bounded for any s, and for both π and π'.

Given a policy π and corresponding Q-function $\tilde{Q}^{\pi}$, we define a new policy π' as

$$\pi'(\cdot|s)\propto\exp\left(\tau^{-1}\tilde{Q}^{\pi}(s,\cdot)\right)$$

where the proportionality is due to a normalization factor written in the proposition statement.

We can rewrite the quantity we seek to maximize as follows

$$\mathbb{E}_{a\sim\pi}\left[\tilde{Q}^{\pi}(s_0,a)-\tau\log\pi(a|s)\right]=\mathbb{E}_{a\sim\pi}\left[\tilde{Q}^{\pi}(s_0,a)\right]-\tau\int\pi(a|s)\left(\log\pi(a|s)-\log\pi'(a|s)\right)\mathrm{d}a$$

$$
\begin{aligned}
&\quad - \tau \int \pi(a|s) \log \pi'(a|s) \mathrm{d}a \\
&= \mathbb{E}_{a\sim\pi}\left[\tilde{Q}^{\pi}(s_0, a)\right] - \tau D_{KL}\left(\pi(\cdot|s) \| \pi'(\cdot|s)\right) \\
&\quad - \tau \int \pi(a|s) \log \left(\frac{\exp\left(\tau^{-1}\tilde{Q}(s,a)\right)}{\int_{\mathcal{A}} \pi(a'|s) \exp\left(\tau^{-1}\tilde{Q}(s,a')\right) \mathrm{d}a'} \right) \\
&= -\tau D_{KL}\left(\pi(\cdot|s) \| \pi'(\cdot|s)\right) + \tau \log \int \exp\left(\tau^{-1}\tilde{Q}(s,a)\right) \mathrm{d}\pi
\end{aligned}
$$

From this, we can see that

$$
\mathbb{E}_{a\sim\pi}\left[\tilde{Q}^{\pi}(s_0, a) - \tau \log \pi_{\vartheta}(a|s)\right] \leq \mathbb{E}_{a\sim\pi'}\left[\tilde{Q}^{\pi}(s_0, a) - \tau \log \pi'(a|s)\right]
$$

Since the any sub-optimal policy can be improved in this way, we know the optimal policy must be of an energy-based form. □

We now consider a restricted family of softmax policies for continuous action spaces. These policies have the form

$$
\pi_{\vartheta}(a|s) = \frac{\exp -\frac{(a - F_{\vartheta}(s))^2}{2\sigma^2}}{\int \exp -\frac{(a' - F_{\vartheta}(s))^2}{2\sigma^2} \mathrm{d}a} = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(a - F_{\vartheta}(s))^2}{2\sigma^2}}
$$

for $F_{\vartheta}(s) = \vartheta s + \vartheta' \Lambda(s)$ and for a certain function Λ to be decided later. The above policies are Gaussian policies around a perturbation of the linear policy $\pi_l(s) = \vartheta s$ (we remain in 1 dimension for the moment). We would like to prove that the policy gradient algorithm converges to a global minimum in this setting, provided that the initial distribution has full support on the real axis. To do so, we must show that the policy gradient dynamics act as a contraction in the space of parameters. In other words, we would like to find a metric, either in the space of parameters or in the space of measures, that "shows" that the policy gradient dynamics converge to the unique minimizer $(\vartheta, \vartheta') = (\vartheta^*, 0)$, where ϑ^* is the optimal policy of the LQR in 1 dimension for the given system matrices (values) A, B, Q, R.

The policy gradient update in the regularized framework reads as follows:

$$
\frac{\mathrm{d}}{\mathrm{d}t}\vartheta = -\nabla_{\vartheta} \mathbb{E}_{s_0\sim\nu}\left[V^{\pi_{\vartheta}}(s_0)\right]
$$

where

$$
\mathbb{E}_{s_0\sim\nu}\left[V^{\pi_{\vartheta}}(s_0)\right] = \mathbb{E}_{s_0\sim\nu}\left[\int \left(Q^{\pi_{\vartheta}}(s_0, a) - \tau \log(\pi_{\vartheta}(\cdot|s))\right) \pi_{\vartheta}(\mathrm{d}a|s_0)\right]
$$

Taking the derivative in ϑ we then obtain

$$
\begin{aligned}
\nabla_{\vartheta} \mathbb{E}_{s_0\sim\nu}\left[V^{\pi_{\vartheta}}(s_0)\right] &= \mathbb{E}_{s_0\sim\nu}\left[\int \left(\nabla_{\vartheta} Q^{\pi_{\vartheta}}(s_0, a) - \tau \frac{\nabla_{\vartheta}\pi_{\vartheta}(a|s)}{\pi_{\vartheta}(a|s)}\right) \pi_{\vartheta}(\mathrm{d}a|s_0)\right] \\
&\quad + \mathbb{E}_{s_0\sim\nu}\left[\int \left(Q^{\pi_{\vartheta}}(s_0, a) - \tau \log(\pi_{\vartheta}(\cdot|s))\right) \nabla_{\vartheta}\pi_{\vartheta}(\mathrm{d}a|s_0)\right] \\
&= \mathbb{E}_{s_0\sim\nu}\left[\int \nabla_{\vartheta} Q^{\pi_{\vartheta}}(s_0, a) \pi_{\vartheta}(\mathrm{d}a|s_0)\right] \\
&\quad + \mathbb{E}_{s_0\sim\nu}\left[\int \left(Q^{\pi_{\vartheta}}(s_0, a) - \tau \log(\pi_{\vartheta}(\cdot|s))\right) \nabla_{\vartheta}\pi_{\vartheta}(\mathrm{d}a|s_0)\right] \\
&= \mathbb{E}_{s_0\sim\nu}\left[\gamma \int \int \nabla_{\vartheta} V^{\pi_{\vartheta}}(s_1) P(s_1|s_0, a) \pi_{\vartheta}(\mathrm{d}a|s_0)\right]
\end{aligned}
$$

$$
\begin{aligned}
&+ \mathbb{E}_{s_0\sim\nu}\left[\int \left(Q^{\pi_\vartheta}(s_0,a) - \tau\log(\pi_\vartheta(\cdot|s))\right)\nabla_\vartheta \pi_\vartheta(\mathrm{d}a|s_0)\right] \\
&= \mathbb{E}_{s_0\sim\varrho_\pi(\nu)}\left[\int \left(Q^{\pi_\vartheta}(s_0,a) - \tau\log(\pi_\vartheta(\cdot|s))\right)\nabla_\vartheta \pi_\vartheta(\mathrm{d}a|s_0)\right] \\
&= \mathbb{E}_{s_0\sim\varrho_\pi(\nu),a\sim\pi}\left[\left(Q^{\pi_\vartheta}(s_0,a) - \tau\log(\pi_\vartheta(\cdot|s))\right)\nabla_\vartheta \log(\pi_\vartheta(a|s_0))\right] \quad \text{(A.1)}
\end{aligned}
$$

where $\varrho_\pi(\nu) = \mathbb{E}_{s_0\sim\nu}\left[\sum_t \gamma^t P_\pi(s_t \in \mathrm{d}s|s_0)\right]$ is the discounted visitation measure under policy π for the initial state distribution ν.

We aim to show that under the above dynamics some energy in the parameter space decreases. We have a few options for such energy:

(1) Defined directly on the parameter space of interest,

$$U(\vartheta) = \|\vartheta - \vartheta^*\|^2$$

(2) on the space of measures,

$$U(\pi_\vartheta) = \int D_{KL}\left(\pi_\vartheta(\cdot,s)\|\pi^*(\cdot,s)\right)\nu(\mathrm{d}s) = \int (F_\vartheta(s) - F_{\vartheta^*}(s))^2\nu(\mathrm{d}s)$$

(3) or a Lojasiewicz-type inequality (*i.e.*, gradient domination).

$$\|\nabla_\vartheta \mathbb{E}_{s_0\sim\nu}\left[V^{\pi_\vartheta}(s_0)\right]\| \geq \mathbb{E}_{s_0\sim\nu}\left[V^{\pi_\vartheta}(s_0) - V^{\pi_{\vartheta^*}}(s_0)\right]$$

Here we go through some computations for two of the possible approaches described above.

(1) Defined directly on the parameter space of interest:

$$U(\vartheta) = \|\vartheta - \vartheta^*\|^2$$

The variation $U(\vartheta)$ under the dynamics (A.1) reads:

$$
\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}U(\vartheta_t) &= \langle \vartheta - \vartheta^*, \frac{\mathrm{d}}{\mathrm{d}t}\vartheta\rangle \\
&= -\langle \vartheta - \vartheta^*, \mathbb{E}_{s_0\sim\varrho_\pi(\nu),a\sim\pi}\left[\left(Q^{\pi_\vartheta}(s_0,a) - \tau\log(\pi_\vartheta(\cdot|s))\right)\nabla_\vartheta \log(\pi_\vartheta(a|s_0))\right]\rangle \\
&= -\langle \vartheta - \vartheta^*, \mathbb{E}_{s_0\sim\varrho_\pi(\nu),a\sim\pi}\left[\left(Q^{\pi_\vartheta}(s_0,a) - \tau\log(\pi_\vartheta(\cdot|s))\right)\nabla_\vartheta \left(-\frac{(a-F_\vartheta(s))^2}{2\sigma^2}\right)\right]\rangle \\
&= \langle \vartheta - \vartheta^*, \mathbb{E}_{s_0\sim\varrho_\pi(\nu),a\sim\pi}\left[\left(Q^{\pi_\vartheta}(s_0,a) - \tau\log(\pi_\vartheta(\cdot|s))\right)\left(-\frac{(a-F_\vartheta(s))}{\sigma^2}\nabla_\vartheta F_\vartheta(s)\right)\right]\rangle \\
&= \mathbb{E}_{s_0\sim\varrho_\pi(\nu),a\sim\pi}\left[\left(Q^{\pi_\vartheta}(s_0,a) - \tau\log(\pi_\vartheta(\cdot|s))\right)\left(-\frac{(a-F_\vartheta(s))}{\sigma^2}(F_\vartheta(s) - F_{\vartheta^*}(s))\right)\right] \\
&= -\sigma^{-2}\mathbb{E}_{s_0\sim\varrho_\pi(\nu)}\left[(F_\vartheta(s) - F_{\vartheta^*}(s))\mathbb{E}_{a\sim\pi}\left[Q^{\pi_\vartheta}(s_0,a)(a - F_\vartheta(s))\right]\right] \quad \text{(A.2)} \\
&= \sigma^{-2}\mathbb{E}_{s_0\sim\varrho_\pi(\nu)}\left[(F_\vartheta(s) - F_{\vartheta^*}(s))\mathbb{E}_{a\sim\pi}\left[\left(V^{\pi_\vartheta}(s_0) - Q^{\pi_\vartheta}(s_0,a)\right)(a - F_\vartheta(s))\right]\right]
\end{aligned}
$$

where in (A.2) we have used that odd central moments of a Gaussian are all equal to 0.

(2) Defined on the space of measures:

$$U(\pi_\vartheta) = \int D_{KL}\left(\pi_\vartheta(\cdot,s)\|\pi^*(\cdot,s)\right)\nu(\mathrm{d}s) = \int (F_\vartheta(s) - F_{\vartheta^*}(s))^2\nu(\mathrm{d}s)$$

We have, similarly to above,

$$
\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}U(\pi_t) &= \int (F_\vartheta(s) - F_{\vartheta^*}(s))\nabla_\vartheta F_\vartheta(s)\nu(\mathrm{d}s)\frac{\mathrm{d}}{\mathrm{d}t}\vartheta \\
&= -\mathbb{E}_{s_0\sim\varrho_\pi(\nu),s\sim\nu,a\sim\pi}\left[(F_\vartheta(s) - F_{\vartheta^*}(s))\langle\nabla_\vartheta F_\vartheta(s), \nabla_\vartheta \log(\pi_\vartheta(a|s_0))\rangle\left(Q^{\pi_\vartheta}(s_0,a) - \tau\log(\pi_\vartheta(\cdot|s))\right)\right] \\
&= \sigma^{-2}\mathbb{E}_{s_0\sim\varrho_\pi(\nu),s\sim\nu,a\sim\pi}\left[(F_\vartheta(s) - F_{\vartheta^*}(s))\langle\nabla_\vartheta F_\vartheta(s), \nabla_\vartheta F_\vartheta(s_0)\rangle(F_\vartheta(s_0) - a)\left(Q^{\pi_\vartheta}(s_0,a) - \tau\log(\pi_\vartheta(\cdot|s))\right)\right]
\end{aligned}
$$

Assuming that the operator $\langle \nabla_\vartheta F_\vartheta(s), \nabla_\vartheta F_\vartheta(s_0) \rangle$ is positive definite with respect to the L^2 product, we then have

$$\frac{\mathrm{d}}{\mathrm{d}t} U(\pi_t) < \lambda_{min} \sigma^{-2} \mathbb{E}_{s_0 \sim \varrho_\pi(\nu), a \sim \pi} \left[(F_\vartheta(s_0) - F_{\vartheta^*}(s_0))(F_\vartheta(s_0) - a) \left(Q^{\pi_\vartheta}(s_0, a) - \tau \log(\pi_\vartheta(\cdot|s)) \right) \right]$$

which corresponds to that we have obtained in (1).