Robust Reinforcement Learning: A Case Study in Linear Quadratic Regulation

Bo Pang, Zhong-Ping Jiang
Department of Electrical and Computer Engineering, New York University, Six Metrotech Center, Brooklyn, NY 11201
{bo.pang, zjiang}@nyu.edu

Abstract

This paper studies the robustness of reinforcement learning algorithms to errors in the learning process. Specifically, we revisit the benchmark problem of discrete-time linear quadratic regulation (LQR) and study the long-standing open question: Under what conditions is the policy iteration method robustly stable from a dynamical systems perspective? Using advanced stability results in control theory, it is shown that policy iteration for LQR is inherently robust to small errors in the learning process and enjoys small-disturbance input-to-state stability: whenever the error in each iteration is bounded and small, the solutions of the policy iteration algorithm are also bounded, and, moreover, enter and stay in a small neighbourhood of the optimal LQR solution. As an application, a novel off-policy optimistic least-squares policy iteration for the LQR problem is proposed, when the system dynamics are subjected to additive stochastic disturbances. The proposed new results in robust reinforcement learning are validated by a numerical example.

Introduction

As an important and popular method in reinforcement learning (RL), policy iteration has been widely studied by researchers and utilized in different kinds of real-life applications by practitioners (Bertsekas 1995; Sutton and Barto 2018). Policy iteration involves two steps, *policy evaluation* and *policy improvement*. In policy evaluation, a given policy is evaluated based on a scalar performance index. Then this performance index is utilized to generate a new control policy in policy improvement. These two steps are iterated in turn, to find the solution of the RL problem at hand. When all the information involved in this process is exactly known, the convergence to the optimal solution can be provably guaranteed, by exploiting the monotonicity property of the policy improvement step. That is, the performance of the newly generated policy is no worse than that of the given policy in each iteration. Over the past decades, various versions of policy iteration have been proposed, for diverse optimal control problems, see (Bertsekas 1995; Sutton and Barto 2018; Lewis, Vrabie, and Syrmos 2012; Jiang and Jiang 2017; Jiang, Bian, and Gao 2020) and the references therein.

In reality, policy evaluation or policy improvement can hardly be implemented precisely, because of the existence of various errors, which may be induced by function approximation, state estimation, sensor noise, external disturbance and so on. Therefore, a natural question to ask is: when is a policy iteration algorithm robust to errors in the learning process? In other words, under what conditions on the errors, does the policy iteration still converge to (a neighbourhood of) the optimal solution? And how to quantify the size of this neighbourhood? In spite of the popularity and empirical successes of policy iteration, its robustness issue has not been fully understood yet in theory, due to the inherent nonlinearity of the process (Bertsekas 2011). The problem becomes more complex when the state and action spaces are unbounded and continuous, which are common in RL problems of physical systems such as robotics and autonomous cars (Lillicrap et al. 2016). Indeed, in this case the stability issue needs to be addressed, to avoid the selection of destabilizing policies that drive the states of the closed-loop system into the infinity or an unsafe region.

In this paper, we investigate the robustness of policy iteration for the discrete-time linear quadratic regulator (LQR) problem, which was firstly proposed in (Hewer 1971). Even if the LQR is the most basic and important optimal control problem with unbounded, continuous state and action spaces (Bertsekas 1995), the robustness of its associated policy iteration to errors in the learning process has not been fully investigated. The main idea of this paper is to regard the policy iteration as a dynamical system, and then utilize the concepts of exponential stability and input-to-state stability in control theory to analyze its robustness (Sontag 2008). To be more specific, we firstly prove that the optimal LQR solution is a locally exponentially stable equilibrium of the exact policy iteration (see Lemma 1). Then based on this observation, we show that the policy iteration with errors is locally input-to-state stable, if the errors are regarded as the disturbance input (see Lemma 2). That is, if the policy iteration starts from an initial solution close to the optimal solution, and the errors are small and bounded, the discrepancies between the solutions generated by the policy iteration and the optimal solution will also be small and bounded. Thirdly, we demonstrate that for any initial stabilizing control gain, as long as the errors are small, the approximate solution given by policy iteration will eventually enter a small neigh-

bourhood of the optimal solution (see Theorem 2). Finally, a novel off-policy model-free RL algorithm, named optimistic least-squares policy iteration (O-LSPI), is proposed for the LQR problem with dynamics perturbed by additive stochastic disturbances. Our robustness result is applied to show the convergence of this off-policy O-LSPI (see Theorem 3). Experiments on a numerical example validate our results.

Our main contributions are two-fold. First, we provide a control-theoretic robustness analysis for the policy iteration of discrete-time LQR. Second, we propose a novel off-policy RL algorithm O-LSPI with provable convergence.

In the rest of this paper, we first present some preliminaries, followed by the robustness analysis and the off-policy O-LSPI. Then we present the experimental results, discuss some related work, and close the paper with some concluding remarks.

Notations

$\mathbb{R}$ ($\mathbb{R}_+$) is the set of all real (nonnegative) numbers; $\mathbb{Z}_+$ denotes the set of nonnegative integers; $\mathbb{S}^n$ is the set of all real symmetric matrices of order n; $\otimes$ denotes the Kronecker product; I_n denotes the identity matrix with dimension n; $\|\cdot\|_F$ is the Frobenius norm; $\|\cdot\|_2$ is the 2-norm for vectors and the induced 2-norm for matrices; for signal $Z:\mathbb{F}\to\mathbb{R}^{n\times m}$, $\|Z\|_\infty$ denotes its l^∞-norm when $\mathbb{F}=\mathbb{Z}_+$, and L^∞-norm when $\mathbb{F}=\mathbb{R}_+$. For matrices $X\in\mathbb{R}^{m\times n}$, $Y\in\mathbb{S}^m$, and vector $v\in\mathbb{R}^n$, define

$$\begin{aligned}\mathrm{vec}(X)&=[\ X_1^T\quad X_2^T\quad\cdots\quad X_n^T\ ]^T,\ \tilde{v}=\mathrm{svec}(vv^T),\\ \mathrm{svec}(Y)&=[y_{11},\sqrt{2}y_{12},\cdots,\sqrt{2}y_{1m},y_{22},\sqrt{2}y_{23},\\ &\quad\cdots,\sqrt{2}y_{m-1,m},y_{m,m}]^T\in\mathbb{R}^{\frac{1}{2}m(m+1)},\end{aligned}$$

where X_i is the ith column of X. For $Z\in\mathbb{R}^{m\times n}$, define $\mathcal{B}_r(Z)=\{X\in\mathbb{R}^{m\times n}|\|X-Z\|_F<r\}$ and $\bar{\mathcal{B}}_r(Z)$ as the closure of $\mathcal{B}_r(Z)$. $Z^\dagger$ is the Moore-Penrose inverse of matrix Z. $\mathrm{blkdiag}(Z_1,Z_2,\cdots,Z_N)$ refers to the block-diagonal matrix that consists of a set of matrices $Z_1,Z_2,\cdots,Z_N$.

Preliminaries

Consider linear time-invariant systems of the form

$$x_{k+1}=Ax_k+Bu_k,\quad x_0=x_{ini}\tag{1}$$

where $x_k\in\mathbb{R}^n$ is the system state, $u_k\in\mathbb{R}^m$ is the control input, $x_{ini}\in\mathbb{R}^n$ is the initial condition, $A\in\mathbb{R}^{n\times n}$ and $B\in\mathbb{R}^{n\times m}$. (A,B) is controllable, that is, $[B,AB,A^2B,\cdots,A^{n-1}B]$ has full row rank. The classic LQR problem is to find a controller u in order to minimize the following cost functional

$$J(x_0,u)=\sum_{k=0}^{\infty}c(x_k,u_k),\tag{2}$$

where $c(x_k,u_k)=x_k^TSx_k+u_k^TRu_k$, $S\in\mathbb{S}^n$ is positive semidefinite and $R\in\mathbb{S}^n$ is positive definite. $(A,S^{1/2})$ is observable, that is, $(A^T,S^{1/2})$ is controllable. It is well-known that in such a setting, the LQR problem admits a unique optimal controller $u^*=-K^*x$, where

$$K^*=(R+B^TP^*B)^{-1}B^TP^*A\tag{3}$$

with $P^*\in\mathbb{S}^n$ the unique positive definite solution of the algebraic Riccati equation (ARE)

$$A^TPA-P-A^TPB(R+B^TPB)^{-1}B^TPA+S=0.\tag{4}$$

In addition, $A-BK^*$ is stable, i.e., the spectral radius $\rho(A-BK^*)<1$. See (Lewis, Vrabie, and Syrmos 2012, Section 2.4) for details. For convenience, a control gain $K\in\mathbb{R}^{m\times n}$ is said to be stabilizing if $A-BK$ is stable.

Policy Iteration for LQR

For any stabilizing control gain $K\in\mathbb{R}^{m\times n}$, the cost (2) with $u_k=-Kx_k$ is a quadratic function of the initial state (Lewis, Vrabie, and Syrmos 2012, Section 2.4). Specifically, $J(x_0,-Kx)=x_0^TP_Kx_0$, where $P_K\in\mathbb{S}^n$ is the unique positive definite solution of the Lyapunov equation

$$(A-BK)^TP_K(A-BK)-P_K+S+K^TRK=0.\tag{5}$$

Define function

$$\begin{aligned}G(P_K)&=\left[\begin{array}{c|c}[G(P_K)]_{xx}&[G(P_K)]_{ux}^T\\\hline[G(P_K)]_{ux}&[G(P_K)]_{uu}\end{array}\right]\\&\triangleq\begin{bmatrix}S+A^TP_KA-P_K&A^TP_KB\\B^TP_KA&R+B^TP_KB\end{bmatrix}.\end{aligned}$$

Then (5) can be rewritten as

$$\mathcal{H}(G(P_K),K)=0,$$

where

$$\mathcal{H}(G(P_K),K)\triangleq\begin{bmatrix}I_n&-K^T\end{bmatrix}G(P_K)\begin{bmatrix}I_n\\-K\end{bmatrix}.$$

The policy iteration for LQR is presented below, which is an equivalent reformulation of the original results in (Hewer 1971).

Procedure 1 (Exact Policy Iteration)**.**

1) *Choose a stabilizing control gain K_1, and let $i=1$.*
2) *(Policy evaluation) Evaluate the performance of control gain K_i, by solving*

$$\mathcal{H}(G_i,K_i)=0\tag{6}$$

for $P_i\in\mathbb{S}^n$, where $G_i\triangleq G(P_i)$.

3) *(Policy improvement) Obtain an improved policy*

$$K_{i+1}=[G_i]_{uu}^{-1}[G_i]_{ux}.\tag{7}$$

4) *Set $i\leftarrow i+1$ and go back to Step 2).*

The following convergence results of Procedure 1 were also provided in (Hewer 1971).

Theorem 1. *In Procedure 1 we have:*

i) $A-BK_i$ is stable for all $i=1,2,\cdots$.
ii) $P_1\geq P_2\geq P_3\geq\cdots\geq P^*$.
iii) $\lim_{i\to\infty}P_i=P^*$, $\lim_{i\to\infty}K_i=K^*$.

Problem Formulation

In Procedure 1, the exact knowledge of A and B is required, as the solution to (6) relies upon A and B. So, the exact policy iteration is model-based. However, in practice, very often we only have access to incomplete information required to solve the problem. In other words, each policy evaluation step will result in inaccurate estimation. Thus we are interested in studying the following problem.

Problem 1. *If G_i is replaced by an approximated matrix $\hat{G}_i$, will the conclusions in Theorem 1 still hold?*

The difference between $\hat{G}_i$ and G_i can be attributed to errors from various sources. One example comes from the problem of using reinforcement learning method to find the optimal solutions for LQR when (1) is subjected to additive external disturbances. Concretely, consider system (1) perturbed by external noise

$$x_{k+1} = Ax_k + Bu_k + Cw_k, \quad x_0 = x_{ini} \tag{8}$$

where $C \in \mathbb{R}^{n\times q}$, $w_k \in \mathbb{R}^q$ is drawn i.i.d. from the standard Gaussian distribution $\mathcal{N}(0, I_q)$, and matrices A, B and C are unknown. Since the information about system matrices is unavailable, we need to implement the policy evaluation using input/state data. Due to the existence of unmeasurable stochastic noise w_k, generally we could only obtain an estimation $\hat{G}_i$ of the true G_i from the noise-corrupted input/state data. Other sources that cause the difference between $\hat{G}_i$ and G_i include but are not limited to: the estimation errors of A and B in indirect adaptive control, system identification and model-based reinforcement learning (Åström and Wittenmark 1995; Ljung 1999; Tu and Recht 2019); the residual caused by an early termination of the iteration to numerically solve ARE (4), in order to save computational efforts (Hylla 2011); approximate values of S and R in inverse optimal control/imitation learning, due to the absence of exact knowledge of the cost function (Levine and Koltun 2012; Monfort, Liu, and Ziebart 2015).

In this work, using the concept of exponential stability and input-to-state stability in control theory, we provide an answer to Problem 1. Moreover, we provide the convergence analysis of the novel O-LSPI when it is applied to solve the LQR problem for uncertain systems (8).

Notions of Exponential and Input-to-State Stability

Consider a dynamical system of the general form

$$x_{k+1} = f(x_k, u_k), \quad x_0 = x_{ini}, \tag{9}$$

where $x_k \in \mathbb{R}^n$, $u_k \in \mathbb{R}^m$, $f : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n$ is continuous, and x^* is an equilibrium of $x_{k+1} = f(x_k, 0)$ when $u_k = 0$ for all $k \in \mathbb{Z}_+$. The concepts of exponential and input-to-state stability for (9) are recalled in this subsection. See (Jiang, Lin, and Wang 2004) for more details.

Definition 1. *For (9) with $u_k = 0$ for all $k \in \mathbb{Z}_+$, x^* is a locally exponentially stable equilibrium if there exists a $\delta > 0$, such that for some $a > 0$ and $0 < b < 1$,*

$$\|x_k - x^*\|_2 \le ab^k \|x_{ini} - x^*\|_2$$

for all $x_{ini} \in \mathcal{B}_\delta(x^*)$. If $\delta = +\infty$, then x^* is a globally exponentially stable equilibrium.

The exponential stability implies not only the convergence, but also the convergence rate of (9). When the input signal is not zero, the input-to-state stability characterizes how the solution of (9) is affected by the input signal.

Definition 2. *A function $\gamma : \mathbb{R}_+ \to \mathbb{R}_+$ is said to be of class $\mathcal{K}$ if it is continuous, strictly increasing and vanishes at the origin. A function $\beta : \mathbb{R}_+ \times \mathbb{R}_+ \to \mathbb{R}_+$ is said to be of class $\mathcal{KL}$ if $\beta(\cdot, t)$ is of class $\mathcal{K}$ for every fixed $t \in \mathbb{R}_+$ and, for every fixed $r \ge 0$, $\beta(r, t)$ decreases to 0 as $t \to \infty$.*

Definition 3. *System (9) is locally input-to-state stable if there exist some $\alpha_1 > 0$, some $\alpha_2 > 0$, some $\beta \in \mathcal{KL}$ and some $\gamma \in \mathcal{K}$, such that for each u and each x_{ini} satisfying $x_{ini} \in \mathcal{B}_{\alpha_1}(x^*)$, $\|u\|_\infty < \alpha_2$, the corresponding solution x_k satisfies*

$$\|x_k - x^*\|_2 \le \beta(\|x_{ini} - x^*\|_2, k) + \gamma(\|u\|_\infty).$$

Literally speaking, the local input-to-state stability implies that the distance from the state to the equilibrium is bounded if the input signal is small and the initial state is close to the equilibrium. In addition, the effect of the initial condition vanishes as time goes to infinity.

Robustness Analysis of Policy Iteration

Consider the policy iteration in the presence of errors.

Procedure 2 (Inexact Policy Iteration)**.**

1) *Choose a stabilizing control gain $\hat{K}_1$, and let $i = 1$.*
2) *(Inexact policy evaluation) Obtain $\hat{G}_i = \tilde{G}_i + \Delta G_i$, where $\Delta G_i \in \mathbb{S}^{m+n}$ is a disturbance, $\tilde{G}_i \triangleq G(\tilde{P}_i)$ and $\tilde{P}_i \in \mathbb{S}^n$ satisfy*

$$\mathcal{H}(\tilde{G}_i, \hat{K}_i) = 0, \tag{10}$$

and $J(x_0, -\hat{K}_i x) = x_0^T \tilde{P}_i x_0$ is the true cost induced by control gain $\hat{K}_i$, if $\hat{K}_i$ is stabilizing.
3) *(Policy update) Construct a new control gain*

$$\hat{K}_{i+1} = [\hat{G}_i]_{uu}^{-1}[\hat{G}_i]_{ux}. \tag{11}$$

4) *Set $i \leftarrow i + 1$ and go back to Step 2).*

We firstly show that the exact policy iteration Procedure 1, viewed as a dynamical system, is locally exponentially stable at P^*. Then based on this result, we show that the inexact policy iteration, viewed as a dynamical system with ΔG_i as the input, is locally input-to-state stable.

For $X \in \mathbb{R}^{n\times n}$, $Y \in \mathbb{R}^{n\times n}$, define

$$\mathscr{A}(X) = X^T \otimes X^T - I_n \otimes I_n,\ \mathcal{L}_X(Y) = X^T Y X - Y,$$
$$\mathscr{K}(Y) = \mathscr{R}^{-1}(Y) B^T Y A,$$
$$\mathscr{R}(Y) = R + B^T Y B, \quad \mathcal{A}(\mathscr{K}(Y)) = A - B\mathscr{K}(Y).$$

Then obviously

$$\operatorname{vec}(\mathcal{L}_X(Y)) = \mathscr{A}(X)\operatorname{vec}(Y). \tag{12}$$

If X is stable, then $\mathscr{A}(X)$ is invertible, by (12) the inverse operator $\mathcal{L}_X^{-1}(\cdot)$ exists on $\mathbb{R}^{n\times n}$.

In Procedure 1, suppose $K_1 = \mathscr{K}(P_0)$, where $P_0 \in \mathbb{S}^n$ is chosen such that K_1 is stabilizing. Such a P_0 always exists. For example, since K^* is stabilizing, one can choose P_0 close to P^* by continuity. Then from (6) and (7), the sequence $\{P_i\}_{i=0}^{\infty}$ generated by Procedure 1 satisfies

$$P_{i+1} = \mathcal{L}^{-1}_{\mathcal{A}(\mathscr{K}(P_i))}\left(-S - \mathscr{K}(P_i)^T R \mathscr{K}(P_i)\right). \quad (13)$$

If P_i is regarded as the state, and the iteration index i is regarded as time, then (13) is a discrete-time dynamical system and P^* is an equilibrium by Theorem 1. The next lemma shows that P^* is actually a locally exponentially stable equilibrium, whose proof is given in Appendix B.

Lemma 1. *For any $\sigma < 1$, there exists a $\delta_0(\sigma) > 0$, such that for any $P_i \in \mathcal{B}_{\delta_0}(P^*)$, $\mathscr{R}(P_i)$ is invertible, $\mathcal{A}(\mathscr{K}(P_i))$ is stable and $\|P_{i+1} - P^*\|_F \le \sigma \|P_i - P^*\|_F$.*

Lemma 1 is inspired by (Hewer 1971, Theorem 2), which states that Procedure 1 has the rate of convergence

$$\|P_{i+1} - P^*\|_F \le c_0 \|P_i - P^*\|_F^2. \quad (14)$$

for any $P_i \ge P^*$, and some $c_0 > 0$. Notice that Lemma 1 does not have the requirement $P_i \ge P^*$.

In Procedure 2, suppose $\hat{K}_1 = \mathscr{K}(\tilde{P}_0)$ and $\Delta G_0 = 0$, where $\tilde{P}_0 \in \mathbb{S}^n$ is chosen such that $\hat{K}_1$ is stabilizing. If $\hat{K}_i$ is stabilizing and $[\hat{G}_i]_{uu}$ is invertible for all $i \in \mathbb{Z}_+$, $i > 0$ (this is possible under certain conditions, see Appendix C), the sequence $\{\tilde{P}_i\}_{i=0}^{\infty}$ generated by Procedure 2 satisfies

$$\begin{aligned}\tilde{P}_{i+1} = {} & \mathcal{L}^{-1}_{\mathcal{A}(\mathscr{K}(\tilde{P}_i))}\left(-S - \mathscr{K}(\tilde{P}_i)^T R \mathscr{K}(\tilde{P}_i)\right) \\ & + \mathcal{E}(\tilde{G}_i, \Delta G_i),\end{aligned} \quad (15)$$

where

$$\begin{aligned}\mathcal{E}(\tilde{G}_i, \Delta G_i) = {} & \mathcal{L}^{-1}_{\mathcal{A}(\hat{K}_{i+1})}\left(-S - \hat{K}_{i+1}^T R \hat{K}_{i+1}\right) \\ & - \mathcal{L}^{-1}_{\mathcal{A}(\mathscr{K}(\tilde{P}_i))}\left(-S - \mathscr{K}(\tilde{P}_i)^T R \mathscr{K}(\tilde{P}_i)\right).\end{aligned}$$

Here, the dependence of $\mathcal{E}$ on $\tilde{G}_i$ and ΔG_i comes from (11). Regarding $\{\Delta G_i\}_{i=0}^{\infty}$ as the disturbance input, the next lemma shows that dynamical system (15) is locally input-to-state stable, whose proof can be found in Appendix C.

Lemma 2. *For σ and its associated δ_0 in Lemma 1, there exists $\delta_1(\delta_0) > 0$, such that if $\|\Delta G\|_\infty < \delta_1$, $\tilde{P}_0 \in \mathcal{B}_{\delta_0}(P^*)$,*

(i) $[\hat{G}_i]_{uu}$ is invertible, $\hat{K}_i$ is stabilizing, $\forall i \in \mathbb{Z}_+$, $i > 0$;

(ii) (15) is locally input-to-state stable (see Definition 3):

$$\|\tilde{P}_i - P^*\|_F \le \beta(\|\tilde{P}_0 - P^*\|_F, i) + \gamma(\|\Delta G\|_\infty),$$

for all $i \in \mathbb{Z}_+$, where $\beta(y, i) = \sigma^i y$, $\gamma(y) = c_3 y/(1-\sigma)$, $y \in \mathbb{R}$ and $c_3(\delta_0) > 0$.

(iii) $\|\hat{K}_i\|_F < \kappa_1$ for some $\kappa_1 \in \mathbb{R}_+$, $\forall i \in \mathbb{Z}_+$, $i > 0$;

(iv) $\lim_{i\to\infty} \|\Delta G_i\|_F = 0$ implies $\lim_{i\to\infty} \|\tilde{P}_i - P^*\|_F = 0$.

To prove Lemma 2, we firstly prove that with the given conditions, by continuity $[\hat{G}_i]_{uu}$ is invertible, $\hat{K}_i$ is stabilizing and $\|\mathcal{E}(\tilde{G}_i, \Delta G_i)\|_F \le c_3 \|\Delta G_i\|_F$. Then by Lemma 1 and (15), if $\tilde{P}_i \in \mathcal{B}_{\delta_0}(P^*)$, δ_1 can be chosen small enough so that

$$\begin{aligned}\|\tilde{P}_{i+1} - P^*\|_F &\le \sigma \|\tilde{P}_i - P^*\|_F + c_3 \|\Delta G\|_\infty \\ &< \sigma\delta_0 + c_3\delta_1 < \delta_0.\end{aligned} \quad (16)$$

By mathematical induction, unrolling (16) completes the proof. In the unrolling process, the coefficient $0 < \sigma < 1$ in the exponential stability of Lemma 1 prevents the accumulated effects of disturbance ΔG_i from driving $\|\tilde{P}_{i+1} - P^*\|_F$ to the infinity.

Intuitively, Lemma 2 implies that in Procedure 2, if $\tilde{P}_0$ is near P^* (thus $\hat{K}_1$ is near K^*), and the disturbance input ΔG is bounded and not too large, then the cost of the generated control policy $\hat{K}_i$ is also bounded, and will ultimately be no larger than a constant proportional to the l^∞-norm of the disturbance. The smaller the disturbance is, the better the ultimately generated policy is. In other words, the algorithm described in Procedure 2 is not sensitive to small disturbances when the initial condition is in a neighbourhood of the optimal solution.

The requirement that the initial condition $\tilde{P}_0$ needs to be in a neighbourhood of P^* in Lemma 2 can be removed, as stated in the following theorem whose proof is given in the Appendix D.

Theorem 2. *For any given stabilizing control gain $\hat{K}_1$ and any $\epsilon > 0$, if $S > 0$, there exist $\delta_2(\epsilon, \hat{K}_1) > 0$, $\Pi(\delta_2) > 0$, and $\kappa(\delta_2) > 0$, such that for all ΔG satisfying $\|\Delta G\|_\infty < \delta_2$, $[\hat{G}_i]_{uu}$ is invertible, $\hat{K}_i$ is stabilizing, $\|\tilde{P}_i\|_F < \Pi$, $\|\hat{K}_i\|_F < \kappa$, $\forall i \in \mathbb{Z}_+$, $i > 0$ and*

$$\limsup_{i\to\infty} \|\tilde{P}_i - P^*\|_F < \epsilon.$$

If in addition $\lim_{i\to\infty} \|\Delta G_i\|_F = 0$, then $\lim_{i\to\infty} \|\tilde{P}_i - P^*\|_F = 0$.

Here are the essential elements of the proof for Theorem 2: It is firstly proved that given any stabilizing control gain $\hat{K}_1$, there exist $\bar{i} \in \mathbb{Z}_+$, $\bar{i} < +\infty$, and $b_{\bar{i}} > 0$, such that if $\|\Delta G_i\|_F < b_{\bar{i}}$ for $i = 1, 2, \cdots, \bar{i}$, then (1) $[\hat{G}_i]_{uu}$ is invertible, $\hat{K}_i$ is stabilizing and bounded, $\tilde{P}_i$ is bounded, $i = 1, 2, \cdots, \bar{i}$, (2) $\tilde{P}_{\bar{i}}$ enters the neighbourhood of P^*, i.e., $\mathcal{B}_{\delta_0}(P^*)$ defined in Lemma 2. Secondly, an application of Lemma 2 completes the proof.

In Theorem 2, $\hat{K}_1$ can be any stabilizing control gain, which is different from that of Lemma 2. When there is no disturbance, Theorem 2 implies the convergence result of Procedure 1 in (Hewer 1971, Theorem 1) (i.e. Theorem 1 in this paper).

Optimistic Least-Squares Policy Iteration

For system (8), due to the presence of stochastic noise w_k, the cost function (2) will not be finite. Thus alternatively the objective is to find a control law in the form of $u = -Kx$ directly from the input/state data, minimizing the cost function

$$J_{Avg}(u) = \lim_{N\to\infty} \frac{1}{N} \mathop{\mathbb{E}}_{\substack{w_k \\ k=0,1,\cdots}} \left\{ \sum_{k=0}^{N-1} c(x_k, u_k) \right\}, \quad (17)$$

where S and R in $c(x_k, u_k)$ are positive definite. It is well-known (Bertsekas 1995, Section 4.4) that this problem shares the same optimal solutions with the standard LQR for system (1) and cost function (2). Specifically, the optimal control gain is given by (3), and the optimal cost is $J^*_{Avg} = \mathrm{tr}(C^T P^* C)$, with P^* the unique positive definite solution of (4). For any stabilizing gain K, the cost it induces is $J_{Avg}(-Kx) = \mathrm{tr}(C^T P_K C)$, with K and P_K satisfying (5) (or equivalently (6)). Note that the assumption that $w_k \sim \mathcal{N}(0, I_q)$ in (8) is not a restriction, since any random variable $X_1 \sim \mathcal{N}(0, \Sigma)$ with $\Sigma \in \mathbb{S}^q$ positive semidefinite, can be represented by $X_1 = DX_2$, where $\Sigma = D^T D$, $D \in \mathbb{R}^{q\times q}$, and $X_2 \sim \mathcal{N}(0, I_q)$. Then D is absorbed into C in (8).

The optimistic least-squares policy iteration (O-LSPI) is based on the following observation: for a stabilizing gain K, its associated P_K is the stable equilibrium of linear dynamical system

$$P_{K,j+1} = \mathcal{H}(Q(P_{K,j}), K), \quad P_{K,0} \in \mathbb{S}^n, \tag{18}$$

where

$$\begin{aligned} Q(P_{K,j}) &= \left[\begin{array}{c|c} [Q(P_{K,j})]_{xx} & [Q(P_{K,j})]_{ux}^T \\ \hline [Q(P_{K,j})]_{ux} & [Q(P_{K,j})]_{uu} \end{array}\right] \\ &\triangleq \left[\begin{array}{cc} S + A^T P_{K,j} A & A^T P_{K,j} B \\ B^T P_{K,j} A & R + B^T P_{K,j} B \end{array}\right]. \end{aligned} \tag{19}$$

This fact can be easily verified by rewriting and vectorizing (18) into its equivalent form

$$\begin{aligned} p_{K,j+1} = &\left((A - BK)^T \otimes (A - BK)^T\right) p_{K,j} \\ &+ \mathrm{vec}(S + K^T R K), \quad p_{K,0} \in \mathbb{R}^{n^2}, \end{aligned} \tag{20}$$

where $p_{K,j} = \mathrm{vec}(P_{K,j})$. Since $(A - BK)$ is stable, $(A - BK)^T \otimes (A - BK)^T$ is also stable. Thus (20) admits a unique stable equilibrium. So does (18) and the unique solution must be P_K because

$$\begin{aligned} Q(P_{K,j}) &= G(P_{K,j}) + \mathrm{blkdiag}(P_{K,j}, 0), \\ \mathcal{H}(G(P_{K,j}), K) &= \mathcal{H}(Q(P_{K,j}), K) - P_{K,j}. \end{aligned} \tag{21}$$

This implies that instead of solving (6), we may utilize iteration (18) to achieve policy evaluation. It is not hard to recognize that (18) is actually the LQR version of the optimistic policy iteration in (Tsitsiklis 2002; Bertsekas 2011) for problems with discrete state and action spaces (thus the name "optimistic" in O-LSPI). Suppose a behavior policy (the policy used to generate data is called the behavior policy, see (Sutton and Barto 2018)) $u_k = -\hat{K}_1 x_k + v_k$ is applied to the system to collect data, where $\hat{K}_1$ is stabilizing and v_k is drawn i.i.d. from Gaussian distribution $\mathcal{N}(0, \sigma_u^2 I_m)$, $\sigma_u \in \mathbb{R}_+$. Then the state-control pair $[x^T, u^T]^T$ admits a unique invariant distribution π. We make the following assumption.

Assumption 1. *$\mathbb{E}_\pi\left[\tilde{z}\tilde{z}^T\right]$ is invertible, where $z = [x^T, u^T, 1]^T$.*

For any $P \in \mathbb{S}^n$, we have

$$\mathbb{E}\left[z_k^T F(P) z_k - x_{k+1}^T P x_{k+1} | x_k, u_k\right] = c(x_k, u_k)$$

where $F(P) = \mathrm{blkdiag}(Q(P), \mathrm{tr}(C^T P C))$. Vectorizing and multiplying the above equation by $\tilde{z}_k$ yields

$$\begin{aligned} &\mathbb{E}\left[\tilde{z}_k \tilde{x}_{k+1}^T | x_k, u_k\right] \mathrm{svec}(P) \\ &= \mathbb{E}\left[\tilde{z}_k \tilde{z}_k^T | x_k, u_k\right] \mathrm{svec}(F(P)) - \tilde{z}_k c(x_k, u_k). \end{aligned}$$

Taking expectation with respect to the invariant distribution π, by Assumption 1 we obtain

$$\mathrm{svec}(F(P)) = \varphi_1^{-1}\left(\varphi_2 \,\mathrm{svec}(P) + \varphi_3\right), \tag{22}$$

where $\varphi_1 = \mathbb{E}_\pi\left[\tilde{z}_k \tilde{z}_k^T\right]$, $\varphi_2 = \mathbb{E}_\pi\left[\tilde{z}_k \tilde{x}_{k+1}^T\right]$, and $\varphi_3 = \mathbb{E}_\pi\left[\tilde{z}_k c(x_k, u_k)\right]$. For known P, $F(P)$ can be estimated using least squares from the collected data

$$\mathrm{svec}(\hat{F}(P)) = \Phi_M^\dagger \Psi_M \,\mathrm{svec}(P) + \Phi_M^\dagger \Xi_M,$$

where $M \in \mathbb{Z}_+$, $M > 0$ and

$$\Phi_M = \frac{1}{M}\sum_{k=0}^{M-1} \tilde{z}_k \tilde{z}_k^T, \; \Psi_M = \frac{1}{M}\sum_{k=0}^{M-1} \tilde{z}_k \tilde{x}_{k+1}^T,$$

$$\Xi_M = \frac{1}{M}\sum_{k=0}^{M-1} \tilde{z}_k c(x_k, u_k).$$

In this way, (18) can be solved approximately and directly from the data by noticing that $Q(P) = \mathcal{H}(F(P), 0)$. The O-LSPI is presented in Algorithm 1. Note that the same data matrices Φ_M, Ψ_M and Ξ_M are reused for all iterations, thus O-LSPI is off-policy. The convergence of O-LSPI is proved in the following theorem.

Theorem 3. *In Algorithm 1, under Assumption 1, for any initial stabilizing control gain $\hat{K}_1$ and any $\epsilon > 0$, there exist $T_0 \in \mathbb{Z}_+$ and $M_0 \in \mathbb{Z}_+$, such that for any $T \geq T_0$ and $M \geq M_0$, almost surely,*

$$\limsup_{N\to\infty} \|\tilde{P}_N - P^*\|_F < \epsilon$$

and $\hat{K}_i$ is stabilizing for all $i = 1, \cdots, N$, where $\tilde{P}_N$ is the unique solution of (5) *for $\hat{K}_N$.*

The proof of Theorem 3 can be found in Appendix E. Let $J_{Avg}(-\hat{K}_i x) = \mathrm{tr}(C^T \tilde{P}_i C)$ denote the true cost induced by $\hat{K}_i$. By Theorem 2, the task is to prove that there exist $T_0 \in \mathbb{Z}_+$ and $M_0 \in \mathbb{Z}_+$, such that for any $T \geq T_0$ and $M \geq M_0$, almost surely, $\|\Delta G\|_\infty < \delta_2$. For Algorithm 1,

$$\hat{G}_i = \hat{Q}_{i,T} - \mathrm{blkdiag}(\hat{P}_{i,T}, 0).$$

Using (21), we have

$$\begin{aligned} \|\Delta G_i\|_F \leq &\|\hat{Q}_{i,T} - Q(\hat{P}_{i,T})\|_F + \|Q(\hat{P}_{i,T}) - Q(\tilde{P}_i)\|_F \\ &+ \|\hat{P}_{i,T} - \tilde{P}_i\|_F. \end{aligned} \tag{23}$$

Since $\hat{K}_1$ is stabilizing, by the Birkhoff ergodic theorem (Koralov and Sinai 2007, Theorem 16.14), almost surely

$$\begin{aligned} &\lim_{M\to\infty} \Phi_M = \varphi_1, \; \lim_{M\to\infty} \Psi_M = \varphi_2, \\ &\lim_{M\to\infty} \Xi_M = \varphi_3. \end{aligned} \tag{24}$$

Using (24), Assumption 1, (18) and (22), we are able to show that there exist T_0 and M_0, independent of iteration index i, such that for any $T \geq T_0$ and $M \geq M_0$, almost surely every term in (23) is less than $\delta_2/3$. Then Theorem 2 completes the proof.

Algorithm 1: O-LSPI

Input: Initial stabilizing control gain $\hat{K}_1$, Number of policy iterations N, Number of iterations for policy evaluation T, Number of rollout M, Exploration variance σ_u^2.

Collect data with input $u_k = -\hat{K}_1 x_k + v_k$, $v_k \sim \mathcal{N}(0, \sigma_u^2 I_m)$, to construct Φ_M, Ψ_M and Ξ_M;
for $i = 1, \cdots, N-1$ **do**
  $\hat{P}_{i,0} \leftarrow 0$;
  for $j = 0, \cdots, T-1$ **do**
    $\operatorname{svec}(\hat{F}_{i,j}) \leftarrow \Phi_M^\dagger \Psi_M \operatorname{svec}(\hat{P}_{i,j}) + \Phi_M^\dagger \Xi_M$;
    $\hat{Q}_{i,j} \leftarrow \mathcal{H}(\hat{F}_{i,j}, 0)$;
    $\hat{P}_{i,j+1} \leftarrow \mathcal{H}(\hat{Q}_{i,j}, \hat{K}_i)$;
  end
  $\operatorname{svec}(\hat{F}_{i,T}) \leftarrow \Phi_M^\dagger \Psi_M \operatorname{svec}(\hat{P}_{i,T}) + \Phi_M^\dagger \Xi_M$;
  $\hat{Q}_{i,T} \leftarrow \mathcal{H}(\hat{F}_{i,T}, 0)$;
  $\hat{K}_{i+1} \leftarrow [\hat{Q}_{i,T}]_{uu}^{-1} [\hat{Q}_{i,T}]_{ux}$;
end
return $\hat{K}_N$.

Experiments

We apply O-LSPI to the LQR problem studied in (Krauth, Tu, and Recht 2019) with

$$A = \begin{bmatrix} 0.95 & 0.01 & 0 \\ 0.01 & 0.95 & 0.01 \\ 0 & 0.01 & 0.95 \end{bmatrix}, \quad B = \begin{bmatrix} 1 & 0.1 \\ 0 & 0.1 \\ 0 & 0.1 \end{bmatrix},$$
$$C = S = I_3, \quad R = I_2.$$

Here A is stable, so we just choose the initial stabilizing control gain to be $\hat{K}_1 = 0_{2\times 3}$. The exploration variance is set to $\sigma_u^2 = 1$. All the experiments are conducted using MATLAB[1] 2017b, on the New York University High Performance Computing Cluster *Prince* with 4 CPUs and 16GB Memory. Algorithm 1 is implemented with increasing values of parameters N, T and M, until the performance of the resulting control gain (almost) does not improve. This yields $N = 5$, $T = 45$ and $M = 10^6$. To investigate the performance of the algorithm with different values of M and T, we conducted two sets of experiments: (a) Fix $N = 5$ and $T = 45$, and implement Algorithm 1 with increasing values of M from 200 to 10^6; (b) Fix $N = 5$ and $M = 10^6$, and implement Algorithm 1 with increasing values of T from 2 to 45. To evaluate the stability, we run Algorithm 1 for 100 times per set of parameters, and compute the fraction of times it produces stable policies in all phases (left column in Figure 1). To evaluate the optimality, the relative error of the cost function $\operatorname{tr}(C^T(\tilde{P}_N - P^*)C)/\operatorname{tr}(C^T P^* C)$ is calculated. The relative errors of 100 stable implementations of Algorithm 1 are collected (i.e., implementation that yields stabilizing control gains in all phases), based on which the sample average (middle column in Figure 1) and sample variance (right column in Figure 1) of the relative error are plotted.

[1] https://www.mathworks.com/

In Figure 1, as the number of rollout M increases, the fraction of stability becomes one, and both the sample average and sample variance of relative error converge to zero. The fraction of stability is not sensitive to the number of iteration for policy evaluation T. But as T increases, the sample average and sample variance of relative error improve and converge to zeros. These observations are consistent with our Theorem 3, thus are also consistent with our robustness analysis for policy iteration, since Theorem 3 is based on Theorem 2.

For comparison, the off-policy least-squares policy iteration algorithm LSPIv1 in (Krauth, Tu, and Recht 2019) is also implemented, using the same setting with the first set of experiments of various M (upper row in Figure 1). The O-LSPI and LSPIv1 have similar performance for $M \geq 10^4$, while the performance of LSPIv1 is slightly better than that of O-LSPI for $M < 10^4$. This may be explained by the fact that the LSPIv1 in (Krauth, Tu, and Recht 2019) assume knowledge of the matrix C in (8), which is not required in O-LSPI.

Finally, the performance of O-LSPI and LSPIv1 with various choices of exploration variance σ_u^2 has been investigated on the same example (see Appendix F). The performance of both O-LSPI and LSPIv1 is best when the exploration variances are large ($\sigma_u^2 \geq 10$). The performance of both O-LSPI and LSPIv1 deteriorates when the exploration noise variances are medium and small ($\sigma_u^2 < 10$). And O-LSPI performs better than LSPIv1 when the exploration noise variances are small ($\sigma_u^2 \leq 10^{-5}$).

Related Work

The investigation of the robustness of policy iteration for problems with continuous state/control spaces is available in previous literature. In (Bertsekas 1995, Proposition 3.6), for discounted optimal control problems of discrete-time systems, it is reported that

$$\limsup_{i\to\infty} \max_{x\in\mathbb{X}\subset\mathbb{R}^n} \left(J_{\mu^i}(x) - J^*(x)\right) \leq \frac{\epsilon + 2\alpha\delta}{(1-\alpha)^2}, \qquad (25)$$

where μ^i is the policy generated in ith iteration, δ and ϵ are the upper bounds of the errors in policy evaluation and policy improvement respectively, $0 < \alpha < 1$ is the discount factor. Bound (25) and our bound in Theorem 2 have the similar styles. However, in our setting the discount factor is $\alpha = 1$ so our bound cannot be implied by (25). Utilizing the fact that Riccati operator is contractive in Thompson's part metric (Thompson 1963), it is shown in (Krauth, Tu, and Recht 2019, Appendix B) that the convergence to the optimal solutions is still achieved in Thompson's part metric, if the errors converge to zero. But it is unclear if this result could imply Theorem 2 in this paper. Sufficient conditions on the errors are given in (Hylla 2011, Chapter 2) and (Boussios 1998, Chapter 2) for continuous-time linear and nonlinear system dynamics respectively, to guarantee that the newly generated control policy is stabilizing and improved. The robustness analysis in this paper is parallel to that in (Pang, Bian, and Jiang 2020). However, since we are dealing with discrete-time systems here, the derivations and proofs are inevitably distinct.

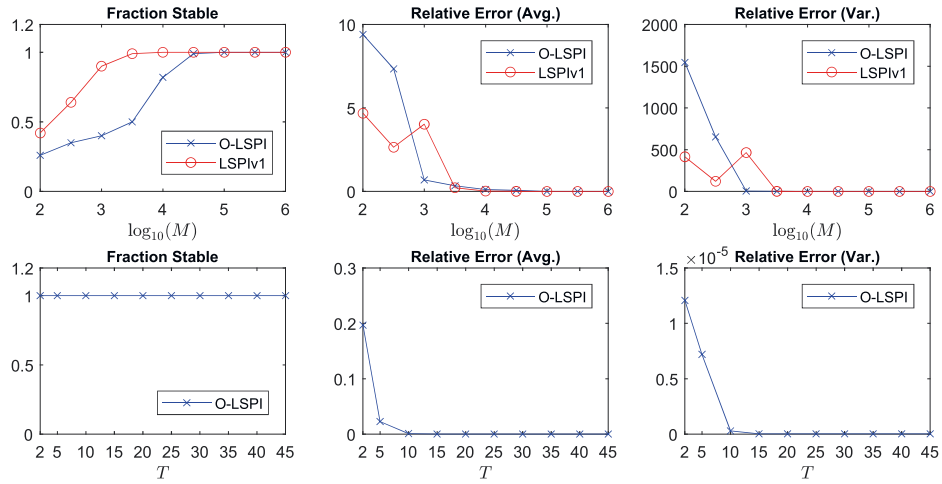

Figure 1: Experimental evaluation on the dynamics of (Krauth, Tu, and Recht 2019).

In recent years, there have been resurgent research interests in LQR problems, about learning the optimal solutions from the input/state/output data. The model-based certainty equivalence methods explicitly estimate the values of A, B and C in (8) from data, and obtain near-optimal solutions based on the estimations, see (Abbasi-Yadkori and Szepesvári 2011; Ouyang, Gagrani, and Jain 2017; Abeille and Lazaric 2018; Dean et al. 2018; Shirani Faradonbeh, Tewari, and Michailidis 2020; Umenberger and Schön 2020; Cassel, Cohen, and Koren 2020; Basei, Guo, and Hu 2020), to name a few. The model-free methods aim at finding the near-optimal solutions directly from the data, without the estimations of system dynamics. Action-value model-free methods learn the value functions of policies, and then generate new (improved) policies based on the estimated value functions, see (Bradtke, Ydstie, and Barto 1994; Tu and Recht 2018; Krauth, Tu, and Recht 2019; Abbasi-Yadkori, Lazic, and Szepesvári 2019; Tu and Recht 2019; Bian and Jiang 2019). Policy-gradient model-free methods directly learn the policies based on the gradient of some scalar performance measure with respect to the policy parameter, see (Fazel et al. 2018; Bu et al. 2019; Preiss et al. 2019; Mohammadi et al. 2019; Yang et al. 2019; Qu et al. 2020; Jansch-Porto, Hu, and Dullerud 2020a,b; Furieri, Zheng, and Kamgarpour 2020). Derivative-free model-free methods randomly search in the parameter space of policies for the near-optimal solutions, without explicitly estimate the gradient, see (Mania, Guy, and Recht 2018; Malik et al. 2019; Li et al. 2020).

Most of the model-free methods for LQR mentioned above are on-policy, fewer theoretical results exist for off-policy methods. Among the off-policy action-value model-free methods for LQR, the most related to our proposed O-LSPI are the LSPIv1 in (Krauth, Tu, and Recht 2019) and the MFLQv1 in (Abbasi-Yadkori, Lazic, and Szepesvári 2019). However, (a) no convergence result is reported for LSPIv1 in (Krauth, Tu, and Recht 2019), and (b) MFLQv1 in (Abbasi-Yadkori, Lazic, and Szepesvári 2019) needs to learn the P_K in (5) first in on-policy fashion, before it can learn the $Q(P_K)$ in (19) in off-policy fashion in each iteration, and (c) both the LSPIv1 and MFLQv1 need the knowledge of matrix C in (8), and (d) both the LSPIv1 and MFLQv1 need to solve a pseudo-inverse problem in each iteration. In contrast, in our O-LSPI, (a) a convergence result is given (Theorem 3), and (b) both the P_K and $Q(P_K)$ are learned in off-policy fashion (Lines 5 to 7 in Algorithm 1), and (c) no knowledge of C is required, and (d) the pseudo-inverse problem only needs to be solved once.

Finally, it is worth mentioning again that the robustness of RL to errors in the learning processes is analyzed in this paper. This is different from the the robustness of the controllers learned by RL to disturbances in the system dynamics, studied in (Gravell, Mohajerin Esfahani, and Summers 2020; Zhang, Hu, and Basar 2020a,b; Turchetta, Krause, and Trimpe 2020; Jiang and Jiang 2017). We also notice that using control theory to study RL algorithms, as we did in this paper, has become popular recently. By regarding the learning processes as dynamical systems, abundant results and techniques in control theory can be applied to obtain better understanding of RL algorithms, see (Srikant and Ying 2019; Gupta, Srikant, and Ying 2019; Hu and Syed 2019; Lee and He 2019; Bian and Jiang 2019).

Concluding Remarks

This paper analyzes the robustness of policy iteration for discrete-time LQR. It is proved that starting from any stabilizing initial policy, the solutions generated by policy iteration with errors are bounded and ultimately enter and stay in a neighbourhood of the optimal solution, as long as the errors are small and bounded. This result in the spirit of small-disturbance input-to-state stability is employed to prove the convergence of the optimistic least-squares policy iteration (O-LSPI), a novel off-policy model-free RL algorithm for discrete-time LQR with additive stochastic noises in the dynamics. The theoretical results are verified by the experiments on a numerical example.

Acknowledgments

This work has been supported in part by the U.S. National Science Foundation under Grants ECCS-1501044 and EPCN-1903781. Bo Pang thanks Dr. Stephen Tu for sharing the code of the least-squares policy iteration algorithms in (Krauth, Tu, and Recht 2019).

References

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

Abbasi-Yadkori, Y.; Lazic, N.; and Szepesvári, C. 2019. Model-free linear quadratic control via reduction to expert prediction. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 3108–3117.

Abeille, M., and Lazaric, A. 2018. Improved regret bounds for Thompson sampling in linear quadratic control problems. In *International Conference on Machine Learning (ICML)*, 1–9.

Basei, M.; Guo, X.; and Hu, A. 2020. Linear quadratic reinforcement learning: Sublinear regret in the episodic continuous-time framework. *arXiv preprint arXiv:2006.15316*.

Bertsekas, D. P. 1995. *Dynamic Programming and Optimal Control*, volume 2. Belmont, MA: Athena Scientific.

Bertsekas, D. P. 2011. Approximate policy iteration: A survey and some new methods. *Journal of Control Theory and Applications* 310–335.

Bian, T., and Jiang, Z. P. 2019. Continuous-time robust dynamic programming. *SIAM Journal on Control and Optimization* 4150–4174.

Boussios, C. I. 1998. *An approach for nonlinear control design via approximate dynamic programming*. Ph.D. Dissertation, Massachusetts Institute of Technology.

Bradtke, S. J.; Ydstie, B. E.; and Barto, A. G. 1994. Adaptive linear quadratic control using policy iteration. In *American Control Conference (ACC)*, 3475–3479.

Bu, J.; Mesbahi, A.; Fazel, M.; and Mesbahi, M. 2019. LQR through the lens of first order methods: Discrete-time case. *arXiv preprint arXiv:1907.08921*.

Cassel, A.; Cohen, A.; and Koren, T. 2020. Logarithmic regret for learning linear quadratic regulators efficiently. *International Conference on Machine Learning (ICML)* 1328–1337.

Dean, S.; Mania, H.; Matni, N.; Recht, B.; and Tu, S. 2018. Regret bounds for robust adaptive control of the linear quadratic regulator. In *Advances in Neural Information Processing Systems (NeurIPS)*, 4192–4201.

Fazel, M.; Ge, R.; Kakade, S.; and Mesbahi, M. 2018. Global convergence of policy gradient methods for the linear quadratic regulator. In *International Conference on Machine Learning (ICML)*, 1467–1476.

Furieri, L.; Zheng, Y.; and Kamgarpour, M. 2020. Learning the globally optimal distributed LQ regulator. In *Annual Conference on Learning for Dynamics and Control (L4DC)*, 287–297.

Gravell, B.; Mohajerin Esfahani, P.; and Summers, T. H. 2020. Learning optimal controllers for linear systems with multiplicative noise via policy gradient. *IEEE Transactions on Automatic Control*.

Gupta, H.; Srikant, R.; and Ying, L. 2019. Finite-time performance bounds and adaptive learning rate selection for two time-scale reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Hewer, G. 1971. An iterative technique for the computation of the steady state gains for the discrete optimal regulator. *IEEE Transactions on Automatic Control* 382–384.

Hu, B., and Syed, U. 2019. Characterizing the exact behaviors of temporal difference learning algorithms using Markov jump linear system theory. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Hylla, T. 2011. *Extension of inexact Kleinman-Newton methods to a general monotonicity preserving convergence theory*. Ph.D. Dissertation, Universität Trier.

Jansch-Porto, J. P.; Hu, B.; and Dullerud, G. 2020a. Convergence guarantees of policy optimization methods for Markovian jump linear systems. *American Control Conference (ACC)* 2882–2887.

Jansch-Porto, J. P.; Hu, B.; and Dullerud, G. 2020b. Policy learning of MDPs with mixed continuous/discrete variables: A case study on model-free control of Markovian jump systems. *Annual Conference on Learning for Dynamics and Control (L4DC)* 947–957.

Jiang, Y., and Jiang, Z. P. 2017. *Robust Adaptive Dynamic Programming*. Hoboken, New Jersey: Wiley-IEEE Press.

Jiang, Z.; Bian, T.; and Gao, W. 2020. Learning-based control: A tutorial and some recent results. *Foundations and Trends in Systems and Control* 176–284.

Jiang, Z. P.; Lin, Y.; and Wang, Y. 2004. Nonlinear small-gain theorems for discrete-time feedback systems and applications. *Automatica* 2129–2136.

Koralov, L., and Sinai, Y. G. 2007. *Theory of Probability and Random Processes*. Berlin: Springer, 2nd edition.

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

Lee, D., and He, N. 2019. A unified switching system perspective and ODE analysis of Q-learning algorithms. *arXiv preprint arXiv:1912.02270*.

Levine, S., and Koltun, V. 2012. Continuous inverse optimal control with locally optimal examples. In *International Conference on Machine Learning (ICML)*, 475–482.

Lewis, F. L.; Vrabie, D.; and Syrmos, V. L. 2012. *Optimal Control*. Hoboken, New Jersey: John Wiley & Sons.

Li, Y.; Tang, Y.; Zhang, R.; and Li, N. 2020. Distributed reinforcement learning for decentralized linear quadratic control: A derivative-free policy optimization approach. *Annual Conference on Learning for Dynamics and Control (L4DC)* 814–814.

Lillicrap, T. P.; Hunt, J. J.; Pritzel, A.; Heess, N.; Erez, T.; Tassa, Y.; Silver, D.; and Wierstra, D. 2016. Continuous control with deep reinforcement learning. In *International Conference on Learning Representations (ICLR)*.

Ljung, L. 1999. *System Identification: Theory for the user*. Upper Saddle River: Prentice Hall PTR, 2nd edition.

Malik, D.; Pananjady, A.; Bhatia, K.; Khamaru, K.; Bartlett, P.; and Wainwright, M. 2019. Derivative-free methods for policy optimization: Guarantees for linear quadratic systems. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2916–2925.

Mania, H.; Guy, A.; and Recht, B. 2018. Simple random search of static linear policies is competitive for reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 1805–1814.

Mohammadi, H.; Zare, A.; Soltanolkotabi, M.; and Jovanović, M. R. 2019. Convergence and sample complexity of gradient methods for the model-free linear quadratic regulator problem. *arXiv preprint arXiv:1912.11899*.

Monfort, M.; Liu, A.; and Ziebart, B. D. 2015. Intent prediction and trajectory forecasting via predictive inverse linear-quadratic regulation. In *AAAI Conference on Artificial Intelligence (AAAI)*, 3672–3678.

Ouyang, Y.; Gagrani, M.; and Jain, R. 2017. Learning-based control of unknown linear systems with Thompson sampling. *arXiv preprint arXiv:1709.04047*.

Pang, B.; Bian, T.; and Jiang, Z.-P. 2020. Robust policy iteration for continuous-time linear quadratic regulation. *arXiv preprint arXiv:2005.09528*.

Preiss, J. A.; Arnold, S. M.; Wei, C.-Y.; and Kloft, M. 2019. Analyzing the variance of policy gradient estimators for the linear-quadratic regulator. *arXiv preprint arXiv:1910.01249*.

Qu, G.; Yu, C.; Low, S.; and Wierman, A. 2020. Combining model-based and model-free methods for nonlinear control: A provably convergent policy gradient approach. *arXiv preprint arXiv:2006.07476*.

Åström, K. J., and Wittenmark, B. 1995. *Adaptive Control*. Reading, Massachusetts: Addison-Wesley, 2nd edition.

Shirani Faradonbeh, M. K.; Tewari, A.; and Michailidis, G. 2020. On adaptive linear–quadratic regulators. *Automatica*.

Sontag, E. D. 2008. Input to state stability: Basic concepts and results. In *Nonlinear and optimal control theory*, volume 1932. Berlin: Springer. 163–220.

Srikant, R., and Ying, L. 2019. Finite-time error bounds for linear stochastic approximation and TD learning. In *Annual Conference on Learning Theory (COLT)*, 2803–2830.

Sutton, R. S., and Barto, A. G. 2018. *Reinforcement Learning: An Introduction*. Cambridge, Massachusetts: MIT Press, 2nd edition.

Thompson, A. C. 1963. On certain contraction mappings in a partially ordered vector space. *Proceedings of the American Mathematical Society* 438 – 443.

Tsitsiklis, J. N. 2002. On the convergence of optimistic policy iteration. *Journal of Machine Learning Research* 59–72.

Tu, S., and Recht, B. 2018. Least-squares temporal difference learning for the linear quadratic regulator. In *International Conference on Machine Learning (ICML)*, 5005–5014.

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

Turchetta, M.; Krause, A.; and Trimpe, S. 2020. Robust model-free reinforcement learning with multi-objective bayesian optimization. In *IEEE International Conference on Robotics and Automation (ICRA)*, 10702–10708.

Umenberger, J., and Schön, T. B. 2020. Optimistic robust linear quadratic dual control. *Annual Conference on Learning for Dynamics and Control (L4DC)* 550–560.

Yang, Z.; Chen, Y.; Hong, M.; and Wang, Z. 2019. On the global convergence of actor-critic: A case for linear quadratic regulator with ergodic cost. *arXiv preprint arXiv:1907.06246*.

Zhang, K.; Hu, B.; and Basar, T. 2020a. On the stability and convergence of robust adversarial reinforcement learning: A case study on linear quadratic systems. *Advances in Neural Information Processing Systems (NeurIPS)*.

Zhang, K.; Hu, B.; and Basar, T. 2020b. Policy optimization for $\mathcal{H}_2$ linear control with $\mathcal{H}_\infty$ robustness guarantee: Implicit regularization and global convergence. *Learning for Dynamics and Control (L4DC)* 179–190.