Least-Squares Temporal Difference Learning for the Linear Quadratic Regulator

Stephen Tu [1] **Benjamin Recht** [1]

Abstract

Reinforcement learning (RL) has been successfully used to solve many continuous control tasks. Despite its impressive results however, fundamental questions regarding the sample complexity of RL on continuous problems remain open. We study the performance of RL in this setting by considering the behavior of the Least-Squares Temporal Difference (LSTD) estimator on the classic Linear Quadratic Regulator (LQR) problem from optimal control. We give the first finite-time analysis of the number of samples needed to estimate the value function for a fixed static state-feedback policy to within ε-relative error. In the process of deriving our result, we give a general characterization for when the minimum eigenvalue of the empirical covariance matrix formed along the sample path of a fast-mixing stochastic process concentrates above zero, extending a result by Koltchinskii and Mendelson (2013) in the independent covariates setting. Finally, we provide experimental evidence indicating that our analysis correctly captures the qualitative behavior of LSTD on several LQR instances.

1. Introduction

Despite excellent performance on locomotion (Kober et al., 2013; Levine & Koltun, 2014; Lillicrap et al., 2016; Schulman et al., 2016; Tedrake et al., 2004) and manipulation (Krishnan et al., 2017; Levine et al., 2016a;b; 2015) tasks, model-free reinforcement learning (RL) is still considered very data intensive. This is especially a problem for learning on robotic systems which requires human supervision, limiting the applicability of RL. While there have been various attempts to improve the sample efficiency of RL in practice (Gu et al., 2017; 2016; Schaul et al., 2016), a theoretical understanding of the issue is still an open question. A more rigorous foundation could help to differentiate between whether RL suffers from fundamental statistical limitations in the continuous setting, or if more sample efficient estimators are possible.

For continuous control tasks, the Linear Quadratic Regulator (LQR) is an ideal benchmark for studying RL, due to a combination of its theoretical tractability combined with its practical application in various engineering domains. Recent work by Dean et al. (2017) adopts this point of view, and studies the problem of designing a stabilizing controller for LQR when the system dynamics are unknown to the practitioner. Here, the authors take a model-based approach, and propose to directly estimate the state-transition matrices that describe the dynamics from observations. In practice however, model-free methods such as Q-learning or policy-gradient type algorithms are preferred over model-based methods due to their flexibility and ease of use. This naturally raises the question of how well do model-free RL methods perform on the LQR problem.

In this paper, we shed light on this question by focusing on the classic Least-Squares Temporal Difference (LSTD) estimator (Boyan, 1999; Bradtke & Barto, 1996). Given a sample trajectory from a Markov Decision Process (MDP) in feedback with a fixed policy π, LSTD computes the value function V^π associated to π. Estimating V^π is the core primitive in value and policy-iteration type algorithms (Sutton & Barto, 1998). The key property exploited by LSTD is the *linear-architecture* assumption, which states that the value function can be expressed as a linear function after applying a known non-linear transformation to the state. To the best of our knowledge, LQR is the simplest continuous problem which exhibits this property.

Our main result regarding the LSTD estimator for LQR is an upper bound on the necessary length of a single trajectory to estimate the value function of a stabilizing state-feedback policy. Letting n denote the dimension of the state and ignoring instance specific factors, we establish that roughly n^2/ε^2 samples are sufficient to estimate the value function up to ε-relative error. Our analysis builds upon the work of Lazaric et al. (2012), which requires bounding the minimum eigenvalue of the sample covariance matrix formed by the

[1]EECS Department, University of California, Berkeley. Correspondence to: Stephen Tu <stephent@berkeley.edu>.

Proceedings of the 35^{th} International Conference on Machine Learning, Stockholm, Sweden, PMLR 80, 2018.

transformed state vectors; the same eigenvalue quantity also appears in many other analyses of the LSTD estimator in the literature (Lazaric et al., 2012; Liu et al., 2015; 2012; Prashanth et al., 2014). We bound this quantity by studying the more general problem of controlling the minimum eigenvalue of the covariance matrix formed from dependent covariates that mix quickly to a stationary distribution. Our analysis extends an elegant technique based on small-ball probabilities from Koltchinskii and Mendelson (2013), and is of independent interest. Specializing to the setting when the covariates are bounded almost surely, our result improves upon the analysis given by Lazaric et al.

We conclude our work with an end-to-end empirical comparison of the model-free Least-Squares Policy Iteration (LSPI) algorithm (Lagoudakis & Parr, 2003) with the model-based methods proposed in Dean et al. Our experiments show that model-free LSPI can be substantially less sample efficient and less robust compared to model-based methods. This corroborates our theoretical results which suggest a factor of state-dimension gap between the number of samples needed to estimate a value function versus the bounds given in Dean et al. for robustly computing a stabilizing controller. We hope that our findings encourage further investigation, both theoretical and empirical, into the performance of RL on continuous control problems.

1.1. Related Work

Least-squares methods for temporal difference learning are well-studied in reinforcement learning, with asymptotic convergence results in a general MDP setting provided by (Tsitsiklis & Van Roy, 1997; Yu & Bertsekas, 2009). More recently, non-asymptotic analyses were given in both the batch setting (Antos et al., 2008; Farahmand et al., 2016; Lazaric et al., 2012) and the online setting (Liu et al., 2015; 2012; Prashanth et al., 2014). The prevailing assumption employed in prior art is that the MDP has uniformly bounded features and rewards, which excludes the LQR problem. We note that earlier results by Bradtke (1993; 1994) studied policy-iteration specifically for LQR, and proved an asymptotic convergence result. To the best of our knowledge, our work is the first to provide finite-time results for temporal difference learning on LQR. Furthermore, our concentration result for the sample covariance matrix drawn from a mixing process specialized to the bounded setting improves upon Lemma 4 of Lazaric et al. (2012), by reducing the necessary trajectory length from $\Omega(d^2)$ to $\Omega(d)$, where d is the dimension of the features.

The problem of estimating the spectra of an empirical covariance matrix formed from independent samples has received much attention in the past decade. Some representative results can be found in (Adamczak et al., 2011; Koltchinskii & Mendelson, 2013; Mendelson & Paouris, 2014; Rudelson & Vershynin, 2009; Srivastava & Vershynin, 2013; Vershynin, 2011) and the references within. Our focus on the result of Koltchinskii and Mendelson in this paper is primarily motivated by the fact that their proof technique is generalizable to the dependent-data setting using standing mixing assumptions. The use of distributional mixing assumptions for proving uniform convergence bounds is by now a well-established technique in the statistics and machine learning literature; see (Mohri & Rostamizadeh, 2008; 2010; Yu, 1994) for some of the earlier results, and (Agarwal & Duchi, 2013; Kuznetsov & Mohri, 2015; 2016; McDonald et al., 2017) for generalizations to time-series and online learning. In this work, our focus is on bounding a very particular empirical process (the minimum eigenvalue of a sample covariance matrix), and not in developing general machinery for empirical process theory on dependent data.

2. A Sample Covariance Bound for Fast-Mixing Processes

In this section, we state our result regarding the minimum eigenvalue of the sample covariance matrix formed along a trajectory of a β-mixing process. We start by fixing notation. Let $(X_k)_{k=1}^{\infty}$ be an $\mathbb{R}^n$-valued discrete-time stochastic process adapted to a filtration $(\mathcal{F}_k)_{k=1}^{\infty}$. For all $k \geq 1$, let ν_k denote the marginal distribution of X_k. We assume that $(X_k)_{k=1}^{\infty}$ admits a unique stationary distribution ν_∞, and we define the β-mixing coefficient $\beta(k)$ with respect to ν_∞ as $\beta(k) := \sup_{t \geq 1} \mathbb{E}_{X_1^t}[\|\mathbb{P}_{X_{t+k}}(\cdot|\mathcal{F}_t) - \nu_\infty\|_{\text{tv}}]$. Here, the notation X_1^t refers to the prefix $X_1^t := (X_1, ..., X_t)$ and $\|\cdot\|_{\text{tv}}$ refers to the total-variation norm on probability measures. Our main assumption in what follows is that $(X_k)_{k=1}^{\infty}$ is β-mixing to its stationary distribution at an exponential decay rate, i.e. $\beta(k) \leq \Gamma \rho^k$ for some fixed $\Gamma > 0$ and $\rho \in (0, 1)$. We note that our analysis is easily amendable to slower (e.g. polynomial) decay rates.

We are now ready to state our generalization of Theorem 2.1 from Koltchinskii and Mendelson (2013) for fast-mixing processes. We note that no attempt was made to optimize the constants appearing in the result.

Theorem 2.1. *Fix a $\delta \in (0, 1)$. Suppose that $(X_k)_{k=1}^{\infty}$ is a discrete-time stochastic process with stationary distribution ν_∞ that satisfies $\beta(a) \leq \Gamma \rho^a$ for some $\Gamma > 0$, $\rho \in (0, 1)$. For any positive $\tau > 0$ define the small-ball probability $Q_\infty(\tau)$ as*

$$Q_\infty(\tau) := \inf_{t \in S^{n-1}} \mathbb{P}_{\nu_\infty}\{|\langle t, X\rangle| \geq \tau\}. \tag{2.1}$$

Suppose that there exists a τ satisfying $Q_\infty(\tau) > 0$. Define

$$\Psi_1 := \frac{1024 \mathbb{E}_{\nu_\infty}[\|X\|^2]}{\tau^2 Q_\infty^2(\tau)},$$

$$\Psi_2(N) := \frac{32}{Q_\infty^2(\tau)} \log\left(\frac{4}{\delta(1-\rho)} \log\left(\frac{2\Gamma N}{\delta}\right)\right).$$

If N satisfies

$$N \geq \frac{1}{1-\rho} \log\left(\frac{2\Gamma N}{\delta}\right) (\max\{\Psi_1, \Psi_2(N)\} + 1) \ , \tag{2.2}$$

then with probability at least $1-\delta$,

$$\lambda_{\min}\left(\frac{1}{N}\sum_{k=1}^{N} X_k X_k^{\mathsf{T}}\right) \geq \frac{\tau^2 Q_\infty(\tau)}{8} \ .$$

Proofs for all the results in this paper can be found in the full version (Tu & Recht, 2017).

Following a similar line of reasoning as in Koltchinskii and Mendelson, we immediately recover a corollary to Theorem 2.1, where the small-ball condition in (2.1) is replaced by a stronger moment contractivity assumption $\sup_{t\in S^{n-1}} \frac{\|\langle X,t\rangle\|_{L^2}}{\|\langle X,t\rangle\|_{L^1}} \leq B$.

3. Fast-Mixing of Linear Dynamical Systems

In order to pave the way for our main result regarding LQR, we need to understand the mixing time of a stable linear, time-invariant (LTI) dynamical system. This will allow us to directly apply the results from Section 2. While it is known that an LTI system mixes at a linear rate (see e.g. (Mokkadem, 1988)), our focus is to derive the specific constants involved. Towards this, consider the LTI system

$$X_{k+1} = AX_k + W_k \ , \ \ W_k \sim \mathcal{N}(0, I) \ , \tag{3.1}$$

with A an $n \times n$ matrix, initial condition $X_0 = 0$, and W_k independent from $W_{k'}$ for all $k \neq k'$.

It is not hard to see that the marginal distribution ν_k of X_k evolving according to (3.1) is $\mathcal{N}(0, P_k)$, where the covariance $P_k := \sum_{t=0}^{k-1} (A^t)(A^t)^{\mathsf{T}}$ is positive-definite. The stability of the linear system (3.1) is equivalent to the spectral radius of A, denoted $\rho(A)$, being strictly less than one. When $\rho(A) < 1$, the stationary distribution ν_∞ of $(X_k)_{k=1}^{\infty}$ is $\mathcal{N}(0, P_\infty)$, where the covariance matrix P_∞ is the unique, positive-definite solution of the discrete-time Lyapunov equation $AP_\infty A^{\mathsf{T}} - P_\infty + I = 0$.

Observe that in the case of a Markov chain, the β-mixing coefficient simplifies to

$$\beta(k) = \sup_{t\geq 1} \mathbb{E}_{x\sim\nu_t}[\|\mathbb{P}_{X_k}(\cdot|X_0{=}x) - \nu_\infty\|_{\mathrm{tv}}] \ . \tag{3.2}$$

The following upper bound on $\mathbb{E}_{x\sim\nu_t}[\|\mathbb{P}_{X_k}(\cdot|X_0{=}x) - \nu_\infty\|_{\mathrm{tv}}]$ uses the assumption of a known decay on the spectral norm of A^k.

Proposition 3.1. *Suppose that $\|A^k\| \leq \Gamma\rho^k$ for all $k \geq 0$, where $\Gamma > 0$ and $\rho \in (0,1)$. Let $\mathbb{P}_{X_k}(\cdot|X_0{=}x)$ denote the conditional distribution of X_k given $X_0 = x$. We have that for all $k \geq 0$ and any distribution ν_0 over x with $\zeta = \mathbb{E}_{x\sim\nu_0}[\|x\|^2]$,*

$$\mathbb{E}_{x\sim\nu_0}[\|\mathbb{P}_{X_k}(\cdot|X_0{=}x) - \nu_\infty\|_{\mathrm{tv}}] \leq \frac{\Gamma}{2}\sqrt{\zeta + \frac{n}{1-\rho^2}}\rho^k \ .$$

Now we turn our attention to obtaining a quantitative handle on the decay rate of the spectral norm of A^k. To do this, we introduce some basic concepts from robust control theory; see (Zhou et al., 1995) for a more thorough treatment. Let $\mathbb{T}$ (resp. $\mathbb{D}$) denote the unit circle (resp. open unit disk) in the complex plane. Let $\mathcal{RH}_\infty$ denote the space of matrix-valued, real-rational functions which are analytic on $\mathbb{D}^c$. For a $G \in \mathcal{RH}_\infty$, we define the $\mathcal{H}_\infty$-norm $\|G\|_{\mathcal{H}_\infty}$ as $\|G\|_{\mathcal{H}_\infty} := \sup_{z\in\mathbb{T}} \|G(z)\|$. Furthermore, given a square matrix A, we define its resolvant $\Phi_A(z)$ as $\Phi_A(z) := (zI - A)^{-1}$. When A is stable, $\Phi_A \in \mathcal{RH}_\infty$, and hence $\|G\|_{\mathcal{H}_\infty} < \infty$. The next proposition characterizes the decay rate in terms of the stability radius $\rho(A)$ and the $\mathcal{H}_\infty$-norm $\|\Phi_A\|_{\mathcal{H}_\infty}$.

Proposition 3.2 (See e.g. Lemma 1 from (Goldenshluger & Zeevi, 2001))**.** *Let A be a stable matrix with spectral radius $\rho(A)$. Fix any $\rho \in (\rho(A), 1)$. For all $k \geq 1$, we have*

$$\|A^k\| \leq \|\Phi_{\rho^{-1}A}\|_{\mathcal{H}_\infty}\rho^k \ .$$

Combining these last two claims with (3.2) and using the fact that $\mathbb{E}_{\nu_t}[\|X\|^2] \leq \mathbb{E}_{\nu_\infty}[\|X\|^2]$ for all $t \geq 1$, we have the following corollary which establishes an exponential decay rate for the mixing-time of a stable linear system.

Corollary 3.3. *Fix any $\rho \in (\rho(A), 1)$. For any $k \geq 1$ we have*

$$\beta(k) \leq \frac{\|\Phi_{\rho^{-1}A}\|_{\mathcal{H}_\infty}}{2}\sqrt{\mathbf{Tr}(P_\infty) + \frac{n}{1-\rho^2}}\rho^k \ .$$

4. Least-Squares Temporal Difference Learning

We turn our attention to the LSTD estimator. The goal of LSTD is to compute the value function V^π associated with a policy π for an MDP. This is an important primitive operation in many RL algorithms, such as policy-iteration.

Consider an MDP $\mathcal{M} = (\mathcal{S}, \mathcal{A}, p, \gamma, r)$, where $\mathcal{S}$ denotes the state-space, $\mathcal{A}$ denotes the action-space, $p : \mathcal{S}\times\mathcal{A} \longrightarrow \mu(S)$ denotes the transition kernel of the dynamics with $\mu(S)$ denoting the space of measures on $\mathcal{S}$, $\gamma \in (0,1)$ is the discount factor, and $r : \mathcal{S}\times\mathcal{A} \longrightarrow \mathbb{R}$ is the reward function. Given a policy $\pi : \mathcal{S} \longrightarrow \mathcal{A}$, its value function $V^\pi : \mathcal{S} \longrightarrow \mathbb{R}$ is defined as $V^\pi(x) := \mathbb{E}\left[\sum_{k=0}^{\infty}\gamma^k r(X_k, \pi(X_k)) \,\middle|\, X_0 = x\right]$ with $X_{k+1} \sim p(\cdot|X_k, \pi(X_k))$.

Bellman's equation for the discounted, infinite-horizon cost (Bertsekas, 2007) states that V^π is the solution to the fixed-point equation

$$V^\pi(x) = r(x, \pi(x)) + \gamma \mathbb{E}_{x' \sim p(\cdot|x,\pi(x))}[V^\pi(x')] \,. \quad (4.1)$$

When $\mathcal{S}$ is finite, dynamic programming can be used to solve (4.1). However, when $\mathcal{S}$ is continuous, solving (4.1) in general is difficult without imposing additional structure. By assuming that V^π admits the representation $V^\pi(x) = \langle \phi(x), v_\pi \rangle$ for some feature map $\phi : \mathcal{S} \longrightarrow \mathbb{R}^d$, one turns (4.1) into a system of linear equations; this is known as the *linear-architecture* assumption. Specifically, if the dynamics $p(\cdot|x,u)$ are known, then V^π can be recovered as the solution to the system of linear equations for v_π,

$$\langle \phi(x) - \gamma \psi(x), v_\pi \rangle = r(x, \pi(x)) \,, \quad (4.2)$$

with $\psi(x) := \mathbb{E}_{x' \sim p(\cdot|x,\pi(x))}[\phi(x')]$.

Of course, we are interested in settings where the dynamics $p(\cdot|x,u)$ are not known, and hence we cannot directly compute $\psi(x)$ in (4.2). This is where the LSTD estimator enters the picture: given a trajectory $\{(X_k, R_k, X_{k+1})\}_{k=1}^N$ of length N, the LSTD estimator $\widehat{v}_{\text{lstd}}$ approximates the solution to (4.2) by solving

$$\widehat{v}_{\text{lstd}} = \left(\sum_{k=1}^N \phi_k (\phi_k - \gamma \phi_{k+1})^\mathsf{T} \right)^\dagger \left(\sum_{k=1}^N \phi_k R_k \right) \,, \quad (4.3)$$

where $\phi_k = \phi(X_k)$ and $(\cdot)^\dagger$ denotes the pseudo-inverse. The curious looking nature of (4.3) accounts for the fact that when $\phi(X_k) - \gamma\phi(X_{k+1})$ is used as an estimate for $\phi(X_k) - \gamma\psi(X_k)$ in (4.2), the noise in the linear measurement is not independent from the covariate; see e.g. (Bradtke & Barto, 1996) for a more detailed discussion of the issue.

We will let the matrix $\Phi \in \mathbb{R}^{N \times d}$ denote the matrix where the k-th row is $\phi(X_k)$. While our main result is a bound on the sample complexity of the LSTD estimator on LQR, we first consider the implications of Theorem 2.1 on LSTD when both the features ϕ and the rewards are bounded, in order to compare to the setting of Lazaric et al. We will then study the LQR problem, which is the simplest non-trivial MDP which relaxes these boundedness assumptions.

4.1. Bounded features and rewards

For this section only we assume that $\sup_{x \in \mathcal{S}} \|\phi(x)\|_\infty^2 \leq \overline{L}$ and $\sup_{x \in \mathcal{S}, a \in \mathcal{A}} |r(s,a)| \leq R_{\max}$. Under these assumptions, we immediately have $\sup_{x \in \mathcal{S}} |V^\pi(x)| \leq \frac{1}{1-\gamma} R_{\max} := V_{\max}$. The following result from Lazaric et al. gives a bound on the in-sample prediction error of the estimator $\widehat{V^\pi}(\cdot) := \langle \phi(\cdot), \widehat{v}_{\text{lstd}} \rangle$.

Theorem 4.1 (Theorem 1, Lazaric et al. (2012))**.** *With probability at least $1 - \delta$, we have*

$$\|\widehat{V^\pi} - V^\pi\|_N \leq \eta V_{\max} \sqrt{\frac{\overline{L} d}{\nu_N}} \left(\sqrt{\frac{8 \log(2d/\delta)}{N}} + \frac{1}{N} \right) \,,$$

where $\eta = \frac{\gamma}{1-\gamma}$ and ν_N is the smallest non-zero eigenvalue of $\frac{1}{N}\Phi^\mathsf{T}\Phi$ and $\|\cdot\|_N$ denotes the L^2-norm w.r.t. the empirical measure $\frac{1}{N}\sum_{k=1}^N \delta_{X_k}$.

Immediately, Theorem 2.1 combined with Theorem 4.1 yield the following corollary.

Corollary 4.2. *Suppose that the stochastic process $\{\phi(X_k)\}_{k=1}^\infty$ mixes to some stationary measure ν_∞ at a rate $\beta(k) \leq \Gamma \rho^k$. Furthermore, suppose that $0 < \ell \leq \lambda_{\min}(\mathbb{E}_{\nu_\infty}[\phi(X)\phi(X)^\mathsf{T}]) \leq L$ and $\sup_{t \in S^{d-1}} \frac{\|\langle \phi(X), t \rangle\|_{L^2(\nu_\infty)}}{\|\langle \phi(X), t \rangle\|_{L^1(\nu_\infty)}} \leq O(1)$. Fix a $\delta \in (0,1)$, and suppose that N satisfies*

$$\frac{N}{\log(\Gamma N/\delta) \log\log(\Gamma N/\delta)} \geq \Omega\left(\frac{1}{1-\rho} \frac{dL}{\ell} \right) \,.$$

Then, with probability at least $1 - \delta$,

$$\|\widehat{V^\pi} - V^\pi\|_N \leq O\left(\eta V_{\max} \sqrt{\frac{\overline{L} d}{\ell}} \left(\sqrt{\frac{\log(d/\delta)}{N}} + \frac{1}{N} \right) \right) \,.$$

We remark that Lemma 4 of Lazaric et al. also provides an analysis of $\lambda_{\min}(\frac{1}{N}\Phi^\mathsf{T}\Phi)$, but under the boundedness assumptions of this section. Let us compare Theorem 4.1 to their Lemma 4. Specializing their result to the case when the mixing is characterized by $\beta(k) \leq (1/2)^k$, they prove that $\lambda_{\min}\left(\frac{1}{N}\Phi^\mathsf{T}\Phi\right) \geq \Omega(\ell)$ where $\ell = \lambda_{\min}(\mathbb{E}_{\nu_\infty}[\phi(X)\phi(X)^\mathsf{T}])$ as long as

$$\frac{N}{\log^2(N/\delta)} \geq \Omega\left(\frac{\overline{L} d^2}{\ell} \right) \,.$$

Under the same setting, as long as the contractivity condition in Corollary 4.2 holds for the stationary distribution, our result relaxes the condition on N to

$$\frac{N}{\log(N/\delta) \log\log(N/\delta)} \geq \Omega\left(\frac{Ld}{\ell} \right) \,,$$

where $L = \lambda_{\max}(\mathbb{E}_{\nu_\infty}[\phi(X)\phi(X)^\mathsf{T}])$. We therefore improve the bound from Lazaric et al. by reducing the minimum trajectory length from $N \geq \widetilde{\Omega}(d^2)$ to $N \geq \widetilde{\Omega}(d)$.

4.2. Linear Quadratic Regulator

We now study the performance of LSTD on LQR. The LQR problem is an MDP with linear dynamics

$$X_{k+1} = AX_k + BU_k + W_k \,, \;\; W_k \sim \mathcal{N}(0, I) \,, \quad (4.4)$$

and quadratic rewards

$$r(x,u) = -(x^\mathsf{T} Q x + u^\mathsf{T} R u)\,,$$

where A is $n \times n$, B is $n \times n_i$, Q and R are positive-definite matrices, and W_k is independent from $W_{k'}$ for all $k \neq k'$. It is well known that the LQR problem can be solved with a linear feedback policy $\pi(x) = Kx$, and hence we will assume linear policies in the sequel. We will further assume that the policy π *stabilizes* the dynamics, i.e. the closed-loop matrix $L := A + BK$ is a stable matrix. This stability assumption ensures that the dynamics mix and the value function is finite. We note that our analysis does not handle the case when L is not stable, but $\sqrt{\gamma}L$ is. In this case, the value function is finite, but the dynamics do not mix.

Under our assumptions, it is straightforward to show by Bellman's equation (4.1) that $V^\pi(x) = -x^\mathsf{T} P_\pi x - \eta\,\mathbf{Tr}(P_\pi)$, where $\eta = \gamma/(1-\gamma)$ and P_π uniquely solves the discrete-time Lyapunov equation,

$$(\gamma^{1/2}L)^\mathsf{T} P_\pi (\gamma^{1/2}L) - P_\pi + (Q + K^\mathsf{T} R K) = 0\,.$$

Furthermore, the stationary distribution of the dynamics is $\nu_\infty = \mathcal{N}(0, P_\infty)$, where P_∞ uniquely solves the Lyapunov equation $LP_\infty L^\mathsf{T} - P_\infty + I = 0$. To cast this problem into the linear-architecture format of LSTD, we define the feature map $\phi(x)$ as $\phi(x) = \mathrm{svec}(xx^\mathsf{T} + \eta I)$. Here, $\mathrm{svec} : \mathrm{Sym}_{n\times n} \longrightarrow \mathbb{R}^{n(n+1)/2}$ is the linear operator mapping the space of $n \times n$ symmetric matrices (denoted $\mathrm{Sym}_{n\times n}$) to vectors while preserving the property that $\langle \mathrm{svec}(M_1), \mathrm{svec}(M_2)\rangle_{\mathbb{R}^{n(n+1)/2}} = \langle M_1, M_2\rangle_{\mathrm{Sym}_{n\times n}}$ for all symmetric M_1, M_2. We will also let $\mathrm{smat} : \mathbb{R}^{n(n+1)/2} \longrightarrow \mathrm{Sym}_{n\times n}$ denote the inverse of svec. Hence in our setting, d (the dimension of the lifted features) is $d = n(n+1)/2$. We will denote $v_\pi = \mathrm{svec}(P_\pi)$.

Our main result is the following theorem which gives a bound on the error of the difference between the LSTD estimator $\widehat{P} = \mathrm{smat}(\widehat{v}_{\mathrm{lstd}})$ and the true value function P_π.

Theorem 4.3. *Fix $\delta \in (0,1)$ and $\rho \in (\rho(L), 1)$. Define $\widetilde{\Gamma} := \|\Phi_{\rho^{-1}L}\|_{\mathcal{H}_\infty}\sqrt{\mathbf{Tr}(P_\infty) + n/(1-\rho^2)}$. Let $\widehat{P}$ denote the LSTD estimator (4.3) for the LQR problem. Suppose that N is large enough to satisfy*

$$\frac{N}{\log(\widetilde{\Gamma}N/\delta)\log\log(\widetilde{\Gamma}N/\delta)} \geq \Omega\left(\frac{\max\{\mathbf{Tr}(P_\infty)^2, \eta^2 n\}}{(1-\rho)\lambda^2_{\min}(P_\infty)}\right).$$

Then, with probability at least $1-\delta$,

$$\frac{\|\widehat{P} - P_\pi\|_F}{\|P_\pi\|_F} \leq \widetilde{O}\left(\frac{\eta\sqrt{\|P_\infty\|}\max\{\mathbf{Tr}(P_\infty), \eta\sqrt{n}\}}{\sqrt{N}\lambda^2_{\min}(P_\infty)}\right), \tag{4.5}$$

where $\widetilde{O}(\cdot)$ hides $\mathrm{polylog}(N, n, 1/\delta)$ factors.

We make several remarks on the behavior of (4.5). Let us first simplify it to ease the exposition, by applying the bound $\mathbf{Tr}(P_\infty) \leq n\|P_\infty\|$ and assuming we are in the regime when $n \gg (\eta/\|P_\infty\|)^2$ so that $n\|P_\infty\|$ dominates $\eta\sqrt{n}$. With these simplifications, (4.5) becomes

$$\frac{\|\widehat{P} - P_\pi\|_F}{\|P_\pi\|_F} \leq \widetilde{O}\left(\frac{n}{(1-\gamma)\sqrt{N}}\frac{\|P_\infty\|^{3/2}}{\lambda^2_{\min}(P_\infty)}\right),$$

which yields the sufficient condition that

$$N \geq \widetilde{\Omega}\left(\frac{n^2}{(1-\gamma)^2\varepsilon^2}\frac{\kappa^3(P_\infty)}{\lambda_{\min}(P_\infty)}\right),\quad \kappa(P_\infty) := \frac{\|P_\infty\|}{\lambda_{\min}(P_\infty)} \tag{4.6}$$

samples ensure the relative error is less than ε.

We first remark on the dependence of (4.6) on the spectral properties of P_∞. In particular, (4.6) suggests that as $\kappa(P_\infty)$ increases, more samples are needed to reach a fixed ε tolerance. In controls parlance, the matrix P_∞ is known as the controllability gramian. A system with large $\kappa(P_\infty)$ is one where different modes exhibit qualitatively different behaviors. The simplest example of this is when the closed-loop matrix is $L = \mathrm{diag}(\rho_1, ..., \rho_n)$ with $\rho_k \in (0,1)$, in which case $P_\infty = \mathrm{diag}(1/(1-\rho_1^2), ..., 1/(1-\rho_n^2))$. Here, as ρ_1 increases towards one, (4.6) predicts that estimating the value function requires more samples. In Section 5.1, we show that this predicted behavior actually occurs in numerical simulations.

Let us compare (4.6) to the setting of Dean et al. (2017), where ordinary least-squares is used to estimate the state-transition matrices (A, B) of (4.4), and a robust control procedure is used to design a controller to stabilize (4.4). Ignoring problem specific parameters, Corollary 4.3 of Dean et al. states that at most $\widetilde{\Omega}(n/\varepsilon^2)$ samples are needed to design a controller which incurs a relative error of at most ε. On the other hand, (4.6) suggests that $\widetilde{\Omega}(n^2/\varepsilon^2)$ samples are needed to estimate a single value function. While this gap in the upper bounds is not directly comparable, it does suggests that for LQR, model-based methods may perform better than policy-iteration methods such as Least-Squares Policy Iteration (LSPI), which require multiple policy evaluation steps. In Section 5.2, we provide empirical evidence that shows this is indeed the case for certain LQR instances. We leave as future work lower bounds to separate the sample complexities of model-free and model-based methods.

The remainder of the section is dedicated to a proof sketch of Theorem 4.3. Because the estimator (4.3) is not a standard least-squares estimator (despite its name), some analysis is needed to manipulate the estimator into a form that is easier to analyze. We follow the development in Lazaric et al. and state the main structural result of their paper below.

Lemma 4.4 (Lazaric et al. (2012))**.** *As long as Φ has full column rank, the LSTD estimator $\widehat{P}$ satisfies the following*

inequality,

$$\|\widehat{P} - P_\pi\|_F \leq \frac{\eta \left\| \sum_{k=1}^N \phi_k(\phi_{k+1} - \mathbb{E}[\phi_{k+1}|X_k])^\mathsf{T} v_\pi \right\|}{\lambda_{\min}\left(\sum_{k=1}^N \phi_k \phi_k^\mathsf{T}\right)} .$$

The proof of Theorem 4.3 proceeds by bounding the terms that appear in Lemma 4.4. Theorem 2.1 from Section 2 combined with the mixing analysis in Section 3 can be directly applied to estimate the minimum eigenvalue of the matrix $\sum_{k=1}^N \phi(X_k)\phi(X_k)^\mathsf{T}$. The term in the numerator can also be dealt with via standard martingale techniques. The two bounds are stated below.

Lemma 4.5. *Suppose that the hypothesis of Theorem 4.3 hold. Then, with probability at least* $1-\delta$*,*

$$\lambda_{\min}\left(\sum_{k=1}^N \phi(X_k)\phi(X_k)^\mathsf{T}\right) \geq \Omega(N\lambda_{\min}^2(P_\infty)) .$$

Lemma 4.6. *With probability at least* $1-\delta$*,*

$$\left\| \sum_{k=1}^N \phi(X_k)(\phi(X_{k+1}) - \mathbb{E}[\phi(X_{k+1})|X_k])^\mathsf{T} v_\pi \right\| \leq \widetilde{O}(\|v_\pi\|\sqrt{N}(\mathbf{Tr}(P_N) + \eta\sqrt{n})\|LP_N^{1/2}\|) .$$

5. Experiments

We conduct numerical experiments on LSTD for value function estimation, and Least-Squares Policy Iteration (LSPI) for an end-to-end comparison with the model-based methods in Dean et al. (2017). Our implementation is carried out in Python using `numpy` for linear algebraic computations and PyWren (Jonas et al., 2017) for parallelization.

In our first set of experiments, we construct synthetic examples where we vary the condition number of the resulting closed-loop controllability gramian matrix. We find that on these instances, as the condition number increases, the required number of samples to estimate the value function to fixed relative error increases, as predicted by our result in Theorem 4.3. In our second set of experiments, we compare model-free policy iteration (LSPI) to two model-based methods: (a) the naïve nominal model controller which uses a controller designed assuming that the nominal model has zero error, and (b) a controller based on a semidefinite relaxation to the non-convex robust control problem with static state-feedback. Our experiments show that model-free policy iteration requires more samples than model-based methods for the instances we consider.

5.1. Synthetic Data

The goal in this section is to showcase the qualitative behavior of LSTD on LQR predicted by Theorem 4.3 as

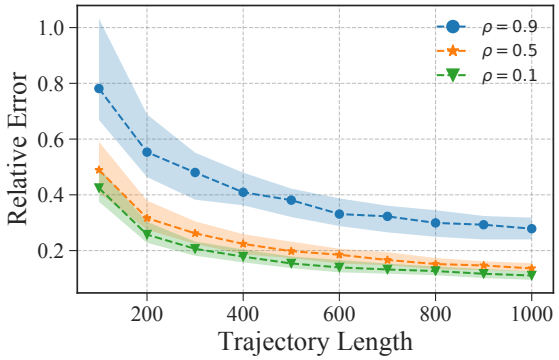

Figure 1: Performance of LSTD on LQR instances where the closed loop response is $L = A + BK = \rho I_5$ for $\rho \in \{0.1, 0.5, 0.9\}$. The dashed line represents the median relative error, and the shaded region covers the 25-th to 75-th percentile of the relative error out of 100 trajectories.

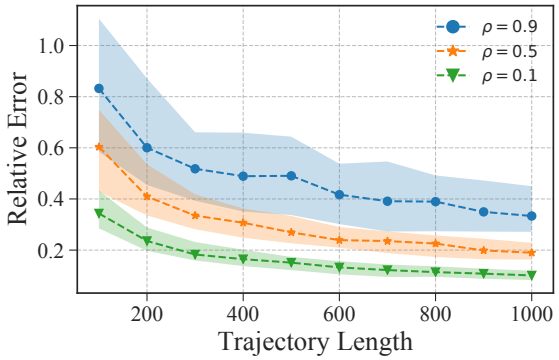

Figure 2: Performance of LSTD on LQR instances where $B = K = 0_{5\times 5}$ and A is generated randomly with spectral radius $\rho(A) \in \{0.1, 0.5, 0.9\}$. The dashed line represents the median relative error, and the shaded region covers the 25-th to 75-th percentile of the relative error out of 100 trajectories.

the conditioning of the closed-loop controllability gramian varies. We consider several instances of LQR with $n = 5$, $Q = R = 0.1I_5$, and $\gamma = 0.9$, where the state transition matrices (A, B) and the policy $\pi(x) = Kx$ will be specified later. For each configuration, we collect 100 trajectories of length $N = 1000$. For each trajectory, we take the first N_p points for $N_p \in \{100, 200, ..., 1000\}$ and compute the LSTD estimator $\widehat{P}_{N_p}$ on the first N_p data points. We then compute the relative error $\frac{\|P_\pi - \widehat{P}_{N_p}\|_F}{\|P_\pi\|_F}$ for each N_p, and report the median and 25-th to 75-th percentile over the 100 trajectories.

In the first experiment, we set $A = B = I_5$, and we vary $K \in \{\text{diag}(-(1-\rho), -(1-\rho), ..., -(1-0.01)) : \rho \in \{0.1, 0.5, 0.9\}\}$ so that $L = A + BK = \text{diag}(\rho, \rho, ..., 0.01)$ and $\kappa(P_\infty) = \frac{1}{1-\rho^2}(1 - 0.01^2)$ for $\rho \in \{0.1, 0.5, 0.9\}$. Theorem 4.3 predicts that as ρ increases towards one, the

number of samples required for ε-relative error increases as well. Figure 1 corroborates this finding.

For our second experiment, we set $B = K = 0_{5\times 5}$ so that the closed-loop response is simply A. We generate random instances of A as follows. For each $\rho \in \{0.1, 0.5, 0.9\}$, we generated 1000 A instances by setting $A_{ii} = \rho$ for all diagonal entries and $A_{ij} \sim$ clip($\mathcal{N}(0,1)$, -1, 1) independently for all upper triangular entries. We order the A instances by $\kappa(P_\infty(A))$, where $AP_\infty(A)A^\mathsf{T} - P_\infty(A) + I = 0$ and take the median. This results in $\kappa \approx 7$, $\kappa \approx 35$, and $\kappa \approx 7 \times 10^5$ for $\rho = 0.1, 0.5, 0.9$, respectively. We then run LSTD on the three median instances, reporting the results in Figure 2. Once again, as $\kappa(P_\infty(A))$ increases, the required trajectory length increases, as suggested by Theorem 4.3. We note, however, that Theorem 4.3 appears to be conservative in predicting the actual scaling behavior with $\kappa(P_\infty(A))$.

5.2. Least-Squares Policy Iteration

We now describe our comparison of the Least-Squares Policy Iteration (LSPI) algorithm from Lagoudakis and Parr (2003) to the model-based approaches of Dean et al. (2017). It is interesting to empirically compare the end-to-end sample complexity of model-free versus model-based methods for LQR in order to reach a specified controller cost, since our theoretical results in Section 4.2 suggest that LSPI can require more samples than the model-based approaches. We look at the same LQR instance from Dean et al., which is described by

$$A = \begin{bmatrix} 1.01 & 0.01 & 0 \\ 0.01 & 1.01 & 0.01 \\ 0 & 0.01 & 1.01 \end{bmatrix}, \quad B = I_3, \qquad (5.1)$$

with cost matrices $Q = 10^{-3} I_3$ and $R = I_3$. We consider both the discounted LQR problem with $\gamma = 0.98$ and the average cost LQR problem. The choice of $\gamma = 0.98$ ensures that the closed-loop system $A + BK$ with K the optimal discounted controller is stable. Our metric of interest is the relative error $\frac{J(K) - J_\star}{J_\star}$, where $J_\star$ is the optimal infinite-horizon cost on either the discounted or average cost objective, and $J(K)$ is the infinite-horizon cost of using the controller K in feedback with the true system (5.1).

We run our experiments as follows. We collect M independent trajectories of the system (5.1) excited by independent Gaussian noise $\mathcal{N}(0, I_3)$ of length $N = 20$. This produces a collection of MN tuples $D = \{(x_k^{(\ell)}, u_k^{(\ell)}, r_k^{(\ell)}, x_{k+1}^{(\ell)})\}_{k=1,\ell=1}^{N,M}$. We repeat this whole process 100 times. In our experiments, we will refer to the value MN as the number of timesteps, and each set D of MN tuples collected will be referred to as a trial. As in the previous experiment, we use the prefix of the data to report different values for the number of timesteps used. We now describe in more detail the different algorithms we evaluate.

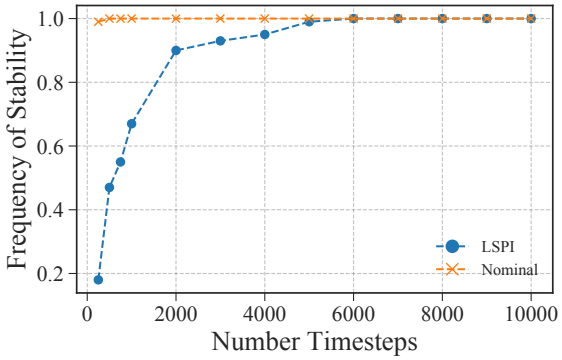

Figure 3: A comparison of how frequently (out of 100 trials) both LSPI and the nominal synthesis procedure were able to produce a controller $\widehat{K}$ such that the matrix $\sqrt{\gamma}(A + B\widehat{K})$ was stable. This condition is necessary and sufficient for the discounted infinite-horizon cost to be finite.

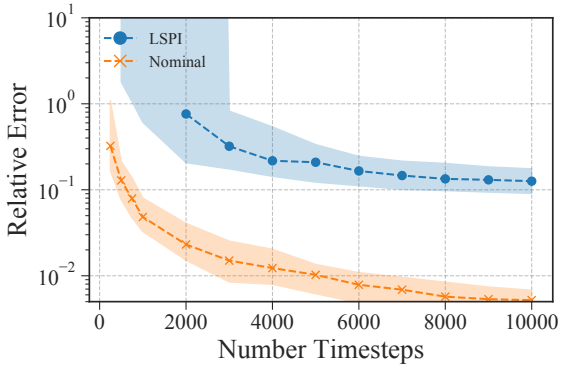

Figure 4: A comparison of the relative error of the controllers produced by both LSPI and the nominal synthesis procedure for the discounted LQR problem. The points along the dashed line denote the median cost, and the shaded region covers the 25-th to 75-th percentile out of 100 trials.

LSPI. To run LSPI, we need a starting controller K_0. The trivial controller $K_0 = 0_{3\times 3}$ is insufficient, since the matrix $\sqrt{\gamma}A$ is not stable and hence does not induce a finite Q-function. This is a drawback of LSPI; a reasonable initialization must be chosen for the algorithm to work. For the purposes of comparison, we set K_0 such that the closed loop matrix $A + BK_0 = \mathrm{diag}(0.6, 0.6, 0.6)$ and is hence a valid starting point for LSPI. Furthermore, the relative error $(J(K_0) - J_\star)/J_\star \approx 6.603$ for the discounted case and $(J(K_0) - J_\star)/J_\star \approx 4.778$ for the average cost case. When running LSPI for discounted cost (resp. average cost), if at any point we estimate a policy K_t such that $\sqrt{\gamma}(A + BK_t)$ (resp. $A + BK_t$) is not stable, we consider the algorithm as having failed and assign it a score of $+\infty$.

Nominal controller. The nominal controller works by first estimating the state-transition matrices $(\widehat{A}, \widehat{B})$ from

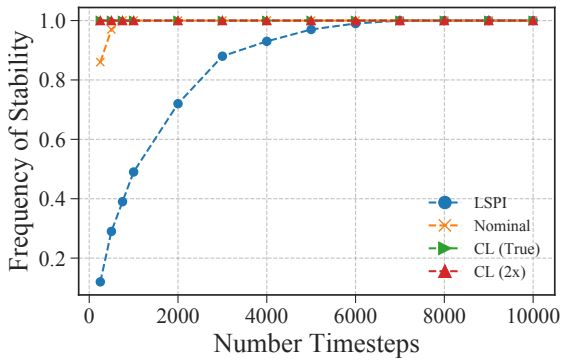

Figure 5: A comparison of how frequently (out of 100 trials) LSPI, the nominal synthesis procedure, and the common Lyapunov (CL) synthesis procedures were able to produce a controller $\widehat{K}$ such that the matrix $A + B\widehat{K}$ was stable. This condition is necessary and sufficient for the average infinite-horizon cost to be finite.

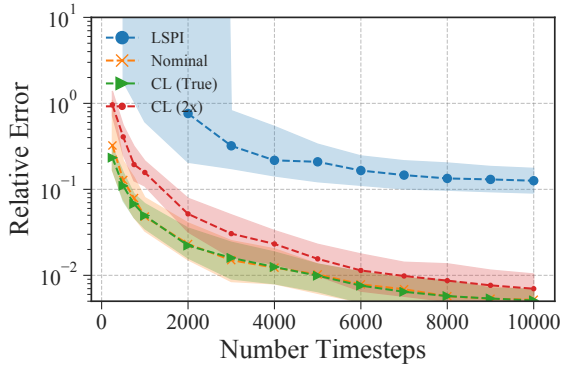

Figure 6: A comparison of the relative error of the controllers produced by LSPI, the nominal synthesis procedure, and the common Lyapunov (CL) procedures for the average cost LQR problem. The points along the dashed line denote the median cost, and the shaded region covers the 25-th to 75-th percentile out of 100 trials.

the given trajectories via ordinary least-squares. With the estimates $(\widehat{A}, \widehat{B})$, we directly solve via algebraic Ricatti equations for the optimal discounted/average cost controllers under the assumption that the dynamics are exactly $(\widehat{A}, \widehat{B})$. We then check to see if the resulting costs with the nominal controller in feedback with the true system are finite, and assign a score of $+\infty$ otherwise.

Common Lyapunov controller. The common Lyapunov synthesis procedure is developed in Dean et al. as a semidefinite (SDP) relaxation to the non-convex robust controller synthesis problem with static state-feedback. If the program succeeds, it provides a certificate that the actual closed-loop system is stable (this is not guaranteed by the nominal controller, nor LSPI). Since the formulation in Dean et al. is for the average cost setting, we only run the procedure in this setting. Because the procedure is a robust synthesis algorithm, it takes as input an upper bound on the estimation errors $\|\widehat{A} - A\| \leq \varepsilon_A$ and $\|\widehat{B} - B\| \leq \varepsilon_B$. We use both the true errors and $2\times$ the true errors as the input bounds. The former showcases the best possible performance, and the latter simulates the non-parametric bootstrap method used in Dean et al. to compute these confidence bounds; their results suggest that the bootstrap over-estimates the true errors by roughly a factor of two. We solve the resulting SDPs using cvxpy (Diamond & Boyd, 2016) with MOSEK (2015).

The results for the discounted LQR problem are shown in Figure 3 and Figure 4, and the results for the average cost LQR problem are shown in Figure 5 and Figure 6. We observe on the discounted problem that LSPI less robust and more sample inefficient than the nominal controller. In Figure 3, we observe that even with 3000 timesteps the frequency of stability for LSPI is worse than that of the nominal controller at 250 timesteps. Similarly, in Figure 4, we see that the relative error achieved by LSPI at 3000 timesteps is comparable to that achieved by the nominal controller at 250 timesteps. The qualitative differences between LSPI and the nominal controller remain the same when we move to the average cost controller. In Figure 6, we see that the nominal controller and the common Lyapunov controller given the actual error bounds perform the best, the common Lyapunov controller given $2\times$ the actual error bound performs slightly worse, and the performance of LSPI is substantially behind the rest, taking for instance over $10\times$ more samples compared to the nominal controller to achieve a relative error of 10^{-1} .

6. Conclusion

We studied the number of samples needed for the LSTD estimator to return a ε-accurate solution in relative error for the value function associated to a fixed policy π for LQR. In the process of deriving our result, we provided a concentration result for the minimum eigenvalue of a sample covariance matrix formed along the trajectory of a β-mixing stochastic process. Empirically, we demonstrated that model-free policy iteration (LSPI) requires substantially more samples on certain LQR instances than the model-based methods from Dean et al.

We hope our results encourage further investigation into the foundations of RL for continuous control problems. In particular, some interesting extensions of our work include providing an end-to-end guarantee for LSPI with noisy policy evaluations, establishing algorithmic and information-theoretic lower bounds for RL algorithms on the LQR problem, and also analyzing other widely used RL algorithms such as policy gradient (Williams, 1992) and Trust Region Policy Optimization (Schulman et al., 2015).

Acknowledgements

We thank Orianna DeMasi, Vitaly Kuznetsov, Horia Mania, Max Simchowitz, Vikas Sindhwani, Xinyan Yan, and the anonymous reviewers for many helpful comments and suggestions. Part of this work was completed when ST was interning at Google Brain, New York, NY. BR is generously supported by NSF award CCF-1359814, ONR awards N00014-14-1-0024 and N00014-17-1-2191, the DARPA Fundamental Limits of Learning (Fun LoL) Program, a Sloan Research Fellowship, and a Google Faculty Award.

References

Adamczak, R., Litvak, A. E., Pajor, A., and Tomczak-Jaegermann, N. Sharp bounds on the rate of convergence of empirical covariance matrix. *C. R., Math., Acad. Sci. Paris*, 349, 2011.

Agarwal, A. and Duchi, J. C. The Generalization Ability of Online Algorithms for Dependent Data. *IEEE Transactions on Information Theory*, 59(1), 2013.

Antos, A., Szepesvári, C., and Munos, R. Learning near-optimal policies with Bellman-residual minimization based fitted policy iteration and a single sample path. *Machine Learning*, 71(1), 2008.

Bertsekas, D. P. *Dynamic Programming and Optimal Control, Vol. II*. 2007.

Boyan, J. Least-Squares Temporal Difference Learning. In *International Conference on Machine Learning*, 1999.

Bradtke, S. J. Reinforcement Learning Applied to Linear Quadratic Regulation. In *Neural Information Processing Systems*, 1993.

Bradtke, S. J. *Incremental Dynamic Programming for On-Line Adaptive Optimal Control*. PhD thesis, University of Massachusetts Amherst, 1994.

Bradtke, S. J. and Barto, A. G. Linear Least-Squares Algorithms for Temporal Difference Learning. *Machine Learning*, 22, 1996.

Dean, S., Mania, H., Matni, N., Recht, B., and Tu, S. On the Sample Complexity of the Linear Quadratic Regulator. *arXiv:1710.01688*, 2017.

Diamond, S. and Boyd, S. CVXPY: A Python-embedded modeling language for convex optimization. *Journal of Machine Learning Research*, 17(83), 2016.

Farahmand, A., Ghavamzadeh, M., Szepesvári, C., and Mannor, S. Regularized Policy Iteration with Nonparametric Function Spaces. *Journal of Machine Learning Research*, 17(139), 2016.

Goldenshluger, A. and Zeevi, A. Nonasymptotic bounds for autoregressive time series modeling. *The Annals of Statistics*, 29(2), 2001.

Gu, S., Lillicrap, T., Sutskever, I., and Levine, S. Continuous Deep Q-Learning with Model-based Acceleration. In *International Conference on Machine Learning*, 2016.

Gu, S., Lillicrap, T., Ghahramani, Z., Turner, R. E., and Levine, S. Q-Prop: Sample-Efficient Policy Gradient with An Off-Policy Critic. In *International Conference on Learning Representations*, 2017.

Jonas, E., Pu, Q., Venkataraman, S., Stoica, I., and Recht, B. Occupy the Cloud: Distributed Computing for the 99%. In *ACM Symposium on Cloud Computing*, 2017.

Kober, J., Bagnell, J. A., and Peters, J. Reinforcement Learning in Robotics: A Survey. *The International Journal of Robotics Research*, 32(11), 2013.

Koltchinskii, V. and Mendelson, S. Bounding the smallest singular value of a random matrix without concentration. *arXiv:1312.3580*, 2013.

Krishnan, S., Fox, R., Stoica, I., and Goldberg, K. DDCO: Discovery of Deep Continuous Options for Robot Learning from Demonstrations. In *Conference on Robot Learning*, 2017.

Kuznetsov, V. and Mohri, M. Learning Theory and Algorithms for Forecasting Non-Stationary Time Series. In *Neural Information Processing Systems*, 2015.

Kuznetsov, V. and Mohri, M. Time Series Prediction and Online Learning. In *Conference on Learning Theory*, 2016.

Lagoudakis, M. G. and Parr, R. Least-Squares Policy Iteration. *Journal of Machine Learning Research*, 4, 2003.

Lazaric, A., Ghavamzadeh, M., and Munos, R. Finite-Sample Analysis of Least-Squares Policy Iteration. *Journal of Machine Learning Research*, 13, 2012.

Levine, S. and Koltun, V. Learning Complex Neural Network Policies with Trajectory Optimization. In *International Conference on Machine Learning*, 2014.

Levine, S., Wagener, N., and Abbeel, P. Learning Contact-Rich Manipulation Skills with Guided Policy Search. In *International Conference on Robotics and Automation*, 2015.

Levine, S., Finn, C., Darrell, T., and Abbeel, P. End-to-End Training of Deep Visuomotor Policies. *Journal of Machine Learning Research*, 17(39), 2016a.

Levine, S., Pastor, P., Krizhevsky, A., and Quillen, D. Learning Hand-Eye Coordination for Robotic Grasping with Deep Learning and Large-Scale Data Collection. *arXiv:1603.02199*, 2016b.

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

Liu, B., Mahadevan, S., and Liu, J. Regularized Off-Policy TD-Learning. In *Neural Information Processing Systems*, 2012.

Liu, B., Liu, J., Ghavamzadeh, M., Mahadevan, S., and Petrik, M. Finite-Sample Analysis of Proximal Gradient TD Algorithms. In *Uncertainty in Artificial Intelligence*, 2015.

McDonald, D. J., Shalizi, C. R., and Schervish, M. Nonparametric Risk Bounds for Time-Series Forecasting. *Journal of Machine Learning Research*, 18(32), 2017.

Mendelson, S. and Paouris, G. On the singular values of random matrices. *Journal of the European Mathematical Society*, 16, 2014.

Mohri, M. and Rostamizadeh, A. Rademacher Complexity Bounds for Non-I.I.D. Processes. In *Neural Information Processing Systems*, 2008.

Mohri, M. and Rostamizadeh, A. Stability Bounds for Stationary ϕ-mixing and β-mixing Processes. *Journal of Machine Learning Research*, 11, 2010.

Mokkadem, A. Mixing properties of ARMA processes. *Stochastic Processes and their Applications*, 29(2), 1988.

MOSEK ApS. *The MOSEK optimization toolbox for MATLAB manual. Version 7.1 (Revision 28).*, 2015. URL http://docs.mosek.com/7.1/toolbox/index.html.

Prashanth, L., Korda, N., and Munos, R. Fast LSTD using stochastic approximation: Finite time analysis and application to traffic control. *arXiv:1306.2557*, 2014.

Rudelson, M. and Vershynin, R. Smallest Singular Value of a Random Rectangular Matrix. *Communications on Pure and Applied Mathematics*, 62(12), 2009.

Schaul, T., Quan, J., Antonoglou, I., and Silver, D. Prioritized Experience Replay. In *International Conference on Learning Representations*, 2016.

Schulman, J., Levine, S., Moritz, P., Jordan, M. I., and Abbeel, P. Trust Region Policy Optimization. In *International Conference on Machine Learning*, 2015.

Schulman, J., Moritz, P., Levine, S., Jordan, M., and Abbeel, P. High-Dimensional Continuous Control Using Generalized Advantage Estimation. In *International Conference on Learning Representations*, 2016.

Srivastava, N. and Vershynin, R. Covariance estimation for distributions with $2 + \varepsilon$ moments. *The Annals of Probability*, 41(5), 2013.

Sutton, R. S. and Barto, A. G. Reinforcement Learning. 1998.

Tedrake, R., Zhang, T. W., and Seung, H. S. Stochastic Policy Gradient Reinforcement Learning on a Simple 3D Biped. In *International Conference on Intelligent Robots and Systems*, 2004.

Tsitsiklis, J. N. and Van Roy, B. An Analysis of Temporal-Difference Learning with Function Approximation. *IEEE Transactions on Automatic Control*, 42(5), 1997.

Tu, S. and Recht, B. Least-Squares Temporal Difference Learning for the Linear Quadratic Regulator. *arXiv:1712.08642*, 2017.

Vershynin, R. Introduction to the non-asymptotic analysis of random matrices. *arXiv:1011.3027*, 2011.

Williams, R. J. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8, 1992.

Yu, B. Rates of Convergence for Empirical Processes of Stationary Mixing Sequences. *The Annals of Probability*, 22(1), 1994.

Yu, H. and Bertsekas, D. P. Convergence Results for Some Temporal Difference Methods Based on Least Squares. *IEEE Transactions on Automatic Control*, 54(7), 2009.

Zhou, K., Doyle, J. C., and Glover, K. *Robust and Optimal Control*. 1995.