Finite-Sample Analysis of LSTD

Alessandro Lazaric ALESSANDRO.LAZARIC@INRIA.FR
Mohammad Ghavamzadeh MOHAMMAD.GHAVAMZADEH@INRIA.FR
Rémi Munos REMI.MUNOS@INRIA.FR
INRIA Lille - Nord Europe, Team SequeL, FRANCE

Abstract

In this paper we consider the problem of policy evaluation in reinforcement learning, i.e., learning the value function of a fixed policy, using the least-squares temporal-difference (LSTD) learning algorithm. We report a finite-sample analysis of LSTD. We first derive a bound on the performance of the LSTD solution evaluated at the states generated by the Markov chain and used by the algorithm to learn an estimate of the value function. This result is general in the sense that no assumption is made on the existence of a stationary distribution for the Markov chain. We then derive generalization bounds in the case when the Markov chain possesses a stationary distribution and is β-mixing.

1. Introduction

Least-squares temporal-difference (LSTD) learning (Bradtke & Barto, 1996; Boyan, 1999) is a widely used algorithm for prediction in general, and in the context of reinforcement learning (RL), for learning the value function V^π of a given policy π. LSTD has been successfully applied to a number of problems especially after the development of the least-squares policy iteration (LSPI) algorithm (Lagoudakis & Parr, 2003), which extends LSTD to control problems. More precisely, LSTD computes the fixed point of the operator $\Pi\mathcal{T}$, where $\mathcal{T}$ is the Bellman operator and Π is the projection operator in a linear function space. Although LSTD and LSPI have been widely used in the RL community, a finite-sample analysis of LSTD (i.e., performance bounds in terms of the number of samples) is still lacking.

Most of the theoretical work analyzing LSTD have been focused on the model-based case, where explicit models of the reward function and the dynamics are available. In particular, Tsitsiklis & Van Roy (1997) showed that the distance between the LSTD solution and the value function V^π is bounded by the distance between V^π and its closest approximation in the linear space, multiplied by a constant which increases as the discount factor approaches 1. In this bound, it is assumed that the Markov chain possesses a stationary distribution ρ^π and the distances are measured according to ρ^π. Bertsekas (2001) reported a similar analysis for the empirical version of LSTD. His analysis reveals a critical dependency on the inverse of the smallest eigenvalue of the LSTD's A matrix (note that the LSTD solution is obtained by solving the system of linear equations $Ax = b$). Nonetheless, Bertsekas (2001) does not provide a finite-sample analysis of the algorithm. On the other hand, Antos et al. (2008) analyzed the modified Bellman residual (MBR) minimization algorithm for a finite number of samples, bounded function spaces, and a μ-norm that might be different from the norm induced by ρ^π. Although MBR minimization was shown to reduce to LSTD in case of linear spaces, it is not straightforward how the finite-sample bounds derived by Antos et al. (2008) can be extended to unbounded linear spaces considered by LSTD.

In this paper, we report a finite-sample analysis of LSTD. To the best of our knowledge, this is the first complete finite-sample analysis of this widely used algorithm. Our analysis is for a specific implementation of LSTD that we call *pathwise LSTD*. Pathwise LSTD has two specific characteristics: **1)** it takes a single trajectory generated by the Markov chain induced by policy π as input, and **2)** it uses the pathwise Bellman operator (will be precisely defined later), which is defined to be a contraction w.r.t. the empirical norm. We first derive a bound on the performance of the pathwise LSTD solution for a setting that we call *Markov design*. In this setting, the performance is evaluated at the points used by the algorithm to learn an estimate of V^π. This bound is general in the sense that no as-

Appearing in *Proceedings of the 27th International Conference on Machine Learning*, Haifa, Israel, 2010.

sumption is made on the existence of a stationary distribution for the Markov chain. Then, in the case the Markov chain admits a stationary distribution ρ^π and is β-mixing, we derive generalization bounds w.r.t. the norm induced by ρ^π.

Besides providing a full finite-sample analysis of LSTD, the major insights gained by the analysis in the paper can be summarized as follows. The first result is about the existence of the LSTD solution and its performance. In Theorem 1 we show that with a slight modification of the empirical Bellman operator $\widehat{\mathcal{T}}$ (leading to the definition of pathwise LSTD), the operator $\widehat{\Pi}\widehat{\mathcal{T}}$ (where $\widehat{\Pi}$ is an empirical projection operator) has always a fixed point $\hat{v}$ even when the sample-based Gram matrix is not invertible and the Markov chain does not admit a stationary distribution. In this very general setting, it is still possible to derive a bound for the performance of $\hat{v}$ evaluated on the states of the trajectory, and an analysis of the bound reveals a critical dependency on the smallest strictly positive eigenvalue ν_n of the sample-based Gram matrix. Then, in the case in which the Markov chain has a stationary distribution ρ^π, it is possible to relate the value of ν_n to the smallest eigenvalue of the Gram matrix defined according to ρ^π. Furthermore, it is possible to generalize the previous performance bound over the entire state space under the measure ρ^π, when the samples are drawn from a stationary β-mixing process (Theorem 2). It is important to note that the asymptotic bound obtained by taking the number of samples, n, to infinity is equal (up to constants) to the bound in Tsitsiklis & Van Roy (1997) for model-based LSTD. Furthermore, a comparison with the bounds in Antos et al. (2008) shows that we successfully leverage on the specific setting of LSTD: **1)** the space of functions is linear, and **2)** the distribution used to evaluate the performance is the stationary distribution of the Markov chain induced by the policy. In particular, we obtain a better bound both in terms of estimation error, a rate of order $O(1/n)$ instead of $O(1/\sqrt{n})$ for the squared error, and in terms of approximation error, the minimal distance between the value function V^π and the space $\mathcal{F}$ instead of the inherent Bellman errors of $\mathcal{F}$. Finally, the extension in Theorem 3 to the case in which the samples belong to a trajectory generated by a fast mixing Markov chain shows that it is possible to achieve the same performance as in the case of stationary β-mixing processes.

2. Preliminaries

For a measurable space with domain $\mathcal{X}$, we let $\mathcal{S}(\mathcal{X})$ and $\mathcal{B}(\mathcal{X}; L)$ denote the set of probability measures over $\mathcal{X}$, and the space of bounded measurable functions with domain $\mathcal{X}$ and bound $0 < L < \infty$, respectively. For a measure $\rho \in \mathcal{S}(\mathcal{X})$ and a measurable function $f : \mathcal{X} \to \mathbb{R}$, we define the $\ell_2(\rho)$-norm of f, $||f||_\rho$, and for a set of n states $X_1, \dots, X_n \in \mathcal{X}$, we define the empirical norm $||f||_n$ as

$$||f||_\rho^2 = \int f(x)^2 \rho(dx) \qquad \text{and} \qquad ||f||_n^2 = \frac{1}{n}\sum_{t=1}^n f(X_t)^2.$$

The supremum norm of f, $||f||_\infty$, is defined as $||f||_\infty = \sup_{x\in\mathcal{X}} |f(x)|$.

We consider the standard RL framework (Sutton & Barto, 1998) in which a learning agent interacts with a stochastic environment by following a policy π and this interaction is modeled as a discrete-time discounted Markov chain (MC). A discounted MC is a tuple $\mathcal{M}^\pi = \langle \mathcal{X}, R^\pi, P^\pi, \gamma \rangle$, where the state space $\mathcal{X}$ is a subset of a Euclidean space, the reward function $R^\pi : \mathcal{X} \to \mathbb{R}$ is uniformly bounded by $R_{\max}$, the transition kernel P^π is such that for all $x \in \mathcal{X}$, $P^\pi(\cdot|x)$ is a distribution over $\mathcal{X}$, and $\gamma \in (0, 1)$ is a discount factor. The value function of a policy π, V^π, is the unique fixed-point of the Bellman operator $\mathcal{T}^\pi : \mathcal{B}(\mathcal{X}; V_{\max} = \frac{R_{\max}}{1-\gamma}) \to \mathcal{B}(\mathcal{X}; V_{\max})$ defined by[1]

$$(\mathcal{T}^\pi V)(x) = R^\pi(x) + \gamma \int_{\mathcal{X}} P^\pi(dy|x) V(y).$$

To approximate the value function V, we use a linear approximation architecture with parameters $\alpha \in \mathbb{R}^d$ and basis functions $\varphi_j \in \mathcal{B}(\mathcal{X}; L)$, $j = 1, \dots, d$. We denote by $\phi : \mathcal{X} \to \mathbb{R}^d$, $\phi(\cdot) = \big(\varphi_1(\cdot), \dots, \varphi_d(\cdot)\big)^\top$ the feature vector, and by $\mathcal{F}$ the linear function space spanned by the basis functions φ_j. Thus $\mathcal{F} = \{f_\alpha,\ \alpha \in \mathbb{R}^d\}$, where $f_\alpha(\cdot) = \phi(\cdot)^\top \alpha$.

Let $(X_1, \dots, X_n)$ be a sample path (or trajectory) of size n generated by the Markov chain $\mathcal{M}$. Let $v \in \mathbb{R}^n$ and $r \in \mathbb{R}^n$ such that $v_t = V(X_t)$ and $r_t = R(X_t)$ be the value vector and the reward vector, respectively. Also, let $\Phi = [\phi(X_1)^\top; \dots; \phi(X_n)^\top]$ be the feature matrix defined at the states, and $\mathcal{F}_n = \{\Phi\alpha,\ \alpha \in \mathbb{R}^d\} \subset \mathbb{R}^n$ be the corresponding vector space. We denote by $\widehat{\Pi} : \mathbb{R}^n \to \mathcal{F}_n$ the orthogonal projection onto $\mathcal{F}_n$, defined as $\widehat{\Pi} y = \arg\min_{z\in\mathcal{F}_n} ||y - z||_n$, where $||y||_n^2 = \frac{1}{n}\sum_{t=1}^n y_t^2$. Note that the orthogonal projection $\widehat{\Pi} y$ for any $y \in \mathbb{R}^n$ exists and is unique.

3. Pathwise LSTD

Pathwise LSTD is a version of LSTD which takes as input a single path $X_1, \dots, X_n$ and returns the fixed-

[1] To simplify the notation, we remove the dependency to the policy π and use $\mathcal{M}$, R, P, V, and $\mathcal{T}$ instead of $\mathcal{M}^\pi$, R^π, P^π, V^π, and $\mathcal{T}^\pi$ throughout the paper.

point of the empirical operator $\widehat{\Pi}\widehat{\mathcal{T}}$, where $\widehat{\mathcal{T}} : \mathbb{R}^n \to \mathbb{R}^n$ is the *pathwise Bellman operator* defined as

$$(\widehat{\mathcal{T}}y)_t = \begin{cases} r_t + \gamma y_{t+1} & 1 \le t < n, \\ r_t & t = n. \end{cases}$$

Note that by defining the operator $\widehat{P} : \mathbb{R}^n \to \mathbb{R}^n$ as $(\widehat{P}y)_t = y_{t+1}$ for $1 \le t < n$ and $(\widehat{P}y)_n = 0$, we have $\widehat{\mathcal{T}}y = r + \gamma\widehat{P}y$. The motivation for using the pathwise Bellman operator is that it is γ-contraction in ℓ_2-norm, i.e., for any $y, z \in \mathbb{R}^n$, we have

$$||\widehat{\mathcal{T}}y - \widehat{\mathcal{T}}z||_n^2 = ||\gamma\widehat{P}(y-z)||_n^2 \le \gamma^2||y-z||_n^2 .$$

Moreover, it can be shown that the orthogonal projection $\widehat{\Pi}$ is non-expansive: since $||\widehat{\Pi}y||_n^2 = \langle y, \widehat{\Pi}y\rangle_n \le ||y||_n||\widehat{\Pi}y||_n$, using the Cauchy-Schwarz inequality we obtain $||\widehat{\Pi}y||_n \le ||y||_n$. Therefore, from Banach fixed point theorem, there exists a unique fixed-point $\hat{v}$ of the mapping $\widehat{\Pi}\widehat{\mathcal{T}}$, i.e., $\hat{v} = \widehat{\Pi}\widehat{\mathcal{T}}\hat{v}$. We call $\hat{v}$ the pathwise LSTD solution. Note that the unicity of $\hat{v}$ does not imply that there exists a unique parameter $\hat{\alpha}$ such that $\hat{v} = \Phi\hat{\alpha}$.

4. Markov Design Bound

Theorem 1. *Let $X_1, \dots, X_n$ be a trajectory of the Markov chain, and $v, \hat{v} \in \mathbb{R}^n$ be the vectors whose components are the value function and the pathwise LSTD solution at $\{X_i\}_{i=1}^n$, respectively. Then with probability $1-\delta$, where the probability is w.r.t. the random trajectory, we have*

$$||\hat{v} - v||_n \le \frac{1}{1-\gamma}\Big[||v - \widehat{\Pi}v||_n + \gamma V_{\max} L\sqrt{\frac{d}{\nu_n}}\Big(\sqrt{\frac{8\log(2d/\delta)}{n}} + \frac{1}{n}\Big)\Big], \quad (1)$$

where the random variable ν_n is the smallest strictly-positive eigenvalue of the sample-based Gram matrix $\frac{1}{n}\Phi^\top\Phi$.

Remark 1 When the eigenvalues of the sample-based Gram matrix $\frac{1}{n}\Phi^\top\Phi$ are non-zero, $\Phi^\top\Phi$ is invertible, and thus, $\widehat{\Pi} = \Phi(\Phi^\top\Phi)^{-1}\Phi^\top$. In this case, the unicity of $\hat{v}$ implies the unicity of $\hat{\alpha}$ since

$$\hat{v} = \Phi\alpha \implies \Phi^\top\hat{v} = \Phi^\top\Phi\alpha \implies \hat{\alpha} = (\Phi^\top\Phi)^{-1}\Phi^\top\hat{v}. \quad (2)$$

Since $\hat{v}$ is the unique fixed-point of $\widehat{\Pi}\widehat{\mathcal{T}}$, it can be replaced by $\widehat{\Pi}\widehat{\mathcal{T}}\hat{v}$ in Eq. 2. Using the definitions of $\widehat{\Pi}$ and $\widehat{\mathcal{T}}$, we obtain $\Phi^\top(I - \gamma\widehat{P})\Phi\hat{\alpha} = \Phi^\top r$. By defining $A = \Phi^\top(I-\gamma\widehat{P})\Phi$ and $b = \Phi^\top r$, $\hat{\alpha}$ can be seen as the unique solution of the $d \times d$ system of linear equations $A\alpha = b$.

Remark 2 Note that in case there exists a constant $\nu > 0$, such that with probability $1-\delta'$ all the eigenvalues of the sample-based Gram matrix are lower bounded by ν, Eq. 1 (with ν_n replaced by ν) holds with probability at least $1-(\delta+\delta')$.

Remark 3 When the sample-based Gram matrix $\frac{1}{n}\Phi^\top\Phi$ is not invertible, the unicity of $\hat{v}$ does not imply the unicity of the solution to the system $\Phi\alpha = \hat{v}$. However, since $\hat{v}$ is the unique fixed point of $\widehat{\Pi}\widehat{\mathcal{T}}$, the vector $\hat{v} - \widehat{\mathcal{T}}\hat{v}$ is perpendicular to the space $\mathcal{F}_n$, and thus, $\Phi^\top(\hat{v} - \widehat{\mathcal{T}}\hat{v}) = 0$. By replacing $\hat{v}$ with $\Phi\alpha$, we obtain $\Phi^\top\Phi\alpha = \Phi^\top(r + \gamma\widehat{P}\Phi\alpha)$ and then $\Phi^\top(I-\gamma\widehat{P})\Phi\hat{\alpha} = \Phi^\top r$. Therefore, we still have the same system of equations $A\alpha = b$ as in Remark 1, with the exact same A and b, but now the system may have many solutions.[2] Among all possible solutions, one may choose the one with minimal norm: $\hat{\alpha} = A^+ b$, where A^+ is the Moore-Penrose pseudo-inverse of A.

Remark 4 Theorem 1 provides a bound without any reference to the stationary distribution of the Markov chain. In fact, the bound of Eq. 1 holds even when the chain does not possess a stationary distribution. For example, consider a Markov chain on the real line where the transitions always move the states to the right, i.e., $p(X_{t+1} \in dy | X_t = x) = 0$ for $y \le x$. For simplicity assume that the value function V is bounded and belongs to $\mathcal{F}$. This Markov chain is not recurrent, and thus, does not have a stationary distribution. We also assume that the feature vectors $\phi(X_1), \dots, \phi(X_n)$ are sufficiently independent, so that the eigenvalues of $\frac{1}{n}\Phi^\top\Phi$ are greater than $\nu > 0$. Then according to Theorem 1, pathwise LSTD is able to estimate the value function at the states at a rate $O(1/\sqrt{n})$. This may seem surprising because at each state X_t the algorithm is only provided with a noisy estimation of the expected value of the next state. However, the estimates are unbiased conditioned on the current state, and we will see in the proof that using a concentration inequality for martingale, pathwise LSTD is able to learn a good estimate of the value function at a state X_t using noisy pieces of information at other states that may be far away from X_t. In other words, learning the value function at a given state does not require making an average over many samples close to that state. This implies that LSTD does not require the Markov chain to possess a stationary distribution.

Remark 5 The most critical part of the bound in Eq. 1 is the inverse dependency on the smallest positive eigenvalue ν_n. A similar dependency is shown in the LSTD analysis of Bertsekas (2001). The main

[2] Note that since the fixed point $\hat{v}$ exists, this system always has at least one solution.

difference is that here we have a more complete finite-sample analysis with an explicit dependency on the number of samples and the other characteristic parameters of the problem. Furthermore, if the Markov chain admits a stationary distribution ρ, we are able to relate the existence of the LSTD solution to the smallest eigenvalue of the Gram matrix defined according to ρ (see Section 5.1).

In order to prove Theorem 1, we first introduce the model of regression with *Markov design* and then state and prove a Lemma about this model.

Definition 1. *The model of regression with **Markov design** is a regression problem where the data $(X_t, Y_t)_{1\le t\le n}$ are generated according to the following model: $X_1, \dots, X_n$ is a sample path generated by a Markov chain, $Y_t = f(X_t) + \xi_t$, where f is the target function, and the noise term ξ_t is a random variable which is adapted to the filtration generated by $X_1, \dots, X_{t+1}$ and is such that*

$$|\xi_t| \le C, \quad \text{and} \quad \mathbb{E}[\xi_t | X_1, \dots, X_t] = 0. \tag{3}$$

Lemma 1 (Regression bound for the Markov design setting)**.** *We consider the model of regression with Markov design from Definition 1. Let $\hat{w} \in \mathcal{F}_n$ be the least-squares estimate of the (noisy) values $Y = \{Y_t\}_1^n$, i.e., $\hat{w} = \widehat{\Pi} Y$, and $w \in \mathcal{F}_n$ be the least-squares estimate of the (noiseless) values $Z = \{Z_t\}_1^n = \{f(X_t)\}_1^n$, i.e., $w = \widehat{\Pi} Z$. Then for any $\delta > 0$, with probability at least $1-\delta$, where the probability is w.r.t. the random sample path $X_1, \dots, X_n$, we have*

$$||\hat{w} - w||_n \le CL \sqrt{\frac{2d \log(2d/\delta)}{n \nu_n}}, \tag{4}$$

where ν_n is the smallest strictly-positive eigenvalue of the sample-based Gram matrix $\frac{1}{n}\Phi^\top \Phi$.

Proof of Lemma 1. We define $\xi \in \mathbb{R}^n$ to be the vector with components ξ_t, and $\hat{\xi} = \hat{w} - w = \widehat{\Pi}(Y - Z) = \widehat{\Pi}\xi$. Since the projection is orthogonal we have $\langle \hat{\xi}, \xi \rangle_n = ||\hat{\xi}||_n^2$. Since $\hat{\xi} \in \mathcal{F}_n$, there exists at least one $\alpha \in \mathbb{R}^d$ such that $\hat{\xi} = \Phi\alpha$, so by Cauchy-Schwarz inequality we have

$$\begin{aligned} ||\hat{\xi}||_n^2 = \langle \hat{\xi}, \xi \rangle_n &= \frac{1}{n} \sum_{i=1}^d \alpha_i \sum_{t=1}^n \xi_t \varphi_i(X_t) \\ &\le \frac{1}{n} ||\alpha||_2 \left[\sum_{i=1}^d \Big(\sum_{t=1}^n \xi_t \varphi_i(X_t) \Big)^2 \right]^{1/2}. \end{aligned} \tag{5}$$

Now among the vectors α such that $\hat{\xi} = \Phi\alpha$, we define $\hat{\alpha}$ to be the one with minimal ℓ_2-norm, i.e., $\hat{\alpha} = \Phi^+ \hat{\xi}$. Let K denote the null space of Φ, which is also the null space of $\frac{1}{n}\Phi^\top\Phi$. Then $\hat{\alpha}$ can be decomposed as $\hat{\alpha} = \hat{\alpha}_K + \hat{\alpha}_{K^\perp}$, where $\hat{\alpha}_K \in K$ and $\hat{\alpha}_{K^\perp} \in K^\perp$, and because the decomposition is orthogonal, we have $||\hat{\alpha}||_2^2 = ||\hat{\alpha}_K||_2^2 + ||\hat{\alpha}_{K^\perp}||_2^2$. Since $\hat{\alpha}$ is of minimal norm among all the vectors α such that $\hat{\xi} = \Phi\alpha$, its component in K must be zero, thus $\hat{\alpha} \in K^\perp$.

The Gram matrix $\frac{1}{n}\Phi^\top\Phi$ is positive-semidefinite, thus its eigenvectors corresponding to zero eigenvalues generate K and the other eigenvectors generate its orthogonal complement $K^\perp$. Therefore, from the assumption that the smallest strictly-positive eigenvalue of $\frac{1}{n}\Phi^\top\Phi$ is ν_n, we deduce that since $\hat{\alpha} \in K^\perp$,

$$||\hat{\xi}||_n^2 = \frac{1}{n} \hat{\alpha}^\top \Phi^\top \Phi \hat{\alpha} \ge \nu_n \hat{\alpha}^\top \hat{\alpha} = \nu_n ||\hat{\alpha}||_2^2. \tag{6}$$

By using the result of Eq. 6 in Eq. 5, we obtain

$$||\hat{\xi}||_n \le \frac{1}{n\sqrt{\nu_n}} \left[\sum_{i=1}^d \Big(\sum_{t=1}^n \xi_t \varphi_i(X_t) \Big)^2 \right]^{1/2}. \tag{7}$$

Now, from Eq. 3, we have that

$$\mathbb{E}[\xi_t \varphi_i(X_t) | X_1, \dots, X_t] = \varphi_i(X_t) \mathbb{E}[\xi_t | X_1, \dots, X_t] = 0,$$

thus $\xi_t \varphi_i(X_t)$ is a martingale difference sequence w.r.t. the filtration generated by the Markov chain, and one may apply Azuma's inequality to deduce that with probability $1-\delta$,

$$\Big| \sum_{t=1}^n \xi_t \varphi_i(X_t) \Big| \le CL \sqrt{2n \log(2/\delta)}\,.$$

By a union bound over all features, we have that with probability $1-\delta$, for all $i = 1 \dots d$,

$$\Big| \sum_{t=1}^n \xi_t \varphi_i(X_t) \Big| \le CL \sqrt{2n \log(2d/\delta)}\,. \tag{8}$$

The results follows by combining Eq. 8 with Eq. 7. □

Remark about Lemma 1 In the Markov design model considered in this lemma, states $\{X_t\}_1^n$ are random variables generated according to the Markov chain and the noise terms ξ_t may depend on the next state X_{t+1} (but should be centered conditioned on the past $X_1, \dots, X_t$). This lemma will be used in order to prove Theorem 1, where we replace the target function f with the value function V, and the noise term ξ_t with the temporal difference $r(X_t) + \gamma V(X_{t+1}) - V(X_t)$.

Note that this lemma is an extension of the bound for the model of regression with deterministic design in which the states $\{X_t\}_1^n$ are fixed and the noise terms, ξ_t's, are independent. In the setting of deterministic design, usual concentration results provide high probability bounds similar to Eq. 4, but without the dependence on ν_n. An open question is whether it is

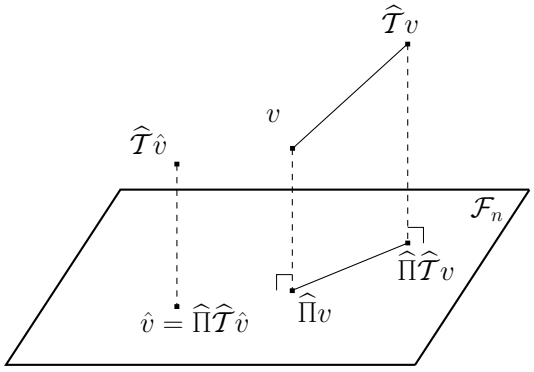

Figure 1. This figure represents the space $\mathbb{R}^n$, the linear vector subspace $\mathcal{F}_n$ and some vectors used in the proof of Theorem 1.

possible to remove ν_n in the bound for the Markov design regression setting.

Proof of Theorem 1. **Step 1:** Using the triangle inequality, we have (see Figure 1)

$$||\hat{v}-v||_n \le ||\hat{v}-\widehat{\Pi}\widehat{\mathcal{T}}v||_n + ||\widehat{\Pi}\widehat{\mathcal{T}}v-\widehat{\Pi}v||_n + ||\widehat{\Pi}v-v||_n. \quad (9)$$

From the γ-contraction of $\widehat{\Pi}\widehat{\mathcal{T}}$ mapping and the fact that $\hat{v}$ is its unique fixed point, we obtain

$$||\hat{v}-\widehat{\Pi}\widehat{\mathcal{T}}v||_n = ||\widehat{\Pi}\widehat{\mathcal{T}}\hat{v}-\widehat{\Pi}\widehat{\mathcal{T}}v||_n \le \gamma||\hat{v}-v||_n, \quad (10)$$

Thus from Eq. 9 and 10, we have

$$||\hat{v}-v||_n \le \frac{1}{1-\gamma}\Big[||\widehat{\Pi}v-v||_n + ||\widehat{\Pi}\widehat{\mathcal{T}}v-\widehat{\Pi}v||_n\Big]. \quad (11)$$

Step 2: We now provide a high probability bound on $||\widehat{\Pi}\widehat{\mathcal{T}}v-\widehat{\Pi}v||_n$. This is a consequence of Lemma 1 applied to the vectors $Y = \widehat{\mathcal{T}}v$ and $Z = v$. Since v is the value function at the points $\{X_t\}_1^n$, from the definition of the pathwise Bellman operator, we have that for $1 \le t \le n-1$,

$$\begin{aligned}\xi_t &= y_t - v_t = r(X_t) + \gamma V(X_{t+1}) - V(X_t)\\ &= \gamma\big[V(X_{t+1}) - \int P(dy|X_t)V(y)\big],\end{aligned}$$

and $\xi_n = y_n - v_n = -\gamma\int P(dy|X_n)V(y)$. Thus, Eq. 3 holds for $1 \le t \le n-1$. Here we may choose $C = 2\gamma V_{\max}$ for a bound on ξ_t, $1 \le t \le n-1$, and $C = \gamma V_{\max}$ for a bound on ξ_n. Azuma's inequality may only be applied to the sequence of $n-1$ terms (the n-th term adds a contribution to the bound), thus instead of Eq. 8, we obtain with probability $1-\delta$

$$\Big|\sum_{t=1}^{n}\xi_t\varphi_i(X_t)\Big| \le \gamma V_{\max}L\big(2\sqrt{2n\log(2d/\delta)}+1\big),$$

for all $1 \le i \le d$. Combining with Eq. 7, we deduce that with probability $1-\delta$, we have

$$||\widehat{\Pi}\widehat{\mathcal{T}}v-\widehat{\Pi}v||_n \le \gamma V_{\max}L\sqrt{\frac{d}{\nu_n}}\Big(\sqrt{\frac{8\log(2d/\delta)}{n}}+\frac{1}{n}\Big), \quad (12)$$

The claim follows by combining Eq. 12 and 11. □

Remark 6 In addition to Eq. 1, one may easily deduce a tighter bound (when γ is close to 1):

$$\begin{aligned}||\hat{v}-v||_n \le{}& \frac{1}{\sqrt{1-\gamma^2}}||v-\widehat{\Pi}v||_n\\ &+\frac{1}{1-\gamma}\Big[\gamma V_{\max}L\sqrt{\frac{d}{\nu_n}}\Big(\sqrt{\frac{8\log(2d/\delta)}{n}}+\frac{1}{n}\Big)\Big]\end{aligned}$$

by using Pytagora's Theorem in Step 1, i.e., $||\hat{v}-v||_n^2 \le (||\hat{v}-\widehat{\Pi}\widehat{\mathcal{T}}v||_n + ||\widehat{\Pi}\widehat{\mathcal{T}}v-\widehat{\Pi}v||_n)^2 + ||\widehat{\Pi}v-v||_n^2$ instead of Eq. 9.

5. Generalization Bounds

The generality of Theorem 1 comes at the cost that the performance is evaluated only at the states visited by the Markov chain. The reason is that no assumption about the existence of the stationary distribution of the Markov chain is made. However in many problems of interest, the Markov chain has a stationary distribution ρ, and thus, the performance can be generalized to the whole state space under the measure ρ. Moreover, if ρ exists, it is possible to derive a condition for the existence of the pathwise LSTD solution depending on the number of samples and the smallest eigenvalue of the Gram matrix defined according to the stationary distribution ρ ; $G \in \mathbb{R}^{d\times d}$, $G_{ij} = \int \phi_i(x)\phi_j(x)\rho(dx)$. In this section, we assume that the Markov chain $\mathcal{M}$ is exponentially fast β-mixing with parameters $\bar{\beta}, b, \kappa$, i.e., its β-mixing coefficients satisfy $\beta_i \le \bar{\beta}\exp(-bi^\kappa)$ (see e.g., Sections 7.2 and 7.3 in Lazaric et al. 2010 for a more detailed definition of β-mixing processes).

Before stating the main results of this section, we introduce some notation. If ρ is the stationary distribution of the Markov chain, we define the orthogonal projection operator $\Pi : \mathcal{B}(\mathcal{X}; V_{\max}) \to \mathcal{F}$ as

$$\Pi V = \arg\min_{f\in\mathcal{F}} ||V-f||_\rho. \quad (13)$$

Furthermore, in the rest of the paper with a little abuse of notation, we replace the empirical norm $||v||_n$ defined on states X_1^n by $||V||_n$, where $V \in \mathcal{B}(\mathcal{X}; V_{\max})$ is such that $V(X_t) = v_t$. Finally, we should guarantee that the pathwise LSTD solution $\hat{V}$ is uniformly bounded on $\mathcal{X}$. For this reason, we move from $\mathcal{F}$ to the truncated space $\widetilde{\mathcal{F}}$. A function $\tilde{f} \in \widetilde{\mathcal{F}}$ is defined as

$$\tilde{f}(x) = \begin{cases} f(x) & \text{if } |f(x)| \le V_{\max}, \\ \operatorname{sgn}\big(f(x)\big) V_{\max} & \text{otherwise.} \end{cases} \tag{14}$$

In the next sections, we present conditions on the existence of the pathwise LSTD solution and derive generalization bounds under different assumptions on the way that the samples $X_1, \dots, X_n$ are generated.

5.1. Existence of Pathwise LSTD Solution

In this section, we assume that all the eigenvalues of G are strictly positive and derive a condition to guarantee that the sample-based Gram matrix $\frac{1}{n}\Phi^\top\Phi$ is invertible. In particular, we show that if a large enough number of samples (depending on the smallest eigenvalue of G) is available, then the smallest eigenvalue of $\frac{1}{n}\Phi^\top\Phi$ is strictly positive with high probability.

Lemma 2. *Let G be the Gram matrix defined according to the distribution ρ and $\omega > 0$ be its smallest eigenvalue. Let $X_1, \dots, X_n$ be a path of length n of a stationary β-mixing process with stationary distribution ρ. If the number of samples n satisfies the following condition*

$$\frac{\Lambda(n,\delta)}{n} \max\left\{\frac{\Lambda(n,\delta)}{b}, 1\right\}^{1/\kappa} < \frac{\omega}{288 L^2} \quad , \tag{15}$$

where $\Lambda(n,\delta) = \log\frac{e}{\delta} + \log\big(\max\{6, n\bar{\beta}\}\big)$, then with probability $1-\delta$, the family of features $(\varphi_1, \dots, \varphi_d)$ is linearly independent on the states $X_1, \dots, X_n$ (i.e., $||f_\alpha||_n = 0$ implies $\alpha = 0$) and the smallest eigenvalue ν_n of the sample-based Gram matrix $\frac{1}{n}\Phi^\top\Phi$ satisifies

$$\sqrt{\nu_n} \ge \frac{\sqrt{\omega}}{2} - \sqrt{72 L^2 \frac{\Lambda(n,\delta)}{n} \max\left\{\frac{\Lambda(n,\delta)}{b}, 1\right\}^{1/\kappa}} > 0 \,. \tag{16}$$

Proof. From the definition of the Gram matrix and the fact that ω is its smallest eigenvalue, for any function $f_\alpha \in \mathcal{F}$, we have

$$||f_\alpha||_\rho^2 = ||\phi^\top\alpha||_\rho^2 = \alpha^\top G \alpha \ge \omega \alpha^\top\alpha = \omega ||\alpha||^2.$$

Using a concentration inequality (see Corollary 4 of Lazaric et al. 2010) and the fact that the basis functions φ_j are bounded by L, thus f_α is bounded by $L||\alpha||$, we have $||f_\alpha||_\rho - 2||f_\alpha||_n \le \epsilon$ with probability $1-\delta$, where

$$\epsilon = \|\alpha\| \sqrt{288 L^2 \frac{\Lambda(n,\delta)}{n} \max\left\{\frac{\Lambda(n,\delta)}{b}, 1\right\}^{1/\kappa}} \,.$$

Thus we obtain

$$2||f_\alpha||_n + \epsilon \ge \sqrt{\omega}||\alpha||. \tag{17}$$

Let α be such that $||f_\alpha||_n = 0$, then from Eq. 17 and the definition of ϵ we deduce that $\alpha = 0$. Thus $\nu_n > 0$ and the inequality in Eq. 16 is obtained by choosing α to be the eigenvector of $\frac{1}{n}\Phi^\top\Phi$ correspond to the smallest eigenvalue ν_n. For this value of α, we have $||f_\alpha||_n = \sqrt{\nu_n}||\alpha||$, and the claim follows using Eq. 17. □

Remark 1 If $\Lambda(n,\delta)/b > 1$ and $n\bar{\beta} \ge 6$, the condition on the number of samples can be rewritten as

$$\frac{n}{\log\left(\frac{e}{\delta} n\bar{\beta}\right)^{\frac{1+\kappa}{\kappa}}} \ge \frac{288 L^2}{\omega b^{1/\kappa}}.$$

As it can be seen, the number of samples needed to have strictly positive eigenvalues in the sample-based Gram matrix has an inverse dependency on the smallest eigenvalue of G. As a consequence, the more G is ill-conditioned the more samples we need for the sample-based Gram matrix $\frac{1}{n}\Phi^\top\Phi$ to be invertible.

5.2. Generalization Bounds for Stationary β-mixing Processes

In this section, we show how Theorem 1 can be generalized to the entire state space $\mathcal{X}$ when the Markov chain $\mathcal{M}$ has a stationary distribution ρ. In particular, we consider the case in which the samples $X_1, \dots, X_n$ are obtained by following a single trajectory in the stationary regime of $\mathcal{M}$, i.e., when we consider that X_1 is drawn from ρ.

Theorem 2. *Let $X_1, \dots, X_n$ be a path generated by a stationary β-mixing process with stationary distribution ρ. Let ω be the smallest eigenvalue of the Gram matrix defined according to ρ and n satisfy the condition in Eq. 15. Let $\tilde{V}$ be the truncation (using Eq. 14) of the pathwise LSTD solution, then*

$$||\tilde{V} - V||_\rho \le \frac{2}{1-\gamma}\Big[2\sqrt{2}||V - \Pi V||_\rho + \varepsilon_2 + \gamma V_{\max} L \sqrt{\frac{d}{\nu}}\Big(\sqrt{\frac{8\log(8d/\delta)}{n}} + \frac{1}{n}\Big)\Big] + \varepsilon_1 \tag{18}$$

with probability $1-\delta$, where ν is a lower bound on the eigenvalues of the sample-based Gram matrix defined by Eq. 16,

$$\varepsilon_1 = \sqrt{\frac{\Lambda(n,d,\delta/4)}{n C_2} \max\left\{\frac{\Lambda(n,d,\delta/4)}{b}, 1\right\}^{1/\kappa}}$$

with $\Lambda(n,d,\delta/4) = 2(d+1)\log n + \log\frac{4e}{\delta} + \log^+\left(\max\{18(C_1C_2)^{2(d+1)}, \bar{\beta}\}\right)$, $C_1 = 6912eV_{\max}^2$, *and* $C_2 = (1152V_{\max}^2)^{-1}$, *and*

$$\varepsilon_2 = \sqrt{288\big(V_{\max} + L||\alpha^*||\big)^2 \frac{\Lambda(n,\delta/4)}{n} \max\left\{\frac{\Lambda(n,\delta/4)}{b}, 1\right\}}^{1/\kappa}$$

where $\Lambda(n,\delta/4) = \log\frac{4e}{\delta} + \log\left(\max\{6, n\bar{\beta}\}\right)$ *and* α^* *is such that* $f_{\alpha^*} = \Pi V$.

Proof. This result is a consequence of applying generalization bounds to both sides of Eq. 1 (Theorem 1). We first bound the left-hand side.

$$2||\hat{V} - V||_n \geq 2||\tilde{V} - V||_n \geq ||\tilde{V} - V||_\rho - \varepsilon_1$$

with probability $1-\delta'$. The first step follows from the definition of the truncation operator, while the second step is a straightforward application of Corollary 3 in Lazaric et al. (2010).

We now bound the term $||V - \widehat{\Pi}V||_n$ in Eq. 1:

$$||V - \widehat{\Pi}V||_n \leq ||V - \Pi V||_n \leq 2\sqrt{2}||V - \Pi V||_\rho + \varepsilon_2$$

with probability $1-\delta'$. The first step follows from the definition of the operator $\widehat{\Pi}$. The second step is an application of the inequality of Corollary 4 in Lazaric et al. (2010) for the function $V - \Pi V$.

From Theorem 1, the two generalization bounds, and the lower bound on ν, each one holding with probability $1-\delta'$, the statement of the Theorem (Eq. 18) holds with probability $1-\delta$ by setting $\delta = 4\delta'$. □

Remark 1 Rewriting the bound in terms of the approximation and estimation error terms (up to constants and logarithmic factors), we obtain

$$||\tilde{V} - V||_\rho \leq O\left(\frac{1}{1-\gamma}||V - \Pi V||_\rho + \frac{1}{1-\gamma}\frac{1}{\sqrt{n}}\right).$$

While the first term (*approximation error*) only depends on the target function V and the function space $\mathcal{F}$, the second term (*estimation error*) primarily depends on the number of samples. Thus, when n goes to infinity, the estimation error goes to zero and we obtain the same performance bound (up to a $4\sqrt{2}$ constant) as for the model-based case reported by Tsitsiklis & Van Roy (1997).

Remark 2 Antos et al. (2008) reported a sample-based analysis for the modified Bellman residual (MBR) minimization algorithm. They consider a general setting in which the function space $\mathcal{F}$ is bounded and the performance of the algorithm is evaluated according to an arbitrary measure μ (possibly different than the stationary distribution of the Markov chain ρ). Since Antos et al. (2008) showed that the MBR minimization algorithm is equivalent to LSTD when $\mathcal{F}$ is a linearly parameterized space, it may be interesting to compare Theorem 2 to the bound in Lemma 11 of Antos et al. (2008). In Theorem 2, similar to Antos et al. (2008), samples are drawn from a stationary β-mixing process, however, $\mathcal{F}$ is a linear space and ρ is the stationary distribution of the Markov chain. It is interesting to note the impact of these two differences in the final bound. The use of linear spaces has a direct effect on the estimation error and leads to a better convergence rate due to the use of improved functional concentration inequalities (Lemma 5 in Lazaric et al. 2010). In fact, while in Antos et al. (2008) the estimation error for the squared error is of order $O(1/\sqrt{n})$, here we achieve a faster convergence rate of order $O(1/n)$. The use of ρ instead of an arbitrary measure μ has a significant impact on the approximation error. The approximation error in Eq. 18 $||V - \Pi V||_\rho$ only depends on how well the space $\mathcal{F}$ can approximate the value function V. On the other hand, the approximation error of Antos et al. (2008) contains terms that are related to more complex properties of the space, such as its capability to approximate any function obtained by applying the Bellman operator $\mathcal{T}$ to any function in $\mathcal{F}$. This term called the inherent Bellman error can be shown to be small only for specific classes of MDPs (e.g., Lipschitz MDPs). Finally, it is interesting to notice that although the solution of MBR minimization reduces to LSTD, its sample-based analysis cannot be directly used for LSTD. In fact, in Antos et al. (2008) the function space $\mathcal{F}$ is assumed to be bounded, while general linear spaces cannot be bounded. Whether the analysis of Antos et al. (2008) can be extended to the truncated solution $\tilde{V}$ of LSTD is an open question that requires further investigation.

5.3. Generalization Bounds for Markov Chains

The main assumption in the previous section is that $X_1, \ldots, X_n$ is generated by a stationary β-mixing process with stationary distribution ρ. This is possible if we consider samples of a Markov chain during its stationary regime, i.e. $X_1 \sim \rho$. However in practice, ρ is not known, and the first sample X_1 is usually drawn from a given initial distribution and the rest of the sequence is obtained by following the Markov chain from X_1 on. As a result, the sequence $X_1, \ldots, X_n$ is no longer a realization of a stationary β-mixing process. Nonetheless, under suitable conditions, after $\tilde{n} < n$ steps, the distribution of $X_{\tilde{n}}$ approaches the stationary distribution ρ. In fact, according to the convergence theorem for fast-mixing Markov chains (see e.g., Proposition 3 in Lazaric et al. 2010), for any initial

distribution $\lambda \in \mathcal{S}(\mathcal{X})$, we have

$$|| \int_{\mathcal{X}} \lambda(dx) P^n(\cdot|x) - \rho(\cdot)||_{TV} \leq \bar{\beta} \exp(-b n^{\kappa}).$$

We now derive a bound for a modification of pathwise LSTD in which the first $\tilde{n}$ samples (that are used to burn the chain) are discarded and the remaining $n - \tilde{n}$ samples are used as training samples for the algorithm.

Theorem 3. *Let $X_1, \ldots, X_n$ be a trajectory generated by a β-mixing Markov chain with stationary distribution ρ. Let $\tilde{n}$ $(1 \leq \tilde{n} < n)$ be such that $n - \tilde{n}$ satisfies the condition of Eq. 15, and $X_{\tilde{n}+1}, \ldots, X_n$ be the samples actually used by the algorithm. Let ω be the smallest eigenvalue of the Gram matrix defined according to ρ and $\alpha^* \in \mathbb{R}^d$ be such that $f_{\alpha^*} = \Pi V$. Let $\tilde{V}$ be the truncation of the pathwise LSTD solution (using Eq. 14), then by setting $\tilde{n} = \left(\frac{1}{b} \log \frac{2e\bar{\beta}n}{\delta}\right)^{1/\kappa}$, we have*

$$||\tilde{V} - V||_{\rho} \leq \frac{2}{1-\gamma}\Big[2\sqrt{2}||V - \Pi V||_{\rho} + \varepsilon_2 + \gamma V_{\max} L \sqrt{\frac{d}{\nu}}\Big(\sqrt{\frac{8 \log(8d/\delta)}{n - \tilde{n}}} + \frac{1}{n - \tilde{n}}\Big)\Big] + \varepsilon_1 \quad (19)$$

with probability $1 - \delta$, where ε_1 and ε_2 are defined as in Theorem 2 (with $n - \tilde{n}$ as the number of training samples).

The proof of this result is a simple consequence of Lemma 8 in Lazaric et al. (2010) applied to Theorem 2.

Remark 1 The bound in Eq. 19 indicates that in the case of β-mixing Markov chains, a similar performance to the one for stationary β-mixing processes is obtained by discarding the first $\tilde{n} = O(\log n)$ samples.

6. Conclusions

In this paper we presented a finite-sample analysis of a natural version of LSTD, called pathwise LSTD. We first considered a general setting where we do not make any assumption about the Markov chain. We derived an empirical performance bound which indicates how close the LSTD solution is to the value function V at the states generated by the Markov chain. The bound is expressed in terms of the best possible approximation of V (approximation error) in the linear approximation space $\mathcal{F}$ and an estimation error term which depends on the number of samples (the quadratic error scales with $O(n^{-1/2})$) and the smallest strictly-positive eigeinvalue of the sample-based Gram matrix. We then showed that when the Markov chain possesses a stationary distribution, then one can deduce generalization performance bounds using the stationary distribution of the chain as our generalization measure. In particular, we considered the cases where the sample trajectory is generated by stationary and non-stationary β-mixing Markov chains and derived the corresponding bounds.

This work raises two open questions: 1) Is it possible to remove the dependence to ν_n in the bound of Theorem 1? 2) Is it possible to extend the current analysis to the general case of LSTD(λ)?

Acknowledgments This work was supported by French National Research Agency (ANR) (project EXPLO-RA n° ANR-08-COSI-004).

References

Antos, A., Szepesvári, Cs., and Munos, R. Learning near-optimal policies with Bellman-residual minimization based fitted policy iteration and a single sample path. *Machine Learning Journal*, 71:89–129, 2008.

Bertsekas, D. *Dynamic Programming and Optimal Control*. Athena Scientific, 2001.

Boyan, J. Least-squares temporal difference learning. *Proceedings of the 16th International Conference on Machine Learning*, pp. 49–56, 1999.

Bradtke, S. and Barto, A. Linear least-squares algorithms for temporal difference learning. *Machine Learning*, 22:33–57, 1996.

Lagoudakis, M. and Parr, R. Least-squares policy iteration. *Journal of Machine Learning Research*, 4: 1107–1149, 2003.

Lazaric, A., Ghavamzadeh, M., and Munos, R. Finite-sample analysis of LSTD. Technical Report inria-00482189, INRIA, 2010.

Sutton, R. and Barto, A. *Reinforcement Learning: An Introduction*. MIP Press, 1998.

Tsitsiklis, J. and Van Roy, B. An analysis of temporal difference learning with function approximation. *IEEE Transactions on Automatic Control*, 42:674–690, 1997.