Finite-Sample Analysis of Least-Squares Policy Iteration

Alessandro Lazaric, Mohammad Ghavamzadeh, Rémi Munos

Journal of Machine Learning Research 13 (2012) 3041-3074 Submitted 8/10; Revised 11/11; Published 10/12

Finite-Sample Analysis of Least-Squares Policy Iteration

Alessandro Lazaric ALESSANDRO.LAZARIC@INRIA.FR
Mohammad Ghavamzadeh MOHAMMAD.GHAVAMZADEH@INRIA.FR
Rémi Munos REMI.MUNOS@INRIA.FR
INRIA Lille - Nord Europe, Team SequeL
40 Avenue Halley
59650 Villeneuve d'Ascq, France

Editor: Ronald Parr

Abstract

In this paper, we report a performance bound for the widely used least-squares policy iteration (LSPI) algorithm. We first consider the problem of policy evaluation in reinforcement learning, that is, learning the value function of a fixed policy, using the least-squares temporal-difference (LSTD) learning method, and report finite-sample analysis for this algorithm. To do so, 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. Finally, we analyze how the error at each policy evaluation step is propagated through the iterations of a policy iteration method, and derive a performance bound for the LSPI algorithm.

Keywords: Markov decision processes, reinforcement learning, least-squares temporal-difference, least-squares policy iteration, generalization bounds, finite-sample analysis

1. Introduction

Least-squares temporal-difference (LSTD) learning (Bradtke and 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 and Parr, 2003), which extends LSTD to control by using it in the policy evaluation step of policy iteration. 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 $\mathcal{F}$. Although LSTD and LSPI have been widely used in the RL community, a finite-sample analysis of LSTD, that is, performance bounds in terms of the number of samples, the space $\mathcal{F}$, and the characteristic parameters of the MDP at hand, is still missing.

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 and 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. Yu (2010) has extended this analysis and derived an asymptotic convergence analysis for off-policy LSTD(λ), that is when the samples are collected following a behavior policy different from the policy π under evaluation. Finally, on-policy empirical LSTD has been analyzed by Bertsekas (2007). 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 a system of linear equations $Ax = b$). Nonetheless, Bertsekas (2007) does not provide a finite-sample analysis of the algorithm. Although these analyses already provide some insights on the behavior of LSTD, asymptotic results do not give a full characterization of the performance of the algorithm when only a finite number of samples is available (which is the most common situation in practice). On the other hand, a finite-sample analysis has a number of important advantages: **1)** unlike in Tsitsiklis and Van Roy (1997), where they assume that model-based LSTD always returns a solution, in a finite-sample analysis we study the characteristics of the actual empirical LSTD fixed point, including its existence, **2)** a finite-sample bound explicitly reveals how the prediction error of LSTD is related to the characteristic parameters of the MDP at hand, such as the discount factor, the dimensionality of the function space $\mathcal{F}$, and the number of samples, **3)** once this dependency is clear, the bound can be used to determine the order of magnitude of the number of samples needed to achieve a desired accuracy.

Recently, several works have been focused on deriving a finite-sample analysis for different RL algorithms. In the following, we review those that are more strictly related to LSTD and to the results reported in this paper. 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 to extend the finite-sample bounds derived by Antos et al. (2008) to unbounded linear spaces considered by LSTD. Farahmand et al. (2008) proposed a ℓ_2-regularized extension of LSPI and provided finite-sample analysis for the algorithm when the function space is a reproducing kernel Hilbert space (RKHS). In this work, the authors consider the optimization formulation of LSTD (instead of the better known fixed-point formulation) and assume that a generative model of the environment is available. Moreover, the analysis is for ℓ_2-regularized LSTD (LSPI) and also for the case that the function space $\mathcal{F}$ is a RKHS. Pires and Szepesvári (2012) also analyzed a regularized version of LSTD reporting performance bounds for both the on-policy and off-policy case. In this paper, we first 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 (precisely defined in Section 3), 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 assumption is made on the existence of a stationary distribution for the Markov chain. Then, in the case that 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. Finally, along the lines of Antos et al. (2008), we show how the LSTD error is propagated through the iterations of LSPI, and under suitable assumptions, derive a performance bound for the LSPI algorithm.

Besides providing a full finite-sample analysis of LSPI, the major insights gained by the analysis in the paper may 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) always has 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 the LSTD solution, $\hat{v}$, evaluated at the states of the trajectory used by the algorithm. Moreover, 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 5). 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 and 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, and obtain a better bound both in terms of **1)** estimation error, a rate of order $O(1/n)$ instead of $O(1/\sqrt{n})$ for the squared error, and **2)** approximation error, the minimal distance between the value function V^π and the space $\mathcal{F}$ instead of the inherent Bellman errors of $\mathcal{F}$. The extension in Theorem 6 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. Finally, the analysis of LSPI reveals the need for several critical assumptions on the stationary distributions of the policies that are greedy w.r.t. to the functions in the linear space $\mathcal{F}$. These assumptions seem unavoidable when an on-policy method is used at each iteration, and whether they can be removed or relaxed in other settings is still an open question. This paper extends and improves over the conference paper by Lazaric et al. (2010) in the following respects: **1)** we report the full proofs and technical tools for all the theoretical results, thus making the paper self-contained, **2)** we extend the LSTD results to LSPI showing how the approximation errors are propagated through iterations.

The rest of the paper is organized as follows. In Section 2, we set the notation used throughout the paper. In Section 3, we introduce pathwise LSTD by a minor modification to the standard LSTD formulation in order to guarantee the existence of at least one solution. In Section 4, we introduce the Markov design setting for regression and report an empirical bound for LSTD. In Section 5, we show how the Markov design bound of Section 4 may be extended when the Markov chain admits a stationary distribution. In Section 6, we analyze how the LSTD error is propagated through the iterations of LSPI and derive a performance bound for the LSPI algorithm. Finally in Section 7, we draw conclusions and discuss some possible directions for future work.

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 points $X_1, \ldots, X_n \in \mathcal{X}$, 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 (Bertsekas and Tsitsiklis, 1996; Sutton and Barto, 1998) in which a learning agent interacts with a stochastic environment and this interaction is modeled as a discrete-time discounted Markov decision process (MDP). A discounted MDP is a tuple $\mathcal{M} = \langle \mathcal{X}, \mathcal{A}, r, P, \gamma \rangle$ where the state space $\mathcal{X}$ is a bounded closed subset of the s-dimensional Euclidean space, $\mathcal{A}$ is a finite ($|\mathcal{A}| < \infty$) action space, the reward function $r : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$ is uniformly bounded by $R_{\max}$, the transition kernel P is such that for all $x \in \mathcal{X}$ and $a \in \mathcal{A}$, $P(\cdot|x,a)$ is a distribution over $\mathcal{X}$, and $\gamma \in (0,1)$ is a discount factor. A deterministic policy $\pi : \mathcal{X} \to \mathcal{A}$ is a mapping from states to actions. For a given policy π, the MDP $\mathcal{M}$ is reduced to a Markov chain $\mathcal{M}^\pi = \langle \mathcal{X}, R^\pi, P^\pi, \gamma \rangle$ with the reward function $R^\pi(x) = r\big(x, \pi(x)\big)$, transition kernel $P^\pi(\cdot|x) = P\big(\cdot\,|x, \pi(x)\big)$, and stationary distribution ρ^π (if it admits one). 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

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

We also define the optimal value function V^* as the unique fixed-point of the optimal Bellman operator $\mathcal{T}^* : \mathcal{B}(\mathcal{X}; V_{\max}) \to \mathcal{B}(\mathcal{X}; V_{\max})$ defined by

$$(\mathcal{T}^* V)(x) = \max_{a\in\mathcal{A}} \left[r(x,a) + \gamma \int_{\mathcal{X}} P(dy|x,a) V(y) \right].$$

In the following sections, to simplify the notation, we remove the dependency to the policy π and use R, P, V, ρ, and $\mathcal{T}$ instead of R^π, P^π, V^π, ρ^π, and $\mathcal{T}^\pi$ whenever the policy π is fixed and clear from the context.

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

Let $(X_1, \ldots, X_n)$ be a sample path (trajectory) of size n generated by the Markov chain $\mathcal{M}^\pi$. Let $v \in \mathbb{R}^n$ and $r \in \mathbb{R}^n$ be such that $v_t = V(X_t)$ and $r_t = R(X_t)$ be the value vector and the reward vector, respectively. Also, let $\Phi = [\phi(X_1)^\top; \ldots; \phi(X_n)^\top]$ 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. Moreover, $\widehat{\Pi}$ is a non-expansive mapping w.r.t. the ℓ_2-norm: since the projection is orthogonal and using the Cauchy-Schwarz inequality $||\widehat{\Pi} y - \widehat{\Pi} z||_n^2 = \langle y - z, \widehat{\Pi} y - \widehat{\Pi} z \rangle_n \le ||y - z||_n ||\widehat{\Pi} y - \widehat{\Pi} z||_n$, and thus, we obtain $||\widehat{\Pi} y - \widehat{\Pi} z||_n \le ||y - z||_n$.

3. Pathwise LSTD

Pathwise LSTD (Algorithm 1) is a version of LSTD that takes as input a linear function space $\mathcal{F}$ and a single trajectory $X_1, \ldots, X_n$ generated by following the policy, and returns the fixed-point

Algorithm 1 A pseudo-code for the batch pathwise LSTD algorithm.

Input: Linear space $\mathcal{F} = \text{span}\{\varphi_i, 1 \le i \le d\}$, sample trajectory $\{(x_t, r_t)\}_{t=1}^n$ of the Markov chain

Build the feature matrix $\Phi = [\phi(x_1)^\top; \ldots; \phi(x_n)^\top]$
Build the empirical transition matrix $\widehat{P} : \widehat{P}_{ij} = \mathbb{I}\{j = i+1, j \neq n\}$
Build matrix $A = \Phi^\top (I - \gamma \widehat{P})\Phi$
Build vector $b = \Phi^\top r$
Return the **pathwise LSTD solution** $\hat{\alpha} = A^+ b$

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, that is, for any $y, z \in \mathbb{R}^n$, we have

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

Since the orthogonal projection $\widehat{\Pi}$ is non-expansive w.r.t. ℓ_2-norm, from Banach fixed point theorem, there exists a unique fixed-point $\hat{v}$ of the mapping $\widehat{\Pi}\widehat{\mathcal{T}}$, that is, $\hat{v} = \widehat{\Pi}\widehat{\mathcal{T}}\hat{v}$. 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 $\Phi^\top (I - \gamma \widehat{P})\Phi\alpha = \Phi^\top r$. Therefore, by setting $A = \Phi^\top (I - \gamma \widehat{P})\Phi$ and $b = \Phi^\top r$, we recover a $d \times d$ system of equations $A\alpha = b$ similar to the one in the original LSTD algorithm. Note that since the fixed point $\hat{v}$ exists, this system always has at least one solution. We call the solution with minimal norm, $\hat{\alpha} = A^+ b$, where A^+ is the Moore-Penrose pseudo-inverse of A, the pathwise LSTD solution.[1]

Finally, notice that the algorithm reported in Figure 1 may be easily extended to the incremental version of LSTD by incrementally building the inverse of the matrix A as the samples are collected.

4. Markov Design Bound

In Section 3, we defined the pathwise Bellman operator with a slight modification in the definition of the empirical Bellman operator $\widehat{\mathcal{T}}$, and showed that the operator $\widehat{\Pi}\widehat{\mathcal{T}}$ always has a unique fixed point $\hat{v}$. In this section, we derive a bound for the performance of $\hat{v}$ evaluated at the states of the trajectory used by the pathwise LSTD algorithm. We first state the main theorem and we discuss it in a number of remarks. The proofs are postponed at the end of the section.

Theorem 1 *Let $X_1, \ldots, X_n$ be a trajectory generated by 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 $\{X_t\}_{t=1}^n$, respectively. Then with probability at least $1 - \delta$ (the probability is w.r.t. the random trajectory), we have*

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

1. Note that whenever the matrix A is invertible $A^+ = A^{-1}$.

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 Theorem 1 provides a bound on the prediction error of the LSTD solution $\hat{v}$ w.r.t. the true value function v on the trajectory $X_1,\ldots,X_n$ used as a training set for pathwise-LSTD. The bound contains two main terms. The first term $||v-\widehat{\Pi}v||_n$ is the *approximation* error and it represents the smallest possible error in approximating v with functions in $\mathcal{F}$. This error cannot be avoided. The second term, of order $O(\sqrt{d/n})$, is the *estimation* error and it accounts for the error due to the use of a finite number of noisy samples and it shows what is the influence of the different elements of the problem (e.g., γ, d, n) on the prediction error and it provides insights about how to tune some parameters. We first notice that the bound suggests that the number of samples n should be significantly bigger than the number of features d in order to achieve a small estimation error. Furthermore, the bound can be used to estimate the number of samples needed to guarantee a desired prediction error ε. In fact, apart from the approximation error, which is unavoidable, we have that $n = O(d/((1-\gamma)^2\varepsilon^2))$ samples are enough to achieve an ε-accurate approximation of the true value function v. We also remark that one might be tempted to reduce the dimensionality d, so as to reduce the sample cost of the algorithm. Nonetheless, this is likely to reduce the approximation capability of $\mathcal{F}$ and thus increase the approximation error.

Remark 2 When the eigenvalues of the sample-based Gram matrix $\frac{1}{n}\Phi^\top\Phi$ are all non-zero, $\Phi^\top\Phi$ is invertible, and thus, $\widehat{\Pi} = \Phi(\Phi^\top\Phi)^{-1}\Phi^\top$. In this case, the uniqueness of $\hat{v}$ implies the uniqueness of $\hat{\alpha}$ since

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

On the other hand, when the sample-based Gram matrix $\frac{1}{n}\Phi^\top\Phi$ is not invertible, the system $Ax = b$ may have many solutions. Among all the possible solutions, one may choose the one with minimal norm: $\hat{\alpha} = A^+ b$.

Remark 3 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, Equation 1 (with ν_n replaced by ν) holds with probability at least $1-(\delta+\delta')$ (see Section 5.1 for a case in which such constant ν can be computed and it is related to the smallest eigenvalue of the model based Gram matrix).

Remark 4 Theorem 1 provides a bound without any reference to the stationary distribution of the Markov chain. In fact, the bound of Equation 1 holds even when the chain does not admit a stationary distribution. For example, consider a Markov chain on the real line where the transitions always move the states to the right, that is, $p(X_{t+1} \in dy | X_t = x) = 0$ 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 $\phi(X_1),\ldots,\phi(X_n)$ are sufficiently independent, so that all 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 samples 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 Equation 1 is the inverse dependency on the smallest positive eigenvalue ν_n. A similar dependency is shown in the LSTD analysis of Bertsekas (2007). The main 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 regression setting with *Markov design* and then state and prove a lemma about this model. Delattre and Gaïffas (2011) recently analyzed a similar setting in the general case of martingale incremental errors.

Definition 2 *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*

$$|\xi_t| \le C \quad \text{and} \quad \mathbb{E}[\xi_t | X_1, \dots, X_t] = 0. \tag{2}$$

The next lemma reports a risk bound for the Markov design setting which is of independent interest.

Lemma 3 (Regression bound for the Markov design setting) *We consider the model of regression with Markov design in Definition 2. Let $\hat{w} \in \mathcal{F}_n$ be the least-squares estimate of the (noisy) values $Y = \{Y_t\}_{t=1}^n$, that is, $\hat{w} = \widehat{\Pi} Y$, and $w \in \mathcal{F}_n$ be the least-squares estimate of the (noiseless) values $Z = \{Z_t = f(X_t)\}_{t=1}^n$, that is, $w = \widehat{\Pi} Z$. Then for any $\delta > 0$, with probability at least $1-\delta$ (the probability is w.r.t. the random sample path $X_1, \dots, X_n$), we have*

$$||\hat{w} - w||_n \le CL \sqrt{\frac{2d \log(2d/\delta)}{n \nu_n}}\,, \tag{3}$$

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

Proof [Lemma 3] We define $\xi \in \mathbb{R}^n$ to be the vector with components $\xi_t = Y_t - Z_t$, and $\hat{\xi} = \hat{w} - w = \widehat{\Pi}(Y - Z) = \widehat{\Pi}\xi$. Since the projection is orthogonal we have $\langle \hat{\xi}, \xi \rangle_n = ||\hat{\xi}||_n^2$ (see Figure 1). 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

$$||\hat{\xi}||_n^2 = \langle \hat{\xi}, \xi \rangle_n = \frac{1}{n} \sum_{i=1}^d \alpha_i \sum_{t=1}^n \xi_t \varphi_i(X_t) \le \frac{1}{n} ||\alpha||_2 \left[\sum_{i=1}^d \Big(\sum_{t=1}^n \xi_t \varphi_i(X_t) \Big)^2 \right]^{1/2}. \tag{4}$$

Now among the vectors α such that $\hat{\xi} = \Phi\alpha$, we define $\hat{\alpha}$ to be the one with minimal ℓ_2-norm, that is, $\hat{\alpha} = \Phi^+ \hat{\xi}$. Let K denote the null-space of Φ, which is also the null-space of $\frac{1}{n}\Phi^\top\Phi$. Then $\hat{\alpha}$ may 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

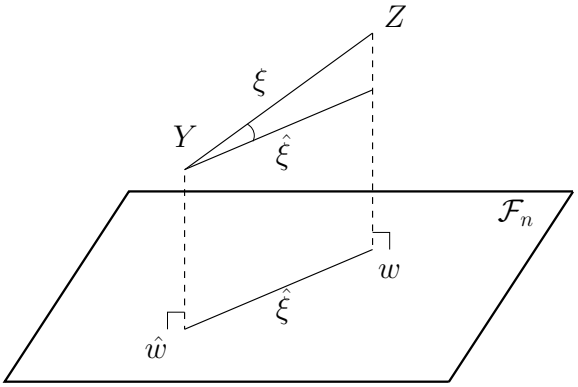

Figure 1: This figure shows the components used in Lemma 3 and its proof such as w, $\hat{w}$, ξ, and $\hat{\xi}$, and the fact that $\langle \hat{\xi}, \xi \rangle_n = ||\hat{\xi}||_n^2$.

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 $\hat{\alpha} \in K^\perp$,

$$||\hat{\xi}||_n^2 = \frac{1}{n}\hat{\alpha}^\top\Phi^\top\Phi\hat{\alpha} \geq \nu_n\hat{\alpha}^\top\hat{\alpha} = \nu_n||\hat{\alpha}||_2^2. \tag{5}$$

By using the result of Equation 5 in Equation 4, we obtain

$$||\hat{\xi}||_n \leq \frac{1}{n\sqrt{\nu_n}}\left[\sum_{i=1}^{d}\Big(\sum_{t=1}^{n}\xi_t\varphi_i(X_t)\Big)^2\right]^{1/2}. \tag{6}$$

Now, from the conditions on the noise in Equation 2, we have that for any $i = 1,\ldots,d$

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

and since $\xi_t\varphi_i(X_t)$ is adapted to the filtration generated by $X_1,\ldots,X_{t+1}$, it is a martingale difference sequence w.r.t. that filtration. Thus one may apply Azuma's inequality to deduce that with probability $1-\delta$,

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

where we used that $|\xi_t\varphi_i(X_t)| \leq CL$ for any i and t. By a union bound over all features, we have that with probability $1-\delta$, for all $1 \leq i \leq d$

$$\Big|\sum_{t=1}^{n}\xi_t\varphi_i(X_t)\Big| \leq CL\sqrt{2n\log(2d/\delta)}\,. \tag{7}$$

The result follows by combining Equations 7 and 6. ■

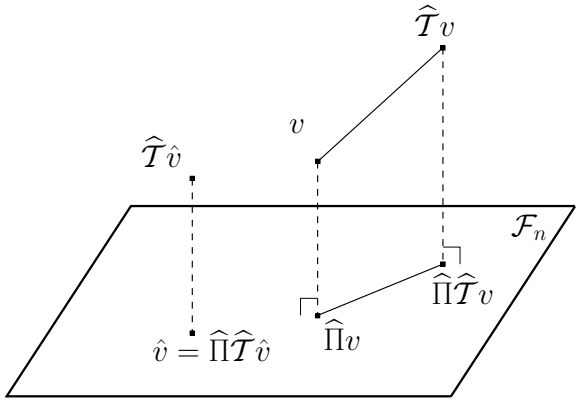

Figure 2: This figure represents the space $\mathbb{R}^n$, the linear vector subspace $\mathcal{F}_n$ and some vectors used in the proof of Theorem 1.

Remark about Lemma 3 Note that this lemma is an extension of the bound for regression with deterministic design in which the states, $\{X_t\}_{t=1}^n$, are fixed and the noise terms, ξ_t's, are independent. In deterministic design, usual concentration results provide high probability bounds similar to Equation 3 (see, e.g., Hsu et al., 2012), but without the dependence on ν_n. An open question is whether it is possible to remove ν_n in the bound for the Markov design regression setting.

In the Markov design model considered in this lemma, states $\{X_t\}_{t=1}^n$ 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 states $X_1,\ldots,X_t$). 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 $r(X_t)+\gamma V(X_{t+1})-V(X_t)$.

Proof [Theorem 1]

Step 1: Using the Pythagorean theorem and the triangle inequality, we have (see Figure 2)

$$||v-\hat{v}||_n^2 = ||v-\widehat{\Pi}v||_n^2 + ||\hat{v}-\widehat{\Pi}v||_n^2 \le ||v-\widehat{\Pi}v||_n^2 + \left(||\hat{v}-\widehat{\Pi}\widehat{\mathcal{T}}v||_n + ||\widehat{\Pi}\widehat{\mathcal{T}}v-\widehat{\Pi}v||_n\right)^2. \quad (8)$$

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

$$||\hat{v}-\widehat{\Pi}\widehat{\mathcal{T}}v||_n = ||\widehat{\Pi}\widehat{\mathcal{T}}\hat{v}-\widehat{\Pi}\widehat{\mathcal{T}}v||_n \le \gamma||\hat{v}-v||_n, \quad (9)$$

Thus from Equation 8 and 9, we have

$$||v-\hat{v}||_n^2 \le ||v-\widehat{\Pi}v||_n^2 + \left(\gamma||v-\hat{v}||_n + ||\widehat{\Pi}\widehat{\mathcal{T}}v-\widehat{\Pi}v||_n\right)^2. \quad (10)$$

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 3 applied to the vectors $Y=\widehat{\mathcal{T}}v$ and $Z=v$. Since v is the value function at the points $\{X_t\}_{t=1}^n$, from the definition of the pathwise Bellman operator, we have that for $1\le t\le n-1$,

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

and $\xi_n = y_n - v_n = -\gamma \int P(dy|X_n)V(y)$. Thus, Equation 2 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 be applied only to the sequence of $n-1$ terms (the n-th term adds a contribution to the bound), thus instead of Equation 7, we obtain

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

with probability $1-\delta$, for all $1 \le i \le d$. Combining with Equation 6, we deduce that with probability $1-\delta$, we have

$$||\widehat{\Pi}\widehat{\mathcal{T}}v - \widehat{\Pi}v||_n \le \gamma V_{\max} L \sqrt{\frac{d}{\nu_n}} \Big(\sqrt{\frac{8\log(2d/\delta)}{n}} + \frac{1}{n} \Big), \tag{11}$$

where ν_n is the smallest strictly-positive eigenvalue of $\frac{1}{n}\Phi^\top \Phi$. The claim follows by solving Equation 10 for $||v - \hat{v}||_n$ and replacing $||\widehat{\Pi}\widehat{\mathcal{T}}v - \widehat{\Pi}v||_n$ from Equation 11. ■

5. Generalization Bounds

As we pointed out earlier, Theorem 1 makes no assumption on the existence of the stationary distribution of the Markov chain. This generality comes at the cost that the performance is evaluated only at the states visited by the Markov chain and no generalization on other states is possible. However in many problems of interest, the Markov chain has a stationary distribution ρ, and thus, the performance may 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 ρ ; $G \in \mathbb{R}^{d\times d}$, $G_{ij} = \int \varphi_i(x)\varphi_j(x)\rho(dx)$. In this section, we assume that the Markov chain $\mathcal{M}^\pi$ is exponentially fast β-mixing with parameters $\bar{\beta}, b, \kappa$, that is, its β-mixing coefficients satisfy $\beta_i \le \bar{\beta}\exp(-bi^\kappa)$ (see Section A.2 in the appendix 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

$$\Pi V = \arg\min_{f\in\mathcal{F}} ||V - f||_\rho \,.$$

Furthermore, in the rest of the paper with a little abuse of notation, we replace the empirical norm $||v||_n$ defined on states $X_1, \ldots, X_n$ 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 $\widehat{V}$ is uniformly bounded on $\mathcal{X}$. For this reason, we move from $\mathcal{F}$ to the truncated space $\widetilde{\mathcal{F}}$ in which for any function $f \in \mathcal{F}$, a truncated function $\tilde{f}$ is defined as

$$\tilde{f}(x) = \begin{cases} f(x) & \text{if } |f(x)| \le V_{\max} \,, \\ \operatorname{sgn}\big(f(x)\big) V_{\max} & \text{otherwise.} \end{cases} \tag{12}$$

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 the samples $X_1, \ldots, X_n$ are generated.

5.1 Uniqueness of Pathwise LSTD Solution

In this section, we assume that all the eigenvalues of G are strictly positive; that is, we assume the existence of the model-based solution of LSTD, and derive a condition to guarantee that the sample-based Gram matrix $\frac{1}{n}\Phi^\top\Phi$ is invertible. More specifically, 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 4 *Let G be the Gram matrix defined according to the distribution ρ and $\omega > 0$ be its smallest eigenvalue. Let $X_1,\ldots,X_n$ be a trajectory of length n of a stationary β-mixing process with parameters $\bar{\beta}, b, \kappa$ and stationary distribution ρ. If the number of samples n satisfies the following condition*

$$n > \frac{288L^2\Lambda(n,d,\delta)}{\omega}\max\left\{\frac{\Lambda(n,d,\delta)}{b},1\right\}^{1/\kappa}, \tag{13}$$

where[2] $\Lambda(n,d,\delta) = 2(d+1)\log n + \log\frac{e}{\delta} + \log^+\left(\max\{18(6e)^{2(d+1)},\bar{\beta}\}\right)$*, then with probability* $1-\delta$*, the family of features $(\varphi_1,\ldots,\varphi_d)$ is linearly independent on the states $X_1,\ldots,X_n$ (i.e., $||f_\alpha||_n = 0$ implies $\alpha = 0$) and the smallest eigenvalue ν_n of the sample-based Gram matrix $\frac{1}{n}\Phi^\top\Phi$ satisfies*

$$\sqrt{\nu_n} \geq \sqrt{\nu} = \frac{\sqrt{\omega}}{2} - 6L\sqrt{\frac{2\Lambda(n,d,\delta)}{n}\max\left\{\frac{\Lambda(n,d,\delta)}{b},1\right\}^{1/\kappa}} > 0\,. \tag{14}$$

Proof From the definition of the Gram matrix and the fact that $\omega > 0$ is its smallest eigenvalue, for any function $f_\alpha \in \mathcal{F}$, we have

$$||f_\alpha||_\rho^2 = ||\phi^\top\alpha||_\rho^2 = \alpha^\top G\alpha \geq \omega\alpha^\top\alpha = \omega||\alpha||^2. \tag{15}$$

Using the concentration inequality from Corollary 18 in the appendix and the fact that the basis functions φ_i are bounded by L, thus f_α is bounded by $L||\alpha||$, we have $||f_\alpha||_\rho - 2||f_\alpha||_n \leq \varepsilon$ with probability $1-\delta$, where

$$\varepsilon = 12L\|\alpha\|\sqrt{\frac{2\Lambda(n,d,\delta)}{n}\max\left\{\frac{\Lambda(n,d,\delta)}{b},1\right\}^{1/\kappa}}\,.$$

Thus we obtain

$$2||f_\alpha||_n + \varepsilon \geq \sqrt{\omega}||\alpha||. \tag{16}$$

Let α be such that $||f_\alpha||_n = 0$, then if the number of samples n satisfies the condition of Equation 13, we may deduce from Equation 16 and the definition of ε that $\alpha = 0$. This indicates that given Equation 13, with probability $1-\delta$, the family of features $(\varphi_1,\ldots,\varphi_d)$ is linearly independent on the states $X_1,\ldots,X_n$, and thus, $\nu_n > 0$. The inequality in Equation 14 is obtained by choosing α to be the eigenvector of $\frac{1}{n}\Phi^\top\Phi$ corresponding to the smallest eigenvalue ν_n. For this value of α, we have $||f_\alpha||_n = \sqrt{\nu_n}||\alpha||$. By using the definition of ε in Equation 16 and reordering we obtain

$$2\sqrt{\nu_n}||\alpha|| + 12L||\alpha||\sqrt{\frac{2\Lambda(n,d,\delta)}{n}\max\left\{\frac{\Lambda(n,d,\delta)}{b},1\right\}^{1/\kappa}} \geq \sqrt{\omega}||\alpha||\,,$$

and the claim follows. ■

2. We define $\log^+ x = \max\{\log x, 0\}$.

Remark 1 In order to make the condition on the number of samples and its dependency on the critical parameters of the problem at hand more explicit, let us consider the case of a stationary process with $b = \beta = \kappa = 1$. Then the condition in Equation 13 becomes (up to constant and logarithmic factors)

$$n \geq \widetilde{O}\left(\frac{288L^2}{\omega}\left((d+1)\log\frac{n}{\delta}\right)^2\right).$$

As 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 are needed 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 may be generalized to the entire state space $\mathcal{X}$ when the Markov chain $\mathcal{M}^\pi$ has a stationary distribution ρ. In particular, we consider the case in which the samples $X_1,\ldots,X_n$ are obtained by following a single trajectory in the stationary regime of $\mathcal{M}^\pi$, that is, when we consider that X_1 is drawn from ρ.

Theorem 5 *Let $X_1,\ldots,X_n$ be a path generated by a stationary β-mixing process with parameters $\bar{\beta}, b, \kappa$ and stationary distribution ρ. Let $\omega > 0$ be the smallest eigenvalue of the Gram matrix defined according to ρ and n satisfy the condition in Equation 13. Let $\widetilde{V}$ be the truncation (using Equation 12) of the pathwise LSTD solution, then*

$$||\widetilde{V} - V||_\rho \leq \frac{2}{\sqrt{1-\gamma^2}}\left(2\sqrt{2}||V-\Pi V||_\rho + \varepsilon_2\right) + \frac{2}{1-\gamma}\left[\gamma V_{\max} L\sqrt{\frac{d}{\nu}}\left(\sqrt{\frac{8\log(8d/\delta)}{n}} + \frac{1}{n}\right)\right] + \varepsilon_1 \tag{17}$$

with probability $1-\delta$, where ν is a lower-bound on the eigenvalues of the sample-based Gram matrix defined by Equation 14,

$$\varepsilon_1 = 24V_{\max}\sqrt{\frac{2\Lambda_1(n,d,\delta/4)}{n}\max\left\{\frac{\Lambda_1(n,d,\delta/4)}{b},1\right\}^{1/\kappa}},$$

with $\Lambda_1(n,d,\delta/4) = 2(d+1)\log n + \log\frac{4e}{\delta} + \log^+\left(\max\{18(6e)^{2(d+1)},\bar{\beta}\}\right)$, and

$$\varepsilon_2 = 12\left(V_{\max} + L||\alpha^*||\right)\sqrt{\frac{2\Lambda_2(n,\delta/4)}{n}\max\left\{\frac{\Lambda_2(n,\delta/4)}{b},1\right\}^{1/\kappa}}, \tag{18}$$

with $\Lambda_2(n,\delta/4) = \log\frac{4e}{\delta} + \log\left(\max\{6,n\bar{\beta}\}\right)$ and α^* is such that $f_{\alpha^*} = \Pi V$.

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

$$2||\widehat{V} - V||_n \geq 2||\widetilde{V} - V||_n \geq ||\widetilde{V} - V||_\rho - \varepsilon_1$$

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 17 in the appendix.

We now bound the term $||V - \widehat{\Pi}V||_n$ in Equation 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 19 in the appendix 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 (Equation 17) 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

$$||\widetilde{V} - V||_\rho \leq \widetilde{O}\left(\frac{1}{\sqrt{1-\gamma^2}}||V - \Pi V||_\rho + \frac{1}{1-\gamma}\frac{1}{\sqrt{n}}\right).$$

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 and Van Roy (1997). The additional multiplicative constant $4\sqrt{2}$ in front of the approximation error is the standard cost to have the improved rate bounds for the squared loss and linear spaces (see, e.g., Györfi et al., 2002). In fact, it is possible to derive a bounds with constant 1 but a worse rate $n^{-1/4}$ instead of $n^{-1/2}$. The bound in Theorem 5 is more accurate whenever the approximation error is small and few samples are available.

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 would be interesting to compare the bound in Theorem 5 to the one in Lemma 11 of Antos et al. (2008). In Theorem 5, 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 16 in the appendix). 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 $O(1/n)$. Moreover, although Antos et al. (2008) showed that the solution of MBR minimization coincides with the LSTD solution, its sample-based analysis cannot be directly applied to 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) may be extended to the truncated solution 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 the trajectory $X_1, \dots, X_n$ 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, that is, $X_1 \sim \rho$. 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 $X_1, \dots, X_n$ 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 20 in the appendix), for any initial distribution $\lambda \in \mathcal{S}(\mathcal{X})$, we have

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

where $||\cdot||_{TV}$ is the total variation.[3]

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 6 *Let $X_1, \dots, X_n$ be a trajectory generated by a β-mixing Markov chain with parameters β, b, κ and stationary distribution ρ. Let $\tilde{n}$ ($1 \leq \tilde{n} < n$) be such that $n - \tilde{n}$ satisfies the condition of Equation 13, and $X_{\tilde{n}+1}, \dots, X_n$ be the samples actually used by the algorithm. Let $\omega > 0$ 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 $\widetilde{V}$ be the truncation of the pathwise LSTD solution (using Equation 12), then by setting $\tilde{n} = \left(\frac{1}{b} \log \frac{2e\bar{\beta}n}{\delta}\right)^{1/\kappa}$, with probability $1 - \delta$, we have*

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

where ε_1 and ε_2 are defined as in Theorem 5 (with $n - \tilde{n}$ as the number of training samples).

The proof of this result is a simple consequence of Lemma 24 in the appendix applied to Theorem 5.

Remark 1 The bound in Equation 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. Finite-Sample Analysis of LSPI

In the previous sections we studied the performance of pathwise-LSTD for policy evaluation. Now we move to the analysis of the least-squares policy iteration (LSPI) algorithm (Lagoudakis and Parr, 2003) in which at each iteration k samples are collected by following a single trajectory of the

3. We recall that for any two distributions $\mu_1, \mu_2 \in \mathcal{S}(\mathcal{X})$, the total variation norm is defined as $||\mu_1 - \mu_2||_{TV} = \sup_{X \subseteq \mathcal{X}} |\mu_1(X) - \mu_2(X)|$.

policy under evaluation, π_k, and LSTD is used to compute an approximation of V^{π_k}. In particular, in the next section we report a performance bound by comparing the value of the policy returned by the algorithm after K iterations, V^{π_K}, and the optimal value function, V^*, w.r.t. an arbitrary target distribution σ. In order to achieve this bound we introduce assumptions on the MDP and the linear space $\mathcal{F}$. In Section 6.2 we show that in some cases one of these assumptions does not hold and the performance of LSPI can be arbitrarily bad.

6.1 Generalization Bound for LSPI

In this section, we provide a performance bound for the LSPI algorithm (Lagoudakis and Parr, 2003). We first introduce the *greedy policy* operator $\mathcal{G}$ that maps value functions to their corresponding greedy policies:

$$\big(\mathcal{G}(V)\big)(x) = \arg\max_{a\in\mathcal{A}} \left[r(x,a) + \gamma \int_{\mathcal{X}} P(dy|x,a)V(y) \right].$$

We use $\mathcal{G}(\mathcal{F})$ to refer to the set of all the greedy policies w.r.t. the functions in $\mathcal{F}$. LSPI is a policy iteration algorithm that uses LSTD for policy evaluation at each iteration. It starts with an arbitrary initial value function $V_{-1} \in \widetilde{\mathcal{F}}$ and its corresponding greedy policy π_0. At the first iteration, it approximates V^{π_0} using LSTD and returns a function V_0 whose truncated version $\widetilde{V}_0$ is used to build the policy π_1 for the second iteration.[4] More precisely, π_1 is the greedy policy w.r.t. $\widetilde{V}_0$, that is, $\pi_1 = \mathcal{G}(\widetilde{V}_0)$. So, at each iteration k of LSPI, a function V_{k-1} is computed as an approximation to $V^{\pi_{k-1}}$, and then truncated, $\widetilde{V}_{k-1}$, and used to build the policy $\pi_k = \mathcal{G}(\widetilde{V}_{k-1})$. Note that the MDP model is needed in order to generate the greedy policy π_k. To avoid the need for the model, we could simply move from LSTD to LSTD-Q. The analysis of LSTD in the previous sections may be easily extended to action-value function, and thus, to LSTD-Q.[5] For simplicity we use value function in the paper and report the LSPI bound in terms of the distance to the optimal value function.

It is important to note that in general the measure used to evaluate the final performance of LSPI, $\sigma \in \mathcal{S}(\mathcal{X})$, might be different than the distribution used to generate the samples at each iteration. Moreover, the LSTD performance bounds of Section 5 require the samples to be collected by following the policy under evaluation. Thus, we make the following assumption.

Assumption 1 (Lower-bounding distribution) *There exists a distribution $\mu \in \mathcal{S}(\mathcal{X})$ such that for any policy π that is greedy w.r.t. a function in the truncated space $\widetilde{\mathcal{F}}$, $\mu \le C\rho^\pi$, where $C < \infty$ is a constant and ρ^π is the stationary distribution of policy π.*

Assumption 2 . (Discounted-average Concentrability of Future-State Distribution [Antos et al., 2008]) *Given the target distribution $\sigma \in \mathcal{S}(\mathcal{X})$ and an arbitrary sequence of policies $\{\pi_m\}_{m\ge1}$, let*

$$c_{\sigma,\mu} = \sup_{\pi_1,\dots,\pi_m} \left\| \frac{d(\mu P^{\pi_1} \dots P^{\pi_m})}{d\sigma} \right\|.$$

4. Unlike in the original formulation of LSPI, here we need to explicitly truncate the function so as to prevent unbounded functions.
5. We point out that moving to LSTD-Q requires the introduction of some exploration to the current policy. In fact, in the on-policy setting, if the policy under evaluation is deterministic, it does not provide any information about the value of actions $a \neq \pi(\cdot)$ and the policy improvement step would always fail. Thus, we need to consider stochastic policies where the current policy is perturbed by an $\varepsilon > 0$ randomization which guarantees that any action has a non-zero probability to be selected in any state.

We define the second-order discounted-average concentrability of future-state distributions as

$$C_{\sigma,\mu} = (1-\gamma)^2 \sum_{m\geq 1} m\gamma^{m-1} c_{\sigma,\mu}(m)$$

and we assume that $C_{\sigma,\mu} < \infty$.

We also need to guarantee that with high probability a unique LSTD solution exists at each iteration of the LSPI algorithm, thus, we make the following assumption.

Assumption 3 (Linear independent features) *Let* $\mu \in \mathcal{S}(\mathcal{X})$ *be the lower-bounding distribution from Assumption 1. We assume that the features* $\phi(\cdot)$ *of the function space* $\mathcal{F}$ *are linearly independent w.r.t.* μ. *In this case, the smallest eigenvalue* ω_μ *of the Gram matrix* $G_\mu \in \mathbb{R}^{d\times d}$ *w.r.t.* μ *is strictly positive.*

Lemma 7 *Under Assumption 3, at each iteration k of LSPI, the smallest eigenvalue* ω_k *of the Gram matrix* G_k *defined according to the stationary distribution* $\rho_k = \rho^{\pi_k}$ *is strictly positive and* $\omega_k \geq \frac{\omega_\mu}{C}$.

Proof Similar to Lemma 4, for any function $f_\alpha \in \mathcal{F}$, we have $||\alpha|| \leq \frac{||f_\alpha||_\mu}{\sqrt{\omega_\mu}}$. Using Assumption 1, $||f_\alpha||_\mu \leq \sqrt{C}\,||f_\alpha||_{\rho_k}$, and thus, $||\alpha|| \leq \sqrt{\frac{C}{\omega_\mu}}\,||f_\alpha||_{\rho_k}$. For the α that is the eigenvector of G_k corresponding to ρ_k, we have $||\alpha|| = \frac{||f_\alpha||_{\rho_k}}{\sqrt{\omega_k}}$. For this value of α, we may write $\frac{||f_\alpha||_{\rho_k}}{\sqrt{\omega_k}} \leq \sqrt{\frac{C}{\omega_\mu}}\,||f_\alpha||_{\rho_k}$, and thus, $\omega_k \geq \frac{\omega_\mu}{C}$, which guarantees that ω_k is strictly positive, because ω_μ is strictly positive according to Assumption 3. ■

Finally, we make the following assumption on the stationary β-mixing processes corresponding to the stationary distributions of the policies encountered at the iterations of the LSPI algorithm.

Assumption 4 (Slower β-mixing process) *We assume that there exists a stationary* β*-mixing process with parameters* β, b, κ, *such that for any policy* π *that is greedy w.r.t. a function in the truncated space* $\widetilde{\mathcal{F}}$, *it is slower than the stationary* β*-mixing process with stationary distribution* ρ^π *(with parameters* $\underline{\beta}_\pi, \underline{b}_\pi, \underline{\kappa}_\pi$*). This means that* β *is larger and b and* κ *are smaller than their counterparts* $\underline{\beta}_\pi$, $\underline{b}_\pi$, *and* $\underline{\kappa}_\pi$ *(see Definition 14).*

Now we may state the main theorem of this section.

Theorem 8 *Let us assume that at each iteration k of the LSPI algorithm, a path of size n is generated from the stationary* β*-mixing process with stationary distribution* $\rho_{k-1} = \rho^{\pi_{k-1}}$. *Let n satisfy the condition in Equation 13 for the slower* β*-mixing process defined in Assumption 4. Let* $V_{-1} \in \widetilde{\mathcal{F}}$ *be an arbitrary initial value function,* $V_0, \ldots, V_{K-1}$ ($\widetilde{V}_0, \ldots, \widetilde{V}_{K-1}$) *be the sequence of value functions (truncated value functions) generated by LSPI after K iterations, and* π_K *be the greedy policy w.r.t. the truncated value function* $\widetilde{V}_{K-1}$. *Then under Assumptions 1- 4, with probability* $1-\delta$ *(w.r.t. the random samples), we have*

$$||V^* - V^{\pi_K}||_\sigma \leq \frac{4\gamma}{(1-\gamma)^2}\left\{(1+\gamma)\sqrt{CC_{\sigma,\mu}}\left[\frac{2}{\sqrt{1-\gamma^2}}\left(2\sqrt{2}E_0(\mathcal{F}) + E_2\right)\right.\right.$$
$$\left.\left. + \frac{2}{1-\gamma}\left(\gamma V_{\max} L\sqrt{\frac{d}{\nu_\mu}}\left(\sqrt{\frac{8\log(8dK/\delta)}{n}} + \frac{1}{n}\right)\right) + E_1\right] + \gamma^{\frac{K-1}{2}} R_{\max}\right\},$$

where

1. $E_0(\mathcal{F}) = \sup_{\pi \in \mathcal{G}(\widetilde{\mathcal{F}})} \inf_{f \in \mathcal{F}} ||f - V^\pi||_{\rho^\pi}$,
2. E_1 *is* ε_1 *from Theorem 5 written for the slower* β*-mixing process defined in Assumption 4,*
3. E_2 *is* ε_2 *from Theorem 5 written for the slower* β*-mixing process defined in Assumption 4 and* $||\alpha^*||$ *replaced by* $\sqrt{\frac{C}{\omega_\mu}}\frac{R_{\max}}{1-\gamma}$*, and*
4. ν_μ *is* ν *from Equation 14 in which* ω *is replaced by* ω_μ *defined in Assumption 3, and the second term is written for the slower* β*-mixing process defined in Assumption 4.*

Remark 1 The previous theorem states a bound on the prediction error when LSPI is stopped after a fixed number K of iterations. The structure of the bound resembles the one in Antos et al. (2008). Unlike policy evaluation, the approximation error $E_0(\mathcal{F})$ now depends on how well the space $\mathcal{F}$ can approximate the target functions V^π obtained in the policy improvement step. While the estimation errors are mostly similar to those in policy evaluation, an additional term of order γ^K is introduced. Finally, we notice that the concentrability terms may significantly amplify the prediction error (see also next remark). Farahmand et al. (2010) recently performed a refined analysis of the propagation of the error in approximate policy iteration and have interesting insights on the concentrability terms.

Remark 2 The most critical issue about Theorem 8 is the validity of Assumptions 1–4. The analysis of LSTD explicitly requires that the samples are collected by following the policy under evaluation, π_k, and the performance is bounded according to its stationary distribution ρ_k. Since the performance of LSPI is assessed w.r.t. a target distribution σ, we need each of the policies encountered through the LSPI process to have a stationary distribution which does not differ too much from σ. Furthermore, since the policies are random (at each iteration k the new policy π_k is greedy w.r.t. the approximation $\widetilde{V}_{k-1}$ which is random because of the sampled trajectory), we need to consider the distance of σ and the stationary distribution of any possible policy generated as greedy w.r.t. a function in the truncated space $\widetilde{\mathcal{F}}$, that is, ρ^π, $\pi \in \mathcal{G}(\widetilde{\mathcal{F}})$. Thus in Assumption 1 we first assume the existence of a distribution μ lower-bounding any possible stationary distribution ρ_k. The existence of μ and the value of the constant C depend on the MDP at hand. In Section 6.2, we provide an example in which the constant C is infinite. In this case, we show that the LSPI performance, when the samples at each iteration are generated according to the stationary distribution of the policy under evaluation, can be arbitrarily bad. A natural way to relax this assumption would be the use of off-policy LSTD in which the samples are collected by following a behavior policy. Nonetheless, we are not aware of any finite-sample analysis for such an algorithm. Another critical term appearing in the bound of LSPI, inherited from Theorem 5, is the maximum of $||\alpha_k^*||$ over the iterations, where α_k^* is such that $f_{\alpha_k^*} = \Pi_{\rho_k} V^{\pi_k}$. Each term $||\alpha_k^*||$ can be bounded whenever the features of the space $\mathcal{F}$ are linearly independent according to the stationary distribution ρ_k. Since α_k^* is a random variable, the features $\{\varphi_i\}_{i=1}^d$ of the space $\mathcal{F}$ should be carefully chosen so as to be linearly independent w.r.t. the lower-bounding distribution μ.

We now prove a lemma that is used in the proof of Theorem 8.

Lemma 9 *Let* π_k *be the greedy policy w.r.t.* $\widetilde{V}_{k-1}$*, that is,* $\pi_k = \mathcal{G}(\widetilde{V}_{k-1})$ *and* ρ^{π_k} *be the stationary distribution of the Markov chain induced by* π_k*. We have*

$$||\widetilde{V}_k - \mathcal{T}^{\pi_k}\widetilde{V}_k||_{\rho^{\pi_k}} \leq (1+\gamma)||\widetilde{V}_k - V^{\pi_k}||_{\rho^{\pi_k}} .$$

Proof [Lemma 9] We first show that $\widetilde{V}_k - \mathcal{T}^{\pi_k}\widetilde{V}_k = (I-\gamma P^{\pi_k})(\widetilde{V}_k - V^{\pi_k})$

$$\begin{aligned}(I-\gamma P^{\pi_k})(\widetilde{V}_k - V^{\pi_k}) &= \widetilde{V}_k - V^{\pi_k} - \gamma P^{\pi_k}\widetilde{V}_k + \gamma P^{\pi_k}V^{\pi_k} = \widetilde{V}_k - V^{\pi_k} - \mathcal{T}^{\pi_k}\widetilde{V}_k + \mathcal{T}^{\pi_k}V^{\pi_k} \\ &= \widetilde{V}_k - V^{\pi_k} - \mathcal{T}^{\pi_k}\widetilde{V}_k + V^{\pi_k} = \widetilde{V}_k - \mathcal{T}^{\pi_k}\widetilde{V}_k\,.\end{aligned}$$

For any distribution $\sigma \in \mathcal{S}(\mathcal{X})$, we may write

$$\begin{aligned}||\widetilde{V}_k - \mathcal{T}^{\pi_k}\widetilde{V}_k||_\sigma &= ||(I-\gamma P^{\pi_k})(\widetilde{V}_k - V^{\pi_k})||_\sigma \le ||I-\gamma P^{\pi_k}||_\sigma ||\widetilde{V}_k - V^{\pi_k}||_\sigma \\ &\le \big(1+\gamma||P^{\pi_k}||_\sigma\big)||\widetilde{V}_k - V^{\pi_k}||_\sigma\end{aligned}$$

If σ is the stationary distribution of π_k, that is, $\sigma = \rho^{\pi_k}$, then $||P^{\pi_k}||_\sigma = 1$ and the claim follows. Note that this theorem holds not only for ℓ_2-norm, but for any ℓ_p-norm, $p \ge 1$. ■

Proof [Theorem 8] Rewriting Lemma 12 in Antos et al. (2008) for V instead of Q, we obtain[6]

$$||V^* - V^{\pi_K}||_\sigma \le \frac{4\gamma}{(1-\gamma)^2}\Big(\sqrt{C_{\sigma,\mu}}\max_{0\le k<K}||\widetilde{V}_k - \mathcal{T}^{\pi_k}\widetilde{V}_k||_\mu + \gamma^{\frac{K-1}{2}}R_{\max}\Big). \tag{20}$$

From Assumption 1, we know that $||\cdot||_\mu \le \sqrt{C}||\cdot||_{\rho_k}$ for any $0 \le k < K$ and thus we may rewrite Equation 20 as

$$||V^* - V^{\pi_K}||_\sigma \le \frac{4\gamma}{(1-\gamma)^2}\Big(\sqrt{CC_{\sigma,\mu}}\max_{0\le k<K}||\widetilde{V}_k - \mathcal{T}^{\pi_k}\widetilde{V}_k||_{\rho_k} + \gamma^{\frac{K-1}{2}}R_{\max}\Big). \tag{21}$$

Using the result of Lemma 9, Equation 21 may be rewritten as

$$||V^* - V^{\pi_K}||_\sigma \le \frac{4\gamma}{(1-\gamma)^2}\Big((1+\gamma)\sqrt{CC_{\sigma,\mu}}\max_{0\le k<K}||\widetilde{V}_k - V^{\pi_k}||_{\rho_k} + \gamma^{\frac{K-1}{2}}R_{\max}\Big). \tag{22}$$

We can now use the result of Theorem 5 (which holds with probability δ/K) and replace $||\widetilde{V}_k - V^{\pi_k}||_{\rho_k}$ with its upper-bound. The next step would be to apply the maximum over k to this upper-bound (the right hand side of Equation 17). There are four terms on the r.h.s. of Equation 17 that depend on k and in following we find a bound for each of them.

1. $||V^{\pi_k} - \Pi_{\rho_k}V^{\pi_k}||_{\rho_k}$: This term can be upper-bounded by $E_0(\mathcal{F})$. This quantity, $E_0(\mathcal{F})$, measures the approximation power of the linear function space $\mathcal{F}$.

2. ε_1: This term only depends on the parameters $\bar{\beta}_k, b_k, \kappa_k$ of the stationary β-mixing process with stationary distribution ρ_k. Using Assumption 4, this term can be upper-bounded by E_1, which is basically ε_1 written for the slower β-mixing process from Assumption 4.

3. ε_2: This term depends on the following k-related terms.

6. The slight difference between Equation 20 and the bound in Lemma 12 of Antos et al. (2008) is due to a small error in Equation 26 of Antos et al. (2008). It can be shown that the r.h.s. of Equation 26 in Antos et al. (2008) is not an upper-bound for the r.h.s. of its previous equation. This can be easily fixed by redefining the coefficients α_k while we make sure that they remain positive and still sum to one. This modification causes two small changes in the final bound: the constant 2 in front of the parenthesis becomes 4 and the power of the γ in front of $R_{\max}$ changes from K/p to $(K-1)/p$.

- The term under the root-square in Equation 18: This term depends on the parameters β_k, b_k, κ_k of the stationary β-mixing process with stationary distribution ρ_k. Similar to ε_1, this term can be upper-bounded by rewriting it for the slower β-mixing process from Assumption 4.
- α_k^*: The coefficient vector α_k^* is such that $f_{\alpha_k^*} = \Pi_{\rho_k} V^{\pi_k}$. This term can be upper-bounded as follows:

$$||\alpha_k^*|| \overset{(a)}{\leq} \frac{||f_{\alpha^*}||_\mu}{\sqrt{\omega_\mu}} \overset{(b)}{\leq} \sqrt{\frac{C}{\omega_\mu}} ||f_{\alpha^*}||_{\rho_k} = \sqrt{\frac{C}{\omega_\mu}} ||\Pi_{\rho_k} V^{\pi_k}||_{\rho_k} \overset{(c)}{\leq} \sqrt{\frac{C}{\omega_\mu}} ||V^{\pi_k}||_{\rho_k}$$
$$\leq \sqrt{\frac{C}{\omega_\mu}} ||V^{\pi_k}||_\infty = \sqrt{\frac{C}{\omega_\mu}}\, V_{\max} = \sqrt{\frac{C}{\omega_\mu}}\, \frac{R_{\max}}{1-\gamma}\,.$$

  (a) Similar to Equation 15, this is true for any function $f_\alpha \in \mathcal{F}$.
  (b) This is an immediate application of Assumption 1.
  (c) We use the fact that the orthogonal projection Π_{ρ_k} is non-expansive for norm $||\cdot||_{\rho_k}$.

4. ν_{ρ_k}: This term depends on the following k-related terms.

   - ω_k: This is the smallest eigenvalue of the Gram matrix G_k defined according to the distribution ρ_k. From Lemma 7, this term can be lower-bounded by ω_μ.
   - The second term on the r.h.s. of Equation 14: This term depends on the parameters β_k, b_k, κ_k of the stationary β-mixing process with stationary distribution ρ_k. Similar to ε_1 and ε_2, this term can be upper-bounded by rewriting it for the slower β-mixing process from Assumption 4.

   By replacing the above lower and upper bounds in Equation 14, we obtain ν_μ which is a lower-bound for any ν_{ρ_k}.

The claim follows by replacing the bounds for the above four terms in Equation 22. ■

6.2 A Negative Result for LSPI

In the previous section we analyzed the performance of LSPI when at each iteration the samples are obtained from a trajectory generated by following the policy under evaluation. In order to bound the performance of LSPI in Theorem 8, we made a strong assumption on all possible stationary distributions that can be obtained at the iterations of the algorithm. Assumption 1 states the existence of a lower-bounding distribution μ for the stationary distribution ρ^π of any policy $\pi \in \mathcal{G}(\widetilde{\mathcal{F}})$. If such a distribution does not exist (C is infinite), the LSPI performance can no longer be bounded. In other words, this result states that in some MDPs, even if at each iteration the target function V^{π_k} is perfectly approximated by $\widehat{V}_k$ under ρ_k-norm, that is, $||V^{\pi_k} - \widehat{V}_k||_{\rho_k} = 0$, the LSPI performance could be arbitrarily bad. In this section we show a very simple MDP in which this is actually the case.

Let consider a finite MDP with $\mathcal{X} = \{x_1, x_2, x_3\}$, $\mathcal{A} = \{a, b\}$, and the reward function r and transition model p as illustrated in Figure 3. As it can be noticed only two policies are available in this MDP: π_a which takes action a in state x_1 and π_b which takes action b in this state. It is easy to verify that the stationary distribution ρ^{π_a} assigns probabilities $\frac{\varepsilon}{1+\varepsilon}$, $\frac{1}{1+\varepsilon}$, and 0 to x_1, x_2, and x_3,

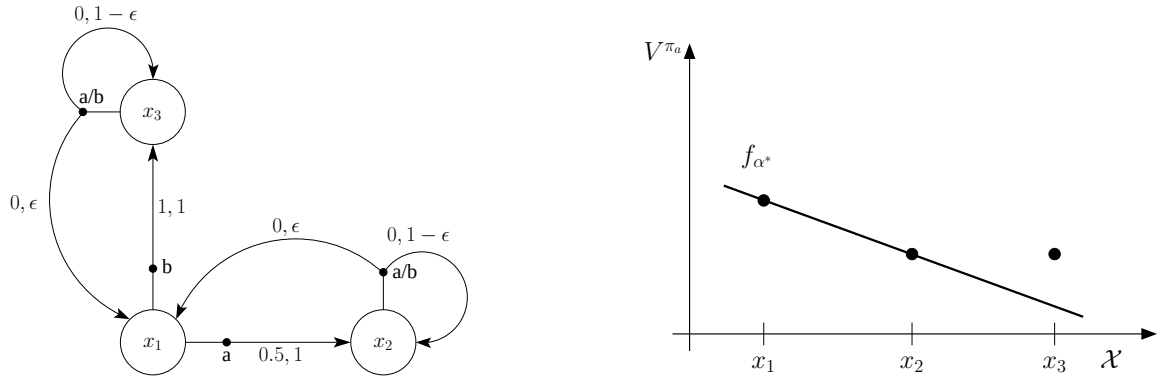

Figure 3: *(left)* The MDP used in the example of Section 6.2 and *(right)* the value function for policy π_a in this MDP.

while ρ^{π_b} has probabilities $\frac{\varepsilon}{1+\varepsilon}$, 0, and $\frac{1}{1+\varepsilon}$. Since ρ^{π_a} and ρ^{π_b} assign a probability 0 to two different states, it is not possible to find a finite constant C such that a distribution μ is lower-bounding both ρ^{π_a} and ρ^{π_b}, thus, $C = \infty$ and according to Theorem 8 LSPI may have an arbitrary bad performance.

Let initialize LSPI with the suboptimal policy π_a. The value function V^{π_a} is shown in Figure 3 (note that the specific values depend on the choice of ε and γ). Let $\mathcal{F} = \{f_\alpha(x) = \alpha_1 x + \alpha_2, \alpha \in \mathbb{R}^2\}$ be the space of lines in dimension 1. Let α^* be the solution to the following minimization problem $\alpha^* = \arg\inf_{\alpha\in\mathbb{R}} ||V^{\pi_a} - f_\alpha||^2_{\rho^{\pi_a}}$ (the projection of V^{π_a} onto space $\mathcal{F}$). Since ρ^{π_a} assigns a probability 0 to state x_3, the f_{α^*} in Figure 3 has a zero loss, that is, $||V^{\pi_a} - f_{\alpha^*}||_{\rho^{\pi_a}} = 0$. Nonetheless, while the greedy policy w.r.t. V^{π_a} is the optimal policy π_b, the policy improvement step w.r.t. f_{α^*} returns the policy π_a. As a result, although at each iteration the function space $\mathcal{F}$ may accurately approximate the value function of the current policy π w.r.t. its stationary distribution ρ^π, LSPI never improves its performance and returns π_a instead of the optimal policy π_b. By properly setting the rewards we could make the performance of π_a arbitrarily worse than π_b.

7. Conclusions

In this paper we presented a finite-sample analysis of the least-squares policy iteration (LSPI) algorithm (Lagoudakis and Parr, 2003). This paper substantially extends the analysis in Lazaric et al. (2010) by reporting all the lemmas used to prove the performance bounds of LSTD in the case of β-mixing and Markov chain processes and by analyzing how the performance of LSTD is propagated through iterations in LSPI.

More in detail, we first studied a version of LSTD, called pathwise LSTD, for policy evaluation. We considered a general setting where we do not make any assumption on the Markov chain. We derived an empirical performance bound that indicates how close the LSTD solution is to the value function at the states along a trajectory generated by following the policy and used by the algorithm. The bound is expressed in terms of the best possible approximation of the value function in the selected linear space (approximation error), and an estimation error which depends on the number of samples and the smallest strictly-positive eigenvalue of the sample-based Gram matrix. We then showed that when the Markov chain possesses a stationary distribution, one may deduce

generalization performance bounds using the stationary distribution of the chain as the generalization measure. In particular, we considered two cases, where the sample trajectory is generated by stationary and non-stationary β-mixing Markov chains, and derived the corresponding bounds. Finally, we considered the whole policy iteration algorithm (LSPI) and showed that under suitable conditions it is possible to bound the error cumulated through the iterations.

The techniques used for the analysis of LSTD have also been recently employed for the development of the finite-sample analysis of a number of novel algorithms such as LSTD with random projections (Ghavamzadeh et al., 2010), LassoTD (Ghavamzadeh et al., 2011), and Classification-based Policy Iteration with a Critic (Gabillon et al., 2011).

Technical issues. From a technical point of view there are two main open issues.

1. *Dependency on ν_n in the bound of Theorem 1.* In Section 4 we introduced the Markov design setting for regression in which the samples are obtained by following a Markov chain and the noise is a zero-mean martingale. By comparing the bound in Lemma 3 with the bounds for least-squares regression in deterministic design (see, e.g., Theorem 11.1 in Györfi et al., 2002), the main difference is the inverse dependency on the eigenvalue ν_n of the empirical Gram matrix. It is not clear whether this dependency is intrinsic in the process generating the samples or whether it can be removed. Abbasi-Yadkori et al. (2011) recently developed improved Azuma's inequalities for self-normalizing process (see also, e.g., de la Peña et al., 2007; de la Peña and Pang, 2009) which suggest that the bound can be improved by removing the dependency from ν_n and, thus, also from the L_∞-norm L of the features.

2. *The $\log n$ dependency in the generalization bounds.* Chaining techniques (Talagrand, 2005) can be successfully applied to remove the $\log n$ dependency in Pollard's inequalities for regression in bounded spaces. An interesting question is whether similar techniques can be applied to the refined analysis for squared losses and linear spaces (see, e.g., Lemma 10) used in our theorems.

Extensions. Some extensions to the current work are possible.

1. *LSTD(λ).* A popular improvement to LSTD is the use of eligibility traces, thus obtaining LSTD(λ). The extension of the results presented in this paper to this setting does not seem to be straightforward since the regression problem solved in LSTD(λ) does not match the Markov design setting introduced in Definition 2. Hence, it is an open question how a finite-sample analysis of LSTD(λ) could be derived.

2. *Off-policy LSTD.* Yu and Bertsekas (2010) derived new bounds for projected linear equations substituting the $\frac{1}{\sqrt{1-\gamma^2}}$ term in front of the approximation error with a much sharper term depending on the spectral radius of some matrices defined by the problem. An open question is whether these new bounds can be effectively reused in the finite-sample analysis derived in this paper, thus obtaining much sharper bounds.

3. *Joint analysis of BRM and LSTD.* Scherrer (2010) recently proposed a unified view of Bellman residual minimization (BRM) (Schweitzer and Seidmann, 1985; Baird, 1995) and temporal difference methods through the notion of oblique projections. This suggests the possibility that the finite-sample analysis of LSTD could be extended to BRM through this unified view over the two methods.

Acknowledgments

We would like to thank Odalric Maillard for useful discussions. This work was supported by French National Research Agency (ANR) through the projects EXPLO-RA n° ANR-08-COSI-004 and LAMPADA n° ANR-09-EMER-007, by Ministry of Higher Education and Research, Nord-Pas de Calais Regional Council and FEDER through the "contrat de projets état region (CPER) 2007–2013", European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 270327, and by PASCAL2 European Network of Excellence through PASCAL2 Pump Priming Programme.

Appendix A.

In this appendix we report a series of lemmata which are used throughout the paper. In particular, we derive concentration of measures inequalities for linear spaces and squared loss when samples are generated from different stochastic processes. We start with the traditional setting of independent and identically distributed samples in Section A.1, then move to samples generated from mixing processes in Section A.2, and finally consider the more general case of samples obtained by simulating a fast mixing Markov chain starting from an arbitrary distribution in Section A.3.

As a general rule, we use *proposition* to indicate results which are copied from other sources, while *lemma* refers to completely or partially new results.

A.1 IID Samples

Although in the setting considered in the paper the samples are non-i.i.d., we first report functional concentration inequalities for i.i.d. samples which will be later extended to stationary and non-stationary β-mixing processes. We first recall the definition of expected and empirical ℓ_2-norms for a function $f : \mathcal{X} \to \mathbb{R}$

$$\|f\|^2_{X_1^n} = \frac{1}{n}\sum_{t=1}^{n} |f(X_t)|^2 \qquad , \qquad \|f\|^2 = \mathbb{E}\left[|f(X_1)|^2\right].$$

Lemma 10 *Let $\mathcal{F}$ be a class of functions $f : \mathcal{X} \to \mathbb{R}$ bounded in absolute value by B. Let $X_1^n = \{X_1, \dots, X_n\}$ be a sequence of i.i.d. samples. For any $\varepsilon > 0$*

$$\mathbb{P}\left[\exists f \in \mathcal{F} : \|f\| - 2\|f\|_{X_1^n} > \varepsilon\right] \le 3\mathbb{E}\left[\mathcal{N}_2\left(\frac{\sqrt{2}}{24}\varepsilon, \mathcal{F}, X_1^{2n}\right)\right] \exp\left(-\frac{n\varepsilon^2}{288B^2}\right),$$

and

$$\mathbb{P}\left[\exists f \in \mathcal{F} : \|f\|_{X_1^n} - 2\|f\| > \varepsilon\right] \le 3\mathbb{E}\left[\mathcal{N}_2\left(\frac{\sqrt{2}}{24}\varepsilon, \mathcal{F}, X_1^{2n}\right)\right] \exp\left(-\frac{n\varepsilon^2}{288B^2}\right),$$

where $\mathcal{N}_2(\varepsilon, \mathcal{F}, X_1^n)$ is the (L_2, ε)-cover number of the function space $\mathcal{F}$ on the samples X_1^n (see Györfi et al. 2002).

Proof The first statement is proved in Györfi et al. (2002) and the second one can be proved similarly. ■

Proposition 11 *Let $\mathcal{F}$ be a class of linear functions $f : \mathcal{X} \to \mathbb{R}$ of dimension d and $\widetilde{\mathcal{F}}$ be the class of functions obtained by truncating functions $f \in \mathcal{F}$ at a threshold B. Then for any sample $X_1^n = \{X_1, \ldots, X_n\}$ and $\varepsilon > 0$*

$$\mathcal{N}_2\left(\varepsilon, \widetilde{\mathcal{F}}, X_1^n\right) \leq 3\left(\frac{3e(2B)^2}{\varepsilon^2}\right)^{2(d+1)}.$$

Proof Using Theorem 9.4. in Györfi et al. (2002) and the fact that the pseudo-dimension of $\widetilde{\mathcal{F}}$ is the same as $\mathcal{F}$, we have

$$\mathcal{N}_2\left(\varepsilon, \widetilde{\mathcal{F}}, X_1^n\right) \leq 3\left(\frac{2e(2B)^2}{\varepsilon^2}\log\frac{3e(2B)^2}{\varepsilon^2}\right)^{d+1} \leq 3\left(\frac{3e(2B)^2}{\varepsilon^2}\right)^{2(d+1)}.$$

■

We now use Proposition 11 to invert the bound in Lemma 10 for truncated linear spaces.

Corollary 12 *Let $\mathcal{F}$ be a class of linear functions $f : \mathcal{X} \to \mathbb{R}$ of dimension d, $\widetilde{\mathcal{F}}$ be the class of functions obtained by truncating functions $f \in \mathcal{F}$ at a threshold B, and $X_1^n = \{X_1, \ldots, X_n\}$ be a sequence of i.i.d. samples. By inverting the bound of Lemma 10, for any $\tilde{f} \in \widetilde{\mathcal{F}}$, we have*

$$\|\tilde{f}\| - 2\|\tilde{f}\|_{X_1^n} \leq \varepsilon(\delta),$$
$$\|\tilde{f}\|_{X_1^n} - 2\|\tilde{f}\| \leq \varepsilon(\delta),$$

with probability $1-\delta$, where

$$\varepsilon(\delta) = 12B\sqrt{\frac{2\Lambda(n,d,\delta)}{n}}, \tag{23}$$

and $\Lambda(n,d,\delta) = 2(d+1)\log n + \log\frac{e}{\delta} + \log\left(9(12e)^{2(d+1)}\right)$.

Proof In order to prove the corollary it is sufficient to verify that the following inequality holds for the ε defined in Equation 23

$$3\mathbb{E}\left[\mathcal{N}_2\left(\frac{\sqrt{2}}{24}\varepsilon, \widetilde{\mathcal{F}}, X_1^{2n}\right)\right]\exp\left(-\frac{n\varepsilon^2}{288B^2}\right) \leq \delta.$$

Using Proposition 11, we bound the first term as

$$\mathbb{E}\left[\mathcal{N}_2\left(\frac{\sqrt{2}}{24}\varepsilon, \widetilde{\mathcal{F}}, X_1^{2n}\right)\right] \leq 3\left(\frac{C_1}{\varepsilon^2}\right)^{2(d+1)},$$

with $C_1 = 3456eB^2$. Next we notice that $\Lambda(n,d,\delta) \geq 1$ and thus $\varepsilon \geq \sqrt{1/(nC_2)}$ with $C_2 = (288B^2)^{-1}$. Using these bounds in the original inequality and some algebra we obtain

$$\begin{aligned}
3\mathbb{E}\left[\mathcal{N}_2\left(\frac{\sqrt{2}}{24}\varepsilon, \widetilde{\mathcal{F}}, X_1^{2n}\right)\right]\exp\left(-\frac{n\varepsilon^2}{288B^2}\right) &\leq 9\left(\frac{C_1}{\varepsilon^2}\right)^{2(d+1)}\exp\left(-nC_2\varepsilon^2\right) \\
&\leq 9\left(nC_1C_2\right)^{2(d+1)}\exp\left(-C_2 n\frac{\Lambda(n,d,\delta)}{nC_2}\right) \\
&= 9\left(nC_1C_2\right)^{2(d+1)} n^{-2(d+1)}\frac{\delta}{e}\frac{1}{9(C_1C_2)^{2(d+1)}} \\
&= \frac{\delta}{e} \leq \delta.
\end{aligned}$$

■

Non-functional versions of Corollary 12 can be simply obtained by removing the covering number from the statement of Lemma 10.

Corollary 13 *Let $f : \mathcal{X} \to \mathbb{R}$ be a function bounded in absolute value by B and $X_1^n = \{X_1, \dots, X_n\}$ be a sequence of i.i.d. samples. Then*

$$\|f\| - 2\|f\|_{X_1^n} \le \varepsilon(\delta),$$

$$\|f\|_{X_1^n} - 2\|f\| \le \varepsilon(\delta),$$

with probability $1-\delta$, where

$$\varepsilon(\delta) = 12B\sqrt{\frac{2}{n}\log\frac{3}{\delta}}\,.$$

A.2 Stationary β-mixing Processes

We first introduce β-mixing stochastic processes and β-mixing coefficients.

Definition 14 *Let $\{X_t\}_{t\ge1}$ be a stochastic process. Let $X_i^j = \{X_i, X_{i+1}, \dots, X_j\}$ and $\sigma(X_i^j)$ denote the sigma-algebra generated by X_i^j. The i-th β-mixing coefficient of the stochastic process is defined by*

$$\beta_i = \sup_{t\ge1} \mathbb{E}\left[\sup_{B\in\sigma(X_{t+i}^{\infty})} |\mathbb{P}(B|X_1^t) - \mathbb{P}(B)|\right].$$

The process $\{X_t\}_{t\ge1}$ is said to be β-mixing if $\beta_i \to 0$ as $i \to \infty$. In particular, $\{X_t\}_{t\ge1}$ mixes at an exponential rate with parameters β, b, κ if $\beta_i \le \beta\exp(-bi^{\kappa})$. Finally, $\{X_t\}_{t\ge1}$ is strictly stationary if $X_t \sim \nu$ for any $t > 0$.

Let $X_1, \dots, X_n$ be a sequence of samples drawn from a stationary β-mixing process with coefficients $\{\beta_i\}$. We first introduce the blocking technique of Yu (1994). Let us divide the sequence of samples into blocks of size k_n. For simplicity we assume $n = 2m_nk_n$ with $2m_n$ be the number of blocks.[7] For any $1 \le j \le m_n$ we define the set of indexes in an odd and even block respectively as

$$H_j = \{t : 2(j-1)k_n + 1 \le t \le (2j-1)k_n\}, \quad \text{and}$$
$$E_j = \{t : (2j-1)k_n + 1 \le t \le (2j)k_n\}.$$

Let $H = \cup_{j=1}^{m_n} H_j$ and $E = \cup_{j=1}^{m_n} E_j$ be the set of all indexes in the odd and even blocks, respectively. We use $X(H_j) = \{X_t : t \in H_j\}$ and $X(H) = \{X_t : t \in H\}$. We now introduce a ghost sample X' (the size of the ghost sample X' is equal to the number of samples in each block k_n) in each of the odd blocks such that the joint distribution of $X'(H_j)$ is the same as $X(H_j)$ but independent from any other block. In the following, we also use another ghost sample X'' independently generated from the same distribution as X'.

7. The extension to the general case is straightforward.

Proposition 15 (Yu, 1994) *Let $X_1,\dots,X_n$ be a sequence of samples drawn from a stationary β-mixing process with coefficients $\{\beta_i\}$. Let Q, Q' be the distributions of $X(H)$ and $X'(H)$, respectively. For any measurable function $h : \mathcal{X}^{m_n k_n} \to \mathbb{R}$ bounded by B*

$$\left|\mathbb{E}_Q\left[h(X(H))\right] - \mathbb{E}_{Q'}\left[h(X'(H))\right]\right| \le B m_n \beta_{k_n}.$$

Before moving to the extension of Propsition 10 to β mixing processes, we report this technical lemma.

Lemma 16 *Let $\mathcal{F}$ be a class of functions $f : \mathcal{X} \to \mathbb{R}$ bounded in absolute value by B and $X_1,\dots,X_n$ be a sequence of samples drawn from a stationary β-mixing process with coefficients $\{\beta_i\}$. For any $\varepsilon > 0$*

$$\mathbb{P}\left[\exists f \in \mathcal{F} : \|f\| - 2\|f\|_{X_1^n} > \varepsilon\right] \le 2\delta(\sqrt{2}\varepsilon) + 2m_n\beta_{k_n}, \tag{24}$$

$$\mathbb{P}\left[\exists f \in \mathcal{F} : \|f\|_{X_1^n} - 2\sqrt{2}\|f\| > \varepsilon\right] \le 2\delta(\sqrt{2}\varepsilon) + 2m_n\beta_{k_n}, \tag{25}$$

where

$$\delta(\varepsilon) = 3\mathbb{E}\left[\mathcal{N}_2\left(\frac{\sqrt{2}}{24}\varepsilon, \mathcal{F}, X'(H)\cup X''(H)\right)\right]\exp\left(-\frac{m_n\varepsilon^2}{288B^2}\right).$$

Proof Similar to Meir (2000), we first introduce $\overline{\mathcal{F}}$ as the class of block functions $\overline{f} : \mathcal{X}^{k_n} \to \mathbb{R}$ defined as

$$\overline{f}\big(X(H_j)\big)^2 = \frac{1}{k_n}\sum_{t\in H_j} f(X_t)^2.$$

It is interesting to notice that block functions have exactly the same norms as the functions in $\mathcal{F}$. In fact

$$\|\overline{f}\|_{X(H)}^2 = \frac{1}{m_n}\sum_{j=1}^{m_n}|\overline{f}(X(H_j))|^2 = \frac{1}{m_n}\sum_{j=1}^{m_n}\frac{1}{k_n}\sum_{t\in H_j}|f(X_t)|^2 = \|f\|_{X(H)}, \tag{26}$$

and

$$\|\overline{f}\|^2 = \mathbb{E}\left[|\overline{f}(X(H_1))|^2\right] = \frac{1}{k_n}\sum_{t\in H_1}\mathbb{E}\left[|f(X_t)|^2\right] = \mathbb{E}\left[|f(X_1)|^2\right] = \|f\|, \tag{27}$$

where in Equation 27, we used the fact that the process is stationary. We now focus on Equation 24

$$\begin{aligned}
&\mathbb{P}\left[\exists f \in \mathcal{F} : \|f\| - 2\|f\|_{X_1^n} > \varepsilon\right] \\
&\overset{(a)}{\le} \mathbb{P}\left[\exists f \in \mathcal{F} : \|f\| - \big(\|f\|_{X(H)} + \|f\|_{X(E)}\big) > \varepsilon\right] \\
&\overset{(b)}{=} \mathbb{P}\left[\exists f \in \mathcal{F} : \frac{1}{2}\big(\|f\| - 2\|f\|_{X(H)}\big) + \frac{1}{2}\big(\|f\| - 2\|f\|_{X(E)}\big) > \varepsilon\right] \\
&\overset{(c)}{\le} \mathbb{P}\left[\exists f \in \mathcal{F} : \|f\| - 2\|f\|_{X(H)} > 2\varepsilon\right] + \mathbb{P}\left[\exists f \in \mathcal{F} : \|f\| - 2\|f\|_{X(E)} > 2\varepsilon\right] \\
&\overset{(d)}{=} 2\mathbb{P}\left[\exists \overline{f} \in \overline{\mathcal{F}} : \|\overline{f}\| - 2\|\overline{f}\|_{X(H)} > 2\varepsilon\right] \\
&\overset{(e)}{\le} 2\left(\mathbb{P}\left[\exists \overline{f} \in \overline{\mathcal{F}} : \|\overline{f}\| - 2\|\overline{f}\|_{X'(H)} > 2\varepsilon\right] + m_n\beta_{k_n}\right) \\
&\overset{(f)}{\le} 2\delta'(2\varepsilon) + 2m_n\beta_{k_n}.
\end{aligned}$$

(a) We used the inequality $\sqrt{a+b} \geq \frac{1}{\sqrt{2}}(\sqrt{a}+\sqrt{b})$ to split the norm $\|f\|_{X_1^n} \geq \frac{1}{2}\big(\|f\|_{X(H)} + \|f\|_{X(E)}\big)$.
(b) Algebra.
(c) Split the probability.
(d) (1) Since the process is stationary the distribution over the even blocks is the same as the distribution over the odd blocks. (2) From Equations 26 and 27.
(e) Using Proposition 15 with h equals to the indicator function of the event inside the bracket, and the fact that the indicator function is bounded by $B = 1$ and its expected value is equal to the probability of the event.
(f) Lemma 10 on space $\overline{\mathcal{F}}$ where

$$\delta'(\varepsilon) = 3\mathbb{E}\left[\mathcal{N}_2\left(\frac{\sqrt{2}}{24}\varepsilon, \overline{\mathcal{F}}, \{X'(H_j), X''(H_j)\}_{j=1}^{m_n}\right)\right] \exp\left(-\frac{m_n\varepsilon^2}{288B^2}\right),$$

where X'' is a ghost sample independently generated from the same distribution as X'. Now we relate the ℓ_2-covering number of $\overline{\mathcal{F}}$ to the covering number of $\mathcal{F}$. Using the definition of $\bar{f}$ we have

$$\begin{aligned}
||\bar{f} - \bar{g}||^2_{X(H)} &= \frac{1}{m_n}\sum_{j=1}^{m_n}\Big(\bar{f}\big(X(H_j)\big) - \bar{g}\big(X(H_j)\big)\Big)^2 \\
&= \frac{1}{m_n k_n}\sum_{j=1}^{m_n}\left[\Big(\sum_{t\in H_j} f(X_t)^2\Big)^{\frac{1}{2}} - \Big(\sum_{t'\in H_j} g(X_{t'})^2\Big)^{\frac{1}{2}}\right]^2.
\end{aligned}$$

Taking the square and using the Cauchy-Schwarz inequality, each element of the outer summation may be written as

$$\begin{aligned}
&\sum_{t\in H_j}\big(f(X_t)^2 + g(X_t)^2\big) - 2\Big(\sum_{t\in H_j} f(X_t)^2\Big)^{\frac{1}{2}}\Big(\sum_{t'\in H_j} g(X_{t'})^2\Big)^{\frac{1}{2}} \\
&\qquad \leq \sum_{t\in H_j}\big(f(X_t)^2 + g(X_t)^2 - 2f(X_t)g(X_t)\big) = \sum_{t\in H_j}\big(f(X_t) - g(X_t)\big)^2.
\end{aligned}$$

By taking the sum over all the odd blocks we obtain

$$||\bar{f} - \bar{g}||^2_{X(H)} \leq ||f - g||^2_{X(H)},$$

which indicates that $\mathcal{N}_2\big(\varepsilon, \overline{\mathcal{F}}, \{X'(H_j), X''(H_j)\}_{j=1}^{m_n}\big) \leq \mathcal{N}_2\big(\varepsilon, \mathcal{F}, X'(H) \cup X''(H)\big)$. Therefore, we have $\delta'(2\varepsilon) \leq \delta(2\varepsilon) \leq \delta(\sqrt{2}\varepsilon)$, which concludes the proof.

With a similar approach, we can prove Equation 25

$$
\begin{aligned}
&\mathbb{P}\left[\exists f \in \mathcal{F} : \|f\|_{X_1^n} - 2\sqrt{2}\|f\| > \varepsilon\right] \\
&\overset{(a)}{\leq} \mathbb{P}\left[\exists f \in \mathcal{F} : \frac{\sqrt{2}}{2}\left(\|f\|_{X(H)} + \|f\|_{X(E)}\right) - 2\sqrt{2}\|f\| > \varepsilon\right] \\
&\overset{(b)}{=} \mathbb{P}\left[\exists f \in \mathcal{F} : \left(\frac{\sqrt{2}}{2}\|f\|_{X(H)} - \sqrt{2}\|f\|\right) + \left(\frac{\sqrt{2}}{2}\|f\|_{X(E)} - \sqrt{2}\|f\|\right) > \varepsilon\right] \\
&\overset{(c)}{\leq} \mathbb{P}\left[\exists f \in \mathcal{F} : \|f\|_{X(H)} - 2\|f\| > \sqrt{2}\varepsilon\right] + \mathbb{P}\left[\exists f \in \mathcal{F} : \|f\|_{X(E)} - 2\|f\| > \sqrt{2}\varepsilon\right] \\
&\overset{(d)}{=} 2\mathbb{P}\left[\exists \bar{f} \in \overline{\mathcal{F}} : \|\bar{f}\|_{X(H)} - 2\|\bar{f}\| > \sqrt{2}\varepsilon\right] \\
&\overset{(e)}{\leq} 2\left(\mathbb{P}\left[\exists \bar{f} \in \overline{\mathcal{F}} : \|\bar{f}\|_{X'(H)} - 2\|\bar{f}\| > \sqrt{2}\varepsilon\right] + m_n\beta_{k_n}\right) \\
&\overset{(f)}{\leq} 2\delta'(\sqrt{2}\varepsilon) + 2m_n\beta_{k_n} \leq 2\delta(\sqrt{2}\varepsilon) + 2m_n\beta_{k_n}.
\end{aligned}
$$

(a) We used the inequality $\sqrt{a+b} \leq (\sqrt{a}+\sqrt{b})$ to split the norm $\|f\|_{X_1^n} \leq \frac{\sqrt{2}}{2}\left(\|f\|_{X(H)} + \|f\|_{X(E)}\right)$.
(b)-(f) use the same arguments as before. ■

Corollary 17 *Let $\mathcal{F}$ be a class of linear functions $f : \mathcal{X} \to \mathbb{R}$ of dimension d, $\widetilde{\mathcal{F}}$ be the class of functions obtained by truncating functions $f \in \mathcal{F}$ at a threshold B, and $X_1^n = \{X_1, \ldots, X_n\}$ be a sequence of samples drawn from a stationary exponentially fast β-mixing process with coefficients $\{\beta_i\}$. By inverting the bound of Lemma 16, for any $\tilde{f} \in \widetilde{\mathcal{F}}$ we have*

$$\|\tilde{f}\| - 2\|\tilde{f}\|_{X_1^n} \leq \varepsilon(\delta),$$

$$\|\tilde{f}\|_{X_1^n} - 2\sqrt{2}\|\tilde{f}\| \leq \varepsilon(\delta),$$

with probability $1-\delta$, where

$$\varepsilon(\delta) = 12B\sqrt{\frac{2\Lambda(n,d,\delta)}{n}\max\left\{\frac{\Lambda(n,d,\delta)}{b}, 1\right\}^{1/\kappa}}, \tag{28}$$

and $\Lambda(n,d,\delta) = 2(d+1)\log n + \log\frac{e}{\delta} + \log^+\left(\max\{18(6e)^{2(d+1)}, \bar{\beta}\}\right)$.

Proof In order to prove the statement, we need to verify that ε in Equation 28 satisfies

$$\delta' = 6\mathbb{E}\left[\mathcal{N}_2\left(\frac{1}{12}\varepsilon, \widetilde{\mathcal{F}}, X'(H) \cup X''(H)\right)\right]\exp\left(-\frac{m_n\varepsilon^2}{144B^2}\right) + 2m_n\beta_{k_n} \leq \delta\,.$$

Using Proposition 11 the covering number can be bounded by

$$\mathbb{E}\left[\mathcal{N}_2\left(\frac{1}{12}\varepsilon, \widetilde{\mathcal{F}}, X'(H) \cup X''(H)\right)\right] \leq 3\left(\frac{1728eB^2}{\varepsilon^2}\right)^{2(d+1)}.$$

By recalling the definition of the β-coefficients $\{\beta_i\}$ and $k_n \geq 1$ we have

$$2m_n\beta_{k_n} \leq \frac{n}{k_n}\bar{\beta}\exp(-bk_n^\kappa) \leq n\bar{\beta}\exp(-bk_n^\kappa)\,.$$

From the last two inequalities, $m_n = n/2k_n$, setting $C_1 = 1728eB^2$ and $D = 2(d+1)$ we obtain

$$\delta' \leq 18\left(\frac{C_1}{\varepsilon^2}\right)^D \exp\left(-\frac{n\varepsilon^2}{144B^2}\frac{1}{2k_n}\right) + n\bar{\beta}\exp(-bk_n^\kappa).$$

By equalizing the arguments of the two exponential we obtain the definition of k_n as

$$k_n = \left\lceil \left(\frac{nC_2\varepsilon^2}{b}\right)^{\frac{1}{\kappa+1}} \right\rceil,$$

where $C_2 = (576B^2)^{-1}$, which implies

$$\max\left\{\left(\frac{nC_2\varepsilon^2}{b}\right)^{\frac{1}{\kappa+1}},1\right\} \leq k_n \leq \max\left\{\left(\frac{2nC_2\varepsilon^2}{b}\right)^{\frac{1}{\kappa+1}},1\right\}.$$

Thus we have the bound

$$\frac{1}{2k_n} \geq \frac{1}{4}\min\left\{\left(\frac{b}{nC_2\varepsilon^2}\right)^{\frac{1}{\kappa+1}},2\right\} \geq \frac{1}{4}\min\left\{\left(\frac{b}{nC_2\varepsilon^2}\right)^{\frac{1}{\kappa+1}},1\right\}.$$

Using the above inequalities, we may write δ' as

$$\delta' \leq 18\left(\frac{C_1}{\varepsilon^2}\right)^D \exp\left(-\min\left\{\frac{b}{nC_2\varepsilon^2},1\right\}^{\frac{1}{\kappa+1}} nC_2\varepsilon^2\right) + n\bar{\beta}\exp\left(-b\max\left\{\frac{nC_2\varepsilon^2}{b},1\right\}^{\frac{\kappa}{\kappa+1}}\right).$$

The objective now is to make the arguments of the two exponential equal. For the second argument we have

$$b\max\left\{\frac{nC_2\varepsilon^2}{b},1\right\}^{\frac{\kappa}{\kappa+1}} = b\max\left\{\frac{nC_2\varepsilon^2}{b},1\right\}\min\left\{\frac{b}{nC_2\varepsilon^2},1\right\}^{\frac{1}{\kappa+1}} \geq nC_2\varepsilon^2\min\left\{\frac{b}{nC_2\varepsilon^2},1\right\}^{\frac{1}{\kappa+1}}.$$

Thus

$$\delta' \leq \left(18\left(\frac{C_1}{\varepsilon^2}\right)^D + n\bar{\beta}\right)\exp\left(-\min\left\{\frac{b}{nC_2\varepsilon^2},1\right\}^{\frac{1}{\kappa+1}} nC_2\varepsilon^2\right).$$

Now we plug in ε from Equation 28. Using the fact that $\Lambda \geq 1$, we know that $\varepsilon^2 \geq (nC_2)^{-1}$, and thus

$$\delta' \leq \left(18\,(nC_1C_2)^D + n\bar{\beta}\right)\exp\left(-\Lambda\right).$$

Using the definition of Λ, we obtain

$$\delta' \leq \left(18\,(nC_1C_2)^D + n\bar{\beta}\right)n^{-D}\max\{18(C_1C_2)^D,\bar{\beta}\}^{-1}\frac{\delta}{e} \leq (1+n^{1-D})\frac{\delta}{e} \leq (1+1)\frac{\delta}{e} \leq \delta\,,$$

which concludes the proof. ■

In order to understand better the shape of the estimation error, we consider a simple β-mixing process with parameters $\bar{\beta} = b = \kappa = 1$. Equation 28 reduces to

$$\varepsilon(\delta) = \sqrt{\frac{288B^2\Lambda(n,d,\delta)}{n}^2},$$

with $\Lambda(n,d,\delta) = 2(d+1)\log n + \log\frac{e}{\delta} + \log\left(18(6e)^{2(d+1)}\right)$. It is interesting to notice that the shape of the bound in this case resembles the structure of the bound in Corollary 12 for i.i.d. samples. Finally, we report the non-functional version of the previous corollary.

Corollary 18 *Let $\mathcal{F}$ be a class of linear functions $f : \mathcal{X} \to \mathbb{R}$ of dimension d such that its features $\varphi_i : \mathcal{X} \to \mathbb{R}$ are bounded in absolute value by L for any $i = 1,\dots,d$ and $X_1^n = \{X_1,\dots,X_n\}$ be a sequence of samples drawn from a stationary exponentially fast β-mixing process with coefficients $\{\beta_i\}$. For any $f \in \mathcal{F}$ we have*

$$\|f\| - 2\|f\|_{X_1^n} \le \varepsilon(\delta),$$

$$\|f\|_{X_1^n} - 2\sqrt{2}\|f\| \le \varepsilon(\delta),$$

with probability $1-\delta$, where

$$\varepsilon(\delta) = 12||\alpha||L\sqrt{\frac{2\Lambda(n,d,\delta)}{n}\max\left\{\frac{\Lambda(n,d,\delta)}{b},1\right\}^{1/\kappa}},$$

and $\Lambda(n,d,\delta) = 2(d+1)\log n + \log\frac{e}{\delta} + \log^+\left(\max\{18(6e)^{2(d+1)},\bar{\beta}\}\right)$.

Proof Let $\mathcal{G} = \left\{g_\alpha = \frac{f_\alpha}{L||\alpha||}\right\}$ so that

$$||g_\alpha||_\infty = \frac{1}{L||\alpha||}||f_\alpha||_\infty \le \frac{1}{L||\alpha||}||\alpha||\sup_i||\varphi_i(x)||_\infty \le 1.$$

We can thus apply Lemma 16 to the bounded space $\mathcal{G}$ with $B = 1$. By using a similar inversion as in Corollary 17, we thus obtain that with probability $1-\delta$, for any function $g_\alpha \in \mathcal{G}$

$$\|g_\alpha\| - 2\|g_\alpha\|_{X_1^n} \le \varepsilon(\delta),$$
$$\|g_\alpha\|_{X_1^n} - 2\sqrt{2}\|g_\alpha\| \le \varepsilon(\delta),$$

with

$$\varepsilon(\delta) = 12\sqrt{\frac{2\Lambda(n,d,\delta)}{n}\max\left\{\frac{\Lambda(n,d,\delta)}{b},1\right\}^{1/\kappa}}.$$

Finally, we notice that $||g_\alpha|| = \frac{1}{L||\alpha||}||f_\alpha||$ and $||g_\alpha||_{X_1^n} = \frac{1}{L||\alpha||}||f_\alpha||_{X_1^n}$ and the statement follows. ■

Corollary 19 *Let $f : \mathcal{X} \to \mathbb{R}$ be a linear function, $\tilde{f}$ be its truncation at a threshold B, and $X_1^n = \{X_1, \dots, X_n\}$ be a sequence of samples drawn from a stationary exponentially fast β-mixing process with coefficients $\{\beta_i\}$. Then*

$$\|\tilde{f}\| - 2\|\tilde{f}\|_{X_1^n} \le \varepsilon(\delta),$$
$$\|\tilde{f}\|_{X_1^n} - 2\sqrt{2}\|\tilde{f}\| \le \varepsilon(\delta),$$

with probability $1-\delta$, where

$$\varepsilon(\delta) = 12B\sqrt{\frac{2\Lambda(n,\delta)}{n}\max\left\{\frac{\Lambda(n,\delta)}{b}, 1\right\}^{1/\kappa}},$$

$\Lambda(n,\delta) = \log\frac{e}{\delta} + \log\left(\max\{6, n\bar{\beta}\}\right)$.

Proof The proof follows the same steps as in Corollary 17. We have the following sequence of inequalities

$$\delta' \le 6\exp\left(-\frac{nC_2\varepsilon^2}{k_n}\right) + \frac{n}{k_n}\bar{\beta}\exp(-bk_n^\kappa) \le (6+n\bar{\beta})\exp(-\Lambda)$$
$$= (6+n\bar{\beta})\max\{6, n\bar{\beta}\}^{-1}\frac{\delta}{e} \le (1+1)\frac{\delta}{e} \le \delta,$$

where $C_2 = (576B^2)^{-1}$. ■

A.3 Markov Chains

We first review the conditions for the convergence of Markov chains (Theorem 13.3.3. in Meyn and Tweedie 1993).

Proposition 20 *Let $\mathcal{M}$ be an ergodic and aperiodic Markov chain defined on $\mathcal{X}$ with stationary distribution ρ. If $P(A|x)$ is the transition kernel of $\mathcal{M}$ with $A \subseteq \mathcal{X}$ and $x \in \mathcal{X}$, then for any initial distribution λ*

$$\lim_{i\to\infty}\left|\left|\int_{\mathcal{X}} \lambda(dx)P^i(\cdot|x) - \rho(\cdot)\right|\right|_{TV} = 0,$$

where $||\cdot||_{TV}$ is the total variation norm.

Definition 21 *Let $\mathcal{M}$ be an ergodic and aperiodic Markov chain with stationary distribution ρ. $\mathcal{M}$ is mixing with an exponential rate with parameters $\bar{\beta}, b, \kappa$, if its β-mixing coefficients $\{\beta_i\}$ satisfy $\beta_i \le \bar{\beta}\exp(-bi^\kappa)$. Then for any initial distribution λ*

$$\left|\left|\int_{\mathcal{X}} \lambda(dx)P^i(\cdot|x) - \rho(\cdot)\right|\right|_{TV} \le \bar{\beta}\exp(-bi^\kappa).$$

Lemma 22 *Let $\mathcal{M}$ be an ergodic and aperiodic Markov chain with a stationary distribution ρ. Let $X_1, \dots, X_n$ be a sequence of samples drawn from the stationary distribution of the Markov chain ρ and $X'_1, \dots, X'_n$ be a sequence of samples such that $X'_1 \sim \rho'$ and $X'_{1<t\le n}$ are generated by simulating $\mathcal{M}$ from X'_1. Let η be an event defined on $\mathcal{X}^n$, then*

$$\left|\mathbb{P}\left[\eta(X_1, \dots, X_n)\right] - \mathbb{P}\left[\eta(X'_1, \dots, X'_n)\right]\right| \le \|\rho' - \rho\|_{TV}$$

Proof We prove one side of the inequality. Let Q be the conditional joint distribution of $(X_{1<t\leq n}|X_1 = x)$ and Q' be the conditional joint distribution of $(X'_{1<t\leq n}|X'_1 = x)$. We first notice that Q is exactly the same as Q'. In fact, the first sequence $(X_{1<t\leq n})$ is generated by drawing X_1 from the stationary distribution ρ and then following the Markov chain. Similarly, the second sequence $(X'_{1<t\leq n})$ is obtained following the Markov chain from $X'_1 \sim \rho'$. As a result, the conditional distributions of the two sequences is exactly the same and just depend on the Markov chain. As a result, we obtain the following sequence of inequalities

$$\begin{aligned}
\mathbb{P}\Big[\eta(X_1,\ldots,X_n)\Big] &= \mathbb{E}_{X_1,\ldots,X_n}\left[\mathbb{I}\left\{\eta(X_1,\ldots,X_n)\right\}\right] \\
&= \mathbb{E}_{X_1\sim\rho}\left[\mathbb{E}_{X_2,\ldots,X_n}\left[\mathbb{I}\left\{\eta(X_1,X_2\ldots,X_n)\right\}|X_1\right]\right] \\
&= \mathbb{E}_{X_1\sim\rho}\left[\mathbb{E}_{X'_2,\ldots,X'_n}\left[\mathbb{I}\left\{\eta(X_1,X'_2\ldots,X'_n)\right\}|X_1\right]\right] \\
&\overset{(a)}{\leq} \mathbb{E}_{X_1\sim\rho'}\left[\mathbb{E}_{X'_2,\ldots,X'_n}\left[\mathbb{I}\left\{\eta(X_1,X'_2\ldots,X'_n)\right\}|X_1\right]\right] + \|\rho'-\rho\|_{TV} \\
&\overset{(b)}{=} \mathbb{E}_{X'_1\sim\rho'}\left[\mathbb{E}_{X'_2,\ldots,X'_n}\left[\mathbb{I}\left\{\eta(X'_1,X'_2\ldots,X'_n)\right\}|X'_1\right]\right] + \|\rho'-\rho\|_{TV} \\
&= \mathbb{P}\Big[\eta(X'_1,\ldots,X'_n)\Big] + \|\rho'-\rho\|_{TV}.
\end{aligned}$$

Note that $\mathbb{I}\{\cdot\}$ is the indicator function.
(a) simply follows from

$$\begin{aligned}
\mathbb{E}_{X\sim\rho}[f(X)] - \mathbb{E}_{X\sim\rho'}[f(X)] &= \int_{\mathcal{X}} f(x)\rho(dx) - \int_{\mathcal{X}} f(x)\rho'(dx) \\
&\leq ||f||_\infty \int_{\mathcal{X}} \big(\rho(dx)-\rho'(dx)\big) \leq ||f||_\infty ||\rho-\rho'||_{TV}.
\end{aligned}$$

(b) From the fact that $X_1 = X'_1 = x$. ■

Lemma 23 *Let $\mathcal{F}$ be a class of functions $f : \mathcal{X} \to \mathbb{R}$ bounded in absolute value by B, $\mathcal{M}$ be a an ergodic and aperiodic Markov chain with a stationary distribution ρ. Let $\mathcal{M}$ be mixing with an exponential rate with parameters β, b, κ. Let λ be an initial distribution over $\mathcal{X}$ and $X_1,\ldots,X_n$ be a sequence of samples such that $X_1 \sim \lambda$ and $X_{1<t\leq n}$ obtained by following $\mathcal{M}$ from X_1. For any $\varepsilon > 0$,*

$$\mathbb{P}\Big[\exists f \in \mathcal{F} : \|f\| - 2\|f\|_{X_1^n} > \varepsilon\Big] \leq \|\lambda-\rho\|_{TV} + 2\delta(\sqrt{2}\varepsilon) + 2m_n\beta_{k_n},$$

and

$$\mathbb{P}\Big[\exists f \in \mathcal{F} : \|f\|_{X_1^n} - 2\sqrt{2}\|f\| > \varepsilon\Big] \leq \|\lambda-\rho\|_{TV} + 2\delta(\sqrt{2}\varepsilon) + 2m_n\beta_{k_n},$$

where

$$\delta(\varepsilon) = 3\mathbb{E}\left[\mathcal{N}_2\left(\frac{\sqrt{2}}{24}\varepsilon, \mathcal{F}, X(H)\cup X'(H)\right)\right]\exp\left(-\frac{m_n\varepsilon^2}{288B^2}\right).$$

Proof The proof is an immediate consequence of Lemma 16 and Lemma 22 by defining $\eta(X_1,\ldots,X_n)$ as

$$\eta(X_1,\ldots,X_n) = \{\exists f \in \mathcal{F} : \|f\| - 2\|f\|_{X_1^n} > \varepsilon\},$$

and

$$\eta(X_1,\dots,X_n) = \{\exists f \in \mathcal{F} : \|f\|_{X_1^n} - 2\sqrt{2}\|f\| > \varepsilon\},$$

respectively. ■

Finally, we consider a special case in which out of the n total number of samples, $\tilde{n}$ ($1 \le \tilde{n} < n$) are used to "burn" the chain and $n-\tilde{n}$ are actually used as training samples.

Lemma 24 *Let $\mathcal{F}$ be a class of linear functions $f : \mathcal{X} \to \mathbb{R}$ of dimension d and $\widetilde{\mathcal{F}}$ be the class of functions obtained by truncating functions $f \in \mathcal{F}$ at a threshold B. Let $\mathcal{M}$ be an ergodic and aperiodic Markov chain with a stationary distribution ρ. Let $\mathcal{M}$ be mixing with an exponential rate with parameters β, b, κ. Let μ be the initial distribution and $X_1,\dots,X_n$ be a sequence of samples such that $X_1 \sim \mu$ and $X_{1<t\le n}$ obtained by following $\mathcal{M}$ from X_1. If the first $\tilde{n}$ ($1 \le \tilde{n} < n$) samples are used to burn the chain and $n-\tilde{n}$ are actually used as training samples, by inverting Lemma 23, for any $\tilde{f} \in \widetilde{\mathcal{F}}$, we obtain*

$$\|\tilde{f}\| - 2\|\tilde{f}\|_{X_1^n} \le \varepsilon(\delta),$$

$$\|\tilde{f}\|_{X_1^n} - 2\sqrt{2}\|\tilde{f}\| \le \varepsilon(\delta),$$

with probability $1-\delta$, where

$$\varepsilon(\delta) = 12B\sqrt{\frac{2\Lambda(n-\tilde{n},d,\delta)}{(n-\tilde{n})}\max\left\{\frac{\Lambda(n-\tilde{n},d,\delta)}{b},1\right\}^{1/\kappa}},$$

and $\Lambda(n,d,\delta) = 2(d+1)\log n + \log\frac{e}{\delta} + \log^+\left(\max\{18(6e)^{2(d+1)},\bar{\beta}\}\right)$, and $\tilde{n} = \left(\frac{1}{b}\log\frac{2e\bar{\beta}n}{\delta}\right)^{1/\kappa}$.

Proof After $\tilde{n}$ steps, the first sample used in the training set ($X_{\tilde{n}+1}$) is drawn from the distribution $\lambda = \mu P^{\tilde{n}}$. Using Proposition 20 and Definition 21 we have

$$||\lambda - \rho||_{TV} \le \bar{\beta}\exp(-b\tilde{n}^{\kappa}). \tag{29}$$

We first substitute the total variation in Lemma 23 with the bound in Equation 29, and then verify that ε in Equation 24 satisfies the following inequality.

$$\begin{aligned}
\delta' &= \|\lambda-\rho\|_{TV} + 2\delta(\sqrt{2}\varepsilon) + 2m_{n-\tilde{n}}\beta_{k_{n-\tilde{n}}} \\
&\le \bar{\beta}\exp(-b\tilde{n}^{\kappa}) + 18\left(\frac{C_1}{\varepsilon^2}\right)^{D}\exp\left(-\frac{(n-\tilde{n})C_2\varepsilon^2}{k_{n-\tilde{n}}}\right) + (n-\tilde{n})\bar{\beta}\exp(-bk_{n-\tilde{n}}^{\kappa}) \\
&\le \left(\frac{1}{2n} + 1 + (n-\tilde{n})^{1-D}\right)\frac{\delta}{e} \le \left(\frac{1}{2}+1+1\right)\frac{\delta}{e} \le \delta,
\end{aligned}$$

where $C_1 = 1728eB^2$ and $C_2 = (288B^2)^{-1}$. The above inequality can be verified by following the same steps as in Corollary 17 and by optimizing the bound for $\tilde{n}$. ■

References

Y. Abbasi-Yadkori, D. Pal, and Cs. Szepesvári. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*, 2011.

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

L. Baird. Residual algorithms: Reinforcement learning with function approximation. In *Proceedings of the Twelfth International Conference on Machine Learning*, pages 30–37, 1995.

D. Bertsekas. *Dynamic Programming and Optimal Control, volume II*. Athena Scientific, 2007.

D. Bertsekas and J. Tsitsiklis. *Neuro-Dynamic Programming*. Athena Scientific, 1996.

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

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

V. de la Peña and G. Pang. Exponential inequalities for self-normlized processes with applications. *Electronic Communications in Probability*, 14:372–381, 2009.

V. de la Peña, M. Klass, and T. Leung Lai. Pseudo-maximization and self-normalized processes. *Propability Surveys*, 4:172–192, 2007.

S. Delattre and S. Gaïffas. Nonparametric regression with martingale increment errors. *Stochastic Processes and their Applications*, 121(12):2899 – 2924, 2011.

A. M. Farahmand, M. Ghavamzadeh, Cs. Szepesvári, and S. Mannor. Regularized policy iteration. In *Advances in Neural Information Processing Systems 21*, pages 441–448. MIT Press, 2008.

A. M. Farahmand, R. Munos, and Cs. Szepesvári. Error propagation for approximate policy and value iteration. In *Advances in Neural Information Processing Systems*, 2010.

V. Gabillon, A. Lazaric, M. Ghavamzadeh, and B. Scherrer. Classification-based policy iteration with a critic. In *Proceedings of the Twenty-Eighth International Conference on Machine Learning*, pages 1049–1056, 2011.

M. Ghavamzadeh, A. Lazaric, O. Maillard, and R. Munos. Lstd with random projections. In *Advances in Neural Information Processing Systems*, pages 721–729, 2010.

M. Ghavamzadeh, A. Lazaric, R. Munos, and M. Hoffman. Finite-sample analysis of lasso-td. In *Proceedings of the 28th International Conference on Machine Learning*, pages 1177–1184, 2011.

L. Györfi, M. Kohler, A. Krzyżak, and H. Walk. *A Distribution-free Theory of Nonparametric Regression*. Springer-Verlag, New York, 2002.

D. Hsu, S. Kakade, and T. Zhang. Random design analysis of ridge regression. In *Proceedings of the 25th Conference on Learning Theory*, 2012.

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

A. Lazaric, M. Ghavamzadeh, and R. Munos. Finite-sample analysis of lstd. In *Proceedings of the 27th International Conference on Machine Learning*, pages 615–622, 2010.

R. Meir. Nonparametric time series prediction through adaptive model selection. *Machine Learning*, 39(1):5–34, April 2000.

S. P. Meyn and R. L. Tweedie. *Markov Chains and Stochastic Stability*. Springer–Verlag, 1993.

B. Ávila Pires and Cs. Szepesvári. Statistical linear estimation with penalized estimators: an application to reinforcement learning. In *Proceedings of the 29th International Conference on Machine Learning*, 2012.

B. Scherrer. Should one compute the temporal difference fix point or minimize the bellman residual? the unified oblique projection view. In *Proceedings of the 27th International Conference on Machine Learning*, pages 959–966, 2010.

P. Schweitzer and A. Seidmann. Generalized polynomial approximations in markovian decision processes. *Journal of Mathematical Analysis and Applications*, 110:568–582, 1985.

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

M. Talagrand. *The Generic Chaining: Upper and Lower Bounds of Stochastic Processes*. Springer–Verlag, 2005.

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

B. Yu. Rates of convergence for empirical processes of stationary mixing sequences. *The Annals of Probability*, 22(1):94–116, January 1994.

H. Yu. Convergence of least squares temporal difference methods under general conditions. In *Proceedings of the 27th International Conference on Machine Learning*, pages 1207–1214, 2010.

H. Yu and D. Bertsekas. Error bounds for approximations from projected linear equations. *Math. Oper. Res.*, 35(2):306–329, 2010.