On the rate of convergence and error bounds for LSTD(λ)

Manel Tagorti, Bruno Scherrer

On the Rate of Convergence and Error Bounds for LSTD(λ)

Manel Tagorti MANEL.TAGORTI@INRIA.FR
Bruno Scherrer BRUNO.SCHERRER@INRIA.FR
Inria, Villers-lès-Nancy, F-54600, France
Université de Lorraine, LORIA, UMR 7503, Vandœuvre-lès-Nancy, F-54506, France

Abstract

We consider LSTD(λ), the least-squares temporal-difference algorithm with eligibility traces algorithm proposed by Boyan (2002). It computes a linear approximation of the value function of a fixed policy in a large Markov Decision Process. Under a β-mixing assumption, we derive, for any value of $\lambda \in (0, 1)$, a high-probability bound on the rate of convergence of this algorithm to its limit. We deduce a high-probability bound on the error of this algorithm, that extends (and slightly improves) that derived by Lazaric et al. (2012) in the specific case where $\lambda = 0$. In the context of temporal-difference algorithms with value function approximation, this analysis is to our knowledge the first to provide insight on the choice of the eligibility-trace parameter λ with respect to the approximation quality of the space and the number of samples.

1. Introduction

In a large Markov Decision Process context, we consider LSTD(λ), the least-squares temporal-difference algorithm with eligibility traces proposed by Boyan (2002). It is a popular algorithm for performing a projection onto a linear space of the value function of a fixed policy. Such a value estimation procedure can for instance be useful in a policy iteration context to eventually estimate an approximately optimal controller (Bertsekas & Tsitsiklis, 1996; Szepesvári, 2010).

The asymptotic almost sure convergence of LSTD(λ) was proved by Nedic & Bertsekas (2002). Under a β-mixing assumption, and given a finite number of samples n, Lazaric et al. (2012) derived a high-probability error bound with a $\tilde{O}(\frac{1}{\sqrt{n}})$ rate[1] in the restricted situation where $\lambda = 0$. Pires & Szepesvári (2012) also sketch an analysis of variations of LSTD(0) with several sorts of regularizations. To our knowledge, however, similar finite-sample error bounds are not known in the literature for $\lambda > 0$. The main goal of this paper is to fill this gap. This is all the more important that it is known that the parameter λ allows to control the quality of the asymptotic solution of the value: by moving λ from 0 to 1, one can continuously move from an oblique projection of the value (Scherrer, 2010) to its orthogonal projection and consequently improve the corresponding guarantee (Tsitsiklis & Roy, 1997) (restated in Theorem 2, Section 3).

The paper is organized as follows. Section 2 starts by describing the necessary background. Section 3 then contains our main results. Theorem 1 shows that unpenalized LSTD(λ) converges to its limit at the rate $\tilde{O}(\frac{1}{\sqrt{n}})$. We then deduce a global error (Corollary 1) that sheds some light on the role of the parameter λ, and discuss some of its practical consequences. Theorem 3 then extends this result to the case of penalized LSTD(λ). Section 4 will go on by providing a detailed proof of our claims. Finally, Section 5 concludes by describing related and potential future work.

2. LSTD(λ) and Related Background

We consider a Markov chain taking its values on a finite or countable state space $\mathcal{X}$, with transition kernel P, and that is ergodic[2]; consequently, it admits a unique stationary distribution μ. For any $K \in \mathbb{R}^+$, we denote $\mathcal{B}(\mathcal{X}, K)$ the set of functions defined on $\mathcal{X}$ and bounded by K. We consider a reward function $r \in \mathcal{B}(\mathcal{X}, R_{\max})$ for some $R_{\max} \in \mathbb{R}$, that provides the quality of being in some state. The value function v related to the Markov chain is defined, for any state i, as the average discounted sum of rewards along infinitely

[1] Throughout the paper, we shall write $f(n) = \tilde{O}(g(n))$ as a shorthand for $f(n) = O(g(n) \log^k g(n))$ for some $k \geq 0$.

[2] We focus on finite/countable state spaces essentially because it eases the presentation. We believe that extensions to more general state spaces is straight-forward.

Proceedings of the 32^{nd} International Conference on Machine Learning, Lille, France, 2015. JMLR: W&CP volume 37.

long trajectories starting from i:

$$\forall i \in \mathcal{X},\ v(i) = \mathbb{E}\left[\sum_{j=0}^{\infty} \gamma^j r(X_j) \big| X_0 = i\right],$$

where $\gamma \in (0,1)$ is a discount factor. It is well-known that the value function v is the unique fixed point of the linear Bellman operator T: $\forall i \in \mathcal{X},\ Tv(i) = r(i) + \gamma \mathbb{E}\left[v(X_1)|X_0 = i\right]$. It can easily be seen that $v \in \mathcal{B}(\mathcal{X}, V_{\max})$ with $V_{\max} = \frac{R_{\max}}{1-\gamma}$.

When the size $|\mathcal{X}|$ of the state space is very large, one may consider approximating v by using *a linear architecture*. Given some d (typically $d \ll |\mathcal{X}|$), we consider a feature matrix $\Phi = (\phi(x))_{x\in\mathcal{X}} = (\phi_1 \dots \phi_d)$ of dimension $|\mathcal{X}| \times d$. For any $x \in \mathcal{X}$, $\phi(x) = (\phi_1(x), ..., \phi_d(x))^T$ is the *feature vector* in state x. For any $j \in \{1, ..., d\}$, we assume that the *feature function* $\phi_j : \mathcal{X} \mapsto \mathbb{R}$ belongs to $\mathcal{B}(\mathcal{X}, L)$ for some finite L. Throughout the paper, we will make the following assumption.

Assumption 1. *The feature vectors $(\phi_j)_{j\in\{1,...,d\}}$ are linearly independent.*

Let $\mathcal{S}$ be the subspace generated by the vectors $(\phi_j)_{1\le j\le d}$. We consider the orthogonal projection Π onto $\mathcal{S}$ with respect to the μ-weighed quadratic norm $\|f\|_\mu = \sqrt{\sum_{x\in\mathcal{X}} f(x)^2 \mu(x)}$. It is well known that this projection has the following closed form

$$\Pi = \Phi(\Phi^T D_\mu \Phi)^{-1} \Phi^T D_\mu, \tag{1}$$

where D_μ is the diagonal matrix with elements of μ on the diagonal, and for all u, u^T denotes the transpose of u.

The goal of LSTD(λ) is to estimate a solution of the equation $v = \Pi T^\lambda v$, where the operator T^λ is defined as a geometric average of the applications of the powers T^i of the Bellman operator T for all $i > 1$:

$$\forall \lambda \in (0,1),\ \forall v,\ T^\lambda v = (1-\lambda)\sum_{i=0}^{\infty} \lambda^i T^{i+1} v. \tag{2}$$

Note in particular that when $\lambda = 0$, one has $T^\lambda = T$. By using the facts that T^i is affine and $\|P\|_\mu = 1$ (Tsitsiklis & Roy, 1997), it has been shown that the operator T^λ is a contraction mapping of modulus $\frac{(1-\lambda)\gamma}{1-\lambda\gamma} \le \gamma$ (Nedic & Bertsekas, 2002). Since the orthogonal projector Π is non-expansive with respect to μ (Tsitsiklis & Roy, 1997), the operator ΠT^λ is contracting and thus the equation $v = \Pi T^\lambda v$ has one and only one solution, which we shall denote $v_{LSTD(\lambda)}$ since it is what the LSTD(λ) algorithm converges to (Nedic & Bertsekas, 2002). As $v_{LSTD(\lambda)}$ belongs to the subspace $\mathcal{S}$, there exists a $\theta \in \mathbb{R}^d$ such that $v_{LSTD(\lambda)} = \Phi\theta = \Pi T^\lambda \Phi\theta$. If we replace Π and T^λ with their expressions (Equations 1 and 2), it can be seen that θ is a solution of the equation $A\theta = b$ (Tsitsiklis & Roy, 1997; Nedic & Bertsekas, 2002) where for any i,

$$A = \Phi^T D_\mu (I - \gamma P)(I - \lambda\gamma P)^{-1}\Phi \tag{3}$$

$$= \mathbb{E}\left[\sum_{k=-\infty}^{i} (\gamma\lambda)^{i-k}\phi(X_k)(\phi(X_i) - \gamma\phi(X_{i+1}))^T\right] \tag{4}$$

$$\text{and}\quad b = \Phi^T D_\mu (I - \gamma\lambda P)^{-1} r$$

$$= \mathbb{E}\left[\sum_{k=-\infty}^{i} (\gamma\lambda)^{i-k}\phi(X_k) r(X_i)\right], \tag{5}$$

where the sum starts from $-\infty$ to ensure that the process (X_k) is in stationary regime. Since for all x, $\phi(x)$ is of dimension d, we see that A is a $d \times d$ matrix and b is a vector of size d. Under Assumption 1, it can be shown (Nedic & Bertsekas, 2002) that the matrix A is invertible, and thus $v_{LSTD(\lambda)} = \Phi A^{-1} b$ is well defined.

The LSTD(λ) algorithm that is the focus of this article is now precisely described. Given one trajectory $X_1, ..., X_n$ generated by the Markov chain, the expectation-based expressions of A and b in Equations (4)-(5) suggest to compute the following estimates:

$$\hat{A} = \frac{1}{n-1}\sum_{i=1}^{n-1} z_i(\phi(X_i) - \gamma\phi(X_{i+1}))^T$$

$$\text{and}\quad \hat{b} = \frac{1}{n-1}\sum_{i=1}^{n-1} z_i r(X_i)$$

$$\text{where}\quad z_i = \sum_{k=1}^{i} (\lambda\gamma)^{i-k}\phi(X_k) \tag{6}$$

is the so-called *eligibility trace*. The algorithm then returns $\hat{v}_{LSTD(\lambda)} = \Phi\hat{\theta}$ with[3] $\hat{\theta} = \hat{A}^{-1}\hat{b}$, which is a (finite sample) approximation of $v_{LSTD(\lambda)}$. Using a variation of the law of large numbers, Nedic & Bertsekas (2002) showed that both $\hat{A}$ and $\hat{b}$ converge almost surely respectively to A and b, which implies that $\hat{v}_{LSTD(\lambda)}$ tends to $v_{LSTD(\lambda)}$. The main goal of this paper is to deepen this analysis: we shall estimate a bound on the rate of convergence of $\hat{v}_{LSTD(\lambda)}$ to $v_{LSTD(\lambda)}$, and bound the error $\|\hat{v}_{LSTD(\lambda)} - v\|_\mu$ of the overall algorithm.

3. Main result

This section contains our main results. Our key assumption for the analysis is that the Markov chain process that

[3] We will see in Theorem 1 that $\hat{A}$ is invertible with high probability for a sufficiently big n.

generates the states has some mixing property[4].

Assumption 2. *The process $(X_n)_{n\geq 1}$ is β-mixing, in the sense that its i^{th} coefficient $\beta_i = \sup_{t\geq 1} \mathbb{E}\left[\sup_{B\in\sigma(X_{t+i}^{\infty})} |P(B|\sigma(X_1^t)) - P(B)|\right]$ tends to 0 when i tends to infinity, where $X_l^j = \{X_l, ..., X_j\}$ for $j \geq l$ and $\sigma(X_l^j)$ is the sigma algebra generated by X_l^j. Furthermore, $(X_n)_{n\geq 1}$ mixes at an exponential decay rate with parameters $\overline{\beta} > 0$, $b > 0$, and $\kappa > 0$ in the sense that $\beta_i \leq \overline{\beta} e^{-bi^{\kappa}}$.*

Intuitively the β_i coefficients measure the degree of dependence of samples separated by i time steps (the smaller the coefficient the more independence). We are now ready to state the main results of the paper, which provides a rate of convergence of LSTD(λ).

Theorem 1. *Let Assumptions 1 and 2 hold and let $X_1 \sim \mu$, where μ is the stationary distribution of the chain. For any $n \geq 1$ and $\delta \in (0,1)$, define $I(n,\delta) = 32\Lambda(n,\delta)\max\left\{\frac{\Lambda(n,\delta)}{b}, 1\right\}^{\frac{1}{\kappa}}$, where $\Lambda(n,\delta) = \log\left(\frac{8n^2}{\delta}\right) + \log(\max\{4e^2, n\overline{\beta}\})$. Also define the positive integer $m_n^{\lambda} = \left\lceil \frac{\log(n-1)}{\log\frac{1}{\lambda\gamma}} \right\rceil$. Let $n_0(\delta)$ be the smallest integer such that for all $n \geq n_0(\delta)$,*

$$\frac{2dL^2}{(1-\gamma)\nu}\left[\frac{2}{\sqrt{n-1}}\sqrt{(m_n^{\lambda}+1)I(n-1,\delta)} + \frac{1}{(n-1)(1-\lambda\gamma)} + \frac{2}{(n-1)}m_n^{\lambda}\right] < 1 \qquad (7)$$

where ν is the smallest eigenvalue of the Gram matrix $\Phi^T D_{\mu}\Phi$. Then, for all δ, with probability at least $1-\delta$, for all $n \geq n_0(\delta)$, $\hat{A}$ is invertible and the distance $\|v_{LSTD(\lambda)} - \hat{v}_{LSTD(\lambda)}\|_{\mu}$ is upper bounded by

$$\frac{4V_{max}dL^2}{\sqrt{n-1}(1-\gamma)\nu}\sqrt{(m_n^{\lambda}+1)\,I(n-1,\delta)} + h(n,\delta)$$

with $h(n,\delta) = \tilde{O}(\frac{1}{n}\log\frac{1}{\delta})$.

The constant ν is positive under Assumption 1. For all δ, it is clear that the finite constant $n_0(\delta)$ exists since the l.h.s. of Equation (7) tends to 0 when n tends to infinity. As m_n^{λ} and $I(n-1,\delta)$ are of order $\tilde{O}(1)$, we can see that LSTD(λ) estimates $v_{LSTD(\lambda)}$ at a rate $\tilde{O}\left(\frac{1}{\sqrt{n}}\right)$. Finally, we can observe that since $\lambda \mapsto m_n^{\lambda}$ is increasing, the rate of convergence deteriorates when λ increases. This negative effect can be balanced by the fact that, as shown by the following result from the literature, the quality of $v_{LSTD(\lambda)}$ improves when λ increases.

[4] A Markov chain that is ergodic and stationary is *always* β-mixing (Bradley, 2005).

Theorem 2 (Tsitsiklis & Roy (1997))**.** *The approximation error satisfies*

$$\|v - v_{LSTD(\lambda)}\|_{\mu} \leq \frac{1-\lambda\gamma}{1-\gamma}\|v - \Pi v\|_{\mu}.$$

Since the constant equals 1 when $\lambda = 1$, one recovers the well-known fact that LSTD(1) computes the orthogonal projection Πv of v. By using the triangle inequality, one deduces from Theorems 1 and 2 the following global error bound.

Corollary 1. *Let the assumptions and notations of Theorem 1 hold. For all δ, with probability at least $1-\delta$, for all $n \geq n_0(\delta)$, the global error of LSTD(λ) satisfies:*

$$\|v - \hat{v}_{LSTD(\lambda)}\|_{\mu} \leq \frac{1-\lambda\gamma}{1-\gamma}\|v - \Pi v\|_{\mu} + \frac{4V_{max}dL^2}{\sqrt{n-1}(1-\gamma)\nu}\sqrt{(m_n^{\lambda}+1)\,I(n-1,\delta)} + h(n,\delta).$$

The bound requires a sufficiently large number of samples n ($n \geq n_0(\delta)$). For a fixed δ, this number increases when λ increases. The existence of such a condition is not surprising since we focus on an unregularized version of LSTD(λ), and thus the estimated matrix $\hat{A}$ may not be invertible when n is too small.

As we have already mentioned, $\lambda = 1$ minimizes the bound on the approximation error $\|v - v_{LSTD(\lambda)}\|_{\mu}$ (the first term in the r.h.s. in Corollary 1) while $\lambda = 0$ minimizes the bound on the estimation error $\|v_{LSTD(\lambda)} - \hat{v}_{LSTD(\lambda)}\|_{\mu}$ (the second term). For any δ and $n \geq n_0(\delta)$, there exists a value λ^* that minimizes the global error bound by making an optimal compromise between the approximation and estimation errors upper-bounds. When the number of samples n tend to infinity, the optimal value λ^* tends to 1. Previous studies on the role of the parameter λ were to our knowledge empirical (Sutton & Barto, 1998; Downey & Sanner, 2010) or dedicated to an exact representation of the value function (Kearns & Singh, 2000). This is the first time a bound on a temporal-difference learning algorithm with value function approximation shows this trade-off explicitely.

The form of the result stated in Corollary 1 is slightly stronger than the one of Lazaric et al. (2012). It has the advantage to make clear the connection with the previous analysis of Nedic & Bertsekas (2002) since our formulation implies the almost sure convergence of $\hat{v}_{LSTD(\lambda)}$ to $v_{LSTD(\lambda)}$: for some property $P(n)$, our result is of the form "$\forall\delta$, $\exists n_0(\delta)$, such that with probability at least $1-\delta, \forall n \geq n_0(\delta)$, $P(n)$ holds" while the result stated by (Lazaric et al., 2012) is of the form "$\forall n$, $\exists\delta(n)$, such that with probability at least $1-\delta(n)$, $P(n)$ holds." In other words, we can fix a real δ such

that the property is true for all $n \geq n_0(\delta)$ with probability at least $1-\delta$, while in (Lazaric et al., 2012), δ depends on the number of samples.

Pires & Szepesvári (2012) studied penalized versions of linear systems estimated with noise, and explained how to apply their approach to LSTD(0). Such a penalization allows to control the magnitude of $\hat{\theta}$ in situations where the matrix $\hat{A}$ is (close to) singular. This has the advantage of removing the need for a condition on the number of samples to ensure the invertibility of $\hat{A}$, and as a side effect this allows to derive bounds that are valid for any value of the probability threshold δ and number of samples n (while in the above mentionned result without penalization, the minimum number of samples $n_0(\delta)$ grows to infinity when δ approaches 0). The most natural penalization that one would like to consider for LSTD(λ) is the one where we add a term ρI to the estimate $\hat{A}$ (Nedic & Bertsekas, 2002). This amounts to solve the following penalized problem: $\hat{\theta}_\rho = \arg\min_\theta \left\{ \|\hat{A}\theta - \hat{b}\|_2^2 + \rho\|\theta\|_2^2 \right\}$. Unfortunately, this very form of regularization—squared error with squared penalty—is not considered by Pires & Szepesvári (2012). It turns out that it is rather straight-forward to bound the residual $\|A\hat{\theta}_\rho - b\|_2$ in this case by following an approach very similar to that described in Pires & Szepesvári (2012). Combined with the analyisis performed for Theorem 1, we can derive the following result.

Theorem 3. *Under Assumptions 1 and 2, for any $\delta \in (0,1)$ and n consider the estimate $\hat{v}_{LSTD(\lambda)}^{\rho_{n,\delta}} = \Phi\hat{\theta}_\rho$ obtained with penalization parameter $\rho_{n,\delta} = 2\Xi^2(n,\delta)$ s.t.*

$$\Xi(n,\delta) = \frac{4dL^2}{(1-\lambda\gamma)\sqrt{n-1}}\sqrt{(m_\lambda^n+1)I\left(n-1,\frac{2n^2\delta}{3}\right)} + \frac{2dL^2}{(n-1)(1-\lambda\gamma)^2} + \frac{4dL^2 m_n^\lambda}{(n-1)(1-\lambda\gamma)}.$$

Then, with probability at least $1-\delta$, for all n, $\|\hat{v}_{LSTD(\lambda)}^{\rho_{n,\delta}} - v_{LSTD(\lambda)}\|_\mu$ is bounded by

$$\frac{4V_{max}\sqrt{d}L(3+\sqrt{d}L)}{\sqrt{n-1}(1-\gamma)\sqrt{\nu}}\sqrt{(m_n^\lambda+1)\,I(n-1,\delta)} + g(n,\delta),$$

where $g(n,\delta)$ and $I(n,\delta)$ and m_n^λ are defined as in Theorem 1.

We defer the proof to Appendix B of the supplementary material.

4. Proof of Theorem 1

This section provides a detailed proof of Theorem 1. The proof is organized in four steps. In the first step, we study the sensitivity of the solution $v_{LSTD(\lambda)}$ to a potential deterministic deviation of the estimates $\hat{A}$ and $\hat{b}$ from their limits A and b. In the second step, we shall derive a general concentration analysis to control with high probability the deviations of processes defined through infinitely-long eligibility traces. Then, in the third step, we will apply this concentration analysis to $\hat{A}$ and $\hat{b}$. Finally, we will gather all elements to deduce the high-probability bound on the distance between $\hat{v}_{LSTD(\lambda)}$ and $v_{LSTD(\lambda)}$.

4.1. Deterministic sensivity of LSTD(λ)

We begin by showing the following lemma on the sensitivity of LSTD(λ).

Lemma 1. *Write $\epsilon_A = \hat{A} - A$, $\epsilon_b = \hat{b} - b$ and ν the smallest eigenvalue of the matrix $\Phi^T D_\mu \Phi$. For all $\lambda \in (0,1)$, the error $\|v_{LSTD(\lambda)} - \hat{v}_{LSTD(\lambda)}\|_\mu$ is upper bounded by[5]:*

$$\frac{1-\lambda\gamma}{(1-\gamma)\sqrt{\nu}}\|(I+\epsilon_A A^{-1})^{-1}\|_2\|\epsilon_A\theta - \epsilon_b\|_2,$$

where $\theta = A^{-1}b$. Furthermore, if for some ϵ and C, $\|\epsilon_A\|_2 \leq \epsilon < C \leq \frac{1}{\|A^{-1}\|_2}$, then $\hat{A}$ is invertible and

$$\|(I+\epsilon_A A^{-1})^{-1}\|_2 \leq \frac{1}{1-\frac{\epsilon}{C}}.$$

Proof. The definitions of $v_{LSTD(\lambda)}$ and $\hat{v}_{LSTD(\lambda)}$ lead to

$$\hat{v}_{LSTD(\lambda)} - v_{LSTD(\lambda)} = \Phi A^{-1}(A\hat{\theta} - b). \tag{8}$$

On the one hand, with the expression of A in Equation (3), writing $M = (1-\lambda)\gamma P(I-\lambda\gamma P)^{-1}$ and $M_\mu = \Phi^T D_\mu \Phi$, we can see that

$$\begin{aligned}\Phi A^{-1} &= \Phi\left[\Phi^T D_\mu (I-\gamma P)(I-\lambda\gamma P)^{-1}\Phi\right]^{-1} \\ &= \Phi\left[\Phi^T D_\mu (I-\lambda\gamma P-(1-\lambda)\gamma P)(I-\lambda\gamma P)^{-1}\Phi\right]^{-1} \\ &= \Phi(M_\mu - \Phi^T D_\mu M\Phi)^{-1}.\end{aligned}$$

Since the matrices A and M_μ are invertible, the matrix $(I - M_\mu^{-1}\Phi^T D_\mu M\Phi)$ is also invertible and

$$\Phi A^{-1} = \Phi(I - M_\mu^{-1}\Phi^T D_\mu M\Phi)^{-1}M_\mu^{-1}.$$

By definition, the projection matrix Π defined in Equation (1) satisfies $\|\Pi\|_\mu = 1$ and we know from Tsitsiklis & Roy (1997) that the stochastic matrix P of the process also satisfies $\|P\|_\mu = 1$. Hence, we have $\|\Pi M\|_\mu = \frac{(1-\lambda)\gamma}{1-\lambda\gamma} < 1$ and the matrix $(I-\Pi M)$ is invertible. We can use the identity $X(I-YX)^{-1} = (I-XY)^{-1}X$ with $X = \Phi$ and $Y = M_\mu^{-1}\Phi^T D_\mu M$, and obtain

$$\Phi A^{-1} = (I-\Pi M)^{-1}\Phi M_\mu^{-1}. \tag{9}$$

[5] When $\hat{A}$ is not invertible, we have $\hat{v}_{LSTD(\lambda)} = \infty$ and the inequality is always satisfied since, as we will see shortly, the invertiblity of $\hat{A}$ is equivalent to that of $(I+\epsilon_A A^{-1})$.

On the other hand, using the facts that $A\theta = b$ and $\hat{A}\hat{\theta} = \hat{b}$, we can see that

$$\begin{aligned} A\hat{\theta} - b &= A\hat{\theta} - b - (\hat{A}\hat{\theta} - \hat{b}) \\ &= \hat{b} - b - (\hat{A} - A)(\hat{\theta} - \theta) - (\hat{A} - A)\theta \\ &= \hat{b} - \hat{A}\theta - (b - A\theta) + \epsilon_A A^{-1}(A\theta - A\hat{\theta}) \\ &= \hat{b} - \hat{A}\theta - \epsilon_A A^{-1}(A\hat{\theta} - b) \\ &= (I + \epsilon_A A^{-1})^{-1}(\hat{b} - \hat{A}\theta) \\ &= (I + \epsilon_A A^{-1})^{-1}(\epsilon_b - \epsilon_A \theta). \end{aligned} \tag{10}$$

Using Equations (9) and (10), Equation (8) can be rewritten as follows:

$$\begin{aligned} &\hat{v}_{LSTD(\lambda)} - v_{LSTD(\lambda)} \\ &= (I - \Pi M)^{-1}\Phi M_\mu^{-1}(I + \epsilon_A A^{-1})^{-1}(\epsilon_b - \epsilon_A\theta). \end{aligned} \tag{11}$$

We shall now bound $\|\Phi M_\mu^{-1}(I + \epsilon_A A^{-1})^{-1}(\epsilon_b - \epsilon_A\theta)\|_\mu$. Notice that for all x,

$$\begin{aligned} \|\Phi M_\mu^{-1} x\|_\mu &= \sqrt{x^T M_\mu^{-1}\Phi^T D_\mu \Phi M_\mu^{-1} x} \\ &= \sqrt{x^T M_\mu^{-1} x} \leq \frac{1}{\sqrt{\nu}}\|x\|_2 \end{aligned} \tag{12}$$

where ν is the smallest (real) eigenvalue of the Gram matrix M_μ. By taking the norm in Equation (11) and using the above relation, we get

$$\begin{aligned} &\|\hat{v}_{LSTD(\lambda)} - v_{LSTD(\lambda)}\|_\mu \\ &\leq \|(I - \Pi M)^{-1}\|_\mu \|\Phi M_\mu^{-1}(I + \epsilon_A A^{-1})^{-1}(\epsilon_b - \epsilon_A\theta)\|_\mu \\ &\leq \|(I - \Pi M)^{-1}\|_\mu \frac{1}{\sqrt{\nu}}\|(I + \epsilon_A A^{-1})^{-1}\|_2 \|\epsilon_A\theta - \epsilon_b\|_2. \end{aligned}$$

The first part of the lemma is obtained by using the fact that $\|\Pi M\|_\mu = \frac{(1-\lambda)\gamma}{1-\lambda\gamma} < 1$, which implies that

$$\begin{aligned} \|(I - \Pi M)^{-1}\|_\mu &= \left\|\sum_{i=0}^{\infty}(\Pi M)^i\right\|_\mu \leq \sum_{i=0}^{\infty}\|\Pi M\|_\mu^i \\ &\leq \frac{1}{1 - \frac{(1-\lambda)\gamma}{1-\lambda\gamma}} = \frac{1-\lambda\gamma}{1-\gamma}. \end{aligned} \tag{13}$$

We are going now to prove the second part of the lemma. Since A is invertible, the matrix $\hat{A}$ is invertible if and only if the matrix $\hat{A}A^{-1} = (A + \epsilon_A)A^{-1} = I + \epsilon_A A^{-1}$ is invertible. Let us denote $\rho(\epsilon_A A^{-1})$ the spectral radius of the matrix $\epsilon_A A^{-1}$. A sufficient condition for $\hat{A}A^{-1}$ to be invertible is that $\rho(\epsilon_A A^{-1}) < 1$. From the inequality $\rho(M) \leq \|M\|_2$ for any square matrix M, we can see that for any C and ϵ that satisfy $\|\epsilon_A\|_2 \leq \epsilon < C < \frac{1}{\|A^{-1}\|_2}$,

$$\rho(\epsilon_A A^{-1}) \leq \|\epsilon_A A^{-1}\|_2 \leq \|\epsilon_A\|_2\|A^{-1}\|_2 \leq \frac{\epsilon}{C} < 1.$$

It follows that the matrix $\hat{A}$ is invertible and

$$\|(I + \epsilon_A A^{-1})^{-1}\|_2 = \left\|\sum_{i=0}^{\infty}(\epsilon_A A^{-1})^i\right\|_2 \leq \sum_{i=0}^{\infty}\left(\frac{\epsilon}{C}\right)^i$$

This concludes the proof of Lemma 1. □

Lemma 1 suggests that we control both terms $\|\epsilon_A\|_2 = \|\hat{A} - A\|_2$ and $\|\epsilon_b\|_2 = \|\hat{b} - b\|_2$. The next subsection shows how to do so with high probability.

4.2. Concentration inequality for infinitely-long trace-based estimates

As both terms $\hat{A}$ and $\hat{b}$ have the same structure, we will consider here a matrix that has the following general form:

$$\hat{G} = \frac{1}{n-1}\sum_{i=1}^{n-1} G_i \quad \text{with} \quad G_i = z_i(\tau(X_i, X_{i+1}))^T$$

where z_i is the trace defined in Equation (6) and $\tau : \mathcal{X}^2 \to \mathbb{R}^k$. Let $\|.\|_F$ denote the Frobenius norm satisfying: for $M \in \mathbb{R}^{d\times k}$, $\|M\|_F^2 = \sum_{l=1}^{d}\sum_{j=1}^{k}(M_{l,j})^2$. The second important element of our analysis is the following concentration inequality for the infinitely-long-trace β-mixing process $\hat{G}$.

Lemma 2. *Let Assumptions 1 and 2 hold and let $X_1 \sim \mu$. Define the $d \times k$ matrix G_i such that*

$$G_i = \sum_{l=1}^{i}(\lambda\gamma)^{i-l}\phi(X_l)(\tau(X_i, X_{i+1}))^T. \tag{14}$$

Recall that $\phi = (\phi_1, \ldots, \phi_d)$ is such that for all j, $\phi_j \in \mathcal{B}(\mathcal{X}, L)$. Assume that for all $1 \leq j \leq d$, $\tau_j \in \mathcal{B}(\mathcal{X}^2, L')$. Let m_n^λ and $I(n,\delta)$ be defined as in Theorem 1. Let $J(n,\delta) = I(n, 4n^2\delta)$. Then, for all δ in $(0,1)$, with probability at least $1 - \delta$,

$$\begin{aligned} &\left\|\frac{1}{n-1}\sum_{i=1}^{n-1} G_i - \frac{1}{n-1}\sum_{i=1}^{n-1}\mathbb{E}[G_i]\right\|_2 \\ &\leq \frac{2\sqrt{d\times k}LL'}{(1-\lambda\gamma)\sqrt{n-1}}\sqrt{(m_n^\lambda + 1)\,J(n-1,\delta)} + \epsilon(n), \end{aligned}$$

where $\epsilon(n) = 2m_n^\lambda \frac{\sqrt{d\times k}LL'}{(n-1)(1-\lambda\gamma)}$.

Proof. The proof of this result is tedious, so we only give a sketch and defer the details to Appendix A in the Supplementary material. There are two main difficulties regarding the estimates G_i used to compute $\hat{G}$: 1) G_i is a $\sigma(\mathcal{X}^{i+1})$ measurable function of the non-stationary vector $(X_1, \ldots, X_{i+1})$, and is consequently *not stationary*; 2) For all i, G_i are computed from one single trajectory of the Markov chain and are consequently *mutually dependent*.

To deal with the first issue (non-stationarity), we shall consider the m-truncated trace,

$$z_i^m = \sum_{k=\max(i-m+1,1)}^{i} (\lambda\gamma)^{i-k}\phi(X_k),$$

and approximate $\hat{G}$ with the process $\hat{G}^m$ defined as:

$$\hat{G}^m = \frac{1}{n-1}\sum_{i=1}^{n-1} G_i^m, \quad \text{with} \quad G_i^m = z_i^m(\tau(X_i, X_{i+1}))^T.$$

Indeed, G_i^m is now a $\sigma(\mathcal{X}^{m+1})$ measurable function of the *stationary* vector $Z_i = (X_{i-m+1}, X_{i-m+2}, \ldots, X_{i+1})$, the vector Z_i being stationary since we assumed $X_1 \sim \mu$.

To deal with the second issue (dependence of samples), for any possible value of the truncation depth m, we shall use the β-mixing assumption (Assumption 2) to transform the dependent samples G_i^m into blocks of independent samples, by using the "blocking technique" of Yu (1994) in a way somewhat similar to—but technically slightly more involved than—what Lazaric et al. (2012) did for LSTD(0). This being done, we will be able to use a concentration inequality for i.i.d. processes from the literature (Lemma 7 in Appendix A in the Supplementary material). In addition to the use of a truncation depth m, a specific ingredient of the analysis of LSTD(λ) with respect to that of LSTD(0) is that we need to prove that the stationary process $(Z_i)_{i\geq 1} = (X_{i-m+1}, X_{i-m+2}, \ldots, X_{i+1})_{i\geq 1}$ on which the m-truncated process G_i^m is defined, inherits the β-mixing property of the original process $(X_i)_{i\geq 1}$. This is the purpose of the following technical lemma.

Lemma 3. *Let $(X_n)_{n\geq 1}$ be a β-mixing process, then $(Z_n)_{n\geq m} = (X_{n-m+1}, X_{n-m+2}, \ldots, X_{n+1})_{n\geq m}$ is a β-mixing process such that its i^{th} β mixing coefficient β_i^Z satisfies $\beta_i^Z \leq \beta_{i-m}^X$.*

Finally, setting m to m_n^λ will ensure that the distance between $\hat{G}$ and $\hat{G}^m$ is bounded by $\epsilon(n)$ (as defined in Lemma 2), and is therefore neglibible with respect to the result of the deviation analysis obtained by the "blocking techinque" of (Yu, 1994). □

Using a very similar proof, we may derive a (simpler and) general-purpose concentration inequality for β-mixing processes:

Lemma 4. *Let $Y = (Y_1, \ldots, Y_n)$ be random variables taking their values in the space $\mathbb{R}^d$, generated from a stationary exponentially β-mixing process with parameters $\overline{\beta}$, b and κ, and such that for all i, $\|Y_i - \mathbb{E}[Y_i]\|_2 \leq B_2$ almost surely. Then for all $\delta > 0$, with probability at least $1-\delta$,*

$$\left\| \frac{1}{n}\sum_{i=1}^{n} Y_i - \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[Y_i] \right\|_2 \leq \frac{B_2}{\sqrt{n}}\sqrt{J(n,\delta)}$$

where $J(n,\delta)$ is defined as in Lemma 2.

If the variables Y_i were independent, we would have $\beta_i = 0$ for all i, that is we could choose $\overline{\beta} = 0$ and $b = \infty$, so that $J(n,\delta)$ reduces to $32\log\frac{8e^2}{\delta} = O(1)$ and we recover standard concentration results for i.i.d. processes (such as the one we describe in Lemma 7 in Appendix A in the Supplementary material). The price to pay for making a β-mixing assumption (instead of simple independence) lies in the extra coefficient $J(n,\delta)$ which is $\tilde{O}(1)$; in other words, it is rather mild.

4.3. Bounding the deviations of $\hat{A}$ and $\hat{b}$

We shall now apply the concentration inequality of Lemma 2 on the quantities of interest of Lemma 1, i.e. on $\|\epsilon_A\|_2$ and $\|\epsilon_A\theta - \epsilon_b\|_2$.

Bounding $\|\epsilon_A\|_2$. By the triangle inequality, we have

$$\|\epsilon_A\|_2 \leq \|\mathbb{E}[\epsilon_A]\|_2 + \|\epsilon_A - \mathbb{E}[\epsilon_A]\|_2. \tag{15}$$

Write $\hat{A}_{n,k} = \phi(X_k)(\phi(X_n) - \gamma\phi(X_{n+1}))^T$. For all n and k, we have: $\|\hat{A}_{n,k}\|_2 \leq 2dL^2$. We can bound the first term of the r.h.s. of Equation (15) by replacing A with its expression in Equation (4):

$$\begin{aligned}
\|\mathbb{E}[\epsilon_A]\|_2 &= \left\| A - \mathbb{E}\left[\frac{1}{n-1}\sum_{i=1}^{n-1}\sum_{k=1}^{i}(\lambda\gamma)^{i-k}\hat{A}_{i,k}\right]\right\|_2 \\
&= \left\| \mathbb{E}\left[\frac{1}{n-1}\sum_{i=1}^{n-1}\left(\sum_{k=-\infty}^{i}(\lambda\gamma)^{i-k}\hat{A}_{i,k} - \sum_{k=1}^{i}(\lambda\gamma)^{i-k}\hat{A}_{i,k}\right)\right]\right\|_2 \\
&= \left\| \mathbb{E}\left[\frac{1}{n-1}\sum_{i=1}^{n-1}(\lambda\gamma)^{i}\sum_{k=-\infty}^{0}(\lambda\gamma)^{-k}\hat{A}_{i,k}\right]\right\|_2 \\
&\leq \frac{1}{n-1}\sum_{i=1}^{n-1}(\lambda\gamma)^i\frac{2dL^2}{1-\lambda\gamma} \leq \frac{1}{n-1}\frac{2dL^2}{(1-\lambda\gamma)^2} \overset{def}{=} \epsilon_0(n).
\end{aligned} \tag{16}$$

Let $(\delta_n)_{n\geq 1}$ be a sequence in $(0,1)$ that we will set later. With $\epsilon(n) = \frac{4dL^2}{(n-1)(1-\lambda\gamma)}m_n^\lambda$ (defined in Lemma 2) and $\epsilon_0(n)$ defined in Equation (16), define:

$$\begin{aligned}
\epsilon_1(n,\delta_n) =& \frac{4dL^2}{(1-\lambda\gamma)\sqrt{n-1}}\sqrt{(m_n^\lambda+1)\,J(n-1,\delta_n)} \\
&+ \epsilon(n) + \epsilon_0(n).
\end{aligned} \tag{17}$$

By using Equation (15), the bound of Equation (16) and Lemma 2 applied to ϵ_A, we get

$$\begin{aligned}
&\mathbb{P}\left\{\|\epsilon_A\|_2 \geq \epsilon_1(n,\delta_n)\right\} \\
&\leq \mathbb{P}\{\|\epsilon_A - \mathbb{E}[\epsilon_A]\|_2 \geq \epsilon_1(n,\delta_n) - \epsilon_0(n)\} \\
&\leq \delta_n.
\end{aligned} \tag{18}$$

Bounding $\|\epsilon_A\theta-\epsilon_b\|_2$. By using the fact that $A\theta = b$, the definitions of $\hat{A}$ and $\hat{b}$, and the fact that $\phi(x)^T\theta = [\phi\theta](x)$, we have

$$\begin{aligned}
&\epsilon_A\theta - \epsilon_b = \hat{A}\theta - \hat{b} \\
&= \frac{1}{n-1}\sum_{i=1}^{n-1} z_i(\phi(X_i) - \gamma\phi(X_{i+1})^T)\theta - \frac{1}{n-1}\sum_{i=1}^{n-1} z_i r(X_i) \\
&= \frac{1}{n-1}\sum_{i=1}^{n-1} z_i([\phi\theta](X_i) - \gamma[\phi\theta](X_{i+1}) - r(X_i)) \\
&= \frac{1}{n-1}\sum_{i=1}^{n-1} z_i\Delta_i
\end{aligned}$$

where, since $v_{LSTD(\lambda)} = \Phi\theta$, Δ_i is the following number:

$$\Delta_i = v_{LSTD(\lambda)}(X_i) - \gamma v_{LSTD(\lambda)}(X_{i+1}) - r(X_i).$$

Let L' be a bound on $\max_{1\le i\le n-1}|\Delta_i|$ (we shall compute L' below). We can control $\|\epsilon_A\theta - \epsilon_b\|_2$ by following the same proof steps as above. In fact we can see that

$$\begin{aligned}
\|\epsilon_A\theta - \epsilon_b\|_2 \;&\le\; \|\epsilon_A\theta - \epsilon_b - \mathbb{E}[\epsilon_A\theta - \epsilon_b]\|_2 \\
&\quad + \|\mathbb{E}[\epsilon_A\theta - \epsilon_b]\|_2, \qquad (19)\\
\text{with}\;\; \|\mathbb{E}[\epsilon_A\theta - \epsilon_b]\|_2 \;&\le\; \|\mathbb{E}[\epsilon_A]\|_2\|\theta\|_2 + \|\mathbb{E}[\epsilon_b]\|_2.
\end{aligned}$$

From what has been developed before we can see that $\|\mathbb{E}[\epsilon_A]\|_2 \le \epsilon_0(n) = \frac{1}{n-1}\frac{2dL^2}{(1-\lambda\gamma)^2}$. Similarly we can show that $\|\mathbb{E}[\epsilon_b]\|_2 \le \frac{1}{n-1}\frac{\sqrt{d}LR_{\max}}{(1-\lambda\gamma)^2}$. We can hence conclude that

$$\begin{aligned}
&\|\mathbb{E}[\epsilon_A\theta - \epsilon_b]\|_2 \\
&\le \frac{1}{n-1}\frac{2dL^2}{(1-\lambda\gamma)^2}\|\theta\|_2 + \frac{1}{n-1}\frac{\sqrt{d}LR_{\max}}{(1-\lambda\gamma)^2} \overset{def}{=} \epsilon_0'(n). \qquad (20)
\end{aligned}$$

With $\epsilon(n) = \frac{4dL^2}{(n-1)(1-\lambda\gamma)}m_n^\lambda$ (defined in Lemma 2) and $\epsilon_0'(n)$ defined in Equation (20), define:

$$\begin{aligned}
\epsilon_2(n,\delta_n) =& \frac{2\sqrt{d}LL'}{(1-\lambda\gamma)\sqrt{n-1}}\sqrt{(m_n^\lambda+1)\,J(n-1,\delta_n)} \\
&+ \epsilon(n) + \epsilon_0'(n). \qquad (21)
\end{aligned}$$

By using Equation (19), Equation (21) and Lemma 2 applied to $\epsilon_a\theta - b$, we get

$$\begin{aligned}
&\mathbb{P}(\|\epsilon_A\theta - \epsilon_b\|_2 \ge \epsilon_2(n,\delta_n)) \\
&\le \mathbb{P}(\|\epsilon_A\theta - \epsilon_b - \mathbb{E}[\epsilon_A\theta - \epsilon_b]\|_2 \ge \epsilon_2(n,\delta_n) - \epsilon_0'(n)) \\
&\le \delta_n. \qquad (22)
\end{aligned}$$

To finish this third part of the proof, it remains to compute the bound L' on $\max_{1\le i\le n-1}|\Delta_i|$. To do so, it suffices to bound $v_{LSTD(\lambda)}(x)$ for all x. For all $x\in\mathcal{X}$, we have

$$|v_{LSTD(\lambda)}(x)| = |\phi^T(x)\theta| \le \|\phi^T(x)\|_2\|\theta\|_2 \le \sqrt{d}L\|\theta\|_2,$$

where the first inequality is obtained from the Cauchy-Schwarz inequality. It remains to bound $\|\theta\|_2$. On the one hand, we have: $\|v_{LSTD(\lambda)}\|_\mu = \|\Phi\theta\|_\mu = \sqrt{\theta^T M_\mu\theta} \ge \sqrt{\nu}\|\theta\|_2$, and on the other hand, we have: $\|v_{LSTD(\lambda)}\|_\mu = \|(I-\Pi M)^{-1}\Pi(I-\lambda\gamma P)^{-1}r\|_\mu \le \frac{R_{\max}}{1-\gamma} = V_{\max}$. Therefore $\|\theta\|_2 \le \frac{V_{\max}}{\sqrt{\nu}}$, and we can deduce that: $\forall x \in \mathcal{X}$, $|v_{LSTD(\lambda)}(x)| \le \frac{\sqrt{d}LV_{\max}}{\sqrt{\nu}}$. Then, for all i we have

$$\begin{aligned}
|\Delta_i| &= |v_{LSTD(\lambda)}(X_i) - \gamma v_{LSTD(\lambda)}(X_{i+1}) - r(X_i)| \\
&\le \frac{\sqrt{d}LV_{\max}}{\sqrt{\nu}} + \gamma\frac{\sqrt{d}LV_{\max}}{\sqrt{\nu}} + (1-\gamma)V_{\max}.
\end{aligned}$$

Since $\Phi^T D_\mu\Phi$ is a symmetric matrix, we have $\nu \le \|\Phi^T D_\mu\Phi\|_2$. We can see that $\|\Phi^T D_\mu\Phi\|_2 \le d\max_{j,k}|\phi_k^T D_\mu\phi_j| = d\max_{j,k}|\phi_k^T D_\mu^{\frac{1}{2}} D_\mu^{\frac{1}{2}}\phi_j| \le d\max_{j,k}\|\phi_k^T\|_\mu\|\phi_j\|_\mu \le dL^2$, so that $\nu \le dL^2$. It follows that, for all i

$$|\Delta_i| \le \frac{\sqrt{d}LV_{\max}}{\sqrt{\nu}} + \gamma\frac{\sqrt{d}LV_{\max}}{\sqrt{\nu}} + \frac{\sqrt{d}L}{\sqrt{\nu}}(1-\gamma)V_{\max},$$

and therefore we can take $L' = 2\frac{\sqrt{d}L}{\sqrt{\nu}}V_{\max}$.

4.4. Conclusion of the proof

Now that we know how to control both terms $\|\epsilon_A\|_2$ and $\|\epsilon_A\theta - \epsilon_b\|_2$, we are ready to conclude the proof. Consider the event

$$\begin{aligned}
E = \big\{\; &\exists n \ge 1,\; \{\|\epsilon_A\|_2 \ge \epsilon_1(n,\delta_n)\} \\
&\cup \{\|\epsilon_A\theta - \epsilon_b)\|_2 \ge \epsilon_2(n,\delta_n)\}\;\big\}.
\end{aligned}$$

Using the analysis of Section 4.3 and in particular Equations (18) and 22, we deduce that

$$\begin{aligned}
\mathbb{P}(E) &\le \sum_{n=1}^{\infty}\mathbb{P}\left\{\|\epsilon_A\|_2 \ge \epsilon_1(n,\delta_n)\right\} \\
&\qquad + \mathbb{P}\left\{\|\epsilon_A\theta - \epsilon_b)\|_2 \ge \epsilon_2(n,\delta_n)\right\} \\
&\le 2\sum_{n=1}^{\infty}\delta_n = \frac{1}{2}\sum_{n=1}^{\infty}\frac{1}{n^2}\delta = \frac{1}{2}\frac{\pi^2}{6}\delta < \delta
\end{aligned}$$

if on the last line we set $\delta_n = \frac{1}{4n^2}\delta$. By the second part of Lemma 1, for all δ, with probability at least $1-\delta$, for all n such that $\epsilon_1(n,\delta_n) < C$, where C is chosen such that $C \le \frac{1}{\|A^{-1}\|_2}$, then $\hat{A}$ is invertible and

$$\begin{aligned}
\|v_{LSTD(\lambda)} - \hat{v}_{LSTD(\lambda)}\|_\mu &\le \frac{1-\lambda\gamma}{(1-\gamma)\sqrt{\nu}}\frac{\epsilon_2(n,\delta_n)}{1-\frac{\epsilon_1(n,\delta_n)}{C}} \\
&= \frac{1-\lambda\gamma}{(1-\gamma)\sqrt{\nu}}\left[\epsilon_2(n,\delta_n) + \frac{\epsilon_1(n,\delta_n)\,\epsilon_2(n,\delta_n)}{C-\epsilon_1(n,\delta_n)}\right].
\end{aligned}$$

The bound of the Theorem 1 is obtained by replacing $\epsilon_1(n,\delta_n)$ and $\epsilon_2(n,\delta_n)$ with their definitions in Equations (17) and (21), in particularly noticing that $\epsilon(n)$, $\epsilon_0(n)$ and $\epsilon'_0(n)$ are $\tilde{O}(\frac{1}{n})$.

To fully complete the proof of Theorem 1, we finally need to show how to pick $C \leq \frac{1}{\|A^{-1}\|_2}$. We have $\forall v \in \mathbb{R}^d, \|\Phi A^{-1} v\|_\mu = \sqrt{(A^{-1}v)^T M_\mu A^{-1} v} \geq \sqrt{\nu}\|A^{-1}v\|_2$. We know that $\|\Phi A^{-1} v\|_\mu = \|(I - \Pi M)^{-1}\Phi M_\mu^{-1} v\|_\mu \leq \frac{1-\lambda\gamma}{(1-\gamma)\sqrt{\nu}}\|v\|_2$ where the inequalities are respectively obtained from Equations (12) and (13). Therefore $\|A^{-1}\|_2 \leq \frac{1-\lambda\gamma}{(1-\gamma)\nu}$, and consequently we can take $C = \frac{(1-\gamma)\nu}{1-\lambda\gamma}$. Note that the condition $\epsilon_1(n,\delta_n) < C$ for this choice of C is equivalent to the one that characterizes the index $n_0(\delta)$ in the theorem. This concludes the proof of Theorem 1.

5. Summary, Related and Future Work

This paper provides high-probability bound on the convergence rate for the standard LSTD(λ) and a penalized variation, in terms of the number of samples n and the parameter λ. Theorems 1 and 3 show that this convergence is at the rate of $\tilde{O}(\frac{1}{\sqrt{n}})$, in the case where the samples are generated from a stationary β-mixing process. Our result is based on two original technical contributions: a) a deterministic sensitivity analysis of LSTD(λ) (Lemma 1) and b) an original vector concentration inequality (Lemma 2) for estimates that are based on eligibility traces. A simplified version of the latter (Lemma 4) is a general-purpose concentration inequality that may apply to general stationary beta-mixing processes, which may be useful in many other contexts where we want to relax the i.i.d. hypothesis on the samples. Corollary 1, which is an immediate consequence of Theorem 1, is to our knowledge the very first analytical result that provides insight on the choice of the eligibility-trace parameter λ of temporal-difference learning algorithm with respect to the approximation quality of the space and the number of samples. Validating empirically the lessons that we can take from this result constitutes immediate interesting future work.

Under the same assumptions, the global error bound obtained by Lazaric et al. (2012) in the restricted case where $\lambda = 0$ has the following form:

$$\|\tilde{v}_{LSTD(0)} - v\|_\mu \leq \frac{4\sqrt{2}}{1-\gamma}\|v - \Pi v\|_\mu + O\left(\sqrt{\frac{d\log d}{\nu n}}\right),$$

where $\tilde{v}_{LSTD(0)}$ is the truncation with thresholds $\{-V_{\max}, V_{\max}\}$ of the estimate $\hat{v}_{LSTD(0)}$. In our analysis, we get for $\lambda = 0$:

$$\|\hat{v}_{LSTD(0)} - v\|_\mu \leq \frac{1}{1-\gamma}\|v - \Pi v\|_\mu + \tilde{O}\left(\frac{d}{\nu\sqrt{n}}\right).$$

On the one hand, the term corresponding to the approximation error is a factor $4\sqrt{2}$ better with our analysis; our bound is thus asymptotically better. Note that, contrary to our approach, the analysis of Lazaric et al. (2012) does not imply a rate of convergence for LSTD(0) (a bound on $\|v_{LSTD(0)} - \hat{v}_{LSTD(0)}\|_\mu$); their arguments, based on a model of regression with Markov design, consists in *directly* bounding the global error. On the other hand, our bound on the estimation error depends linearly on the features space dimension d and on $\frac{1}{\nu}$ while the one obtained by Lazaric et al. (2012) takes the form of $O\left(\sqrt{d\log d/(n\nu)}\right)$. Thus our bound seems suboptimal on d and ν. A technical element for explaining such a difference is the fact, mentionned above, that Lazaric et al. (2012) consider the truncated version of $v_{LSTD(0)}$. Indeed, a close examination shows that the extra term $\sqrt{d/\nu}$ in our bound results from a bound (uniform on x) on $v_{LSTD(\lambda)}(x)$.

A critical condition in the analysis of LSTD(0) previously done by Lazaric et al. (2012) is that the noise term in the Markov Regression model is a Martingale difference sequence with respect to the filtration generated by the Markov chain. As soon as $\lambda > 0$, this property stops to hold and it has not been clear how one may fix this issue. We believe that the techniques we used for the proof of our concentration inequality (Lemma 2)—the truncation of the trace at some depth m and the focus on the "block" chain $(Z_n) = (X_{i-m+1}, X_{i-m}, \ldots, X_{i+1})$—constitutes a potential track for addressing these issues. If successful, note however that an extension to $\lambda > 0$ of the work of Lazaric et al. (2012) would still contain a suboptimal $4\sqrt{2}$ extra factor in the final bound.

Regarding the dependence with respect to the parameters d and ν, it is worth mentionning that the bound obtained by Pires & Szepesvári (2012) for a regularized version of LSTD(0) depends also linearly on d and $\|\theta\|_2$ (which in turn can be bounded by $V_{\max}/\sqrt{\nu}$). In (Antos et al., 2006) the bound does not depend on ν but the convergence rate is of order $\tilde{O}\left(1/n^{\frac{1}{4}}\right)$ which is a slower rate than the one we get. In the deterministic design and pure regression setting—pure regression corresponds to value function learning with $\gamma = 0$—, the corresponding bound does not also involve the parameter ν (Györfi et al., 2002). We do not know whether one could have the best of all worlds: the best asymptotic bound without the $4\sqrt{2}$ coefficient, and the best rate with respect to n, d and ν. This constitutes interesting future work.

More generally, in the future, we plan to instantiate our new bound in a Policy Iteration context like Lazaric et al. (2012) did for LSTD(0). An interesting follow-up work would also be to extend our analysis of LSTD(λ) to the situation where one considers non-stationary policies, as Scherrer & Lesner (2012) showed that it allows to improve

the overall performance of the Policy Iteration Scheme. Finally, a challenging problem would be to consider convergence rate LSTD(λ) in the off-policy case, for which the convergence has recently been proved by Yu (2010).

Acknowledgments. We thank the anonymous reviewers, whose comments helped to improve the paper. This work was supported by the French National Research Agency (ANR) through the project BARQ.

References

Antos, András, Szepesvári, Csaba, and Munos, Rémi. Learning near-optimal policies with bellman-residual minimization based fitted policy iteration and a single sample path. In *In COLT-19*, pp. 574–588. Springer-Verlag, 2006.

Bertsekas, Dimitri P. and Tsitsiklis, John N. *Neuro-Dynamic Programming*. Athena Scientific, 1996.

Boyan, Justin A. Technical update: Least-squares temporal difference learning. *Machine Learning*, 49(2–3):233–246, 2002. ISSN 0885-6125.

Bradley, Richard. Basic properties of strong mixing conditions. a survey and some open questions. *Probability Survey*, 2:107–144, 2005.

Downey, Carlton and Sanner, Scott. Temporal difference bayesian model averaging: A bayesian perspective on adapting lambda. In Fürnkranz, Johannes and Joachims, Thorsten (eds.), *ICML*, pp. 311–318. Omnipress, 2010.

Györfi, László, Kholer, Michael, Krzyzak, Adam, and Walk, Harro. *A Distribution-Free Theory of Nonparametric Regression*. Springer-Verlag, New York, 2002.

Hayes, Thomas P. A large-deviation inequality for vector-valued martingales, 2005. Technical report.

Kearns, M.J. and Singh, S.P. Bias-variance error bounds for temporal difference updates. In Cesa-Bianchi, Nicolò and Goldman, Sally A. (eds.), *COLT*, pp. 142–147. Morgan Kaufmann, 2000.

Lazaric, Alessandro, Ghavamzadeh, Mohammad, and Munos, Rémi. Finite-sample analysis of least-squares policy iteration. *Journal of Machine Learning Research*, 13:3041–3074, October 2012.

Nedic, Angelia and Bertsekas, Dimitri P. Least squares policy evaluation algorithms with linear function approximation. *Theory and Applications*, 13:79–110, 2002.

Pires, Bernardo A. and Szepesvári, Csaba. Statistical linear estimation with penalized estimators: an application to reinforcement learning. In *ICML*, pp. 1535–1542, 2012.

Scherrer, Bruno. Should one compute the temporal difference fix point or minimize the Bellman residual? the unified oblique projection view. In *ICML*, pp. 959–966, 2010.

Scherrer, Bruno and Lesner, Boris. On the use of non-stationary policies for stationary infinite-horizon Markov decision processes. In *NIPS 2012 Adv.in Neural Information Processing*, December 2012.

Sutton, Richard S. and Barto, Andrew G. Reinforcement learning i: Introduction, 1998.

Szepesvári, Csaba. *Algorithms for Reinforcement Learning*. Morgan and Claypool, 2010.

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

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

Yu, Huizhen. Convergence of least-squares temporal difference methods under general conditions. In *ICML*, pp. 1207–1214, 2010.

Supplementary Material

A. Proof of Lemma 2

We begin by bounding, for any value of m, the distance between $\hat{G}$ and $\hat{G}^m$. Set m to any integer greater or equal to 1. Writing

$$\epsilon_1 = \frac{1}{n-1}\sum_{i=1}^{m-1} G_i - \mathbb{E}[G_i]$$
$$\text{and} \quad \epsilon_2 = \frac{1}{n-1}\sum_{i=m}^{n-1}(z_i - z_i^m)\tau(X_i, X_{i+1})^T - \mathbb{E}[(z_i - z_i^m)\tau(X_i, X_{i+1})^T],$$

we have

$$\begin{aligned}
\frac{1}{n-1}\sum_{i=1}^{n-1} G_i - \mathbb{E}[G_i] &= \frac{1}{n-1}\sum_{i=m}^{n-1} G_i - \mathbb{E}[G_i] + \epsilon_1 \\
&= \frac{1}{n-1}\sum_{i=m}^{n-1} z_i\tau(X_i, X_{i+1})^T - \mathbb{E}[z_i\tau(X_i, X_{i+1})^T] + \epsilon_1 \\
&= \frac{1}{n-1}\sum_{i=m}^{n-1} z_i^m\tau(X_i, X_{i+1})^T - \mathbb{E}[z_i^m\tau(X_i, X_{i+1})^T] + (\epsilon_1 + \epsilon_2) \\
&= \frac{1}{n-1}\sum_{i=m}^{n-1}(G_i^m - \mathbb{E}[G_i^m]) + (\epsilon_1 + \epsilon_2). \qquad (23)
\end{aligned}$$

For all i, we have $\|z_i\|_\infty \leq \frac{L}{1-\lambda\gamma}$, $\|G_i\|_\infty \leq \frac{LL'}{1-\lambda\gamma}$, and $\|z_i - z_i^m\|_\infty \leq \frac{(\lambda\gamma)^m L}{1-\lambda\gamma}$. As a consequence—using $\|M\|_2 \leq \|M\|_F = \sqrt{d\times k}\|x\|_\infty$ for $M \in \mathbb{R}^{d\times k}$ with x the vector obtained by concatenating all M columns—, we can see that

$$\|\epsilon_1 + \epsilon_2\|_2 \leq \frac{2(m-1)\sqrt{d\times k}LL'}{(n-1)(1-\lambda\gamma)} + \frac{2(\lambda\gamma)^m\sqrt{d\times k}LL'}{(1-\lambda\gamma)} \qquad (24)$$

By concatenating all its columns, the $d \times k$ matrix G_i^m may be seen a single vector U_i^m of size dk. Then, for all $\epsilon > 0$,

$$\begin{aligned}
\mathbb{P}\left(\left\|\frac{1}{n-m}\sum_{i=m}^{n-1}(G_i^m - \mathbb{E}[G_i^m])\right\|_2 \geq \epsilon\right) &\leq \mathbb{P}\left(\left\|\frac{1}{n-m}\sum_{i=m}^{n-1}(G_i^m - \mathbb{E}[G_i^m])\right\|_F \geq \epsilon\right) \\
&= \mathbb{P}\left(\left\|\frac{1}{n-m}\sum_{i=m}^{n-1}(U_i^m - \mathbb{E}[U_i^m])\right\|_2 \geq \epsilon\right). \qquad (25)
\end{aligned}$$

The process $(U_n^m)_{n\geq m}$, defined as a function of the process $(Z_n)_{n\geq m} = (X_{n-m+1}, X_{n-m+2}, \ldots, X_{n+1})_{n\geq m}$, is stationary. By using the next lemma, we can see that it inherits in some sense the β-mixing property of the process $(X_i)_{i\geq 1}$ (Assumption 2).

Lemma 5 (originally stated as Lemma 3)**.** *Let $(X_n)_{n\geq 1}$ be a β-mixing process, then $(Z_n)_{n\geq m} = (X_{n-m+1}, X_{n-m+2}, \ldots, X_{n+1})_{n\geq m}$ is a β-mixing process such that its i^{th} β mixing coefficient β_i^Z satisfies $\beta_i^Z \leq \beta_{i-m}^X$.*

Proof. Let $\Gamma = \sigma(Z_m, ..., Z_t)$, by definition we have

$$\Gamma = \sigma(Z_j^{-1}(B) : j \in \{m, ..., t\}, B \in \sigma(\mathcal{X}^{m+1})).$$

For all $j \in \{m, ..., t\}$ we have

$$Z_j^{-1}(B) = \{\omega \in \Omega, Z_j(\omega) \in B\}.$$

For $B = B_0 \times ... \times B_m$, we observe that

$$Z_j^{-1}(B) = \{\omega \in \Omega, X_{j-m+1}(\omega) \in B_0, ..., X_{j+1}(\omega) \in B_m\}.$$

Then we have

$$\Gamma = \sigma(X_j^{-1}(B) : j \in \{m, ..., t\}, B \in \sigma(\mathcal{X})) = \sigma(X_1, ..., X_{t+1}).$$

Similarly we can prove that $\sigma(Z_{t+i}^{\infty}) = \sigma(X_{t+i-m+1}^{\infty})$. Then let β_i^X be the i^{th} β-mixing coefficient of the process $(X_n)_{n\geq 1}$, we have

$$\beta_i^X = \sup_{t\geq 1} \mathbb{E}\left[\sup_{B\in\sigma(X_{t+i}^{\infty})} |P(B|\sigma(X_1, ..., X_t)) - P(B)|\right].$$

Similarly for the process $(Z_n)_{n\geq m}$ we can see that

$$\beta_i^Z = \sup_{t\geq m} \mathbb{E}\left[\sup_{B\in\sigma(Z_{t+i}^{\infty})} |P(B|\sigma(Z_m, ..., Z_t)) - P(B)|\right].$$

By applying what we developed above we obtain

$$\beta_i^Z = \sup_{t\geq m} \mathbb{E}\left[\sup_{B\in\sigma(X_{t+i-m+1}^{\infty})} |P(B|\sigma(X_1, ..., X_{t+1})) - P(B)|\right].$$

Denote $u = t + 1$ we have

$$\beta_i^Z = \sup_{u\geq m+1} \mathbb{E}\left[\sup_{B\in\sigma(X_{u+i-m}^{\infty})} |P(B|\sigma(X_1, ..., X_u)) - P(B)|\right]$$

Then for $i > m$

$$\beta_i^Z \leq \beta_{i-m}^X.$$

□

Now that we know that $(U_n^m)_{n\geq m}$ is a β-mixing stationary process, we shall use the decomposition technique proposed by Yu (1994) that consists in dividing the sequence $U_m^m, \ldots, U_{n-1}^m$ into $2\mu_{n-m}$ blocks of length a_{n-m} (we assume here that $n - m = 2a_{n-m}\mu_{n-m}$). The blocks are of two kinds: those which contains the even indexes $E = \cup_{l=1}^{\mu_{n-m}} E_l$ and those with odd indexes $H = \cup_{l=1}^{\mu_{n-m}} H_l$. Thus, by grouping the variables into blocks we get

$$\begin{aligned}
&\mathbb{P}\left(\left\|\frac{1}{n-m}\sum_{i=m}^{n-1} U_i^m - \mathbb{E}[U_i^m]\right\|_2 \geq \epsilon\right) \\
\leq &\mathbb{P}\left(\left\|\sum_{i\in H} U_i^m - \mathbb{E}[U_i^m]\right\|_2 + \left\|\sum_{i\in E} U_i^m - \mathbb{E}[U_i^m]\right\|_2 \geq (n-m)\frac{\epsilon}{2}\right) &(26)\\
\leq &\mathbb{P}\left(\left\|\sum_{i\in H} U_i^m - \mathbb{E}[U_i^m]\right\|_2 \geq \frac{(n-m)\epsilon}{4}\right) + \mathbb{P}\left(\left\|\sum_{i\in E} U_i^m - \mathbb{E}[U_i^m]\right\|_2 \geq \frac{(n-m)\epsilon}{4}\right) &(27)\\
= &2\mathbb{P}\left(\left\|\sum_{i\in H} U_i^m - \mathbb{E}[U_i^m]\right\|_2 \geq \frac{(n-m)\epsilon}{4}\right) &(28)
\end{aligned}$$

where Equation (26) follows from the triangle inequality, Equation (27) from the fact that the event $\{X+Y \geq a\}$ implies $\{X \geq \frac{a}{2}\}$ or $\{Y \geq \frac{a}{2}\}$, and Equation (28) from the assumption that the process is stationary. Since $H = \cup_{l=1}^{\mu_{n-m}} H_l$ we have

$$\begin{aligned}\mathbb{P}\left(\left\|\frac{1}{n-m}\sum_{i=m}^{n-1} U_i^m - \mathbb{E}[U_i^m]\right\|_2 \geq \epsilon\right) &\leq 2\mathbb{P}\left(\left\|\sum_{l=1}^{\mu_{n-m}}\sum_{i\in H_l} U_i^m - \mathbb{E}[U_i^m]\right\|_2 \geq \frac{(n-m)\epsilon}{4}\right) \\ &= 2\mathbb{P}\left(\left\|\sum_{l=1}^{\mu_{n-m}} U(H_l) - \mathbb{E}[U(H_l)]\right\|_2 \geq \frac{(n-m)\epsilon}{4}\right) \end{aligned} \tag{29}$$

where we defined $U(H_l) = \sum_{i\in H_l} U_i^m$. Now consider the sequence of identically distributed independent blocks $(U'(H_l))_{l=1,\ldots,\mu_{n-m}}$ such that each block $U'(H_l)$ has the same distribution as $U(H_l)$. We are going to use the following technical result.

Lemma 6. *(Yu, 1994) Let $X_1,\ldots,X_n$ be a sequence of samples drawn from a stationary β-mixing process with coefficients $\{\beta_i\}$. Let $X(H) = (X(H_1),\ldots,X(H_{\mu_{n-m}}))$ where for all j $X(H_j) = (X_i)_{i\in H_j}$. Let $X'(H) = (X'(H_1),\ldots,X'(H_{\mu_{n-m}}))$ with $X'(H_j)$ independent and such that for all j, $X'(H_j)$ has same distribution as $X(H_j)$. Let Q and Q' be the distribution of $X(H)$ and $X'(H)$ respectively. For any measurable function $h : \mathcal{X}^{a_n\mu_n} \to \mathbb{R}$ bounded by B, we have*

$$|\mathbb{E}_Q[h(X(H)] - \mathbb{E}_{Q'}[h(X'(H)]| \leq B\mu_n\beta_{a_n}.$$

By applying Lemma 6, Equation (29) leads to:

$$\begin{aligned}\mathbb{P}\left(\left\|\frac{1}{n-m}\sum_{i=m}^{n-1} U_i^m - \mathbb{E}[U_i^m]\right\|_2 \geq \epsilon\right) \leq& 2\mathbb{P}\left(\left\|\sum_{l=1}^{\mu_{n-m}} U'(H_l) - \mathbb{E}[U'(H_l)]\right\|_2 \geq \frac{(n-m)\epsilon}{4}\right) \\ &+ 2\mu_{n-m}\beta_{a_{n-m}}. \end{aligned} \tag{30}$$

The variables $U'(H_l)$ are independent. Furthermore, it can be seen that $(\sum_{l=1}^{\mu_{n-m}} U'(H_l) - \mathbb{E}[U'(H_l)])_{\mu_{n-m}}$ is a $\sigma(U'(H_1),\ldots,U'(H_{\mu_{n-m}}))$ martingale:

$$\begin{aligned}&\mathbb{E}\left[\sum_{l=1}^{\mu_{n-m}} U'(H_l) - \mathbb{E}[U'(H_l)] \,\middle|\, U'(H_1),\ldots,U'(H_{\mu_{n-m}-1})\right] \\ =& \sum_{l=1}^{\mu_{n-m}-1} U'(H_l) - \mathbb{E}[U'(H_l)] + \mathbb{E}[U'_{H_{\mu_{n-m}}} - \mathbb{E}[U'_{H_{\mu_{n-m}}}]] \\ =& \sum_{l=1}^{\mu_{n-m}-1} U'(H_l) - \mathbb{E}[U'(H_l)].\end{aligned}$$

We can now use the following concentration result for martingales.

Lemma 7 ((Hayes, 2005))**.** *Let $X = (X_0,\ldots,X_n)$ be a discrete time martingale taking values in an Euclidean space such that $X_0 = 0$ and for all i, $\|X_i - X_{i-1}\|_2 \leq B_2$ almost surely. Then for all ϵ,*

$$P\{\|X_n\|_2 \geq \epsilon\} < 2e^2 e^{-\frac{\epsilon^2}{2n(B_2)^2}}.$$

Indeed, taking $X_{\mu_{n-m}} = \sum_{l=1}^{\mu_{n-m}} U'(H_l) - \mathbb{E}[U'(H_l)]$, and observing that $\|X_i - X_{i-1}\| = \|U'(H_l) - \mathbb{E}[U'(H_l)]\|_2 \leq a_{n-m}C$ with $C = \frac{2\sqrt{dk}LL'}{1-\lambda\gamma}$, the lemma leads to

$$\begin{aligned}\mathbb{P}\left(\left\|\sum_{l=1}^{\mu_{n-m}} U'(H_l) - \mathbb{E}[U'(H_l)]\right\|_2 \geq \frac{(n-m)\epsilon}{4}\right) &\leq 2e^2 e^{-\frac{(n-m)^2\epsilon^2}{32\mu_{n-m}(a_{n-m}C)^2}} \\ &= 2e^2 e^{-\frac{(n-m)\epsilon^2}{16a_{n-m}C^2}}.\end{aligned}$$

where the second line is obtained by using the fact that $2a_{n-m}\mu_{n-m} = n - m$. With Equations (29) and (30), we finally obtain

$$\mathbb{P}\left(\left\|\frac{1}{n-m}\sum_{i=m}^{n-1} U_i^m - \mathbb{E}[U_i^m]\right\|_2 \geq \epsilon\right) \leq 4e^2 e^{-\frac{(n-m)\epsilon^2}{16 a_{n-m} C^2}} + 2(n-m)\beta^U_{a_{n-m}}.$$

The vector U_i^m is a function of $Z_i = (X_{i-m+1}, \ldots, X_{i+1})$, and Lemma 3 tells us that for all $j > m$,

$$\beta_j^U \leq \beta_j^Z \leq \beta^X_{j-m} \leq \overline{\beta} e^{-b(j-m)^\kappa}.$$

So the equation above may be re-written as

$$\mathbb{P}\left(\left\|\frac{1}{n-m}\sum_{i=m}^{n-1} U_i^m - \mathbb{E}[U_i^m]\right\|_2 \geq \epsilon\right) \leq 4e^2 e^{-\frac{(n-m)\epsilon^2}{16 a_{n-m} C^2}} + 2(n-m)\overline{\beta} e^{-b(a_{n-m}-m)^\kappa} = \delta'. \tag{31}$$

We now follow a reasoning similar to that of (Lazaric et al., 2012) in order to get the same exponent in both of the above exponentials. Taking $a_{n-m} - m = \left\lceil \frac{C_2(n-m)\epsilon^2}{b} \right\rceil^{\frac{1}{\kappa+1}}$ with $C_2 = (16C^2\zeta)^{-1}$, and $\zeta = \frac{a_{n-m}}{a_{n-m}-m}$, we have

$$\delta' \leq (4e^2 + (n-m)\overline{\beta}) \exp\left(-\min\left\{\left(\frac{b}{(n-m)\epsilon^2 C_2}\right), 1\right\}^{\frac{1}{k+1}} \frac{1}{2}(n-m)C_2\epsilon^2\right). \tag{32}$$

Define

$$\Lambda(n,\delta) = \log\left(\frac{2}{\delta}\right) + \log(\max\{4e^2, n\overline{\beta}\}),$$

and

$$\epsilon(\delta) = \sqrt{2\frac{\Lambda(n-m,\delta)}{C_2(n-m)} \max\left\{\frac{\Lambda(n-m,\delta)}{b}, 1\right\}^{\frac{1}{\kappa}}}.$$

It can be shown that

$$\exp\left(-\min\left\{\left(\frac{b}{(n-m)(\epsilon(\delta))^2 C_2}\right), 1\right\}^{\frac{1}{k+1}} \frac{1}{2}(n-m)C_2(\epsilon(\delta))^2\right) \leq \exp\left(-\Lambda(n-m,\delta)\right). \tag{33}$$

Indeed[6], there are two cases:

1. Suppose that $\min\left\{\left(\frac{b}{(n-m)(\epsilon(\delta))^2 C_2}\right), 1\right\} = 1$. Then

$$\begin{aligned}
&\exp\left(-\min\left\{\left(\frac{b}{(n-m)(\epsilon(\delta))^2 C_2}\right), 1\right\}^{\frac{1}{k+1}} \frac{1}{2}(n-m)C_2(\epsilon(\delta))^2\right) \\
&= \exp\left(-\Lambda(n-m,\delta)\max\left\{\frac{\Lambda(n-m,\delta)}{b}, 1\right\}^{\frac{1}{k}}\right) \\
&\leq \exp\left(-\Lambda(n-m,\delta)\right).
\end{aligned}$$

[6]This inequality exists in (Lazaric et al., 2012), and is developped here for completeness.

2. Suppose now that $\min\left\{\left(\frac{b}{(n-m)(\epsilon(\delta))^2C_2}\right),1\right\}=\left(\frac{b}{(n-m)(\epsilon(\delta))^2C_2}\right)$. Then

$$\begin{aligned}
&\exp\left(-\min\left\{\left(\frac{b}{(n-m)(\epsilon(\delta))^2C_2}\right),1\right\}^{\frac{1}{k+1}}\frac{1}{2}(n-m)C_2(\epsilon(\delta))^2\right)\\
=&\exp\left(-\frac{1}{2}b^{\frac{1}{k+1}}((n-m)C_2(\epsilon(\delta))^2)^{\frac{k}{k+1}}\right)\\
=&\exp\left(-\frac{1}{2}b^{\frac{1}{k+1}}(\Lambda(n-m,\delta)^{\frac{k}{k+1}}\max\left\{\frac{\Lambda(n-m,\delta)}{b},1\right\}^{\frac{1}{k+1}}\right)\\
=&\exp\left(-\frac{1}{2}\Lambda(n-m,\delta)^{\frac{k}{k+1}}\max\left\{\Lambda(n-m,\delta),b\right\}^{\frac{1}{k+1}}\right)\\
\leq&\exp\left(-\Lambda(n-m,\delta)\right).
\end{aligned}$$

By combining Equations (32) and (33), we get

$$\delta'\leq(4e^2+(n-m)\overline{\beta})\exp\left(-\Lambda(n-m,\delta)\right).$$

If we replace $\Lambda(n-m,\delta)$ with its expression, we obtain

$$\exp\left(-\Lambda(n-m,\delta)\right)=\frac{\delta}{2}\max\{4e^2,(n-m)\overline{\beta}\}^{-1}.$$

Since $4e^2\max\{4e^2,(n-m)\overline{\beta}\}^{-1}\leq 1$ and $(n-m)\overline{\beta}\max\{4e^2,(n-m)\overline{\beta}\}^{-1}\leq 1$, we consequently have

$$\delta'\leq 2\frac{\delta}{2}\leq\delta.$$

Now, note that since $a_{n-m}-m\geq 1$, we have

$$\zeta=\frac{a_{n-m}}{a_{n-m}-m}=\frac{a_{n-m}-m+m}{a_{n-m}-m}\leq 1+m.$$

Let $J(n,\delta)=32\Lambda(n,\delta)\max\left\{\frac{\Lambda(n,\delta)}{b},1\right\}^{\frac{1}{\kappa}}$. Then Equation (31) is reduced to

$$\mathbb{P}\left(\left\|\frac{1}{n-m}\sum_{i=m}^{n-1}(U_i^m-\mathbb{E}[U_i^m])\right\|_2\geq\frac{C}{\sqrt{n-m}}\left(\zeta J(n-m,\delta)\right)^{\frac{1}{2}}\right)\leq\delta. \tag{34}$$

Since $J(n,\delta)$ is an increasing function on n, and $\frac{n-1}{\sqrt{n-1}(n-m)}=\frac{1}{\sqrt{n-m}}\sqrt{\frac{n-1}{n-m}}\geq\frac{1}{\sqrt{n-m}}$, we have

$$\begin{aligned}
&\mathbb{P}\left(\left\|\frac{1}{n-1}\sum_{i=m}^{n-1}(G_i^m-\mathbb{E}[G_i^m])\right\|_2\geq\frac{C}{\sqrt{n-1}}\left(\zeta J(n-1,\delta)\right)^{\frac{1}{2}}\right)\\
\leq\ &\mathbb{P}\left(\left\|\frac{1}{n-m}\sum_{i=m}^{n-1}(G_i^m-\mathbb{E}[G_i^m])\right\|_2\geq\frac{C}{\sqrt{n-1}}\frac{n-1}{n-m}\left((m+1)J(n-1,\delta)\right)^{\frac{1}{2}}\right)\\
\leq\ &\mathbb{P}\left(\left\|\frac{1}{n-m}\sum_{i=m}^{n-1}(G_i^m-\mathbb{E}[G_i^m])\right\|_2\geq\frac{C}{\sqrt{n-m}}\left((m+1)J(n-m,\delta)\right)^{\frac{1}{2}}\right).
\end{aligned}$$

By using Equations (25) and (34), we deduce that

$$\mathbb{P}\left(\left\|\frac{1}{n-1}\sum_{i=m}^{n-1}(G_i^m-\mathbb{E}[G_i^m])\right\|_2\geq\frac{C}{\sqrt{n-1}}\left((m+1)J(n-1,\delta)\right)^{\frac{1}{2}}\right)\leq\delta. \tag{35}$$

By combining Equations (23), (24) and (35), plugging the value of $C=\frac{2\sqrt{dk}LL'}{1-\lambda\gamma}$, and taking $m=\left\lceil\frac{\log(n-1)}{\log\frac{1}{\lambda\gamma}}\right\rceil$—so that $\|\epsilon_1+\epsilon_2\|_2\leq\epsilon(n)$—, we get the announced result.

B. Proof of Theorem 3

We prove here the following result: for any $\delta \in (0,1)$, for all $n \geq 1$, consider $\hat{v}^{\rho}_{LSTD(\lambda)} = \Phi\hat{\theta}_{\rho}$ with penalization parameter $\rho = 2\Xi^2(n,\delta)$. Then, with at least probability $1-\delta$, for all n,

$$\|\hat{v}^{\rho}_{LSTD(\lambda)} - v_{LSTD(\lambda)}\|_{\mu} \leq \frac{4V_{\max}\sqrt{d}L(3+\sqrt{d}L)}{\sqrt{n-1}(1-\gamma)\sqrt{\nu}}\sqrt{(m_n^{\lambda}+1)\,I(n-1,\delta)} + g(n,\delta),$$

where $g(n,\delta)$ and $I(n,\delta)$ are defined as in Theorem 1.

Proof. Let $\hat{\theta}_{\rho}$ be the vector that satisfies

$$\hat{\theta}_{\rho} = \arg\min_{\theta\in\mathbb{R}^d}\left\{\|\hat{A}\theta_{\rho} - \hat{b}\|_2^2 + \rho\|\theta_{\rho}\|_2^2\right\}. \tag{36}$$

We have

$$\|A\hat{\theta}_{\rho} - b\|_2 \leq \|\epsilon_A\|_2\|\hat{\theta}_{\rho}\|_2 + \|\epsilon_b\|_2 + \|\hat{A}\hat{\theta}_{\rho} - \hat{b}\|_2.$$

Then by using the inequality $(a+b)^2 \leq 2(a^2+b^2)$ twice on $\underbrace{\|\epsilon_A\|_2\|\hat{\theta}\|_2 + \|\epsilon_b\|_2}_{a} + \underbrace{\|\hat{A}\hat{\theta} - \hat{b}\|_2}_{b}$ and then on $\underbrace{\|\epsilon_A\|_2\|\hat{\theta}\|_2}_{a} + \underbrace{\|\epsilon_b\|_2}_{b}$ we have

$$\|A\hat{\theta}_{\rho} - b\|_2^2 \leq 4\|\epsilon_A\|_2^2\|\hat{\theta}_{\rho}\|_2^2 + 4\|\epsilon_b\|_2^2 + 2\|\hat{A}\hat{\theta}_{\rho} - \hat{b}\|_2^2.$$

From Equation (36) we can write that

$$\left\{\|\hat{A}\hat{\theta}_{\rho} - \hat{b}\|_2^2 + \rho\|\hat{\theta}_{\rho}\|_2^2\right\} = \min_{\theta\in\mathbb{R}^d}\left\{\|\hat{A}\theta_{\rho} - \hat{b}\|_2^2 + \rho\|\theta\|_2^2\right\}$$

$$\|\hat{\theta}_{\rho}\|_2^2 = \frac{1}{\rho}\min_{\theta\in\mathbb{R}^d}\left\{\|\hat{A}\theta - \hat{b}\|_2^2 + \rho\|\theta\|_2^2 - \|\hat{A}\hat{\theta}_{\rho} - \hat{b}\|_2^2\right\},$$

and

$$\begin{aligned}\|\hat{A}\hat{\theta}_{\rho} - \hat{b}\|_2^2 &= \min_{\theta\in\mathbb{R}^d}\left\{\|\hat{A}\theta - \hat{b}\|_2^2 + \rho(\|\theta\|_2^2 - \|\hat{\theta}_{\rho}\|_2^2)\right\}\\ &\leq \min_{\theta\in\mathbb{R}^d}\left\{\|\hat{A}\theta - \hat{b}\|_2^2 + \rho\|\theta\|_2^2\right\}.\end{aligned}$$

So that

$$\begin{aligned}\|A\hat{\theta}_{\rho} - b\|_2^2 &\leq 4\frac{\|\epsilon_A\|_2^2}{\rho}\min_{\theta\in\mathbb{R}^d}\left\{\|\hat{A}\theta - \hat{b}\|_2^2 + \rho\|\theta\|_2^2 - \|\hat{A}\hat{\theta} - \hat{b}\|_2^2\right\} + 4\|\epsilon_b\|_2^2 + 2\|\hat{A}\hat{\theta} - \hat{b}\|_M^2\\ &\leq 4\frac{\|\epsilon_A\|_2^2}{\rho}\min_{\theta\in\mathbb{R}^d}\left\{\|\hat{A}\theta - \hat{b}\|_2^2 + \rho\|\theta\|_2^2\right\} + \max\left(0, 2 - 4\frac{\|\epsilon_A\|_2^2}{\rho}\right)\|\hat{A}\hat{\theta} - \hat{b}\|_2^2 + 4\|\epsilon_b\|_2^2\\ &\leq \max\left(4\frac{\|\epsilon_A\|_2^2}{\rho}, 2\right)\min_{\theta\in\mathbb{R}^d}\left\{\|\hat{A}\theta - \hat{b}\|_2^2 + \rho\|\theta\|_2^2\right\} + 4\|\epsilon_b\|_2^2.\end{aligned}$$

In Section 4.3, we derived high-probability bounds on $\|\epsilon_A\|_2$ and $\|\hat{A}\theta^* - \hat{b}\|_2 = \|\epsilon_A\theta^* - \epsilon_b\|_2$ with $\theta^* = A^{-1}b$. It is easy to also derive a high-probability bound on $\|\epsilon_b\|_2^2$. More precisely, with the definitions of ϵ_1 and ϵ_2 given in Equations (17) and (21), and with $\epsilon_3(n,\delta_n) = \frac{2\sqrt{d}L^2}{(1-\lambda\gamma)\sqrt{n-1}}\sqrt{(m_n^{\lambda}+1)\,J(n-1,\delta_n)} + \tilde{O}(\frac{1}{n})$, we know that with probability at least $1-\delta$,

$$\|\epsilon_A\|_2 \leq \epsilon_1(n,\delta_n), \quad \|\epsilon_A\theta^* - \epsilon_b\|_2 \leq \epsilon_2(n,\delta_n) \quad \text{and} \quad \|\epsilon_b\|_2 \leq \epsilon_3(n,\delta_n).$$

As a consequence,

$$\|A\hat{\theta}_{\rho} - b\|_2^2 \leq \max\left(4\frac{\|\epsilon_A\|_2^2}{\rho}, 2\right)\left\{(\epsilon_2(n,\delta_n)^2 + \rho)\|\theta^*\|_2^2\right\} + 4\|\epsilon_b\|_2^2.$$

With $\rho = 2(\epsilon_1(n, \delta_n))^2$, we obtain with probability at $1 - \delta$,

$$\|A\hat{\theta}_\rho - b\|_2^2 \leq 2(2(\epsilon_1(n, \delta_n))^2 + (\epsilon_2(n, \delta_n))^2)\|\theta^*\|_2^2 + 4(\epsilon_3(n, \delta_n))^2$$

By using the fact that $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$, this implies

$$\|A\hat{\theta}_\rho - b\|_2 \leq \sqrt{2(2\epsilon_1(n, \delta_n) + \epsilon_2(n, \delta_n))}\|\theta^*\|_2 + 2(\epsilon_3(n, \delta_n))$$

We conclude by using Equation (8) in which we take the norm, by bounding $\|\Phi A^{-1}\|_\mu$ in the same way as we did in the proof of Lemma 1, and finish in the way similar to the unregularized proof with $\delta_n = \frac{\delta}{6n^2}$. □