HAL
open science

Rate of Convergence and Error Bounds for LSTD(λ)

Manel Tagorti, Bruno Scherrer

To cite this version:

Manel Tagorti, Bruno Scherrer. Rate of Convergence and Error Bounds for LSTD(λ). [Research Report] 2014. hal-00990525

HAL Id: hal-00990525

https://inria.hal.science/hal-00990525v1

Submitted on 13 May 2014

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Rate of Convergence and Error Bounds for LSTD(λ)

Manel Tagorti and Bruno Scherrer
INRIA Nancy Grand Est, Team MAIA
manel.tagorti@inria.fr, bruno.scherrer@inria.fr

May 13, 2014

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 estimate of 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. (2010) in the specific case where $\lambda = 0$. In particular, our analysis sheds some light on the choice of λ with respect to the quality of the chosen linear space and the number of samples, that complies with simulations.

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 estimating 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 and Tsitsiklis, 1996; Szepesvári, 2010).

The asymptotic almost sure convergence of LSTD(λ) was proved by Nedic and 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 $\lambda = 0$. 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 and Roy, 1997) (restated in Theorem 2, Section 3).

The paper is organized as follows. Section 2 starts by describing the LSTD(λ) algorithm and the necessary background. Section 3 then contains our main result (Theorem 1): for all $\lambda \in (0,1)$, we will show that LSTD(λ) converges to its limit at a rate $\tilde{O}(\frac{1}{\sqrt{n}})$. We shall then deduce a global error (Corollary 1) that sheds some light on the role of the parameter λ and discuss some of its interesting practical consequences. Section 4 will go on by providing a detailed proof of our claims. Finally, Section 5 concludes and describes potential future work.

2 LSTD(λ) and Related background

We consider a Markov chain $\mathcal{M}$ taking its values on a finite or countable state space[2] $\mathcal{X}$, with transition kernel P. We assume $\mathcal{M}$ ergodic[3]; consequently, it admits a unique stationary distribution μ. For any

[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 restrict our focus to finite/countable mainly because it eases the presentation of our analysis. Though this requires some extra work, we believe the analysis we make here can be extended to more general state spaces.

[3] In our countable state space situation, ergodicity holds if and only if the chain is aperiodic and irreducible, that is formally if and only if: $\forall (x,y) \in \mathcal{X}^2,\ \exists n_0,\ \forall n \geq n_0,\ P^n(x,y) > 0$.

$K \in \mathbb{R}$, we denote $\mathcal{B}(\mathcal{X}, K)$ the set of measurable 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 $\mathcal{M}$ is defined, for any state i, as the average discounted sum of rewards along infinitely long trajectories starting from i:

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

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 $d \ll |\mathcal{X}|$, we consider a feature matrix Φ 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, and without loss of generality[4] we will make the following assumption.

Assumption 1. *The feature functions $(\phi_j)_{j\in\{1,...,d\}}$ 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.

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 weighted arithmetic mean 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 and Roy, 1997; Nedic and Bertsekas, 2002), it can be seen that the operator T^λ is a contraction mapping of modulus $\frac{(1-\lambda)\gamma}{1-\lambda\gamma} \le \gamma$; indeed, for any vectors u, v:

$$\begin{aligned}
\|T^\lambda u - T^\lambda v\|_\mu &\le (1-\lambda)\|\sum_{i=0}^{\infty} \lambda^i (T^{i+1}u - T^{i+1}v)\|_\mu \\
&= (1-\lambda)\|\sum_{i=0}^{\infty} \lambda^i (\gamma^{i+1}P^{i+1}u - \gamma^{i+1}P^{i+1}v)\|_\mu \\
&\le (1-\lambda)\sum_{i=0}^{\infty} \lambda^i \gamma^{i+1}\|u - v\|_\mu \\
&= \frac{(1-\lambda)\gamma}{1-\lambda\gamma}\|u - v\|_\mu.
\end{aligned}$$

[4]This assumption is not fundamental: in theory, we can remove any set of features that makes the family linearly dependent; in practice, the algorithm we are going to describe can use the pseudo-inverse instead of the inverse.

Since the orthogonal projector Π is non-expansive with respect to μ (Tsitsiklis and 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 and 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 $A\theta = b$ (Nedic and Bertsekas, 2002), such that for any i,

$$A = \Phi^T D_\mu (I - \gamma P)(I - \lambda\gamma P)^{-1}\Phi = \mathbb{E}_{X_{-\infty}\sim\mu}\left[\sum_{k=-\infty}^{i} (\gamma\lambda)^{i-k}\phi(X_k)(\phi(X_i) - \gamma\phi(X_{i+1}))^T\right] \quad (3)$$

$$\text{and} \quad b = \Phi^T D_\mu (I - \gamma\lambda P)^{-1} r = \mathbb{E}_{X_{-\infty}\sim\mu}\left[\sum_{k=-\infty}^{i} (\gamma\lambda)^{i-k}\phi(X_k) r(X_i)\right], \quad (4)$$

where u^T is the transpose of u. 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 and 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 (3)-(4) suggest to compute the following estimates:

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

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

$$\text{where} \quad z_i = \sum_{k=1}^{i} (\lambda\gamma)^{i-k}\phi(X_k) \quad (5)$$

is the so-called *eligibility trace*. The algorithm then returns $\hat{v}_{LSTD(\lambda)} = \Phi\hat{\theta}$ with[5] $\hat{\theta} = \hat{A}^{-1}\hat{b}$, which is a (finite sample) approximation of $v_{LSTD(\lambda)}$. Using a variation of the law of large numbers, Nedic and 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 the remaining of the paper is to deepen this analysis: we shall estimate the rate of convergence of $\hat{v}_{LSTD(\lambda)}$ to $v_{LSTD(\lambda)}$, and bound the approximation error $\|\hat{v}_{LSTD(\lambda)} - v\|_\mu$ of the overall algorithm.

3 Main results

This section contains our main results. Our key assumption for the analysis is that the Markov chain process that generates the states has some mixing property[6].

Assumption 2. *The process $(X_n)_{n\geq 1}$ is β-mixing, in the sense that its ith coefficient*

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

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 $\beta_i \leq \overline{\beta} e^{-bi^\kappa}$.

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

[6] A stationary ergodic Markov chain is *always* β-mixing.

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

Theorem 1. *Let Assumptions 1 and 2 hold and let $X_1 \sim \mu$. For any $n \geq 1$ and $\delta \in (0,1)$, define:*

$$I(n,\delta) = 32\Lambda(n,\delta)\max\left\{\frac{\Lambda(n,\delta)}{b}, 1\right\}^{\frac{1}{\kappa}}$$

$$\textit{where}\quad \Lambda(n,\delta) = \log\left(\frac{8n^2}{\delta}\right) + \log(\max\{4e^2, n\overline{\beta}\}).$$

Let $n_0(\delta)$ be the smallest integer such that

$$\forall n \geq n_0(\delta),\ \frac{2dL^2}{(1-\gamma)\nu}\left[\frac{2}{\sqrt{n-1}}\sqrt{\left(\left\lceil\frac{\log(n-1)}{\log\left(\frac{1}{\lambda\gamma}\right)}\right\rceil + 1\right)I(n-1,\delta)} + \frac{1}{(n-1)(1-\lambda\gamma)} + \frac{2}{(n-1)}\left\lceil\frac{\log(n-1)}{\log\left(\frac{1}{\lambda\gamma}\right)}\right\rceil\right] < 1 \tag{6}$$

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 we have:

$$\|v_{LSTD(\lambda)} - \hat{v}_{LSTD(\lambda)}\|_\mu \leq \frac{4V_{max}dL^2}{\sqrt{n-1}(1-\gamma)\nu}\sqrt{\left(1 + \left\lceil\frac{\log(n-1)}{\log\left(\frac{1}{\lambda\gamma}\right)}\right\rceil\right)I(n-1,\delta)} + h(n,\delta)$$

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

The constant ν is strictly 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 (6) tends to 0 when n tends to infinity. As $\left(1 + \left\lceil\frac{\log(n-1)}{\log\left(\frac{1}{\lambda\gamma}\right)}\right\rceil\right) I(n-1, \frac{\delta}{n^2}) = \tilde{O}(1)$, we can see that LSTD($\lambda$) estimates $v_{LSTD(\lambda)}$ at the rate $\tilde{O}\left(\frac{1}{\sqrt{n}}\right)$. Finally, we can observe that since the function $\lambda \mapsto \frac{1}{\log\left(\frac{1}{\lambda\gamma}\right)}$ 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.

Theorem 2 (Tsitsiklis and Roy (1997))**.** *The approximation error satisfies*[7]*:*

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

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 $1-\delta$, for all $n \geq n_0(\delta)$, the global error of LSTD(λ) satisfies:*

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

[7] As suggested by V. Papavassilou (Tsitsiklis and Roy, 1997), this bound can in fact be improved by using the Pythagorean theorem to

$$\|v - v_{LSTD(\lambda)}\|_\mu \leq \frac{1-\lambda\gamma}{\sqrt{(1-\gamma)(1+\gamma-2\lambda\gamma)}}\|v - \Pi v\|_\mu.$$

We keep the simple form of Theorem 2 for simplicity.

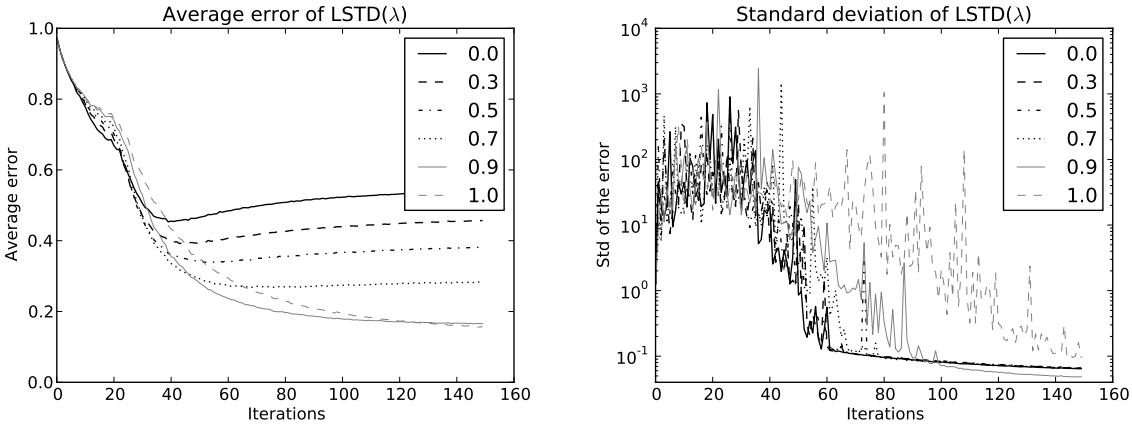

Figure 1: **Learning curves for different values of** λ. We generated 1000 random Garnet MDPs (Archibald *et al.*, 1995) with 100 states, random uniform rewards and $\gamma = 0.99$. We also generated 1000 random feature spaces of dimension 20 (by taking random matrices with random uniform entries). For all values of $\lambda \in \{0.0, 0.3, 0.5, 0.7, 0.9, 1.0\}$, we display (left) the average of the *real* error and (right) the standard deviation with respect to the number of samples. Empirically, the best value of λ appears to be a monotonic function of the number of samples n, that tends to 1 asymptotically. This is in accordance with our results in Corollary 1.

Remark 1. *The form of the result stated in Corollary 1 is slightly stronger than the one of Lazaric* et al. *(2012): for some property $P(n)$, our result if of the form "$\forall\delta,\ \exists n_0(\delta)$, such that $\forall n > n_0(\delta),\ P(n)$ holds with probability $1-\delta$" while theirs is of the form "$\forall n,\ \forall\delta,\ P(n)$ holds with probability $1-\delta$". Furthermore, 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:*

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

where $\tilde{v}_{LSTD(0)}$ is the truncation (with V_{max}) of the pathwise LSTD solution[8]*, while we get in this analysis*

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

The term corresponding to the approximation error is a factor $4\sqrt{2}$ better with our analysis. Moreover, contrary to what we do here, the analysis of Lazaric et al. *(2012) does not imply a rate of convergence for LSTD(λ) (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. Our two-step argument (bounding the estimation error with respect to $\|\cdot\|_\mu$, and then the approximation error with respect to $\|\cdot\|_\mu$) allows us to get a tighter result.*

As we have already mentioned, $\lambda = 1$ minimizes the bound on the approximation error $\|v - v_{LSTD(\lambda)}\|$ (the first term in the r.h.s. in Corollary 1) while $\lambda = 0$ minimizes the bound on the estimation error $\|v_{LSTD(\lambda)} - \hat{v}_{LSTD(\lambda)}\|$ (the second term). For any n, and for any δ, there exists hence a value λ^* that minimizes the global error bound by making an optimal compromise between the approximation and estimation errors. Figure 1 illustrates through simulations the interplay between λ and n. The optimal value λ^* depends on the process mixing parameters (b, κ and $\overline{\beta}$) as well as on the quality of the policy space $\|v - \Pi v\|_\mu$, which are quantities that are usually unknown in practice. However, when the number of samples n tends to infinity, it is clear that this optimal value λ^* tends to 1.

The next section contains a detailed proof of Theorem 1.

[8] See (Lazaric *et al.*, 2012) for more details.

4 Proof of Theorem 1

In this section, we develop the arguments underlying the results of the previous section. The proof is organized in two parts. In a first preliminary part, we prove a concentration inequality for vector processes: a general result that is based on infinitely-long eligibility traces. Then, in a second part, we actually prove Theorem 1: we apply this result to the error on estimating A and b, and relate these errors with that on $v_{LSTD(\lambda)}$.

4.1 Concentration inequality for infinitely-long trace-based estimates

One of the first difficulties for the analysis of LSTD(λ) is that the variables $A_i = z_i(\phi(X_i) - \gamma\phi(X_{i+1}))^T$ (respectively $b_i = z_i r(X_i)$) are not independent. Thus standard concentration results (like Lemma 6 we will describe in the Appendix A) for quantifying the speed at which the estimates converge to their limit cannot be used. As both terms $\hat{A}$ and $\hat{b}$ have the same structure, we will consider here a matrix that has the following general form:

$$\hat{G} = \frac{1}{n-1}\sum_{i=1}^{n-1} G_i \tag{7}$$

$$\text{with} \quad G_i = z_i(\tau(X_i, X_{i+1}))^T \tag{8}$$

with z_i, defined in Equation (5), satisfies $z_i = \sum_{k=1}^{i} (\lambda\gamma)^{i-k}\phi(X_k)$ and $\tau : \mathcal{X}^2 \mapsto \mathbb{R}^k$ is such that for $1 \leq i \leq k$, τ_i belongs to $\mathcal{B}(\mathcal{X}^2, L')$ for some finite L' [9]. The variables G_i are computed from one single trajectory, they are then significantly dependent. Nevertheless with the mixing assumption (Assumption 2), we can overcome this difficulty, and this by using a blocking technique due to Yu (1994). This technique leads us back to the independent case. However the transition from the mixing case to the independent one requires stationarity (Lemma 5) while G_i as a $\sigma(\mathcal{X}^{i+1})$ measurable function of the *non-stationary* vector $(X_1, \ldots, X_{i+1})$ does not define a *stationary* process. In order to satisfy the *stationarity* condition we will approximate G_i by it truncated stationary version G_i^m. This is possible if we approximate z_i by its m-truncated version:

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

Since the function ϕ is bounded by some constant L and the influence of the old events are controlled by some power of $\lambda\gamma < 1$, it is easy to check that $\|z_i - z_i^m\|_\infty \leq \frac{L}{1-\lambda\gamma}(\lambda\gamma)^m$. If we choose m such that $m > \frac{\log(n-1)}{\log\frac{1}{\lambda\gamma}}$, we obtain $\|z_i - z_i^m\|_2 = O\left(\frac{1}{n}\right)$. Therefore it seems reasonable to approximate $\hat{G}$ with the process $\hat{G}^m$ satisfying

$$\hat{G}^m = \frac{1}{n-1}\sum_{i=1}^{n-1} G_i^m, \tag{9}$$

$$\text{with} \quad G_i^m = z_i^m(\tau(X_i, X_{i+1}))^T. \tag{10}$$

For all $i \geq m$, G_i^m is a $\sigma(\mathcal{X}^{m+1})$ measurable function of the *stationary* vector $Z_i = (X_{i-m+1}, X_{i-m+2}, \ldots, X_{i+1})$. So we can apply the blocking technique of Yu (1994) to G_i^m, but before to do so we have to check out whether G_i^m well defines a β-mixing process. It can be shown (Yu, 1994) that any mesurable function f of a β-mixing process is a β^f-mixing process with $\beta^f \leq \beta$, so we only have to prove that the process Z_i is a β-mixing process. For that we need to relate its β coefficients to those of $(X_i)_{i\geq 1}$ on which Assumption 2 is made. This is the purpose of the following Lemma.

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

[9] We denote $\mathcal{X}^i = \underbrace{X \times \mathcal{X} \ldots \times \mathcal{X}}_{\text{i times}}$ for $i \geq 1$.

Proof. Let $\Gamma = \sigma(Z_1, ..., Z_t)$, by definition we have

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

For all $j \in \{1, ..., t\}$ 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

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

Then we have

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

Similarly we can prove that $\sigma(Z_{t+i}^{\infty}) = \sigma(X_{t+i}^{\infty})$. Then let β_i^X be the ith 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 $(Z_n)_{n\geq 1}$ we can see that

$$\beta_i^Z = \sup_{t\geq 1} \mathbb{E}\left[\sup_{B\in\sigma(Z_{t+i}^{\infty})} |P(B|\sigma(Z_1, ..., Z_t)) - P(B)|\right].$$

By applying what we developped above we obtain

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

Denote $t' = t + m$ then for $i > m$ we have

$$\begin{aligned}\beta_i^Z &= \sup_{t'\geq m+1} \mathbb{E}\left[\sup_{B\in\sigma(X_{t'+i-m}^{\infty})} |P(B|\sigma(X_1, ..., X_{t'})) - P(B)|\right] \\ &\leq \beta_{i-m}^X.\end{aligned}$$

□

Let $\|.\|_F$ denote the Frobenius norm satisfying : for $M \in \mathbb{R}^{d\times k}$, $\|M\|_F^2 = \sum_{l=1}^d \sum_{j=1}^k (M_{l,j})^2$. We are now ready to prove the 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*

$$G_i = \sum_{k=1}^{i} (\lambda\gamma)^{i-k} \phi(X_k)(\tau(X_i, X_{i+1}))^T. \tag{11}$$

Recall that $\phi = (\phi_1, \ldots, \phi_d)$ is such that for all j, $\phi_j \in \mathcal{B}(\mathcal{X}, L)$, and that $\tau \in \mathcal{B}(\mathcal{X}^2, L')$. Then for all δ in $(0, 1)$, with probability $1 - \delta$,

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

where

$$J(n, \delta) = 32\Gamma(n, \delta) \max\left\{\frac{\Gamma(n, \delta)}{b}, 1\right\}^{\frac{1}{\kappa}},$$

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

$$\epsilon(n) = 2\left\lceil \frac{\log(n-1)}{\log\left(\frac{1}{\lambda\gamma}\right)}\right\rceil \frac{\sqrt{d\times k}LL'}{(n-1)(1-\lambda\gamma)}.$$

Note that with respect to the quantities I and Λ introduced in Theorem 1, the quantities we introduce here are such that $J(n,\delta) = I(n, 4n^2\delta)$ and $\Gamma(n,\delta) = \Lambda(n, 4n^2\delta)$.

Proof. The proof amounts to show that i) the approximation due to considering the estimate $\hat{G}^m$ with truncated traces instead of $\hat{G}$ is bounded by $\epsilon(n)$, and then ii) to apply the block 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). We defer the technical arguments to Appendix A for readability. □

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

Lemma 3. *Let $Y = (Y_1, \dots, Y_n)$ 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 $\delta > 0$,*

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

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

Remark 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 results such as the one we describe in Lemma 6 we will describe in the Appendix A. Furthermore, the price to pay for having 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.*

4.2 Proof of Theorem 1

After having introduced the corresponding concentration inequality for infinitely-long trace-based estimates we are ready to prove Theorem 1. The first important step to Theorem 1 proof consists in deriving the following lemma.

Lemma 4. *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 estimate $\hat{v}_{LSTD(\lambda)}$ satisfies[10]:*

$$\|v_{LSTD(\lambda)} - \hat{v}_{LSTD(\lambda)}\|_\mu \leq \frac{1-\lambda\gamma}{(1-\gamma)\sqrt{\nu}} \|(I + \epsilon_A A^{-1})^{-1}\|_2 \|\epsilon_A \theta - \epsilon_b\|_2.$$

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. Starting from the definitions of $v_{LSTD(\lambda)}$ and $\hat{v}_{LSTD(\lambda)}$, we have

$$\begin{aligned} \hat{v}_{LSTD(\lambda)} - v_{LSTD(\lambda)} &= \Phi\hat{\theta} - \Phi\theta \\ &= \Phi A^{-1}(A\hat{\theta} - b). \end{aligned} \tag{12}$$

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

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

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

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

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

We know from Tsitsiklis and Roy (1997) that $\|\Pi\|_\mu = 1$—the projection matrix Π is defined in Equation (1)—and $\|P\|_\mu = 1$. 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

$$\Phi A^{-1} = (I - \Pi M)^{-1}\Phi M_\mu^{-1}. \tag{13}$$

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

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

Then we have

$$A\hat{\theta} - b = \hat{b} - \hat{A}\theta - \epsilon_A A^{-1}(b - A\hat{\theta}).$$

Consequently

$$\begin{aligned}
A\hat{\theta} - b &= (I + \epsilon_A A^{-1})^{-1}(\hat{b} - \hat{A}\theta) \\
&= (I + \epsilon_A A^{-1})^{-1}(\epsilon_b - \epsilon_A\theta)
\end{aligned} \tag{14}$$

where the last equality follows from the identity $A\theta = b$. Using Equations (13) and (14), Equation (12) can be rewritten as follows:

$$\hat{v}_{LSTD(\lambda)} - v_{LSTD(\lambda)} = (I - \Pi M)^{-1}\Phi M_\mu^{-1}(I + \epsilon_A A^{-1})^{-1}(\epsilon_b - \epsilon_A\theta). \tag{15}$$

Now we will try to bound $\|\Phi M_\mu^{-1}(I + \epsilon_A A^{-1})^{-1}(\epsilon_b - \epsilon_A\theta)\|_\mu$. Notice that for all x,

$$\|\Phi M_\mu^{-1}x\|_\mu = \sqrt{x^T M_\mu^{-1}\Phi^T D_\mu \Phi M_\mu^{-1}x} = \sqrt{x^T M_\mu^{-1}x} \leq \frac{1}{\sqrt{\nu}}\|x\|_2 \tag{16}$$

where ν is the smallest (real) eigenvalue of the Gram matrix M_μ. By taking the norm in Equation (15) and using the above relation, we get

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

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 imply that

$$\|(I - \Pi M)^{-1}\|_\mu = \left\|\sum_{i=0}^{\infty}(\Pi M)^i\right\|_\mu \leq \sum_{i=0}^{\infty}\|\Pi M\|_\mu^i \leq \frac{1}{1 - \frac{(1-\lambda)\gamma}{1-\lambda\gamma}} = \frac{1-\lambda\gamma}{1-\gamma}. \tag{17}$$

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 $\|\epsilon_A\|_2 \leq \epsilon < C < \frac{1}{\|A^{-1}\|_2}$, we have

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

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

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

This concludes the proof of Lemma 4. □

To finish the proof of Theorem 1, Lemma 4 suggests that we should control both terms $\|\epsilon_A\|_2$ and $\|\epsilon_A \theta - \epsilon_b\|_2$ with high probability. This is what we do now.

Controlling $\|\epsilon_A\|_2$. By the triangle inequality, we can see that

$$\|\epsilon_A\|_2 \leq \|\mathbb{E}[\epsilon_A]\|_2 + \|\epsilon_A - \mathbb{E}[\epsilon_A]\|_2. \tag{18}$$

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 (18) as follows, by replacing A with its expression in (3):

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

Let (δ_n) a parameter in $(0,1)$ depending on n, that we will fix later, a consequence of Equation (18) and the just derived bound is that:

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

if we choose $\epsilon_1(n,\delta_n)$ such that (cf. Lemma 2)

$$\epsilon_1(n,\delta_n) - \epsilon_0(n) = \frac{4dL^2}{(1-\lambda\gamma)\sqrt{n-1}}\sqrt{\left(\left\lceil \frac{\log(n-1)}{\log\left(\frac{1}{\lambda\gamma}\right)}\right\rceil + 1\right) J(n-1,\delta_n) + \epsilon(n)}$$

where $\epsilon(n) = \frac{4mdL^2}{(n-1)(1-\lambda\gamma)}$, that is if

$$\epsilon_1(n,\delta_n) = \frac{4dL^2}{(1-\lambda\gamma)\sqrt{n-1}}\sqrt{\left(\left\lceil \frac{\log(n-1)}{\log\left(\frac{1}{\lambda\gamma}\right)}\right\rceil + 1\right) J(n-1,\delta_n) + \epsilon(n)} + \epsilon_0(n). \tag{19}$$

Controlling $\|\epsilon_A\theta - \epsilon_b\|_2$. 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

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

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

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

We can control $\|\epsilon_A\theta - \epsilon_b\|_2$ by following the same proof steps as above. In fact we have

$$\begin{aligned}
&\|\epsilon_A\theta - \epsilon_b\|_2 \leq \|\epsilon_A\theta - \epsilon_b - \mathbb{E}[\epsilon_A\theta - \epsilon_b]\|_2 + \|\mathbb{E}[\epsilon_A\theta - \epsilon_b]\|_2, \\
\text{and} \quad &\|\mathbb{E}[\epsilon_A\theta - \epsilon_b]\|_2 \leq \|\mathbb{E}[\epsilon_A]\|_2\|\theta\|_2 + \|\mathbb{E}[\epsilon_b]\|_2.
\end{aligned} \tag{20}$$

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

$$\|\mathbb{E}[\epsilon_A\theta - \epsilon_b]\|_2 \leq \frac{1}{n-1}\frac{2dL^2}{(1-\lambda\gamma)^2}\|\theta\|_2 + \frac{1}{n-1}\frac{\sqrt{d}LR_{\max}}{(1-\lambda\gamma)^2} = \epsilon_0'(n).$$

As a consequence of Equation (20) and the just derived bound we have

$$\mathbb{P}(\|\epsilon_A\theta - \epsilon_b\|_2 \geq \epsilon_2(\delta_n)) \leq \mathbb{P}(\|\epsilon_A\theta - \epsilon_b - \mathbb{E}[\epsilon_A\theta - \epsilon_b]\|_2 \geq \epsilon_2(\delta_n) - \epsilon_0'(n)) \leq \delta_n$$

if we choose $\epsilon_2(\delta_n)$ such that (cf Lemma 2)

$$\epsilon_2(\delta_n) = \frac{2\sqrt{d}L\|\Delta_i\|_\infty}{(1-\lambda\gamma)\sqrt{n-1}}\sqrt{\left(\left\lceil\frac{\log(n-1)}{\log\left(\frac{1}{\lambda\gamma}\right)}\right\rceil + 1\right)J(n-1,\delta_n)} + \frac{2\sqrt{d}L\|\Delta_i\|_\infty}{(n-1)(1-\lambda\gamma)}\left\lceil\frac{\log(n-1)}{\log\left(\frac{1}{\lambda\gamma}\right)}\right\rceil + \epsilon_0'(n). \tag{21}$$

It remains to compute a bound on $\|\Delta_i\|_\infty$. To do so, it suffices to bound $v_{LSTD(\lambda)}$. For all $x \in \mathcal{X}$, we have

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

where the first inequality is obtained from the Cauchy-Schwarz inequality. We thus need to bound $\|\theta\|_2$. On the one hand, we have

$$\|v_{LSTD(\lambda)}\|_\mu = \|\Phi\theta\|_\mu = \sqrt{\theta^T M_\mu\theta} \geq \sqrt{\nu}\|\theta\|_2,$$

and on the other hand, we have

$$\|v_{LSTD(\lambda)}\|_\mu = \|(I - \Pi M)^{-1}\Pi(I - \lambda\gamma P)^{-1}r\|_\mu \leq \frac{R_{\max}}{1-\gamma} = V_{\max}.$$

Therefore

$$\|\theta\|_2 \leq \frac{V_{\max}}{\sqrt{\nu}}.$$

We can conclude that

$$\forall x \in \mathcal{X},\ |v_{LSTD(\lambda)}(x)| \leq \frac{\sqrt{d}LV_{\max}}{\sqrt{\nu}}.$$

Then for all i we have

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

Since $\Phi^T D_\mu \Phi$ is a symmetric matrix, we have $\nu \leq \|\Phi^T D_\mu \Phi\|_2$. We can see that

$$\|\Phi^T D_\mu \Phi\|_2 \leq d \max_{j,k} |\phi_k^t D_\mu \phi_j| = d \max_{j,k} |\phi_k^t D_\mu^{\frac{1}{2}} D_\mu^{\frac{1}{2}} \phi_j| \leq d \max_{j,k} \|\phi_k^t\|_\mu \|\phi_j\|_\mu \leq dL^2,$$

so that $\nu \leq dL^2$. It follows that, for all i

$$|\Delta_i| \leq \frac{\sqrt{d}LV_{\max}}{\sqrt{\nu}} + \gamma\frac{\sqrt{d}LV_{\max}}{\sqrt{\nu}} + \frac{\sqrt{d}L}{\sqrt{\nu}}(1-\gamma)V_{\max} = 2\frac{\sqrt{d}L}{\sqrt{\nu}}V_{\max}.$$

Conclusion of the proof. We are ready to conclude the proof. Now that we know how to control both terms $\|\epsilon_A\|_2$ and $\|\epsilon_A\theta - \epsilon_b\|_2$, we can see that

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

if we choose $\delta_n = \frac{1}{4n^2}\delta$. By the second part of Lemma 4, for all δ, with probability at least $1-\delta$, for all n such that $\epsilon_1(n,\delta_n) < C$, $\hat{A}$ is invertible and

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

We get the bound of the Theorem by replacing $\epsilon_1(n,\delta_n)$ and $\epsilon_2(n,\delta_n)$ with their definitions in Equations (19) and (21).

To complete the proof of Theorem 1, we now need to show how to pick C, which will allow to show that the condition $\epsilon_1(n,\delta_n) < C \leq \frac{1}{\|A^{-1}\|_2}$ is equivalent to the one that characterizes the index $n_0(\delta)$ in the Theorem. Indeed 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

$$\|\Phi A^{-1} v\|_\mu = \|(I - \Pi M)^{-1}\Phi M_\mu^{-1} v\|_\mu \leq \frac{1-\lambda\gamma}{1-\gamma}\|\Phi M_\mu^{-1} v\|_\mu \leq \frac{1-\lambda\gamma}{(1-\gamma)\sqrt{\nu}}\|v\|_2$$

where the last inequality is obtained from Equation (16). Then

$$\|A^{-1}\|_2 \leq \frac{1-\lambda\gamma}{(1-\gamma)\nu},$$

and consequently we can take $C = \frac{(1-\gamma)\nu}{1-\lambda\gamma}$. This concludes the proof of Theorem 1.

5 Conclusion and Future Work

This paper introduces a high-probability convergence rate for the algorithm LSTD(λ) in terms of the number of samples n and the parameter λ. We have shown 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. To do so, we introduced an original vector concentration inequality (Lemma 2) for estimates that are based on eligibility traces. A simplified version of this concentration inequality (Lemma 3), that applies to general stationary beta-mixing processes, may be useful in many other contexts where we want to relax the i.i.d. hypothesis on the samples.

The performance bound that we deduced is more accurate than the one from Lazaric *et al.* (2012), restricted to the case $\lambda = 0$. The analysis that they proposed was based on a Markov design regression model. By using the trace truncation technique we have employed, we believe it is possible to extend the proof of Lazaric *et al.* (2012) to the general case λ in $(0, 1)$. However we would still pay a $4\sqrt{2}$ extra factor in the final bound.

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 and Lesner (2012) showed that it allows to improve the overall performance of the Policy Iteration Scheme. Finally, a challenging question would be to consider LSTD(λ) in the off-policy case, for which the convergence has recently been proved by Yu (2010).

A Proof of Lemma 2

Writing for a given integer $m > 1$

$$\epsilon_1 = \frac{1}{n-1}\sum_{i=1}^{m-1} G_i - \mathbb{E}[G_i]$$

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

we have

$$\begin{aligned}
\frac{1}{n-1}\sum_{i=1}^{n-1} G_i - \mathbb{E}[G_i] &= \frac{1}{n-1}\sum_{i=m}^{n-1} G_i - \mathbb{E}[G_i] + \epsilon_1 \\
&= \frac{1}{n-1}\sum_{i=m}^{n-1} z_i\tau(X_i, X_{i+1})^T - \mathbb{E}[z_i\tau(X_i, X_{i+1})^T] + \epsilon_1 \\
&= \frac{1}{n-1}\sum_{i=m}^{n-1} z_i^m\tau(X_i, X_{i+1})^T - \mathbb{E}[z_i^m\tau(X_i, X_{i+1})^T] + \epsilon_1 + \epsilon_2 \\
&= \frac{1}{n-1}\sum_{i=m}^{n-1}(G_i^m - \mathbb{E}[G_i^m]) + \epsilon_1 + \epsilon_2. \qquad (22)
\end{aligned}$$

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

$$\|\epsilon_1 + \epsilon_2\|_2 \leq \frac{2(m-1)\sqrt{d \times k}LL'}{(n-1)(1-\lambda\gamma)} + \frac{2(\lambda\gamma)^m\sqrt{d \times k}LL'}{(1-\lambda\gamma)} \qquad (23)$$

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 $\epsilon > 0$,

$$\mathbb{P}\left(\left\|\frac{1}{n-m}\sum_{i=m}^{n-1}(G_i^m - \mathbb{E}[G_i^m])\right\|_2 \geq \epsilon\right) \leq \mathbb{P}\left(\left\|\frac{1}{n-m}\sum_{i=m}^{n-1}(G_i^m - \mathbb{E}[G_i^m])\right\|_F \geq \epsilon\right)$$
$$= \mathbb{P}\left(\left\|\frac{1}{n-m}\sum_{i=m}^{n-1}(U_i^m - \mathbb{E}[U_i^m])\right\|_2 \geq \epsilon\right). \tag{24}$$

The variables U_i^m define a stationary β-mixing process (Lemma 1). To deal with the β-mixing assumption, we use the decomposition technique proposed by Yu (1994) that consists in dividing the stationary 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

$$\mathbb{P}\left(\left\|\frac{1}{n-m}\sum_{i=m}^{n-1} U_i^m - \mathbb{E}[U_i^m]\right\|_2 \geq \epsilon\right) \leq \mathbb{P}\left(\left\|\sum_{i\in H} U_i^m - \mathbb{E}[U_i^m]\right\|_2 + \left\|\sum_{i\in E} U_i^m - \mathbb{E}[U_i^m]\right\|_2 \geq (n-m)\frac{\epsilon}{2}\right) \tag{25}$$

$$\leq \mathbb{P}\left(\left\|\sum_{i\in H} U_i^m - \mathbb{E}[U_i^m]\right\|_2 \geq \frac{(n-m)\epsilon}{4}\right) + \mathbb{P}\left(\left\|\sum_{i\in E} U_i^m - \mathbb{E}[U_i^m]\right\|_2 \geq \frac{(n-m)\epsilon}{4}\right) \tag{26}$$

$$= 2\mathbb{P}\left(\left\|\sum_{i\in H} U_i^m - \mathbb{E}[U_i^m]\right\|_2 \geq \frac{(n-m)\epsilon}{4}\right) \tag{27}$$

where Equation (25) follows from the triangle inequality, Equation (26) from the fact that the event $\{X + Y \geq a\}$ implies $\{X \geq \frac{a}{2}\}$ or $\{Y \geq \frac{a}{2}\}$, and Equation (27) from the assumption that the process is stationary. Since $H = \cup_{l=1}^{\mu_{n-m}} H_l$ we have

$$\mathbb{P}\left(\left\|\frac{1}{n-m}\sum_{i=m}^{n-1} U_i^m - \mathbb{E}[U_i^m]\right\|_2 \geq \epsilon\right) \leq 2\mathbb{P}\left(\left\|\sum_{l=1}^{\mu_{n-m}}\sum_{i\in H_l} U_i^m - \mathbb{E}[U_i^m]\right\|_2 \geq \frac{(n-m)\epsilon}{4}\right)$$
$$= 2\mathbb{P}\left(\left\|\sum_{l=1}^{\mu_{n-m}} U(H_l) - \mathbb{E}[U(H_l)]\right\|_2 \geq \frac{(n-m)\epsilon}{4}\right) \tag{28}$$

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 5. *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 5, Equation (28) leads to:

$$\mathbb{P}\left(\left\|\frac{1}{n-m}\sum_{i=m}^{n-1} U_i^m - \mathbb{E}[U_i^m]\right\|_2 \geq \epsilon\right) \leq 2\mathbb{P}\left(\left\|\sum_{l=1}^{\mu_{n-m}} U'(H_l) - \mathbb{E}[U'(H_l)]\right\|_2 \geq \frac{(n-m)\epsilon}{4}\right) + 2\mu_{n-m}\beta_{a_{n-m}}. \tag{29}$$

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 6 (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,*

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

Indeed, taking $X_{\mu_{n-m}} = \sum_{l=}^{\mu_{n-m}} U'(H_l) - \mathbb{E}[U'(H_l)]$, 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 (28) and (29), we finally obtain

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

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

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

So the equation above may be re-written as

$$\mathbb{P}\left(\left\|\frac{1}{n-m}\sum_{i=m}^{n-1} U_i^m - \mathbb{E}[U_i^m]\right\|_2 \geq \epsilon\right) \leq 4e^2 e^{-\frac{(n-m)\epsilon^2}{16a_{n-m}C^2}} + 2(n-m)\overline{\beta} e^{-b(a_{n-m}-m)^\kappa} = \delta'. \tag{30}$$

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

$$\delta' \leq (4e^2 + (n-m)\overline{\beta}) \exp\left(-\min\left\{\left(\frac{b}{(n-m)\epsilon^2 C_2}\right), 1\right\}^{\frac{1}{k+1}} \frac{1}{2}(n-m)C_2\epsilon^2\right). \tag{31}$$

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

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

Indeed[11], there are two cases:

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

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

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

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

By combining Equations (31) and (32), 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 (30) is reduced to

$$\mathbb{P}\left(\left\|\frac{1}{n-m}\sum_{i=m}^{n-1}(U_i^m-\mathbb{E}[U_i^m])\right\|_2\geq\frac{C}{\sqrt{n-m}}\left(\zeta J(n-m,\delta)\right)^{\frac{1}{2}}\right)\leq\delta. \tag{33}$$

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

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

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

By using Equations (24) and (33), we deduce that

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

By combining Equations (22), (23),(34), plugging the value of $C = \frac{2\sqrt{dk}LL'}{1-\lambda\gamma}$, and taking $m = \left\lceil \frac{\log(n-1)}{\log\frac{1}{\lambda\gamma}} \right\rceil$, we get the announced result.

References

Archibald, T., McKinnon, K., and Thomas, L. (1995). On the generation of Markov decision processes. *Journal of the Operational Research Society*, **46**, 354–361.

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

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

Hayes, T. P. (2005). A large-deviation inequality for vector-valued martingales. Manuscript.

Lazaric, A., Ghavamzadeh, M., and Munos, R. (2012). Finite-sample analysis of least-squares policy iteration. *Journal of Machine Learning Research*, **13**, 3041–3074.

Nedic, A. and Bertsekas, D. P. (2002). Least squares policy evaluation algorithms with linear function approximation. *Theory and Applications*, **13**, 79–110.

Scherrer, B. (2010). Should one compute the temporal difference fix point or minimize the bellman residual? the unified oblique projection view. In *ICML*.

Scherrer, B. and Lesner, B. (2012). On the use of non-stationary policies for stationary infinite-horizon Markov decision processes. In *NIPS 2012 Adv.in Neural Information Processing Systems*, South Lake Tahoe, United States.

Szepesvári, C. (2010). *Algorithms for Reinforcement Learning*. Morgan and Claypool.

Tsitsiklis, J. N. and Roy, B. V. (1997). An analysis of temporal-difference learning with function approximation. Technical report, IEEE Transactions on Automatic Control.

Yu, B. (1994). Rates of convergence for empirical processes stationnary mixing consequences. *The Annals of Probability*, **19**, 3041–3074.

Yu, H. (2010). Convergence of least-squares temporal difference methods under general conditions. In *ICML*.