Master MVA: Apprentissage par renforcement	**Lecture: 5**

Sample complexity en apprentissage par renforcement

Professeur: Rémi Munos *http://researchers.lille.inria.fr/~munos/master-mva/*

Références bibliographiques: [LGM10b, MS08, MMLG10, ASM08]

Plan:

1. Inégalité d'Azuma
2. Sample complexity of LSTD
3. Other results

1 Inégalité d'Azuma

Etend l'inégalité de Chernoff-Hoeffding à des variables aléatoires qui peuvent être dépendantes mais qui forment une Martingale.

Proposition 1. Soient $X_i \in [a_i, b_i]$ variables aléatoires telles que $\mathbb{E}[X_i|X_1, \dots, X_{i-1}] = 0$. Alors

$$\mathbb{P}\big(\big|\sum_{i=1}^n X_i\big| \geq \epsilon\big) \leq 2e^{-\frac{2\epsilon^2}{\sum_{i=1}^n (b_i-a_i)^2}}. \tag{1}$$

Autrement dit, pour tout $\delta \in (0,1]$, on a avec probabilité au moins $1-\delta$,

$$\frac{1}{n}\big|\sum_{i=1}^n X_i\big| \leq \sqrt{\frac{1}{n}\sum_{i=1}^n (b_i-a_i)^2}\sqrt{\frac{\log 2/\delta}{2n}}.$$

Preuve. On a:

$$\begin{aligned}
\mathbb{P}(\sum_{i=1}^n X_i \geq \epsilon) &= \mathbb{P}(e^{s\sum_{i=1}^n X_i} \geq e^{s\epsilon}) \\
&\leq e^{-s\epsilon}\mathbb{E}[e^{s\sum_{i=1}^n X_i}], \text{par Markov} \\
&\leq e^{-s\epsilon}\mathbb{E}_{X_1,\dots,X_{n-1}}\Big[\mathbb{E}_{X_n}\big[e^{s\sum_{i=1}^n X_i}\big|X_1,\dots,X_{n-1}\big]\Big], \\
&\leq e^{-s\epsilon}\mathbb{E}_{X_1,\dots,X_{n-1}}\Big[e^{s\sum_{i=1}^{n-1} X_i}\mathbb{E}_{X_n}\big[e^{sX_n}\big|X_1,\dots,X_{n-1}\big]\Big], \\
&\leq e^{-s\epsilon+s^2(b_n-a_n)^2/8}\mathbb{E}_{X_1,\dots,X_{n-1}}\big[e^{s\sum_{i=1}^{n-1} X_i}\big], \text{ par Hoeffding} \\
&\leq e^{-s\epsilon+s^2\sum_{i=1}^n (b_i-a_i)^2/8}
\end{aligned}$$

En choisissant $s = 4\epsilon/\sum_{i=1}^n (b_i-a_i)^2$ on déduit $\mathbb{P}\big(\sum_{i=1}^n X_i - \mu_i \geq \epsilon\big) \leq e^{-\frac{2\epsilon^2}{\sum_{i=1}^n (b_i-a_i)^2}}$. En refaisant le même calcul pour $\mathbb{P}\big(\sum_{i=1}^n X_i - \mu_i \leq -\epsilon\big)$ on déduit (1). □

2 Sample complexity of LSTD

2.1 Pathwise LSTD

We follow a fixed policy π. Our goal is to approximate the value function V^π (written V removing reference to π to simplify notations). We use a linear approximation space $\mathcal{F}$ spanned by a set of d basis functions $\varphi_i : \mathcal{X} \to \mathbb{R}$. We denote by $\phi : \mathcal{X} \to \mathbb{R}^d$, $\phi(\cdot) = \big(\varphi_1(\cdot), \ldots, \varphi_d(\cdot)\big)^\top$ the feature vector. Thus

$$\mathcal{F} = \big\{ f_\alpha \mid \alpha \in \mathbb{R}^d \text{ and } f_\alpha(\cdot) = \phi(\cdot)^\top \alpha \big\}.$$

Let $(X_1, \ldots, X_n)$ be a sample path (trajectory) of size n generated by following policy π. Let $v \in \mathbb{R}^n$ and $r \in \mathbb{R}^n$ such that $v_t = V(X_t)$ and $r_t = R(X_t)$ be the value vector and the reward vector, respectively. Also, let $\Phi = [\phi(X_1)^\top; \ldots; \phi(X_n)^\top]$ be the feature matrix defined at the states, and $\mathcal{F}_n = \{\Phi\alpha,\ \alpha \in \mathbb{R}^d\} \subset \mathbb{R}^n$ be the corresponding vector space. We denote by $\widehat{\Pi} : \mathbb{R}^n \to \mathcal{F}_n$ the **empirical orthogonal projection** onto $\mathcal{F}_n$, defined as

$$\widehat{\Pi} y = \arg\min_{z \in \mathcal{F}_n} ||y - z||_n,$$

where $||y||_n^2 = \frac{1}{n} \sum_{t=1}^n y_t^2$. Note that $\widehat{\Pi}$ is a non-expansive mapping w.r.t. the ℓ_2-norm: $||\widehat{\Pi} y - \widehat{\Pi} z||_n \leq ||y - z||_n$.

Define the *empirical Bellman operator* $\widehat{\mathcal{T}} : \mathbb{R}^n \to \mathbb{R}^n$ as

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

Proposition 2. The operator $\widehat{\Pi}\widehat{\mathcal{T}}$ is a contraction in ℓ_2-norm, thus possesses a unique fixed point $\hat{v}$.

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

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

Now, since the orthogonal projection $\widehat{\Pi}$ is non-expansive w.r.t. ℓ_2-norm, from Banach fixed point theorem, there exists a unique fixed-point $\hat{v}$ of the mapping $\widehat{\Pi}\widehat{\mathcal{T}}$, i.e., $\hat{v} = \widehat{\Pi}\widehat{\mathcal{T}}\hat{v}$. □

Since $\hat{v}$ is the unique fixed point of $\widehat{\Pi}\widehat{\mathcal{T}}$, the vector $\hat{v} - \widehat{\mathcal{T}}\hat{v}$ is perpendicular to the space $\mathcal{F}_n$, and thus, $\Phi^\top(\hat{v} - \widehat{\mathcal{T}}\hat{v}) = 0$. By replacing $\hat{v}$ with $\Phi\alpha$, we obtain $\Phi^\top\Phi\alpha = \Phi^\top(r + \gamma\widehat{P}\Phi\alpha)$ and then $\underbrace{\Phi^\top(I - \gamma\widehat{P})\Phi}_{A}\,\alpha = \underbrace{\Phi^\top r}_{b}$.

Therefore, by setting

$$\begin{aligned} A_{i,j} &= \sum_{t=1}^{n-1} \phi_i(x_t)\big[\phi_j(x_t) - \gamma\phi_j(x_{t+1})\big] + \phi_i(x_n)\phi_j(x_n), \\ b_i &= \sum_{t=1}^{n} \phi_i(x_t) r_t, \end{aligned}$$

we have that the system $A\alpha = b$ always has at least one solution (since the fixed point $\hat{v}$ exists) and we call the solution with minimal norm, $\hat{\alpha} = A^+ b$, where A^+ is the Moore-Penrose pseudo-inverse of A, the pathwise LSTD solution.

2.2 Performance Bound

Here we derive a bound for the performance of $\hat{v}$ evaluated on the states of the trajectory used by the pathwise LSTD algorithm.

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

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

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

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

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

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

Remark 3 Theorem 1 provides a bound without any reference to the stationary distribution of the Markov chain. In fact, the bound of Equation 2 holds even when the chain does not possess a stationary distribution. For example, consider a Markov chain on the real line where the transitions always move the states to the right, i.e., $p(X_{t+1} \in dy | X_t = x) = 0$ for $y \leq x$. For simplicity assume that the value function V is bounded and belongs to $\mathcal{F}$. This Markov chain is not recurrent, and thus, does not have a stationary distribution. We also assume that the feature vectors $\phi(X_1), \ldots, \phi(X_n)$ are sufficiently independent, so that the eigenvalues of $\frac{1}{n}\Phi^\top\Phi$ are greater than $\nu > 0$. Then according to Theorem 1, pathwise LSTD is able to estimate the value function at the samples at a rate $O(1/\sqrt{n})$. This may seem surprising because at each state X_t the algorithm is only provided with a noisy estimation of the expected value of the next state. However, the estimates are unbiased conditioned on the current state, and we will see in the proof that using a concentration inequality for martingale, pathwise LSTD is able to learn a good estimate of the value function at a state X_t using noisy pieces of information at other states that may be far away from X_t. In other words, learning the value function at a given state does not require making an average over many samples close to that state. This implies that LSTD does not require the Markov chain to possess a stationary distribution.

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

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

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

The next lemma reports a risk bound for the Markov design setting.

Lemme (Regression bound for the Markov design setting). Let $\hat{w} \in \mathcal{F}_n$ be the least-squares estimate of the (noisy) values $Y = \{Y_t\}_1^n$, i.e., $\hat{w} = \widehat{\Pi} Y$, and $w \in \mathcal{F}_n$ be the least-squares estimate of the (noiseless) values

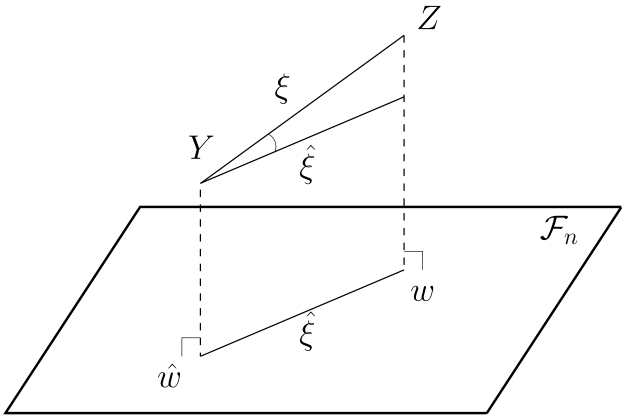

Figure 1: This figure shows the components used in Lemma 2.2 and its proof such as w, $\hat{w}$, ξ, and $\hat{\xi}$, and the fact that $\langle\hat{\xi},\xi\rangle_n = ||\hat{\xi}||_n^2$.

$Z = \{Z_t = f(X_t)\}_1^n$, i.e., $w = \widehat{\Pi} Z$. Then for any $\delta > 0$, with probability at least $1-\delta$ (the probability is w.r.t. the random sample path $X_1,\dots,X_n$), we have

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

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

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

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

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

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

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

By using the result of Equation 6 in Equation 5, we obtain

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

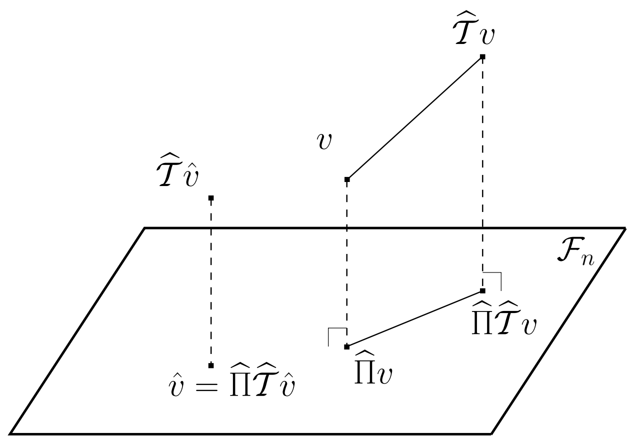

Figure 2: This figure represents the space $\mathbb{R}^n$, the linear vector subspace $\mathcal{F}_n$ and some vectors used in the proof of Theorem 1.

Now, from Equation 3, we have that for any $i = 1, \ldots, d$

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

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

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

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

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

The result follows by combining Equations 8 and 7. □

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

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

Preuve. [Théorème 1]

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

$$||\hat{v}-v||_n^2 = ||v-\widehat{\Pi}v||_n^2 + ||\hat{v}-\widehat{\Pi}v||_n^2 \leq ||\hat{v}-\widehat{\Pi}v||_n^2 + \big(||\hat{v}-\widehat{\Pi}\widehat{\mathcal{T}}v||_n + ||\widehat{\Pi}\widehat{\mathcal{T}}v-\widehat{\Pi}v||_n\big)^2. \tag{9}$$

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

$$||\hat{v}-\widehat{\Pi}\widehat{\mathcal{T}}v||_n = ||\widehat{\Pi}\widehat{\mathcal{T}}\hat{v}-\widehat{\Pi}\widehat{\mathcal{T}}v||_n \leq \gamma||\hat{v}-v||_n, \tag{10}$$

Thus from Equation 9 and 10, we have

$$||\hat{v}-v||_n^2 \leq ||\hat{v}-\widehat{\Pi}v||_n^2 + \big(\gamma||\hat{v}-v||_n + ||\widehat{\Pi}\widehat{\mathcal{T}}v-\widehat{\Pi}v||_n\big)^2. \tag{11}$$

Step 2: We now provide a high probability bound on $||\widehat{\Pi}\widehat{\mathcal{T}}v-\widehat{\Pi}v||_n$. This is a consequence of Lemma 2.2 applied to the vectors $Y=\widehat{\mathcal{T}}v$ and $Z=v$. Since v is the value function at the points $\{X_t\}_1^n$, from the definition of the pathwise Bellman operator, we have that for $1\leq t\leq n-1$,

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

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

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

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

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

where ν_n is the smallest strictly-positive eigenvalue of $\frac{1}{n}\Phi^\top\Phi$. The claim follows by combining Equations 12 and 11, and solving the result for $||\hat{v}-v||_n$. □

2.3 Generalization bound

When the Markov chain is ergodic (say β-mixing) and possesses a stationnary distribution μ, then it is possible to derive generalization bounds of the form: with probability $1-\delta$,

$$||\hat{V}-V||_\mu \leq \frac{c}{\sqrt{1-\gamma^2}}\inf_{f\in\mathcal{F}}||V-f||_\mu + O\Big(\sqrt{\frac{d\log(d/\delta)}{n\nu}}\Big),$$

which provides a bound expressed in terms of

- the best possible approximation of V in $\mathcal{F}$ measured with μ
- the smallest eigenvalue ν of the Gram matrix $\big(\int\phi_i\phi_j d\mu\big)_{i,j}$
- β-mixing coefficients of the chain (hidden in O).

(see [Lazaric, Ghavamzadeh, Munos, *Finite-sample analysis of LSTD*, 2010]).

3 Other results

Similar results have been obtained for different algorithms:

- Approximate Value iteration [MS08]
- Policy iteration with Bellman residual minimization [MMLG10]
- Policy iteration with modified Bellman residual minimization [ASM08]
- Classification based policy iteration algorithm [LGM10a]

But there remains many open problems...

References

[ASM08] A. Antos, Cs. Szepesvari, and R. Munos. Learning near-optimal policies with Bellman-residual minimization based fitted policy iteration and a single sample path. *Machine Learning Journal*, 71:89–129, 2008.

[LGM10a] A. Lazaric, M. Ghavamzadeh, and R. Munos. Analysis of a classification-based policy iteration algorithm. In *International Conference on Machine Learning*, pages 607–614, 2010.

[LGM10b] A. Lazaric, M. Ghavamzadeh, and R. Munos. Finite-sample analysis of LSTD. In *International Conference on Machine Learning*, pages 615–622, 2010.

[MMLG10] O. A. Maillard, R. Munos, A. Lazaric, and M. Ghavamzadeh. Finite sample analysis of bellman residual minimization. In Masashi Sugiyama and Qiang Yang, editors, *Asian Conference on Machine Learning. JMLR: Workshop and Conference Proceedings*, volume 13, pages 309–324, 2010.

[MS08] R. Munos and Cs. Szepesvári. Finite time bounds for sampling based fitted value iteration. *Journal of Machine Learning Research*, 9:815–857, 2008.