JMLR: Workshop and Conference Proceedings 13: 299-314
2nd Asian Conference on Machine Learning (ACML2010), Tokyo, Japan, Nov. 8–10, 2010.

Finite-Sample Analysis of Bellman Residual Minimization

Odalric-Ambrym Maillard ODALRIC.MAILLARD@INRIA.FR

Rémi Munos REMI.MUNOS@INRIA.FR

Alessandro Lazaric ALESSANDRO.LAZARIC@INRIA.FR

Mohammad Ghavamzadeh MOHAMMAD.GHAVAMZADEH@INRIA.FR
All authors are from Sequel team, INRIA Lille - Nord Europe

Editor: Masashi Sugiyama and Qiang Yang

Abstract

We consider the Bellman residual minimization approach for solving discounted Markov decision problems, where we assume that a generative model of the dynamics and rewards is available. At each policy iteration step, an approximation of the value function for the current policy is obtained by minimizing an empirical Bellman residual defined on a set of n states drawn i.i.d. from a distribution μ, the immediate rewards, and the next states sampled from the model. Our main result is a generalization bound for the Bellman residual in linear approximation spaces. In particular, we prove that the empirical Bellman residual approaches the true (quadratic) Bellman residual in μ-norm with a rate of order $O(1/\sqrt{n})$. This result implies that minimizing the empirical residual is indeed a sound approach for the minimization of the true Bellman residual which guarantees a good approximation of the value function for each policy. Finally, we derive performance bounds for the resulting approximate policy iteration algorithm in terms of the number of samples n and a measure of how well the function space is able to approximate the sequence of value functions.

Keywords: Markov decision processes, reinforcement learning, Bellman residual minimization, generalization bounds, finite sample analysis

1. Introduction

In this paper we consider the problem of solving a Markov decision problem (MDP) (Bertsekas and Shreve, 1978; Puterman, 1994) by means of an approximate policy iteration algorithm (Bertsekas and Tsitsiklis, 1996b; Si et al., 2004; Powell, 2007) with a linear approximation space $\mathcal{F}$. In particular, we focus on the Bellman residual minimization approach (Schweitzer and Seidmann, 1985; Baird, 1995; Munos, 2003; Lagoudakis and Parr, 2003; Scherrer, 2010) when a generative model is available, that is, for any state-action pair it is possible to obtain the immediate reward and an independent sample of the next state drawn from the transition distribution.

More in details, at each iteration k, in order to evaluate the current policy π_k, we build an approximation $V_k \in \mathcal{F}$ of the value function V^{π_k} by solving an empirical Bellman residual minimization problem: $V_k = \arg\min_{f \in \mathcal{F}} \mathcal{B}_n(f)$, where $\mathcal{B}_n(f)$ is the empirical Bellman

residual. The specific definition of $\mathcal{B}_n$ is critical since, as observed in several previous works (see e.g., Sutton and Barto 1998; Lagoudakis and Parr 2003; Antos et al. 2008), the squared temporal difference between successive states (e.g., states obtained following a single trajectory), gives rise to a biased estimate of the (true) Bellman residual $\mathcal{B}(f) = ||f - \mathcal{T}^\pi f||_\mu^2$. In this paper, in order to build an unbiased estimate of $\mathcal{B}(f)$ we take advantage of the generative model and build $\mathcal{B}_n$ on n states drawn i.i.d. from a given distribution μ, as well as the immediate rewards and two next states independently sampled from the generative model (i.e., the double-sampling technique suggested in Sutton and Barto 1998, p. 220).

Motivation. The idea of minimizing the Bellman residual is natural (see e.g., Schweitzer and Seidmann 1985; Baird 1995) and it is based on the property that for any policy π the value function V^π has a zero residual, i.e., $\mathcal{B}(V^\pi) = 0$. As a result, it is reasonable to expect that the minimization of the Bellman residual $\mathcal{B}(f)$ in a given function space $\mathcal{F}$ leads to a function which is close to the value function. Williams and Baird (1994) and Munos (2007) proved that indeed the residual $||\mathcal{T}^\pi f - f||$ (in sup-norm and L_p-norms, respectively) of a function f is related to its distance (in the same norm) to the value function V^π, $||V^\pi - f||$. Thus, minimizing the Bellman residual leads to a small approximation error. However, those results concern the (true) Bellman residual $\mathcal{B}(f)$ but not its empirical estimate $\mathcal{B}_n(f)$, which is the quantity that is actually minimized by real algorithms.

Although it is often believed that the minimization of the empirical residual $\mathcal{B}_n(f)$ is "approximately" equivalent to minimizing the (true) residual $B(f)$, no such result is available in the literature so far. The closest work in this direction is by Antos et al. (2008), who provides a finite-sample analysis of a variant of the Bellman-residual minimization, called Modified Bellman residual, which reduces to Least Squares Temporal Differences (LSTD) in the case of linear function spaces. A finite sample analysis of LSTD is also reported in Lazaric et al. (2010), and a regularized version of those algorithms is described in Farahmand et al. (2008). However, these works analyze algorithms that are related but different from the empirical Bellman residual minimization considered here.

Contribution. Our main contribution in this paper is to address this question: does minimizing the empirical Bellman residual $\mathcal{B}_n$ implies that we also minimize the true Bellman residual at all states w.r.t. a distribution μ? In other terms, is it possible to control the true Bellman residual $\mathcal{B}(f)$ in terms of the empirical Bellman residual $\mathcal{B}_n(f)$?

We show that the answer to those questions is actually not obvious but is positive. It is not obvious because we show that the usual generalization results for regression cannot be trivially adopted in bounding the difference between the true Bellman residual and its empirical counterpart. In fact, in Bellman residual minimization we are not trying to minimize an empirical distance to a given target function, but we are directly searching for an approximate fixed-point (in $\mathcal{F}$) of an empirical version of the Bellman operator $\mathcal{T}^\pi$. As a result, it might be possible that a function with very low empirical residual (even possibly zero) at the sampled states has a large (true) Bellman residual at other states and even at the same states. However, we show that this problem does not occurs when the empirical Bellman residual minimizer belongs to a set of controlled sized (e.g. measured in terms of the norm of its parameter). More precisely, we show that for functions $f_\alpha \in \mathcal{F}$ with bounded parameter $||\alpha||$, the difference between $\mathcal{B}(f_\alpha)$ and $\mathcal{B}_n(f_\alpha)$ decreases as the number of samples n increases. Then, we prove that when the number of samples n is large enough, the norm $||\hat{\alpha}||$ of the empirical Bellman residual minimizer $f_{\hat{\alpha}} = \arg\min_{f_\alpha \in \mathcal{F}} \mathcal{B}_n(f)$

is indeed upper-bounded, provided that the set of features defining the linear space $\mathcal{F}$ are linearly independent under the distribution μ. Thus we deduce that the Bellman residual $\mathcal{B}(f_{\hat{\alpha}})$ of the empirical Bellman minimizer $f_{\hat{\alpha}}$ is bounded by the empirical Bellman residual $\mathcal{B}_n(f_{\hat{\alpha}})$ plus an estimation error term of order $O(1/\sqrt{n})$. In other terms, we provide a generalization result for the Bellman residual in linear approximation spaces. This result implies that minimizing the empirical residual is indeed a sound approach for deriving a good approximation of the value function for each policy.

The paper is organized as follows. In Section 2 we introduce the notation. Section 3 reports the main contribution of this paper, that is the finite-sample analysis of Bellman residual minimization for policy evaluation. Finally, in Section 4 we extend the policy evaluation result to the whole policy iteration algorithm.

2. Preliminaries

In this section, we introduce the main notations used in the paper. For a measurable space with domain $\mathcal{X}$, we let $\mathcal{S}(\mathcal{X})$ and $\mathcal{B}(\mathcal{X}; L)$ denote the set of probability measures over $\mathcal{X}$ and the space of bounded measurable functions with domain $\mathcal{X}$ and bound $0 < L < \infty$, respectively. For a measure $\mu \in \mathcal{S}(\mathcal{X})$ and a measurable function $f : \mathcal{X} \to \mathbb{R}$, we define the $\ell_2(\mu)$-norm of f as $||f||_\mu^2 = \int f(x)^2 \mu(dx)$, the supremum norm of f as $||f||_\infty = \sup_{x \in \mathcal{X}} |f(x)|$. Moreover, for a vector $u \in \mathbb{R}^d$, we write its ℓ_2-norm as $||u||^2 = \sum_{i=1}^d u_i^2$.

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

Value functions. The value function of a policy π, V^π, is the unique fixed-point of the Bellman operator $\mathcal{T}^\pi : \mathcal{B}(\mathcal{X}; V_{\max} = \frac{R_{\max}}{1-\gamma}) \to \mathcal{B}(\mathcal{X}; V_{\max})$ defined by

$$(\mathcal{T}^\pi V)(x) = R^\pi(x) + \gamma \int_{\mathcal{X}} P^\pi(dy|x) V(y). \tag{1}$$

We also define the optimal value function V^* as the unique fixed-point of the optimal Bellman operator $\mathcal{T}^* : \mathcal{B}(\mathcal{X}; V_{\max} = \frac{R_{\max}}{1-\gamma}) \to \mathcal{B}(\mathcal{X}; V_{\max})$ defined by

$$(\mathcal{T}^* V)(x) = \max_{a \in \mathcal{A}} \left[r(x, a) + \gamma \int_{\mathcal{X}} p(dy|x, a) V(y) \right]. \tag{2}$$

Approximation space. We consider a linear function space $\mathcal{F}$ defined as the span of d basis functions $\varphi_i : \mathcal{X} \mapsto \mathbb{R}$, $i = 1, \ldots, d$, i.e.,

$$\mathcal{F} = \{ f_\alpha(\cdot) = \phi(\cdot)^\top \alpha, \alpha \in \mathbb{R}^d \},$$

where $\phi(\cdot) = \big(\varphi_1(\cdot), \dots, \varphi_d(\cdot)\big)^\top$ is the feature vector. We define the Gram matrix $G \in \mathbb{R}^{d\times d}$ with respect to a distribution $\mu \in \mathcal{S}(\mathcal{X})$ as

$$G_{ij} = \int_{\mathcal{X}} \varphi_i(x)\varphi_j(x)\mu(dx), \tag{3}$$

with $i, j = 1, \dots, d$. Finally, we write $L_{\max} = \sup_{x\in\mathcal{X}} ||\phi(x)||$ and assume that $L_{\max} < \infty$.

3. Bellman Residual Minimization for Policy Evaluation

In this section, we consider the Bellman Residual Minimization (BRM) algorithm for the evaluation of a fixed policy π, using the double sampling technique (see e.g., Sutton and Barto 1998). We assume that a generative model of the MDP is available, and that for each state x and action a a call to the generative model returns the reward $r(x, a)$ and two independent samples drawn from the distribution $P(\cdot|x, a)$.

3.1 The Empirical Bellman Residual Solution

We build a dataset $\mathcal{D} = \{(X_i, R_i, Y_i, Y_i')_{1\le i\le n}\}$ where for all $i = 1, \dots, n$, we sample a state $X_i \overset{iid}{\sim} \mu$ and make a call to the generative model to obtain the reward $R_i = r(X_i, \pi(X_i))$ and two independent next-state samples Y_i and Y_i' drawn from $P^\pi(\cdot|X_i)$. The **empirical Bellman residual** (EBR) is defined for any $f \in \mathcal{F}$ as

$$\mathcal{B}_n(f) = \frac{1}{n}\sum_{i=1}^{n} \big[f(X_i) - \gamma f(Y_i) - R_i\big]\big[f(X_i) - \gamma f(Y_i') - R_i\big]. \tag{4}$$

The EBR minimizer $f_{\hat\alpha}$ is defined, whenever it exists, as the minimizer of $\mathcal{B}_n(f_\alpha)$ in $\mathcal{F}$:

$$f_{\hat\alpha} = \arg\min_{f_\alpha\in\mathcal{F}} \mathcal{B}_n(f_\alpha), \tag{5}$$

and $\hat\alpha$ is the parameter of the EBR minimizer. Using matrix notations, by defining the $n \times d$-matrices Ψ and Ψ' as $\Psi_{ij} = \varphi_j(X_i) - \gamma\varphi_j(Y_i)$ and $\Psi'_{ij} = \varphi_j(X_i) - \gamma\varphi_j(Y_i')$, $\mathcal{B}_n(f_\alpha)$ may be written as

$$\mathcal{B}_n(f_\alpha) = \frac{1}{n}\Big[\alpha^\top\Psi^\top\Psi'\alpha - R^\top(\Psi + \Psi')\alpha + R^\top R\Big],$$

where $R \in \mathbb{R}^n$ is the vector of components R_i. Thus, by defining the $d \times d$ empirical Gram matrix $A = \frac{1}{n}(\Psi^\top\Psi' + \Psi'^\top\Psi)$, the d-vector $b = \frac{1}{n}(\Psi + \Psi')^\top R$, and the constant $c = \frac{1}{n}R^\top R$, we have

$$\mathcal{B}_n(f_\alpha) = \frac{1}{2}\alpha^\top A\alpha - b^\top\alpha + c. \tag{6}$$

Using this notation, the gradient of $\mathcal{B}_n$ is $\nabla_\alpha\mathcal{B}_n(f_\alpha) = A\alpha - b$, thus whenever the EBR minimizer exists, its parameter $\hat\alpha$ is the solution to the linear system $A\alpha = b$.

Although the empirical Bellman residual $\mathcal{B}_n(f_\alpha)$ is a quadratic function of α, with A a symmetric matrix, A may not be definite positive. A may even possess negative eigenvalues, thus $\mathcal{B}_n(f_\alpha)$ may not have any minimizer. However we will see in the next section that when n is large enough then the EBR minimizer exists and is unique.

3.2 Finite-Sample Analysis

Defining $\mathcal{B}(f) = ||f - \mathcal{T}^\pi f||_\mu^2$ the true squared Bellman residual in μ-norm, we have the property that for any f, $\mathcal{B}_n(f)$ is an unbiased estimate of $\mathcal{B}(f)$. In fact,

$$\mathbb{E}_{Y_i, Y_i' \overset{iid}{\sim} P^\pi(\cdot|X_i)} \Big[\left[f(X_i) - \gamma f(Y_i) - R_i \right] \left[f(X_i) - \gamma f(Y_i') - R_i \right] | X_i \Big] = \left[f(X_i) - \mathcal{T}^\pi f(X_i) \right]^2,$$

thus, since $X_i \overset{iid}{\sim} \mu$, it follows that $\mathbb{E}_\mathcal{D}[\mathcal{B}_n(f)] = \mathcal{B}(f)$.

The main issue is to show that by minimizing the empirical Bellman residual $\mathcal{B}_n$, we actually obtain a function $f_{\hat{\alpha}}$ whose (true) residual $f_{\hat{\alpha}} - \mathcal{T}^\pi f_{\hat{\alpha}}$ is small at the states $(X_1, \ldots, X_n)$ and at other states measured by μ (i.e., it has a small $\mathcal{B}$). This property would hold if we could have a generalization result for the Bellman residual, like in the regression setting.

In regression, generalization bounds for spaces bounded in sup-norm are applied to the result of the truncation (at a threshold which depends on a sup-norm of the target function) of the empirical risk minimizer (Györfi et al., 2002). However, this approach does not work for BRM, because the truncation $\bar{f}_{\hat{\alpha}}$ of the EBR minimizer $f_{\hat{\alpha}}$ may amplify the residual (i.e., $\mathcal{B}(\bar{f}_{\hat{\alpha}})$ may not be smaller than $\mathcal{B}(f_{\hat{\alpha}})$). Thus, we follow another direction by considering spaces of functions $\mathcal{F}(C) \subset \mathcal{F}$ with bounded parameter: $\mathcal{F}(C) = \{f_\alpha \in \mathcal{F}, ||\alpha|| \leq C\}$, and provide a generalization bound for Bellman residual for functions $f_\alpha \in \mathcal{F}(C)$ (the proof is in Appendix).

Lemma 1 *For any $\delta > 0$, we have that with probability at least $1 - \delta$,*

$$\sup_{f_\alpha \in \mathcal{F}(C)} |\mathcal{B}(f_\alpha) - \mathcal{B}_n(f_\alpha)| \leq c_1 \sqrt{\frac{2d \log(2) + 6 \log(8/\delta)}{n}},$$

where $c_1 = 96\sqrt{2}[C(1+\gamma)L_{\max} + R_{\max}]^2$.

Unfortunately, this result cannot be immediately applied to the EBR minimizer $f_{\hat{\alpha}}$ since we do not have a bound on the norm $||\hat{\alpha}||$. In fact, when we solve the minimization problem (5), we do not have any control on the norm of the solution (if it exists) $||\hat{\alpha}||$. For instance, if we consider the case in which two features φ_1 and φ_2 are identical, then $\alpha_1\varphi_1 + \alpha_2\varphi_2 = 0$ whenever $\alpha_1 = -\alpha_2$, thus $||\alpha||$ can be made arbitrarily large without changing the value of f_α simply by playing on the values of α_1 and α_2. In order to avoid such degenerate situations, we introduce the following assumption on the linear independence of the features $(\varphi_i)_{1 \leq i \leq d}$ w.r.t. the distribution μ.

Assumption 1 *The smallest eigenvalue ν of the Gram matrix G (defined in (3)) is strictly positive, i.e., $\nu > 0$.* [1]

We show in the following that Assumption 1 is a sufficient condition to derive a bound on the norm $||\hat{\alpha}||$ for any $\hat{\alpha}$ solution of the EBR minimization problem. Before moving to the analysis of the EBR minimizer with linear independent features, we first introduce some additional notation. Let $\mathcal{L}(f) = ||(I - \gamma P^\pi)f||_\mu^2$ be the quadratic part of $\mathcal{B}(f)$, and

$$\mathcal{L}_n(f) = \frac{1}{n} \sum_{i=1}^n \left[f(X_i) - \gamma f(Y_i) \right] \left[f(X_i) - \gamma f(Y_i') \right],$$

1. Note that this condition implies the linear independence of the features in μ-norm.

be its empirical version. Thus $\mathcal{L}_n(f_\alpha) = \frac{1}{2}\alpha^\top A\alpha$. Now, whenever the EBR minimizer $f_{\hat{\alpha}}$ exists, since by definition $\hat{\alpha}$ satisfies $A\hat{\alpha} = b$, we can write

$$\mathcal{B}_n(f_\alpha) = \frac{1}{2}(\alpha - \hat{\alpha})^\top A(\alpha - \hat{\alpha}) - \frac{1}{2}\hat{\alpha}^\top A\hat{\alpha} + c = \mathcal{L}_n(f_{\alpha-\hat{\alpha}}) - \mathcal{L}_n(f_{\hat{\alpha}}) + c, \tag{7}$$

and deduce that $\mathcal{B}_n(f_{\hat{\alpha}}) = c - \mathcal{L}_n(f_{\hat{\alpha}}) = c - \frac{1}{2}b^\top\hat{\alpha}$.

Bounding $||\hat{\alpha}||$. In order to deduce a bound on the parameter of the EBR minimizer $\hat{\alpha}$, in the next three lemmas, we relate $||\alpha||$ to respectively $\mathcal{L}(f_\alpha)$ and $\mathcal{L}_n(f_\alpha)$. For that purpose, let us write

$$C^\pi(\mu) = (1-\gamma)||(I - \gamma P^\pi)^{-1}||_\mu,$$

which is related to the concentrability coefficient (see e.g., Antos et al. 2008) of the discounted future state distribution starting from μ and following policy π, i.e., $(1-\gamma)\mu(I - \gamma^\pi)^{-1}$ w.r.t. μ. Note that if the discounted future state distribution is not absolutely continuous w.r.t. μ, then $C^\pi(\mu) = \infty$.

Lemma 2 *Under Assumption 1, for any $\alpha \in \mathbb{R}^d$*

$$||\alpha||^2 \leq \frac{1}{\nu}||f_\alpha||_\mu^2 \leq \frac{C^\pi(\mu)^2}{\nu(1-\gamma)^2}\mathcal{L}(f_\alpha).$$

This indicates that the eigenvalues of the Gram matrix $\widetilde{G}$ defined by $\widetilde{G}_{ij} = \int_{\mathcal{X}} \psi_i\psi_j d\mu$, where $\psi_i = (I - \gamma P^\pi)\varphi_i$, are lower bounded by $\xi = \frac{\nu(1-\gamma)^2}{C^\pi(\mu)^2}$.

Proof From the definition that ν is the smallest eigenvalue of G, we have $\alpha^\top\alpha \leq \frac{1}{\nu}\alpha^\top G\alpha = \frac{1}{\nu}||f_\alpha||_\mu^2$. Now since $(I - \gamma P^\pi)$ is an invertible operator (the eigenvalues of any stochastic kernel P^π have a modulus less than 1), we have $||f_\alpha||_\mu^2 \leq ||(I-\gamma P^\pi)^{-1}||_\mu^2||(I-\gamma P^\pi)f_\alpha||_\mu^2 = \left(\frac{C^\pi(\mu)}{1-\gamma}\right)^2\mathcal{L}(f_\alpha)$, and the lemma follows. ■

This lemma provides a bound on $||\hat{\alpha}||$ in terms of $\mathcal{L}(f_{\hat{\alpha}})$. However $\mathcal{L}(f_{\hat{\alpha}})$ is not known, and we would like to relate it to its empirical counterpart $\mathcal{L}_n(\hat{\alpha})$. The next lemma (the proof is in the Appendix) provides a generalization bound for $\mathcal{L}$, which enables to bound the difference between $\mathcal{L}$ and $\mathcal{L}_n$.

Lemma 3 *For any $\delta > 0$, we have that with probability at least $1 - \delta$,*

$$\forall\alpha \in \mathbb{R}^d, \quad |\mathcal{L}(f_\alpha) - \mathcal{L}_n(f_\alpha)| \leq c_2||\alpha||^2\sqrt{\frac{2d\log(2) + \log(4/\delta)}{n}},$$

where $c_2 = 96\sqrt{2}(1+\gamma)^2L_{\max}^2$.

Combining Lemmas 2 and 3 we deduce that when n is large enough (as a function of ν and $C^\pi(\mu)$), then all the eigenvalues of the empirical Gram matrix A are strictly positive, and thus the EBR minimizer exists and is unique.

Lemma 4 *For any $\delta > 0$, whenever $n \geq n^\pi(\nu, \delta) = \frac{4c_2^2 C^\pi(\mu)^4}{\nu^2(1-\gamma)^4}(2d\log 2 + \log 4/\delta)$, with probability $1-\delta$ we have for all $\alpha \in \mathbb{R}^d$, $||\alpha||^2 \leq \frac{2}{\xi}\mathcal{L}_n(f_\alpha)$.*

We deduce that all the eigenvalues of the empirical Gram matrix A are strictly positive, and thus the EBR minimizer exists and is unique.

Proof From Lemmas 2 and 3,

$$||\alpha||^2 \leq \frac{1}{\xi}\mathcal{L}(f_\alpha) \leq \frac{1}{\xi}\big(\mathcal{L}_n(f_\alpha) + c_2||\alpha||^2\sqrt{\frac{2d\log(2)+\log(4/\delta)}{n}}\big),$$

thus whenever $c_2\sqrt{\frac{2d\log(2)+\log(4/\delta)}{n}} \leq \frac{\xi}{2}$, i.e., $n \geq n^\pi(\nu,\delta)$, we have $||\alpha||^2 \leq \frac{2}{\xi}\mathcal{L}_n(f_\alpha)$. The claim about the eigenvalues of the empirical Gram matrix simply follows from the statement of the Lemma, the inequality $\alpha^\top\alpha \leq \frac{1}{\chi}\alpha^\top A\alpha$, where χ is the smallest eigenvalue of A, and the definition of $\mathcal{L}_n(f_\alpha) = \frac{1}{2}\alpha^\top A\alpha$. ■

From this result we immediately deduce a bound on $||\hat{\alpha}||$.

Corollary 5 *For any $\delta > 0$, whenever $n \geq n^\pi(\nu,\delta)$, with probability $1-\delta$ we have*

$$||\hat{\alpha}|| \leq \frac{2}{\xi}(1+\gamma)L_{\max}R_{\max}.$$

Proof From Lemma 4, using Cauchy-Schwarz's inequality, and recalling the definition of Ψ in Section 3.1

$$\begin{aligned}
||\hat{\alpha}||^2 &\leq \frac{2}{\xi}\frac{1}{2}b^\top\hat{\alpha} = \frac{1}{\xi}\sum_{j=1}^d\left(\frac{1}{n}\sum_{i=1}^n R_i(\Psi_{i,j}+\Psi'_{i,j})\hat{\alpha}_j\right)\\
&\leq \frac{1}{\xi}\frac{1}{n}\sum_{i=1}^n R_{\max}\Big(2\Big|\sum_{j=1}^d\hat{\alpha}_j\phi_j(X_i)\Big| + \gamma\Big|\sum_{j=1}^d\hat{\alpha}_j\phi_j(Y_i)\Big| + \gamma\Big|\sum_{j=1}^d\hat{\alpha}_j\phi_j(Y'_i)\Big|\Big)\\
&\leq \frac{2}{\xi}R_{\max}||\hat{\alpha}||\sup_x||\phi(x)||(1+\gamma)
\end{aligned}$$

from which the result follows. ■

We now state our main result which bounds the Bellman residual of the EBR minimizer.

Theorem 6 *For any $\delta > 0$, whenever $n \geq n^\pi(\nu,\delta/2)$, with probability $1-\delta$ we have*

$$\mathcal{B}(f_{\hat{\alpha}}) \leq \mathcal{B}_n(f_{\hat{\alpha}}) + c_3\sqrt{\frac{2d\log(2)+6\log(8/\delta)}{n}},$$

where $c_3 = 96\sqrt{2}[\frac{2}{\xi}(1+\gamma)^2L_{\max}^2+1]^2R_{\max}^2$.

Proof When $n \geq n^\pi(\nu,\delta)$, Corollary 5 states that $||\hat{\alpha}|| \leq C$ is bounded and the results follows from a direct consequence of Lemma 1. ■

Thus, the true residual $\mathcal{B}(f_{\hat{\alpha}})$ of the EBR minimizer $f_{\hat{\alpha}}$ is upper-bounded by the empirical residual $\mathcal{B}_n(f_{\hat{\alpha}})$ plus an estimation error term, which is of order $O(1/\sqrt{n})$. We deduce that minimizing the empirical residual is indeed a sound method for deriving a function with small (true) Bellman residual $\mathcal{B}$.

Remark 1 The obtained estimation error term is of order $O(1/\sqrt{n})$, which is worse than the estimation error of order $O(\log n/n)$ deduced in linear regression with a quadratic loss (see e.g., Györfi et al. 2002). This is due to the fact that although $\mathcal{B}(f)$ is positive for any f, this is not the case for $\mathcal{B}_n(f)$, which may be negative (e.g., think of $\mathcal{B}_n(V^\pi)$ which is an unbiased estimate of $\mathcal{B}(V^\pi) = 0$). Thus the usual argument described in Györfi et al. (2002), where one would derive $\sqrt{\mathcal{B}(f)} \leq 2\sqrt{\mathcal{B}_n(f)} + O(1/\sqrt{n})$ does not directly apply here. One could also think of applying this argument to $\mathcal{L}_n$, since $\mathcal{L}_n$ is positive for sufficiently large n. However, this does not work either, since $\mathcal{L}_n$ is the sum of terms which are not individually positive, independently of the value of n. Therefore, it remains an open question to whether it is possible to obtain a bound of the form $\mathcal{B}(f_{\hat{\alpha}}) \leq c\mathcal{B}_n(f_{\hat{\alpha}}) + O(\log n/n)$ (with an additional multiplicative factor $c > 1$). This could be particularly interesting when $\mathcal{B}_n(f_{\hat{\alpha}})$ is small.

Remark 2 The dependence to the dimension d of the function space $\mathcal{F}$ is of order $L_{\max}^4\sqrt{d}$. This is due to the fact that we cannot use truncation in this Bellman residual setting (see the first paragraph of Section 3.2), which would give us an order $L_{\max}^2\sqrt{d}$. We use instead a covering of the function space $\mathcal{F}(C)$ (see Theorem 7) with C (which itself depends on $L_{\max}$) being a bound on $||\hat{\alpha}||$. This explains the additional $L_{\max}^2$ factor.

Remark 3 It is interesting to notice that although we derived Corollary 5 specifically for the case of Bellman residual minimization, a similar result can be obtained in the traditional regression setting. The bound on the norm of $\hat{\alpha}$ solution of the least-squares problem may be used to derive an excess risk bound for the empirical risk minimizer in an unbounded space without truncation, at the price of a weaker dependence on $L_{\max}$, as discussed in Remark 2.

3.3 Bellman Residual Minimization and Approximation of V^π

We are now interested to relate the Bellman residual of $f_{\hat{\alpha}}$ to the minimum Bellman residual in $\mathcal{F}$, i.e., $\inf_{f\in\mathcal{F}} \mathcal{B}(f)$, and to the approximation error (in μ-norm) of the value function V^π w.r.t. the function space $\mathcal{F}$, i.e., $\inf_{f\in\mathcal{F}} ||V^\pi - f||_\mu$. In fact, these two quantities are related since for any function $f \in \mathcal{F}$, we have $\mathcal{T}^\pi f - f = (I - \gamma P^\pi)(V^\pi - f)$. Thus, by defining

$$f_{\tilde{\alpha}} = \arg\min_{f\in\mathcal{F}} \mathcal{B}(f), \quad \text{and} \quad f_{\alpha^*} = \arg\min_{f\in\mathcal{F}} ||V^\pi - f||_\mu,$$

we have

$$||V^\pi - f_{\alpha^*}||_\mu \leq ||V^\pi - f_{\tilde{\alpha}}||_\mu \leq \frac{C^\pi(\mu)}{1-\gamma}\sqrt{\mathcal{B}(f_{\tilde{\alpha}})} \leq \frac{C^\pi(\mu)}{1-\gamma}\sqrt{\mathcal{B}(f_{\alpha^*})}. \tag{8}$$

We can now relate both the Bellman residual of $f_{\hat{\alpha}}$, $\mathcal{B}(f_{\hat{\alpha}})$, and its approximation error, $||V^\pi - f_{\hat{\alpha}}||_\mu$, to the minimum possible Bellman residual in $\mathcal{F}$ and the distance between V^π and $\mathcal{F}$.

Theorem 7 *For any $\delta > 0$, whenever $n \geq n^\pi(\nu, \delta/2)$, with probability $1-\delta$, the Bellman residual of the EBR minimizer $f_{\hat{\alpha}}$ is bounded as*

$$\mathcal{B}(f_{\hat{\alpha}}) \;\; \leq \inf_{f \in \mathcal{F}} \mathcal{B}(f) + c_4 \sqrt{\frac{2d\log(2) + 6\log(16/\delta)}{n}},$$

with $c_4 = (96\sqrt{2}+1)[\frac{2}{\xi}(1+\gamma)^2 L_{\max}^2 + 1]^2 R_{\max}^2$, and the approximation error of V^π is bounded as $||V^\pi - f_{\hat{\alpha}}||_\mu^2 \leq \left(\frac{C^\pi(\mu)}{1-\gamma}\right)^2 \mathcal{B}(f_{\hat{\alpha}})$. Moreover, since $\inf_{f\in\mathcal{F}} \mathcal{B}(f) \leq (1+\gamma||P^\pi||_\mu)^2 \inf_{f\in\mathcal{F}} ||V^\pi - f||_\mu^2$, we obtain an alternative bound

$$||V^\pi - f_{\hat{\alpha}}||_\mu^2 \;\; \leq \left(\frac{C^\pi(\mu)}{1-\gamma}\right)^2 \left((1+\gamma||P^\pi||_\mu)^2 \inf_{f\in\mathcal{F}} ||V^\pi - f||_\mu^2 + c_4 \sqrt{\frac{2d\log(2) + 6\log(16/\delta)}{n}} \right).$$

Proof From the definition of $\tilde{\alpha}$ (the minimum of $\mathcal{B}$), we have $\mathcal{L}(f_{\tilde{\alpha}}) = 2\langle R^\pi, (I - \gamma P^\pi)\phi^\top \tilde{\alpha}\rangle_\mu$. Thus, from Lemma 2, we obtain

$$||\tilde{\alpha}||^2 \leq \frac{1}{\xi}\mathcal{L}(f_{\tilde{\alpha}}) \leq \frac{2}{\xi}(1+\gamma) L_{\max} R_{\max} ||\tilde{\alpha}||, \text{ thus } ||\tilde{\alpha}|| \leq \frac{2}{\xi}(1+\gamma) L_{\max} R_{\max}.$$

Now using Chernoff Hoeffding's inequality, we have with probability $1-\delta/2$,

$$\mathcal{B}_n(f_{\tilde{\alpha}}) \leq \mathcal{B}(f_{\tilde{\alpha}}) + \left[\frac{2}{\xi}(1+\gamma)^2 L_{\max}^2 + 1\right]^2 R_{\max}^2 \sqrt{\frac{2\log(2/\delta)}{n}}\,. \tag{9}$$

We may write

$$\mathcal{B}(f_{\hat{\alpha}}) \leq (\mathcal{B}(f_{\hat{\alpha}}) - \mathcal{B}_n(f_{\hat{\alpha}})) + \mathcal{B}_n(f_{\tilde{\alpha}}) \leq \inf_{f\in\mathcal{F}} \mathcal{B}(f) + (\mathcal{B}(f_{\hat{\alpha}}) - \mathcal{B}_n(f_{\hat{\alpha}})) + (\mathcal{B}_n(f_{\tilde{\alpha}}) - \mathcal{B}(f_{\tilde{\alpha}})).$$

The claim follows by applying Theorem 6 (with probability $\delta/2$) and (9) for the second and third terms on the right hand, respectively, and a union bound so that both events hold simultaneously with probability at least $1-\delta$. The other inequalities are deduced from the definition of $\tilde{\alpha}$ and α^* and (8). ■

This result means that whenever the space $\mathcal{F}$ is such that it contains a function with a small Bellman residual or that it can well approximate V^π, then the residual of the EBR minimizer $f_{\hat{\alpha}}$ is small. In addition, assuming that $C^\pi(\mu)$ is small, $f_{\hat{\alpha}}$ is also a good approximation of the value function V^π.

4. Bellman Residual Minimization for Policy Iteration

We now move to the full analysis of the policy iteration algorithm where at each iteration k, the policy π_k is approximated by the solution of an empirical Bellman residual minimization. The Bellman Residual Minimization Policy Iteration (BRM-PI) algorithm is described in Figure 1. At each iteration k, BRM-PI generates a new dataset $\mathcal{D}_k = \{(X_i^{(k)}, R_i^{(k)}, Y_i^{(k)}, Y_i^{'(k)})\}_{i=1}^n$ where $X_i^{(k)} \overset{\text{iid}}{\sim} \mu$, $R_i^{(k)} = r(X_i^{(k)}, \pi_k(X_i^{(k)}))$, and $Y_i^{(k)}$

Input: Function space $\mathcal{F}$, state distribution μ, number of samples n, number of iterations K
Initialize: Let $V_0 \in \mathcal{B}(\mathcal{X}; V_{\max})$ be an arbitrary value function
for $k = 1, 2, \ldots, K$ **do**
(1) Let π_k be the greedy policy w.r.t. V_{k-1} (see Eq. 13).
(2) Build a new dataset $\mathcal{D}_k = \{(X_i^{(k)}, R_i^{(k)}, Y_i^{(k)}, Y_i^{'(k)})\}_{i=1}^n$, where $X_i^{(k)} \overset{\text{iid}}{\sim} \mu$, $R_i^{(k)} = r(X_i^{(k)}, \pi_k(X_i^{(k)}))$, and use the generative model to draw two independent samples $Y_i^{(k)}$ and $Y_i^{'(k)}$ from $P^{\pi_k}(\cdot|X_i^{(k)})$.
(3) Let $\hat{\alpha}_k$ be the solution to the linear system $A_k \alpha = b_k$, where A_k and b_k are defined by (10) and (11).
(4) Let $V_k = f_{\hat{\alpha}_k}$.
end for
Return policy π_K.

Figure 1: The Bellman Residual Minimization Policy Iteration (BRM-PI) algorithm.

and $Y_i^{'(k)}$ are two independent samples drawn from $P^{\pi_k}(\cdot|X_i^{(k)})$. The $d \times d$-matrix A_k and d-vector b_k are defined as

$$A_k = \frac{1}{n}(\Psi_k^\top \Psi'_k + \Psi_k^{'\top} \Psi_k) \tag{10}$$

$$b_k = \frac{1}{n}(\Psi_k + \Psi'_k)^\top R^{(k)} \tag{11}$$

where $(\Psi_k)_{ij} = \varphi_j(X_i^{(k)}) - \gamma \varphi_j(Y_i^{(k)})$ and $(\Psi'_k)_{ij} = \varphi_j(X_i^{(k)}) - \gamma \varphi_j(Y_i^{'(k)})$. Then $\hat{\alpha}_k$ is defined as the solution of

$$A_k \alpha = b_k \tag{12}$$

(the next theorem will provide conditions under which this system has a solution), which defines the approximation $V_k = f_{\hat{\alpha}_k}$ of the current value function V^{π_k}. Finally, the approximation V_k is used to generate the policy π_{k+1} for the next iteration $k+1$

$$\pi_{k+1}(x) = \arg\max_{a \in \mathcal{A}} \Big[r(x,a) + \gamma \int_{\mathcal{X}} P(dy|x,a) V_k(y) \Big]. \tag{13}$$

Note that in order to compute the expectation we can use the generative model and replace the expectation by an average over a sufficiently large number of samples. However this is not convenient and a usual technique used to avoid computing the expectations for deriving the greedy policy is to use action-value functions Q instead of value functions V (see e.g., Watkins 1989; Lagoudakis and Parr 2003; Antos et al. 2008), or functions defined over post-decision states (Powell, 2007). We do not further develop this point here but we simply mention that all the finite-sample analysis derived in the previous section for the setting of value functions can be easily extended to action-value functions.

Now following the analysis of Munos (2003) and Antos et al. (2008), we relate the performance of the policy π_K returned by the algorithm to the optimal policy $||V^* - V^{\pi_K}||_\rho$ (where ρ is a distribution chosen by the user), in terms of the Bellman residuals of the EBR minimizers $f_{\hat{\alpha}_k}$ at all the iterations $k < K$ of the BRM-PI algorithm. In order to do so, we make use of the concentrability coefficients, $C_{\rho,\mu}$, defined for any couple of distributions ρ

and μ in Antos et al. (2008) and Munos and Szepesvári (2008) (A refined analysis can be found in Farahmand et al. (2010)).

Let us also define $n(\delta) = \sup_\pi n^\pi(\nu^\pi, \delta)$ and write $\mathcal{B}^\pi(f) = ||f - \mathcal{T}^\pi f||_\mu^2$ the Bellman residual of f under policy π. We can now state the main result which provides a performance bound for BRM-PI.

Theorem 8 *For any $\delta > 0$, whenever $n \geq n(\delta/K)$, with probability $1-\delta$, the EBR minimizer $f_{\hat{\alpha}_k}$, where $\hat{\alpha}_k$ is the solution of the linear system (12), exists for all iterations $1 \leq k < K$, thus the BRM-PI algorithm is well defined, and the performance V^{π_K} of the policy π_K returned by the algorithm is such that*

$$||V^* - V^{\pi_K}||_\rho^2 \leq \Big(\frac{2\gamma}{(1-\gamma)^2}\Big)^2 \left[C_{\rho,\mu} \sup_{1\leq k<K} \left(\inf_{f\in\mathcal{F}} \mathcal{B}^{\pi_k}(f) + c_k \sqrt{\frac{2d\log(2) + 6\log(16K/\delta)}{n}} \right) + \gamma^K R_{\max}^2 \right],$$

where $c_k = (96\sqrt{2}+1)[\frac{2}{\xi_k}(1+\gamma)^2 L_{\max}^2 + 1]^2 R_{\max}^2$, with ξ_k defined similarly as ξ in Lemma 2 for the policy π_k. A bound using the distances between the sequence of value functions and $\mathcal{F}$ can be obtained using the fact that $\inf_{f\in\mathcal{F}} \mathcal{B}^{\pi_k}(f) \leq (1+\gamma||P^{\pi_k}||_\mu)^2 \inf_{f\in\mathcal{F}} ||V^{\pi_k} - f||_\mu^2$.

Proof From Antos et al. (2008, Lemma 12) we have

$$||V^* - V^{\pi_K}||_\rho^2 \quad \leq \quad \big(\frac{2\gamma}{(1-\gamma)^2}\big)^2 \big(C_{\rho,\mu} \max_{0\leq k<K} \mathcal{B}^{\pi_k}(f_{\hat{\alpha}_k}) + \gamma^K R_{\max}^2\big). \tag{14}$$

Now from Lemma 4, we have that at each step $k < K$, whenever $n \geq n(\delta/K) \geq n^{\pi_k}(\nu^{\pi_k}, \delta/K)$, with probability $1 - \delta/K$, the EBR minimizer $f_{\hat{\alpha}_k}$ exists and from Theorem 7, the Bellman residual of $f_{\hat{\alpha}_k}$ is bounded as

$$\mathcal{B}^{\pi_k}(f_{\hat{\alpha}_k}) \quad \leq \inf_{f\in\mathcal{F}} \mathcal{B}^{\pi_k}(f) + c_k \sqrt{\frac{2d\log(2) + 6\log(16K/\delta)}{n}},$$

where we used a union bound that guarantees that these bounds hold for all K iterations. ■

The performance bounds reported in Theorem 8 are composed of the sum of three terms. The first term is an approximation error term, which indicates how well the function space $\mathcal{F}$ is adapted to the problem, either in terms of containing functions with low Bellman residuals (for the sequence of policies) $\inf_{f\in\mathcal{F}} \mathcal{B}^{\pi_k}(f)$, or in terms of well approximating the corresponding value functions $\inf_{f\in\mathcal{F}} ||V^{\pi_k} - f||_\mu$. The second term is an estimation error term, which decreases as $O(1/\sqrt{n})$, and the third term is decreasing exponentially fast with K, the number of policy iterations.

Remark: In the current description of the BRM-PI algorithm, we regenerate a new dataset $\mathcal{D}_k$ at each policy evaluation step. However, we could generate once for all n samples $(X_1, \ldots, X_n)$ and all actions $a \in \mathcal{A}$, the corresponding rewards $R_i(a) = r(X_i, a)$ and $2n$ independent next states $Y_i(a)$ and $Y_i'(a)$ sampled from $P(\cdot|X_i, a)$. Then at each iteration k, we use these samples and build the dataset $\mathcal{D}_k = \big\{\big(X_i, R_i(\pi_k(X_i)), Y_i(\pi_k(X_i)), Y_i'(\pi_k(X_i))\big)\big\}_{i=1}^n$. This sampling strategy requires generating $2n \times |\mathcal{A}|$ samples instead of $2n \times K$ for the previous method, which is advantageous when $|\mathcal{A}| \leq K$. In terms of performance, this version

attains a similar performance as in Theorem 8. The main difference is that at each iteration k, the target function V^{π_k} depends on the samples because the policy π_k is greedy w.r.t. the function $f_{\alpha_{k-1}}$ learned at the previous iteration. As a result, Lemma 1 should be restated by taking a supremum over all the possible policies that can be generated as greedy policies of the functions in $\mathcal{F}$. The complexity of this space of policies depends on the number of actions $|\mathcal{A}|$ and the dimension d. Finally, the complexity of the joint space obtained by $\mathcal{F}$ and the space of policies would appear in the final bound which would differ from the one in Theorem 8 only in constant factors.

5. Conclusion and comparison with LSTD

We provided a generalization bound for Bellman residuals and used it to provide performance bounds for an approximate policy iteration algorithm in which an empirical Bellman residual minimization is used at each policy evaluation step.

Compared to the LSTD approach analyzed in Lazaric et al. (2010) we have a poorer estimation rate of $O(1/\sqrt{n})$ instead of $O(1/n)$ and it is an open question to whether an improved rate for Bellman residuals can be obtained (see Remark 1). The assumptions are also different: in this BRM approach we assumed that we have a generative model and thus performance bounds can be obtained under any sampling distribution μ, whereas since LSTD only requires the observation of a single trajectory (following a given policy) it can only provide performance bounds under the stationary distribution of that policy. However in a policy iteration scheme it is not enough to accurately approximate the current policy under the stationary distribution since the greedy policy w.r.t. that approximation can be arbitrarily poor. Thus the performance of BRM are better controlled than that of LSTD, which is reflected in the fact that the concentrability coefficients $C(\rho, \mu)$ (used in Theorem 8) can be controlled in the BRM approach (such as by choosing a uniform distribution μ) but not in LSTD unless we make additional (usually strong) assumptions on the stationary distributions (such as being lower-bounded by a uniform distribution, like in (Munos, 2003)).

Acknowledgments

This work was supported by French National Research Agency (ANR) through the projects EXPLO-RA n° ANR-08-COSI-004 and LAMPADA n° ANR-09-EMER-007, by Ministry of Higher Education and Research, Nord-Pas de Calais Regional Council and FEDER through the "contrat de projets état region (CPER) 2007–2013", and by PASCAL2 European Network of Excellence.

References

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.

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

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

D. P. Bertsekas and S.E. Shreve. *Stochastic Optimal Control (The Discrete Time Case).* Academic Press, New York, 1978.

D. P. Bertsekas and J. N. Tsitsiklis. *Neuro-Dynamic Programming.* Athena Scientific, Belmont, MA, 1996b.

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

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

L. Györfi, M. Kohler, A. Krzyżak, and H. Walk. *A distribution-free theory of nonparametric regression.* Springer-Verlag, New York, 2002.

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

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

R. Munos. Error bounds for approximate policy iteration. In *19th International Conference on Machine Learning*, pages 560–567, 2003.

R. Munos. Performance bounds in Lp norm for approximate value iteration. *SIAM J. Control and Optimization*, 2007.

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

Warren B. Powell. *Approximate Dynamic Programming: Solving the Curses of Dimensionality (Wiley Series in Probability and Statistics).* Wiley-Interscience, 2007.

M.L. Puterman. *Markov Decision Processes — Discrete Stochastic Dynamic Programming.* John Wiley & Sons, Inc., New York, NY, 1994.

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

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

Jennie Si, Andrew G. Barto, Warren Buckler Powell, and Don Wunsch. *Handbook of Learning and Approximate Dynamic Programming (IEEE Press Series on Computational Intelligence).* Wiley-IEEE Press, 2004.

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

C. J. C. H. Watkins. *Learning from Delayed Rewards.* PhD thesis, King's College, Cambridge, UK, 1989.

R. J. Williams and L.C. Baird, III. Tight performance bounds on greedy policies based on imperfect value functions. In *Proceedings of the Tenth Yale Workshop on Adaptive and Learning Systems*, 1994.

Appendix A. Proof of Lemma 3

Step 1: Introduce the empirical process. Let $\mathcal{J}(C)$ be the class of functions induced by $\mathcal{L}_n$ from $\mathcal{F}(C)$ defined as

$$\mathcal{J}(C) = \{j_\alpha : (x,y,z) \mapsto (f_\alpha(x) - \gamma f_\alpha(y))(f_\alpha(x) - \gamma f_\alpha(z)); ||\alpha||_2 \leq C\}.$$

Note that this is the product of two linear spaces of dimension d. Furthermore, we can now rewrite $\mathcal{L}_n(f_\alpha) = P_n j_\alpha$ and $\mathcal{L}(f_\alpha) = P j_\alpha$, where P_n is the empirical measure w.r.t. X_i, Y_i, Y_i' and P is the measure according to which the samples are distributed. As a result both $\mathcal{L}_n(f_\alpha)$ and $\mathcal{L}(f_\alpha)$ are linear w.r.t. j_α. Note also that for any $(x,y,z) \in \mathcal{X}^3$, $|j_\alpha(x,y,z)| \leq ||\alpha||_2^2 (1+\gamma)^2 \sup_{x\in\mathcal{X}} ||\phi(x)||_2^2 = C^2(1+\gamma)^2 L_{\max}^2$, using Cauchy-Schwartz's inequality.

Step 2: Bound the covering number. We want to bound the ϵ-covering number of the class of functions $\mathcal{J}(C)$ in norm $||.||_\infty$. Since each function j_α can be written as $j_\alpha(x,y,z) = g_\alpha(x,y) g_\alpha(x,z)$, where $g_\alpha(x,y) = \sum_{i=1}^d \alpha_i(\phi_i(x) - \gamma\phi_i(y))$, we can relate the covering number of $\mathcal{J}(C)$ to the covering number of the space of functions g_α. Indeed, let us consider an ϵ-cover G_0 for the space of functions g_α such that $||\alpha||_2 \leq C$. Thus for a given α there exists $g_{\alpha_0} \in G_0$ such that $||g_\alpha - g_{\alpha_0}|| \leq \epsilon$. Now, we can build a cover for $\mathcal{J}(C)$. We have

$$\begin{aligned}
|j_\alpha(x,y,z) - j_{\alpha_0}(x,y,z)| &\leq |g_\alpha(x,y)g_\alpha(x,z) - g_{\alpha_0}(x,y)g_\alpha(x,z)| \\
&\quad + |g_{\alpha_0}(x,y)g_\alpha(x,z) - g_{\alpha_0}(x,y)g_{\alpha_0}(x,z)| \\
&\leq ||g_\alpha||_\infty ||g_\alpha - g_{\alpha_0}||_\infty + ||g_{\alpha_0}||_\infty ||g_\alpha - g_{\alpha_0}||_\infty \\
&\leq 2C(1+\gamma)L_{\max}\epsilon,
\end{aligned}$$

which enables us to deduce that

$$\begin{aligned}
\mathcal{N}(\epsilon, \mathcal{J}(C), ||.||_\infty) &\leq \mathcal{N}\Big(\frac{\epsilon}{2C(1+\gamma)L_{\max}}, \{g_\alpha; ||\alpha||_2 \leq C\}, ||.||_\infty\Big) \\
&\leq \mathcal{N}\Big(\frac{\epsilon}{2C(1+\gamma)L_{\max}}, \{g_\alpha; ||g||_n \leq C(1+\gamma)L_{\max}\}, ||.||_n\Big) \\
&\leq \Big(\frac{6C^2(1+\gamma)^2 L_{\max}^2}{\epsilon}\Big)^d
\end{aligned}$$

where we used the fact that $||g||_n \leq ||g||_\infty$ and $||g||_n \leq ||\alpha||_2(1+\gamma)L_{\max}$.

Step 3: Use chaining technique. Let us consider ϵ_l-covers $\mathcal{J}_l$ of $\mathcal{J}(C)$, for $l = 0, \ldots, \infty$, with $\mathcal{J}_0 = j_{\alpha_0}$. We moreover assume that $\mathcal{J}_{l+1}$ is a refinement of $\mathcal{J}_l$ and that $\epsilon_{l+1} \leq \epsilon_l$. Then for a given $j \in \mathcal{J}(C)$, we define $j_l = \Pi(j, \mathcal{J}_l)$ the projection of j into $\mathcal{J}_l$, for the norm $||j||_\infty$. Thus, $j = (j - j_L) + \sum_{l=1}^L (j_l - j_{l-1}) + j_0$. Since $0 \in \mathcal{J}(C)$, we consider $j_{\alpha_0} = 0$. Note that by definition, we need $||j||_\infty \leq \epsilon_0$. Thus we define $\epsilon_0 = C^2(1+\gamma)^2 L_{\max}^2$.

Moreover, we have for any $j \in \mathcal{J}(C)$,

$$|(P - P_n)(j)| \leq |(P - P_n)(j - j_L)| + \sum_{l=1}^L |(P - P_n)(j_l - j_{l-1})| \leq 2\epsilon_L + \sum_{l=1}^L |(P - P_n)(j_l - j_{l-1})|$$

We introduce for convenience the following notation: $\rho(t) = \Pr(\exists f \in \mathcal{F}(C), \quad |\mathcal{L}(f_\alpha) - \mathcal{L}_n(f_\alpha)| > t)$. Thus if we now introduce η and $(\eta_l)_{l \leq L}$ such that $\sum_{l=1}^{L} \eta_l \leq \eta$, then for L large enough such that $2\epsilon_L \leq t_2$, we have:

$$\begin{aligned}
\rho(\eta t_1 + t_2) &\leq \Pr(\exists f \in \mathcal{F}(C), \quad 2\epsilon_L + \sum_{l=1}^{L} |(P - P_n)(j_l - j_{l-1})| > \sum_{l=1}^{L} \eta_l t + t_2) \\
&\leq \sum_{l=1}^{L} \Pr(\exists j \in \mathcal{J}(C), \quad |(P - P_n)(j_{\alpha_l} - j_{\alpha_{l-1}})| > \eta_l t_1) \\
&\leq \sum_{l=1}^{L} N_l N_{l-1} \sup_{j \in \mathcal{J}(C)} \Pr(|(P - P_n)(j_{\alpha_l} - j_{\alpha_{l-1}})| > \eta_l t_1) \\
&\leq \sum_{l=1}^{L} 2N_l^2 \exp(-\frac{n t_1^2 \eta_l^2}{2(4\epsilon_l)^2})
\end{aligned}$$

where $N_l = \mathcal{N}(\epsilon_l, \mathcal{J}(C), ||.||_\infty)$, and where the last inequality comes from the fact that

$$|j_{\alpha_l}(X_i, Y_i, Y_i') - j_{\alpha_{l-1}}(X_i, Y_i, Y_i') - Pj_{\alpha_l} + Pj_{\alpha_{l-1}}| \leq 2||j_{\alpha_l} - j_{\alpha_{l-1}}||_\infty \leq 4||j_{\alpha_l} - j_\alpha||_\infty \leq 4\epsilon_l.$$

Step 4: Define the free parameters. Thus, if we define, for all $l \geq 1$, $\eta_l \stackrel{\text{def}}{=} \frac{8\epsilon_l}{t_1}\sqrt{\frac{2\log(N_l)}{n}}$, then we deduce the following inequality: $\rho(\eta t_1 + t_2) \leq 2\sum_{l=1}^{L} N_l^{-2}$.

Now, since $N_l \leq (\frac{6C^2(1+\gamma)^2 L_{\max}^2}{\epsilon_l})^d$, let $\epsilon_l = 6C^2(1+\gamma)^2 L_{\max}^2 2^{-l}(\delta/2)^{1/2d}(2^{2d}-1)^{1/2d}$ for $l \geq 1$. Thus we deduce that $\sum_{l=1}^{L} N_l^{-2} \leq \delta/2$. We finally get:

$$\begin{aligned}
\eta t_1 + t_2 &= \sum_{l=1}^{L} 8\epsilon_l \sqrt{\frac{2\log(N_l)}{n}} + 2\epsilon_L \\
&\leq 48C^2(1+\gamma)^2 L_{\max}^2 (\delta/2)^{1/2d}(2^{2d}-1)^{1/2d} \sum_{l=1}^{L} 2^{-l}\sqrt{\frac{2\log(N_l)}{n}} + 2\epsilon_L \\
&\leq \frac{96C^2(1+\gamma)^2 L_{\max}^2}{\sqrt{n}} \sum_{l=1}^{L} 2^{-l}\sqrt{2dl\log(2) + \log(2/\delta) - \log(2^{2d}-1)} + 2\epsilon_L \\
&\leq \frac{96C^2(1+\gamma)^2 L_{\max}^2}{\sqrt{n}} \sum_{l=1}^{L} 2^{-l}\sqrt{2d(l-1)\log(2) + \log(4/\delta)} + 2\epsilon_L
\end{aligned}$$

Thus, when $L \to \infty$, we get:

$$\eta t_1 + t_2 \leq \frac{96C^2(1+\gamma)^2 L_{\max}^2}{\sqrt{n}} \sum_{l=1}^{\infty} 2^{-l}\sqrt{2d(l-1)\log(2) + \log(4/\delta)}$$

We deduce that with probability higher than $1-\delta$, the following holds true:

$$\sup_{f \in \mathcal{F}(C)} |\mathcal{L}(f_\alpha) - \mathcal{L}_n(f_\alpha)| \leq 96C^2 L_{\max}^2 (\sqrt{\frac{2d\log(2)}{n}} + \sqrt{\frac{\log(4/\delta)}{n}})$$

Then we use the fact that $\mathcal{L}(f_\alpha) = \mathcal{L}(f_{\frac{\alpha}{||\alpha||}})||\alpha||^2$ and similarly $\mathcal{L}_n(f_\alpha) = \mathcal{L}_n(f_{\frac{\alpha}{||\alpha||}})||\alpha||^2$ to deduce that with the same probability, for all α,

$$|\mathcal{L}(f_\alpha) - \mathcal{L}_n(f_\alpha)| \leq ||\alpha||^2 (\sup_{f \in \mathcal{F}(1)} |\mathcal{L}(f) - \mathcal{L}_n(f)|)$$

The final results follows by aesthetics simplifications.

Appendix B. Proof of Lemma 1

Step 1: Introduce the empirical process. The proof for B_n follows the same lines as for L_n using the following class of functions, induced by $\mathcal{B}_n$ from $\mathcal{F}(C)$ and defined as:

$$\mathcal{J}(C) = \{j_\alpha : (x,y,z) \mapsto (f_\alpha(x) - \gamma f_\alpha(y) + r(x))(f_\alpha(x) - \gamma f_\alpha(z) + r(x)); ||\alpha||_2 \leq C\}.$$

Then we have $\mathcal{B}_n(f_\alpha) = P_n j_\alpha$ and $\mathcal{B}(f_\alpha) = P j_\alpha$. Now, we have $|j_\alpha(X_i, Y_i, Y_i')| \leq (||\alpha||_2(1+\gamma)\sup_x ||\phi(x)||_2 + R_{\max})^2 = [C(1+\gamma)L_{\max} + R_{\max}]^2$. Note that the function 0 does not a priori belongs to $\mathcal{J}(C)$, thus we have an additional term to control corresponding to the decomposition of $j = (j - j_L) + \sum_{l=1}^{L}(j_l - j_{l-1}) + j_0$ for some nonzero $j_0 \in J(C)$.

Step 2: Bound the covering number. With this new definition of $\mathcal{J}(C)$, we have:

$$\mathcal{N}(\epsilon, \mathcal{J}(C), ||.||_\infty) \leq \Big(\frac{6(C(1+\gamma)L_{\max} + R_{\max})^2}{\epsilon_l}\Big)^d$$

Step 3: Use chaining technique. Then using chaining technique, we get the corresponding upper bound:

$$\begin{aligned}
\rho(\eta t_1 + t_2 + t_3) &= \Pr(\exists f \in \mathcal{F}(C) | \mathcal{L}(f_\alpha) - \mathcal{L}_n(f_\alpha)| > \eta t_1 + t_2 + t_3) \\
&\leq 2\sum_{l=1}^{L} N_l^{-2} + 2\exp(-\frac{n t_3^2}{2[C(1+\gamma)L_{\max} + R]^4})
\end{aligned}$$

where the last term comes from the bound on $\Pr(|(P - P_n)(j_0) \geq t_3)$.

Step 4: Define the free parameters. We define $\epsilon_l = 9(C(1+\gamma)L_{\max}+R)^2 2^{-l}(\delta/4)^{1/2d}(2^{2d}-1)^{1/2d}$ for $l \geq 1$, set $t_3 = [C(1+\gamma)L_{\max} + R]^2\sqrt{\frac{2\log(4/\delta)}{n}}$ and derive that with probability higher than $1-\delta$,

$$\begin{aligned}
\sup_{f \in \mathcal{F}(C)} |\mathcal{B}(f_\alpha) - \mathcal{B}_n(f_\alpha)| &\leq 96[C(1+\gamma)L_{\max} + R]^2 (\sqrt{\frac{2d\log(2)}{n}} + \sqrt{\frac{\log(8/\delta)}{n}}) \\
&\quad + [C(1+\gamma)L_{\max} + R]^2 \sqrt{\frac{2\log(4/\delta)}{n}}.
\end{aligned}$$

The final result follows after some aesthetics simplifications.