**Machine Learning manuscript No.**
(will be inserted by the editor)

# Learning near-optimal policies with Bellman-residual minimization based fitted policy iteration and a single sample path

**András Antos**[1], **Csaba Szepesvári**[1*], **Rémi Munos**[2]

[1] Computer and Automation Research Inst.
of the Hungarian Academy of Sciences
Kende u. 13-17, Budapest 1111, Hungary
e-mail: {antos, szcsaba}@sztaki.hu

[2] Centre de Mathématiques Appliquées
Ecole Polytechnique
91128 Palaiseau Cedex, France
e-mail: remi.munos@polytechnique.fr

The date of receipt and acceptance will be inserted by the editor

**Abstract** We consider the problem of finding a near-optimal policy in continuous space, discounted Markovian Decision Problems given the trajectory of some behaviour policy. We study the policy iteration algorithm where in successive iterations the action-value functions of the intermediate policies are obtained by picking a function from some fixed function set (chosen by the user) that minimizes an unbiased finite-sample approximation to a novel loss function that upper-bounds the unmodified Bellman-residual criterion. The main result is a finite-sample, high-probability bound on the performance of the resulting policy that depends on the mixing rate of the trajectory, the capacity of the function set as measured by a novel capacity concept that we call the VC-crossing dimension, the approximation power of the function set and the discounted-average concentrability of the future-state distribution. To the best of our knowledge this is the first theoretical reinforcement learning result for off-policy control learning over continuous state-spaces using a single trajectory.

* Now at the Department of Computing Science, University of Alberta, Edmonton, AB, Canada

**Key words** Reinforcement learning, policy iteration, Bellman-residual minimization, off-policy learning, nonparametric regression, least-squares regression, finite-sample bounds

## 1 Introduction

In industrial control problems gathering data of the controlled system is often separated from the learning phase: The data is gathered via "field-experiments", whence it is taken to the laboratory where it is used to design a new optimized controller. A crucial feature of these problems is that the data is fixed and new samples cannot be generated at will. Often, the data is obtained by observing the controlled system operated using a existing controller, or policy, henceforth, following standard terminology in reinforcement learning, called the behaviour policy (Sutton and Barto, 1987, Chapter 5.6).

In this paper we are interested in designing algorithms and proving bounds on the achievable performance in such settings, where specifically we assume that the control task can be modelled as a discounted Markovian Decision Problem with continuous state-variables and a finite number of actions. We shall denote the number of actions by $L$.

The algorithm studied here is an instance of fitted policy iteration. Like in policy iteration, the algorithm repeatedly computes an evaluation function of the policy of the previous step and then uses this evaluation function to compute the next improved policy. In order to avoid the need of learning a model, action-value evaluation functions are computed, making the policy improvement step trivial, just like in the least-squares policy iteration (LSPI) algorithm of Lagoudakis and Parr (2003). However, unlike LSPI which builds on least-squares temporal difference learning (LSTD) of **?**, we build our algorithm on the idea of minimizing Bellman-residuals. The idea of using Bellman-residuals in policy iteration goes back at least to **?** who proposed it for computing approximate state-value functions given the model of a finite-state and action MDP. The idea is that a small Bellman-error may yield a good approximation to the policy evaluation function, which in turn may imply a good final performance. One major obstacle of using Bellman-residual minimization when learning without model is that the trivial sample-based approximation of the loss function that uses data from a single trajectory of the behaviour policy is not an unbiased estimate of the population Bellman-residual loss function (e.g. Sutton and Barto, 1987, pp. 200). Here we propose to overcome this problem by using a novel loss function that upper bounds the loss Bellman-residual loss that does not suffer from this problem.

Our main result shows that if the trajectory used as the input is sufficiently representative then the performance of the policy returned by our algorithm improves at a rate of $1/N^{1/4}$ (where $N$ is the length of the trajectory) up to a limit set by the choice of the function set $\mathcal{F}^L$. To the best of our knowledge this is the first result in the literature where finite-sample error bounds are obtained for an algorithm that works for continuous state-space MDPs, uses function approximators and considers control learning in an off-policy setting, i.e., learning from a single trajectory of some fixed behaviour policy.

One major technical difficulty of the proof is that we have to deal with dependent samples. The main condition here is that the trajectory should be sufficiently representative and rapidly mixing. For the sake of simplicity, we also require that the states in the trajectory follow a stationary distribution, though we believe that with some additional work this condition could be removed. The mixing condition, on the other hand, seems to be essential for efficient learning. The particular mixing condition that we use is exponential $\beta$-mixing, used earlier e.g. by Meir (2000) for analyzing nonparametric time-series prediction or by Baraud et al. (2001) for analyzing penalized least-squares regression. The particular mixing condition assumed allows us to derive polynomial decay rates for the estimation error as a function of the sample size. If we were to relax this condition to algebraic $\beta$-mixing (i.e., mixing at a slower rate), the estimation error-bound would decay with the logarithm of the number of samples, i.e., at a sub-polynomial rate. Hence, learning is still possible, but could be very slow. Let us finally note that for Markov processes, geometric ergodicity implies exponential $\beta$-mixing (see Davidov, 1973; or Doukhan, 1994, Chap. 2.4), hence for such processes there is no loss of generality in assuming exponential $\beta$-mixing.

In order to arrive at our bound, we introduce a new capacity concept which we call the VC-crossing dimension. The VC-crossing dimension of $\mathcal{F}$ is defined as the VC-dimension of a set-system that consists of the zero-level sets of the pairwise differences of functions from $\mathcal{F}$. A short intuitive explanation of the need of this concept is that in one step of policy iteration we attempt to evaluate a policy the is greedy to the action-value function computed in the previous step. Computing the greedy actions involves comparing functions of $\mathcal{F}$ (an action-value evaluation function can be given as a list $(f_1, \ldots, f_L)$ of functions from $\mathcal{F}$, each corresponding to some action from $\mathcal{A}$), and the set of states when a function $f$ from $\mathcal{F}$ majorizes another function $f'$ from $\mathcal{F}$ can be given as the set of states where the difference $f - f'$ is greater than zero.

Similarly to bounds of regression, our bounds depend on the approximation power of the function set, too. One major difference, though, is that in our case the natural way to measure the approximation power of a function set is different from how it is done in regression. Whilst in regression, the approximation power is measured as the minimum distance to the target (regression) function, we use error measures that reflect whether the function set fits the policy evaluation operators underlying the MDP in a sense that will be made precise later.

The bound also depends on the number of steps ($K$) of policy iteration. As expected, there are two terms involving $K$ that behave conversely: One of the terms us the usual one that decays at a geometric rate (the base being

$\gamma$, the discount factor of the MDP). The other term comes from the reuse of the data throughout all the iterations and this scales with the logarithm of the number of iterations. Hence, the reuse of data in the iterations makes the performance degrade at a slow rate, a point that was made just recently by Munos and Szepesvári (2006) in the analysis of approximate value iteration. Optimizing the bound in $K$ numerically is straightforward. As a matter of fact, the optimal value of $K$ will depend e.g. on the capacity of the function set, the mixing rate and the number of samples. Interestingly, it will not depend on the approximation-power of the function set.

In order to arrive at our results, we also need to make some assumptions on the controlled system. In particular, we assume that the state space is compact and the action space is finite. The compactness condition is purely technical and can be relaxed in various ways (e.g. by making assumptions about the stability of the system). The finiteness condition on the action space, on the other hand, seems to be essential for our analysis to go through. We also need to make a certain controllability (or rather uncontrollability) assumption. This particular assumption is used in the method proposed by Munos (2003) that is used to bound the final *weighted-norm* error as a function of the weighted-norm errors made in the intermediate steps of the algorithm. If we were to use $L^\infty$ analysis then the controllability assumption would not be needed. The difficulty is that since the policy evaluation-functions are obtained via a least-squares approach, it would be difficult to derive good $L^\infty$ bounds on the errors of the intermediate steps.

The particular assumption studied here requires that the rate at which the future-state distribution can be concentrated (by selecting a non-stationary Markov policy) as compared with the state-distribution $\nu$ should be sub-exponential. In general, this holds for "noisy" systems, but, as argued by Munos and Szepesvári (2006) it holds for certain deterministic systems, as well.

The paper is organized as follows: In the next section (Section 2) we introduce the basic concepts, definitions and symbols needed in the rest of the paper. The algorithm along with its motivation is given in Section 3. This is followed by some additional definitions necessary for the presentation of the main result, which is done at the beginning of Section 4. The rest of this section is divided into three parts, each devoted to one major step of the proof. In particular, in Section 4.1 a finite-sample bound is given on the error of the particular policy evaluation procedure proposed here. The bound makes the dependence on the complexity of the function space, the mixing rate of the trajectory and the number of samples explicit. In Section 4.2 we prove a bound on how errors propagate throughout the iterations of the procedure. The proof of the main result is finished in Section 4.3. We discuss the main result, in the context of previous work in Section 5. Finally, our conclusions are drawn and possible directions for future work are outlined in Section 6.

## 2 Definitions

As we shall work with continuous spaces we will need some simple measure theoretic concepts. These are introduced first. Next, Markovian Decision Problems (MDPs) and various MDP concepts are introduced along with the symbols, notation, operators, etc. needed throughout the paper.

For a measurable space with domain $S$ we let $M(S)$ denote the set of all probability measures over $S$. Fix $p \geq 1$. For a measure $\nu \in M(S)$ and a measurable function $f : S \to \mathbb{R}$ we let $\|f\|_{p,\nu}$ denote the $L^p(\nu)$-norm of $f$:

$$\|f\|_{p,\nu}^p = \int |f(s)|^p \nu(ds).$$

We shall also write $\|f\|_\nu$ to denote the $L^2(\nu)$-norm of $f$. We denote the space of bounded measurable functions with domain $\mathcal{X}$ by $B(\mathcal{X})$, and the space of measurable functions with bound $0 < K < \infty$ by $B(\mathcal{X}; K)$. We let $\|f\|_\infty$ denote the supremum norm: $\|f\|_\infty = \sup_{x \in \mathcal{X}} |f(x)|$. The symbol $\mathbb{I}_{\{E\}}$ shall denote the indicator function: For an event $E$, $\mathbb{I}_{\{E\}} = 1$ iff $E$ holds and $\mathbb{I}_{\{E\}} = 0$, otherwise. We use $\mathbf{1}$ to denote the function that takes on the constant value one everywhere over its domain and use $\mathbf{0}$ to denote the likewise function that takes zero everywhere.

A discounted MDP is defined by a quintuple $(\mathcal{X}, \mathcal{A}, P, S, \gamma)$, where $\mathcal{X}$ is the (possibly infinite) *state space*, $\mathcal{A} = \{a_1, a_2, \ldots, a_L\}$ is the set of *actions*, $P : \mathcal{X} \times \mathcal{A} \to M(\mathcal{X})$ is the *transition probability kernel*, $P(\cdot|x, a)$ defining the next-state distribution upon taking action $a$ from state $x$, $S(\cdot|x, a)$ gives the corresponding distribution of *immediate rewards*, and $\gamma \in (0, 1)$ is the discount factor.

We make the following assumptions on the MDP:

**Assumption 1 (MDP Regularity)** *$\mathcal{X}$ is a compact subspace of the s-dimensional Euclidean space. We assume that the random immediate rewards are bounded by $\hat{R}_{\max}$ and the expected immediate rewards $r(x, a) = \int r S(dr|x, a)$ are bounded by $R_{\max}$: $\|r\|_\infty \leq R_{\max}$. (Note that $R_{\max} \leq \hat{R}_{\max}$.)*

A policy is defined as a (measurable) mapping from past observations to a distribution over the set of actions (for details, see Bertsekas and Shreve, 1978). A policy is called Markov if the distribution depends only on the last state of the observation sequence. A policy is called stationary Markov if this dependency does not change by time. For a stationary Markov policy, the probability distribution over the actions given some state $x$ will be denoted by $\pi(\cdot|x)$. A policy is deterministic if the probability distribution concentrates on a single action for all histories. Such policies will be identified by mappings from the states to actions, i.e., functions of the form $\pi : \mathcal{X} \to \mathcal{A}$.

The value of a policy $\pi$ when it is started from a state $x$ is defined as the total expected discounted reward that is encountered while the policy is executed:

$$V^\pi(x) = \mathbb{E}_\pi\left[\sum_{t=0}^{\infty} \gamma^t R_t | X_0 = x\right].$$

Here $R_t$ denotes the reward received at time step $t$; $R_t \sim S(\cdot|X_t, A_t)$ and $X_t$ is assumed to evolve according to $X_{t+1} \sim P(\cdot|X_t, A_t)$ where $A_t$ is sampled from the distribution assigned to the past observations by $\pi$. For a stationary Markov policy $\pi$, $A_t \sim \pi(\cdot|X_t)$, whilst if $\pi$ is deterministic stationary Markov then by our previous remark we write $A_t = \pi'_t(X_t)$. We introduce $Q^\pi : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$, the action-value function of policy $\pi$:

$$Q^\pi(x,a) = \mathbb{E}_\pi\left[\sum_{t=0}^{\infty} \gamma^t R_t | X_0 = x, A_0 = a\right].$$

The goal is to find a policy that attains the best possible values, $V^*(x) = \sup_\pi V^\pi(x)$ for all states $x \in \mathcal{X}$. Function $V^*$ is called the optimal value function. A policy is called optimal if it attains the optimal values $V^*(x)$ for *any* state $x \in \mathcal{X}$, i.e., if $V^\pi(x) = V^*(x)$ for all $x \in \mathcal{X}$. The function $Q^*(x,a)$ is defined analogously: $Q^*(x,a) = \sup_\pi Q^\pi(x,a)$. It is known that for any policy $\pi$, the functions $V^\pi$,$Q^\pi$ are bounded by $R_{\max}/(1-\gamma)$, as are $Q^*$ and $V^*$. We say that a (deterministic stationary) policy $\pi$ is *greedy* w.r.t. an action-value function $Q \in B(\mathcal{X} \times \mathcal{A})$ and write

$$\pi = \hat{\pi}(\cdot; Q),$$

if, for all $x \in \mathcal{X}$ and $a \in \mathcal{A}$,

$$\pi(x) \in \operatorname*{argmax}_{a \in \mathcal{A}} Q(x,a).$$

Since $\mathcal{A}$ is finite, such a greedy policy always exists. It is known that under mild conditions the greedy policy w.r.t. $Q^*$ is optimal (e.g. Bertsekas and Shreve, 1978). Hence, in what follows, without the loss of generality we restrict the search for a good policy to the set of deterministic, stationary Markov policies and by the word 'policy' we shall mean such policies.

For a (deterministic stationary Markov) policy $\pi$, we define the operator $T^\pi : B(\mathcal{X} \times \mathcal{A}) \to B(\mathcal{X} \times \mathcal{A})$ by

$$(T^\pi Q)(x,a) = r(x,a) + \gamma \int Q(y, \pi(y)) P(dy|x,a).$$

It is easy to see that $T^\pi$ is a contraction operator w.r.t. the supremum-norm and the action-value value-function of $\pi$ is the unique fixed point of $T^\pi$:

$$T^\pi Q^\pi = Q^\pi. \tag{1}$$

We define the projection operator $E^\pi : B(\mathcal{X} \times \mathcal{A}) \to B(\mathcal{X})$ by

$$(E^\pi Q)(x) = Q(x, \pi(x)), \quad Q \in B(\mathcal{X} \times \mathcal{A}).$$

We define two operators corresponding to the transition probability kernel $P$ as follows: A right-linear operator, $P\cdot : B(\mathcal{X}) \to B(\mathcal{X} \times \mathcal{A})$, is defined by

$$(PV)(x,a) = \int V(y) P(dy|x,a),$$

whilst a left-linear operator, $\cdot P : M(\mathcal{X} \times \mathcal{A}) \to M(\mathcal{X})$, is defined by

$$(\rho P)(dy) = \int P(dy|x,a)\rho(dx,da).$$

This operator is also extended to act on measures over $\mathcal{X}$ via

$$(\rho P)(dy) = \frac{1}{L} \sum_{a \in \mathcal{A}} \int P(dy|x,a)\rho(dx).$$

By composing $P$ and $E^\pi$, we define $P^\pi$:

$$P^\pi = P E^\pi.$$

Note that this equation defines *two* operators: a right- and a left-linear one.

Throughout the paper $\mathcal{F} \subset \{f : \mathcal{X} \to \mathbb{R}\}$ will denote some subset of real-valued functions over the state-space $\mathcal{X}$. For convenience, we will treat elements of $\mathcal{F}^L$ as real-valued functions $f$ defined over $\mathcal{X} \times \mathcal{A}$ with the obvious identification $f \equiv (f_1, \ldots, f_L)$, $f(x, a_j) = f_j(x)$, $j = 1, \ldots, L$. The set $\mathcal{F}^L$ will define the set of admissible functions used in the optimization step of our algorithm.

Finally, for $\nu \in M(\mathcal{X})$, we extend $\|\cdot\|_{p,\nu}$ ($p \geq 1$) to $\mathcal{F}^L$ by

$$\|f\|_{p,\nu}^p = \frac{1}{L} \sum_{j=1}^{L} \|f_j\|_{p,\nu}^p .$$

Alternatively, we define $\nu(dx, da)$, the extension of $\nu$ to $\mathcal{X} \times \mathcal{A}$ via

$$\int Q(x,a)\nu(dx,da) = \frac{1}{L} \sum_{j=1}^{L} \int Q(x,a_j)\nu(dx). \tag{2}$$

For real numbers $a$ and $b$, $a \vee b$ shall denote the maximum of $a$ and $b$. Similarly, $a \wedge b$ shall denote the minimum of $a$ and $b$.

## 3 Algorithm

The algorithm studied in this paper is an instance of the generic fitted policy iteration method, whose pseudo-code is shown in Figure 1. By assumption, the training sample, $D$, used by the algorithm consists of a finite trajectory

$$\{(X_t, A_t, R_t)\}_{1 \leq t \leq N}$$

**FittedPolicyQ**(D,K,$Q_{-1}$,PEval)
// D: samples (e.g. trajectory)
// K: number of iterations
// $Q_{-1}$: Initial action-value function
// PEval: Policy evaluation routine
$Q \leftarrow Q_{-1}$ // Initialization
**for** $k = 0$ to $K - 1$ **do**
  $Q' \leftarrow Q$
  $Q \leftarrow$PEval($\hat{\pi}(\cdot; Q'), D$)
**end for**
**return** $Q$ // or $\hat{\pi}(\cdot; Q)$, the greedy policy w.r.t. $Q$

**Fig. 1** Model-free Fitted Policy Iteration

of some stochastic stationary policy $\pi$: $A_t \sim \pi(\cdot|X_t)$, $X_{t+1} \sim P(\cdot|X_t, A_t)$, $R_t \sim S(\cdot|X_t, A_t)$. We assume that this trajectory is sufficiently representative in a sense that will be made precise in the next section. For now, let us make the assumption that $X_t$ is stationary and is distributed according to some (unknown) distribution $\nu$. The action-evaluation function $Q_{-1}$ is used to initialize the first policy (alternatively, one may start with an arbitrary initial policy at the price of making the algorithm somewhat more complicated). Procedure *PEval* takes data in the form of a long trajectory and some policy. In this case the policy is just the greedy policy with respect to $Q'$, $\hat{\pi} = \hat{\pi}(\cdot; Q')$. Based on $\hat{\pi}$, *PEval* should return an approximation to the action-value function $Q^{\hat{\pi}}$.

There are many possibilities to design *PEval*. In this paper we consider an approach based on Bellman-residual minimization (BRM). The basic idea of BRM comes from rewriting the fixed point equation (1) for $Q^{\hat{\pi}}$ in the form $Q^{\hat{\pi}} - T^{\hat{\pi}} Q^{\hat{\pi}} = 0$. The left-hand-side of this equation is called the *Bellman-residual* of $Q^{\hat{\pi}}$. For $Q \neq Q^{\hat{\pi}}$, the Bellman-residual of $Q$ is nonzero: $Q - T^{\hat{\pi}} Q \neq 0$. Hence it is expected that a smaller risk, $\left\| Q - T^{\hat{\pi}} Q \right\|$, yields better estimates. Here $\|\cdot\|$ could be any norm. From our point of view, the $L^2$-norm is attractive as it leads to well-studied optimization problems and makes the connection to regression-estimation easier. Hence, let us consider the loss function

$$L(Q; \hat{\pi}) = \left\| Q - T^{\hat{\pi}} Q \right\|_{\nu}^2 ,$$

where $\nu$, the stationary distribution underlying the states in the input data is selected to facilitate a sample-based approximation (remember that $\|Q\|_{\nu}^2 = 1/L \sum_{j=1}^{L} \|Q(\cdot, a_j)\|_{\nu}^2$).

We chase $Q = \operatorname{argmin}_{f \in \mathcal{F}^L} L(f; \hat{\pi})$.[1] At a first sight it may seem that

$$\hat{L}_N(f; \hat{\pi}) = \frac{1}{NL} \sum_{t=1}^{N} \sum_{j=1}^{L} \frac{\mathbb{I}_{\{A_t = a_j\}}}{\pi(a_j|X_t)} \left( f(X_t, a_j) - (R_t + \gamma f(X_{t+1}, \hat{\pi}(X_{t+1}))) \right)^2 \tag{3}$$

is an appropriate sample-based approximation to $L(f; \hat{\pi})$. Indeed, for any given $X_t, A_t$ and $f$, $R_t + \gamma f(X_{t+1}, \hat{\pi}(X_{t+1})$ is an unbiased estimate of $(T^{\hat{\pi}} f)(X_t, A_t)$. However, as it is well known (see Sutton and Barto, 1987, pp. 200, Munos, 2003 or Lagoudakis and Parr, 2003 for discussions), $\hat{L}_N$ is *not* a "proper" approximation to the corresponding $L^2$ Bellman-error: $\mathbb{E}\left[\hat{L}_N(f; \hat{\pi})\right] \neq L(f; \hat{\pi})$. Indeed, elementary calculus shows that for $Y \sim P(\cdot|x, a)$, $R \sim S(\cdot|x, a)$,

$$\begin{aligned} &\mathbb{E}\left[ \Big( f(x, a) - (R + \gamma f(Y, \hat{\pi}(Y))) \Big)^2 \right] \\ &= (f(x, a) - (T^{\hat{\pi}} f)(x, a))^2 + \operatorname{Var}\left[R + \gamma f(Y, \hat{\pi}(Y))\right]. \end{aligned}$$

It follows that minimizing $\hat{L}_N(f; \hat{\pi})$ in the limit when $N \to \infty$ is equivalent to minimizing the sum of $\gamma^2 \mathbb{E}\left[\operatorname{Var}\left[f(Y, \hat{\pi}(Y))\right] |X\right]$ and $L(f; \hat{\pi})$. The unwanted variance term acts like a penalty factor, favoring smooth solutions (if $f$ is constant then $\operatorname{Var}\left[f(Y, \hat{\pi}(Y))|X\right] = 0$). Although smoothness penalties are often used as a means of complexity regularization, in order to arrive at a consistent procedure one needs a way to control the influence of the penalty. Here we do not have such a control and hence the procedure will yield biased estimates even as the number of samples grows without a limit. Hence, we need to look for alternative ways to approximate the loss $L$.

A common suggestion is to use uncorrelated, or "double" samples with $\hat{L}_N$ as defined by (3). According to this proposal, for each state and action in the sample at least two next states should be generated (see e.g. Sutton and Barto, 1987, pp. 200). This is however ruled out by our assumption that we have a sample generated by a fixed policy. Another possibility, motivated by the double-sample proposal, would be to reuse samples that are close in space (e.g., use nearest neighbors). The difficulty with this approach is that it requires a definition of 'proximity'. Here, we pursue an alternative approach that avoids the need to define such a notion. The idea underlying our proposal is that if we knew $T^{\hat{\pi}} f$, we could use it to cancel the variance term that causes the problems. Hence, we propose to introduce an auxiliary function $h$ to be used as the approximation to $T^{\hat{\pi}} f$. Define

$$L(f, h; \hat{\pi}) = L(f; \hat{\pi}) - \left\| h - T^{\hat{\pi}} f \right\|_{\nu}^2 . \tag{4}$$

[1] In order to simplify the presentation we assume sufficient regularity of $\mathcal{F}$ so that we do not need to worry about the existence of a minimizer which can be guaranteed under fairly mild conditions, such as the compactness of $\mathcal{F}$ w.r.t. $\|\cdot\|_{\nu}$, or if $\mathcal{F}$ is finite dimensional (Cheney, 1966).

We propose to solve for

$$\hat{f} = \operatorname*{argmin}_{f\in\mathcal{F}^L} \sup_{h\in\mathcal{F}^L} L(f,h;\hat{\pi}). \tag{5}$$

where the supremum in $h$ comes from the negative sign of the term that involves $h$ (we want to make $\left\|h - T^{\hat{\pi}} f\right\|_\nu^2$ small, or $-\left\|h - T^{\hat{\pi}} f\right\|_\nu^2$ large).

Notice that $L(f,h;\hat{\pi})$ can be used to upper-bound $L(f;\hat{\pi})$ as follows: Let $h_f^* \in \mathcal{F}^L$ be a function that minimizes $\left\|h - T^{\hat{\pi}} f\right\|_\nu^2$. Then

$$L(f;\hat{\pi}) = L(f,h_f^*;\hat{\pi}) + \left\|h_f^* - T^{\hat{\pi}} f\right\|_\nu^2 .$$

Hence, if $\inf_{h\in\mathcal{F}^L} \left\|h - T^{\hat{\pi}} f\right\|_\nu^2$ is uniformly small (say, smaller than $\varepsilon > 0$) then we can guarantee that the solution of (5) will admit a small Bellman-residual, since

$$L(f;\hat{\pi}) \le L(f,h_f^*;\hat{\pi}) + \varepsilon.$$

We now argue that the sample-based approximation of $L(f,h;\hat{\pi})$, defined by

$$\begin{aligned}\hat{L}_N(f,h;\hat{\pi}) = \frac{1}{NL}\sum_{t=1}^{N}\sum_{j=1}^{L} \frac{\mathbb{I}_{\{A_t=a_j\}}}{\pi(a_j|X_t)} \Big( &(f(X_t,a_j) - (R_t + \gamma f(X_{t+1},\hat{\pi}(X_{t+1}))))^2 \\ &-(h(X_t,a_j) - (R_t + \gamma f(X_{t+1},\hat{\pi}(X_{t+1}))))^2 \Big)\end{aligned} \tag{6}$$

is unbiased, hence this loss-function overcomes the problem of the original Bellman-residual loss function. This key result of the paper is stated formally in the following lemma:

**Lemma 1 (Unbiased Loss Approximation)** *Assume that the behaviour policy $\pi$ samples all actions in all states with positive probability. Then for any $f$,$h \in \mathcal{F}^L$, policy $\hat{\pi}$, $\hat{L}_N(f,h;\hat{\pi})$ as defined by (6) provides an unbiased estimate to $L(f,h;\hat{\pi})$:*

$$\mathbb{E}\left[\hat{L}_N(f,h;\hat{\pi})\right] = L(f,h;\hat{\pi}). \tag{7}$$

*Proof* Let us define $C_{tj} = \frac{\mathbb{I}_{\{A_t=a_j\}}}{\pi(a_j|X_t)}$ and $\hat{Q}_{f,t} = R_t + \gamma f(X_{t+1},\hat{\pi}(X_{t+1}))$. Then, by (6), the $t$th term of $\hat{L}_N(f,h;\hat{\pi})$ can be written as

$$L^{(t)} = \frac{1}{L}\sum_{j=1}^{L} C_{tj} \left( (f_j(X_t) - \hat{Q}_{f,t})^2 - (h_j(X_t) - \hat{Q}_{f,t})^2 \right). \tag{8}$$

Note that

$$\begin{aligned}\mathbb{E}\left[C_{tj}\hat{Q}_{f,t}|X_t\right] &= \mathbb{E}\left[\hat{Q}_{f,t}|X_t, A_t = a_j\right] \\ &= r(X_t,a_j) + \gamma\int_y f(y,\hat{\pi}(y))\, dP(y|X_t,a_j) = (T^{\hat{\pi}} f)_j(X_t)\end{aligned} \tag{9}$$

and $\mathbb{E}\left[C_{tj}|X_t\right] = 1$ since all actions are sampled with positive probability in any state. Taking expectations,

$$\begin{aligned}\mathbb{E}\left[L^{(1)}\right] &= \mathbb{E}\left[\mathbb{E}\left[L^{(1)}\,|\,X_1\right]\right] \\ &= \frac{1}{L}\sum_{j=1}^{L}\mathbb{E}\left[\mathbb{E}\left[C_{1j}\left((f_j(X_1)-\hat{Q}_{f,1})^2-(h_j(X_1)-\hat{Q}_{f,1})^2\right)|X_1\right]\right].\end{aligned}$$

Now, by using the elementary identity $(a-x)^2-(b-x)^2 = (a-b)(a+b-2x)$ twice we get

$$\begin{aligned}&\mathbb{E}\left[C_{1j}\left((f_j(X_1)-\hat{Q}_{f,1})^2-(h_j(X_1)-\hat{Q}_{f,1})^2\right)|X_1\right] \\ &= \mathbb{E}\left[C_{1j}\left(f_j(X_1)-h_j(X_1)\right)\left(f_j(X_1)+h_j(X_1)-2\hat{Q}_{f,1}\right)|\,X_1\right] \\ &= \left(f_j(X_1)-h_j(X_1)\right)\left(f_j(X_1)+h_j(X_1)-2\mathbb{E}\left[C_{1j}\hat{Q}_{f,1}|X_1\right]\right) \\ &\qquad \text{(since } f_j(X_1), h_j(X_1) \text{ are } X_1-\text{measurable and } \mathbb{E}\left[C_{1j}|X_1\right]=1) \\ &= \left(f_j(X_1)-h_j(X_1)\right)\left(f_j(X_1)+h_j(X_1)-2(T^{\hat{\pi}}f)_j(X_1)\right) \quad \text{(by (9))} \\ &= \left(f_j(X_1)-(T^{\hat{\pi}}f)_j(X_1)\right)^2-\left(h_j(X_1)-(T^{\hat{\pi}}f)_j(X_1)\right)^2.\end{aligned}$$

Taking expectations of both sides we get that

$$\begin{aligned}\mathbb{E}\left[L^{(1)}\right] &= \frac{1}{L}\sum_{j=1}^{L}\left(\left\|f_j-(T^{\hat{\pi}}f)_j\right\|_\nu^2-\left\|h_j-(T^{\hat{\pi}}f)_j\right\|_\nu^2\right) \\ &= L(f;Q')-\left\|h-T^{\hat{\pi}}f\right\|_\nu^2 \\ &= L(f,h;Q').\end{aligned}$$

Because of stationarity this holds for $\mathbb{E}\left[L^{(t)}\right]$ for any $t$, thus finishing the proof of (7). □

It can be observed that the key step here is that the quadratic terms $\hat{Q}_{f,t}^2$ and $(T^{\hat{\pi}}f)_j^2$ are cancelled in the new loss functions (both in the sample based and the population based versions).

Hence we let *PEval* solve for

$$Q = \underset{f\in\mathcal{F}^L}{\operatorname{argmin}}\ \sup_{h\in\mathcal{F}^L}\hat{L}_N(f,h;\hat{\pi}). \tag{10}$$

Note that for linearly parameterized function classes the solution of the optimization problem (10) can be obtained in a closed form. In general, one may expect that the number of parameters doubles as a result of the introduction of the auxiliary function. Although this may represent a considerable additional computational burden on the algorithm, given the possible merits of the Bellman-residual minimization approach over the least-squares fixed point approach (cf. the discussion by Munos, 2003), we think that the extra effort may well pay off.

A slight optimization over (6) is to modify the term that includes $h$ by removing $R_t$. This helps one to reduce the variance of the samples that $h$ is regressed onto and hence we suspect that it helps increasing the efficiency of the procedure. The main statements of the paper remain essentially intact, with the constants appearing in the bound improving slightly. However, for the sake of keeping the paper compact we do not pursue this direction here.

## 4 Main Result

Before describing the main result we need some additional definitions.

We start with a mixing-property of stochastic processes. Informally, a process is mixing if future depends weakly on the past. The particular mixing concept we use here is called $\beta$-mixing:

**Definition 1 ($\beta$-mixing)** *Let $\{Z_t\}_{t=1,2,\ldots}$ be a stochastic process. Denote by $Z^{1:n}$ the collection $(Z_1,\ldots,Z_n)$, where we allow $n=\infty$. Let $\sigma(Z^{i:j})$ denote the sigma-algebra generated by $Z^{i:j}$ ($i \le j$). The $m$-th $\beta$-mixing coefficient of $\{Z_t\}$, $\beta_m$, is defined by*

$$\beta_m = \sup_{t\ge 1} \mathbb{E}\left[\sup_{B\in\sigma(Z^{t+m:\infty})} |P(B|Z^{1:t}) - P(B)|\right].$$

*A stochastic process is said to be $\beta$-mixing if $\beta_m \to 0$ as $m\to\infty$. In particular, we say that a $\beta$-mixing process mixes at an* exponential *rate with parameters $\overline{\beta}, b, \kappa > 0$ if $\beta_m \le \overline{\beta}\exp(-bm^{\kappa})$ holds for all $m \ge 0$.*

Note that besides $\beta$-mixing, many other definitions of mixing exist in the literature (see, e.g. Doukhan, 1994). The weakest among those most commonly used is called $\alpha$-mixing. Another commonly used one is $\phi$-mixing which is stronger than $\beta$-mixing (see Meyn and Tweedie, 1993).

Let us now state the main assumptions regarding the sample path:

**Assumption 2 (Sample Path Properties)** *Assume that*

$$\{(X_t, A_t, R_t)\}_{t=1,\ldots,N}$$

*is the sample path of $\pi$, a stochastic stationary policy. Further, assume that $\{X_t\}$ is strictly stationary ($X_t \sim \nu \in M(\mathcal{X})$) and exponentially $\beta$-mixing with the actual rate given by the parameters $(\overline{\beta}, b, \kappa)$. We further assume that the sampling policy $\pi$ satisfies $\pi_0 \stackrel{\text{def}}{=} \min_{a\in\mathcal{A}} \inf_{x\in\mathcal{X}} \pi(a|x) > 0$.*

The $\beta$-mixing property will be used to establish tail inequalities for certain empirical processes. Note that if $X_t$ is $\beta$-mixing then the hidden-Markov process $\{(X_t,(A_t,R_t))\}$ is also $\beta$-mixing with the same rate (see, e.g., the proof of Proposition 4 by Carrasco and Chen (2002) for an argument that can be used to prove this).

As discussed in the introduction, our bounds will depend on the average concentrability of the future-state distribution. This quantity that we introduce relates two distributions: $\nu$, the stationary distribution underlying

$\{X_t\}$, and $\rho$, the distribution used to assess the performance of the procedure (chosen by the user). It turns out that in the technique that we use to bound the final error as a function of the intermediate errors we need to change distributions between future state-distributions started from $\rho$ and $\nu$. An easy way to bound the amplification factor of changing from measure $\alpha$ to measure $\beta$ is to use the Radon-Nykodim derivative of $\alpha$ w.r.t. $\beta$. Denoting this derivative (density) by $d\alpha/d\beta$, we have that for any nonnegative measurable function $f$, $\int f\, d\alpha = \int f \frac{d\alpha}{d\beta}\, d\beta \leq \|\frac{d\alpha}{d\beta}\|_\infty \int f\, d\beta$. This motivates the following definition introduced in Munos and Szepesvári (2006):

**Definition 2 (Discounted-average Concentrability of Future-State Distribution)** *Given $\rho$, $\nu$, $m \geq 0$ and an arbitrary sequence of stationary policies $\{\pi_m\}_{m\geq 1}$, let*

$$c_{\rho,\nu}(m) = \sup_{\pi_1,\ldots,\pi_m} \left\| \frac{d(\rho P^{\pi_1} P^{\pi_2} \ldots P^{\pi_m})}{d\nu} \right\|_\infty, \tag{11}$$

*with the understanding that if the future state distribution $\rho P^{\pi_1} P^{\pi_2} \ldots P^{\pi_m}$ is not absolutely continuous w.r.t. $\nu$ then we take $c_{\rho,\nu}(m) = \infty$. The second-order discounted-average concentrability of future-state distributions is defined by*

$$C_{\rho,\nu} = (1-\gamma)^2 \sum_{m\geq 1} m\gamma^{m-1} c_{\rho,\nu}(m).$$

In general $c_{\rho,\nu}(m)$ diverges to infinity as $m \to \infty$. However, thanks to the discounting, $C_{\rho,\nu}$ will still be finite whenever $\gamma^m$ converges to zero faster than $c_{\rho,\nu}(m)$ converges to $\infty$. In particular, if the rate of divergence of $c_{\rho,\nu}(m)$ is sub-exponential, i.e., if $\Gamma = \limsup_{m\to\infty} 1/m \log c_{\rho,\nu}(m) \leq 0$ then $C_{\rho,\nu}$ will be finite. In the stochastic process literature, $\Gamma$ is called the top-Lyapunov exponent of the system and the condition $\Gamma \leq 0$ is interpreted as a stability condition. Hence, our condition on the finiteness of the discounted-average concentrability coefficient $C_{\rho,\nu}$ can also be interpreted as a stability condition. Further discussion of this concept and some examples of how to estimate $C_{\rho,\nu}$ for various system classes can be found in the report by Munos and Szepesvári (2006).

The concentrability coefficient $C_{\rho,\nu}$ will enter our bound on the weighted error of the algorithm. In addition to these weighted-error bounds, we shall also derive a bound on the $L^\infty$-norm error of the algorithm. This bound requires a stronger controllability assumption. In fact, the bound will depend on

$$C_\nu = \sup_{x\in\mathcal{X}, a\in\mathcal{A}} \frac{dP(\cdot|x,a)}{d\nu},$$

i.e., the supremum of the density of the transition kernel w.r.t. the state-distribution $\nu$. Again, if the system is "noisy" then $C_\nu$ is finite: In fact, the noisier is the dynamics (the less control we have), the smaller is $C_\nu$. As a side-note, let us remark that $C_{\rho,\nu} \leq C_\nu$ holds for any measures $\rho, \nu$. This follows directly from their respective definitions.

Our bounds also depend on the capacity of the function set $\mathcal{F}$. Let us now develop the necessary concepts. We assume that the reader is familiar with the concept of VC-dimension.[2] The VC-dimension of a set system $\mathcal{C}$ shall be denoted by $V_{\mathcal{C}}$. To avoid any confusions we introduce the definition of covering numbers:

**Definition 3 (Covering Numbers)** *Fix $\varepsilon > 0$ and a semi-metric space $\mathcal{M} = (\mathcal{M}, d)$. We say that $\mathcal{M}$ is covered by $m$ discs $D_1, \ldots, D_m$ if $\mathcal{M} \subset \cup_j D_j$. We define the* covering number *$\mathcal{N}(\varepsilon, \mathcal{M}, d)$ of $\mathcal{M}$ as the smallest integer $m$ such that $\mathcal{M}$ can be covered by $m$ discs each of which having a radius less than $\varepsilon$. If no such finite $m$ exists then we let $\mathcal{N}(\varepsilon, \mathcal{M}, d) = \infty$.*

In particular, for a class $\mathcal{F}$ of real-valued functions with domain $\mathcal{X}$ and points $x^{1:N} = (x_1, x_2, \ldots, x_N)$ in $\mathcal{X}$, we use the *empirical covering numbers*, i.e., the covering number of $\mathcal{F}$ equipped with the empirical $L^1$ semi-metric

$$l_{x^{1:N}}(f, g) = \frac{1}{N} \sum_{t=1}^{N} |f(x_t) - g(x_t)|.$$

In this case $\mathcal{N}(\varepsilon, \mathcal{F}, l_{x^{1:N}})$ shall be denoted by $\mathcal{N}_1(\varepsilon, \mathcal{F}, x^{1:N})$.

Another widely used capacity measure in the nonparametric statistics literature that we will need is the *pseudo-dimension* of function sets:

**Definition 4 (Pseudo-dimension)** *The* pseudo-dimension *$V_{\mathcal{F}^+}$ of $\mathcal{F}$ is defined as the* VC*-dimension of the subgraphs of functions in $\mathcal{F}$ (hence it is also called the* VC-subgraph dimension *of $\mathcal{F}$).*

In addition to the pseudo-dimension, we will need a new capacity concept:

**Definition 5 (**VC**-crossing Dimension)** *Let $\mathcal{C}_2 = \{\{x \in \mathcal{X} : f_1(x) \geq f_2(x)\} : f_1, f_2 \in \mathcal{F}\}$. The* VC-crossing dimension *of $\mathcal{F}$, denoted by $V_{\mathcal{F}^\times}$, is defined as the* VC*-dimension of $\mathcal{C}_2$: $V_{\mathcal{F}^\times} \stackrel{\text{def}}{=} V_{\mathcal{C}_2}$.*

The rationale of this definition is as follows: Remember that in the $k$th iteration of the algorithm we want to compute an approximate (action-value) evaluation of the policy greedy w.r.t. a previously computed action-value function $Q'$. Thus, in such a step we will jointly select $L$ functions (one for each action of $\mathcal{A}$) from $\mathcal{F}$ for a policy $\hat{\pi}$ that is greedy w.r.t. $Q'$ through (6) and (10). It follows that we will ultimately need a covering number bound for the set

$$\mathcal{F}_{\hat{\pi}}^{\vee} = \{\, f \,:\, f(\cdot) = Q(\cdot, \hat{\pi}(\cdot)) \text{ and } Q \in \mathcal{F}^L \,\}.$$

[2] Readers not familiar with VC-dimension are suggested to consult a book, such as the one by Anthony and Bartlett (1999).

Since $Q'$ depends on the data, the above set is random. In order to deal with this, we consider the following, non-random superset of $\mathcal{F}^{\vee}_{\hat{\pi}}$:

$$\begin{aligned}\mathcal{F}^{\vee} &= \bigcup_{Q' \in \mathcal{F}^L} \mathcal{F}^{\vee}_{\hat{\pi}(\cdot;Q')} \\ &= \{\, f \,:\, f(\cdot) = Q(\cdot, \hat{\pi}(\cdot)),\ \hat{\pi} = \hat{\pi}(\cdot; Q') \text{ and } Q, Q' \in \mathcal{F}^L \,\}.\end{aligned}$$

Ultimately, we will bound the estimation error of the procedure using the capacity of this class. Note that $\mathcal{F}^{\vee}$ can be written in the equivalent form:

$$\mathcal{F}^{\vee} = \left\{ \sum_{j=1}^{L} \mathbb{I}_{\{f_j(x) = \max_{1 \le k \le L} f_k(x)\}} g_j(x) \,:\, f_j, g_j \in \mathcal{F} \right\}$$

(with ties broken in a systematic, but otherwise arbitrary way). If we define the set of partitions of $\mathcal{X}$ induced by elements of $\mathcal{F}$ as

$$\Xi_{\mathcal{F},L} = \left\{ \xi : \xi = \{A_j\}_{j=1}^{L}, A_j \subset \mathcal{X}, x \in A_j \Leftrightarrow f_j(x) = \max_{1 \le k \le L} f_k(x), f_j \in \mathcal{F} \right\} \tag{12}$$

then we see that

$$\mathcal{F}^{\vee} = \left\{ \sum_{j=1}^{L} \mathbb{I}_{\{A_j\}}\, g_j \,:\, \{A_k\} = \xi \in \Xi_{\mathcal{F},L},\ g_j \in \mathcal{F} \right\}. \tag{13}$$

It turns out that the capacity of this class ultimately depends on the capacity (i.e., VC-dimension) of the set-system $\mathcal{C}_2$ defined above. The form (13) suggests to view the elements of the set $\mathcal{F}^{\vee}$ as regression trees defined by the partition system $\Xi_{\mathcal{F},L}$ and set $\mathcal{F}$. Actually, as the starting point for our capacity bounds we will use a result from the regression tree literature due to Nobel (1996).

Having introduced this new capacity measure, the first question is if it is really different from previous measures. The next lemma, listing basic properties of VC-crossing dimension answers this question affirmatively.

**Lemma 2 (Properties of VC-crossing Dimension)** *For any class $\mathcal{F}$ of $\mathcal{X} \to \mathbb{R}$ functions the following statements hold:*
*a) $V_{\mathcal{F}^+} \le V_{\mathcal{F}^\times}$. In particular, if $V_{\mathcal{F}^\times} < \infty$ then $V_{\mathcal{F}^+} < \infty$.*
*b) If $\mathcal{F}$ is a vector space then $V_{\mathcal{F}^+} = V_{\mathcal{F}^\times} = \dim(\mathcal{F})$. In particular, if $\mathcal{F}$ is a subset of a finite dimensional vector space then $V_{\mathcal{F}^\times} < \infty$.*
*c) There exists $\mathcal{F}$ with $V_{\mathcal{F}^\times} < \infty$ which is not a subset of any finite dimensional vector space.*
*d) There exists $\mathcal{F}$ with $\mathcal{X} = [0,1]$, $V_{\mathcal{F}^+} < \infty$ but $V_{\mathcal{F}^\times} = \infty$. In particular, there exists $\mathcal{F}$ with these properties such that the following properties also hold for $\mathcal{F}$:* (i) *$\mathcal{F}$ is countable,* (ii) $\{\{x \in \mathcal{X} : f(x) \ge a\} : f \in \mathcal{F}, a \in \mathbb{R}\}$ *is a* VC*-class (i.e., $\mathcal{F}$ is* VC-major *class),* (iii) *each $f \in \mathcal{F}$ is monotonous, bounded, and continuously differentiable with uniformly bounded derivatives.*

The proof of this lemma is given in the Appendix. Now we are ready to state our assumptions on the function set $\mathcal{F}$:

**Assumption 3 (Assumptions on the Function Set)** *Assume that* $\mathcal{F} \subset B(\mathcal{X}; Q_{\max})$ *for* $Q_{\max} > 0$ *and* $V_{\mathcal{F}^\times} < +\infty$.

Let us now turn to the definition of the quantities measuring the approximation power of $\mathcal{F}$. Like in regression, we need $\mathcal{F}$ to be sufficiently powerful in order to be able to approximate the evaluation functions of the policies encountered during the iterations closely. We shall define the approximation power of the function space in terms of two measures, its *inherent Bellman-error* and its *inherent one-step Bellman-error.*

The Bellman-error of an action-value function $Q$ w.r.t. a policy evaluation operator $T^{\hat{\pi}}$ is commonly defined as the supremum norm of the difference $Q - T^{\hat{\pi}}Q$ in analogy with the definition where the operators act on state-value functions. As it is widely known, if the Bellman-error is small then $Q$ is close to the fixed point of $T^{\hat{\pi}}$ thanks to $T^{\hat{\pi}}$ being a contraction. Hence, it is natural to expect that the final error of fitted policy iteration will be small if for all policies $\hat{\pi}$ encountered during the run of the algorithm, we can find some admissible action-value function $Q \in \mathcal{F}^L$ such that $Q - T^{\hat{\pi}}Q$ is small. For a fixed policy $\hat{\pi}$, the quantity

$$E_\infty(\mathcal{F}^L; \hat{\pi}) = \inf_{Q \in \mathcal{F}^L} \left\| Q - T^{\hat{\pi}}Q \right\|_\nu$$

can be used to measure the power of $\mathcal{F}$ in this respect. Since we do not know in advance which policies will be encountered during the execution of the algorithm, taking a pessimistic approach, we bound the final error in terms of

$$E_\infty(\mathcal{F}^L) \stackrel{\text{def}}{=} \sup_{Q' \in \mathcal{F}^L} E_\infty(\mathcal{F}^L; \hat{\pi}(\cdot; Q')),$$

called the *inherent Bellman-error of* $\mathcal{F}$. The subindex '$\infty$' is meant to convey the view that the fixed points of an operator can be obtained by repeating it an infinite number of times.

Another related quantity is the *inherent one-step Bellman-error of* $\mathcal{F}$. For a fixed policy $\hat{\pi}$, the one-step Bellman-error of $\mathcal{F}$ w.r.t. $T^{\hat{\pi}}$ is defined as the one-sided Hausdorff distance w.r.t. the $L^2(\nu)$-norm of the sets $\mathcal{F}^L$ and $T^{\hat{\pi}}\mathcal{F}^L$:

$$E_1(\mathcal{F}^L; \hat{\pi}) = d_\nu(T^{\hat{\pi}}\mathcal{F}^L, \mathcal{F}^L) \left( = \sup_{Q \in \mathcal{F}^L} \inf_{Q' \in \mathcal{F}^L} \left\| Q' - T^{\hat{\pi}}Q \right\|_\nu \right).$$

Taking again a pessimistic approach, the inherent one-step Bellman-error of $\mathcal{F}$ is defined as

$$E_1(\mathcal{F}^L) = \sup_{Q'' \in \mathcal{F}^L} E_1(\mathcal{F}^L; \hat{\pi}(\cdot; Q'')).$$

The rationale of the 'one-step' qualifier is that $T^{\hat{\pi}}$ is applied only once and then we look at how well the function in the resulting one-step image-space can be approximated by elements of $\mathcal{F}^L$.

The final error will actually depend on the squared sum of the inherent Bellman-error and the inherent one-step Bellman-error of $\mathcal{F}$:

$$E^2(\mathcal{F}^L) = E^2_\infty(\mathcal{F}^L) + E^2_1(\mathcal{F}^L),$$

$E(\mathcal{F}^L)$ is called the *total inherent Bellman-error* of $\mathcal{F}$. It is the additional term, $-\left\|h - T^{\hat{\pi}} f\right\|_\nu$ that we added in (4) to the unmodified Bellman-residual that causes the inherent one-step Bellman-error to enter our bounds.

We are now ready to give the main result of the paper:

**Theorem 3 (Finite-sample Error Bounds)** *Let $(\mathcal{X}, \mathcal{A}, P, S, \gamma)$ be a discounted MDP satisfying Assumption 1. In particular, let $R_{\max}$ denote a bound on the expected immediate rewards and let $\hat{R}_{\max}$ denote a bound on the random immediate rewards. Fix the set of admissible functions $\mathcal{F}$ satisfying Assumption 3 with $Q_{\max} \le R_{\max}/(1-\gamma)$. Consider the fitted policy iteration algorithm with the modified Bellman-residual minimization criterion defined by* (10) *and the input $\{(X_t, A_t, R_t)\}$, satisfying the mixing assumption, Assumption 2. Let $Q_k \in \mathcal{F}^L$ be the kth iterate ($k = -1, 0, 1, 2, \ldots$) and let $\pi_{k+1}$ be greedy w.r.t. $Q_k$. Choose $\rho \in M(\mathcal{X})$, a measure used to evaluate the performance of the algorithm and let $0 < \delta \le 1$. Then*

$$\|Q^* - Q^{\pi_K}\|_\rho \le$$
$$\frac{2\gamma}{(1-\gamma)^2}\left(C^{1/2}_{\rho,\nu}\left(E(\mathcal{F}^L) + \left(\frac{\Lambda_N(\frac{\delta}{K})\,(\Lambda_N(\frac{\delta}{K})/b \vee 1)^{1/\kappa}}{C_2 N}\right)^{1/4}\right) + \gamma^{K/2}\, R_{\max}\right) \tag{14}$$

*holds with probability at least $1-\delta$. Here $E(\mathcal{F}^L)$ is the total inherent Bellman-error of $\mathcal{F}$, $\Lambda_N(\delta)$ quantifies the dependence of the estimation error on $N$, $\delta$, and the capacity of the function set $\mathcal{F}$:*

$$\Lambda_N(\delta) = \tfrac{V}{2}\left(\log N + \log^+ C_2\right) + \log^+ C_1 + \log(e/\delta) + \log^+ \overline{\beta},$$

*$V$ being the "effective" dimension of $\mathcal{F}$:*

$$V = 3LV_{\mathcal{F}^+} + L_2 V_{\mathcal{F}^\times},$$

$L_2 = L(L-1)$,

$$\log C_1 = V\,\log\left(\frac{512eQ_{\max}\tilde{R}_{\max}}{L\pi_0}\right) + V_{\mathcal{F}^\times}\,L_2\log L_2 + V_{\mathcal{F}^+}\,L\log 2 + L^2$$
$$+ L_2\log(V_{\mathcal{F}^\times}+1) + L\log(V_{\mathcal{F}^+}+1) + 2\log(LV_{\mathcal{F}^+}+1) + 2\log(4e),$$

$$C_2 = \frac{1}{2}\left(\frac{L\pi_0}{32\tilde{R}^2_{\max}}\right)^2,$$

*and*

$$\tilde{R}_{\max} = (1+\gamma)Q_{\max} + \hat{R}_{\max}.$$

*Further, $\|Q^* - Q^{\pi_K}\|_\infty$ can be bounded with probability at least $1-\delta$ by a bound identical to* (14)*, except that $C^{1/2}_{\rho,\nu}$ has to be replaced by $C^{1/2}_\nu$.*

Before developing the proof, let us make some comments on the form of the bound (14). The bound has three terms, the first two of which are similar to terms that should be familiar from regression-estimation: In particular, the first term that depends on the total inherent Bellman-error of $\mathcal{F}$, $E(\mathcal{F}^L)$, quantifies the approximation power of $\mathcal{F}$ as discussed beforehand. The next term, apart from logarithmic and constant factors and terms and after some simplifications can be written in the form

$$\left(\frac{(V \log N + \log(K/\delta))^{1+1/\kappa}}{N}\right)^{1/4}.$$

This term bounds the estimation error. Note that the rate obtained (as a function of the number of samples, $N$) is worse than the best rates available in the regression literature. However, we think that this is only a proof artifact. Just like in regression, using a different proof technique (cf. Chapter 11 of Györfi et al., 2002), it seems possible to get a bound that scales with the reciprocal of the square-root of $N$, though this has the price that $E(\mathcal{F}^L)$ is replaced by $(1+\alpha)E(\mathcal{F}^L)$ with $\alpha > 0$. The last term does not have a counterpart in regression settings, as it is a bound on the error remaining after running the policy iteration algorithm for a finite number ($K$) of iterations. It can be readily observed that the optimal value of $K$ will depend amongst other factors on the capacity of the function set, the mixing rate, and the number of samples. However, it will not depend on the approximation-power of the function set.

Finally, let us comment on the multipliers of the bound. The multiplier $2\gamma/(1-\gamma)^2$ appears in previous $L^\infty$-performance bounds for policy iteration, too (cf. Bertsekas and Tsitsiklis, 1996b). As discussed previously, the concentrability coefficient, $C_{\rho,\nu}^{1/2}$, enters the bound due to the change-of-measure argument that we use when we propagate the error bounds through the iterations.

Note that a bound on the difference of the optimal action-value function, $Q^*$, and the action-value function of $\pi_K$, $Q^{\pi_K}$, does not immediately yield a bound on the difference of $V^*$ and $V^{\pi_K}$. However, with some additional work (by using similar techniques to the ones used below) it is possible to derive such a bound by starting with the point-wise bound

$$\begin{aligned}|V^* - V^{\pi_K}| \le E^{\pi^*}&(Q^* - Q^{\pi_{K-1}} + Q^{\pi_{K-1}} - Q_{K-1}) \\ &+ E^{\pi_K}(Q_{K-1} - Q^{\pi_{K-1}} + Q^{\pi_{K-1}} - Q^* + Q^* - Q^{\pi_K}),\end{aligned}$$

which follows by elementary arguments. For the sake of compactness this bound is not explored here in further details.

The following sections are devoted to develop the proof of the above theorem.

*4.1 Bounds on the Error of the Fitting Procedure*

The goal of this section is to derive a bound on the error introduced due to using a finite sample in the main optimization routine minimizing the (modified) sample-based Bellman-residual criterion defined by (6). If the samples were identically distributed and independent of each other, we could use the results developed for empirical processes (e.g. Pollard's inequality) to arrive at such a bound. However, since the samples are dependent these tools cannot be used. Instead, we will use the blocking device of Yu (1994). For simplicity assume that $N = 2m_N k_N$ for appropriate positive integers $m_N$, $k_N$ (the general case can be taken care of as was done by Yu, 1994). The technique of Yu partitions the samples into $2m_N$ blocks, each having $k_N$ samples. The samples in every second block are replaced by "ghost" samples whose joint marginal distribution is kept the same as that of the original samples (for the same block). However, these new random variables are constructed such that the new blocks are independent of each other. In order to keep the the flow of the developments continuous, the proofs of the statements of these results are given in the Appendix.

We start with the following lemma, which refines a previous result of Meir (2000):

**Lemma 4** *Suppose that $Z_0, \ldots, Z_N \in \mathcal{Z}$ is a stationary $\beta$-mixing process with mixing coefficients $\{\beta_m\}$, $Z'_t \in \mathcal{Z}$ $(t \in H)$ are the block-independent "ghost" samples as done by Yu (1994), and $H = \{2ik_N + j : 0 \le i < m_N, 1 \le j \le k_N\}$, and that $\mathcal{F}$ is a permissible class of $\mathcal{Z} \to [-K, K]$ functions. Then*

$$\mathbb{P}\left(\sup_{f\in\mathcal{F}}\left|\frac{1}{N}\sum_{t=1}^{N} f(Z_t) - \mathbb{E}\left[f(Z_0)\right]\right| > \varepsilon\right) \le 16\mathbb{E}\left[\mathcal{N}_1(\varepsilon/8, \mathcal{F}, (Z'_t; t \in H))\right] e^{-\frac{m_N \varepsilon^2}{128K^2}} + 2m_N\beta_{k_N+1}.$$

Note that this lemma is based on the following form of a lemma due to Yu (1994). This lemma is stated without a proof:[3]

**Lemma 5 (Yu, 1994, 4.2 Lemma)** *Suppose that $H_i = \{2k_N(i-1) + j : 1 \le j \le k_N\}$, $\{Z_t\}$, $\{Z'_t\}$, and $H = \bigcup_{i=1}^{m_N} H_i$ are as in Lemma 4, and that $\mathcal{F}$ is a permissible class of bounded $\mathcal{Z} \to \mathbb{R}$ functions. Then*

$$\mathbb{P}\left(\sup_{f\in\mathcal{F}}\left|\frac{1}{N}\sum_{t=1}^{N} f(Z_t)\right| > \varepsilon\right) \le 2\mathbb{P}\left(\sup_{f\in\mathcal{F}}\left|\frac{1}{N}\sum_{i=1}^{m_N}\sum_{t\in H_i} f(Z'_t)\right| > \frac{\varepsilon}{2}\right) + 2m_N\beta_{k_N+1}.$$

Let us now develop the tools used to bound the capacity of the function set of interest. For this, let $\Xi$ be a family of partitions of $\mathcal{X}$. By a partition

[3] Note that both Yu (1994) and Meir (2000) give a bound that contains $\beta_{k_N}$ instead of $\beta_{k_N+1}$ which we have here. Actually, a careful investigation of the original proof of Yu (1994) leads to the bound that is presented here.

of $\mathcal{X}$ we mean an ordered list of disjoint subsets of $\mathcal{X}$ whose union covers $\mathcal{X}$. Note that the empty set may enter multiple times the list. Following Nobel (1996), we define the *cell count* of a partition family $\Xi$ by

$$m(\Xi) = \max_{\xi \in \Xi} |\{A \in \xi : A \neq \emptyset\}|.$$

We will work with partition families that have finite cell counts. Note that we may always achieve that all partitions have the same number of cells by introducing the necessary number of empty sets. Hence, in what follows we will always assume that all partitions have the same number of elements. For $x^{1:N} \in \mathcal{X}^N$, let $\Delta(x^{1:N}, \Xi)$ be the number of distinct partitions (regardless the order) of $x^{1:N}$ that are induced by the elements of $\Xi$. The *partitioning number* of $\Xi$, $\Delta_N^*(\Xi)$, is defined as $\max\{\Delta(x^{1:N}, \Xi) : x^{1:N} \in \mathcal{X}^N\}$. Note that the partitioning number is a generalization of shatter-coefficient.

Given a class $\mathcal{G}$ of real-valued functions on $\mathcal{X}$ and a partition family $\Xi$ over $\mathcal{X}$, define the set of *$\Xi$-patched functions of $\mathcal{G}$* as follows:

$$\mathcal{G} \circ \Xi = \left\{ f = \sum_{A_j \in \xi} g_j \mathbb{I}_{\{A_j\}} : \xi = \{A_j\} \in \Xi, g_j \in \mathcal{G} \right\}.$$

Note that from this, (12) and (13), we have $\mathcal{F}^\vee = \mathcal{F} \circ \Xi_{\mathcal{F},L}$. We quote here a result of Nobel (with any domain $\mathcal{X}$ instead of $\mathbb{R}^s$ and with minimized premise):

**Proposition 6 (Nobel, 1996, Proposition 1)** *Let $\Xi$ be any partition family with $m(\Xi) < \infty$, $\mathcal{G}$ be a class of real-valued functions on $\mathcal{X}$, $x^{1:N} \in \mathcal{X}^N$. Let $\phi_N : \mathbb{R}^+ \to \mathbb{R}^+$ be a function that upper-bounds the empirical covering numbers of $\mathcal{G}$ on all subsets of the multi-set $[x_1, \ldots, x_N]$ at all scales:*

$$\mathcal{N}_1(\varepsilon, \mathcal{G}, A) \leq \phi_N(\varepsilon), \quad A \subset [x_1, \ldots, x_N], \varepsilon > 0.$$

*Then, for any $\varepsilon > 0$,*

$$\mathcal{N}_1(\varepsilon, \mathcal{G} \circ \Xi, x^{1:N}) \leq \Delta(x^{1:N}, \Xi)\phi_N(\varepsilon)^{m(\Xi)} \leq \Delta_N^*(\Xi)\phi_N(\varepsilon)^{m(\Xi)}. \quad (15)$$

In our next result we refine this bound by replacing the partitioning number by the covering number of the partition family:

**Lemma 7** *Let $\Xi$, $\mathcal{G}$, $x^{1:N}$, $\phi_N : \mathbb{R}^+ \to \mathbb{R}^+$ be as in Proposition 6. Moreover, let $\mathcal{G}$ be bounded: $\forall g \in \mathcal{G}$, $|g| \leq K$. For $\xi = \{A_j\}$, $\xi' = \{A'_j\} \in \Xi$, introduce the semi-metric*

$$d(\xi, \xi') = d_{x^{1:N}}(\xi, \xi') = \mu_N(\xi \bigtriangleup \xi'),$$

*where*

$$\xi \bigtriangleup \xi' = \{x \in \mathcal{X} : \exists j \neq j'; x \in A_j \cap A'_{j'}\} = \bigcup_{j=1}^{m(\Xi)} A_j \bigtriangleup A'_j,$$

*and where* $\mu_N$ *is the empirical measure corresponding to* $x^{1:N}$ *defined by* $\mu_N(A) = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}_{\{x_i \in A\}}$ *(here* $A$ *is any measurable subset of* $\mathcal{X}$*). Then, for any* $\varepsilon > 0$*,* $\alpha \in (0,1)$*,*

$$\mathcal{N}_1(\varepsilon, \mathcal{G} \circ \Xi, x^{1:N}) \leq \mathcal{N}\left(\frac{\alpha\varepsilon}{2K}, \Xi, d_{x^{1:N}}\right) \phi_N((1-\alpha)\varepsilon)^{m(\Xi)}.$$

Note that from this latter bound, provided that $\phi_N$ is left-continuous, the conclusion of Proposition 6 follows in the following limiting sense: Since $\mathcal{N}(\varepsilon, \Xi, d_{x^{1:N}}) \leq \Delta(x^{1:N}, \Xi)$ holds for any $\epsilon > 0$, we have

$$\mathcal{N}_1(\varepsilon, \mathcal{G} \circ \Xi, x^{1:N}) \leq \Delta(x^{1:N}, \Xi)\phi_N((1-\alpha)\varepsilon)^{m(\Xi)}.$$

Thus, letting $\alpha \to 0$ yields the bound (15).

Lemma 7 is used by the following result that develops a capacity bound on the function set of interest:

**Lemma 8** *Let* $\mathcal{F}$ *be a class of uniformly bounded functions on* $\mathcal{X}$ *(*$\forall f \in \mathcal{F}$*,* $|f| \leq K$*),* $x^{1:N} \in \mathcal{X}^N$*,* $\phi_N : \mathbb{R}^+ \to \mathbb{R}^+$ *be an upper-bound on the empirical covering numbers of* $\mathcal{F}$ *on all subsets of the multi-set* $[x_1, \ldots, x_N]$ *at all scales as in Proposition 6. Let* $\mathcal{G}_2^1$ *denote the class of indicator functions* $\mathbb{I}_{\{f_1(x) \geq f_2(x)\}} : \mathcal{X} \to \{0,1\}$ *for any* $f_1, f_2 \in \mathcal{F}$*. Then for* $\mathcal{F}^\vee$ *defined in* (13)*, for every* $\varepsilon > 0$*,* $\alpha \in (0,1)$*,*

$$\mathcal{N}(\varepsilon, \mathcal{F}^\vee, x^{1:N}) \leq \mathcal{N}_1\left(\frac{\alpha\varepsilon}{L(L-1)K}, \mathcal{G}_2^1, x^{1:N}\right)^{L(L-1)} \phi_N((1-\alpha)\varepsilon)^L.$$

We shall use the following lemma due to Haussler (1995) (see also, Anthony and Bartlett, 1999, Theorem 18.4) to bound the empirical covering numbers of our function sets in terms of their pseudo-dimensions:

**Proposition 9 (Haussler, 1995, Corollary 3)** *For any set* $\mathcal{X}$*, any points* $x^{1:N} \in \mathcal{X}^N$*, any class* $\mathcal{F}$ *of functions on* $\mathcal{X}$ *taking values in* $[0, K]$ *with pseudo-dimension* $V_{\mathcal{F}^+} < \infty$*, and any* $\varepsilon > 0$*,*

$$\mathcal{N}_1(\varepsilon, \mathcal{F}, x^{1:N}) \leq e(V_{\mathcal{F}^+} + 1)\left(\frac{2eK}{\varepsilon}\right)^{V_{\mathcal{F}^+}}.$$

Define

$$\tilde{E}_1^2(\mathcal{F}^L; \hat{\pi}) = E_1^2(\mathcal{F}^L; \hat{\pi}) - \inf_{f,h \in \mathcal{F}^L} \left\| h - T^{\hat{\pi}} f \right\|_\nu^2.$$

Certainly, $\tilde{E}_1^2(\mathcal{F}^L; \hat{\pi}) \leq E_1^2(\mathcal{F}^L; \hat{\pi})$. The following lemma is the main result of this section:

**Lemma 10** *Let Assumption 1 and 2 hold, and fix the set of admissible functions* $\mathcal{F}$ *satisfying Assumption 3. Let* $Q'$ *be a real-valued random function over* $\mathcal{X} \times \mathcal{A}$*,* $Q'(\omega) \in \mathcal{F}^L$ *(possibly not independent from the sample path). Let* $\hat{\pi} = \hat{\pi}(\cdot; Q')$ *be a policy that is greedy w.r.t.* $Q'$*. Let* $f'$ *be defined by*

$$f' = \operatorname*{argmin}_{f \in \mathcal{F}^L} \sup_{h \in \mathcal{F}^L} \hat{L}_N(f, h; \hat{\pi}).$$

*For* $0 < \delta \le 1$, $N \ge 1$, *with probability at least* $1-\delta$,

$$\left\|f' - T^{\hat{\pi}} f'\right\|_\nu^2 \le E_\infty^2(\mathcal{F}^L;\hat{\pi}) + \tilde{E}_1^2(\mathcal{F}^L;\hat{\pi}) + \sqrt{\frac{\Lambda_N(\delta)(\Lambda_N(\delta)/b \vee 1)^{1/\kappa}}{C_2 N}},$$

*where* $\Lambda_N(\delta)$ *and* $C_2$ *are defined as in Theorem 3. Further, the bound remains true if* $E_\infty^2(\mathcal{F}^L;\hat{\pi}) + \tilde{E}_1^2(\mathcal{F}^L;\hat{\pi})$ *above is replaced by* $E^2(\mathcal{F}^L)$.

By considering the case when $\gamma = 0$ and $L = 1$ we get an interesting side-result for regression function estimation (we use $r = r(x)$ since there are no actions):

**Corollary 11** *Let Assumption 1 hold. Assume that* $\{(X_t, R_t)\}_{t=1,\ldots,N}$ *is the sample path,* $\{X_t\}$ *is strictly stationary* ($X_t \sim \nu \in M(\mathcal{X})$) *and* $\beta$*-mixing with exponential rate* $(\overline{\beta}, b, \kappa)$. *Assume that* $\mathcal{F} \subset B(\mathcal{X}; Q_{\max})$ *for* $Q_{\max} \ge 0$ *and* $V_{\mathcal{F}^+} < \infty$. *Let* $f'$ *be defined by*

$$f' = \operatorname*{argmin}_{f\in\mathcal{F}} \frac{1}{N}\sum_{t=1}^{N} (f(X_t) - R_t)^2.$$

*Then, for* $0 < \delta \le 1$, $N \ge 1$, *with probability at least* $1-\delta$,

$$\|f' - r\|_\nu^2 \le \inf_{f\in\mathcal{F}} \|f - r\|_\nu^2 + \sqrt{\frac{\Lambda_N(\delta)(\Lambda_N(\delta)/b \vee 1)^{1/\kappa}}{C_2 N}},$$

*where* $\Lambda_N(\delta) = (V_{\mathcal{F}^+}/2 \vee 1)\log N + (V_{\mathcal{F}^+}/2)\log^+ C_2 + \log^+ C_1 + \log(e/\delta) + \log^+ \overline{\beta}$, $C_1 = 16e(V_{\mathcal{F}^+}+1)(128eQ_{\max}\tilde{R}_{\max})^{V_{\mathcal{F}^+}}$, $C_2 = \left(\frac{1}{32\tilde{R}_{\max}^2}\right)^2$, $\tilde{R}_{\max} = Q_{\max} + \hat{R}_{\max}$.

### *4.2 Propagation of Errors*

The main result of the previous section shows that if the approximation power of $\mathcal{F}$ is good enough and the number of samples is high then for any policy $\pi$ the optimization procedure will return a function $Q$ with small weighted error. Now, let $Q_0, Q_1, Q_2, \ldots$ denote the iterates returned by our algorithm, with $Q_{-1}$ being the initial action-value function:

$$Q_k = \operatorname*{argmin}_{Q\in\mathcal{F}^L} \sup_{h\in\mathcal{F}^L} \hat{L}_N(Q, h; \pi_k), \quad k = 0, 1, 2, \ldots,$$
$$\pi_k = \hat{\pi}(\cdot; Q_{k-1}), \quad k = 0, 1, 2, \ldots.$$

Further, let

$$\varepsilon_k = Q_k - T^{\pi_k} Q_k, \quad k = 0, 1, 2, \ldots \tag{16}$$

denote the Bellman-residual of the $k$th step. By the main result of the previous section, in any iteration step $k$ the optimization procedure will find with high probability a function $Q_k$ such that $\|\varepsilon_k\|_\nu^2$ is small. The purpose of this section is to bound the final error as a function of the intermediate errors. This is done in the following lemma without actually making any assumptions about how the sequence $Q_k$ is generated:

**Lemma 12** *Let $p \geq 1$, and let $K$ be a positive integer, $Q_{\max} \leq R_{\max}/(1-\gamma)$. Then, for any sequence of functions $\{Q_k\} \subset B(\mathcal{X}; Q_{\max})$, $0 \leq k < K$ and $\varepsilon_k$ defined by* (16) *the following inequalities hold:*

$$\|Q^* - Q^{\pi_K}\|_{p,\rho} \leq \frac{2\gamma}{(1-\gamma)^2} \left( C_{\rho,\nu}^{1/p} \max_{0 \leq k < K} \|\varepsilon_k\|_{p,\nu} + \gamma^{K/p} R_{\max} \right), \quad (17)$$

$$\|Q^* - Q^{\pi_K}\|_{\infty} \leq \frac{2\gamma}{(1-\gamma)^2} \left( C_{\nu}^{1/p} \max_{0 \leq k < K} \|\varepsilon_k\|_{p,\nu} + \gamma^{K/p} R_{\max} \right). \quad (18)$$

*Proof* We have $C_\nu \geq C_{\rho,\nu}$ for any $\rho$. Thus, if the bound (17) holds for any $\rho$, choosing $\rho$ to be a Dirac at each state implies that (18) also holds. Therefore, we only need to prove (17).

Let

$$E_k = P^{\pi_{k+1}}(I - \gamma P^{\pi_{k+1}})^{-1} - P^{\pi^*}(I - \gamma P^{\pi_k})^{-1}.$$

Closely following the proof of Lemma 4 in (Munos, 2003) we get

$$Q^* - Q^{\pi_{k+1}} \leq \gamma P^{\pi^*}(Q^* - Q^{\pi_k}) + \gamma E_k \epsilon_k.$$

Thus, by induction,

$$Q^* - Q^{\pi_K} \leq \gamma \sum_{k=0}^{K-1} (\gamma P^{\pi^*})^{K-k-1} E_k \epsilon_k + (\gamma P^{\pi^*})^K (Q^* - Q^{\pi_0}). \quad (19)$$

Now, let

$$F_k = P^{\pi_{k+1}}(I - \gamma P^{\pi_{k+1}})^{-1} + P^{\pi^*}(I - \gamma P^{\pi_k})^{-1}.$$

By taking the absolute value point-wise in (19) we get

$$Q^* - Q^{\pi_K} \leq \gamma \sum_{k=0}^{K-1} (\gamma P^{\pi^*})^{K-k-1} F_k |\varepsilon_k| + (\gamma P^{\pi^*})^K |Q^* - Q^{\pi_0}|.$$

From this, using the fact that $Q^* - Q^{\pi_0} \leq \frac{2}{1-\gamma} R_{\max} \mathbf{1}$, we arrive at

$$|Q^* - Q^{\pi_K}| \leq \frac{2\gamma(1-\gamma^{K+1})}{(1-\gamma)^2} \left[ \sum_{k=0}^{K-1} \alpha_k A_k |\varepsilon_k| + \alpha_K A_K R_{\max} \mathbf{1} \right]. \quad (20)$$

Here we introduced the positive coefficients

$$\alpha_k = \frac{(1-\gamma)\gamma^{K-k-1}}{1-\gamma^{K+1}}, \text{ for } 0 \leq k < K, \text{ and } \alpha_K = \frac{(1-\gamma)\gamma^K}{1-\gamma^{K+1}},$$

and the operators

$$A_k = \frac{1-\gamma}{2}(P^{\pi^*})^{K-k-1} F_k, \text{ for } 0 \leq k < K, \text{ and } A_K = (P^{\pi^*})^K.$$

Note that $\sum_{k=0}^{K} \alpha_k = 1$ and the operators $A_k$ are stochastic when considered as a right-linear operators. It is clear that $A_k$ are non-negative: $A_k Q \geq 0$

whenever $Q \geq 0$. It is also clear that $A_k$ are linear operators. It remains to see that they are stochastic, i.e., that $(A_k \mathbf{1})(x,a) = 1$ holds for all $(x,a) \in \mathcal{X} \times \mathcal{A}$. From the definition of $A_k$ it is easy to see that it suffices to check that $(1-\gamma)/2F_k$ is stochastic. For this, it suffices to notice that $(1-\gamma)(I - \gamma P^{\pi_{k+1}})^{-1}$ and $(1-\gamma)(I-\gamma P^{\pi_k})^{-1}$ are stochastic. This follows, however, by e.g. the Neumann-series expansion of these inverse operators. It is known that Jensen's inequality holds for stochastic operators: If $A$ is a stochastic operator and $g$ is a convex function then $g(A_k Q) \leq A_k(g \circ Q)$, where $g$ is applied point-wise, as is done the comparison between the two sides.

Let $\lambda_K = \left[\frac{2\gamma(1-\gamma^{K+1})}{(1-\gamma)^2}\right]^p$. Taking the $p$th power of both sides of (20), using Jensen's inequality twice and then integrating both sides w.r.t. $\rho(x,a)$ (with using $\rho$'s extension to $\mathcal{X} \times \mathcal{A}$ defined by (2)) we get

$$\begin{aligned}\|Q^* - Q^{\pi_K}\|_{p,\rho}^p &= \frac{1}{L} \sum_{a \in \mathcal{A}} \int \rho(dx)\, |Q^*(x,a) - Q^{\pi_K}(x,a)|^p \\ &\leq \lambda_K\, \rho \left[ \sum_{k=0}^{K-1} \alpha_k A_k |\varepsilon_k|^p + \alpha_K A_K (R_{\max})^p \mathbf{1} \right].\end{aligned}$$

From the definition of the coefficients $c_{\rho,\nu}(m)$,

$$\rho A_k \leq (1-\gamma) \sum_{m \geq 0} \gamma^m c_{\rho,\nu}(m+K-k)\nu$$

and hence

$$\begin{aligned}&\|Q^* - Q^{\pi_K}\|_{p,\rho}^p \\ &\quad \leq \lambda_K \left[ (1-\gamma) \sum_{k=0}^{K-1} \alpha_k \sum_{m \geq 0} \gamma^m c_{\rho,\nu}(m+K-k)\, \|\varepsilon_k\|_{p,\nu}^p + \alpha_K\, (R_{\max})^p \right].\end{aligned}$$

Let $\epsilon \stackrel{\text{def}}{=} \max_{0 \leq k < K} \|\varepsilon_k\|_{p,\nu}$. Using the definition of $\alpha_k$, $C_{\rho,\nu}$ and $\lambda_K$ we get

$$\begin{aligned}\|Q^* - Q^{\pi_K}\|_{p,\rho}^p &\leq \lambda_K \left[ \frac{1}{1-\gamma^{K+1}}\, C_{\rho,\nu}\, \epsilon^p + \frac{(1-\gamma)\gamma^K}{1-\gamma^{K+1}}\, (R_{\max})^p \right] \\ &\leq \lambda_K \left[ C_{\rho,\nu}\, \epsilon^p + \gamma^K\, (R_{\max})^p \right] \\ &\leq \left[\tfrac{2\gamma}{(1-\gamma)^2}\right]^p \left[ C_{\rho,\nu}\, \epsilon^p + \gamma^K\, (R_{\max})^p \right],\end{aligned}$$

leading to the desired bound:

$$\|Q^* - Q^{\pi_K}\|_{p,\rho} \leq \frac{2\gamma}{(1-\gamma)^2} C_{\rho,\nu}^{1/p}\, \epsilon + \gamma^{K/p}\, R_{\max}. \square$$

*4.3 Proof of the Main Result*

Now we are ready to prove Theorem 3.

*Proof* As in the case of the previous proof, we only need to prove the statement for the weighted $\rho$-norm.

Fix $N, K > 0$, and let $\rho$ and $\mathcal{F}$ be as in the statement of Theorem 3. Consider the iterates $Q_k$ generated by model-free policy iteration with *PEval* defined by (10), when running on the trajectory $\{(X_t, A_t, R_t)\}$ generated by some stochastic stationary policy $\pi$. Let $\nu$ be the invariant measure underlying the stationary process $\{X_t\}$. Let $\pi_K$ be a policy greedy w.r.t. $Q_K$. Our aim is to derive a bound on the distance of $Q^{\pi_K}$ and $Q^*$. For this, we use Lemma 12. Indeed, if one defines $\varepsilon_k = Q_k - T^{\pi_k} Q_k$ then by Lemma 12 with $p = 2$,

$$\|Q^* - Q^{\pi_K}\|_\rho \le \frac{2\gamma}{(1-\gamma)^2} \left( C_{\rho,\nu}^{1/2} \max_{0 \le k < K} \|\varepsilon_k\|_\nu + \gamma^{K/2} R_{\max} \right). \tag{21}$$

Now, from Lemma 10, we conclude that for any fixed integer $0 \le k < K$ and for any $\delta' > 0$,

$$\|\varepsilon_k\|_\nu \le E(\mathcal{F}^L) + \left( \frac{\Lambda_N(\delta')\,(\Lambda_N(\delta')/b \vee 1)^{1/\kappa}}{C_2 N} \right)^{1/4} \tag{22}$$

holds everywhere except on a set of probability at most $\delta'$. ($\Lambda_N(\delta')$ and $C_2$ are defined as in the theorem.) Take $\delta' = \delta/K$. By the choice of $\delta'$, the total probability of the set of exceptional events for $0 \le k < K$ is at most $\delta$. Outside of this failure set, we have that Equation (22) holds for all $0 \le k < K$. Combining this with (21), we get

$$\|Q^* - Q^{\pi_K}\|_\rho \le$$

$$\frac{2\gamma}{(1-\gamma)^2} \left( C_{\rho,\nu}^{1/2} \left( E(\mathcal{F}^L) + \left( \frac{\Lambda_N(\frac{\delta}{K}) \left( \frac{\Lambda_N(\frac{\delta}{K})}{b} \vee 1 \right)^{1/\kappa}}{C_2 N} \right)^{1/4} \right) + \gamma^{\frac{K}{2}} R_{\max} \right),$$

thus finishing the proof of the weighted-norm bound. □

## 5 Related Work

The idea of using value function approximation goes back to the early days of dynamic programming (Samuel, 1959; Bellman and Dreyfus, 1959). With the recent growth of interest in reinforcement learning, work on value function approximation methods flourished (Bertsekas and Tsitsiklis, 1996a; Sutton and Barto, 1998). Recent theoretical results mostly concern supremum-norm approximation errors (Gordon, 1995; Tsitsiklis and Van Roy, 1996), where the main condition on the way intermediate iterates are mapped (projected) to the function space is that the corresponding operator, $\Pi$, must be a non-expansion. Practical examples when $\Pi$ satisfies the said property include certain kernel-based methods, see e.g. the works by Gordon (1995); Tsitsiklis and Van Roy (1996); Guestrin et al. (2001); Ernst et al. (2005). However, the restriction imposed on $\Pi$ rules out many popular algorithms, such as regression-based approaches that were found, however, to behave well in practice (e.g. Wang and Dietterich, 1999; Dietterich and Wang, 2002; Lagoudakis and Parr, 2003). The need for analyzing the behaviour of such algorithms provided the basic motivation for this work.

To the best of our knowledge there are no previous theoretical results on the finite-sample performance of off-policy control-learning algorithms for infinite horizon problems that use function-approximation and learn from a single trajectory. In fact, the only paper where finite-sample bounds are derived in an off-policy setting and which uses function approximators is the paper by Murphy (2005) who considered fitted Q-iteration in *finite-horizon*, undiscounted problems. A major relief that comes from the finite-horizon assumption is that the training data consists of multiple *independent* trajectories. As a result the samples for any *fixed* stage are independent of each other. Proceeding backwards via a stage-wise analysis it is then possible to eliminate the complications resulting from working with dependent samples completely.

Another interesting theoretical development concerning off-policy control learning with value-function approximation is the paper by Ormoneit and Sen (2002) who considered kernel-regression in conjunction with Q-learning and obtained asymptotic rates on weak-convergence. Q-learning with interpolative function approximation was considered by Szepesvári and Smart (2004), where only asymptotic convergence and performance bounds were given. Both these works carry out the analysis with respect to the $L^\infty$ norm and exploit that the function-approximation operator $\Pi$ is a non-expansion. Precup et al. (2001) considers the use of likelihood ratios to evaluate policies and arrive at asymptotic convergence results, though only for policy evaluation.

As to the methods, the closest to the present work is the paper of Szepesvári and Munos (2005). However, unlike there here we dealt with a fitted policy iteration algorithm and worked with dependent samples and a single sample-path. All these resulted in a much complex analysis and the need to develop new tools: For dealing with dependant data, we used the blocking device originally proposed by Yu (1994). We had to introduce a new capacity concept to deal with the complications arising from the use of policy iteration. The error propagation technique used in Section 4.2 is an extension of a similar technique due to Munos (2003). However, whilst the analysis in Munos (2003) was restricted to the case when the transition probability kernel is point-wise absolute continuous w.r.t. the stationary distribution of the states (i.e., under the assumption $C_\nu < +\infty$), here the analysis was carried out under a weaker condition (namely, $C_{\rho,\nu} < \infty$). Although this condition was studied earlier by Szepesvári and Munos (2005), but only for analyzing approximate value iteration.

## 6 Conclusions and Future Work

We have considered fitted policy iteration with Bellman-residual minimization. To our best knowledge this is the first theoretical paper where high-probability finite-sample bounds are derived on the performance of a reinforcement learning algorithm for infinite-horizon control learning in an off-policy setting, using function approximators over a continuous state-space. In order to derive our results we had to introduce a novel sample-based approximation to the Bellman-residual criterion, a capacity concept, deal with dependent samples and work out a method to propagate weighted norm errors in a policy iteration setting. Our main result quantifies the dependency of the final error on the number of samples, the mixing rate of the process, the average-discounted concentrability of the future-state distribution, the number of iterations, the capacity and the approximation power of the function set used in the embedded least-squares problem.

Although we believe that the present work represents a significant step towards understanding what makes efficient reinforcement learning possible, it appears that much remains to be done.

Although we made some initial steps towards finding out the properties of VC-crossing dimensions, bounds on the VC-crossing dimension of popular function classes, such as regression trees or neural networks are yet to be seen. The present work also leaves open the question of how to design appropriate function sets that have controlled capacity but large approximation power. When the MDP is noisy and the dynamics is "smooth" then it is known that the class of value functions of all stationary policies will be uniformly smooth. Hence, for such MDPs, at least in theory, as the sample size growth to infinity by choosing a sequence of increasing function sets whose union covers the space of smooth functions (like in the method of sieves in regression) it is possible to recover the optimal policy with the presented method. One open question is how to design a method that adaptively chooses the function set so as to fit the actual smoothness of the system. One idea, borrowed from the regression literature, is to use penalized least-squares. It remains to be seen if this method is indeed capable to achieve adaptation to unknown smoothness.

Another possibility is to use different function sets for the representation of the fixed point candidates and the auxiliary function candidates, or in the successive iterations of the algorithm. How to choose these function sets? Also, at many points in the analysis we took a pessimistic approach (e.g. in the derandomization of $\mathcal{F}_{\hat{\pi}}^{\vee}$ or when bounding the approximation error). It might be possible to improve our bounds by a great extent by avoiding these pessimistic steps.

One major challenge is to extend our results to continuous action spaces as the present analysis heavily builds on the finiteness of the action set.

It would also be desirable to remove the condition that the function set must admit a bounded envelope. One idea is to use the truncation technique of Chapter 11 by (Györfi et al., 2002) for this purpose. The technique presented there could also be used to try to improve the rate of our current estimate. Borrowing further ideas from the regression literature, it might be possible to achieve even greater improvement by, e.g., using localization techniques or data-dependent bounds.

Although in this paper we considered Bellman-residual minimization, the techniques developed could be applied to least-squares fixed point approximation based approaches such as the LSPI algorithm of Lagoudakis and Parr (2003), or least-squares fitted Q-iteration considered recently by Ernst et al. (2005). Another direction is to relax the condition that the states are observable. Indeed, this assumption can be lifted easily since the algorithm never works directly with the states. The assumption that the trajectory is sufficient representative certainly fails when the behaviour policy does not sample all actions with positive probability in all states. Still, the result can be extended to this case, but the statement has to be modified appropriately since it is clear that in this case convergence to near-optimality cannot be guaranteed.

Finally, it would be interesting to compare the result that we obtained with $\gamma = 0$ and $L = 1$ for the regression-case (Corollary 11) with similar results available in the regression literature. In connection to this, let us remark that our method applies and can be used to derive bounds to the solution of inverse problems of the form $Pf = r$, $f =?$ with $P$ being a stochastic operator and when the data consists of samples from $r$ and $P$.

## Appendix

### *6.1 Proofs of the Auxiliary Lemmata*

Proof of Lemma 2. a) Since $V_{\mathcal{F}^+}$ is the VC-dimension of the subgraphs of functions in $\mathcal{F}$, there exist $V_{\mathcal{F}^+}$ points, $z_1,\ldots,z_{V_{\mathcal{F}^+}}$ in $\mathcal{X} \times \mathbb{R}$ that are shattered by these subgraphs (see, e.g., Devroye et al., 1996 or Anthony and Bartlett, 1999). This can happen only if the projections, $x_1,\ldots,x_{V_{\mathcal{F}^+}}$, of these points to $\mathcal{X} \times \{0\}$ are all distinct. Now, for any $A \subseteq \{x_1,\ldots,x_{V_{\mathcal{F}^+}}\}$, there is an $f_1 \in \mathcal{F}$ such that $f_1(x_i) > z_i$ for $x_i \in A$ and $f_1(x_i) \le z_i$ for $x_i \notin A$, and also there is an $f_2 \in \mathcal{F}$ such that $f_2(x_i) \le z_i$ for $x_i \in A$ and $f_2(x_i) > z_i$ for $x_i \notin A$. That is, $f_1(x_i) > f_2(x_i)$ for $x_i \in A$ and $f_1(x_i) < f_2(x_i)$ for $x_i \notin A$. Thus, the set in $\mathcal{C}_2$ corresponding to $(f_1, f_2)$ contains exactly the same $x_i$'s as $A$ does. This means that $x_1,\ldots,x_{V_{\mathcal{F}^+}}$ is shattered by $\mathcal{C}_2$, that is, $V_{\mathcal{F}^\times} = V_{\mathcal{C}_2} \ge V_{\mathcal{F}^+}$. The second part of the statement is obvious.

b) According to Theorem 11.4 of Anthony and Bartlett (1999), $V_{\mathcal{F}^+} = \dim(\mathcal{F})$. On the other hand, since now for $f_1,f_2 \in \mathcal{F}$ also $f_1 - f_2 \in \mathcal{F}$, it is easy to see that $\mathcal{C}_2 = \{\{x \in \mathcal{X} : f(x) \ge 0\} : f \in \mathcal{F}\}$. By taking

$g \equiv 0$ in Theorem 3.5 of Anthony and Bartlett (1999), we get the desired $V_{\mathcal{F}^\times} = V_{\mathcal{C}_2} = \dim(\mathcal{F})$. The second statement follows obviously.

c) Let $\mathcal{F} = \{\mathbb{I}_{\{(a,\infty)\}} : a \in \mathbb{R}\}$. Then $V_{\mathcal{F}^\times} = 2$ and $\mathcal{F}$ generates an infinite dimensional vector space.

d) Let $\mathcal{X} = [0,1]$. Let $\{a_j\}$ be monotonously decreasing with $\sum_{j=1}^{\infty} a_j = 1$, $0 \le a_j \le 1/\log_2 j$, and $3a_{j+1} > a_j$. For an integer $n \ge 2$, let $k \ge 1$ and $0 \le i \le 2^k - 1$ be the unique integers defined by $n = 2^k + i$. Define

$$f_n(x) = x + \sum_{j=1}^{n} a_j \qquad \text{and}$$
$$\tilde{f}_n(x) = x + \sum_{j=1}^{n} a_j + \frac{a_n}{4}(-1)^{\lfloor i/2^{\lfloor kx \rfloor} \rfloor} \sin^2(k\pi x).$$

Certainly, $f_n$ and $\tilde{f}_n$ are both differentiable. Note that $a_n \le a_{2^k} \le 1/k$, thus the gradient of the last term of $\tilde{f}_n(x)$ is bounded in absolute value by $k\pi/(4k) < 1$. Hence the functions $\tilde{f}_n$ (and obviously $f_n$) are strictly monotonously increasing, and have range in $[0,2]$. Let $\mathcal{F}_1 = \{f_n : n \ge 2\}$, $\tilde{\mathcal{F}}_1 = \{\tilde{f}_n : n \ge 2\}$, and $\mathcal{F} = \mathcal{F}_1 \cup \tilde{\mathcal{F}}_1$. $\mathcal{F}$ is certainly countable. By the monotonicity of $f_n$ and $\tilde{f}_n$, the VC-dimension of $\{\{x \in \mathcal{X} : f(x) \ge a\} : f \in \mathcal{F}, a \in \mathbb{R}\}$ is 1. Observe that the sequence $f_n$ is point-wise monotonously increasing also in $n$, and this remains true also for $\tilde{f}_n$, since the last modifying term is negligible (less than $a_n/4$ in absolute value). (Moreover, for any $n$,$n'$, $n > n'$, $f_n > \tilde{f}_{n'}$ and $\tilde{f}_n > f_{n'}$ everywhere.) This point-wise monotonicity implies that $V_{\mathcal{F}_1^+} = V_{\tilde{\mathcal{F}}_1^+} = 1$, and thus $V_{\mathcal{F}^+} \le 3$. On the other hand, since $\{x \in \mathcal{X} : \tilde{f}_n(x) \ge f_n(x)\} = \{x \in \mathcal{X} : (-1)^{\lfloor i/2^{\lfloor kx \rfloor} \rfloor} \ge 0\} = \{x \in \mathcal{X} : \lfloor i/2^{\lfloor kx \rfloor} \rfloor \text{ is even}\}$, $\mathcal{C}_2 \supseteq \{\{x \in \mathcal{X} : \tilde{f}_n(x) \ge f_n(x)\} : n \ge 2\} = \{\{x \in \mathcal{X} : \lfloor i/2^{\lfloor kx \rfloor} \rfloor \text{ is even}\} : n \ge 2\}$, and this class contains the unions of $\{1\}$ and any of the intervals $\{[0,1/k), [1/k,2/k), \ldots [1-1/k,1)\}$ for any $k$. Thus it shatters the points $\{0, 1/k, 2/k, \ldots 1-1/k\}$, hence $V_{\mathcal{F}^\times} = V_{\mathcal{C}_2} = \infty$. □

Proof of Lemma 4. Define the block-wise functions $\bar{f} : \mathcal{Z}^{k_N} \to \mathbb{R}$ as

$$\bar{f}(z^{1:k_N}) = \bar{f}(z_1, \ldots, z_{k_N}) \stackrel{\text{def}}{=} \sum_{t=1}^{k_N} f(z_t)$$

for $f \in \mathcal{F}$ and $z^{1:k_N} = (z_1, \ldots, z_{k_N})$ and let $\bar{\mathcal{F}} \stackrel{\text{def}}{=} \{\bar{f} : f \in \mathcal{F}\}$.

We use Lemma 5 of Yu to replace the original process by the block-independent one, implying

$$\begin{aligned}
&\mathbb{P}\left(\sup_{f\in\mathcal{F}}\left|\frac{1}{N}\sum_{t=1}^{N} f(Z_t)-\mathbb{E}\left[f(Z_0)\right]\right|>\varepsilon\right)\\
&=\mathbb{P}\left(\sup_{f\in\mathcal{F}}\left|\frac{1}{N}\sum_{t=1}^{N}(f(Z_t)-\mathbb{E}\left[f(Z_0)\right])\right|>\varepsilon\right)\\
&\le 2\mathbb{P}\left(\sup_{f\in\mathcal{F}}\left|\frac{1}{N}\sum_{i=1}^{m_N}(\bar{f}(Z'^{(i)})-k_N\mathbb{E}\left[f(Z_0)\right])\right|>\frac{\varepsilon}{2}\right)+2m_N\beta_{k_N+1}\\
&=2\mathbb{P}\left(\sup_{f\in\mathcal{F}}\left|\frac{1}{m_N}\sum_{i=1}^{m_N}\bar{f}(Z'^{(i)})-k_N\mathbb{E}\left[f(Z_0)\right]\right|>k_N\varepsilon\right)+2m_N\beta_{k_N+1}. \qquad (23)
\end{aligned}$$

Here $Z'^{(i)} \stackrel{\text{def}}{=} \{Z'_t\}_{t\in H_i} = (Z'_{2k_N(i-1)+1},\ldots,Z'_{2k_N(i-1)+k_N})$.

Now, since any $\bar{f} \in \bar{\mathcal{F}}$ is bounded by $k_N K$, Pollard's inequality (cf. Pollard, 1984) applied to the independent blocks implies the bound

$$\begin{aligned}
&\mathbb{P}\left(\sup_{f\in\mathcal{F}}\left|\frac{1}{m_N}\sum_{i=1}^{m_N}\bar{f}(Z'^{(i)})-k_N\mathbb{E}\left[f(Z_0)\right]\right|>k_N\varepsilon\right)\\
&\qquad\le 8\mathbb{E}\left[\mathcal{N}_1(k_N\varepsilon/8,\bar{\mathcal{F}},(Z'^{(1)},\ldots,Z'^{(m_N)}))\right]e^{-\frac{m_N\varepsilon^2}{128K^2}}. \qquad (24)
\end{aligned}$$

Following Lemma 5.1 by Meir (2000) (or the proof of part (i) of 4.3 Lemma of Yu (1994)), we get that for any $f,\tilde{f}\in\mathcal{F}$, the distance of $\bar{f}$ and $\bar{\tilde{f}}$ can be bounded as follows:

$$\begin{aligned}
\frac{1}{m_N}\sum_{i=1}^{m_N}|\bar{f}(Z'^{(i)})-\bar{\tilde{f}}(Z'^{(i)})| &= \frac{1}{m_N}\sum_{i=1}^{m_N}\left|\sum_{t\in H_i} f(Z'_t)-\sum_{t\in H_i}\tilde{f}(Z'_t)\right|\\
&\le \frac{1}{m_N}\sum_{i=1}^{m_N}\sum_{t\in H_i}|f(Z'_t)-\tilde{f}(Z'_t)|\\
&= \frac{k_N}{N/2}\sum_{t\in H}|f(Z'_t)-\tilde{f}(Z'_t)|,
\end{aligned}$$

implying[4]

$$\mathcal{N}_1(k_N\varepsilon/8,\bar{\mathcal{F}},(Z'^{(1)},\ldots,Z'^{(m_N)}))\le\mathcal{N}_1(\varepsilon/8,\mathcal{F},(Z'_t;t\in H)).$$

This, together with (23) and (24) gives the desired bound. □

[4] Note that neither Meir (2000), nor Yu (1994) exploit that it is enough to use half of the ghost samples in the upper bound above. Also Meir (2000) makes a slight mistake of considering $(Z'_t; t \in H)$ below as having $N$ (instead of $N/2$) variables.

PROOF OF LEMMA 7. Fix $x_1, \ldots, x_N \in \mathcal{X}$ and $\varepsilon > 0$. Let $\widehat{\Xi}$ be an $\alpha\varepsilon/(2K)$-cover for $\Xi$ according to $d$ such that $|\widehat{\Xi}| = \mathcal{N}\left(\frac{\alpha\varepsilon}{2K}, \Xi, d\right)$. If $f \in \mathcal{G} \circ \Xi$, then there is a partition $\xi = \{A_j\} \in \Xi$ and functions $g_j \in \mathcal{G}$ such that

$$f = \sum_{A_j \in \xi} g_j \mathbb{I}_{\{A_j\}}. \tag{25}$$

Let $\xi' \in \widehat{\Xi}$ such that $d(\xi, \xi') < \frac{\alpha\varepsilon}{2K}$, and let $f' = \sum_{A'_j \in \xi'} g_j \mathbb{I}_{\{A'_j\}}$. Then

$$\begin{aligned}
&\frac{1}{N}\sum_{i=1}^{N} |f(x_i) - f'(x_i)| \\
&= \frac{1}{N}\sum_{i=1}^{N} \left| \sum_{A_j \in \xi} g_j(x_i)\mathbb{I}_{\{x_i \in A_j\}} - \sum_{A'_j \in \xi'} g_j(x_i)\mathbb{I}_{\{x_i \in A'_j\}} \right| \\
&= \frac{1}{N}\sum_{i: x_i \in \xi \triangle \xi'} \left| \sum_{A_j \in \xi} g_j(x_i)\mathbb{I}_{\{x_i \in A_j\}} - \sum_{A'_j \in \xi'} g_j(x_i)\mathbb{I}_{\{x_i \in A'_j\}} \right| \\
&\le \frac{2K}{N} |\{i : x_i \in \xi \triangle \xi'\}| = 2K d(\xi, \xi') \\
&< \alpha\varepsilon.
\end{aligned}$$

Let $\mathcal{F}_j$ be an $(1-\alpha)\varepsilon$-cover for $\mathcal{G}$ on $\widehat{A}_j = \{x_1, \ldots, x_N\} \cap A'_j$ such that $|\mathcal{F}_j| \le \phi_N((1-\alpha)\varepsilon)$. To each function $g_j$ appearing in (25) there corresponds an approximating function $f_j \in \mathcal{F}_j$ such that

$$\frac{1}{N_j} \sum_{x_i \in \widehat{A}_j} |g_j(x_i) - f_j(x_i)| < (1-\alpha)\varepsilon,$$

where $N_j = |\widehat{A}_j|$. If we define $f'' = \sum_{A'_j \in \xi'} f_j \mathbb{I}_{\{A'_j\}}$, then it is easy to see that

$$\frac{1}{N}\sum_{i=1}^{N} |f'(x_i) - f''(x_i)| < (1-\alpha)\varepsilon.$$

Hence

$$\frac{1}{N}\sum_{i=1}^{N} |f(x_i) - f''(x_i)| < \varepsilon.$$

When the functions $f_j \in \mathcal{F}_j$ are suitably chosen, every function $\tilde{f} \in \mathcal{G} \circ \Xi$ defined in terms of a partition closer to $\xi'$ than $\varepsilon$ in $d$-metric can be approximated by a similar estimate $f''$. Thus the collection of all such functions $\tilde{f}$ can be covered on $x^{1:N}$ by no more than $\prod_{j=1}^{|\xi'|} |\mathcal{F}_j| \le \phi_N((1-\alpha)\varepsilon)^{|\xi'|}$ approximating functions. As $\xi'$ is chosen from $\mathcal{N}\left(\frac{\alpha\varepsilon}{2K}, \Xi, d\right)$ partitions, the result follows. $\square$

Proof of Lemma 8. Since $\mathcal{F}^\vee = \mathcal{F} \circ \Xi$ for $\Xi = \Xi_{\mathcal{F},L}$ defined in (12),

$$\mathcal{N}_1(\varepsilon, \mathcal{F}^\vee, x^{1:N}) = \mathcal{N}_1(\varepsilon, \mathcal{F} \circ \Xi, x^{1:N}).$$

We apply Lemma 7 to bound this by

$$\mathcal{N}\left(\frac{\alpha\varepsilon}{2K}, \Xi, d_{x^{1:N}}\right) \phi_N((1-\alpha)\varepsilon)^L,$$

where $\mathcal{N}(\varepsilon, \Xi, d_{x^{1:N}})$ is the $\varepsilon$-covering number of $\Xi$ regarding the metric $d_{x^{1:N}}$ defined in Lemma 7.

For $f : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$ ($f \in \mathcal{F}^L$), define the indicator function $I_f : \mathcal{X} \times \mathcal{A} \to \{0, 1\}$

$$I_f(x, a) = \mathbb{I}_{\{\max_{a' \in \mathcal{A}} f(x,a') = f(x,a)\}}$$

(ties should be broken in an arbitrary systematic way) and their class $\mathcal{G} = \{I_f : f \in \mathcal{F}^L\}$.

Now the distance $d_{x^{1:N}}$ of two partitions in $\Xi$ is $L/2$-times the $L^1$-distance of the corresponding two indicator functions in $\mathcal{G}$ regarding to the empirical measure supported on the $NL$ points $x^{1:N} \times \mathcal{A}$. Hence the metric $d_{x^{1:N}}$ on $\Xi$ corresponds to this $L^1$-metric on $\mathcal{G}$. So

$$\mathcal{N}(\varepsilon, \Xi, d_{x^{1:N}}) = \mathcal{N}_1\left(\frac{2\varepsilon}{L}, \mathcal{G}, x^{1:N} \times \mathcal{A}\right).$$

Furthermore, if $\mathcal{G}_L^1$ denotes the class of indicator functions $\mathbb{I}_{\{\max_{a' \in \mathcal{A}} f(x,a') = f_1(x)\}} : \mathcal{X} \to \{0, 1\}$ for any $f : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$ ($f \in \mathcal{F}^L$), then, since the support of a function from $\mathcal{G}$ is the disjoint union of the supports (on different instances of $\mathcal{X}$) of $L$ functions from $\mathcal{G}_L^1$, it is easy to see that (cf., e.g., Devroye *et al.* Devroye et al. (1996, Theorem 29.6))

$$\mathcal{N}_1(\varepsilon, \mathcal{G}, x^{1:N} \times \mathcal{A}) \le \mathcal{N}_1(\varepsilon, \mathcal{G}_L^1, x^{1:N})^L.$$

Now, since a function from $\mathcal{G}_L^1$ is the product of $L-1$ indicator functions from $\mathcal{G}_2^1$, it is easy to see that (cf., e.g., the generalization of Devroye et al., 1996, Theorem 29.7, Pollard, 1990)

$$\mathcal{N}_1(\varepsilon, \mathcal{G}_L^1, x^{1:N}) \le \mathcal{N}_1\left(\frac{\varepsilon}{L-1}, \mathcal{G}_2^1, x^{1:N}\right)^{L-1}.$$

The equations above together give the bound of the lemma. □

We shall need the following technical lemma in the next proof:

**Lemma 13** *Let* $\beta_m \le \overline{\beta} \exp(-b m^\kappa)$, $N \ge 1$, $k_N = \lceil (C_2 N \varepsilon^2 / b)^{\frac{1}{1+\kappa}} \rceil$, $m_N = N/(2k_N)$, $0 < \delta \le 1$, $V \ge 2$, *and* $C_1$, $C_2$, $\overline{\beta}$, $b$, $\kappa > 0$. *Further define* $\varepsilon$ *and* $\Lambda$ *by*

$$\varepsilon = \sqrt{\frac{\Lambda(\Lambda/b \vee 1)^{1/\kappa}}{C_2 N}} \tag{26}$$

*with* $\Lambda = (V/2)(\log N + \log^+ C_2) + \log^+ C_1 + \log(e/\delta) + \log^+ \overline{\beta}$. *Then*

$$C_1 \left(\frac{1}{\varepsilon}\right)^V e^{-4 C_2 m_N \varepsilon^2} + 2 m_N \beta_{k_N} < \delta.$$

PROOF OF LEMMA 13. We have

$$\max((C_2N\varepsilon^2/b)^{\frac{1}{1+\kappa}}, 1) \le k_N \le \max(2(C_2N\varepsilon^2/b)^{\frac{1}{1+\kappa}}, 1)$$

and so

$$\frac{N}{4}\min\left(\frac{b}{C_2N\varepsilon^2}, 1\right)^{\frac{1}{1+\kappa}} \le \frac{N}{4}\min\left(\left(\frac{b}{C_2N\varepsilon^2}\right)^{\frac{1}{1+\kappa}}, 2\right) \le m_N = \frac{N}{2k_N} \le \frac{N}{2}.$$

Obviously, $\Lambda \ge 1$ and from (26),

$$\varepsilon \ge \sqrt{\Lambda/(C_2N)} \ge \sqrt{1/(C_2N)} \qquad \text{and} \qquad C_2N\varepsilon^2 = \Lambda(\Lambda/b \vee 1)^{1/\kappa}. \quad (27)$$

Substituting the proper bounds for $\beta_m$, $k_N$, and $m_N$, we get

$$\begin{aligned}
&C_1\left(\frac{1}{\varepsilon}\right)^V e^{-4C_2m_N\varepsilon^2} + 2m_N\beta_{k_N} \\
&\le C_1\left(\frac{1}{\varepsilon}\right)^V e^{-(\frac{b}{C_2N\varepsilon^2}\wedge 1)^{\frac{1}{1+\kappa}}C_2N\varepsilon^2} + N\overline{\beta}e^{-b(\frac{C_2N\varepsilon^2}{b}\vee 1)^{\frac{\kappa}{1+\kappa}}} \\
&= C_1\left(\frac{1}{\varepsilon}\right)^V e^{-(\frac{b}{C_2N\varepsilon^2}\wedge 1)^{\frac{1}{1+\kappa}}C_2N\varepsilon^2} + N\overline{\beta}e^{-b(\frac{C_2N\varepsilon^2}{b}\vee 1)(\frac{b}{C_2N\varepsilon^2}\wedge 1)^{\frac{1}{1+\kappa}}} \\
&\le \left(C_1\left(\frac{1}{\varepsilon}\right)^V + N\overline{\beta}\right) e^{-(\frac{b}{C_2N\varepsilon^2}\wedge 1)^{\frac{1}{1+\kappa}}C_2N\varepsilon^2},
\end{aligned}$$

which, by (27), is upper bounded by

$$\left(C_1(C_2N)^{V/2} + N\overline{\beta}\right) e^{-(\frac{b}{\Lambda(\Lambda/b\vee 1)^{1/\kappa}}\wedge 1)^{\frac{1}{1+\kappa}}\Lambda(\Lambda/b\vee 1)^{1/\kappa}}.$$

It is easy to check that the exponent of $e$ in the last factor is just $-\Lambda$. Thus, substituting $\Lambda$, this factor is $N^{-V/2}e^{-(V/2)\log^+ C_2}\delta/(e(\overline{\beta}\vee 1)(C_1\vee 1))$, and our bound becomes

$$\begin{aligned}
&\left((C_2N)^{V/2} + N\right) N^{-V/2}e^{-(V/2)\log^+ C_2}\frac{\delta}{e} \\
&\le \left(\left(e^{\log C_2 - \log^+ C_2}\right)^{V/2} + e^{-(V/2)\log^+ C_2}\right)\frac{\delta}{e} \\
&\le (1+1)\frac{\delta}{e} \\
&< \delta. \square
\end{aligned}$$

*6.2 Proof of Lemma 10*

*Proof* Recall that (see the proof of Lemma 1) $\hat{Q}_{f,t} = R_t + \gamma f(X_{t+1}, \hat{\pi}(X_{t+1}))$, and that, for fixed, deterministic $f$ and $\hat{\pi}$,

$$\mathbb{E}\left[\hat{Q}_{f,t}|X_t, A_t\right] = (T^{\hat{\pi}} f)(X_t, A_t),$$

that is, $T^{\hat{\pi}} f$ is the regression function of $\hat{Q}_{f,t}$ given $(X_t, A_t)$. What we have to show is that the chosen $f'$ is close to the corresponding $T^{\hat{\pi}(\cdot;Q')} f'$ with high probability, noting that $Q'$ may not be independent from the sample path.

We can assume that $|\mathcal{F}| \geq 2$ (otherwise the bound is obvious). This implies $V_{\mathcal{F}^+}, V_{\mathcal{F}^\times} \geq 1$, and thus $V \geq L(L+2) \geq 3$. Let $\varepsilon$ and $\Lambda_N(\delta)$ be chosen as in (26):

$$\varepsilon = \sqrt{\frac{\Lambda_N(\delta)(\Lambda_N(\delta)/b \vee 1)^{1/\kappa}}{C_2 N}}$$

with $\Lambda_N(\delta) = (V/2)(\log N + \log^+ C_2) + \log^+ C_1 + \log(e/\delta) + \log^+ \overline{\beta} \geq 1$. Define

$$P_0 \stackrel{\text{def}}{=} \mathbb{P}\left(\left\|f' - T^{\hat{\pi}(\cdot;Q')} f'\right\|_\nu^2 - E_\infty^2(\mathcal{F}^L; \hat{\pi}) - \tilde{E}_1^2(\mathcal{F}^L; \hat{\pi}) > \varepsilon\right).$$

It follows that it is sufficient to prove that $P_0 < \delta$.

Remember that for $\hat{\pi}$ arbitrary, we defined the following losses:

$$\begin{aligned} L(f; \hat{\pi}) &= \left\|f - T^{\hat{\pi}} f\right\|_\nu^2, \\ L(f, h; \hat{\pi}) &= L(f; \hat{\pi}) - \left\|h - T^{\hat{\pi}} f\right\|_\nu^2. \end{aligned}$$

Let us now introduce the following additional shorthand notations:

$$\begin{aligned} L(f; Q') &= L(f; \hat{\pi}(\cdot; Q')), \\ L(f, h; Q') &= L(f, h; \hat{\pi}(\cdot; Q')), \\ \hat{L}_N(f, h; Q') &= \hat{L}_N(f, h; \hat{\pi}(\cdot; Q')) \end{aligned}$$

where $\hat{L}_N$ was defined in (3). Further, define

$$\bar{L}(f; Q') \stackrel{\text{def}}{=} \sup_{h \in \mathcal{F}^L} L(f, h; Q') = L(f; Q') - \inf_{h \in \mathcal{F}^L} \left\|h - T^{\hat{\pi}} f\right\|_\nu^2,$$

Now,

$$
\begin{aligned}
&\left\|f' - T^{\hat\pi(\cdot;Q')} f'\right\|_\nu^2 - E_\infty^2(\mathcal{F}^L;\hat\pi) - \tilde{E}_1^2(\mathcal{F}^L;\hat\pi) \\
&= L(f';Q') - \inf_{f\in\mathcal{F}^L} L(f;Q') - \tilde{E}_1^2(\mathcal{F}^L;\hat\pi) \\
&= \bar{L}(f';Q') + \inf_{h\in\mathcal{F}^L} \left\|h - T^{\hat\pi} f'\right\|_\nu^2 \\
&\qquad - \inf_{f\in\mathcal{F}^L} \left(\bar{L}(f;Q') + \inf_{h\in\mathcal{F}^L} \left\|h - T^{\hat\pi} f\right\|_\nu^2\right) - \tilde{E}_1^2(\mathcal{F}^L;\hat\pi) \\
&\le \bar{L}(f';Q') + \inf_{h\in\mathcal{F}^L} \left\|h - T^{\hat\pi} f'\right\|_\nu^2 \\
&\qquad - \inf_{f\in\mathcal{F}^L} \bar{L}(f;Q') - \inf_{f,h\in\mathcal{F}^L} \left\|h - T^{\hat\pi} f\right\|_\nu^2 - \tilde{E}_1^2(\mathcal{F}^L;\hat\pi) \\
&= \bar{L}(f';Q') - \bar{L}_{\mathcal{F},Q'} + \inf_{h\in\mathcal{F}^L} \left\|h - T^{\hat\pi} f'\right\|_\nu^2 - \sup_{f\in\mathcal{F}^L} \inf_{h\in\mathcal{F}^L} \left\|h - T^{\hat\pi} f\right\|_\nu^2 \\
&\le \bar{L}(f';Q') - \bar{L}_{\mathcal{F},Q'},
\end{aligned}
$$

where $\bar{L}_{\mathcal{F},Q'} = \inf_{f\in\mathcal{F}^L} \bar{L}(f;Q')$ is the error of the function with minimum loss in our class. Define also

$$
\bar{\hat{L}}_N(f;Q') \stackrel{\text{def}}{=} \sup_{h\in\mathcal{F}^L} \hat{L}_N(f,h;Q').
$$

Now, since $f' = \operatorname{argmin}_{f\in\mathcal{F}^L} \bar{\hat{L}}_N(f;Q')$,

$$
\begin{aligned}
&\bar{L}(f';Q') - \bar{L}_{\mathcal{F},Q'} \\
&\qquad = \bar{L}(f';Q') - \bar{\hat{L}}_N(f';Q') + \bar{\hat{L}}_N(f';Q') - \inf_{f\in\mathcal{F}^L} \bar{L}(f;Q') \\
&\qquad \le |\bar{\hat{L}}_N(f';Q') - \bar{L}(f';Q')| + \inf_{f\in\mathcal{F}^L} \bar{\hat{L}}_N(f;Q') - \inf_{f\in\mathcal{F}^L} \bar{L}(f;Q') \\
&\qquad \quad \text{(by the definition of } f'\text{)} \\
&\qquad \le 2 \sup_{f\in\mathcal{F}^L} |\bar{\hat{L}}_N(f;Q') - \bar{L}(f;Q')| \\
&\qquad = 2 \sup_{f\in\mathcal{F}^L} |\sup_{h\in\mathcal{F}^L} \hat{L}_N(f,h;Q') - \sup_{h\in\mathcal{F}^L} L(f,h;Q')| \\
&\qquad \le 2 \sup_{f,h\in\mathcal{F}^L} |\hat{L}_N(f,h;Q') - L(f,h;Q')| \\
&\qquad \le 2 \sup_{Q',f,h\in\mathcal{F}^L} |\hat{L}_N(f,h;Q') - L(f,h;Q')|.
\end{aligned}
$$

Thus we get

$$
P_0 \le \mathbb{P}\left(\sup_{Q',f,h\in\mathcal{F}^L} |\hat{L}_N(f,h;Q') - L(f,h;Q')| > \varepsilon/2\right).
$$

Hence, in the subsequent statements, $Q'$ denotes an arbitrary (deterministic) function in $\mathcal{F}^L$.

We follow the line of proof due to Meir (2000). For any $f$,$h$,$Q' \in \mathcal{F}^L$, define the loss function $l_{f,h,Q'} : \mathcal{X} \times \mathcal{A} \times [-\hat{R}_{\max}, \hat{R}_{\max}] \times \mathcal{X} \to \mathbb{R}$ in accordance with (6) as

$$\begin{aligned} l_{f,h,Q'}(z) &= l_{f,h,Q'}(x,a,r,y) \\ &\stackrel{\text{def}}{=} \frac{1}{L} \sum_{j=1}^{L} \frac{\mathbb{I}_{\{a=a_j\}}}{\pi(a_j|x)} \Big( |f_j(x) - r - \gamma f(y, \hat{\pi}(y;Q'))|^2 \\ &\qquad - |h_j(x) - r - \gamma f(y, \hat{\pi}(y;Q'))|^2 \Big) \end{aligned}$$

for $z = (x,a,r,y)$ and $\mathcal{L}_{\mathcal{F}} \stackrel{\text{def}}{=} \{l_{f,h,Q'} : f,h,Q' \in \mathcal{F}^L\}$. Introduce $Z_t = (X_t, A_t, R_t, X_{t+1})$ for $t = 1,\ldots,N$. Note that the process $\{Z_t\}$ is $\beta$-mixing with mixing coefficients $\{\beta_{m-1}\}$.

Observe that by (8)

$$l_{f,h,Q'}(Z_t) = \frac{1}{L} \sum_{j=1}^{L} \frac{\mathbb{I}_{\{A_t=a_j\}}}{\pi(a_j|X_t)} ((f_j(X_t) - \hat{Q}_{f,t})^2 - (h_j(X_t) - \hat{Q}_{f,t})^2) = L^{(t)},$$

hence we have for any $f$,$h$,$Q' \in \mathcal{F}^L$

$$\frac{1}{N} \sum_{t=1}^{N} l_{f,h,Q'}(Z_t) = \hat{L}_N(f,h;Q'),$$

and

$$\mathbb{E}\left[l_{f,h,Q'}(Z_t)\right] = \mathbb{E}\left[L^{(t)}\right] = L(f,h;Q')$$

(coincidently with (7), but note that $\mathbb{E}\left[\hat{\bar{L}}_N(f;Q')\right] \neq \bar{L}(f;Q')$). This reduces the bound to a uniform tail probability of an empirical process over $\mathcal{L}_{\mathcal{F}}$:

$$P_0 \leq \mathbb{P}\left( \sup_{Q',f,h \in \mathcal{F}^L} \left| \frac{1}{N} \sum_{t=1}^{N} l_{f,h,Q'}(Z_t) - \mathbb{E}\left[l_{f,h,Q'}(Z_1)\right] \right| > \varepsilon/2 \right).$$

Since the samples are correlated, Pollard's tail inequality cannot be used directly. Hence we use the method of Yu (1994), as suggested beforehand. For this we split the $N$ samples into $2m_N$ blocks that come in pairs (for simplicity we assume that splitting can be done exactly), i.e., $N = 2m_N k_N$. Introduce the following blocks, each having the same length, $k_N$:

$$\underbrace{Z_1,\ldots,Z_{k_N}}_{H_1}, \underbrace{Z_{k_N+1},\ldots,Z_{2k_N}}_{T_1}, \underbrace{Z_{2k_N+1},\ldots,Z_{3k_N}}_{H_2}, \underbrace{Z_{3k_N+1},\ldots,Z_{4k_N}}_{T_2}, \ldots$$
$$\ldots, \underbrace{Z_{(2m_N-2)k_N+1},\ldots,Z_{(2m_N-1)k_N}}_{H_{m_N}}, \underbrace{Z_{(2m_N-1)k_N+1},\ldots,Z_{2m_N k_N}}_{T_{m_N}}.$$

Here $H_i \stackrel{\text{def}}{=} \{2k_N(i-1)+1, \ldots, 2k_N(i-1)+k_N\}$ and $T_i \stackrel{\text{def}}{=} \{2ik_N-(k_N-1), \ldots, 2ik_N\}$. Next, we introduce the block-independent "ghost" samples as it was done by Yu (1994) and Meir (2000):

$$\underbrace{Z'_1, \ldots, Z'_{k_N}}_{H_1}, \quad \underbrace{Z'_{2k_N+1}, \ldots, Z'_{3k_N}}_{H_2}, \quad \ldots \quad \underbrace{Z'_{(2m_N-2)k_N+1}, \ldots, Z'_{(2m_N-1)k_N}}_{H_{m_N}},$$

where any particular block has the same marginal distribution as originally, but the $m_N$ blocks are independent of one another. Introduce $H = \bigcup_{i=1}^{m_N} H_i$.

For this ansatz we use Lemma 4 above with $\mathcal{Z} = \mathcal{X} \times \mathcal{A} \times \mathbb{R} \times \mathcal{X}$, $\mathcal{F} = \mathcal{L}_\mathcal{F}$ noting that any $l_{f,h,Q'} \in \mathcal{L}_\mathcal{F}$ is bounded by

$$K = \frac{\tilde{R}^2_{\max}}{L\pi_0}$$

with $\tilde{R}_{\max} = (1+\gamma)Q_{\max} + \hat{R}_{\max}$, to get the bound

$$\mathbb{P}\left(\sup_{Q',f,h \in \mathcal{F}^L} \left| \frac{1}{N} \sum_{t=1}^{N} l_{f,h,Q'}(Z_t) - \mathbb{E}\left[l_{f,h,Q'}(Z_1)\right] \right| > \varepsilon/2 \right)$$
$$\le 16\mathbb{E}\left[\mathcal{N}_1(\varepsilon/16, \mathcal{L}_\mathcal{F}, (Z'_t; t \in H))\right] e^{-\frac{m_N}{2}\left(\frac{L\pi_0\varepsilon}{16\tilde{R}^2_{\max}}\right)^2} + 2m_N\beta_{k_N}.$$

By some calculation, the distance in $\mathcal{L}_\mathcal{F}$ can be bounded as follows:

$$\frac{2}{N}\sum_{t\in H} |l_{f,h,Q'}(Z'_t) - l_{g,\tilde{h},\tilde{Q}'}(Z'_t)|$$
$$\le \frac{2\tilde{R}_{\max}}{L\pi_0}\left( \frac{2}{N}\sum_{t\in H} |f(X'_t, A'_t) - g(X'_t, A'_t)| + \frac{2}{N}\sum_{t\in H} |\tilde{h}(X'_t, A'_t) - h(X'_t, A'_t)| \right.$$
$$\left. + 2\frac{2}{N}\sum_{t\in H} |f(X'_{t+1}, \hat{\pi}(X'_{t+1}; Q')) - g(X'_{t+1}, \hat{\pi}(X'_{t+1}; \tilde{Q}'))| \right).$$

Note that the first and second terms are $\mathcal{D}' = ((X'_t, A'_t); t \in H)$-based $L^1$-distances of functions in $\mathcal{F}^L$, while the last term is just twice the $\mathcal{D}'_+ = (X'_{t+1}; t \in H)$-based $L^1$-distance of two functions in $\mathcal{F}^\vee$ corresponding to $(f, Q')$ and $(g, \tilde{Q}')$. This leads to

$$\mathcal{N}_1\left(\frac{8\tilde{R}_{\max}}{L\pi_0}\varepsilon', \mathcal{L}_\mathcal{F}, (Z'_t; t \in H)\right) \le \mathcal{N}_1^2(\varepsilon', \mathcal{F}^L, \mathcal{D}')\mathcal{N}_1(\varepsilon', \mathcal{F}^\vee, \mathcal{D}'_+).$$

Applying now Lemma 8 with $\alpha = 1/2$,[5] the covering number of $\mathcal{F}^\vee$ is bounded by

$$\mathcal{N}_1\left(\frac{\varepsilon'}{2L_2Q_{\max}}, \mathcal{G}_2^1, \mathcal{D}'_+\right)^{L_2} \phi_{N/2}(\varepsilon'/2)^L,$$

[5] The optimal choice $\alpha = V_{\mathcal{F}^\times}/(V_{\mathcal{F}^\times} + V_{\mathcal{F}^+}/(L-1))$ would give slightly better constants.

where $L_2 = L(L-1)$, $\mathcal{G}_2^1$ is the class of the indicator functions of the sets from $\mathcal{C}_2$, and the empirical covering numbers of $\mathcal{F}$ on all subsets of $\mathcal{D}'_+$ are majorized by $\phi_{N/2}(\cdot)$.

To bound these factors, we use Corollary 3 from Haussler (1995) that was cited here as Proposition 9. The pseudo-dimensions of $\mathcal{F}$ and $\mathcal{G}_2^1$ are $V_{\mathcal{F}^+}$, $V_{\mathcal{F}^\times} < \infty$, respectively, and the range of functions from $\mathcal{F}$ has length $2Q_{\max}$. By the pigeonhole principle, it is easy to see that that the pseudo-dimension of $\mathcal{F}^L$ cannot exceed $LV_{\mathcal{F}^+}$. Thus

$$\begin{aligned}
\mathcal{N}_1\left(\frac{8\tilde{R}_{\max}}{L\pi_0}\varepsilon', \mathcal{L}_{\mathcal{F}}, (Z'_t; t \in H)\right) &\le \left(e(LV_{\mathcal{F}^+}+1)\left(\frac{4eQ_{\max}}{\varepsilon'}\right)^{LV_{\mathcal{F}^+}}\right)^2 \\
&\cdot \left(e(V_{\mathcal{F}^\times}+1)\left(\frac{4eL_2Q_{\max}}{\varepsilon'}\right)^{V_{\mathcal{F}^\times}}\right)^{L_2} \left(e(V_{\mathcal{F}^+}+1)\left(\frac{8eQ_{\max}}{\varepsilon'}\right)^{V_{\mathcal{F}^+}}\right)^L \\
&= e^{L^2+2}(LV_{\mathcal{F}^+}+1)^2(V_{\mathcal{F}^+}+1)^L(V_{\mathcal{F}^\times}+1)^{L_2}2^{LV_{\mathcal{F}^+}}L_2^{L_2V_{\mathcal{F}^\times}}\left(\frac{4eQ_{\max}}{\varepsilon'}\right)^V,
\end{aligned}$$

where $V = 3LV_{\mathcal{F}^+} + L_2V_{\mathcal{F}^\times}$ is the "effective" dimension, and thus

$$\begin{aligned}
\mathcal{N}_1(\varepsilon/16, \mathcal{L}_{\mathcal{F}}, (Z'_t; t \in H)) &\le e^{L^2+2}(LV_{\mathcal{F}^+}+1)^2(V_{\mathcal{F}^+}+1)^L(V_{\mathcal{F}^\times}+1)^{L_2} \cdot \\
&\cdot 2^{LV_{\mathcal{F}^+}}L_2^{L_2V_{\mathcal{F}^\times}}\left(\frac{512eQ_{\max}\tilde{R}_{\max}}{L\pi_0\varepsilon}\right)^V = \frac{C_1}{16}\left(\frac{1}{\varepsilon}\right)^V,
\end{aligned}$$

with $C_1 = C_1(L, V_{\mathcal{F}^+}, V_{\mathcal{F}^\times}, Q_{\max}, \hat{R}_{\max}, \gamma, \pi_0)$. It can be easily checked that $\log C_1$ matches the corresponding expression given in the text of the theorem.

Putting together the above bounds we get

$$P_0 \le C_1\left(\frac{1}{\varepsilon}\right)^V e^{-4C_2m_N\varepsilon^2} + 2m_N\beta_{k_N}, \tag{28}$$

where $C_2 = \frac{1}{2}\left(\frac{L\pi_0}{32\tilde{R}_{\max}^2}\right)^2$. Defining $k_N = \lceil (C_2N\varepsilon^2/b)^{\frac{1}{1+\kappa}} \rceil$ and $m_N = N/(2k_N)$, the proof is finished by Lemma 13, which, together with (28), implies $P_0 < \delta$.

The last statement follows obviously from $Q' \in \mathcal{F}^L$ and the definitions of $E(\mathcal{F}^L)$, $E_\infty(\mathcal{F}^L)$, $E_1(\mathcal{F}^L)$, and $\tilde{E}_1(\mathcal{F}^L; \hat{\pi})$. $\square$

## 7 Acknowledgements

We would like to acknowledge support for this project from the Hungarian National Science Foundation (OTKA), Grant No. T047193 (Cs. Szepesvári) and from the Hungarian Academy of Sciences (Cs. Szepesvári, Bolyai Fellowship and A. Antos, Bolyai Fellowship). We also would like to thank Balázs Csanád Csáji, György András, Levente Kocsis, and Rich Sutton for the friendly discussions that helped to improve the paper to a great extent.

## References


Anthony, M. and P. L. Bartlett: 1999, *Neural Network Learning: Theoretical Foundations*. Cambridge University Press.

Baraud, Y., F. Comte, and G. Viennet: 2001, 'Adaptive estimation in autoregression or $\beta$-mixing regression via model selection'. *Annals of Statistics* **29**, 839–875.

Bellman, R. and S. Dreyfus: 1959, 'Functional Approximation and Dynamic Programming'. *Math. Tables and other Aids Comp.* **13**, 247–251.

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

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

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

Carrasco, M. and X. Chen: 2002, 'Mixing and moment properties of various GARCH and stochastic volatility models'. *Econometric Theory* **18**, 17–39.

Cheney, E.: 1966, *Introduction to approximation theory*. London, New York: McGraw-Hill.

Davidov, Y.: 1973, 'Mixing conditions for Markov chains'. *Theory of Probability and its Applications* **18**, 312–328.

Devroye, L., L. Györfi, and G. Lugosi: 1996, *A Probabilistic Theory of Pattern Recognition*, Applications of Mathematics: Stochastic Modelling and Applied Probability. Springer-Verlag New York.

Dietterich, T. G. and X. Wang: 2002, 'Batch Value Function Approximation via Support Vectors'. In: T. G. Dietterich, S. Becker, and Z. Ghahramani (eds.): *Advances in Neural Information Processing Systems 14*. Cambridge, MA, MIT Press.

Doukhan, P.: 1994, *Mixing Properties and Examples Lecture Notes in Statistics*, Vol. 85 of *Lecture Notes in Statistics*. Berlin: Springer-Verlag.

Ernst, D., P. Geurts, and L. Wehenkel: 2005, 'Tree-Based Batch Mode Reinforcement Learning'. *Journal of Machine Learning Research* **6**, 503–556.

Gordon, G. J.: 1995, 'Stable function approximation in dynamic programming'. In: A. Prieditis and S. Russell (eds.): *Proceedings of the Twelfth International Conference on Machine Learning*. San Francisco, CA, pp. 261–268, Morgan Kaufmann.

Guestrin, C., D. Koller, and R. Parr: 2001, 'Max-norm Projections for Factored MDPs'. *Proceedings of the International Joint Conference on Artificial Intelligence*.

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

Haussler, D.: 1995, 'Sphere packing numbers for subsets of the Boolean $n$-cube with bounded Vapnik-Chervonenkis dimension'. *Journal of Combinatorial Theory, Series A* **69**(2), 217–232.

Lagoudakis, M. and R. Parr: 2003, 'Least-Squares Policy Iteration'. *Journal of Machine Learning Research* **4**, 1107–1149.

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

Meyn, S. and R. Tweedie: 1993, *Markov Chains and Stochastic Stability*. New York: Springer-Verlag.

Munos, R.: 2003, 'Error Bounds for Approximate Policy Iteration'. *19th International Conference on Machine Learning* pp. 560–567.

Munos, R. and C. Szepesvári: 2006, 'Finite Time Bounds for Sampling Based Fitted Value Iteration'. Technical report, Computer and Automation Research Institute of the Hungarian Academy of Sciences, Kende u. 13-17, Budapest 1111, Hungary.

Murphy, S.: 2005, 'A Generalization Error for Q-Learning'. *Journal of Machine Learning Research* **6**, 1073–1097.

Nobel, A.: 1996, 'Histogram regression estimation using data-dependent partitions'. *Annals of Statistics* **24**(3), 1084–1105.

Ormoneit, D. and S. Sen: 2002, 'Kernel-Based Reinforcement Learning'. *Machine Learning* **49**, 161–178.

Pollard, D.: 1984, *Convergence of Stochastic Processes*. Springer Verlag, New York.

Pollard, D.: 1990, *Empirical Processes: Theory and Applications*. NSF-CBMS Regional Conference Series in Probability and Statistics, Institute of Mathematical Statistics, Hayward, CA.

Precup, D., R. Sutton, and S. Dasgupta: 2001, 'Off-Policy Temporal Difference Learning with Function Approximation'. In: *Proceedings of the Eighteenth International Conference on Machine Learning (ICML 2001)*. pp. 417–424.

Samuel, A.: 1959, 'Some studies in machine learning using the game of checkers'. *IBM Journal on Research and Development* pp. 210–229. Reprinted in *Computers and Thought*, E.A. Feigenbaum and J. Feldman, editors, McGraw-Hill, New York, 1963.

Sutton, R. and A. Barto: 1987, 'Toward a modern theory of adaptive networks: Expectation and prediction'. In: *Proc. of the Ninth Annual Conference of Cognitive Science Society*. Erlbaum, Hillsdale, NJ, USA.

Sutton, R. S. and A. G. Barto: 1998, 'Reinforcement Learning: An Introduction'. *Bradford Book*.

Szepesvári, C. and R. Munos: 2005, 'Finite Time Bounds for Sampling Based Fitted Value Iteration'. In: *ICML'2005*. pp. 881–886.

Szepesvári, C. and W. Smart: 2004, 'Interpolation-based Q-learning'. In: D. S. R. Greiner (ed.): *Proceedings of the International Conference on Machine Learning*. pp. 791–798.

Tsitsiklis, J. N. and B. Van Roy: 1996, 'Feature-Based Methods for Large Scale Dynamic Programming'. *Machine Learning* **22**, 59–94.

Wang, X. and T. Dietterich: 1999, 'Efficient Value Function Approximation Using Regression Trees'. In: *Proceedings of the IJCAI Workshop on Statistical Machine Learning for Large-Scale Optimization*. Stockholm,

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