Mach Learn (2008) 71: 89–129
DOI 10.1007/s10994-007-5038-2

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

András Antos · Csaba Szepesvári · Rémi Munos

Received: 9 September 2006 / Revised: 4 September 2007 / Accepted: 15 October 2007 /
Published online: 14 November 2007

Abstract In this paper we consider the problem of finding a near-optimal policy in a continuous space, discounted Markovian Decision Problem (MDP) by employing value-function-based methods when only a single trajectory of a fixed policy is available as the input. We study a policy-iteration algorithm where the iterates are obtained via empirical risk minimization with a risk function that penalizes high magnitudes of the Bellman-residual. Our main result is a finite-sample, high-probability bound on the performance of the computed policy that depends on the mixing rate of the trajectory, the capacity of the function set as measured by a novel capacity concept (the VC-crossing dimension), the approximation power of the function set and the controllability properties of the MDP. Moreover, we prove that when a linear parameterization is used the new algorithm is equivalent to Least-Squares Policy Iteration. To the best of our knowledge this is the first theoretical result for off-policy control learning over continuous state-spaces using a single trajectory.

Editors: Hans Ulrich Simon, Gabor Lugosi, Avrim Blum.

This paper appeared in a preliminary form at COLT2007 (Antos, et al. in LNCS/LNAI, vol. 4005, pp. 574–588, 2006).

A. Antos · C. Szepesvári (✉)
Computer and Automation Research Inst. of the Hungarian Academy of Sciences, Kende u. 13-17, Budapest 1111, Hungary
e-mail: szcsaba@sztaki.hu

A. Antos
e-mail: antos@sztaki.hu

Present address:
C. Szepesvári
Department of Computing Science, University of Alberta, Edmonton, AB, Canada

R. Munos
Institut National de Recherche en Informatique et en Automatique, INRIA Lille, 40 avenue Hally, 59650 Villeneuve d'Ascq, France
e-mail: remi.munos@inria.fr

Keywords Reinforcement learning · Policy iteration · Bellman-residual minimization · Least-squares temporal difference learning · Off-policy learning · Nonparametric regression · Least-squares regression · Finite-sample bounds

1 Introduction

In many industrial control problems collecting data of the controlled system is often separated from the learning phase: The data is collected in "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 while it is operated using an existing controller, also called the *behavior policy* (Sutton and Barto 1998, Chap. 5.6).

In this paper we are interested in designing learning algorithms with provable performance guarantees for infinite-horizon expected total discounted reward Markovian Decision Problems with continuous state-variables and finite action-spaces.

The algorithm that we study is an instance of fitted policy iteration: the main loop computes the evaluation function of the policy of the previous step in an approximate manner by minimizing some risk functional. This new function is then used to compute the next, improved policy. To avoid the need of learning a model, action-value evaluation functions are employed making the policy improvement step trivial, just like how it is done in the Least-Squares Policy Iteration (LSPI) algorithm of Lagoudakis and Parr (2003). However, while LSPI builds on Least-Squares Temporal Difference (LSTD) learning due to Bradtke and Barto (1996), we base our algorithm on the idea of minimizing Bellman-residuals. The idea of minimizing Bellman-residuals is not new by any means. In fact, it goes back at least to the work of Schweitzer and Seidmann (1985), who proposed this idea for computing approximate state-value functions assuming the full knowledge of a finite-state, finite-action MDP.

For both LSTD and Bellman-residual minimization (BRM) the user must select a function class to represent the potential action-value functions that can be used in the algorithm. Obviously, the space of all potential value-functions is too big when the state space is continuous and so one typically works with a small subset of this space. One popular choice is to use linearly parameterized functions, but the results of this paper apply equally to other, richer classes.

In the BRM approach one aims at picking a function that minimizes the so-called Bellman-residual. The Bellman-residual arises when the fixed-point equation for the policy's value function is rewritten so that one side of the equation equals zero. Formally, the fixed point equation then reads $T^\pi Q^\pi - Q^\pi = 0$, where Q^π is the policy's action-value function and T^π is the policy's evaluation operator (T^π, Q^π, just like the other MDP objects used below will be fully defined in the next section). If Q^π is replaced by some other function Q, the left-hand side, $T^\pi Q - Q$, becomes non-zero. A reasonable goal to get a good approximation to Q^π is to control the magnitude of the Bellman-residual, such as its weighted squared 2-norm.

While in BRM one aims directly at minimizing such a term, LSTD does this in an indirect manner. In order to explain the idea underlying LSTD let us rewrite the above reordered fixed-point equation in terms of the samples: If $\{X_t\}$ is a sequence of states, $\{R_t\}$ is a sequence of rewards and $\{A_t\}$ is a sequence of actions encountered while following policy π then the reordered fixed-point equation can be written as

$$\mathbb{E}[R_t + \gamma Q^\pi(X_{t+1}, \pi(X_{t+1})) - Q^\pi(X_t, A_t) | (X_t, A_t) = (x, a)] = 0,$$

where (x, a) is any state-action pair. Here $R_t + \gamma Q(X_{t+1}, \pi(X_{t+1})) - Q(X_t, A_t)$ is called the t^{th} *temporal difference*. In LSTD one works with linear parameterizations of the action-value functions and the idea is to find a function Q such that the average of the temporal differences when correlated with the basis functions underlying the linear parameterization is zero. This is expected to work since averages approximate expectations and if the correlation of a function with a sufficiently rich set of functions is zero then the function must be zero (almost everywhere).

However, from the statistical point of view this algorithm is not straightforward to analyze. This is because unlike most machine learning algorithms, LSTD is not derived from a risk minimization principle, hence existing tools of machine learning and statistics, most of which are geared towards the analysis of risk minimization algorithms, cannot be applied directly to its analysis. This makes the Bellman-residual minimization approach more attractive, at least for the first sight. However, one major obstacle to (direct) Bellman-residual minimization is that the natural candidate of the empirical risk, the average of the squared temporal differences computed along a trajectory does not give rise to an unbiased estimate of the squared L^2-norm of the Bellman-residual (e.g., Sutton and Barto 1998, p. 220).

Here we propose to overcome the biasedness of this empirical risk by modifying the loss function minimized. The novel loss function depends on an auxiliary function whose job is to make sure that the empirical loss is an unbiased estimate of the population-based loss (Lemma 1). In addition, it turns out that in the case of a linear parameterization, the minimizer of the new loss function and the solution returned by LSTD coincide (Proposition 2). In this sense our new BRM minimization algorithm generalizes LSTD, while the new policy iteration algorithm generalizes LSPI to a richer set of functions.

The main result of the paper (Theorem 4) shows that if the input trajectory is sufficiently representative then the performance of the policy returned by our algorithm improves at a rate of $1/N^{1/4}$ up to a limit set by the choice of the function set (here N is the length of the trajectory). 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 behavior 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 relaxed. The mixing condition, on the other hand, seems to be essential for efficient learning. The particular mixing condition that we use is exponential β-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. This mixing condition 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, e.g., algebraic β-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 it could be very slow. Let us finally note that for Markov processes, geometric ergodicity implies exponential β-mixing (see Davidov 1973; or Doukhan 1994, Chap. 2.4), hence for such processes there is no loss of generality in assuming exponential β-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 a function set $\mathcal{F}$ is defined as the VC-dimension of a set-system that consists of the zero-level sets of the pairwise differences

of functions from $\mathcal{F}$. The intuitive explanation is that in policy iteration the action taken by the next policy at some state is obtained by selecting the action that yields the best action-value. When solving the fixed-point equation for this policy, the policy (as a function of states to actions) is composed with the action-value function candidates. In order to control variance, one needs to control the capacity of the resulting function set. The composite functions can be rewritten in terms of the zero-level sets mentioned above, and this is where the VC-dimension of this set-system comes into play. The new concept is compared to previous capacity concepts and is found to be significantly different from them, except for the case of a set of linearly parameterized functions whose VC-crossing dimension equals the number of parameters, as usual (Proposition 3).

Similarly to bounds of regression, our bounds depend on the approximation power of the function set, too. One major difference is that in our case the approximation power of a function set is measured differently from how it is done in regression. While in regression, the approximation power of a function set is characterized by the deviation of the target class from the considered set of functions, we use error measures that quantify the extent to which the function set is invariant with respect to the policy evaluation operators underlying the policies in the MDP. This should come as no surprise: If for some policy encountered while executing the algorithm no function in the chosen set has a small Bellman-residual, the performance of the final policy could be very poor.

The bounds also depend on the number of steps (K) of policy iteration. As expected, there are two terms involving K that behave inversely: One term, that is familiar from previous results, decays at a geometric rate (the base being γ, the discount factor of the MDP). The other term increases proportionally to the logarithm of the number of iterations. This term comes from the reuse of the data throughout all the iterations: Hence we see that data reuse causes only a slow degradation of performance, a point that was made just recently by Munos and Szepesvári (2006) in the analysis of approximate value iteration. Interestingly, the optimal value of K depends on, e.g., the capacity of the function set, the mixing rate, and the number of samples, but it does not depend on the approximation-power of the function set.

In order to arrive at our results, we 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, 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) for bounding 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 an L^∞-analysis then the controllability assumption would not be needed, but since in the intermediate steps the algorithm targets to minimize the L^2-norm of errors, we may expect difficulties in controlling the final error.

The particular controllability assumption studied here requires that the maximum rate at which the future-state distribution can be concentrated by selecting some non-stationary Markov policy is sub-exponential. In general, this holds for systems with "noisy" transitions, but the condition can also hold for deterministic systems (Munos and Szepesvári 2006).

The organization of the paper is as follows: In the next section (Sect. 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 Sect. 3. This is followed by some additional definitions necessary for the presentation of the main result. The main result is given at the beginning of

Sect. 4. The rest of this section is divided into three parts, each devoted to one major step of the proof. In particular, in Sect. 4.1 a finite-sample bound is given on the error of the policy evaluation procedure. This bound makes the dependence on the complexity of the function space, the mixing rate of the trajectory, and the number of samples explicit. In Sect. 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 Sect. 4.3. We discuss the main result, in the context of previous work in Sect. 5. Finally, our conclusions are drawn and possible directions for future work are outlined in Sect. 6.

2 Definitions

As we shall work with continuous spaces we will need a few simple concepts of analysis. These are introduced first. This is followed by the introduction of Markovian Decision Problems (MDPs) and the associated concepts and the necessary notation. The unattributed statements of this section can be found in the book of Bertsekas and Shreve (1978).

For a measurable space with domain S, we let $\mathcal{M}(S)$ denote the set of probability measures over S. For $p \geq 1$, a measure $\nu \in \mathcal{M}(S)$, 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, $\mathbb{I}_{\{E\}} = 1$ if and only if 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*, $\mathcal{A} = \{a_1, a_2, \ldots, a_M\}$ is the set of *actions*, $P : \mathcal{X} \times \mathcal{A} \to \mathcal{M}(\mathcal{X})$ is the *transition probability kernel*, $P(\cdot|x, a)$ defining the next-state distribution upon taking action a in state x, $S(\cdot|x, a)$ gives the corresponding distribution of *immediate rewards*, and $\gamma \in (0, 1)$ is the discount factor. Note that M denotes the number of actions in the MDP.

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. 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. Deterministic stationary Markov policies will be identified with mappings from states to actions, i.e., functions of the form $\pi : \mathcal{X} \to \mathcal{A}$.

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

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

Here R_t denotes the reward received at time step t; $R_t \sim S(\cdot|X_t, A_t)$ and X_t evolves 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 π. For a stationary Markov policy π, $A_t \sim \pi(\cdot|X_t)$, while if π is deterministic stationary Markov then we write $A_t = \pi(X_t)$. The function V^π is also called the state-value function of policy π. Closely related to the state-value functions are the action-value functions, defined by

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

In words, the action-value function underlying π assigns to the pair (x, a) the total expected discounted return encountered when the decision process is started in state x, the first action is a while all the subsequent actions are determined by the policy π. It is easy to see that for any policy π, the functions V^π, Q^π are bounded by $R_{\max}/(1-\gamma)$.

Given an MDP, 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 $x \in \mathcal{X}$.

In order to characterize optimal policies it will be useful to define the optimal action-value function, $Q^*(x, a)$:

$$Q^*(x, a) = \sup_\pi Q^\pi(x, a).$$

Further, we say that a (deterministic stationary) policy π 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, a greedy policy always exists no matter how Q is chosen. Greedy policies are important because any greedy policy w.r.t. Q^* is optimal. Hence, to find an optimal policy it suffices to determine Q^* and the search for optimal policies can be restricted to deterministic, stationary, Markov policies. In what follows we shall use the word 'policy' to mean policies of this class only.

In the policy iteration algorithm (Howard 1960), Q^* is found by computing a series of policies, each policy being greedy w.r.t. the action-value function of the previous policy. The algorithm converges at a geometric rate. The action-value function of a policy can be

found by solving a fixed-point equation. For a (deterministic stationary Markov) policy π, 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^π is a contraction operator w.r.t. the supremum-norm with index γ: $\|T^\pi Q - T^\pi Q'\|_\infty \le \gamma \|Q - Q'\|_\infty$. Moreover, the action-value function of π is the unique fixed point of T^π:

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

For our analysis we shall need a few more operators. We define the projection operator $E^\pi : B(\mathcal{X} \times \mathcal{A}) \to B(\mathcal{X})$ by

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

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

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

Hence, for a function V, PV represents an action-value function such that $(PV)(x,a)$ is the expected value of choosing action a in state x if the future states are evaluated via V and there are no immediate rewards. The left-linear operator, $\cdot P : \mathcal{M}(\mathcal{X} \times \mathcal{A}) \to \mathcal{M}(\mathcal{X})$, is defined by

$$(\rho P)(dy) = \int P(dy|x,a)\rho(dx,da). \tag{2}$$

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

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

For a measure ρ defined over the set of state-action pairs, ρP represents the distribution of the state of the process after one step in the MDP if the initial state and action are sampled from ρ.

By composing P and E^π, we define P^π:

$$P^\pi = PE^\pi.$$

Note that this equation defines *two* operators: a right- and a left-linear one. The interpretation of the right-linear operator is as follows: For the action-value function Q, $PE^\pi Q$ gives the expected values of future states when the future values of the actions are given by the action-value function Q and after the first step policy π is followed. The left-linear operator, $\cdot P^\pi : \mathcal{M}(\mathcal{X} \times \mathcal{A}) \to \mathcal{M}(\mathcal{X} \times \mathcal{A})$, is defined as follows: Let U be a measurable subset of $\mathcal{X} \times \mathcal{A}$. Given $\rho \in \mathcal{M}(\mathcal{X} \times \mathcal{A})$, $(\rho P^\pi)(U) = \rho PE^\pi \mathbb{I}_{\{U\}}$. This operator can be given a probabilistic interpretation, too, but as we have not found this interpretation sufficiently intuitive, it is omitted.

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 $\mathcal{F}^M$ as real-valued functions f defined over $\mathcal{X} \times \mathcal{A}$ with the obvious identification $f \equiv (f_1, \ldots, f_M)$,

$f(x, a_j) = f_j(x)$, $j = 1, \ldots, M$. The set $\mathcal{F}^M$ will denote the set of admissible functions used in the optimization step of our algorithm.

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

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

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

$$\int Q(x, a)\nu(dx, da) = \frac{1}{M} \sum_{j=1}^{M} \int Q(x, a_j)\nu(dx). \tag{3}$$

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. The ceiling value of a real number a is denoted by $\lceil a \rceil$, while for $x > 0$, $\log^+(x) = 0 \vee \log(x)$.

3 Algorithm

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

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

underlying some stochastic stationary policy π_b: $A_t \sim \pi_b(\cdot|X_t)$, $X_{t+1} \sim P(\cdot|X_t, A_t)$, $R_t \sim S(\cdot|X_t, A_t)$. We assume that this trajectory is sufficiently rich 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 ν. The action-evaluation function Q_{-1} is used to initialize the first policy (alternatively, one may start with an arbitrary initial policy). The procedure *PEval* takes data in the form of a long trajectory and some policy $\hat{\pi}$ and should return an approximation to the action-value function of $\hat{\pi}$. In this case the policy is just the greedy policy with respect to Q': $\hat{\pi} = \hat{\pi}(\cdot; Q')$.

There are many possibilities to design *PEval*. In this paper we consider an approach based on Bellman-residual minimization (BRM). Let π denote the policy to be evaluated.

FittedPolicyQ(D,K,Q_{-1},PEval,π_b)
// 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, π_b)
end for
return Q // or $\hat{\pi}(\cdot; Q)$, the greedy policy w.r.t. Q

Fig. 1 Model-free fitted policy iteration

The basic idea of BRM comes from rewriting the fixed-point equation (1) for Q^π in the form $Q^\pi - T^\pi Q^\pi = 0$. When Q^π is replaced by some other function Q, the left-hand side becomes non-zero. The resulting quantity, $Q - T^\pi Q$, is called the *Bellman-residual* of Q. If the magnitude, $\|Q - T^\pi Q\|$, of the Bellman-residual is small then Q can be expected to be a good approximation to Q^π (for an analysis using supremum norms see, e.g., the work of Williams and Baird (1994)). Here we choose a weighted L^2-norm to measure the magnitude of the Bellman-residual as it leads to an optimization problem with favorable characteristics and enables an easy connection to regression function estimation to be made. Hence, define the loss function

$$L(Q;\pi) = \|Q - T^\pi Q\|_\nu^2,$$

where the weighting is determined by ν, the stationary distribution underlying the states in the input data and the uniform distribution over the actions. (Remember that $\|Q\|_\nu^2 = 1/M \sum_{j=1}^M \|Q(\cdot, a_j)\|_\nu^2$.) Since X_t follows the distribution ν, the choice of ν in the loss function facilitates its sample-based approximation. The choice of the uniform distribution over the actions instead of the distribution underlying the sample $\{A_t\}$ expresses our *a priori disbelief* in the action-choices made by the behavior policy: Since the behavior policy may well prefer suboptimal actions over the optimal ones, we have no reason to give more weight to the actions that are sampled more often. Of course, the same issue arises in connection to the state distribution, or the joint state-action distribution. However, correcting for the bias involving the states would be possible only if ν had a known density (a very unrealistic assumption) or if this density was learnt from the samples. Thus, while the correction for the sampling "bias" of actions requires only the (mild) assumption of the knowledge of the behavior policy and is very cheap (as we shall see below), the correction for the states' bias would be quite expensive and risky. Hence, to simplify the presentation we do not consider such a correction here. Another alternative would be to use the joint distribution of (X_t, A_t) in the above norm; the results would not change significantly in this case.

In light with the above remarks, given the loss L, we expect $Q = \operatorname{argmin}_{f \in \mathcal{F}^M} L(f;\pi)$ to be a good approximation to the evaluation function of π.[1] In order to obtain a practical procedure one needs a sample-based approximation to L. To arrive at such an approximation one may first try to replace $\|\cdot\|_\nu$ by its empirical counterpart,

$$\|f\|_{\nu,N}^2 \stackrel{\text{def}}{=} \frac{1}{NM} \sum_{t=1}^N \sum_{j=1}^M \frac{\mathbb{I}_{\{A_t = a_j\}}}{\pi_b(a_j|X_t)} f_j(X_t)^2 = \frac{1}{NM} \sum_{t=1}^N \frac{f(X_t, A_t)^2}{\pi_b(A_t|X_t)}$$

(since $X_t \sim \nu$ and if, e.g., $\{X_t\}$ is ergodic, $\|f\|_{\nu,N}^2$ converges to $\|f\|_\nu^2$ as $N \to \infty$) and then plug in $R_t + \gamma f(X_{t+1}, \pi(X_{t+1}))$ in place of $(T^\pi f)(X_t, A_t)$ (since $\mathbb{E}[R_t + \gamma f(X_{t+1}, \pi(X_{t+1}))|X_t, A_t] = (T^\pi f)(X_t, A_t)$). This results in the loss function

$$\hat{L}_N(f;\pi) = \frac{1}{NM} \sum_{t=1}^N \frac{1}{\pi_b(A_t|X_t)} \times (f(X_t, A_t) - (R_t + \gamma f(X_{t+1}, \pi(X_{t+1}))))^2. \quad (4)$$

[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).

However, as it is well known (see, e.g., Sutton and Barto 1998, p. 220; Munos 2003; or Lagoudakis and Parr 2003), $\hat{L}_N$ is *not* an unbiased estimate of the L^2 Bellman-error: $\mathbb{E}[\hat{L}_N(f;\pi)] \neq L(f;\pi)$. Indeed, elementary calculus shows that for $Y \sim P(\cdot|x,a)$, $R \sim S(\cdot|x,a)$,

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

It follows that minimizing $\hat{L}_N(f;\pi)$ in the limit when $N \to \infty$ is equivalent to minimizing the sum of $L(f;\pi)$ and $\gamma^2 \frac{1}{M} \sum_{j=1}^M \mathbb{E}[\mathrm{Var}[f(Y,\pi(Y))|X, A = a_j]]$ with $Y \sim P(\cdot|X,A)$. The unwanted variance term acts like a penalty factor, favoring smooth solutions (if f is constant then the variance term $\mathrm{Var}[f(Y,\pi(Y))|X, A = a_j]$ becomes zero). 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 in $\hat{L}_N$. 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 1998, p. 220). However, this is neither realistic nor sample efficient unless there is a (cheap) way to generate samples—a possibility that we do not consider here. Another option, 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 these pitfalls and looks simpler.

The trick is to introduce an auxiliary function h to cancel the unwanted variance term. The new loss function is

$$L(f,h;\pi) = L(f;\pi) - \|h - T^\pi f\|_\nu^2 \tag{5}$$

and we propose to solve for

$$\hat{f} = \operatorname*{argmin}_{f \in \mathcal{F}^M} \sup_{h \in \mathcal{F}^M} L(f,h;\pi), \tag{6}$$

where the supremum comes from the negative sign of $\|h - T^\pi f\|_\nu^2$ (our aim is to push h close to $T^\pi f$). There are two issues to worry about: One is if the optimization of this new loss function still makes sense and the other is if the empirical version of this loss is unbiased. A quick heuristic explanation of why the second issue is resolved is as follows: In the sample based estimate of $\|h - T^\pi f\|_\nu^2$ the same variance term, as the one that we wanted to get rid of, appears. Since $\|h - T^\pi f\|_\nu^2$ is subtracted from the original loss function, when considering the empirical loss the unwanted terms cancel each other. A precise reasoning will be given below in Lemma 1.

Now let us consider the issue if optimizing the new loss makes sense. Let $h_f^* \in \mathcal{F}^M$ be a function that minimizes $\|h - T^\pi f\|_\nu^2$. Then

$$L(f;\pi) = L(f,h_f^*;\pi) + \|h_f^* - T^\pi f\|^2.$$

Thus, if $\|h_f^* - T^\pi f\|_\nu^2$ is "small" independently of the choice of f then minimizing $L(f,h_f^*;\pi)$ should give a solution whose loss as measured by $L(f;\pi)$ is small, too.

Before returning to the unbiasedness issue let us note that for $f \in \mathcal{F}^M$, $L(f, h^*_f; \pi) \geq 0$. This inequality holds because by the definition of h^*_f, $L(f, h^*_f; \pi) \geq L(f, h; \pi)$ holds for any $h \in \mathcal{F}^M$. Thus substituting $h = f$ we get $L(f, h^*_f; \pi) \geq L(f, f; \pi) = 0$.

Let us now define the empirical version of $L(f, h; \pi)$ by

$$\hat{L}_N(f, h; \pi) = \frac{1}{NM} \sum_{t=1}^{N} \frac{1}{\pi_b(A_t|X_t)} [(f(X_t, A_t) - (R_t + \gamma f(X_{t+1}, \pi(X_{t+1}))))^2 - (h(X_t, A_t) - (R_t + \gamma f(X_{t+1}, \pi(X_{t+1}))))^2]. \tag{7}$$

Thus, we let *PEval* solve for

$$Q = \operatorname*{argmin}_{f \in \mathcal{F}^M} \sup_{h \in \mathcal{F}^M} \hat{L}_N(f, h; \pi). \tag{8}$$

The key attribute of the new loss function is that its empirical version is unbiased:

Lemma 1 (Unbiased empirical loss) *Assume that the behavior policy π_b samples all actions in all states with positive probability. Then for any $f, h \in \mathcal{F}^M$, policy π, $\hat{L}_N(f, h; \pi)$ as defined by (7) provides an unbiased estimate of $L(f, h; \pi)$:*

$$\mathbb{E}[\hat{L}_N(f, h; \pi)] = L(f, h; \pi). \tag{9}$$

Proof Let us define $C_{tj} = \frac{\mathbb{I}_{\{A_t = a_j\}}}{\pi_b(a_j|X_t)}$ and $\hat{Q}_{f,t} = R_t + \gamma f(X_{t+1}, \pi(X_{t+1}))$. By (7), the t^{th} term of $\hat{L}_N(f, h; \pi)$ can be written as

$$L^{(t)} = \frac{1}{M} \sum_{j=1}^{M} C_{tj}((f_j(X_t) - \hat{Q}_{f,t})^2 - (h_j(X_t) - \hat{Q}_{f,t})^2). \tag{10}$$

Note that $\mathbb{E}[C_{tj}|X_t] = 1$ and

$$\mathbb{E}[C_{tj}\hat{Q}_{f,t}|X_t] = \mathbb{E}[\hat{Q}_{f,t}|X_t, A_t = a_j] = r_j(X_t) + \gamma \int_y f(y, \pi(y)) dP(y|X_t, a_j) = (T^\pi f)_j(X_t) \tag{11}$$

since all actions are sampled with positive probability in any state. (In (10) and (11), we use the convention $f(x, a_j) = f_j(x)$ introduced earlier.)

Consider now $t = 1$ and $L^{(1)}$. Taking expectations,

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

By the bias-variance decomposition formula, if U, V are conditionally uncorrelated given W then,

$$\mathbb{E}[(U - V)^2|W] = \mathbb{E}[(U - \mathbb{E}[V|W])^2|W] + \mathrm{Var}[V|W],$$

where $\mathrm{Var}[V|W] = \mathbb{E}[(V - \mathbb{E}[V|W])^2|W]$. Using this and (11), we get

$$\begin{aligned}
&\mathbb{E}[C_{1j}((f_j(X_1) - \hat{Q}_{f,1})^2 - (h_j(X_1) - \hat{Q}_{f,1})^2)|X_1] \\
&\quad = (f_j(X_1) - (T^\pi f)_j(X_1))^2 + \mathrm{Var}[\hat{Q}_{f,1}|X_1, A_1 = a_j] \\
&\qquad - ((h_j(X_1) - (T^\pi f)_j(X_1))^2 + \mathrm{Var}[\hat{Q}_{f,1}|X_1, A_1 = a_j]) \\
&\quad = (f_j(X_1) - (T^\pi f)_j(X_1))^2 - (h_j(X_1) - (T^\pi f)_j(X_1))^2.
\end{aligned}$$

Taking expectations of both sides we get that

$$\begin{aligned}
\mathbb{E}[L^{(1)}] &= \frac{1}{M}\sum_{j=1}^{M}(\|f_j - (T^\pi f)_j\|_\nu^2 - \|h_j - (T^\pi f)_j\|_\nu^2) \\
&= L(f;\pi) - \|h - T^\pi f\|_\nu^2 \\
&= L(f,h;\pi). \qquad (12)
\end{aligned}$$

Because of stationarity, $\mathbb{E}[L^{(t)}] = \mathbb{E}[L^{(1)}]$ holds for any t, thus finishing the proof of (9). □

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

For linearly parameterized function classes the solution of the optimization problem (8) can be obtained in closed form. Perhaps surprisingly, even more is true in this case: The new method gives the same solutions as LSTD! In order to formally state this result let us first review the LSTD procedure.[2]

Instead of minimizing the distance of Q and $T^\pi Q$, in LSTD one looks for a value function Q in the space of admissible functions $\mathcal{F}^M$ such that the back-projection of the image of Q under T^π onto $\mathcal{F}^M$ comes the closest to Q (see Fig. 2). Formally, this means that we want to minimize $\|Q - \Pi T^\pi Q\|^2$, where $\|f\|$ is a norm compatible with some inner product: $\|f\|^2 = \langle f, f\rangle$. Here the projection operator $\Pi : B(\mathcal{X}\times\mathcal{A}) \to B(\mathcal{X}\times\mathcal{A})$ is defined by $\Pi Q = \mathrm{argmin}_{Q'\in\mathcal{F}^M}\|Q - Q'\|$. In order to make the minimization problem practical it is customary to assume a linear parameterization of the value functions: $\mathcal{F}^M = \{w^T\phi : w \in \mathbb{R}^p\}$, where $\phi : \mathcal{X}\times\mathcal{A} \to \mathbb{R}^p$ is some function extracting features of state-action pairs. Note that $\mathcal{F}^M$ is a linear subspace (hyperplane) of $B(\mathcal{X}\times\mathcal{A})$. Denote by w^π the weights of the solution of the minimization problem and let $Q_{w^\pi} = (w^\pi)^T\phi$. Then due to the properties of projection, $Q_{w^\pi} - T^\pi Q_{w^\pi}$ must be perpendicular to the space $\mathcal{F}^M$ with respect to the inner product underlying the chosen norm.[3] Formally, this means that

[2] We introduce LSTD quite differently from how it is normally done in the literature (the description given in the introduction follows the "normal" pattern). In fact, our treatment is influenced by Lagoudakis and Parr (2003).

[3] This is because the projection of a vector to a linear subspace is the unique element of the subspace such that the vector connecting the element and the projected vector is perpendicular to the subspace. Hence if for some $Q \in \mathcal{F}^M$, $Q - T^\pi Q$ happens to be perpendicular to $\mathcal{F}^M$ then (since $Q \in \mathcal{F}^M$) Q must be the projection of $T^\pi Q$ onto $\mathcal{F}^M$, i.e., Q and $\Pi T^\pi Q$ must coincide. Further, since it is always possible to find $Q \in \mathcal{F}^M$ such that $Q - T^\pi Q$ is perpendicular to $\mathcal{F}^M$, and for such a Q, $Q = \Pi T^\pi Q$, the minimal value of the LSTD loss is always zero. Hence, if Q is the minimizer of the LSTD loss then $Q = \Pi T^\pi Q$. Then $Q - T^\pi Q = \Pi T^\pi Q - T^\pi Q$ must be perpendicular to $\mathcal{F}^M$.

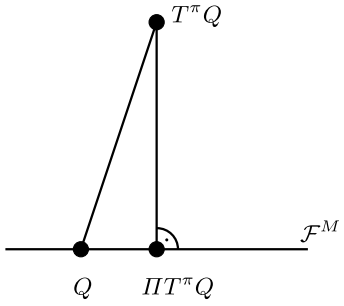

Fig. 2 Comparing the modified Bellman-error and the LSTD criterion. The function space, $\mathcal{F}^M$, is represented by the horizontal line. Under the operator, T^π, a value function, $Q \in \mathcal{F}^M$, is mapped to a function, $T^\pi Q$. The vector connecting $T^\pi Q$ and its back-projection to $\mathcal{F}^M$, $\Pi T^\pi Q$, is orthogonal to the function space $\mathcal{F}^M$. The Bellman-error is the distance of Q and $T^\pi Q$. In order to get the modified Bellman-error loss, the squared distance of $T^\pi Q$ and $\Pi T^\pi Q$ is subtracted from the squared Bellman-error. LSTD aims at picking a function Q such that its distance to $\Pi T^\pi Q$ is minimal. For a linear space, $\mathcal{F}^M$, the solution of this is $Q = \Pi T^\pi Q$, which simultaneously minimizes the modified Bellman-error loss function

$\langle Q_{w^\pi} - T^\pi Q_{w^\pi}, w^T \phi \rangle = 0$ must hold for any weight-vector w. However, this can hold only if for any $i \in \{1, \ldots, p\}$,

$$\langle Q_{w^\pi} - T^\pi Q_{w^\pi}, \phi_i \rangle = 0. \tag{13}$$

These are the so-called *normal equations* and the linearity of the inner product can be used to solve them for w^π.

When LSTD is used in practice, T^π and the inner product are approximated based on the samples. Then (13) becomes

$$0 = \frac{1}{NM} \sum_{t=1}^{N} \frac{\phi_i(X_t, A_t)}{\pi_b(A_t|X_t)} (Q_{w^\pi}(X_t, A_t) - [R_t + \gamma Q_{w^\pi}(X_{t+1}, \pi(X_{t+1}))]), \tag{14}$$

where the normalization factors $(1/M)/\pi_b(A_t|X_t)$ are introduced to remain consistent with our previous convention to normalize with the action frequencies. Note that unlike in the case of the straightforward empirical loss (4), there is no biasedness issue here and hence asymptotic consistency is easy to obtain (Bradtke and Barto 1996).

For our purposes it is important to note that (14) can be derived from a loss-minimization principle with a reasoning that is entirely analogous to the argument used to derive (13). To see this, define $S_N : B(\mathcal{X} \times \mathcal{A}) \to \mathbb{R}^N$, $\hat{T}_N^\pi : B(\mathcal{X} \times \mathcal{A}) \to \mathbb{R}^N$ and $\langle \cdot, \cdot \rangle_N$ by

$$\begin{aligned}
S_N Q &= (Q(X_1, A_1), \ldots, Q(X_N, A_N))^T, \\
\hat{T}_N^\pi Q &= (R_1 + \gamma Q(X_2, \pi(X_2)), \ldots, R_N + \gamma Q(X_{N+1}, \pi(X_{N+1})))^T, \\
\langle q, q' \rangle_N &= \frac{1}{NM} \sum_{t=1}^{N} \frac{q_t q'_t}{\pi_b(A_t|X_t)},
\end{aligned}$$

where $q, q' \in \mathbb{R}^N$. Further, let $\|\cdot\|_N$ denote the ℓ^2-norm on $\mathbb{R}^N$ that corresponds to $\langle \cdot, \cdot \rangle_N$ and let $S_N \mathcal{F}^M = \{S_N Q : Q \in \mathcal{F}^M\}$. Note that $S_N \mathcal{F}^M$ is a linear subspace of $\mathbb{R}^N$. Then (14) can be written in the compact form $0 = \langle S_N Q_{w^\pi} - \hat{T}_N^\pi Q_{w^\pi}, \phi_i \rangle_N$. Further, the solution of these equations minimizes $\|S_N Q - \Pi_N \hat{T}_N^\pi Q\|_N$, where the projection operator $\Pi_N : \mathbb{R}^N \to \mathbb{R}^N$ is defined by $\Pi_N q = \operatorname{argmin}_{q' \in S_N \mathcal{F}^M} \|q - q'\|_N$.

Now we are ready to state our equivalence result:

Proposition 2 *When linearly parameterized functions are used, the solution of (8) and that of LSTD coincide and the algorithm proposed here becomes equivalent to LSPI.*

Proof We prove the statement for the population based losses, $L_{\mathrm{LSTD}}(Q;\pi) = \|Q - \Pi T^\pi Q\|^2$, $L_{\mathrm{BRM}}(Q;\pi) = \|Q - T^\pi Q\|^2 - \inf_{h\in\mathcal{F}^M} \|h - T^\pi Q\|^2$, where $\|\cdot\|$ is any norm derived from some inner product $\langle\cdot,\cdot\rangle$. The argument for the empirical losses is an exact parallel of the argument presented here, one just must use S_N, Π_N and $\langle\cdot,\cdot\rangle_N$ as defined above.

Let $Q \in \mathcal{F}^M$ solve the equations $\langle Q - T^\pi Q, \phi_i\rangle = 0$ simultaneously for all i. We know that all minimizers of L_{LSTD} can be obtained this way and that the value of L_{LSTD} at a minimizer is zero. Since the Pythagorean identity $\|Q - T^\pi Q\|^2 = \|Q - \Pi T^\pi Q\|^2 + \|\Pi T^\pi Q - T^\pi Q\|^2$ holds for any Q, from $L_{\mathrm{LSTD}}(Q;\pi) = 0$ we conclude that $\|Q - T^\pi Q\|^2 = \|\Pi T^\pi Q - T^\pi Q\|^2$. Hence, $L_{\mathrm{BRM}}(Q;\pi) = \|Q - T^\pi Q\|^2 - \|\Pi T^\pi Q - T^\pi Q\|^2 = 0$. Since L_{BRM} is non-negative on $\mathcal{F}^M$, this shows that Q is a minimizer of L_{BRM}.

Now, let Q be the minimizer of L_{BRM}. Using again the Pythagorean identity, we immediately get that $\|Q - \Pi T^\pi Q\|^2 = 0$, which together with the non-negativity of L_{LSTD} gives that Q is a minimizer of L_{LSTD}. □

As a consequence of this equivalence, when a linear function class is used all our results derived for the BRM loss transfer to LSTD/LSPI.

One problem with the LSTD loss is that it is defined in terms of the projection Π which makes its optimization quite involved when a *non-linear* parameterization is used (e.g., when a neural network is used to represent the action-value functions). On the other hand, the BRM criterion proposed here avoids the direct use of the projection operator and hence it is easier to use it with non-linear parameterizations. This can be advantageous when there is a reason to believe that a non-linear parameterization is useful. Of course, for such parameterizations the optimization problem may be difficult to solve anyway.

4 Main result

Before describing the main result we need some more 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 β-mixing:

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

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

$\{Z_t\}$ is said to be β-mixing if $\beta_m \to 0$ as $m\to\infty$. In particular, we say that a β-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 "mixing" can be defined in a large number of ways (see, e.g., Doukhan 1994). The weakest among the most commonly used mixing concepts is called α-mixing. Another commonly used mixing concept is ϕ-mixing, which is stronger than β-mixing (see Meyn and Tweedie 1993).

Our assumptions regarding the sample path are as follows:

Assumption 2 (Sample path properties) Assume that

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

is the sample path of some stochastic stationary policy, π_b. Further, assume that $\{X_t\}$ is strictly stationary ($X_t \sim \nu \in \mathcal{M}(\mathcal{X})$) and exponentially β-mixing with a rate defined by the parameters $(\overline{\beta}, b, \kappa)$. We further assume that the sampling policy π_b satisfies $\pi_{b0} \stackrel{\text{def}}{=} \min_{a\in\mathcal{A}} \inf_{x\in\mathcal{X}} \pi_b(a|x) > 0$.

The β-mixing property will be used to establish tail inequalities for certain empirical processes. Note that if X_t is β-mixing then the hidden-Markov process $\{(X_t, (A_t, R_t))\}$ is also β-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).

Our next assumption concerns the controllability of the MDP. Remember that ν denotes the stationary distribution underlying $\{X_t\}$. For the sake of flexibility, we allow the user to choose another distribution, ρ, to be used in assessing the procedure's performance. 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 ρ and ν. A natural way to bound the effect of changing from measure α to measure β is to use the Radon-Nikodym derivative of α w.r.t. β:[4] for any nonnegative measurable function f, $\int f d\alpha = \int f \frac{d\alpha}{d\beta} d\beta \le \|\frac{d\alpha}{d\beta}\|_\infty \int f d\beta$. This motivates the following definition introduced in (Munos and Szepesvári 2006):

Definition 2 (Discounted-average concentrability of future-state distribution) Given ρ, ν, $m \ge 0$ and an arbitrary sequence of stationary policies $\{\pi_m\}_{m\ge 1}$ let

$$c_{\rho,\nu}(m) = \sup_{\pi_1,\dots,\pi_m} \left\| \frac{d(\rho P^{\pi_1} P^{\pi_2} \dots P^{\pi_m})}{d\nu} \right\|_\infty, \tag{15}$$

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

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

In general $c_{\rho,\nu}(m)$ diverges to infinity as $m \to \infty$. However, thanks to discounting, $C_{\rho,\nu}$ will still be finite whenever γ^m converges to zero faster than $c_{\rho,\nu}(m)$ converges to ∞. In particular, if the rate of divergence of $c_{\rho,\nu}(m)$ is sub-exponential, i.e., if $\Gamma =$

[4] The Radon-Nikodym (RN) derivative is a generalization of the notion of probability densities. According to the Radon-Nikodym Theorem, $d\alpha/d\beta$, the RN derivative of α w.r.t. β is well-defined if β is σ-finite and if α is absolute continuous w.r.t. β. In our case β is a probability measure, so it is actually finite.

$\limsup_{m\to\infty} 1/m \log c_{\rho,\nu}(m) \le 0$ then $C_{\rho,\nu}$ will be finite. In the stochastic process literature, Γ is called the top-Lyapunov exponent of the system and the condition $\Gamma \le 0$ 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^∞-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 ν. Again, if the system is "noisy" then C_ν is finite: In fact, the noisier the dynamics is (the less control we have), the smaller C_ν is. As a side-note, let us remark that $C_{\rho,\nu} \le C_\nu$ holds for any measures ρ, ν. (This follows directly from the 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.[5] 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 pseudo-metric space $\mathcal{M} = (\mathcal{M}, d)$.[6] We say that $\mathcal{M}$ is covered by m discs $D_1, \dots, D_m$ if $\mathcal{M} \subset \bigcup_j D_j$. 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 ε. 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 $x^{1:N} = (x_1, x_2, \dots, x_N)$ in $\mathcal{X}$, we use the *empirical covering numbers*, i.e., the covering number of $\mathcal{F}$ equipped with the empirical ℓ^1 pseudo-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 capacity measure widely used in the nonparametric statistics literature 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:

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

[6] A pseudo-metric satisfies all the properties of a metric except that the requirement of distinguishability is removed.

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 a policy that is greedy w.r.t. a previously computed action-value function Q'. Thus, if $\hat{\pi}$ denotes the chosen greedy policy, then we will jointly select M functions (one for each action of $\mathcal{A}$) from $\mathcal{F}$ through (7) and (8). It follows that we will ultimately need a covering number bound for the set

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

Since Q' depends on the data (a collection of random variables), Q' is random, hence $\hat{\pi}$ is random, and thus the above set is random, too. In order to deal with this, we consider the following, non-random superset of $\mathcal{F}^\vee_{\hat{\pi}}$:

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

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:

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

(ties should be 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

$$\Xi_{\mathcal{F},M} = \left\{ \xi : \xi = \{A_j\}_{j=1}^M, A_j \subset \mathcal{X}, x \in A_j \Leftrightarrow f_j(x) = \max_{1 \leq k \leq M} f_k(x), f_j \in \mathcal{F} \right\} \tag{16}$$

then we see that

$$\mathcal{F}^\vee = \left\{ \sum_{j=1}^M \mathbb{I}_{\{A_j\}} g_j : \{A_k\} = \xi \in \Xi_{\mathcal{F},M}, g_j \in \mathcal{F} \right\}. \tag{17}$$

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 (17) suggests to view the elements of the set $\mathcal{F}^\vee$ as regression trees defined by the partition system $\Xi_{\mathcal{F},M}$ 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 statement, listing basic properties of VC-crossing dimension, answers this question affirmatively.

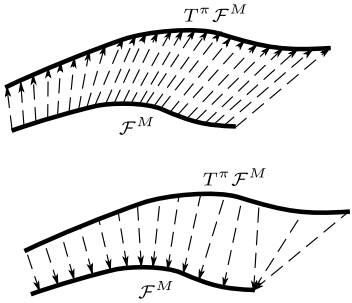

Fig. 3 Illustration of the concepts used to measure the approximation power of the function space $\mathcal{F}^M$. In the *top subfigure* the vectors represent the mapping T^π. On this figure, the measure $E_\infty(\mathcal{F}^M; \pi)$ is the length of the shortest vector. In the *bottom subfigure* the vectors represent the shortest distances of selected points of $T^\pi \mathcal{F}^M$ to $\mathcal{F}^M$. The measure $E_1(\mathcal{F}^M; \pi)$ is the length of the longest of such vectors

Proposition 3 (Properties of VC-crossing dimension) *For any class $\mathcal{F}$ of $\mathcal{X} \to \mathbb{R}$ functions the following statements hold*:

(a) $V_{\mathcal{F}^+} \leq V_{\mathcal{F}^\times}$. *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) $\{\{x \in \mathcal{X} : f(x) \geq a\} : f \in \mathcal{F}, a \in \mathbb{R}\}$ *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 proposition is given in the Appendix. We can now 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 $\mathcal{F}$. We will use two quantities for this purpose, the function space's *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^π is commonly defined as the norm of the difference $Q - T^\pi Q$. If the Bellman-error is small one expects Q to be close to the fixed point of T^π. Hence, it is natural to expect that the final error of fitted policy iteration will be small if for all policies π encountered during the run of the algorithm, we can find some action-value function $Q \in \mathcal{F}^M$ such that $Q - T^\pi Q$ is small. For a fixed policy π, we thus introduce

$$E_\infty(\mathcal{F}^M; \pi) \stackrel{\text{def}}{=} \inf_{Q \in \mathcal{F}^M} \|Q - T^\pi Q\|_\nu.$$

(For an illustration of this quantity see the top subfigure of Fig. 3.) Since we do not know in advance what greedy policies will be seen during the execution of the algorithm, taking a

pessimistic approach, we introduce

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

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

The other quantity, the *inherent one-step Bellman-error of* $\mathcal{F}$, is defined as follows: For a fixed policy π, the one-step Bellman-error of $\mathcal{F}$ w.r.t. T^π is defined as the deviation of $\mathcal{F}^M$ from $T^\pi \mathcal{F}^M$:

$$E_1(\mathcal{F}^M; \pi) \stackrel{\text{def}}{=} \sup_{Q \in \mathcal{F}^M} \inf_{Q' \in \mathcal{F}^M} \|Q' - T^\pi Q\|_\nu.$$

The bottom subfigure of Fig. 3 illustrates this concept. Taking again a pessimistic approach, the *inherent one-step Bellman-error* of $\mathcal{F}$ is defined as

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

The rationale of the 'one-step' qualifier is that T^π 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 $\mathcal{F}^M$. It is the additional term, $\|h - T^\pi f\|_\nu$ that we subtracted in (5) from the unmodified Bellman-error that causes the inherent one-step Bellman-error to enter our bounds.

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

$$E(\mathcal{F}^M) \stackrel{\text{def}}{=} \left(E_\infty^2(\mathcal{F}^M) + E_1^2(\mathcal{F}^M)\right)^{1/2}.$$

$E(\mathcal{F}^M)$ is called the *total inherent Bellman-error* of $\mathcal{F}$.

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

Theorem 4 (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 (8) and the input $\{(X_t, A_t, R_t)\}$, satisfying the mixing assumption, Assumption 2. Let $Q_k \in \mathcal{F}^M$ be the k^{th} iterate $(k = -1, 0, 1, 2, \ldots)$ and let π_{k+1} be greedy w.r.t. Q_k. Choose $\rho \in \mathcal{M}(\mathcal{X})$, a measure used to evaluate the performance of the algorithm and let $0 < \delta \le 1$. Then*

$$\|Q^* - Q^{\pi_K}\|_\rho \le \frac{2\gamma}{(1-\gamma)^2}\left(C_{\rho,\nu}^{1/2}\left(E(\mathcal{F}^M) + \left(\frac{\Lambda_N(\frac{\delta}{K})(\Lambda_N(\frac{\delta}{K})/b \vee 1)^{1/\kappa}}{C_2 N}\right)^{1/4}\right) + \gamma^{K/2} R_{\max}\right) \quad (18)$$

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

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

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

$$V = 3MV_{\mathcal{F}^+} + M_2 V_{\mathcal{F}^\times},$$

$$M_2 = M(M-1),$$

$$\log C_1 = V \log\left(\frac{512 e Q_{\max} \tilde{R}_{\max}}{M\pi_{b0}}\right) + V_{\mathcal{F}^\times} M_2 \log M_2 + V_{\mathcal{F}^+} M \log 2 + M^2 + M_2 \log(V_{\mathcal{F}^\times} + 1) + M \log(V_{\mathcal{F}^+} + 1) + 2\log(MV_{\mathcal{F}^+} + 1) + 2\log(4e),$$

$$C_2 = \frac{1}{2}\left(\frac{M\pi_{b0}}{32\tilde{R}_{\max}^2}\right)^2,$$

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 (18)*, except that in that bound $C_{\rho,\nu}^{1/2}$ has to be replaced by $C_\nu^{1/2}$.*

Before developing the proof, let us make some comments on the form of the bound (18). The bound has three terms, the first two of which are similar to terms that often appear in bounds of the regression literature: In particular, the first term that depends on the total inherent Bellman-error of $\mathcal{F}$, $E(\mathcal{F}^M)$, quantifies the approximation power of $\mathcal{F}$ as discussed beforehand. The next term, after some simplifications and if constant and logarithmic terms are omitted, takes 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. Chap. 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 comes at the price of replacing $E(\mathcal{F}^M)$ in the bound by $(1+\alpha)E(\mathcal{F}^M)$ with some $\alpha > 0$. The last term does not have a counterpart in regression settings, as it comes from 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 in front of the bound. The multiplier $2\gamma/(1-\gamma)^2$ appears in previous L^∞-performance bounds for policy iteration, too (cf. Bertsekas and Tsitsiklis 1996). 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.

As a final remark note that a bound on the difference of the optimal action-value function, Q^*, and the action-value function of π_K, Q^{π_K}, does not immediately yield a bound on the difference of V^* and V^{π_K}. However, with some additional work (by using similar techniques

to the ones used in the proof of Theorem 4) it is possible to derive such a bound by starting with the elementary point-wise bound

$$V^* - V^{\pi_K} \le E^{\pi^*}(Q^* - Q^{\pi_{K-1}} + Q^{\pi_{K-1}} - Q_{K-1}) + E^{\pi_K}(Q_{K-1} - Q^{\pi_{K-1}} + Q^{\pi_{K-1}} - Q^* + Q^* - Q^{\pi_K})$$

which yields to the L^2 bound:

$$1/M\|V^* - V^{\pi_K}\|_\rho^2 \le 2\|Q^* - Q^{\pi_{K-1}}\|_\rho^2 + \|Q^* - Q^{\pi_K}\|_\rho^2 + 2\|Q^{\pi_{K-1}} - Q_{K-1}\|_\rho^2,$$

and where a bound on $\|Q^{\pi_k} - Q_k\|_\rho$ may be derived (similarly as what is done in Lemma 12 by using the point-wise equality: $Q^{\pi_k} - Q_k = (I - \gamma P^{\pi_k})^{-1}(T^{\pi_k} Q_k - Q_k)$) in terms of $\|T^{\pi_k} Q_k - Q_k\|_\nu$ up to a constant (defined similarly as $C_{\rho,\nu}$) times a $1/(1-\gamma)$ factor. The Bellman residual $\|T^{\pi_k} Q_k - Q_k\|_\nu$ being controlled by the BRM algorithm (see e.g. Lemma 10), a bound on $\|V^* - V^{\pi_K}\|_\rho$ then follows. For the sake of compactness, however, we do not explore this bound any further here.

The following sections are devoted to develop the proof of the main 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 risk (7). If the samples were identically distributed and independent of one another, we could use the results developed for empirical processes (e.g., Pollard's inequality). However, in our case the samples are dependent. To deal with this situation, we will use the *blocking device* of Yu (1994) that we introduce now: 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 N samples into $2m_N$ blocks which come in pairs, each having k_N samples:

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

Here $\{Z_t\}$ is the dependent process used by the learning algorithm,

$$H_i \stackrel{\text{def}}{=} \{2(i-1)k_N + j : 1 \le j \le k_N\} \quad \text{and}$$
$$T_i \stackrel{\text{def}}{=} \{(2i-1)k_N + j : 1 \le j \le k_N\}.$$

Next, corresponding to every second block (H_i), we introduce block-independent "ghost" samples as it was done by Yu (1994) and Meir (2000):

$$\underbrace{Z'_1, \dots, Z'_{k_N}}_{H_1}, \underbrace{Z'_{2k_N+1}, \dots, Z'_{3k_N}}_{H_2}, \dots, \underbrace{Z'_{(2m_N-2)k_N+1}, \dots, Z'_{(2m_N-1)k_N}}_{H_{m_N}}. \tag{19}$$

Here any particular block has the same joint marginal distribution as originally, however, these new random variables are constructed such that the m_N new blocks are independent of one another. Introduce $H = \bigcup_{i=1}^{m_N} H_i$.

Our first result generalizes Pollard's tail inequality to β-mixing sequences via the help of this blocking device. This result refines a previous result of Meir (2000). (In order to keep the flow of the developments continuous, the proofs of the statements of these results are given in the Appendix.)

Lemma 5 *Suppose that $Z_1, \ldots, Z_N \in \mathcal{Z}$ is a stationary β-mixing process with mixing coefficients $\{\beta_m\}$ and that $\mathcal{F}$ is a permissible class of $\mathcal{Z} \to [-K, K]$ functions. Then*

$$\mathbb{P}\left(\sup_{f\in\mathcal{F}}\left|\frac{1}{N}\sum_{t=1}^{N} f(Z_t) - \mathbb{E}[f(Z_1)]\right| > \varepsilon\right)$$

$$\leq 16\mathbb{E}[\mathcal{N}_1(\varepsilon/8, \mathcal{F}, (Z'_t; t \in H))]e^{-\frac{m_N\varepsilon^2}{128K^2}} + 2m_N\beta_{k_N+1},$$

where the "ghost" samples $Z'_t \in \mathcal{Z}$ and $H = \bigcup_{i=1}^{m_N} H_i$ are defined above in (19).

Compared with Pollard's tail inequality the main differences are that the exponential term now depends on the number of blocks, and we also have a term that depends on the coefficient sequence $\{\beta_m\}$. For exponential β-mixing both terms decay at an exponential rate. The block size (equivalently the number of blocks) then has to be such that the two terms are balanced. Another difference is that the empirical covering numbers are evaluated on the ghost samples.

Let us now turn to the development of bounds on the covering numbers that we will need. Let Ξ be a family of partitions of $\mathcal{X}$. By a partition 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 Ξ 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 Ξ. The *partitioning number* of Ξ, $\Delta_N^*(\Xi)$, 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 the shatter-coefficient.

Given a class $\mathcal{G}$ of real-valued functions on $\mathcal{X}$ and a partition family Ξ over $\mathcal{X}$, define the set of *Ξ-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, (16), and (17), we have $\mathcal{F}^\vee = \mathcal{F} \circ \Xi_{\mathcal{F},M}$. 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 Ξ 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 $\varepsilon > 0$,

$$\mathcal{N}_1(\varepsilon, \mathcal{G} \circ \Xi, x^{1:N}) \leq \Delta(x^{1:N}, \Xi)\phi_N(\varepsilon)^{m(\Xi)} \leq \Delta_N^*(\Xi)\phi_N(\varepsilon)^{m(\Xi)}. \tag{20}$$

In our next result we refine this bound by replacing the partitioning number by the covering number of the partition family. For arbitrary sets A, B, let $A \triangle B$ denote the symmetric difference of A and B.

Lemma 7 *Let* Ξ, $\mathcal{G}$, $x^{1:N}$, $\phi_N : \mathbb{R}^+ \to \mathbb{R}^+$ *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 pseudo-metric*

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

where

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

and where μ_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 ϕ_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 $\varepsilon > 0$, 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 (20).

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* (17), $M_2 = M(M-1)$, *for every* $\varepsilon > 0$, $\alpha \in (0, 1)$,

$$\mathcal{N}_1(\varepsilon, \mathcal{F}^\vee, x^{1:N}) \leq \mathcal{N}_1\left(\frac{\alpha\varepsilon}{M_2 K}, \mathcal{G}_2^1, x^{1:N}\right)^{M_2}\phi_N((1-\alpha)\varepsilon)^M.$$

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

$$\tilde{E}_1^2(\mathcal{F}^M; \pi) = E_1^2(\mathcal{F}^M; \pi) - \inf_{f,h \in \mathcal{F}^M} \|h - T^\pi f\|_\nu^2. \tag{21}$$

Certainly, $\tilde{E}_1^2(\mathcal{F}^M; \pi) \leq E_1^2(\mathcal{F}^M; \pi)$. 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 $\mathcal{X} \times \mathcal{A}$, $Q'(\omega) \in \mathcal{F}^M$ (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*

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

For $0 < \delta \leq 1$, $N \geq 1$, with probability at least $1 - \delta$,

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

where $\Lambda_N(\delta)$ and C_2 are defined as in Theorem 4. Further, the bound remains true if $E_\infty^2(\mathcal{F}^M; \hat{\pi}) + \tilde{E}_1^2(\mathcal{F}^M; \hat{\pi})$ above is replaced by $E^2(\mathcal{F}^M)$.

By considering the case when $\gamma = 0$ and $M = 1$ 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 ($X_t \sim \nu \in \mathcal{M}(\mathcal{X})$) and β-mixing with exponential rate $(\overline{\beta}, b, \kappa)$. Assume that $\mathcal{F} \subset B(\mathcal{X}; Q_{\max})$ for $Q_{\max} \geq 0$ 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 $0 < \delta \leq 1$, $N \geq 1$, with probability at least $1 - \delta$,

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

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

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 π 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:

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

$$\pi_k = \hat{\pi}(\cdot; Q_{k-1}), \quad k = 0, 1, 2, \ldots.$$

Further, let

$$\varepsilon_k = Q_k - T^{\pi_k} Q_k, \quad k = 0, 1, 2, \ldots \tag{22}$$

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. Note that in the lemma no assumptions are made about how the sequence Q_k is generated:

Lemma 12 *Let $p \geq 1$ be a real, K be a positive integer, and $Q_{\max} \leq R_{\max}/(1-\gamma)$. Then, for any sequence of functions $\{Q_k\} \subset B(\mathcal{X}; Q_{\max})$, $0 \leq k < K$ and ε_k defined by (22) the following inequalities hold*:

$$\|Q^* - Q^{\pi_K}\|_{p,\rho} \leq \frac{2\gamma}{(1-\gamma)^2} \left(C_{\rho,\nu}^{1/p} \max_{0 \leq k < K} \|\varepsilon_k\|_{p,\nu} + \gamma^{K/p} R_{\max} \right), \tag{23}$$

$$\|Q^* - Q^{\pi_K}\|_{\infty} \leq \frac{2\gamma}{(1-\gamma)^2} \left(C_{\nu}^{1/p} \max_{0 \leq k < K} \|\varepsilon_k\|_{p,\nu} + \gamma^{K/p} R_{\max} \right). \tag{24}$$

Proof We have $C_\nu \geq C_{\rho,\nu}$ for any ρ. Thus, if the bound (23) holds for any ρ, choosing ρ to be a Dirac at each state implies that (24) also holds. Therefore, we only need to prove (23).

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

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

Thus, by induction,

$$Q^* - Q^{\pi_K} \leq \gamma \sum_{k=0}^{K-1} (\gamma P^{\pi^*})^{K-k-1} E_k \varepsilon_k + (\gamma P^{\pi^*})^K (Q^* - Q^{\pi_0}). \tag{25}$$

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 of both sides point-wise in (25) 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 $Q^* - Q^{\pi_0} \le \frac{2}{1-\gamma} R_{\max} \mathbf{1}$, we arrive at

$$|Q^* - Q^{\pi_K}| \le \frac{2\gamma(1-\gamma^{K+1})}{(1-\gamma)^2} \left[\sum_{k=0}^{K-1} \alpha_k A_k |\varepsilon_k| + \alpha_K A_K R_{\max} \mathbf{1} \right]. \tag{26}$$

Here we introduced the positive coefficients

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

and the operators

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

Note that $\sum_{k=0}^{K} \alpha_k = 1$. Further, we claim that the operators $A_k \cdot : B(\mathcal{X} \times \mathcal{A}) \to B(\mathcal{X} \times \mathcal{A})$ are positive linear operators and satisfy $A_k \mathbf{1} = \mathbf{1}$. Fix an index k. It is clear that A_k is positive: $A_k Q \ge 0$ whenever $Q \ge 0$. It is also clear that A_k is linear, so it remains to see that A_k leaves $\mathbf{1}$ invariant. From the definition of A_k it is easy to see that it suffices to check that $\frac{1-\gamma}{2} F_k$ possesses this property. For this, it suffices to notice that $(1-\gamma)(I - \gamma P^{\pi_{k+1}})^{-1}$ and $(1-\gamma)(I - \gamma P^{\pi_k})^{-1}$ also possess this property. This follows, however, by, e.g., the Neumann-series expansion of these inverses. Now let us remark that Jensen's inequality holds for positive operators that leave the unity invariant (Kuczma 1985): If A is such an operator and g is a convex function then $g(A_k Q) \le A_k(g \circ Q)$, where g is applied pointwise, as is done the comparison between the two sides.

Let $\lambda_K = [\frac{2\gamma(1-\gamma^{K+1})}{(1-\gamma)^2}]^p$. Taking the p^{th} power of both sides of (26), using Jensen's inequality twice and then integrating both sides w.r.t. $\rho(x,a)$ (using ρ's extension to $\mathcal{X} \times \mathcal{A}$ defined by (3)) we get

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

where we used the shorthand notation introduced in (2). From the definition of the coefficients $c_{\rho,\nu}(m)$,

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

and hence

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

Let $\varepsilon \stackrel{\text{def}}{=} \max_{0\le k<K} \|\varepsilon_k\|_{p,\nu}$. Using the definition of α_k, $C_{\rho,\nu}$ and λ_K we get

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

leading to the desired bound:

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

4.3 Proof of the main result

Now, we are ready to prove Theorem 4.

Proof As in the case of the previous proof, we need to prove only part of the statement that concerns the weighted ρ-norm.

Fix $N, K > 0$, and let ρ and $\mathcal{F}$ be as in the statement of Theorem 4. Consider the iterates Q_k generated by model-free policy iteration with *PEval* defined by (8), when running on the trajectory $\{(X_t, A_t, R_t)\}$ generated by some stochastic stationary policy π_b. Let ν be the invariant measure underlying the stationary process $\{X_t\}$. Let π_K be a policy greedy w.r.t. Q_K. Our aim is to derive a bound on the distance of Q^{π_K} and Q^*. For this, we use Lemma 12. Indeed, if one defines $\varepsilon_k = Q_k - T^{\pi_k} Q_k$ then Lemma 12 with $p = 2$ gives

$$\|Q^* - Q^{\pi_K}\|_\rho \le \frac{2\gamma}{(1-\gamma)^2}\left(C_{\rho,\nu}^{1/2} \max_{0\le k<K} \|\varepsilon_k\|_\nu + \gamma^{K/2} R_{\max}\right). \tag{27}$$

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

$$\|\varepsilon_k\|_\nu \le E(\mathcal{F}^M) + \left(\frac{\Lambda_N(\delta')(\Lambda_N(\delta')/b \vee 1)^{1/\kappa}}{C_2 N}\right)^{1/4} \tag{28}$$

holds everywhere except on a set of probability at most δ'. ($\Lambda_N(\delta')$ and C_2 are defined as in the text of the theorem.) Take $\delta' = \delta/K$. By the choice of δ', the total probability of the set of exceptional events for $0 \le k < K$ is at most δ. Outside of this failure set, we have that (28) holds for all $0 \le k < K$. Combining this with (27), we get

$$\begin{aligned}&\|Q^* - Q^{\pi_K}\|_\rho \\ &\le \frac{2\gamma}{(1-\gamma)^2}\left(C_{\rho,\nu}^{1/2}\left(E(\mathcal{F}^M) + \left(\frac{\Lambda_N(\frac{\delta}{K})(\frac{\Lambda_N(\frac{\delta}{K})}{b} \vee 1)^{1/\kappa}}{C_2 N}\right)^{1/4}\right) + \gamma^{\frac{K}{2}} R_{\max}\right),\end{aligned}$$

thus finishing the proof of the weighted-norm bound. $\square$

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 successes in reinforcement learning, work on value function approximation methods flourished, resulting in

the publication of the books (Bertsekas and Tsitsiklis 1996; Sutton and Barto 1998). Existing theoretical results mostly concern supremum-norm approximation errors (Gordon 1995; Tsitsiklis and Van Roy 1996) and require that the operator projecting the intermediate iterates to the function space chosen by the user to be a non-expansion w.r.t. the supremum-norm. This holds when kernel-smoothing estimation is used such as in the works of Gordon (1995), Tsitsiklis and Van Roy (1996), Guestrin et al. (2001) or Ernst et al. (2005). However, when risk-minimization is involved, the required non-expansion property is lost. Yet these approaches were often found to perform quite satisfactorily in practice (e.g., Wang and Dietterich 1999; Dietterich and Wang 2002; Lagoudakis and Parr 2003). The lack of theoretical results for these approaches served as the main motivation for this work.

To the best of our knowledge this work is unique from yet another point of view: We know of 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 for *finite-horizon*, undiscounted problems. Notice that the finite horizon setting is substantially different from the infinite horizon setting: The algorithm can work backwards (avoiding the use of fixed point equations and arguments!) and thus the learning of an action-value function at any stage becomes a slightly twisted regression problem. In particular, the samples available for learning at any stage will be *independent* of each other since in the finite-horizon framework one naturally assumes that the training data is available as a sequence of independent trajectories.

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 approximate value iteration over action-value functions and obtained asymptotic rates on weak-convergence. Q-learning with interpolative function approximation was considered by Szepesvári and Smart (2004), who derived asymptotic convergence and performance bound guarantees. Both these works carry out the analysis with respect to L^∞-norms and exploit that the function-approximation operator Π 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 analysis methods, the closest to the present work is the work 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. This resulted in a more complex analysis and the need to develop new tools. For dealing with dependent 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 Sect. 4.2 is an extension of a similar technique due to Munos (2003). However, while 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), they did so for analyzing approximate value iteration only.

6 Conclusions and future work

We have considered fitted policy iteration with Bellman-residual minimization and gave high-probability finite-sample bounds on the performance of a policy iteration algorithm for infinite-horizon control learning in an off-policy setting, using function approximators over a continuous state-space. We have also shown that when linearly parameterized value functions are used the new procedure is equivalent to LSPI of Lagoudakis and Parr (2003). Hence, our results apply directly to LSPI, as well.

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

One open question is if the finiteness of the VC-crossing dimension of the function space chosen is really necessary (or finiteness of e.g. the pseudo-dimension suffices). If the finiteness of the VC-crossing dimension proves necessary then it will be important to derive bounds on it for popular function classes, such as regression trees or neural networks.

We have not touched issues such as 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, 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 (a detailed argument along these lines is presented in Munos and Szepesvári 2006). 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, would be to use penalized least-squares. It remains to be seen if this can be done in a reinforcement learning context.

Another possibility is to use different function sets for the representation of the fixed-point candidates and the auxiliary function candidates, or just in the successive iterations of the algorithm.

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. Antos et al. (2007a) makes the first steps in this direction.

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 Chap. 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 bound. Borrowing further ideas from the regression literature, it might be possible to achieve even larger improvement by, e.g., using localization techniques or data-dependent bounds.

As noted beforehand, our results apply to the LSPI algorithm of Lagoudakis and Parr (2003). As to other batch algorithms, the first results for fitted Q-iteration (Ernst et al. 2005) are derived by Antos et al. (2007a), while Antos et al. (2007b) derive results value-iteration based fitted policy iteration. As a next step it would be interesting to unify and compare these results.

Another direction is to lift the condition that the states are observable. If the observations depend on the states in a deterministic way, the results go through except that the achievable performance will be limited by what is observed: The mapping from the states to the observed quantities can be viewed as a restriction on what function approximators can be used. An interesting direction is to add a component to the algorithm that, as the sample size grows, enriches the inputs so that in the limit one still recovers the optimal performance. Less clear is what happens if the observations are noisy because this corresponds to

the "noise in the variables" setting of regression. Another assumption that can be relaxed is the one that requires that all actions are sampled with positive probability in all states. The results still extend to this case, with the only change being that convergence to optimality would be lost (from the point of the learner, the actions that are never sampled are non-existent).

Finally, it would be interesting to compare the result that we obtained with $\gamma = 0$ and $M = 1$ for the regression-case (Corollary 11) with similar results available in the regression literature.

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, András György, Levente Kocsis, and Rich Sutton for the friendly discussions that helped to improve the paper to a great extent. Special thanks to Mohammad Ghavamzadeh who called to our attention that LSPI and our procedure yield the exact same results in the special case when linear function approximation is used.

Appendix

7.1 Proofs of the auxiliary lemmas

Proof of Proposition 3. (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 $f_1(x_i) \leq z_i$ for $x_i \notin A$, and also there is an $f_2 \in \mathcal{F}$ such that $f_2(x_i) \leq z_i$ 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, $V_{\mathcal{F}^\times} = V_{\mathcal{C}_2} \geq V_{\mathcal{F}^+}$. 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 $\mathcal{C}_2 = \{\{x \in \mathcal{X} : f(x) \geq 0\} : f \in \mathcal{F}\}$. 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 $\mathcal{F} = \{I_{\{(a,\infty)\}} : a \in \mathbb{R}\}$. 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 $\sum_{j=1}^{\infty} a_j = 1$, $0 \leq a_j \leq 1/\log_2 j$, and $3a_{j+1} > a_j$. For an integer $n \geq 2$, let $k \geq 1$ and $0 \leq i \leq 2^k - 1$ be the unique integers defined by $n = 2^k + i$. Define

$$f_n(x) = x + \sum_{j=1}^{n} a_j \quad \text{and}$$

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

where $\pi = 3.14159\ldots$ is Ludolf's number. Certainly, f_n and $\tilde{f}_n$ are both differentiable. Note that $a_n \leq a_{2^k} \leq 1/k$, 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 $\mathcal{F}_1 = \{f_n : n \geq 2\}$, $\tilde{\mathcal{F}}_1 = \{\tilde{f}_n : n \geq 2\}$, 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 $\{\{x \in \mathcal{X} : f(x) \geq a\} : f \in \mathcal{F}, a \in \mathbb{R}\}$ 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 $V_{\mathcal{F}^+} \leq 3$. On the other hand, since

$$\begin{aligned}&\{x \in \mathcal{X} : \tilde{f}_n(x) \geq f_n(x)\}\\ &\quad = \{x \in \mathcal{X} : (-1)^{\lfloor i/2^{\lfloor kx \rfloor} \rfloor} \geq 0\} = \{x \in \mathcal{X} : \lfloor i/2^{\lfloor kx \rfloor} \rfloor \text{ is even}\},\end{aligned}$$

so

$$\begin{aligned}\mathcal{C}_2 &\supseteq \{\{x \in \mathcal{X} : \tilde{f}_n(x) \geq f_n(x)\} : n \geq 2\}\\ &= \{\{x \in \mathcal{X} : \lfloor i/2^{\lfloor kx \rfloor} \rfloor \text{ is even}\} : n \geq 2\}.\end{aligned}$$

As this class contains the unions of $\{1\}$ and any of the intervals $[0, 1/k)$, $[1/k, 2/k)$, …, $[1 - 1/k, 1)$ for any k, thus it shatters the points 0, $1/k$, $2/k$, …, $1 - 1/k$, and hence $V_{\mathcal{F}^\times} = V_{\mathcal{C}_2} = \infty$. □

Proof of Lemma 5. The proof uses the following lemma, essentially due to Yu (1994). The lemma is stated without a proof:[7]

Lemma 13 (Yu 1994, 4.2 Lemma) *Suppose that* $\{Z_t\}$, $\mathcal{F}$, $\{Z'_t\}$, $\{H_i\}$, *and* H *are as in Lemma 5. Then*

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

Now, let us turn to the proof of Lemma 5. 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 $\bar{\mathcal{F}} \stackrel{\text{def}}{=} \left\{\bar{f} : f \in \mathcal{F}\right\}$.

We use Lemma 13 to replace the original process by the block-independent one, implying

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

$$\begin{aligned}
&\mathbb{P}\left(\sup_{f\in\mathcal{F}}\left|\frac{1}{N}\sum_{t=1}^{N} f(Z_t)-\mathbb{E}[f(Z_1)]\right|>\varepsilon\right)\\
&\quad=\mathbb{P}\left(\sup_{f\in\mathcal{F}}\left|\frac{1}{N}\sum_{t=1}^{N}(f(Z_t)-\mathbb{E}[f(Z_1)])\right|>\varepsilon\right)\\
&\quad\le 2\mathbb{P}\left(\sup_{f\in\mathcal{F}}\left|\frac{1}{N}\sum_{i=1}^{m_N}(\bar{f}(Z'^{(i)})-k_N\mathbb{E}[f(Z_1)])\right|>\frac{\varepsilon}{2}\right)+2m_N\beta_{k_N+1}\\
&\quad=2\mathbb{P}\left(\sup_{f\in\mathcal{F}}\left|\frac{1}{m_N}\sum_{i=1}^{m_N}\bar{f}(Z'^{(i)})-k_N\mathbb{E}[f(Z_1)]\right|>k_N\varepsilon\right)+2m_N\beta_{k_N+1}. \qquad (29)
\end{aligned}$$

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

$$\begin{aligned}
&\mathbb{P}\left(\sup_{f\in\mathcal{F}}\left|\frac{1}{m_N}\sum_{i=1}^{m_N}\bar{f}(Z'^{(i)})-k_N\mathbb{E}[f(Z_1)]\right|>k_N\varepsilon\right)\\
&\qquad\le 8\mathbb{E}[\mathcal{N}_1(k_N\varepsilon/8,\bar{\mathcal{F}},(Z'^{(1)},\ldots,Z'^{(m_N)}))]e^{-\frac{m_N\varepsilon^2}{128K^2}}. \qquad (30)
\end{aligned}$$

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[8]

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

This, together with (29) and (30) gives the desired bound. □

Proof of Lemma 7. Fix $x_1,\ldots,x_N\in\mathcal{X}$ and $\varepsilon>0$. First note that d is indeed a pseudo-metric as it follows from an elementary argument. Let $\widehat{\Xi}$ be an $\alpha\varepsilon/(2K)$-cover for Ξ according to d such that $|\widehat{\Xi}|=\mathcal{N}(\frac{\alpha\varepsilon}{2K},\Xi,d)$. If $f\in\mathcal{G}\circ\Xi$, then there is a partition $\xi=\{A_j\}\in\Xi$ and

[8]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.

functions $g_j \in \mathcal{G}$ such that

$$f = \sum_{A_j \in \xi} g_j \mathbb{I}_{\{A_j\}}. \tag{31}$$

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 $\mathcal{F}_j = \mathcal{F}_j(\xi')$ 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 (31) 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

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

Hence, $\{\sum_j f_j \mathbb{I}_{\{A'_j\}} : f_j \in F_j(\xi'), \xi' \in \widehat{\Xi}\}$ gives an ε-cover of $G \circ \Xi$. The cardinality of this set, $\sum_{\xi' \in \widehat{\Xi}} \prod_{j=1}^{|\xi'|} |\mathcal{F}_j(\xi')|$, is bounded by $\sum_{\xi' \in \widehat{\Xi}} \phi_N((1-\alpha)\varepsilon)^{|\xi'|} \le \mathcal{N}(\frac{\alpha\varepsilon}{2K}, \Xi, d)\phi_N((1-\alpha)\varepsilon)^{m(\Xi)}$, finishing the proof. □

Proof of Lemma 8. Since $\mathcal{F}^\vee = \mathcal{F} \circ \Xi$ for $\Xi = \Xi_{\mathcal{F},M}$ defined in (16),

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

By Lemma 7 this is bounded by

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

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

For $f : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$ ($f \in \mathcal{F}^M$), 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, but systematic way) and their class $\mathcal{G} = \{I_f : f \in \mathcal{F}^M\}$.

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

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

Furthermore, if $\mathcal{G}_M^1$ 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 $f : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$ ($f \in \mathcal{F}^M$), then, since the support of a function from $\mathcal{G}$ is the disjoint union of the supports (on different instances of $\mathcal{X}$) of M functions from $\mathcal{G}_M^1$, it is easy to see that (cf., e.g., Devroye et al. 1996, Theorem 29.6)

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

Now, since a function from $\mathcal{G}_M^1$ is the product of $M - 1$ 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)

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

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

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

Lemma 14 *Let* $\beta_m \le \overline{\beta} \exp(-bm^\kappa)$, $N \ge 1$, $k_N = \lceil (C_2 N\varepsilon^2/b)^{\frac{1}{1+\kappa}} \rceil$, $m_N = N/(2k_N)$, $0 < \delta \le 1$, $V \ge 2$, *and* C_1, C_2, $\overline{\beta}$, b, $\kappa > 0$. *Further, define* ε *and* Λ *by*

$$\varepsilon = \sqrt{\frac{\Lambda(\Lambda/b \vee 1)^{1/\kappa}}{C_2 N}} \tag{32}$$

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

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

Proof of Lemma 14. We have

$$\max((C_2 N\varepsilon^2/b)^{\frac{1}{1+\kappa}}, 1) \le k_N \le \max(2(C_2 N\varepsilon^2/b)^{\frac{1}{1+\kappa}}, 1)$$

and so

$$\frac{N}{4}\min\left(\frac{b}{C_2 N\varepsilon^2}, 1\right)^{\frac{1}{1+\kappa}} \le \frac{N}{4}\min\left(\left(\frac{b}{C_2 N\varepsilon^2}\right)^{\frac{1}{1+\kappa}}, 2\right) \le m_N = \frac{N}{2k_N} \le \frac{N}{2}.$$

Obviously, $\Lambda \geq 1$ and from (32),

$$\varepsilon \geq \sqrt{\Lambda/(C_2 N)} \geq \sqrt{1/(C_2 N)} \quad \text{and} \quad C_2 N \varepsilon^2 = \Lambda(\Lambda/b \vee 1)^{1/\kappa}. \tag{33}$$

Substituting the proper bounds for β_m, k_N, and m_N, we get

$$\begin{aligned}
&C_1 \left(\frac{1}{\varepsilon}\right)^V e^{-4C_2 m_N \varepsilon^2} + 2m_N \beta_{k_N} \\
&\leq C_1 \left(\frac{1}{\varepsilon}\right)^V e^{-(\frac{b}{C_2 N\varepsilon^2} \wedge 1)^{\frac{1}{1+\kappa}} C_2 N\varepsilon^2} + N\overline{\beta} e^{-b(\frac{C_2 N\varepsilon^2}{b} \vee 1)^{\frac{\kappa}{1+\kappa}}} \\
&= C_1 \left(\frac{1}{\varepsilon}\right)^V e^{-(\frac{b}{C_2 N\varepsilon^2} \wedge 1)^{\frac{1}{1+\kappa}} C_2 N\varepsilon^2} + N\overline{\beta} e^{-b(\frac{C_2 N\varepsilon^2}{b} \vee 1)(\frac{b}{C_2 N\varepsilon^2} \wedge 1)^{\frac{1}{1+\kappa}}} \\
&\leq \left(C_1 \left(\frac{1}{\varepsilon}\right)^V + N\overline{\beta}\right) e^{-(\frac{b}{C_2 N\varepsilon^2} \wedge 1)^{\frac{1}{1+\kappa}} C_2 N\varepsilon^2},
\end{aligned}$$

which, by (33), is upper bounded by

$$(C_1 (C_2 N)^{V/2} + N\overline{\beta}) e^{-(\frac{b}{\Lambda(\Lambda/b\vee 1)^{1/\kappa}} \wedge 1)^{\frac{1}{1+\kappa}} \Lambda(\Lambda/b\vee 1)^{1/\kappa}}.$$

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

$$(C_1 (C_2 N)^{V/2} + N\overline{\beta}) N^{-V/2} \frac{\delta}{e(C_1 C_2^{V/2} \vee \overline{\beta} \vee 1)} \leq (1+1)\frac{\delta}{e} < \delta. \qquad \square$$

7.2 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 $\hat{\pi}$,

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

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 $T^{\hat{\pi}(\cdot;Q')} f'$ with high probability, noting that Q' may *not* be independent of 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 $V \geq M(M+2) \geq 3$. Let ε and $\Lambda_N(\delta)$ be chosen as in (32):

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

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

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

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

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

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

Let us now introduce the following additional shorthand notations:

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

where $\hat{L}_N$ was defined in (7). Further, define

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

Now,

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

where in the second last line we used the definition of $\tilde{E}_1^2$ (cf. (21)) and where $\bar{L}_{\mathcal{F},Q'} = \inf_{f\in\mathcal{F}^M} \bar{L}(f;Q')$ is the error of the function with minimum loss in our class. Define also

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

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

$$\begin{aligned}
&\bar{L}(f';Q') - \bar{L}_{\mathcal{F},Q'} \\
&\quad = \bar{L}(f';Q') - \bar{\hat{L}}_N(f';Q') + \bar{\hat{L}}_N(f';Q') - \inf_{f\in\mathcal{F}^M} \bar{L}(f;Q') \\
&\quad \le |\bar{\hat{L}}_N(f';Q') - \bar{L}(f';Q')| + \inf_{f\in\mathcal{F}^M} \bar{\hat{L}}_N(f;Q') - \inf_{f\in\mathcal{F}^M} \bar{L}(f;Q')
\end{aligned}$$

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

Thus we get

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

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

We follow the line of proof due to Meir (2000). For any $f, h, Q' \in \mathcal{F}^M$, 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 (7) as

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

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

Observe that by (10),

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

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

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

and (by (12))

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

(coincidentally with (9), but note that $\mathbb{E}[\bar{\hat{L}}_N(f;Q')] \ne \bar{L}(f;Q')$). This reduces the bound to a uniform tail probability of an empirical process over $\mathcal{L}_{\mathcal{F}}$:

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

Now we make use of the blocking device mentioned previously: Recall that the "ghost" samples $\{Z'_t\}$ and H are defined above at (19). We use Lemma 5 with $\mathcal{Z} = \mathcal{X} \times \mathcal{A} \times \mathbb{R} \times \mathcal{X}$ and $\mathcal{F} = \mathcal{L}_{\mathcal{F}}$ noting that any $l_{f,h,Q'} \in \mathcal{L}_{\mathcal{F}}$ is bounded by

$$K = \frac{\tilde{R}_{\max}^2}{M\pi_{b0}}$$

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

$$\mathbb{P}\left(\sup_{Q',f,h \in \mathcal{F}^M} \left| \frac{1}{N} \sum_{t=1}^{N} l_{f,h,Q'}(Z_t) - \mathbb{E}[l_{f,h,Q'}(Z_1)] \right| > \varepsilon/2 \right)$$
$$\leq 16\mathbb{E}[\mathcal{N}_1(\varepsilon/16, \mathcal{L}_{\mathcal{F}}, (Z'_t; t \in H))] e^{-\frac{m_N}{2}(\frac{M\pi_{b0}\varepsilon}{16\tilde{R}_{\max}^2})^2} + 2m_N\beta_{k_N}.$$

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

$$\frac{2}{N}\sum_{t\in H} |l_{f,h,Q'}(Z'_t) - l_{g,\tilde{h},\tilde{Q}'}(Z'_t)|$$
$$\leq \frac{2\tilde{R}_{\max}}{M\pi_{b0}} \left(\frac{2}{N}\sum_{t\in H} |f(X'_t, A'_t) - g(X'_t, A'_t)| + \frac{2}{N}\sum_{t\in H} |\tilde{h}(X'_t, A'_t) - h(X'_t, A'_t)| \right.$$
$$\left. + 2\frac{2}{N}\sum_{t\in H} |f(X'_{t+1}, \hat{\pi}(X'_{t+1}; Q')) - g(X'_{t+1}, \hat{\pi}(X'_{t+1}; \tilde{Q}'))| \right).$$

Note that the first and second terms are $\mathcal{D}' = ((X'_t, A'_t); t \in H)$-based ℓ^1-distances of functions in $\mathcal{F}^M$, 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

$$\mathcal{N}_1\left(\frac{8\tilde{R}_{\max}}{M\pi_{b0}}\varepsilon', \mathcal{L}_{\mathcal{F}}, (Z'_t; t \in H) \right) \leq \mathcal{N}_1^2(\varepsilon', \mathcal{F}^M, \mathcal{D}')\mathcal{N}_1(\varepsilon', \mathcal{F}^\vee, \mathcal{D}'_+).$$

Applying now Lemma 8 with $\alpha = 1/2$,[9] the covering number of $\mathcal{F}^\vee$ is bounded by

$$\mathcal{N}_1\left(\frac{\varepsilon'}{2M_2 Q_{\max}}, \mathcal{G}_2^1, \mathcal{D}'_+ \right)^{M_2} \phi_{N/2}(\varepsilon'/2)^M,$$

where $M_2 = M(M-1)$, $\mathcal{G}_2^1$ 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 $V_{\mathcal{F}^+}$ and $V_{\mathcal{F}^\times} < \infty$, respectively, and the range of functions from $\mathcal{F}$ has length $2Q_{\max}$. By the pigeonhole principle, it is easy to see that the pseudo-dimension of $\mathcal{F}^M$ cannot exceed $MV_{\mathcal{F}^+}$. Thus

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

$$
\begin{aligned}
&\mathcal{N}_1\left(\frac{8\tilde{R}_{\max}}{M\pi_{b0}}\varepsilon', \mathcal{L}_{\mathcal{F}}, (Z'_t; t \in H)\right) \\
&\quad \le \left(e(MV_{\mathcal{F}^+}+1)\left(\frac{4eQ_{\max}}{\varepsilon'}\right)^{MV_{\mathcal{F}^+}}\right)^2 \\
&\qquad \times \left(e(V_{\mathcal{F}^\times}+1)\left(\frac{4eM_2Q_{\max}}{\varepsilon'}\right)^{V_{\mathcal{F}^\times}}\right)^{M_2}\left(e(V_{\mathcal{F}^+}+1)\left(\frac{8eQ_{\max}}{\varepsilon'}\right)^{V_{\mathcal{F}^+}}\right)^{M} \\
&\quad = e^{M^2+2}(MV_{\mathcal{F}^+}+1)^2(V_{\mathcal{F}^+}+1)^M(V_{\mathcal{F}^\times}+1)^{M_2}2^{MV_{\mathcal{F}^+}}M_2^{M_2V_{\mathcal{F}^\times}}\left(\frac{4eQ_{\max}}{\varepsilon'}\right)^V,
\end{aligned}
$$

where $V = 3MV_{\mathcal{F}^+} + M_2V_{\mathcal{F}^\times}$ is the "effective" dimension, and thus

$$
\begin{aligned}
&\mathcal{N}_1(\varepsilon/16, \mathcal{L}_{\mathcal{F}}, (Z'_t; t \in H)) \\
&\quad \le e^{M^2+2}(MV_{\mathcal{F}^+}+1)^2(V_{\mathcal{F}^+}+1)^M(V_{\mathcal{F}^\times}+1)^{M_2} \cdot \\
&\qquad \times 2^{MV_{\mathcal{F}^+}}M_2^{M_2V_{\mathcal{F}^\times}}\left(\frac{512eQ_{\max}\tilde{R}_{\max}}{M\pi_{b0}\varepsilon}\right)^V = \frac{C_1}{16}\left(\frac{1}{\varepsilon}\right)^V,
\end{aligned}
$$

with $C_1 = C_1(M, V_{\mathcal{F}^+}, V_{\mathcal{F}^\times}, Q_{\max}, \hat{R}_{\max}, \gamma, \pi_{b0})$. 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

$$
P_0 \le C_1\left(\frac{1}{\varepsilon}\right)^V e^{-4C_2m_N\varepsilon^2} + 2m_N\beta_{k_N}, \tag{34}
$$

where $C_2 = \frac{1}{2}(\frac{M\pi_{b0}}{32\tilde{R}^2_{\max}})^2$. 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 14, which, together with (34), implies $P_0 < \delta$.

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

References

Anthony, M., & Bartlett, P. L. (1999). *Neural network learning: theoretical foundations*. Cambridge: Cambridge University Press.

Antos, A., Szepesvári, C., & Munos, R. (2006). Learning near-optimal policies with Bellman-residual minimization based fitted policy iteration and a single sample path. In G. Lugosi & H. Simon (Eds.), *LNCS/LNAI: Vol. 4005. Proceedings of the nineteenth annual conference on learning theory, COLT 2006* (pp. 574–588), Pittsburgh, PA, USA, 22–25 June 2006. Berlin: Springer.

Antos, A., Munos, R., & Szepesvári, C. (2007a, in press). Fitted Q-iteration in continuous action-space MDPs. In *Advances in neural information processing systems*.

Antos, A., Szepesvári, C., & Munos, R. (2007b). Value-iteration based fitted policy iteration: learning with a single trajectory. In *IEEE symposium on approximate dynamic programming and reinforcement learning (ADPRL 2007)* (pp. 330–337), Honolulu, HI, 1–5 April 2007. New York: IEEE.

Baraud, Y., Comte, F., & Viennet, G. (2001). Adaptive estimation in autoregression or β-mixing regression via model selection. *Annals of Statistics*, *29*, 839–875.

Bellman, R., & Dreyfus, S. (1959). Functional approximation and dynamic programming. *Mathematical Tables and Other Aids to Computation*, *13*, 247–251.

Bertsekas, D. P., & Shreve, S. (1978). *Stochastic optimal control (the discrete time case)*. New York: Academic Press.

Bertsekas, D. P., & Tsitsiklis, J. N. (1996). *Neuro-dynamic programming*. Belmont: Athena Scientific.

Bradtke, S., & Barto, A. (1996). Linear least-squares algorithms for temporal difference learning. *Machine Learning*, *22*, 33–57.

Carrasco, M., & Chen, X. (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: McGraw-Hill.

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

Devroye, L., Györfi, L., & Lugosi, G. (1996). A probabilistic theory of pattern recognition. In *Applications of mathematics: stochastic modelling and applied probability*. New York: Springer.

Dietterich, T. G., & Wang, X. (2002). Batch value function approximation via support vectors. In T. G. Dietterich, S. Becker, & Z. Ghahramani (Eds.), *Advances in neural information processing systems* (Vol. 14). MIT Press: Cambridge.

Doukhan, P. (1994). *Lecture notes in statistics: Vol. 85. Mixing properties and examples lecture notes in statistics*. Berlin: Springer.

Ernst, D., Geurts, P., & Wehenkel, L. (2005). Tree-based batch mode reinforcement learning. *Journal of Machine Learning Research*, *6*, 503–556.

Gordon, G. (1995). Stable function approximation in dynamic programming. In A. Prieditis & S. Russell (Eds.), *Proceedings of the twelfth international conference on machine learning* (pp. 261–268). Kaufmann: San Francisco.

Guestrin, C., Koller, D., & Parr, R. (2001). Max-norm projections for factored MDPs. In *Proceedings of the international joint conference on artificial intelligence*.

Györfi, L., Kohler, M., Krzyżak, A., & Walk, H. (2002). *A distribution-free theory of nonparametric regression*. Berlin: Springer.

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.

Howard, R. A. (1960). *Dynamic programming and Markov processes*. MIT Press: Cambridge.

Kuczma, M. (1985). *An introduction to the theory of functional equations and inequalities*. Katowice: Silesian University Press.

Lagoudakis, M., & Parr, R. (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., & Tweedie, R. (1993). *Markov chains and stochastic stability*. New York: Springer.

Munos, R. (2003). Error bounds for approximate policy iteration. In *19th International conference on machine learning* (pp. 560–567)

Munos, R., & Szepesvári, C. (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., & Sen, S. (2002). Kernel-based reinforcement learning. *Machine Learning*, *49*, 161–178.

Pollard, D. (1984). *Convergence of stochastic processes*. New York: Springer.

Pollard, D. (1990). Empirical processes: theory and applications. In *NSF-CBMS regional conference series in probability and statistics*. Institute of Mathematical Statistics, Hayward, CA.

Precup, D., Sutton, R., & Dasgupta, S. (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 E. A. Feigenbaum & J. Feldman (Eds.) *Computers and thought*. New York: McGraw-Hill (1963)

Schweitzer, P., & Seidmann, A. (1985). Generalized polynomial approximations in Markovian decision processes. *Journal of Mathematical Analysis and Applications*, *110*, 568–582.

Sutton, R., & Barto, A. (1998). *Reinforcement learning: an introduction*. MIT Press: Bradford Book.

Szepesvári, C., & Munos, R. (2005). Finite time bounds for sampling based fitted value iteration. In *ICML'2005* (pp. 881–886).

Szepesvári, C., & Smart, W. (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., & Van Roy, B. (1996). Feature-based methods for large scale dynamic programming. *Machine Learning*, *22*, 59–94.

Wang, X., & Dietterich, T. (1999). Efficient value function approximation using regression trees. In *Proceedings of the IJCAI workshop on statistical machine learning for large-scale optimization*, Stockholm, Sweden.

Williams, R. J., & Baird III L. (1994). Tight performance bounds on greedy policies based on imperfect value functions. In *Proceedings of the tenth yale workshop on adaptive and learning systems*.

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