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

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

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

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

[2] Institut National de Recherche en Informatique et en Automatique
Sequential Learning (SequeL) project, INRIA Futurs Lille, France
e-mail: remi.munos@inria.fr

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

Abstract We consider the problem of finding a near-optimal policy using value-function methods in continuous space, discounted Markovian Decision Problems (MDP) when only a single trajectory underlying some policy can be used as the input. Since the state-space is continuous, one must resort to the use of function approximation. In this paper we study a policy iteration algorithm iterating over action-value functions where the iterates are obtained by empirical risk minimization, where the loss function used penalizes high magnitudes of the Bellman-residual. It turns out that when a linear parameterization is used the algorithm is equivalent to least-squares policy iteration. 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. 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.

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

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

1 Introduction

In 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 *behaviour policy* (Sutton and Barto, 1987, Chapter 5.6).

In this paper we are interested in designing algorithms and proving bounds on the achievable performance for this setting. More specifically, we assume that the control task can be modelled as a discounted Markovian Decision Problem with continuous state-variables and a finite number of actions.

The algorithm studied is an instance of fitted policy iteration: in its main loop the algorithm computes an evaluation function of the policy of the previous step and then uses this evaluation function to compute the next improved policy. In order to avoid the need of learning a model, action-value evaluation functions are employed, making the policy improvement step trivial, just like in the least-squares policy iteration (LSPI) algorithm of Lagoudakis and Parr (2003). However, unlike LSPI which builds on least-squares temporal difference learning (LSTD) due to Bradtke and Barto (1996), we build our algorithm on the idea of minimizing Bellman-residuals. The idea of using Bellman-residuals in policy iteration goes back at least to Schweitzer and Seidmann (1985), who proposed it for computing approximate state-value functions given the model of a finite-state and action MDP.

Both LSTD and Bellman-residual minimization (BRM) assume that the user selects a function class to represent action-value functions and both aim at solving Bellman's fixed-point equation for the current policy in an approximate manner over the chosen set of functions. A popular choice is to use linearly parameterized functions. 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, $T^\pi Q^\pi - Q^\pi = 0$, where Q^π is the policy's action-value function and T^π is the policy's evaluation operator (these will be fully defined in the next section). Then the Bellman-residual of a function Q is $T^\pi Q - Q$. When this residual function is evaluated at a point, we call the resulting value the Bellman-error. A reasonable goal is then 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. One major obstacle for (direct) Bellman-residual minimization is that the average of the individual squared Bellman-errors when computed along a trajectory does not give rise to an unbiased estimate of the squared 2-norm of the Bellman-residual (e.g., Sutton and Barto, 1987, pp. 200). Since LSTD does not suffer from this problem, recently the focus shifted to algorithms that use LSTD.

Here we propose to overcome the biasedness issue by modifying the loss function. The novel loss function depends on an auxiliary function that makes sure that the empirical loss is an unbiased estimate of the population-based loss (Lemma 1). It turns out that when the functions available in the optimization step use a linear parameterization the minimizers of the new loss functions and the solution returned by LSTD coincide (Proposition 2). In this sense the novel loss function generalizes LSTD and the new policy iteration algorithm generalizes LSPI.

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 $1/N^{1/4}$ (where N is the length of the trajectory) up to a limit set by the choice of the function set chosen. 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 of-policy setting, i.e., learning from a single trajectory of some fixed behaviour policy.

One major technical difficulty of the proof is that we have to deal with dependent samples. The main condition here is that the trajectory should be sufficiently representative and rapidly mixing. For the sake of simplicity, we also require that the states in the trajectory follow a stationary distribution, though we believe that with some additional work this condition could be 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 the case of a set of linearly parameterized functions whose VC-crossing dimension just 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, though, 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 be of no surprise: If for some policy encountered while executing the algorithm no function in the chosen set has a small Bellman-residual, the quality of the final solution might well degrade.

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. The difficulty is that since the policy evaluation-functions are obtained via a least-squares approach, it can be difficult to derive good L^∞-bounds on the errors of the intermediate steps.

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 should be sub-exponential. In general, this holds for systems with "noisy" transitions, but, as argued by Munos and Szepesvári (2006), under certain conditions even deterministic systems can meet this condition.

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

2 Definitions

As we shall work with continuous spaces we will need some simple measure theoretic concepts. These are introduced first. This is followed by the introduction of Markovian Decision Problems (MDPs) and the associated concepts and the necessary notation.

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

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

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

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

We make the following assumptions on the MDP:

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

A policy is defined as a (measurable) mapping from past observations to a distribution over the set of actions (for details, see Bertsekas and Shreve, 1978). A policy is called Markov if the distribution depends only on the last state of the observation sequence. A policy is called stationary Markov if this dependency does not change by time. For a stationary Markov policy, the probability distribution over the actions given some state x will be denoted by $\pi(\cdot|x)$. A policy is deterministic if the probability distribution concentrates on a single action for all histories. Deterministic stationary Markov policies will be identified by 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, by our previous remark, we write $A_t = \pi(X_t)$. The function V^π is also called the (state) value function of policy π. Closely related to 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 from state x, the first action is a while all the subsequent actions are determined by the policy π. It is known that for any policy π, the functions V^π,Q^π are bounded by $R_{\max}/(1 - \gamma)$.

Given and 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, $Q^*(x,a)$:

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

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. Greedy policies are important because the greedy policy w.r.t. Q^* is optimal (e.g., Bertsekas and Shreve, 1978). Hence it suffices to determine Q^*. Further, without the loss of generality we can restrict our attention to the set of deterministic, stationary Markov policies. In what follows we shall use the word 'policy' to mean such policies.

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

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

We define two operators corresponding to 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 and given that future states are evaluated by V and there are no immediate rewards. The left-linear operator, $\cdot P : M(\mathcal{X} \times \mathcal{A}) \to M(\mathcal{X})$, is defined by

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

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

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

For a measure ρ defined over the set of state-action pairs, ρP represents the measure of the future state sampled from the transition probability kernel P and given that the initial state and action is 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, $\cdot P^\pi : M(\mathcal{X} \times \mathcal{A}) \to M(\mathcal{X} \times \mathcal{A})$, is defined as follows: Let U be a measurable subset of $\mathcal{X} \times \mathcal{A}$. Given $\rho \in M(\mathcal{X} \times \mathcal{A})$, $(\rho P^\pi)(U) = \rho PE^\pi \mathbb{I}_{\{U\}}$. This can be given a probabilistic interpretation, too, but we have not found this interpretation very intuitive and hence 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 $\mathcal{F}^L$ as real-valued functions f defined over $\mathcal{X} \times \mathcal{A}$ with the obvious identification $f \equiv (f_1, \ldots, f_L)$, $f(x, a_j) = f_j(x)$, $j = 1, \ldots, L$. The set $\mathcal{F}^L$ will denote the set of admissible functions used in the optimization step of our algorithm.

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

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

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

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

```
**FittedPolicyQ**(D,K,$Q_{-1}$,PEval,$\pi$)
// 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($\hat{\pi}(\cdot; Q'), D, \pi$)
end for
return $Q$ // or $\hat{\pi}(\cdot; Q)$, the greedy policy w.r.t. $Q$
```

Fig. 1 Model-free Fitted Policy Iteration

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 ceil 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 Figure 1. By assumption, the training sample, D, used by the algorithm consists of a finite trajectory

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

of some stochastic stationary policy π: $A_t \sim \pi(\cdot|X_t)$, $X_{t+1} \sim P(\cdot|X_t, A_t)$, $R_t \sim S(\cdot|X_t, A_t)$. We assume that this trajectory is sufficiently representative in a sense that will be made precise in the next section. For now, let us make the assumption that X_t is stationary and is distributed according to some (unknown) distribution ν. The action-evaluation function Q_{-1} is used to initialize the first policy (alternatively, one may start with an arbitrary initial policy at the price of making the algorithm somewhat more complicated). Procedure *PEval* takes data in the form of a long trajectory and some policy $\hat{\pi}$ 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. 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, $Q - T^\pi Q$ becomes non-zero. This quantity is called the *Bellman-residual* of Q. It is known that if the magnitude of $Q - T^\pi Q$ is small then Q is a good approximation to Q^π (for an analysis using supremum norms see, e.g., Williams and Baird, 1994). Hence it is expected that a smaller risk, $\|Q - T^\pi Q\|$, yields better estimates.

where $\|\cdot\|$ is some norm chosen appropriately. We choose here the L^2-norm as it leads to an optimization problem with favourable characteristics and makes the connection to regression function estimation easier. Hence, let us consider 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 $\|Q\|_\nu^2 = 1/L \sum_{j=1}^L \|Q(\cdot,a_j)\|_\nu$.) 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 over 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 weights to the actions that are sampled more often. Of course, the same issue exists for 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 due to the additional learning component. To simplify the presentation and the analysis we do not consider such a correction here. In fact, if the behavior policy were not known, we could still use the joint distribution of (X_t, A_t) in the above norm without changing much in our algorithm and results.

When evaluating π, we chase $Q = \operatorname{argmin}_{f\in\mathcal{F}^L} L(f;\pi)$.[1] A simple idea to derive a sample based loss to $L(f;\pi)$ could be to first replace $\|\cdot\|_\nu$ by its empirical counterpart,

$$\|f\|_{\nu,N} = \frac{1}{NL}\sum_{t=1}^N\sum_{j=1}^L \frac{\mathbb{I}_{\{A_t=a_j\}}}{\pi(a_j|X_t)} f(X_t,A_t),$$

(since the states in the trajectory $\{X_t\}$ follow ν, $\|f\|_{\nu,N}$ is expected to converge to $\|f\|_\nu$ 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 $\mathbb{E}\left[R_t+\gamma f(X_{t+1},\pi(X_{t+1}))|X_t,A_t\right] = (T^\pi f)(X_t,A_t)$). This results in the loss function

$$\hat{L}_N(f;\pi) = \frac{1}{NL}\sum_{t=1}^N\sum_{j=1}^L \frac{\mathbb{I}_{\{A_t=a_j\}}}{\pi(a_j|X_t)} \left(f(X_t,a_j) - (R_t+\gamma f(X_{t+1},\pi(X_{t+1})))\right)^2 . \tag{4}$$

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

However, as it is well known (see Sutton and Barto, 1987, pp. 200, Munos, 2003, or Lagoudakis and Parr, 2003 for discussions), $\hat{L}_N$ is *not* an unbiased estimate of the L^2 Bellman-error: $\mathbb{E}\left[\hat{L}_N(f;\pi)\right] \neq L(f;\pi)$. Indeed, elementary calculus shows that for $Y \sim P(\cdot|x,a)$, $R \sim S(\cdot|x,a)$,

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

It follows that minimizing $\hat{L}_N(f;\pi)$ in the limit when $N \to \infty$ is equivalent to minimizing the sum of $\gamma^2 \frac{1}{L}\sum_{j=1}^{L} \mathbb{E}\left[\mathrm{Var}\left[f(Y,\pi(Y))|X, A = a_j\right]\right]$ and $L(f;\pi)$ 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 $\mathrm{Var}\left[f(Y,\pi(Y))|X, A = a_j\right]$ 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, 1987, pp. 200). However, this is neither realistic nor sample efficient unless there is a (cheap) way to generate samples – a framework that we do not consider here. Another possibility, motivated by the double-sample proposal, would be to reuse samples that are close in space (e.g., use nearest neighbors). The difficulty with this approach is that it requires a definition of 'proximity'. Hence, we pursue here an alternative approach that avoids these pitfalls.

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

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

and we propose to solve for

$$\hat{f} = \underset{f \in \mathcal{F}^L}{\mathrm{argmin}} \sup_{h \in \mathcal{F}^L} L(f,h;\pi), \qquad (6)$$

where the supremum in h 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 appears that we wanted to get rid of. 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 $h_f^* \in \mathcal{F}^L$ be a function that minimizes $\|h - T^\pi f\|_\nu^2$. Then

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

Thus if $\left\|h_f^* - T^\pi f\right\|_\nu^2$ 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 just note that for $f \in \mathcal{F}^L$, $L(f, h_f^*;\pi) \geq 0$. This inequality holds because by the definition of h_f^*, $L(f, h_f^*;\pi) \geq L(f, h;\pi)$ holds for any $h \in \mathcal{F}^L$. 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

$$\begin{aligned}\hat{L}_N(f,h;\pi) = \frac{1}{NL}\sum_{t=1}^{N}\sum_{j=1}^{L}\frac{\mathbb{I}_{\{A_t=a_j\}}}{\pi(a_j|X_t)}\Big(&(f(X_t,a_j) - (R_t + \gamma f(X_{t+1},\pi(X_{t+1}))))^2 \\ &-(h(X_t,a_j) - (R_t + \gamma f(X_{t+1},\pi(X_{t+1}))))^2\Big). \qquad (7)\end{aligned}$$

and we shall let *PEval* solve for

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

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

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

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

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

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

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

$$\begin{aligned}\mathbb{E}\left[C_{tj}\hat{Q}_{f,t}\middle| X_t\right] &= \mathbb{E}\left[\hat{Q}_{f,t}\middle| X_t, A_t = a_j\right] \qquad (11)\\ &= r_j(X_t) + \gamma\int_y f(y,\pi(y))\,dP(y|X_t,a_j) = (T^\pi f)_j(X_t)\end{aligned}$$

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,

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

By Steiner's rule (which is, essentially, the bias-variance decomposition) if U,V are conditionally independent given W,

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

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

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

Taking expectations of both sides we get that

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

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

It can be observed that the unbiasedness is achieved because the quadratic terms $\hat{Q}_{f,t}^2$ and $(T^\pi f)_j^2$ are cancelled 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, 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. (We introduce LSTD quite differently from how it is done in the literature, though our treatment is close and is strongly influenced by the description provided in Lagoudakis and Parr (2003).)

Instead aiming at minimizing the distance of Q and $T^\pi Q$, one may alternatively look for a value function Q in the space of admissible functions $\mathcal{F}^L$ such that the backprojection of the image of Q under T^π onto $\mathcal{F}^L$ comes

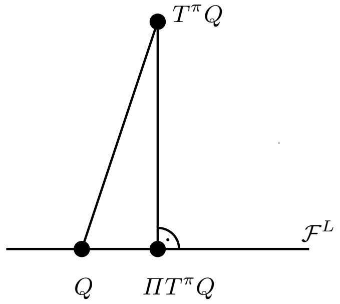

Fig. 2 Comparing the modified Bellman-error and the LSTD criterion. The function space, $\mathcal{F}^L$, is represented by the horizontal line. Under the operator, T^π, a value function, $Q \in \mathcal{F}^L$, is mapped to a function, $T^\pi Q$. The vector connecting $T^\pi Q$ and its back-projection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

the closest to Q. Formally, we want to minimize $\|Q - \Pi T^\pi Q\|$ and this is the criterion used by LSTD (see Figure 2). Here the projection operator $\Pi : B(\mathcal{X} \times \mathcal{A}) \to B(\mathcal{X} \times \mathcal{A})$ is defined by $\Pi Q = \mathrm{argmin}_{Q' \in \mathcal{F}^L} \|Q - Q'\|$. In order to make the minimization problem practical it is customary to assume a linear parameterization of the value functions: $\mathcal{F}^L = \left\{ w^T \phi \,:\, w \in \mathbb{R}^p \right\}$, where $\phi : \mathcal{X} \times \mathcal{A} \to \mathbb{R}^p$ is some function extracting features of state-action pairs. Note that $\mathcal{F}^L$ 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 $\mathcal{F}^L$ with respect to the inner product underlying the chosen norm.[2] Formally this means that $\langle\, Q_{w^\pi} - T^\pi Q_{w^\pi}\,,\, w^T \phi \,\rangle = 0$ must hold for any weight-vector w. However, this can hold only if for any $j \in \{1, \ldots, L\}$,

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

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

[2] 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 $Q \in \mathcal{F}^L$, $Q - T^\pi Q$ happens to be perpendicular to $\mathcal{F}^L$ then Q and $\Pi T^\pi Q$ must coincide.

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

$$0 = \frac{1}{NL} \sum_{t=1,j=1}^{N,L} C_{tj}\, \phi_j(X_t, A_t) \left(Q_{w^\pi}(X_t, A_t) - \Big[R_t + \gamma Q_{w^\pi}(X_{t+1}, \pi(X_{t+1})) \Big] \right), \tag{14}$$

where $C_{tj} = \frac{\mathbb{I}_{\{A_t = a_j\}}}{\pi_b(a_j|X_t)}$ is used to stay in par with our previous convention to normalize with respect to the non-uniform 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 in a reasoning that is entirely analogous to the argument used to derive (13). For 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

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

where $q, q' \in \mathbb{R}^N$. Further, let $\|\cdot\|_N$ denote the ℓ^2-norm on $\mathbb{R}^N$ that corresponds to $\langle \cdot\,, \cdot \rangle_N$.

Then (14) can be written in the compact form $0 = \langle\, S_N Q_{w^\pi} - \hat{T}_N^\pi Q_{w^\pi}\,, \phi_j \,\rangle_N$. Further, the solution minimizes $\left\| S_N Q - \Pi_N \hat{T}_N^\pi Q_{w^\pi} \right\|_N$, where the projection operator $\Pi_N : \mathbb{R}^N \to \mathbb{R}^N$ is defined by $\Pi_N q = \operatorname{argmin}_{q' \in S_N \mathcal{F}^L} \|q - q'\|_N$, where $S_N \mathcal{F}^L = \left\{ S_N Q \,:\, Q \in \mathcal{F}^L \right\}$ is a linear space as before. The reasoning is the same as previously.

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 becomes equivalent to LSPI.*

Proof We use the same notation as above: ϕ_j is the j^{th} component of the basis function ϕ that generates $\mathcal{F}^L$. We prove the statement for the population based losses, $L_{\text{LSTD}}(Q; \pi) = \|Q - \Pi T^\pi Q\|$, $L_{\text{BRM}}(Q; \pi) = \|Q - T^\pi Q\| - \inf h \in \mathcal{F}^L \|h - T^\pi Q\|$, where $\|\cdot\|$ is an norm derived from some inner product $\langle \cdot\,, \cdot \rangle$. The argument for the empirical losses is an exact parallel of this argument, just one must use S_N, Π_N and $\langle \cdot\,, \cdot \rangle_N$ as defined above.

Let $Q \in \mathcal{F}^L$ solve the equations $\langle\, Q - T^\pi Q\,, \phi_j \,\rangle = 0$ simultaneously for all j. For this Q, both $L_{\text{LSTD}}(Q; \pi) = \|Q - \Pi T^\pi Q\|$ and $L_{\text{BRM}}(Q; \pi) = \|Q - T^\pi Q\| - \|\Pi T^\pi Q - T^\pi Q\| = \|Q - T^\pi Q\| - \inf_{h \in \mathcal{F}^L} \|h - T^\pi Q\|$ are zero. Since both are nonnegative, Q minimizes both of them. In order to finish the proof we still need to show that both losses have a unique minima. This is evident for the LSTD loss function. To see that the statement holds for the BRM loss function let us remark that Q is a minimizer for it if and only if $\|Q - T^\pi Q\| = \|\Pi T^\pi Q - T^\pi Q\|$. Since projection minimizes the distance to $\mathcal{F}^L$ and $Q \in \mathcal{F}^L$, we must then have $Q - \Pi T^\pi Q = 0$. But this means that Q is the unique minimizer of the LSTD loss, finishing the proof. □

As a consequence of this equivalence 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 numerical minimization difficult 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 a parameterizations the optimization problem can be difficult to solve.

It is interesting to note that the proposed modification to the BRM loss is not the only one that allows one to achieve the unbiasedness property. In fact, one can replace R_t by any function of X_t,A_t (such as, e.g., zero!). However, this way the equivalence property would be lost. For the sake of compactness we do not pursue this direction any further here.

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 $Z^{1:n}$ the collection $(Z_1, \ldots, Z_n)$, where we allow $n = \infty$. Let $\sigma(Z^{i:j})$ denote the sigma-algebra generated by $Z^{i:j}$ ($i \le j$). The m-th β-mixing coefficient of $\{Z_t\}$, β_m, is defined by*

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

$\{Z_t\}$ 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 besides β-mixing, many other definitions of mixing exist in the literature (see, e.g., Doukhan, 1994). The weakest among those most commonly used is called α-mixing. Another commonly used one 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*

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

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

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 average concentrability of the future-state distribution. Remember that ν denotes the stationary distribution underlying $\{X_t\}$. We shall also need a distribution, chosen by the user, that is used when assessing the performance. We shall denote this distribution by ρ. 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. Now, an easy 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:[3] 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 $\{\pi_m\}_{m\ge1}$, let*

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

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

$$C_{\rho,\nu} = (1-\gamma)^2 \sum_{m\ge1} m\gamma^{m-1} c_{\rho,\nu}(m).$$

In general $c_{\rho,\nu}(m)$ diverges to infinity as $m \to \infty$. However, thanks to the discounting, $C_{\rho,\nu}$ will still be finite whenever 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 $\Gamma = \limsup_{m\to\infty} 1/m \log c_{\rho,\nu}(m) \le 0$

[3] 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 finite.

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 $\Gamma \leq 0$ is interpreted as a stability condition. Hence, our condition on the finiteness of the discounted-average concentrability coefficient $C_{\rho,\nu}$ can also be interpreted as a stability condition. Further discussion of this concept and some examples of how to estimate $C_{\rho,\nu}$ for various system classes can be found in the report by Munos and Szepesvári (2006).

The concentrability coefficient $C_{\rho,\nu}$ will enter our bound on the weighted error of the algorithm. In addition to these weighted-error bounds, we shall also derive a bound on the L^∞-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 $C_{\rho,\nu} \leq C_\nu$ 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.[4] The VC-dimension of a set system $\mathcal{C}$ shall be denoted by $V_{\mathcal{C}}$. To avoid any confusions we introduce the definition of covering numbers:

Definition 3 (Covering Numbers) *Fix $\varepsilon > 0$ and a semi-metric space $\mathcal{M} = (\mathcal{M}, d)$. We say that $\mathcal{M}$ is covered by m discs $D_1, \ldots, D_m$ if $\mathcal{M} \subset \cup_j D_j$. We define the* covering number *$\mathcal{N}(\varepsilon, \mathcal{M}, d)$ of $\mathcal{M}$ as the smallest integer m such that $\mathcal{M}$ can be covered by m discs each of which having a radius less than 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 $x^{1:N} = (x_1, x_2, \ldots, x_N)$ in $\mathcal{X}$, we use the *empirical covering numbers*, i.e., the covering number of $\mathcal{F}$ equipped with the empirical L^1 semi-metric

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

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

Another 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 $\mathcal{F}$).*

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

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

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

The rationale of this definition is as follows: Remember that in the k^{th} iteration of the algorithm we want to compute an approximate (action-value) evaluation of the policy greedy w.r.t. a previously computed action-value function Q'. Thus, if $\hat{\pi}$ denotes the chosen greedy policy, then we will jointly select L 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

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

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 $\mathcal{F}^{\vee}_{\hat{\pi}}$:

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

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

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

(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

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

then we see that

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

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 $\Xi_{\mathcal{F},L}$ and set $\mathcal{F}$. Actually, as the starting point for our capacity bounds we will use a result from the regression tree literature due to Nobel (1996).

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

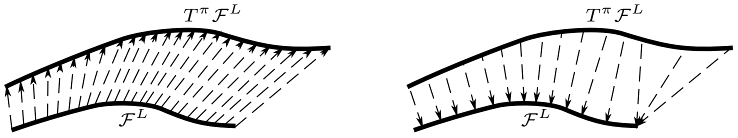

Fig. 3 Illustration of the concepts used to measure the approximation power of the function space $\mathcal{F}^L$. On the left side the vectors represent the mapping T^π. On this figure, the measure $E_\infty(\mathcal{F}^L;\pi)$ is the length of the shortest vector. On the right side the vectors represent the shortest distances of selected points of $T^\pi\mathcal{F}^L$ to $\mathcal{F}^L$. The measure $E_1(\mathcal{F}^L;\pi)$ is the length of the shortest 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) $V_{\mathcal{F}^+} \le V_{\mathcal{F}^\times}$. In particular, if $V_{\mathcal{F}^\times} < \infty$ then $V_{\mathcal{F}^+} < \infty$.
b) If $\mathcal{F}$ is a vector space then $V_{\mathcal{F}^+} = V_{\mathcal{F}^\times} = \dim(\mathcal{F})$. In particular, if $\mathcal{F}$ is a subset of a finite dimensional vector space then $V_{\mathcal{F}^\times} < \infty$.
c) There exists $\mathcal{F}$ with $V_{\mathcal{F}^\times} < \infty$ which is not a subset of any finite dimensional vector space.
d) There exists $\mathcal{F}$ with $\mathcal{X} = [0,1]$, $V_{\mathcal{F}^+} < \infty$ but $V_{\mathcal{F}^\times} = \infty$. In particular, there exists $\mathcal{F}$ with these properties such that the following properties also hold for $\mathcal{F}$: (i) *$\mathcal{F}$ is countable,* (ii) *$\{\, \{\, x \in \mathcal{X} : f(x) \ge a \,\} : f \in \mathcal{F}, a \in \mathbb{R} \,\}$ is a* VC*-class (i.e., $\mathcal{F}$ is* VC-major *class),* (iii) *each $f \in \mathcal{F}$ is monotonous, bounded, and continuously differentiable with uniformly bounded derivatives.*

The proof of this proposition is given in the Appendix. Our assumptions on the function set $\mathcal{F}$ are as follows:

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

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

The Bellman-error of an action-value function Q w.r.t. a policy evaluation operator T^π is commonly defined as the supremum norm of the difference $Q - T^\pi Q$ in analogy with the definition where the operators act on state-value functions. If the Bellman-error is small then Q is close to the fixed point of T^π thanks to T^π being a contraction. 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 admissible action-value function $Q \in \mathcal{F}^L$ such that $Q - T^\pi Q$ is small. For a fixed policy π, the quantity

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

can be used to measure the power of $\mathcal{F}$ in this respect (see Figure 4). Since we do not know in advance the policies seen during the execution of the algorithm, taking a pessimistic approach, we characterize the approximation power of $\mathcal{F}$ in terms of

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

called the *inherent Bellman-error of* $\mathcal{F}$. The subindex '∞' 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.

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

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

The right-hand subfigure of Figure 4 illustrates this concept. Taking again a pessimistic approach, the *inherent one-step Bellman-error* of $\mathcal{F}$ is defined as

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

The rationale of the 'one-step' qualifier is that T^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 $\mathcal{F}^L$. It is the additional term, $\|h - T^\pi f\|_\nu$ that we subtracted in (5) to 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 $\mathcal{F}$:

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

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

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 $Q_k \in \mathcal{F}^L$ be the k^{th} iterate ($k = -1, 0, 1, 2, \ldots$) and let π_{k+1} be greedy w.r.t. Q_k. Choose $\rho \in M(\mathcal{X})$, a measure used to evaluate the performance of the algorithm and let $0 < \delta \le 1$. Then*

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

$$\frac{2\gamma}{(1-\gamma)^2}\left(C^{1/2}_{\rho,\nu}\left(E(\mathcal{F}^L) + \left(\frac{\Lambda_N(\frac{\delta}{K})\,(\Lambda_N(\frac{\delta}{K})/b \vee 1)^{1/\kappa}}{C_2 N}\right)^{1/4}\right) + \gamma^{K/2}\, R_{\max}\right) \tag{18}$$

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

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

$L_2 = L(L-1)$,

$$\log C_1 = V \,\log\left(\frac{512eQ_{\max}\tilde{R}_{\max}}{L\pi_{b0}}\right) + V_{\mathcal{F}^\times}\, L_2 \log L_2 + V_{\mathcal{F}^+}\, L\log 2 + L^2$$
$$+L_2\log(V_{\mathcal{F}^\times}+1) + L\log(V_{\mathcal{F}^+}+1) + 2\log(LV_{\mathcal{F}^+}+1) + 2\log(4e),$$

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

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* $C_{\rho,\nu}^{1/2}$ *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 should be familiar from regression function estimation: In particular, the first term that depends on the total inherent Bellman-error of $\mathcal{F}$, $E(\mathcal{F}^L)$, quantifies the approximation power of $\mathcal{F}$ as discussed beforehand. The next term, apart from logarithmic and constant factors and terms and after some simplifications can be written in the form

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

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

Finally, let us comment on the multipliers of the bound. The multiplier $2\gamma/(1-\gamma)^2$ appears in previous L^∞-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.

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

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

For the sake of compactness this bound is not explored here in further details.

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

4.1 Bounds on the Error of the Fitting Procedure

The goal of this section is to derive a bound on the error introduced due to using a finite sample in the main optimization routine minimizing the (modified) sample-based Bellman-residual criterion defined by (7). If the samples were identically distributed and independent of each other, we could use the results developed for empirical processes (e.g., Pollard's inequality) to arrive at such a bound. However, since the samples are dependent these tools cannot be used. Instead, we will use the blocking device of Yu (1994). For simplicity assume that $N = 2m_N k_N$ for appropriate positive integers m_N, k_N (the general case can be taken care of as was done by Yu, 1994). The technique of Yu partitions the samples into $2m_N$ blocks, each having k_N samples. The samples in every second block are replaced by "ghost" samples whose joint marginal distribution is kept the same as that of the original samples (for the same block). However, these new random variables are constructed such that the new blocks are independent of each other. In order to keep the flow of the developments continuous, the proofs of the statements of these results are given in the Appendix.

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

Lemma 5 *Suppose that $Z_1, \ldots, Z_N \in \mathcal{Z}$ is a stationary β-mixing process with mixing coefficients $\{\beta_m\}$, $Z'_t \in \mathcal{Z}$ ($t \in H$) are the block-independent "ghost" samples as done by Yu (1994), and $H = \{\, 2ik_N + j \,:\, 0 \le i < m_N, 1 \le j \le k_N \,\}$,* *and that $\mathcal{F}$ is a permissible class of $\mathcal{Z} \to [-K, K]$ functions. Then*

$$\mathbb{P}\left(\sup_{f\in\mathcal{F}}\left|\frac{1}{N}\sum_{t=1}^{N} f(Z_t) - \mathbb{E}\left[f(Z_1)\right]\right| > \varepsilon\right)$$
$$\leq 16\mathbb{E}\left[\mathcal{N}_1(\varepsilon/8, \mathcal{F}, (Z'_t; t\in H))\right] e^{-\frac{m_N\varepsilon^2}{128K^2}} + 2m_N\beta_{k_N+1}.$$

Note that this lemma is based on the following form of a lemma due to Yu (1994). This lemma is stated without a proof:[5]

Lemma 6 (Yu, 1994, 4.2 Lemma) *Suppose that $H_i = \{\, 2k_N(i-1)+j \,:\, 1\leq j\leq k_N \,\}$, $\{Z_t\}$, $\{Z'_t\}$, and $H = \bigcup_{i=1}^{m_N} H_i$ are as in Lemma 5, and that $\mathcal{F}$ is a permissible class of bounded $\mathcal{Z}\to\mathbb{R}$ functions. Then*

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

Let us now develop the tools used to bound the capacity of the function set of interest. For this, 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 $\max\left\{\, \Delta(x^{1:N}, \Xi) \,:\, x^{1:N}\in\mathcal{X}^N \,\right\}$. Note that the partitioning number is a generalization of shatter-coefficient.

Given a class $\mathcal{G}$ of real-valued functions on $\mathcal{X}$ and a partition family 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 $\mathcal{F}^{\vee} = \mathcal{F}\circ\Xi_{\mathcal{F},L}$. We quote here a result of Nobel (with any domain $\mathcal{X}$ instead of $\mathbb{R}^s$ and with minimized premise):

[5] 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. Actually, a careful investigation of the original proof of Yu (1994) leads to the bound that is presented here.

Proposition 7 (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:*

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

Then, for any $\varepsilon > 0$,

$$\mathcal{N}_1(\varepsilon, \mathcal{G} \circ \Xi, x^{1:N}) \le \Delta(x^{1:N}, \Xi)\phi_N(\varepsilon)^{m(\Xi)} \le \Delta_N^*(\Xi)\phi_N(\varepsilon)^{m(\Xi)}. \tag{19}$$

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

Lemma 8 *Let Ξ, $\mathcal{G}$, $x^{1:N}$, $\phi_N : \mathbb{R}^+ \to \mathbb{R}^+$ be as in Proposition 7. Moreover, let $\mathcal{G}$ be bounded: $\forall g \in \mathcal{G}$, $|g| \le K$. For $\xi = \{A_j\}$, $\xi' = \{A'_j\} \in \Xi$, introduce the semi-metric*

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

where

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

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 $\varepsilon > 0$, $\alpha \in (0,1)$,

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

Note that from this latter bound, provided that ϕ_N is left-continuous, the conclusion of Proposition 7 follows in the following limiting sense: Since $\mathcal{N}(\varepsilon, \Xi, d_{x^{1:N}}) \le \Delta(x^{1:N}, \Xi)$ holds for any $\varepsilon > 0$, we have

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

Thus, letting $\alpha \to 0$ yields the bound (19).

Lemma 8 is used by the following result that develops a capacity bound on the function set of interest:

Lemma 9 *Let $\mathcal{F}$ be a class of uniformly bounded functions on $\mathcal{X}$ ($\forall f \in \mathcal{F}$, $|f| \le K$), $x^{1:N} \in \mathcal{X}^N$, $\phi_N : \mathbb{R}^+ \to \mathbb{R}^+$ 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 7. Let $\mathcal{G}_2^1$ denote the class of indicator functions $\mathbb{I}_{\{f_1(x) \ge f_2(x)\}} : \mathcal{X} \to \{0,1\}$ for any $f_1, f_2 \in \mathcal{F}$. Then for $\mathcal{F}^\vee$ defined in (17), $L_2 = L(L-1)$, for every $\varepsilon > 0$, $\alpha \in (0,1)$,*

$$\mathcal{N}_1(\varepsilon, \mathcal{F}^\vee, x^{1:N}) \le \mathcal{N}_1\left(\frac{\alpha\varepsilon}{L_2 K}, \mathcal{G}_2^1, x^{1:N}\right)^{L_2} \phi_N((1-\alpha)\varepsilon)^L.$$

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

Proposition 10 (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 $\varepsilon > 0$,*

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

Define

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

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

Lemma 11 *Let Assumption 1 and 2 hold, and fix the set of admissible functions $\mathcal{F}$ satisfying Assumption 3. Let Q' be a real-valued random function over $\mathcal{X} \times \mathcal{A}$, $Q'(\omega) \in \mathcal{F}^L$ (possibly not independent from the sample path). Let $\hat{\pi} = \hat{\pi}(\cdot; Q')$ be a policy that is greedy w.r.t. Q'. Let f' be defined by*

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

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

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

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

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

Corollary 12 *Let Assumption 1 hold. Assume that $\{(X_t, R_t)\}_{t=1,\dots,N}$ is the sample path, $\{X_t\}$ is strictly stationary ($X_t \sim \nu \in M(\mathcal{X})$) and β-mixing with exponential rate $(\overline{\beta}, b, \kappa)$. Assume that $\mathcal{F} \subset B(\mathcal{X}; Q_{\max})$ for $Q_{\max} \ge 0$ and $V_{\mathcal{F}^+} < \infty$. Let f' be defined by*

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

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

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

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

4.2 Propagation of Errors

The main result of the previous section shows that if the approximation power of $\mathcal{F}$ is good enough and the number of samples is high then for any policy 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:

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

Further, let

$$\varepsilon_k = Q_k - T^{\pi_k}Q_k, \quad k=0,1,2,\ldots \tag{20}$$

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

Lemma 13 *Let $p \geq 1$, and let K be a positive integer, $Q_{\max} \leq R_{\max}/(1-\gamma)$. Then, for any sequence of functions $\{Q_k\} \subset B(\mathcal{X}; Q_{\max})$, $0 \leq k < K$ and ε_k defined by* (20) *the following inequalities hold:*

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

$$\|Q^* - Q^{\pi_K}\|_{\infty} \leq \frac{2\gamma}{(1-\gamma)^2}\left(C_{\nu}^{1/p} \max_{0\leq k<K} \|\varepsilon_k\|_{p,\nu} + \gamma^{K/p} R_{\max}\right). \tag{22}$$

Proof We have $C_\nu \geq C_{\rho,\nu}$ for any ρ. Thus, if the bound (21) holds for any ρ, choosing ρ to be a Dirac at each state implies that (22) also holds. Therefore, we only need to prove (21).

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,

$$Q^* - Q^{\pi_K} \leq \gamma \sum_{k=0}^{K-1} (\gamma P^{\pi^*})^{K-k-1} E_k \varepsilon_k + (\gamma P^{\pi^*})^K (Q^* - Q^{\pi_0}). \tag{23}$$

Now, let

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

By taking the absolute value point-wise in (23) we get

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

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

$$|Q^* - Q^{\pi_K}| \le \frac{2\gamma(1-\gamma^{K+1})}{(1-\gamma)^2} \left[\sum_{k=0}^{K-1} \alpha_k A_k |\varepsilon_k| + \alpha_K A_K R_{\max}\mathbf{1} \right]. \qquad (24)$$

Here we introduced the positive coefficients

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

and the operators

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

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

Let $\lambda_K = \left[\frac{2\gamma(1-\gamma^{K+1})}{(1-\gamma)^2}\right]^p$. Taking the p^{th} power of both sides of (24), 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

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

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

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

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

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

leading to the desired bound:

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

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 only need to prove the statement for 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 13. Indeed, if one defines $\varepsilon_k = Q_k - T^{\pi_k} Q_k$ then by Lemma 13 with $p = 2$,

$$\|Q^* - Q^{\pi_K}\|_{\rho} \leq \frac{2\gamma}{(1-\gamma)^2} \left(C_{\rho,\nu}^{1/2} \max_{0 \leq k < K} \|\varepsilon_k\|_{\nu} + \gamma^{K/2} \, R_{\max} \right). \quad (25)$$

Now, from Lemma 11, we conclude that for any fixed integer $0 \leq k < K$ and for any $\delta' > 0$,

$$\|\varepsilon_k\|_{\nu} \leq E(\mathcal{F}^L) + \left(\frac{\Lambda_N(\delta') \, (\Lambda_N(\delta')/b \vee 1)^{1/\kappa}}{C_2 N} \right)^{1/4} \quad (26)$$

holds everywhere except on a set of probability at most δ'. ($\Lambda_N(\delta')$ and C_2 are defined as in the theorem.) Take $\delta' = \delta/K$. By the choice of δ', the total probability of the set of exceptional events for $0 \leq k < K$ is at most δ. Outside of this failure set, we have that Equation (26) holds for all $0 \le k < K$. Combining this with (25), we get

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

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

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

5 Related Work

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

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

Another interesting theoretical development concerning off-policy control learning with value-function approximation is the paper by Ormoneit and Sen (2002) who considered kernel-regression in conjunction with Q-learning and obtained asymptotic rates on weak-convergence. Q-learning with interpolative function approximation was considered by Szepesvári and Smart (2004), where only asymptotic convergence and performance bounds were given. Both these works carry out the analysis with respect to the L^∞ norm 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 methods, the closest to the present work is the paper of Szepesvári and Munos (2005). However, unlike there here we dealt with a fitted policy iteration algorithm and worked with dependent samples and a single sample-path. All these resulted in a much more complex analysis and the need to develop new tools: For dealing with dependant data, we used the blocking device originally proposed by Yu (1994). We had to introduce a new capacity concept to deal with the complications arising from the use of policy iteration. The error propagation technique used in Section 4.2 is an extension of a similar technique due to Munos (2003). However, 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), but only for analyzing approximate value iteration.

6 Conclusions and Future Work

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

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

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

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

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

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

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

Finally, it would be interesting to compare the result that we obtained with $\gamma = 0$ and $L = 1$ for the regression-case (Corollary 12) with similar results available in the regression literature. In connection to this, let us remark that our method applies and can be used to derive bounds to the solution of inverse problems of the form $Pf = r$, $f =?$ with P being a stochastic operator and when the data consists of samples from r and P.

Appendix

6.1 Proofs of the Auxiliary Lemmata

PROOF OF 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 $f_1(x_i) \le z_i$ for $x_i \notin A$, and also there is an $f_2 \in \mathcal{F}$ such that $f_2(x_i) \le z_i$ for $x_i \in A$ and $f_2(x_i) > z_i$ for $x_i \notin A$. That is, $f_1(x_i) > f_2(x_i)$ for $x_i \in A$ and $f_1(x_i) < f_2(x_i)$ for $x_i \notin A$. Thus, the set in $\mathcal{C}_2$ corresponding to (f_1, f_2) contains exactly the same x_i's as A does. This means that $x_1,\ldots,x_{V_{\mathcal{F}^+}}$ is shattered by $\mathcal{C}_2$, that is, $V_{\mathcal{F}^\times} = V_{\mathcal{C}_2} \ge V_{\mathcal{F}^+}$. The second part of the statement is obvious.

b) According to Theorem 11.4 of Anthony and Bartlett (1999), $V_{\mathcal{F}^+} = \dim(\mathcal{F})$. On the other hand, since now for $f_1,f_2 \in \mathcal{F}$ also $f_1 - f_2 \in \mathcal{F}$, it is easy to see that $\mathcal{C}_2 = \{\, \{\, x \in \mathcal{X} \,:\, f(x) \ge 0 \,\} \,:\, f \in \mathcal{F} \,\}$. By taking $g \equiv 0$ in Theorem 3.5 of Anthony and Bartlett (1999), we get the desired $V_{\mathcal{F}^\times} = V_{\mathcal{C}_2} = \dim(\mathcal{F})$. The second statement follows obviously.

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

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

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

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

where $\pi = 3.14159..$ is Ludolf's number. Certainly, f_n and $\tilde{f}_n$ are both differentiable. Note that $a_n \le a_{2^k} \le 1/k$, thus the gradient of the last term of $\tilde{f}_n(x)$ is bounded in absolute value by $k\pi/(4k) < 1$. Hence the functions $\tilde{f}_n$ (and obviously f_n) are strictly monotonously increasing, and have range in $[0,2]$. Let $\mathcal{F}_1 = \{\, f_n \,:\, n \ge 2 \,\}$, $\tilde{\mathcal{F}}_1 = \left\{ \, \tilde{f}_n \,:\, n \ge 2 \, \right\}$, and $\mathcal{F} = \mathcal{F}_1 \cup \tilde{\mathcal{F}}_1$. $\mathcal{F}$ is certainly countable. By the monotonicity of f_n and $\tilde{f}_n$, the VC-dimension of $\{\, \{\, x \in \mathcal{X} \,:\, f(x) \ge a \,\} \,:\, f \in \mathcal{F}, a \in \mathbb{R} \,\}$ is 1. Observe that the sequence f_n is point-wise monotonously increasing also in n, and this remains true also for $\tilde{f}_n$, since the last modifying term is negligible (less than $a_n/4$ in absolute value). (Moreover, for any n,n', $n > n'$, $f_n > \tilde{f}_{n'}$ and $\tilde{f}_n > f_{n'}$ everywhere.) This point-wise monotonicity implies that $V_{\mathcal{F}_1^+} = V_{\tilde{\mathcal{F}}_1^+} = 1$, and thus $V_{\mathcal{F}^+} \le 3$. On the other hand, since $\left\{ x \in \mathcal{X} \,:\, \tilde{f}_n(x) \ge f_n(x) \right\} =$

$\Big\{ x \in \mathcal{X} \,:\, (-1)^{\lfloor i/2^{\lfloor kx \rfloor} \rfloor} \geq 0 \Big\} = \big\{ x \in \mathcal{X} \,:\, \lfloor i/2^{\lfloor kx \rfloor} \rfloor \text{ is even} \big\}, \mathcal{C}_2 \supseteq \Big\{ \Big\{ x \in \mathcal{X} \,:\, \tilde{f}_n(x) \geq f_n(x) \Big\} \,:\, n \geq 2 \Big\} = \big\{ \big\{ x \in \mathcal{X} \,:\, \lfloor i/2^{\lfloor kx \rfloor} \rfloor \text{ is even} \big\} \,:\, n \geq 2 \big\}$, and this class contains the unions of $\{1\}$ and any of the intervals $\{[0,1/k), [1/k,2/k), \ldots [1-1/k,1)\}$ for any k. Thus it shatters the points $\{0, 1/k, 2/k, \ldots 1-1/k\}$, hence $V_{\mathcal{F}^\times} = V_{\mathcal{C}_2} = \infty$. □

Proof of Lemma 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 $\bar{\mathcal{F}} \stackrel{\text{def}}{=} \big\{ \bar{f} \,:\, f \in \mathcal{F} \big\}$.

We use Lemma 6 of Yu to replace the original process by the block-independent one, implying

$$\begin{aligned}
&\mathbb{P}\left(\sup_{f \in \mathcal{F}} \left| \frac{1}{N} \sum_{t=1}^{N} f(Z_t) - \mathbb{E}\left[f(Z_1)\right] \right| > \varepsilon \right) \\
&= \mathbb{P}\left(\sup_{f \in \mathcal{F}} \left| \frac{1}{N} \sum_{t=1}^{N} (f(Z_t) - \mathbb{E}\left[f(Z_1)\right]) \right| > \varepsilon \right) \\
&\leq 2\mathbb{P}\left(\sup_{f \in \mathcal{F}} \left| \frac{1}{N} \sum_{i=1}^{m_N} (\bar{f}(Z'^{(i)}) - k_N \mathbb{E}\left[f(Z_1)\right]) \right| > \frac{\varepsilon}{2} \right) + 2 m_N \beta_{k_N+1} \\
&= 2\mathbb{P}\left(\sup_{f \in \mathcal{F}} \left| \frac{1}{m_N} \sum_{i=1}^{m_N} \bar{f}(Z'^{(i)}) - k_N \mathbb{E}\left[f(Z_1)\right] \right| > k_N \varepsilon \right) + 2 m_N \beta_{k_N+1}. \qquad (27)
\end{aligned}$$

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

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

$$\begin{aligned}
&\mathbb{P}\left(\sup_{f \in \mathcal{F}} \left| \frac{1}{m_N} \sum_{i=1}^{m_N} \bar{f}(Z'^{(i)}) - k_N \mathbb{E}\left[f(Z_1)\right] \right| > k_N \varepsilon \right) \\
&\qquad \leq 8 \mathbb{E}\left[\mathcal{N}_1(k_N \varepsilon/8, \bar{\mathcal{F}}, (Z'^{(1)}, \ldots, Z'^{(m_N)})) \right] e^{-\frac{m_N \varepsilon^2}{128 K^2}}.
\end{aligned} \qquad (28)$$

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:

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

implying[6]

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

This, together with (27) and (28) gives the desired bound. □

Proof of Lemma 8. Fix $x_1, \ldots, x_N \in \mathcal{X}$ and $\varepsilon > 0$. Let $\widehat{\Xi}$ be an $\alpha\varepsilon/(2K)$-cover for Ξ according to d such that $|\widehat{\Xi}| = \mathcal{N}\left(\frac{\alpha\varepsilon}{2K}, \Xi, d\right)$. If $f \in \mathcal{G} \circ \Xi$, then there is a partition $\xi = \{A_j\} \in \Xi$ and functions $g_j \in \mathcal{G}$ such that

$$f = \sum_{A_j \in \xi} g_j \mathbb{I}_{\{A_j\}}. \tag{29}$$

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

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

Let $\mathcal{F}_j$ be an $(1-\alpha)\varepsilon$-cover for $\mathcal{G}$ on $\widehat{A}_j = \{x_1, \ldots, x_N\} \cap A'_j$ such that $|\mathcal{F}_j| \le \phi_N((1-\alpha)\varepsilon)$. To each function g_j appearing in (29) 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

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

[6] 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.

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

Proof of Lemma 9. Since $\mathcal{F}^{\vee} = \mathcal{F} \circ \Xi$ for $\Xi = \Xi_{\mathcal{F},L}$ defined in (16),

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

We apply Lemma 8 to bound this by

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

where $\mathcal{N}(\varepsilon, \Xi, d_{x^{1:N}})$ is the ε-covering number of Ξ regarding the metric $d_{x^{1:N}}$ defined in Lemma 8.

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

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

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

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

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

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

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

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

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

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

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

Lemma 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*

$$\varepsilon = \sqrt{\frac{\Lambda(\Lambda/b \vee 1)^{1/\kappa}}{C_2 N}} \tag{30}$$

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

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

and so

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

Obviously, $\Lambda \ge 1$ and from (30),

$$\varepsilon \ge \sqrt{\Lambda/(C_2N)} \ge \sqrt{1/(C_2N)} \qquad \text{and} \qquad C_2N\varepsilon^2 = \Lambda(\Lambda/b \vee 1)^{1/\kappa}. \tag{31}$$

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

$$\begin{aligned}
&C_1\left(\frac{1}{\varepsilon}\right)^V e^{-4C_2 m_N \varepsilon^2} + 2m_N\beta_{k_N} \\
&\le C_1\left(\frac{1}{\varepsilon}\right)^V e^{-(\frac{b}{C_2N\varepsilon^2}\wedge 1)^{\frac{1}{1+\kappa}} C_2N\varepsilon^2} + N\overline{\beta}e^{-b(\frac{C_2N\varepsilon^2}{b}\vee 1)^{\frac{\kappa}{1+\kappa}}} \\
&= C_1\left(\frac{1}{\varepsilon}\right)^V e^{-(\frac{b}{C_2N\varepsilon^2}\wedge 1)^{\frac{1}{1+\kappa}} C_2N\varepsilon^2} + N\overline{\beta}e^{-b(\frac{C_2N\varepsilon^2}{b}\vee 1)(\frac{b}{C_2N\varepsilon^2}\wedge 1)^{\frac{1}{1+\kappa}}} \\
&\le \left(C_1\left(\frac{1}{\varepsilon}\right)^V + N\overline{\beta}\right) e^{-(\frac{b}{C_2N\varepsilon^2}\wedge 1)^{\frac{1}{1+\kappa}} C_2N\varepsilon^2},
\end{aligned}$$

which, by (31), is upper bounded by

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

It is easy to check that the exponent of e in the last factor is just $-\Lambda$. Thus, substituting Λ, this factor is $N^{-V/2}\delta/(e(C_1C_2^{V/2} \vee \overline{\beta} \vee 1))$, and our bound becomes

$$\left(C_1(C_2N)^{V/2} + N\overline{\beta}\right) N^{-V/2}\frac{\delta}{e(C_1C_2^{V/2} \vee \overline{\beta} \vee 1)} \le (1+1)\frac{\delta}{e} < \delta. \square$$

6.2 Proof of Lemma 11

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

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

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

We can assume that $|\mathcal{F}| \geq 2$ (otherwise the bound is obvious). This implies $V_{\mathcal{F}^+}$, $V_{\mathcal{F}^\times} \geq 1$, and thus $V \geq L(L+2) \geq 3$. Let ε and $\Lambda_N(\delta)$ be chosen as in (30):

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

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

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

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

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

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

Let us now introduce the following additional shorthand notations:

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

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

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

Now,

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

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

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

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

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

Thus we get

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

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

We follow the line of proof due to Meir (2000). For any f,h,$Q' \in \mathcal{F}^L$, define the loss function $l_{f,h,Q'} : \mathcal{X} \times \mathcal{A} \times [-\hat{R}_{\max}, \hat{R}_{\max}] \times \mathcal{X} \to \mathbb{R}$ in accordance with (7) as

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

for $z = (x,a,r,y)$ and $\mathcal{L}_{\mathcal{F}} \overset{\text{def}}{=} \left\{ l_{f,h,Q'} : f,h,Q' \in \mathcal{F}^L \right\}$. 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)

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

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

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

and (by (12))

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

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

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

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

$$\underbrace{Z_1,\ldots,Z_{k_N}}_{H_1}, \underbrace{Z_{k_N+1},\ldots,Z_{2k_N}}_{T_1}, \underbrace{Z_{2k_N+1},\ldots,Z_{3k_N}}_{H_2}, \underbrace{Z_{3k_N+1},\ldots,Z_{4k_N}}_{T_2}, \ldots$$

$$\ldots, \underbrace{Z_{(2m_N-2)k_N+1},\ldots,Z_{(2m_N-1)k_N}}_{H_{m_N}}, \underbrace{Z_{(2m_N-1)k_N+1},\ldots,Z_{2m_N k_N}}_{T_{m_N}}.$$

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

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

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

For this ansatz we use Lemma 5 above with $\mathcal{Z} = \mathcal{X}\times\mathcal{A}\times\mathbb{R}\times\mathcal{X}$, $\mathcal{F} = \mathcal{L}_{\mathcal{F}}$ noting that any $l_{f,h,Q'} \in \mathcal{L}_{\mathcal{F}}$ is bounded by

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

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

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

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

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

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

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

Applying now Lemma 9 with $\alpha = 1/2$,[7] the covering number of $\mathcal{F}^\vee$ is bounded by

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

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

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

To bound these factors, we use Corollary 3 from Haussler (1995) that was cited here as Proposition 10. The pseudo-dimensions of $\mathcal{F}$ and $\mathcal{G}_2^1$ are $V_{\mathcal{F}^+}$, $V_{\mathcal{F}^\times} < \infty$, respectively, and the range of functions from $\mathcal{F}$ has length $2Q_{\max}$. By the pigeonhole principle, it is easy to see that the pseudo-dimension of $\mathcal{F}^L$ cannot exceed $LV_{\mathcal{F}^+}$. Thus

$$\begin{aligned}\mathcal{N}_1\left(\frac{8\tilde{R}_{\max}}{L\pi_{b0}}\varepsilon', \mathcal{L}_{\mathcal{F}}, (Z'_t; t \in H)\right) &\le \left(e(LV_{\mathcal{F}^+}+1)\left(\frac{4eQ_{\max}}{\varepsilon'}\right)^{LV_{\mathcal{F}^+}}\right)^2 \\ &\cdot \left(e(V_{\mathcal{F}^\times}+1)\left(\frac{4eL_2Q_{\max}}{\varepsilon'}\right)^{V_{\mathcal{F}^\times}}\right)^{L_2}\left(e(V_{\mathcal{F}^+}+1)\left(\frac{8eQ_{\max}}{\varepsilon'}\right)^{V_{\mathcal{F}^+}}\right)^{L} \\ &= e^{L^2+2}(LV_{\mathcal{F}^+}+1)^2(V_{\mathcal{F}^+}+1)^L(V_{\mathcal{F}^\times}+1)^{L_2}2^{LV_{\mathcal{F}^+}}L_2^{L_2V_{\mathcal{F}^\times}}\left(\frac{4eQ_{\max}}{\varepsilon'}\right)^V,\end{aligned}$$

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

$$\begin{aligned}\mathcal{N}_1(\varepsilon/16, \mathcal{L}_{\mathcal{F}}, (Z'_t; t \in H)) &\le e^{L^2+2}(LV_{\mathcal{F}^+}+1)^2(V_{\mathcal{F}^+}+1)^L(V_{\mathcal{F}^\times}+1)^{L_2} \cdot \\ &\cdot 2^{LV_{\mathcal{F}^+}}L_2^{L_2V_{\mathcal{F}^\times}}\left(\frac{512eQ_{\max}\tilde{R}_{\max}}{L\pi_{b0}\varepsilon}\right)^V = \frac{C_1}{16}\left(\frac{1}{\varepsilon}\right)^V,\end{aligned}$$

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

$$P_0 \le C_1\left(\frac{1}{\varepsilon}\right)^V e^{-4C_2m_N\varepsilon^2} + 2m_N\beta_{k_N}, \tag{32}$$

where $C_2 = \frac{1}{2}\left(\frac{L\pi_{b0}}{32\tilde{R}^2_{\max}}\right)^2$. Defining $k_N = \lceil(C_2N\varepsilon^2/b)^{\frac{1}{1+\kappa}}\rceil$ and $m_N = N/(2k_N)$, the proof is finished by Lemma 14, which, together with (32), implies $P_0 < \delta$.

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

7 Acknowledgements

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

References

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

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

Bellman, R. and S. Dreyfus: 1959, 'Functional approximation and dynamic programming'. *Math. Tables and other Aids Comp.* **13**, 247–251.

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

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

Bradtke, S. and A. Barto: 1996, 'Linear least-squares algorithms for temporal difference learning'. *Machine Learning* **22**, 33–57.

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

Cheney, E.: 1966, *Introduction to Approximation Theory.* London, New York: McGraw-Hill.

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

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

Dietterich, T. G. and X. Wang: 2002, 'Batch value function approximation via support vectors'. In: T. G. Dietterich, S. Becker, and Z. Ghahramani (eds.): *Advances in Neural Information Processing Systems 14.* Cambridge, MA, MIT Press.

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

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

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

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

Györfi, L., M. Kohler, A. Krzyżak, and H. Walk: 2002, *A Distribution-Free Theory of Nonparametric Regression.* Springer-Verlag.

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

Howard, R. A.: 1960, *Dynamic Programming and Markov Processes.* Cambridge, MA: The MIT Press.

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

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

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

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

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

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

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

Ormoneit, D. and S. Sen: 2002, 'Kernel-based reinforcement learning'. *Machine Learning* **49**, 161–178.

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

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

Precup, D., R. Sutton, and S. Dasgupta: 2001, 'Off-policy temporal difference learning with function approximation'. In: *Proceedings of the Eighteenth International Conference on Machine Learning (ICML 2001).* pp. 417–424.

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

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

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

Sutton, R. and A. Barto: 1998, *Reinforcement Learning: An Introduction*, Bradford Book. MIT Press.

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

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

Tsitsiklis, J. N. and B. Van Roy: 1996, 'Feature-based methods for large scale dynamic programming'. *Machine Learning* **22**, 59–94.

Wang, X. and T. Dietterich: 1999, 'Efficient value function approximation using regression trees'. In: *Proceedings of the IJCAI Workshop on Statistical Machine Learning for Large-Scale Optimization*. Stockholm, Sweden.

Williams, R. J. and L. Baird, III: 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.