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

Andras Antos, Csaba Szepesvari, Rémi Munos

Learning Near-Optimal Policies with Bellman-Residual Minimization Based Fitted Policy Iteration and a Single Sample Path

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

[1] Computer and Automation Research Inst.
of the Hungarian Academy of Sciences
Kende u. 13-17, Budapest 1111, Hungary
{antos, szcsaba}@sztaki.hu
[2] Centre de Mathématiques Appliquées
Ecole Polytechnique
91128 Palaiseau Cedex, France
remi.munos@polytechnique.fr

Abstract. We consider batch reinforcement learning problems in continuous space, expected total discounted-reward Markovian Decision Problems. As opposed to previous theoretical work, we consider the case when the training data consists of a single sample path (trajectory) of some behaviour policy. In particular, we do not assume access to a generative model of the environment. The algorithm studied is policy iteration where in successive iterations the Q-functions of the intermediate policies are obtained by means of minimizing a novel Bellman-residual type error. PAC-style polynomial bounds are derived on the number of samples needed to guarantee near-optimal performance where the bound depends on the mixing rate of the trajectory, the smoothness properties of the underlying Markovian Decision Problem, the approximation power and capacity of the function set used.

1 Introduction

Consider the problem of optimizing a controller for an industrial environment. In many cases the data is collected on the field by running a fixed controller and then taken to the laboratory for optimization. The goal is to derive an optimized controller that improves upon the performance of the controller generating the data.

In this paper we are interested in the performance improvement that can be guaranteed given a finite amount of data. In particular, we are interested in how performance scales as a function of the amount of data available. We study Bellman-residual minimization based policy iteration assuming that the environment is stochastic and the state is observable and continuous valued. The algorithm considered is an iterative procedure where each iteration involves solving a least-squares problem, similar to the Least-Squares Policy Iteration algorithm of Lagoudakis and Parr [1]. However, whilst Lagoudakis and Parr considered the so-called least-squares fixed-point approximation to avoid problems with Bellman-residual minimization in the case of correlated samples, we

G. Lugosi and H.U. Simon (Eds.): COLT 2006, LNAI 4005, pp. 574–588, 2006.

modify the original Bellman-residual objective. In a forthcoming paper we study policy iteration with approximate iterative policy evaluation [2].

The main conditions of our results can be grouped into three parts: Conditions on the system, conditions on the trajectory (and the behaviour policy used to generate the trajectory) and conditions on the algorithm. The most important conditions on the system are that the state space should be compact, the action space should be finite and the dynamics should be smooth in a sense to be defined later. The major condition on the trajectory is that it should be rapidly mixing. This mixing property plays a crucial role in deriving a PAC-bound on the probability of obtaining suboptimal solutions in the proposed Bellman-residual minimization subroutine. The major conditions on the algorithm are that an appropriate number of iterations should be used and the function space used should have a finite capacity and be sufficiently rich at the same time. It follows that these conditions, as usual, require a good balance between the power of the approximation architecture (we want large power to get good approximation of the action-value functions of the policies encountered during the algorithm) and the number of samples: If the power of the approximation architecture is increased the algorithm will suffer from overfitting, as it also happens in supervised learning. Although the presence of the tradeoff between generalization error and model complexity should be of no surprise, this tradeoff is somewhat underrepresented in the reinforcement literature, presumably because most results where function approximators are involved are asymptotic.

The organization of the paper is as follows: In the next section (Section 2) we introduce the necessary symbols and notation. The algorithm is given in Section 3. The main results are presented in Section 4. This section, just like the proof, is broken into three parts: In Section 4.1 we prove our basic PAC-style lemma that relates the complexity of the function space, the mixing rate of the trajectory and the number of samples. In Section 4.2 we prove a bound on the propagation of errors during the course of the procedure. Here the smoothness properties of the MDP are used to bound the 'final' approximation error as a function of the individual errors. The proof of the main result is finished Section 4.3. In Section 5 some related work is discussed. Our conclusions are drawn in Section 6.

2 Notation

For a measurable space with domain S we let $M(S)$ denote the set of all probability measures over S. For $\nu \in M(S)$ and $f : S \to \mathbb{R}$ measurable we let $\|f\|_\nu$ denote the $L^2(\nu)$-norm of f: $\|f\|_\nu^2 = \int f^2(s)\nu(ds)$. We denote the space of bounded measurable functions with domain $\mathcal{X}$ by $B(\mathcal{X})$, the space of measurable functions bounded by $0 < K < \infty$ by $B(\mathcal{X}; K)$. We let $\|f\|_\infty$ denote the supremum norm: $\|f\|_\infty = \sup_{x\in\mathcal{X}} |f(x)|$. $\mathbb{I}_E$ denotes the indicator function of event E, whilst $\mathbf{1}$ denotes the function that takes on the constant value 1 everywhere over the domain of interest.

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

We make the following assumptions on the MDP:

Assumption 1 (MDP Regularity). *$\mathcal{X}$ is a compact subspace of the s-dimensional Euclidean space. We assume that the random immediate rewards are bounded by $\hat{R}_{\max}$, the conditional expectations $r(x,a) = \int rS(dr|x,a)$ and conditional variances $v(x,a) = \int (r - r(x,a))^2 S(dr|x,a)$ of the immediate rewards are both uniformly bounded as functions of $(x,a) \in \mathcal{X} \times \mathcal{A}$. We let $R_{\max}$ denote the bound on the expected immediate rewards: $\|r\|_\infty \le R_{\max}$.*

A policy is defined as a mapping from past observations to a distribution over the set of actions. A policy is deterministic if the probability distribution concentrates on a single action for all histories. A policy is called stationary if the distribution depends only on the last state of the observation sequence.

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: $V^\pi(x) = \mathbb{E}_\pi \left[\sum_{t=0}^\infty \gamma^t R_t | X_0 = x\right]$. Here R_t is the reward received at time step t, $R_t \sim S(\cdot|X_t, A_t)$, 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. We introduce $Q^\pi : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$, the action-value function, or simply the Q-function of policy π: $Q^\pi(x,a) = \mathbb{E}_\pi \left[\sum_{t=0}^\infty \gamma^t R_t | X_0 = x, A_0 = a\right]$.

The goal is to find a policy that attains the best possible values, $V^*(x) = \sup_\pi V^\pi(x)$ for all states $x \in \mathcal{X}$. V^* is called the optimal value function. A policy is called optimal if it attains the optimal values $V^*(x)$ for *any* state $x \in \mathcal{X}$, i.e., if $V_\pi(x) = V^*(x)$ for all $x \in \mathcal{X}$. The function $Q^*(x,a)$ is defined analogously: $Q^*(x,a) = \sup_\pi Q^\pi(x,a)$. It is known that for any policy π, V^π, Q^π are bounded by $R_{\max}/(1-\gamma)$, just like Q^* and V^*. We say that a (deterministic stationary) policy π is *greedy* w.r.t. an action-value function $Q \in B(\mathcal{X} \times \mathcal{A})$ if, for all $x \in \mathcal{X}$,$a \in \mathcal{A}$, $\pi(x) \in \operatorname{argmax}_{a \in \mathcal{A}} Q(x,a)$. Since $\mathcal{A}$ is finite, such a greedy policy always exist. It is known that under mild conditions the greedy policy w.r.t. Q^* is optimal [3]. For a deterministic stationary policy π define the operator $T^\pi : B(\mathcal{X} \times \mathcal{A}) \to B(\mathcal{X} \times \mathcal{A})$ by $(T^\pi Q)(x,a) = r(x,a) + \gamma \int Q(y, \pi(y)) P(dy|x,a)$.

For any deterministic stationary policy $\pi : \mathcal{X} \to \mathcal{A}$ let the operator $E^\pi : B(\mathcal{X} \times \mathcal{A}) \to B(\mathcal{X})$ be defined by $(E^\pi Q)(x) = Q(x, \pi(x))$; $Q \in B(\mathcal{X} \times \mathcal{A})$. We define two operators corresponding to the transition probability kernel P as follows: A right-linear operator is defined by $P\cdot : B(\mathcal{X}) \to B(\mathcal{X} \times \mathcal{A})$ and $(PV)(x,a) = \int V(y) P(dy|x,a)$, whilst a left-linear operator is defined by $\cdot P : M(\mathcal{X} \times \mathcal{A}) \to M(\mathcal{X})$ with $(\rho P)(dy) = \int P(dy|x,a)\rho(dx,da)$. This operator is also extended to act on measures over $\mathcal{X}$ via $(\rho P)(dy) = \frac{1}{L} \sum_{a \in \mathcal{A}} \int P(dy|x,a)\rho(dx)$.

FittedPolicyQ(D,K,Q_0,PEval)
// D: samples (e.g. trajectory)
// K: number of iterations
// Q_0: Initial Q-function
// PEval: Policy evaluation routine
$Q \leftarrow Q_0$ // Initialization
for $k = 0$ to $K - 1$ **do**
  $Q' \leftarrow Q$
  $Q \leftarrow$PEval$(\hat{\pi}(\cdot; Q'), D)$
end for
return Q // or $\hat{\pi}(\cdot; Q)$, the greedy policy w.r.t. Q

Fig. 1. Model-free Policy Iteration

By composing P and E^π, we define $P^\pi = PE^\pi$. Note that this equation defines two operators: a right- and a left-linear one.

Throughout the paper $\mathcal{F} \subset \{f : \mathcal{X} \to \mathbb{R}\}$ will denote some subset of real-valued functions. 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 $f \equiv (f_1, \ldots, f_L)$, $f(x, a_j) = f_j(x)$, $j = 1, \ldots, L$. For $\nu \in M(\mathcal{X})$, we extend $\|\cdot\|_\nu$ to $\mathcal{F}^L$ by $\|f\|_\nu = \left(\frac{1}{L}\sum_{j=1}^L \|f_j\|_\nu^2\right)^{1/2}$.

3 Algorithm

Assume that we are given a finite but long trajectory $\{(X_t, A_t, R_t)\}_{1 \le t \le N}$ generated by 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 shall assume that π is 'persistently exciting' in the sense that $\{(X_t, A_t, R_t)\}$ mixes fast (this will be made precise in the next section).

The algorithm studied in this paper is shown in Figure 1. It is an instance of policy iteration, where policies are only implicitly represented via action-value functions. In the figure D denotes the sample $\{(X_t, A_t, R_t)\}_{1 \le t \le N}$, K is the number of iterations, Q_0 is the initial action-value function. *PEval* is a procedure that takes data in the form of a long trajectory and a policy $\hat{\pi} = \hat{\pi}(\cdot; Q')$, the greedy policy with respect to Q'. Based on $\hat{\pi}$, *PEval* should return an estimate of the action-value function $Q^{\hat{\pi}}$. There are many possibilities to approximate $Q^{\hat{\pi}}$. In this paper we consider Bellman-residual minimization (BRM). The basic idea of BRM is that $Q^{\hat{\pi}}$ is the fixed point of the operator $T^{\hat{\pi}}$: $Q^{\hat{\pi}} - T^{\hat{\pi}} Q^{\hat{\pi}} = 0$. Hence, given some function class $\mathcal{F}^L$, functions $Q \in \mathcal{F}^L$ with small Bellman-residual $L(Q; \hat{\pi}) = \left\|Q - T^{\hat{\pi}} Q\right\|^2$ (with some norm $\|\cdot\|$) should be close to $Q^{\hat{\pi}}$, provided that $\mathcal{F}$ is sufficiently rich (more precisely, the hope is that the performance of the greedy policy w.r.t. the obtained function will be close to the performance of the policy greedy w.r.t. $Q^{\hat{\pi}}$). The most widely used norm is the L^2-norm, so let $L(Q; \hat{\pi}) = \left\|Q - T^{\hat{\pi}} Q\right\|_\nu^2$. We chase $Q = \operatorname{argmin}_{f \in \mathcal{F}^L} L(f; \hat{\pi})$. In the sample based version the minimization of the norm $L(f; \hat{\pi})$ is replaced by minimizing a sample based approximation to it: If we let

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

then the most straightforward way to compute an approximation to $Q^{\hat{\pi}}$ seems to use $Q = \operatorname{argmin}_{f\in\mathcal{F}^L}\hat{L}_N(f;\hat{\pi})$. At a first sight, the choice of $\hat{L}_N$ seems to be logical as for any given X_t, A_t and f, $R_t+\gamma f(X_{t+1},\hat{\pi}(X_{t+1})$ is an unbiased estimate of $(T^{\hat{\pi}}f)(X_t,A_t)$. However, as it is well known (see e.g. [4][pp. 220], [5, 1]), $\hat{L}_N$ is not a "proper" approximation to the corresponding L^2 Bellman-error: $\mathbb{E}\left[\hat{L}_N(f;\hat{\pi})\right] \neq L(f;\hat{\pi})$. In fact, an elementary calculus shows that for $Y\sim P(\cdot|x,a)$, $R\sim S(\cdot|x,a)$,

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

It follows that minimizing $\hat{L}_N(f;\hat{\pi})$ involves minimizing the term $\mathrm{Var}\left[f(Y,\hat{\pi}(Y))\right]$ in addition to minimizing the 'desired term' $L(f;\hat{\pi})$. The unwanted term acts like a penalty factor, favouring smooth solutions (if f is constant then $\mathrm{Var}\left[f(Y,\hat{\pi}(Y))\right] = 0$). Although in some cases smooth functions are preferable, in general it is better to control smoothness penalties in a direct way.

The common suggestion to overcome this problem is to use "double" (uncorrelated) samples. In our setup, however, this is not an option. Another possibility is to reuse samples that are close in space (e.g., use nearest neighbours). The difficulty with that approach is that it requires a definition of what it means for samples being close. Here, we pursue an alternative approach based on the introduction of an auxiliary function that is used to cancel the variance penalty. The idea is to select h to 'match' $(T^{\hat{\pi}}f)(x,a) = \mathbb{E}\left[R+\gamma f(Y,\hat{\pi}(Y))\right]$ and use it to cancel the unwanted term. Define $L(f,h;\hat{\pi}) = L(f;\hat{\pi}) - \left\|h - T^{\hat{\pi}}f\right\|_\nu^2$ and

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

Then, $\mathbb{E}\left[\hat{L}_N(f,h;\hat{\pi})\right] = L(f,h;\hat{\pi})$ and $L(f,T^{\hat{\pi}}f;\hat{\pi}) = L(f;\hat{\pi})$. Hence we let *PE-val* solve for $Q = \operatorname{argmin}_{f\in\mathcal{F}^L}\sup_{h\in\mathcal{F}^L}\hat{L}_N(f,h;\hat{\pi})$. Note that for linearly parameterized function classes the solution can be obtained in a closed form. In general, one may expect that the number of parameters doubles as a result of the introduction of the auxiliary function. Although this may represent a considerable additional computational burden on the algorithm, given the merits of the Bellman-residual minimization approach over the least-squares fixed point approach [5] we think that the potential gain in the performance of the final policy might well worth the extra effort. However, the verification of this claim is left for future work.

Our main result can be formulated as follows: Let $\epsilon,\delta>0$ be given and choose some target distribution ρ that will be used to measure performance. Regarding the function set $\mathcal{F}$ we need the following essential assumptions: $\mathcal{F}$ has finite pseudo-dimension (similarly to the VC-dimension, the pseudo-dimension of a function

class also describes the 'complexity' of the class) and the set of 0 level-sets of the differences of pairs of functions from $\mathcal{F}$ should be a VC-class. Further, we assume that the set $\mathcal{F}^L$ is $\varepsilon/2$-invariant under operators from $\mathcal{T} = \{T^{\hat{\pi}(\cdot;Q)} : Q \in \mathcal{F}^L\}$ with respect to the $\|\cdot\|_\nu$ norm (cf. Definition 3) and that $\mathcal{F}^L$ approximates the fixed-points of the operators of $\mathcal{T}$ well (cf. Definition 4). The MDP has to be regular (satisfying Assumption 1), the dynamics has to satisfy some smoothness properties and the sample path has to be fast mixing. Then for large enough values of N, K the value-function V^{π_K} of the policy π_K returned by fitted policy iteration with the modified BRM criterion will satisfy

$$\|V^{\pi_K} - V^*\|_\rho \leq \epsilon$$

with probability larger than $1 - \delta$. In particular, if the rate of mixing of the trajectory is exponential with parameters (b, κ), then $N, K \sim \text{poly}(L, \hat{R}_{\max}/(1-\gamma), 1/b, V, 1/\varepsilon, \log(1/\delta))$, where V is a VC-dimension like quantity characterizing the complexity of the function class $\mathcal{F}$ and the degree of the polynomial is $1 + 1/\kappa$.

The main steps of the proof are the followings:

1. *PAC-Bounds for BRM:* Starting from a (random) policy that is derived from a random action-value function, we show that BRM is "PAC-consistent", i.e., one can guarantee small Bellman-error with high confidence provided that the number of samples N is large enough.
2. *Error propagation:* If for approximate policy iteration the Bellman-error is small for K steps, then the final error will be small, too (this requires the smoothness conditions).
3. *Final steps:* The error of the whole procedure is small with high probability provided that the Bellman-error is small throughout all the steps with high probability.

4 Main Result

Before describing the main result we need some definitions.

We start with a mixing-property of stochastic processes. Informally, a process is mixing if future depends only weakly on the past, in a sense that we now make precise:

Definition 1. *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 $Z^{i:j}$ $(i \leq j)$. The m-th β-mixing coefficient of $\{Z_t\}$, β_m, is defined by*

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

A stochastic process is said to be β-mixing if $\beta_m \to 0$ as $m \to \infty$.

Note that there exist many other definitions of mixing. 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 [6]). A β-mixing process is said to mix at an *exponential* rate with parameters $b, \kappa > 0$ if $\beta_m = O(\exp(-bm^\kappa))$.

Assumption 2 (Sample Path Properties). *Assume that* $\{(X_t, A_t, R_t)\}_{t=1,\dots,N}$ *is the sample path of* π*,* X_t *is strictly stationary, and* $X_t \sim \nu \in M(\mathcal{X})$*. Further, we assume that* $\{(X_t, A_t, R_t, X_{t+1})\}$ *is* β*-mixing with exponential-rate* (b, κ)*. We further assume that the sampling policy* π *satisfies* $\pi_0 \stackrel{\text{def}}{=} \min_{a\in\mathcal{A}} \inf_{x\in\mathcal{X}} \pi(a|x) > 0$.

The β-mixing property will be used to establish tail inequalities for certain empirical processes.

Let us now define some smoothness constants $C(\nu)$ and $C(\rho, \nu)$, that depend on the MDP. Remember that ν is the stationary distribution of the samples X_t and ρ is the distribution that is used to evaluate the performance of the algorithm.

Definition 2. *We call* $C(\nu) \in \mathbb{R}^+ \cup \{+\infty\}$ *the* ***transition probabilities smoothness constant****, defined as the smallest constant such that for* $x \in \mathcal{X}$*,* $B \subset \mathcal{X}$ *measurable,* $a \in \mathcal{A}$*,*

$$P(B|x, a) \le C(\nu)\nu(B),$$

(if no such constant exists, we set $C(\nu) = \infty$*). Now, for all integer* $m \ge 1$*, we define* $c(m) \in \mathbb{R}^+ \cup \{+\infty\}$ *to be the smallest constant such that, for any* m *stationary policies* $\pi_1, \pi_2, \dots, \pi_m$*,*

$$\rho P^{\pi_1} P^{\pi_2} \dots P^{\pi_m} \le c(m)\nu, \tag{2}$$

and write $c(0) = 1$.[1] *Note that these constants depend on* ρ *and* ν*.*

We let $C(\rho, \nu)$*, the* ***second order discounted future state distribution smoothness constant****, be defined by the equation*

$$C(\rho, \nu) = (1-\gamma)^2 \sum_{m\ge 1} m\gamma^{m-1} c(m). \tag{3}$$

One of the major restriction on the MDP's dynamics will be that $C(\rho, \nu) < \infty$ is finite. In fact, one can show that if $C(\nu) < \infty$ then $C(\rho, \nu) < \infty$ holds for any distribution ρ. Hence, the condition $C(\rho, \nu) < \infty$ is less restrictive than $C(\nu) < \infty$. $C(\nu) < \infty$ is satisfied whenever the transition density kernel is absolute continuous w.r.t. ν.[2]

During the course of the proof, we will need several capacity concepts of function sets. Here we assume that the reader is familiar with the concept of VC-dimension (see, e.g. [7]), but we introduce covering numbers because slightly different definitions of it exist in the literature:

For a semi-metric space $(\mathcal{M}, d)$ and for each $\varepsilon > 0$, define the covering number $\mathcal{N}(\varepsilon, \mathcal{M}, d)$ as the smallest value of m for which there exist $g_1, g_2, \dots, g_m \in \mathcal{M}$

[1] Again, if there exists no such constants, we simply set $c(m) = \infty$. Note that in (2) $\le$ is used to compare two operators. The meaning of $\le$ in comparing operators H, G is the usual: $H \le G$ iff $Hf \le Gf$ holds for all $f \in \mathrm{Dom}(H)$. Here ν is viewed as an operator acting on $B(\mathcal{X} \times \mathcal{A})$.

[2] Further discussion of this condition can be found in the forthcoming paper [2] where these smoothness constants are related to the top-Lyapunov exponent of the system's dynamics.

such that for every $f \in \mathcal{M}$, $\min_j d(f, g_j) < \varepsilon$. If no such finite m exists then $\mathcal{N}(\varepsilon, \mathcal{M}, d) = \infty$. In particular, for a class $\mathcal{F}$ of $\mathcal{X} \to \mathbb{R}$ functions 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 $\mathcal{F}$ with respect to the empirical L_1 distance

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

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

Assumption 3 (Capacity Assumptions on the Function Set). *Assume that $\mathcal{F} \subset B(\mathcal{X}; Q_{\max})$ and that the pseudo-dimension (*VC*-subgraph dimension) $V_{\mathcal{F}^+}$ of $\mathcal{F}$ is finite.*[3] *Let $\mathcal{C}_2 = \{\{x \in \mathcal{X} : f_1(x) \geq f_2(x)\} : f_1, f_2 \in \mathcal{F}\}$. Assume also that the* VC*-dimension, $V_{\mathcal{C}_2}$, of $\mathcal{C}_2$ is finite.*

We shall also need that $\mathcal{F}^L$ is almost-invariant with respect to (certain) policy-evaluation operators:

Definition 3. *$\mathcal{F}$, a subset of a normed function-space is said to be ϵ-invariant with respect to the set of operators $\mathcal{T}$ acting on the function-space if $\inf_{g \in \mathcal{F}} \|g - Tf\| \leq \epsilon$ holds for any $T \in \mathcal{T}$ and $f \in \mathcal{F}$.*

Similarly, we need that $\mathcal{F}^L$ contains ϵ-fixed points of (certain) policy-evaluation operators:

Definition 4. *f is an ϵ-fixed point of T w.r.t. the norm $\|\cdot\|$ if $\|Tf - f\| \leq \epsilon$.*

Our main result is the following:

Theorem 1. *Choose $\rho \in M(\mathcal{X})$ and let $\epsilon, \delta > 0$ be fixed. Let Assumption 1 and 2 hold and let $Q_{\max} \geq R_{\max}/(1-\gamma)$. Fix $\mathcal{F} \subset B(\mathcal{X}; Q_{\max})$. Let $\mathcal{T}$ be the set of policy evaluation operators $\{T^{\hat{\pi}(\cdot;Q)} : Q \in \mathcal{F}^L\}$. Assume that $\mathcal{F}^L$ is $\epsilon/2$-invariant with respect to $\|\cdot\|_\nu$ and $\mathcal{T}$ and contains the $\epsilon/2$-fixed points of $\mathcal{T}$. Further, assume that $\mathcal{F}$ satisfies Assumption 3. Then there exists integers N, K that are polynomials in L, $Q_{\max}$, $1/b$, $1/\pi_0$, $V_{\mathcal{F}^+}$, $V_{\mathcal{C}_2}$, $1/\epsilon$, $\log(1/\delta)$, $1/(1-\gamma)$ and $C(\nu)$ such that $\mathbb{P}\left(\|V^* - V^{\pi_K}\|_\infty > \epsilon\right) \leq \delta$.*

Similarly, there exists integers N, K that are polynomials of the same quantities except that $C(\nu)$ is replaced by $C(\rho, \nu)$ such that $\mathbb{P}\left(\|V^* - V^{\pi_K}\|_\rho > \epsilon\right) \leq \delta$.

4.1 Bounds on the Error of the Fitting Procedure

We first introduce some auxiliary results required for the proof of the main result of this section. For simplicity assume that $N = 2m_N k_N$ for appropriate positive integers m_N, k_N. We start with the following lemmata:

[3] The VC-subgraph dimension of $\mathcal{F}$ is defined as the VC-dimension of the subgraphs of functions in $\mathcal{F}$.

Lemma 2. *Suppose that $Z_0, \dots, Z_N \in \mathcal{Z}$ 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 in [8], and $H = \{2ik_N + j : 0 \le i < m_N, 0 \le j < k_N\}$, and that $\mathcal{F}$ is a permissible class of $\mathcal{Z} \to [-K, K]$ functions. Then*

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

Note that this lemma is based on the following form of a lemma due to Yu [8]:

Lemma 3 (Yu [8] 4.2 Lemma). *Suppose that $\{Z_t\}$, $\{Z'_t\}$, and H are as in Lemma 2 and that $\mathcal{F}$ is a permissible class of bounded $\mathcal{Z} \to \mathbb{R}$ functions. Then*

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

Let Π be a family of partitions of $\mathcal{X}$. Define the *cell count* $m(\Pi) = \max_{\pi\in\Pi} |\{A \in \pi : A \neq \emptyset\}|$. For $x^{1:N} \in \mathcal{X}^N$, let $\Delta(x^{1:N}, \Pi)$ be the number of distinct partitions (regardless the order) of $x^{1:N}$ that are induced by the elements of Π. The *partitioning number* (generalization of shatter-coefficient) $\Delta^*_N(\Pi)$ equals to $\max\{\Delta(x^{1:N}, \Pi) : x^{1:N} \in \mathcal{X}^N\}$.

Given a class $\mathcal{G}$ of functions on $\mathcal{X}$ and a partition family Π, define

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

We quote here a result of Nobel (with any domain $\mathcal{X}$ instead of $\mathbb{R}^s$ and with minimised premise):

Proposition 4 (Nobel [9] Proposition 1). *Let Π be any partition family with $m(\Pi) < \infty$, $\mathcal{G}$ be a class of functions on $\mathcal{X}$, $x^{1:N} \in \mathcal{X}^N$. Let $\phi_N(\cdot)$ be such that for any $\varepsilon > 0$, the empirical ε-covering numbers of G on all subsets of the multiset $[x_1, \dots, X_N]$ are majorized by $\phi_N(\varepsilon)$. Then, for any $\varepsilon > 0$,*

$$\mathcal{N}_1(\varepsilon, \mathcal{G} \circ \Pi, x^{1:N}) \le \Delta(x^{1:N}, \Pi)\phi_N(\varepsilon)^{m(\Pi)} \le \Delta^*_N(\Pi)\phi_N(\varepsilon)^{m(\Pi)}.$$

We extend this result to a refined bound in terms of the covering number of the partition family instead of its partitioning number:

Lemma 5. *Let $\Pi, G, x^{1:N}, \phi_N$ be as in Lemma 4. For $\pi = \{A_j\}, \pi' = \{A'_j\} \in \Pi$, introduce the metric $d(\pi, \pi') = d_{x^{1:N}}(\pi, \pi') = \mu_N(\pi \triangle \pi')$, where*

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

and 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* $\mathcal{X}$*). For every* $\varepsilon > 0, \alpha \in (0,1)$

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

Lemma 5 is used by the following lemma:

Lemma 6. *Let* $\mathcal{F}$ *be a class of uniformly bounded functions on* $\mathcal{X}$ *(*$\forall f \in \mathcal{F} : |f| \leq K$*),* $x^{1:N} \in \mathcal{X}^N$*,* ϕ_N *be such that the* ε*-empirical covering numbers of* $\mathcal{F}$ *on all subsets of the multiset* $[x_1, \ldots, x_N]$ *are majorized by* $\phi_N(\varepsilon)$*. Let* $\mathcal{G}_2^1$ *denote the class of indicator functions* $\mathbb{I}_{\{f_1(x) \geq f_2(x)\}} : \mathcal{X} \to \{0,1\}$ *for any* $f_1, f_2 \in \mathcal{F}$*. Then for every* $\varepsilon > 0$*,*

$$\mathcal{N}(\varepsilon, \mathcal{F}^L \times \mathcal{F}^L, x^{1:N}) \leq \mathcal{N}_1\left(\frac{\varepsilon}{2L(L-1)K}, \mathcal{G}_2^1, x^{1:N}\right)^{L(L-1)} \phi_N(\varepsilon/2)^L,$$

where the distance of (f, Q') *and* $(g, \tilde{Q}') \in \mathcal{F}^L \times \mathcal{F}^L$ *in the left-hand-side covering number is defined in the unusual way*

$$l_{x^{1:N}}((f,Q'),(g,\tilde{Q}')) = \frac{1}{N}\sum_{t=1}^{N} |f(x_t, \hat{\pi}(x_t; Q')) - g(x_t, \hat{\pi}(x_t; \tilde{Q}'))|.$$

Finally, see Haussler [10] (and Anthony and Bartlett [7, Theorem 18.4]) for

Proposition 7 (Haussler [10] Corollary 3). *For any set* $\mathcal{X}$*, any points* $x^{1:N} \in \mathcal{X}^N$*, any class* $\mathcal{F}$ *of functions on* $\mathcal{X}$ *taking values in* $[0, K]$ *with pseudo-dimension* $V_{\mathcal{F}^+} < \infty$*, and any* $\varepsilon > 0$*,* $\mathcal{N}_1(\varepsilon, \mathcal{F}, x^{1:N}) \leq e(V_{\mathcal{F}^+} + 1)\left(\frac{2eK}{\varepsilon}\right)^{V_{\mathcal{F}^+}}$*.*

The following is the main result of this section:

Lemma 8. *Let Assumption 1,2, and 3 hold and let* $Q_{\max} \geq \hat{R}_{\max}/(1-\gamma)$*. 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. to* Q'*. Let* f' *be defined by* $f' = \operatorname{argmin}_{f \in \mathcal{F}^L} \sup_{h \in \mathcal{F}^L} \hat{L}_N(f, h; \hat{\pi})$*. Fix* $\varepsilon, \delta > 0$ *and assume that* $\mathcal{F}^L$ $\varepsilon/2$*-approximates the fixed point of* $T^{\hat{\pi}(\cdot;Q')}$*:*

$$\tilde{E}(\mathcal{F}) \stackrel{\text{def}}{=} \sup_{Q' \in \mathcal{F}^L} \inf_{f \in \mathcal{F}^L} \left\| f - T^{\hat{\pi}(\cdot;Q')} f \right\|_\nu \leq \varepsilon/2 \tag{4}$$

and that $\mathcal{F}^L$ *is* $\varepsilon/2$*-invariant w.r.t.* $\mathcal{T}$*:*

$$E(\mathcal{F}) \stackrel{\text{def}}{=} \sup_{f, Q' \in \mathcal{F}^L} \inf_{h \in \mathcal{F}^L} \left\| h - T^{\hat{\pi}(\cdot;Q')} f \right\|_\nu \leq \varepsilon/2. \tag{5}$$

If $N = \text{poly}(L, Q_{\max}, 1/b, 1/\pi_0, V_{\mathcal{F}^+}, V_{\mathcal{C}_2}, 1/\varepsilon, \log(1/\delta))$*, where the degree of the polynomial is* $O(1 + 1/\kappa)$*, then* $\mathbb{P}\left(\left\|f' - T^{\hat{\pi}} f'\right\|_\nu > \varepsilon\right) \leq \delta$*.*

Proof. (Sketch) We have to show that 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. By (4), it suffices to show that $L(f';Q') \stackrel{\text{def}}{=} \left\|f' - T^{\hat{\pi}(\cdot;Q')}f'\right\|_\nu^2$ is close to $\inf_{f\in\mathcal{F}^L} L(f;Q')$. Denote the difference of these two quantities by $\Delta(f',Q')$. Note that $\Delta(f',Q')$ is increased by taking its supremum over Q'. By (5), $L(f;Q')$ and $\bar{L}(f;Q') \stackrel{\text{def}}{=} \sup_{h\in\mathcal{F}^L} L(f,h;\hat{\pi}(\cdot;Q'))$, as functions of f and Q', are uniformly close to each other. This reduces the problem to bounding $\sup_{Q'}(\bar{L}(f';Q') - \inf_{f\in\mathcal{F}^L}\bar{L}(f;Q'))$. Since $\mathbb{E}\left[\hat{L}_N(f,h;\hat{\pi})\right] = L(f,h;\hat{\pi})$ holds for any $f,h \in \mathcal{F}^L$ and policy $\hat{\pi}$, by defining a suitable error criterion $l_{f,h,Q'}(x,a,r,y)$ in accordance with (1), the problem can be reduced to a usual uniform deviation problem over $\mathcal{L}_\mathcal{F} = \{l_{f,h,Q'} : f,h,Q' \in \mathcal{F}^L\}$. Since the samples are correlated, Pollard's tail inequality cannot be used directly. Instead, we use the method of Yu [8]: We split the samples into m_N pairs of blocks $\{(H_i,T_i)|i=1,\dots,m_N\}$, each block compromised of k_N samples (for simplicity we assume $N = 2m_Nk_N$) and then use Lemma 2 with $\mathcal{Z} = \mathcal{X}\times\mathcal{A}\times\mathbb{R}\times\mathcal{X}$, $\mathcal{F} = \mathcal{L}_\mathcal{F}$. The covering numbers of $\mathcal{L}_\mathcal{F}$ can be bounded by those of $\mathcal{F}^L$ and $\mathcal{F}^L\times\mathcal{F}^L$, where in the latter the distance is defined as in Lemma 6. Next we apply Lemma 6 and then Proposition 7 to bound the resulting three covering numbers in terms of $V_{\mathcal{F}^+}$ and $V_{\mathcal{C}_2}$ (note that the pseudo-dimension of $\mathcal{F}^L$ cannot exceed $LV_{\mathcal{F}^+}$). Defining $k_N = N^{\frac{1}{1+\kappa}}+1$, $m_N = N/(2k_N)$ and substituting $\beta_m \le e^{-bm^\kappa}$, we get the desired polynomial bound on the number of samples after some tedious calculations. □

4.2 Propagation of Errors

Let Q_k denote the kth iterate of (some) approximate policy iteration algorithm where the next iterates are computed by means of some Bellman-residual minimization procedure. Let π_k be the kth policy. Our aim here is to relate the performance of the policy π_K to the magnitude of the Bellman-residuals $\varepsilon_k \stackrel{\text{def}}{=} Q_k - T^{\pi_k}Q_k$, $0 \le k < K$.

Lemma 9. *Let $p \ge 1$. For any $\eta > 0$, there exists K that is linear in $\log(1/\eta)$ and $\log R_{\max}$ such that, if the $L_{p,\nu}$ norm of the Bellman-residuals are bounded by some constant ϵ, i.e. $\|\varepsilon_k\|_{p,\nu} \le \epsilon$ for all $0 \le k < K$, then*

$$\|Q^* - Q^{\pi_K}\|_\infty \le \frac{2\gamma}{(1-\gamma)^2}[C(\nu)]^{1/p}\epsilon + \eta \tag{6}$$

and

$$\|Q^* - Q^{\pi_K}\|_{p,\rho} \le \frac{2\gamma}{(1-\gamma)^2}[C(\rho,\nu)]^{1/p}\epsilon + \eta. \tag{7}$$

Proof. We have $C(\nu) \ge C(\rho,\nu)$ for any ρ. Thus, if the bound (7) holds for any ρ, choosing ρ to be a Dirac at each state implies that (6) also holds. Therefore, we only need to prove (7).

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 [5][Lemma 4], we get $Q^* - Q^{\pi_{k+1}} \le \gamma P^{\pi^*}(Q^* - Q^{\pi_k}) + \gamma E_k\epsilon_k$. Thus, by

induction, $Q^* - Q^{\pi_K} \le \gamma \sum_{k=0}^{K-1} (\gamma P^{\pi^*})^{K-k-1} E_k \epsilon_k + \eta_K$ with $\eta_K = (\gamma P^{\pi^*})^K (Q^* - Q^{\pi_0})$. Hence, $\|\eta_K\|_\infty \le 2Q_{\max}\gamma^K$.

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 pointwise in the above bound on $Q^* - Q^{\pi_K}$ 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

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

Here we introduced the positive coefficients $\alpha_k = \frac{(1-\gamma)\gamma^{K-k-1}}{1-\gamma^{K+1}}$, for $0 \le k < K$, and $\alpha_K = \frac{(1-\gamma)\gamma^K}{1-\gamma^{K+1}}$, and the operators $A_k = \frac{1-\gamma}{2}(P^{\pi^*})^{K-k-1} F_k$, for $0 \le k < K$, $A_K = (P^{\pi^*})^K$. Note that $\sum_{k=0}^{K} \alpha_k = 1$ and the operators A_k are stochastic when considered as a right-linear operators: for any $(x,a) \in \mathcal{X} \times \mathcal{A}$, $\lambda_{(x,a)}^{(k)}(B) = (A_k \chi_B)(x,a)$ is a probability measure and $(A_k Q)(x,a) = \int \lambda_{(x,a)}^{(k)}(dy) Q(y, \pi(y))$. Here $\chi_B : B(\mathcal{X} \times \mathcal{A}) \to [0,1]$ is defined by $\chi_B(x,a) = \mathbb{I}_{\{x \in B\}}$.

Let $\lambda_K = \left[\frac{2\gamma(1-\gamma^{K+1})}{(1-\gamma)^2} \right]^p$. Now, by using two times Jensen's inequality we get

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

From the definition of the coefficients $c(m)$, $\rho A_k \le (1-\gamma) \sum_{m \ge 0} \gamma^m c(m+K-k) \nu$ and we deduce

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

Replace α_k by their values, and from the definition of $C(\rho,\nu)$, and since $\|\varepsilon_k\|_{p,\nu} \le \epsilon$, we have

$$\|Q^* - Q^{\pi_K}\|_{p,\rho}^p \le \lambda_K \left[\frac{1}{1-\gamma^{K+1}} C(\rho,\nu) \epsilon^p + \frac{(1-\gamma)\gamma^K}{1-\gamma^{K+1}} (R_{\max})^p \right].$$

Thus there is K linear in $\log(1/\eta)$ and $\log R_{\max}$, e.g. such that $\gamma^K < \left[\frac{(1-\gamma)^2}{2\gamma R_{\max}} \eta \right]^p$ so that the second term is bounded by η^p. Thus, $\|Q^* - Q^{\pi_K}\|_{p,\rho}^p \le \left[\frac{2\gamma}{(1-\gamma)^2} \right]^p C(\rho,\nu) \epsilon^p + \eta^p$ and hence $\|Q^* - Q^{\pi_K}\|_{p,\rho} \le \frac{2\gamma}{(1-\gamma)^2} [C(\rho,\nu)]^{1/p} \epsilon + \eta$, finishing the proof. □

4.3 Proof of the Main Result

Proof. Consider the kth iteration of the algorithm. Let $\varepsilon_k = Q_k - T^{\pi_k} Q_k$. By the reasoning used in the proof of Lemma 9, we only need to prove the second part of the result. Since $0 < \gamma < 1$, there exists K that is linear in $\log(1/\epsilon)$ and $\log R_{\max}$ such that $\gamma^K < \left[\frac{(1-\gamma)^2}{2\gamma R_{\max}} \frac{\epsilon}{2}\right]^p$. Now, from Lemma 8, there exists N that is $\mathrm{poly}(L, Q_{\max}, 1/b, 1/\pi_0, V_{\mathcal{F}^+}, V_{\mathcal{C}_2}, 1/\varepsilon, \log(1/\delta))$, such that, for each $0 \le k < K$, $\mathbb{P}\left(\|\varepsilon_k\|_{p,\nu} > \frac{(1-\gamma)^2}{2C^{1/p}} \frac{\epsilon}{2}\right) < \delta/K$. Thus, $\mathbb{P}\left(\|\varepsilon_k\|_{p,\nu} > \frac{(1-\gamma)^2}{2C^{1/p}} \frac{\epsilon}{2}, \text{ for all } 0 \le k < K\right) < \delta$. Applying Lemma 9 with $\eta = \epsilon/2$ ends the proof. □

5 Discussion and Related Work

The idea of using value function approximation goes back to the early days of dynamic programming [11, 12]. With the recent growth of interest in reinforcement learning, work on value function approximation methods flourished [13, 14]. Recent theoretical results mostly concern supremum-norm approximation errors [15, 16], 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. [15, 16, 17, 18]. However, the growth-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. [19, 20, 1]). The need for analysing the behaviour of such algorithms provided the basic motivation for this work.

One of the main novelties of our paper is that we introduced a modified Bellman-residual that guarantees asymptotic consistency even with a single sample path.

The closest to the present work is the paper of Szepesvári and Munos [21]. However, as opposed to paper [21], here we dealt with a fitted policy iteration algorithm and unlike previously, we worked with dependent samples. The technique used to deal with dependent samples was to introduce (strong) mixing conditions on the trajectory and extending Pollard's inequality along the lines of Meir [22].

Also, the bounds developed in Section 4.2 are closely related to those developed in [5]. However, there only the case $C(\nu) < \infty$ was considered, whilst in this paper the analysis was extended to the significantly weaker condition $C(\nu, \rho) < \infty$. Although in [21] the authors considered a similar condition, there the propagation of the approximation errors was considered only in a value iteration context. Note that approximate value iteration *per se* is not suitable for learning from a single trajectory since approximate value iteration requires at least one successor state sample per action per sampled state.

That we had to work with fitted policy iteration significantly added to the complexity of the analysis, as the policy to be evaluated at stage k became dependent on the whole set of samples, introducing non-trivial correlations between successive approximants. In order to show that these correlations do not spoil convergence, we had to introduce problem-specific capacity conditions on the function class involved. Although these constraints are satisfied by many popular function

classes (e.g., regression trees, neural networks, etc.), when violated unstable behaviour may arise (i.e., increasing the sample size does not improves the performance).

Note that the conditions that dictate that $\mathcal{F}$ should be rich (namely, that $\mathcal{F}$ should be "almost invariant" under the family of operators $\mathcal{T} = \{T^{\hat{\pi}(\cdot;Q)} : Q \in \mathcal{F}\}$ and that $\mathcal{F}$ should be close to the set of fixed points of $\mathcal{T}$) are non-trivial to guarantee. One possibility is to put smoothness constraints on the transition dynamics and the immediate rewards. It is important to note, however, that both conditions are defined with respect to weighted L^2-norm. This is much less restrictive than if supremum-norm were used here. This observation suggests that one should probably look at frequency domain representations of systems in order to guarantee these properties. However, this is well out of the scope of the present work.

6 Conclusions

We have considered fitted policy iteration with Bellman-residual minimization. We modified the objective function to allow the procedure to work with a single (but long) trajectory. Our results show that the number of samples needed to achieve a small approximation error depend polynomially on the pseudo-dimension of the function class used in the empirical loss minimization step and the smoothness of the dynamics of the system. Future work should concern the evaluation of the proposed procedure in practice. The theoretical results can be extended in many directions: Continuous actions spaces will require substantial additional work as the present analysis relies crucially on the finiteness of the action set. The exploration of interplay between the MDPs dynamics and the approximability of the fixed points and the invariance of function sets with respect to policy evaluation operators also requires substantial further work.

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).

References

1. M. Lagoudakis and R. Parr. Least-squares policy iteration. *Journal of Machine Learning Research*, 4:1107–1149, 2003.
2. A. Antos, Cs. Szepesvári, and R. Munos. Learning near-optimal policies with fitted policy iteration and a single sample path: approximate iterative policy evaluation. (submitted to ICML'2006, 2006.
3. D. P. Bertsekas and S.E. Shreve. *Stochastic Optimal Control (The Discrete Time Case)*. Academic Press, New York, 1978.
4. R.S. Sutton and A.G. Barto. 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, 1987.

5. R. Munos. Error bounds for approximate policy iteration. *19th International Conference on Machine Learning*, pages 560–567, 2003.
6. S.P. Meyn and R. Tweedie. *Markov Chains and Stochastic Stability*. Springer-Verlag, New York, 1993.
7. M. Anthony and P. L. Bartlett. *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, 1999.
8. B. Yu. Rates of convergence for empirical processes of stationary mixing sequences. *The Annals of Probability*, 22(1):94–116, January 1994.
9. A. Nobel. Histogram regression estimation using data-dependent partitions. *Annals of Statistics*, 24(3):1084–1105, 1996.
10. D. Haussler. Sphere packing numbers for subsets of the boolean n-cube with bounded Vapnik-Chervonenkis dimension. *Journal of Combinatorial Theory Series A*, 69:217–232, 1995.
11. A.L. Samuel. Some studies in machine learning using the game of checkers. *IBM Journal on Research and Development*, pages 210–229, 1959. Reprinted in *Computers and Thought*, E.A. Feigenbaum and J. Feldman, editors, McGraw-Hill, New York, 1963.
12. R.E. Bellman and S.E. Dreyfus. Functional approximation and dynamic programming. *Math. Tables and other Aids Comp.*, 13:247–251, 1959.
13. Dimitri P. Bertsekas and J. Tsitsiklis. *Neuro-Dynamic Programming*. Athena Scientific, 1996.
14. Richard S. Sutton and Andrew G. Barto. Reinforcement learning: An introduction. *Bradford Book*, 1998.
15. Geoffrey J. Gordon. Stable function approximation in dynamic programming. In Armand Prieditis and Stuart Russell, editors, *Proceedings of the Twelfth International Conference on Machine Learning*, pages 261–268, San Francisco, CA, 1995. Morgan Kaufmann.
16. J. N. Tsitsiklis and B. Van Roy. Feature-based methods for large scale dynamic programming. *Machine Learning*, 22:59–94, 1996.
17. Carlos Guestrin, Daphne Koller, and Ronald Parr. Max-norm projections for factored mdps. *Proceedings of the International Joint Conference on Artificial Intelligence*, 2001.
18. D. Ernst, P. Geurts, and L. Wehenkel. Tree-based batch mode reinforcement learning. *Journal of Machine Learning Research*, 6:503–556, 2005.
19. X. Wang and T.G. Dietterich. Efficient value function approximation using regression trees. In *Proceedings of the IJCAI Workshop on Statistical Machine Learning for Large-Scale Optimization*, Stockholm, Sweden, 1999.
20. T. G. Dietterich and X. Wang. Batch value function approximation via support vectors. In T. G. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems 14*, Cambridge, MA, 2002. MIT Press.
21. Cs. Szepesvári and R. Munos. Finite time bounds for sampling based fitted value iteration. In *ICML'2005*, 2005.
22. R. Meir. Nonparametric time series prediction through adaptive model selection. *Machine Learning*, 39(1):5–34, April 2000.