Value-Iteration Based Fitted Policy Iteration: Learning with a Single Trajectory

Andras Antos, Csaba Szepesvari, Rémi Munos

Value-Iteration Based Fitted Policy Iteration: Learning with a Single Trajectory

András Antos
* Computer and Automation Research Inst. of the Hungarian Academy of Sciences
Kende u. 13-17, Budapest 1111, Hungary
Email: antos@sztaki.hu

Csaba Szepesvári*
Department of Computing Science
University of Alberta
Edmonton T6G 2E8, Canada
Email: szepesva@cs.ualberta.ca

Rémi Munos
SequeL team, INRIA Futurs
University of Lille
59653 Villeneuve d'Ascq, France
Email: remi.munos@inria.fr

***Abstract*— We consider batch reinforcement learning problems in continuous space, expected total discounted-reward Markovian Decision Problems when the training data is composed of the trajectory of some fixed behaviour policy. The algorithm studied is policy iteration where in successive iterations the action-value functions of the intermediate policies are obtained by means of approximate value iteration. PAC-style polynomial bounds are derived on the number of samples needed to guarantee near-optimal performance. The bounds depend 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. One of the main novelties of the paper is that new smoothness constraints are introduced thereby significantly extending the scope of previous results.**

I. Introduction

We consider batch reinforcement learning (RL) in continuous state space, finite action space expected total discounted-reward Markovian Decision Problems (MDP). The defining feature of batch reinforcement learning is that the training data is gathered 'off-line', typically by collecting samples whilst a fixed behaviour policy is controlling the system. In this paper we consider the problem of learning a good controller given a finite amount of training data. In fact, we are interested in the information content of such a finite-sample: The question we ask is 'what performance can be achieved given such a finite sample?'.

A natural algorithm for batch learning is fitted policy iteration used in quite a few previous empirical works (see, e.g., [1], [2]). In this paper we develop the theory for fitted policy iteration when the policies encountered are evaluated using trajectory-based approximation value iteration (TBAVI). TBAVI mimics approximate value iteration, with action-values replacing state-values and works by combining interleaving regression and with the steps of value iteration: Given a policy to evaluate, in the value iteration loop the next iterate is obtained by solving a regression problem where the targets are computed using one-step predictions based on the data, the most recent iterate and the policy. When this loop finishes, the next policy is formed as the greedy policy w.r.t. the action-value function returned. Hence, the whole procedure can be thought of as a sample-based approximate policy iteration algorithm, where the user may control the number of iterations and the regression procedure.

The main contribution of the paper is the finite-sample analysis of this algorithm. In particular, for the first time, we derive Probably Approximately Correctness (PAC) finite-sample performance bounds on the policy returned by this algorithm. The most important factors influencing this performance are the mixing-rate of the trajectory, some smoothness (stochasticity) parameters of the dynamics of the associated MDP, the capacity of the function space $\mathcal{F}$ and the number of policy improvement and evaluation steps.

One difficulty of the analysis is that the same training data is used throughout all the iterations of the algorithm. The motivation for using all the data is that we observed empirically that it is more efficient to use all data in all iterations than e.g. processing data in disjoint blocks. However, when all data are used throughout all iterations the previous iterates become dependent on the entire data set. As a result, the straightforward application of supervised-learning PAC-techniques fails and one may suspect that without additional restrictions (compared to the usual capacity assumptions) the algorithm may degenerate. We work out a set of conditions on the function space, under which no such degeneracy happens. Another contribution is that a new set of smoothness conditions are worked out on the MDP's dynamics that might be easier to verify than similar previous conditions, such as those of e.g. by [3] and that we relate previous conditions on the MDP's dynamics to bounds on the MDP's top-Lyapunov exponent.

The organization of the paper is as follows: In the next section (Section II) we introduce the necessary symbols and notation. The algorithm is given in Section III. Our main result is presented in Section IV along with a sketch of its proof. In Section V the conditions of the main result and related work are discussed. Our conclusions are drawn in Section VI.

II. Notation

For a measurable space with domain $\mathcal{X}$ we let $M(\mathcal{X})$ denote the set of all probability measures over $\mathcal{X}$ and let $\mathcal{B}(\mathcal{X})$ be the system of measurable sets of $\mathcal{X}$. For $\nu \in M(\mathcal{X})$ and $f : \mathcal{X} \to \mathbb{R}$ measurable we let $\|f\|_{p,\nu}$ (for $p \geq 1$) denote the $L^p(\nu)$-norm of f: $\|f\|_{p,\nu}^p = \int |f(s)|^p \nu(ds)$. We simply write $\|f\|_\nu$ for the L^2-norm of f. We denote the space of bounded measurable functions with domain $\mathcal{X}$ by $B(\mathcal{X})$. Further, the

space of measurable functions bounded by $0 < K < \infty$ shall be denoted 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 of event E, $\mathbf{1}$ denotes the function that takes on the constant value one everywhere over its domain.

A discounted 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* with $P(\cdot|x, a)$ 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 assumption 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 expected immediate reward function, $r(x, a) = \int rS(dr|x, a)$, is uniformly bounded. 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 non-stationary Markovian if the distribution depends only on the last state of the observation sequence and the length of the history. A policy is called stationary (Markovian) if the distribution depends only on the last state of the observation sequence (and not on the length of the history).

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 π^* is called optimal if $V^{\pi^*}(x) = V^*(x)$ for all $x \in \mathcal{X}$. The function $Q^*(x, a)$ is defined analogously: $Q^*(x, a) = \sup_\pi Q^\pi(x, a)$. 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 we write $\pi = \hat{\pi}(\cdot; Q)$, 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 [4]. For a deterministic stationary 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 known that $V^\pi, Q^\pi(\cdot, a)$ are bounded by $R_{\max}/(1 - \gamma)$, just like Q^* and V^*.

We assign to any deterministic stationary policy $\pi : \mathcal{X} \to \mathcal{A}$ the operator $E^\pi : B(\mathcal{X} \times \mathcal{A}) \to B(\mathcal{X})$ defined by $(E^\pi Q)(x) = Q(x, \pi(x))$. Further, we define two operators corresponding to the transition probability kernel P. The 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 the 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 $\mathcal{X}$ with the definition $(\rho P)(dy) = \frac{1}{L}\sum_{a\in\mathcal{A}} \int P(dy|x, a)\rho(dx)$. 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 a 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 $f \equiv (f_1, \ldots, f_L)$, $f(x, a_j) = f_j(x)$, $j = 1, \ldots, L$. For $\nu \in M(\mathcal{X})$, we extend $\|\cdot\|_{p,\nu}$ ($p \ge 1$) to $\mathcal{F}^L$ by $\|f\|_{p,\nu} = (\frac{1}{L}\sum_{j=1}^L \|f_j\|_{p,\nu}^p)^{1/p}$.

FittedPolicyQ(D,K,Q_0,PEval,π)
// D: samples (trajectory)
// K: number of iterations
// Q_0: Initial Q-function
// PEval: Policy evaluation routine
// π: Behaviour policy that generated the samples
$Q \leftarrow Q_0$ // 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 Policy Iteration

III. 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 a sense that $\{(X_t, A_t, R_t)\}$ mixes fast (this will be made precise in Section IV) and that X_t is stationary.

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, π is the behaviour policy used to generate the samples, *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 implement *PEval*. In this paper we consider an iterative procedure that can be thought of as a fitted value iteration algorithm applied to action-value functions and a single sample path.

Define the functions $(Q^m)_{1\le m\le M}$ over $\mathcal{X} \times \mathcal{A}$ as follows.

Start with $Q^0 = 0$ and solve M times the fitting problem:

$$Q^m = \operatorname*{arginf}_{f \in \mathcal{F}^L} \hat{L}_N(f; Q^{m-1}, \hat{\pi}), \tag{1}$$

where $\hat{L}_N(f; Q, \hat{\pi}) = \frac{1}{LN} \sum_{t=1}^{N} \frac{d_t^2(f(\cdot, A_t); Q, \hat{\pi})}{\pi(A_t|X_t)}$, with

$$d_t(f; Q, \hat{\pi}) = R_t + \gamma Q(X_{t+1}, \hat{\pi}(X_{t+1})) - f(X_t)$$

being the t^{th} TD-error. In the following we shall call this procedure the *trajectory-based approximate value iteration (TBAVI)* algorithm. Note that in (1) the minimization can be computed in a componentwise fashion:

$$Q^m(\cdot, a_j) = \operatorname*{arginf}_{f \in \mathcal{F}} \hat{L}_{N,j}(f; Q^{m-1}, \hat{\pi}),$$

where

$$\hat{L}_{N,j}(f; Q, \hat{\pi}) = \frac{1}{N} \sum_{t=1}^{N} \frac{\mathbb{I}_{\{A_t = a_j\}}}{\pi(a_j|X_t)} d_t^2(f; Q, \hat{\pi}). \tag{2}$$

(Strictly speaking, the normalization with the behaviour policy π is not needed for the soundness of the algorithm. It is included for simplifying the analysis only.)

For any Q, $\hat{\pi}$, let

$$L_j(f; Q, \hat{\pi}) \stackrel{\text{def}}{=} \mathbb{E}\left[\hat{L}_{N,j}(f; Q, \hat{\pi})\right].$$

TBAVI is motivated by the identity

$$L_j(f; Q, \hat{\pi}) = \left\| f - (T^{\hat{\pi}} Q)_j \right\|_\nu^2 + L_j^*(Q, \hat{\pi}), \tag{3}$$

where

$$L_j^*(Q, \hat{\pi}) = \mathbb{E}\left[\mathrm{Var}\left[d_0(0, Q; \hat{\pi}) | X_0, A_0 = a_j\right]\right]$$

is a constant that is independent of f. Note that this identity holds for any fixed Q and $\hat{\pi}$. Hence, $\operatorname{arginf}_{f \in \mathcal{F}} L_j(f; Q, \hat{\pi}) = \operatorname{arginf}_{f \in \mathcal{F}} \left\| f - (T^{\hat{\pi}} Q)_j \right\|_\nu^2$ and we may think of the procedure as approximate value iteration (projecting $(T^{\hat{\pi}} Q)_j$ on the space $\mathcal{F}$ w.r.t. $\|\cdot\|_\nu$ distances) using empirical risk minimization.

However, this analogue is not entirely valid. First, the samples used in the procedure come from a single-trajectory underlying the behaviour policy and hence are correlated. We deal with this using a 'blocking' technique due to Yu [5], similarly to the work of Meir [6] who used it to analyse time-series prediction. The other difficulty is caused by the inherent structure of the algorithm: New iterates depend on previous ones. Denoting the k^{th} policy by π_k and the m^{th} iterate of the k^{th} call of TBAVI by Q_k^m, we find that in general $\mathbb{E}\left[\hat{L}_{N,j}(f; Q_k^m, \pi_k)\right] \neq \|f - (T^{\pi_k} Q_k^m)_j\|_\nu^2 + L_j^*(Q_k^m, \pi_k)$ since Q_k^m and π_k are random. One may try to fix this by replacing the first quantity by $\mathbb{E}\left[\hat{L}_{N,j}(f; Q_k^m, \pi_k) | Q_k^m, \pi_k\right]$. However, since Q_k^m and π_k depend on the same set of samples as used in the definition of the loss function $\hat{L}_{N,j}$, this change does not help either. At this point one might wonder if TBAVI is indeed a sound procedure.

In the next section we will show that it is, provided that some (reasonable) assumptions are made on the function space $\mathcal{F}$. Clearly, if $\mathcal{F}$ is very restricted, i.e., when it consists of a single function only (admittedly not a very interesting case when it comes to the optimization step), then the analogue to empirical risk minimization revives. At this point one may conjecture that if $\mathcal{F}$ is sufficiently restricted then Q_k^m and, more importantly, π_k will be sufficiently restricted (since π_k is by definition the greedy policy with respect to $Q_{k-1} \in \mathcal{F}^L$) the analogue will continue to hold. In fact, it turns out that in addition to requiring (the usual condition) that the pseudo-dimension of $\mathcal{F}$ is finite, it suffices to assume that the VC-dimension of the 0 level-sets corresponding to functions of the form $f_1 - f_2$ ($f_1, f_2 \in \mathcal{F}$) is finite. We call this quantity the *VC-crossing dimension* of $\mathcal{F}$ [7].

IV. 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

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

$\{Z_t\}$ is said to be β-mixing if $\beta_m \to 0$ as $m \to \infty$.

For a β-mixing process that mixes at an *exponential* rate we let b, κ be defined by $\beta_m = O(\exp(-bm^\kappa))$. For our results we will need the following conditions on the training data:

Assumption 2 (Sample Path Properties): Let $\{(X_t, A_t, R_t)\}_t$ be the sample path of π. We assume that 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) and 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 that depend on the MDP. Let ρ be the distribution 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, $a \in \mathcal{A}$, $P(B|x, a) \leq C(\nu)\nu(B)$ (if no such constant exists, we set $C(\nu) = \infty$). Now, for all integer $m \geq 1$, we define $c(m) \in \mathbb{R}^+ \cup \{+\infty\}$ to be the smallest constant such that, for any m stationary policies $\pi_1, \pi_2, \ldots, \pi_m$,

$$\rho P^{\pi_1} P^{\pi_2} \ldots P^{\pi_m} \leq c(m)\nu, \tag{4}$$

Note that these constants[1] depend on ρ and ν.

[1]Again, if there exists no such constants, we simply set $c(m) = \infty$. Note that in (4) $\leq$ is used to compare two operators. The meaning of $\leq$ in comparing operators H, G is the usual: $H \leq G$ iff $Hf \leq Gf$ holds for all $f \in \mathrm{Dom}(H)$. Here ν is viewed as an operator acting on $B(\mathcal{X} \times \mathcal{A})$.

We let $C_2(\rho,\nu)$ and $C_3(\rho,\nu)$ denote the *second* and *third order future state distribution smoothness constants*, defined by

$$C_2(\rho,\nu) = (1-\gamma)^2 \sum_{m\geq 1} m\gamma^{m-1} c(m), \quad (5)$$

$$C_3(\rho,\nu) = (1-\gamma)^3 \sum_{m\geq 1} \frac{m(m+1)}{2}\gamma^{m-1} c(m). \quad (6)$$

Moreover, we define the *transition probability ρ-smoothness constant* $K(\rho)$ to be the smallest constant such that, for all stationary policy π,

$$\rho P^\pi \leq K(\rho)\rho,$$

and the *discounted future state distribution ρ-smoothness constant* $K_\gamma(\rho)$ to be the smallest constant such that, for all stationary policy π,

$$\rho(1-\gamma)(I-\gamma P^\pi)^{-1} \leq K_\gamma(\rho)\rho.$$

The results are stated below under the conditions that $C_i(\rho,\nu)$, $i=2,3$ are finite, or that $K(\rho), K_\gamma(\rho)$ are finite. A discussion of these conditions will be provided in Section V-A.

During the course of the proof, we will need several capacity concepts of function sets. We assume that the reader is familiar with concepts of VC-dimension (see, e.g. [8]). We introduce covering numbers because slightly different definitions of it exist in the literature:

For a semi-metric space $(\mathcal{M},d)$ and for $\varepsilon>0$, define the covering number $\mathcal{N}(\varepsilon,\mathcal{M},d)$ as the smallest value of m for which there exist g_1,g_2,...,$g_m \in \mathcal{M}$ 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 $\mathcal{N}(\varepsilon,\mathcal{F},l_{x^{1:N}})$ will 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.[2] 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. This latter quantity is called the *VC-crossing dimension* of $\mathcal{F}$ [7].

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 $f\in\mathcal{F}$.

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 $O(\epsilon/(L\log(2/\epsilon)))$-invariant for operators from $\mathcal{T}$ and with respect to the norm $\|\cdot\|_\nu$ and that $\mathcal{F}$ satisfies Assumption 3. Then there exists integers N,M,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,M,K that are polynomials of the same quantities, except that $C(\nu)$ is replaced by either $\max(C_2(\rho,\nu),C_3(\rho,\nu))$ or $\max(K(\rho),K_\gamma(\rho))$, such that

$$\mathbb{P}\left(\|Q^*-Q^{\pi_K}\|_\rho>\epsilon\right)\leq\delta.$$

Note that from $\|Q^*-Q^{\pi_K}\|_\rho$ bound on $\|V^*-V^{\pi_K}\|_\rho$ does not follow immediately. However, the techniques used in this paper can be used to get such a bound (see [9]).

At this point one might also wonder if a function space required in the statement of the problem exists at all. In fact, if no restriction is put on the MDP's dynamics (or the state space) then this seems hard to guarantee. However, continuity of the MDP's dynamics is sufficient to ensure the existence of these spaces:

Definition 4: If there exists finite α, L_p, L_r such that the conditions

$$\sup_{\substack{(B,x,x',a)\in\\ \mathcal{B}(\mathcal{X})\times\mathcal{X}^2\times\mathcal{A}}} |P(B|x,a)-P(B|x',a)| \leq L_P\|x-x'\|^\alpha,$$

$$\sup_{(x,x',a)\in\mathcal{X}^2\times\mathcal{A}} |r(x,a)-r(x',a)| \leq L_r\|x-x'\|^\alpha.$$

are satisfied then we say that the MDP is (α,L_p,L_r)-Lipschitzian.

Pick any stationary policy π and assume that the MDP is (α,L_p,L_r)-Lipschitzian. It is easy to check that if $Q\in B(\mathcal{X}\times\mathcal{A};Q_{\max})$ then $T^\pi Q$ is $L=(L_r+\gamma Q_{\max}L_P)$-Lipschitzian; i.e., for any $x,x'\in\mathcal{X}$,

$$|(T^\pi Q)(x,a)-(T^\pi Q)(x',a)|\leq(L_r+\gamma Q_{\max}L_P)|x-x'|^\alpha.$$

Note that L is independent of the policy chosen. Denote the space of (L,α)-Lipschitz action-value functions by $\mathrm{Lip}(\alpha;L)$. Consider a nested sequence of function spaces $\{\mathcal{F}_n\}_n$ such that $\mathcal{F}_n\subset B(\mathcal{X}\times\mathcal{A};Q_{\max})$. Then $T^\pi\mathcal{F}_n\stackrel{\text{def}}{=}\{T^\pi Q\,|Q\in\mathcal{F}_n\}$ will only contain L-Lipschitz functions:

$$T^\pi\mathcal{F}_n\subset\mathrm{Lip}(\alpha;L,Q_{\max})\stackrel{\text{def}}{=}\mathrm{Lip}(\alpha;L)\cap B(\mathcal{X}\times\mathcal{A};Q_{\max}).$$

Let $d_\nu(\mathcal{F},\mathcal{G})=\sup_{f\in\mathcal{F}}\inf_{g\in\mathcal{G}}\|g-Tf\|_\nu$. Then

$$d_\nu(T^\pi\mathcal{F}_n,\mathcal{F}_n)\leq d_\nu(\mathrm{Lip}(\alpha;L,Q_{\max}),\mathcal{F}_n).$$

If we make the right-hand side converge to zero then so will do left-hand side. Now, $d_\nu(\mathrm{Lip}(\alpha;L,Q_{\max}),\mathcal{F}_n)\leq d_\nu(\mathrm{Lip}(\alpha;L),\mathcal{F}_n)$, where we exploited that $\mathrm{Lip}(\alpha,L)=\cup_{Q_{\max}>0}\mathrm{Lip}(\alpha;L,Q_{\max})$ which holds thanks to the compactness of $\mathcal{X}$. The idea is to select $\{\mathcal{F}_n\}$ such that for any $\alpha,L>0$, $\lim_{n\to\infty}d_{p,\mu}(\mathrm{Lip}(\alpha;L),\mathcal{F}_n)=0$. Such function sequences are called *universal*. Their existence follows by standard results of approximation theory [10] and in fact many

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

popular choices (neural nets, regression trees, wavelets, etc.) satisfy this requirement. The VC-dimension of these function spaces is typically finite. Moreover, amongst them those that are linear have finite VC-crossing dimension, as well.

One remaining issue is that classical approximation spaces are not uniformly bounded (i.e., the functions in them do not assume a uniform bound). One solution is to use truncations: Let Γ_q be the truncation operator: $\Gamma_q r = r$ iff $|r| \leq q$ and $\Gamma_q r = \mathrm{sign}(r)q$ otherwise ($q, r \in \mathbb{R}$). Then $\Gamma_{Q_{\max}} \mathrm{Lip}(\alpha; L) = \mathrm{Lip}(\alpha; L, Q_{\max})$ and thus $d_\nu(\mathrm{Lip}(\alpha; L, Q_{\max}), \Gamma_{Q_{\max}} \mathcal{F}_n) = d_\nu(\Gamma_{Q_{\max}} \mathrm{Lip}(\alpha; L), \Gamma_{Q_{\max}} \mathcal{F}_n) \leq d_\nu(\mathrm{Lip}(\alpha; L), \mathcal{F}_n)$,. Since neither the VC-dimension nor the VC-crossing dimension is increased by truncation, if $\{\mathcal{F}_n\}$ is universal with finite VC- and VC-crossing dimensions and since π was arbitrary, for large enough n $\{\Gamma_{Q_{\max}} \mathcal{F}_n\}_n$ will satisfy the conditions of the theorem.

However, fitting using these truncated spaces might be a difficult optimization problem. From the point of view of implementations, it might be a better choice to do the fitting first and then truncate the results. Our theorem can be extended to such a procedure by following standard techniques (cf. Chapter 10 of [11]).

A. 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. We start with the following lemmata:

Lemma 2: Suppose that $Z_0, \ldots, Z_N \in \mathcal{Z}$ 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 in [5], and $H = \bigcup_{i=1}^{m_N} H_i$ as in the proof of Lemma 6 below, and that $\mathcal{F}$ is a permissible class of $\mathcal{Z} \to [-K, K]$ functions. Then

$$\mathbb{P}\Big(\sup_{f \in \mathcal{F}} \Big| \frac{1}{N} \sum_{t=1}^{N} f(Z_t) - \mathbb{E}\left[f(Z_0)\right] \Big| > \varepsilon \Big)$$

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

(This lemma is based on Lemma 4.2 in [5].)

By a partition of $\mathcal{X}$ we mean an ordered list of disjoint subsets of $\mathcal{X}$ whose union covers $\mathcal{X}$ and a partition family is just a set of partitions. Let Π be such a family of partitions of $\mathcal{X}$. Define the cell count of Π as follows:

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

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

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

We refine Proposition 1 in [12] to a bound in terms of the covering number of the partition family: :

Lemma 3: Let $x^{1:N} \in \mathcal{X}^N$, let $\mathcal{G}$ be a class of uniformly bounded functions on $\mathcal{X}$ ($\forall g \in \mathcal{G} : |g| \leq K$) whose empirical covering numbers on all subsets of $x^{1:N}$ are majorized by $\phi_N(\cdot)$, and let Π be any partition family with $m(\Pi) < \infty$. 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 $\mu_N(A) = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}_{\{x_i \in A\}}$ is the empirical measure corresponding to $x^{1:N}$ for every Borel set $A \subset \mathcal{X}$. 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 3 is used by the following lemma:

Lemma 4: Let $x^{1:N} \in \mathcal{X}^N$, let $\mathcal{F}$ be a class of uniformly bounded functions on $\mathcal{X}$ ($\forall f \in \mathcal{F} : |f| \leq K$) whose empirical covering numbers on all subsets of $x^{1:N}$ are majorized by $\phi_N(\cdot)$, and 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 $\varepsilon > 0$,

$$\mathcal{N}(\varepsilon, \mathcal{F}^L \times \mathcal{F}^L, x^{1:N})$$

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

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

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

Finally, see [13] (and [8, Theorem 18.4]) for

Proposition 5 ([13] 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 6 (PAC-bound for TBAVI): Let Assumption 1,2, and 3 hold and let $Q_{\max} \geq \hat{R}_{\max}/(1-\gamma)$. Fix $1 \leq j \leq L$. Let Q, Q' be real-valued random functions over $\mathcal{X} \times \mathcal{A}$, $Q(\omega)$,$Q'(\omega) \in \mathcal{F}^L$ (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'_j be defined by $f'_j = \mathrm{arginf}_{f \in \mathcal{F}} \hat{L}_{N,j}(f; Q, \hat{\pi})$. Fix $\varepsilon, \delta > 0$ and assume that $\mathcal{F}^L$ is $\varepsilon/(2L)$-invariant w.r.t. $\mathcal{T}$, implying

$$E_j(\mathcal{F}) \stackrel{\text{def}}{=} \sup_{Q, Q' \in \mathcal{F}^L} \inf_{f \in \mathcal{F}} \left\| f - (T^{\hat{\pi}(\cdot; Q')} Q)_j \right\|_\nu \leq \varepsilon/2. \quad (7)$$

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

Proof: (Sketch) We have to show that f'_j is close to $(T^{\hat{\pi}(\cdot; Q')} Q)_j$ with high probability, noting that Q and Q' may not be independent from the sample path. By (7), it suffices to show that $\|f'_j - (T^{\hat{\pi}(\cdot; Q')} Q)_j\|_\nu^2$

is close to $\inf_{f\in\mathcal{F}}\|f-(T^{\hat\pi(\cdot;Q')}Q)_j\|_\nu^2$. Denote the difference of these two quantities by $\Delta(f_j',Q,Q')$. Note that $\Delta(f_j',Q,Q')$ is increased by taking its supremum over Q and Q'. In fact, $\sup_{Q,Q'}\Delta(f_j',Q,Q') = \sup_{Q,Q'}(L_j(f_j';Q,\hat\pi(\cdot;Q')) - \inf_{f\in\mathcal{F}}L_j(f;Q,\hat\pi(\cdot;Q')))$, thanks to (3). Since $\mathbb{E}[\hat L_{N,j}(f;Q,\hat\pi)] = L_j(f;Q,\hat\pi)$ holds for any $f\in\mathcal{F}$, $Q\in\mathcal{F}^L$ and policy $\hat\pi$, by defining a suitable error criterion $l^{(j)}_{f,Q,Q'}(x,a,r,y)$ in accordance with (2), the problem can be reduced to a usual uniform deviation problem over $\mathcal{L}_{\mathcal{F},j} = \{l^{(j)}_{f,Q,Q'} : f\in\mathcal{F}, Q,Q'\in\mathcal{F}^L\}$. Since the samples are correlated, Pollard's tail inequality cannot be used directly. Instead, we use the method of [5]: 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 $\mathcal{Z}=\mathcal{X}\times\mathcal{A}\times\mathbb{R}\times\mathcal{X}$, $\mathcal{F}=\mathcal{L}_{\mathcal{F},j}$. The covering numbers of $\mathcal{L}_{\mathcal{F},j}$ can be bounded by those of $\mathcal{F}$ and $\mathcal{F}^L\times\mathcal{F}^L$, where in the latter the distance is defined as in Lemma 4. Next we apply Lemma 4 and then Proposition 5 to bound the resulting three covering numbers in terms of $V_{\mathcal{F}^+}$ and $V_{\mathcal{C}_2}$. 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. ■

B. Propagation of Errors

In this section we analyse how the errors made in the innermost loop propagate through FPI. We us the following notation in this section: Let the initial policy be π_0. Let Q_k^m denote the m^{th} approximation to the action-value function Q^{π_k} of $\pi_k=\hat\pi(\cdot;Q_{k-1})$. Let the error made in the m^{th} step of TBAVI be ε_k^m: $\varepsilon_k^m \stackrel{\text{def}}{=} Q_k^{m+1}-T^{\pi_k}Q_k^m$, $0\le m\le M-1$. Further, let $Q_k = Q_k^M$. Note that in this section we deal with pointwise bounds (i.e., for a fixed element ω of the probability space).

Lemma 7: Let $p\ge 1$. For any $\eta>0$, there exists K and M that are linear in $\log(1/\eta)$ (and $\log R_{\max}$) such that, if the $L_{p,\nu}$ norm of the approximation errors ε_k^m is bounded by some ϵ, i.e., $\|\varepsilon_k^m\|_{p,\nu}\le\epsilon$ for all $0\le k<K, 0\le m<M$, then the following bounds hold:

$$\|Q^*-Q^{\pi_K}\|_\infty \le \frac{2\gamma}{(1-\gamma)^3}[C(\nu)]^{1/p}\epsilon+\eta, \quad (8)$$

$$\|Q^*-Q^{\pi_K}\|_{p,\rho} \le \frac{2\gamma}{(1-\gamma)^3}[C_{2,3}(\rho,\nu)]^{1/p}\epsilon+\eta, \quad (9)$$

$$\|Q^*-Q^{\pi_K}\|_{p,\rho} \le \frac{2\gamma}{(1-\gamma)^3}[K'(\rho)]^{1/p}\epsilon+\eta, \quad (10)$$

for arbitrary distribution ρ and where $C_{2,3}(\rho,\nu) = \max\{C_2(\rho,\nu),C_3(\rho,\nu)\}$ and $K'(\rho) = KK(\rho)K_\gamma(\rho)^2[K_\gamma(\rho)(1+\gamma K(\rho))+1-\gamma]/(2\gamma)$.

The message of this result is that if the individual errors of the iterates are called then the final error can be controlled, too.

Proof: For $i=2$ or 3, we have $C(\nu)\ge C_i(\rho,\nu)$ for any ρ. Thus, if (9) holds for any ρ, choosing ρ to be a Dirac at each state implies that (8) also holds. Therefore, (9) implies (8).

Let $E_k = P^{\pi_{k+1}}(I-\gamma P^{\pi_{k+1}})^{-1}(I-\gamma P^{\pi_k})-P^{\pi^*}$, $Z_k = Q^*-Q^{\pi_k}$, $U_k^m = Q_k^m - Q^{\pi_k}$, $W_k = Q_k - Q^{\pi_k}$. Closely following the proof of Lemma 4 of [14] (applied to Q-functions), we have: $Z_{k+1}\le\gamma P^{\pi^*}Z_k+\gamma E_kW_k$, for all $0\le k\le K-1$. Hence, by induction

$$Z_K\le\gamma\sum_{k=0}^{K-1}(\gamma P^{\pi^*})^{K-k-1}E_kW_k+(\gamma P^{\pi^*})^KZ_0.$$

On the other hand, $U_k^{m+1} = \varepsilon_k^m+\gamma P^{\pi_k}U_k^m$, thus by induction (on m), $W_k = U_k^M = \sum_{m=0}^{M-1}(\gamma P^{\pi_k})^{M-1-m}\varepsilon_k^m + (\gamma P^{\pi_k})^MU_k^0$. Hence,

$$\begin{aligned} Z_K \le\ & \gamma\sum_{k=0}^{K-1}\sum_{m=0}^{M-1}(\gamma P^{\pi^*})^{K-k-1}E_k(\gamma P^{\pi_k})^{M-1-m}\varepsilon_k^m \\ & +\gamma\sum_{k=0}^{K-1}(\gamma P^{\pi^*})^{K-k-1}E_k(\gamma P^{\pi_k})^MU_k^0 \\ & +(\gamma P^{\pi^*})^KZ_0. \end{aligned} \quad (11)$$

By taking the absolute value pointwise in this latter inequality, we get

$$\begin{aligned} Z_K \le\ & \gamma\sum_{k=0}^{K-1}\sum_{m=0}^{M-1}(\gamma P^{\pi^*})^{K-k-1}F_k(\gamma P^{\pi_k})^{M-1-m}|\varepsilon_k^m| \\ & +\gamma\sum_{k=0}^{K-1}(\gamma P^{\pi^*})^{K-k-1}F_k(\gamma P^{\pi_k})^M|U_k^0|+(\gamma P^{\pi^*})^KZ_0 \end{aligned}$$

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

Now, given that $||U_k^0||_\infty, ||Z_0||_\infty \le 2R_{\max}/(1-\gamma)$, we rewrite the above inequality as

$$\begin{aligned} Z_K \le\ & \sigma\Big[\sum_{k=0}^{K-1}\sum_{m=0}^{M-1}\frac{\alpha_k^m}{\sigma}A_k^m|\varepsilon_k^m|+\sum_{k=0}^{K-1}\frac{\alpha_k^M}{\sigma}A_k^M2R_{\max}\mathbf{1} \\ & +\frac{\alpha_K}{\sigma}A_K\frac{(1-\gamma)^2}{\gamma}R_{\max}\mathbf{1}\Big] \end{aligned}$$

where $\mathbf{1}$ is the vector with components 1, and α_k^m, α_k^M, α_K, σ, ($0\le k<K$, $0\le m<M$) are real numbers defined by:

$$\begin{aligned} \alpha_k^m &= \frac{2}{1-\gamma}\gamma^{K+M-k-m-1}, \\ \alpha_k^M &= \frac{2}{(1-\gamma)^2}\gamma^{K+M-k}, \\ \alpha_K &= \frac{2\gamma}{(1-\gamma)^3}\gamma^K \end{aligned}$$

which satisfy $\sigma = \sum_{k=0}^{K-1}(\sum_{m=0}^{M-1}\alpha_k^m+\alpha_k^M)+\alpha_K = \frac{2\gamma}{(1-\gamma)^3}$. And the stochastic (right-linear) operators A_k^m, A_k^M, A_K ($0\le k<K$, $0\le m<M$) are defined by:

$$\begin{aligned} A_k^m &= \frac{1-\gamma}{2}(P^{\pi^*})^{K-k-1}F_k(P^{\pi_k})^{M-m-1}, \\ A_k^M &= \frac{1-\gamma}{2}(P^{\pi^*})^{K-k-1}F_k(P^{\pi_k})^M, \\ A_K &= (P^{\pi^*})^K. \end{aligned}$$

By definition, $\|Z_K\|_{p,\rho}^p = \rho|Z_K|^p$. Plugging in the bound on Z_K and using Jensen's inequality twice, we get

$$\begin{aligned}\|Z_K\|_{p,\rho}^p &= \rho|Z_K|^p \\ &\leq \sigma^p\Big[\rho\sum_{k=0}^{K-1}\sum_{m=0}^{M-1}\frac{\alpha_k^m}{\sigma}A_k^m|\varepsilon_k^m|^p + \sum_{k=0}^{K-1}\frac{\alpha_k^M}{\sigma}(2R_{\max})^p \\ &\quad +\frac{\alpha_K}{\sigma}\big[\frac{(1-\gamma)^2}{\gamma}R_{\max}\big]^p\Big]. \qquad (12)\end{aligned}$$

Let r_K^M be the sum of the last two terms in the previous bound. Then

$$\begin{aligned}r_K^M &= \sigma^{p-1}\left(\sum_{k=0}^{K-1}\alpha_k^M(2R_{\max})^p + \alpha_K\big[\frac{(1-\gamma)^2}{\gamma}R_{\max}\big]^p\right) \\ &\leq \frac{2\gamma}{(1-\gamma)^3}\big(2^p+\gamma^{-p}\big)\gamma^{\min\{K,M\}}R_{\max}^p\end{aligned}$$

is smaller than η^p for K and M linear in $\log(1/\eta)$ (and $\log R_{\max}$), i.e. whenever $\gamma^{\min\{K,M\}} \leq \frac{1}{2^p+\gamma^{-p}}\Big[\frac{(1-\gamma)^3}{2\gamma}\frac{\eta}{\mathbb{R}_{\max}}\Big]^p$.

Let us bound the first term in (12) in two different manners to deduce (9) and (10). First, we use the definition of the $c(m)$ and $C_i(\rho,\nu)$ constants, thus

$$\begin{aligned}&\sigma^p\rho\sum_{k=0}^{K-1}\sum_{m=0}^{M-1}\frac{\alpha_k^m}{\sigma}A_k^m \\ &\leq \sigma^{p-1}\sum_{k=0}^{K-1}\sum_{m=0}^{M-1}\gamma^{K+M-k-m-1} \\ &\quad \big[\sum_{n\geq 0}\gamma^n\big(c(K+M-k-m-1+n) \\ &\quad +\gamma c(K+M-k-m+n)\big) \\ &\quad +c(K+M-k-m-1)\big]\nu \\ &\leq \sigma^{p-1}\Big[\frac{\gamma}{(1-\gamma)^2}C_2(\rho,\nu)+\frac{(1+\gamma)\gamma}{(1-\gamma)^3}C_3(\rho,\nu)\Big]\nu \\ &\leq \Big[\frac{2\gamma}{(1-\gamma)^3}\Big]^p C_{2,3}(\rho,\nu)\nu,\end{aligned}$$

which, together with the bound on r_K^M, gives (9).

Now, we derive similar bounds using the $K(\rho)$ and $K_\gamma(\rho)$ constants: the first term in (12), divided by σ^{p-1}, satisfies

$$\begin{aligned}&\rho\sum_{k=0}^{K-1}\sum_{m=0}^{M-1}\alpha_k^m A_k^m \\ &\leq \rho\sum_{k=0}^{K-1}\gamma^{K-k-1}(P^{\pi^*})^{K-k-1}F_k\sum_{m=0}^{M-1}\gamma^{M-m-1}(P^{\pi_k})^{M-m-1} \\ &\leq \rho\sum_{k=0}^{K-1}\gamma^{K-k-1}(P^{\pi^*})^{K-k-1}F_k(I-\gamma P^{\pi_k})^{-1} \\ &\leq \rho(I-\gamma P^{\pi^*})^{-1}\sum_{l=0}^{K-1}F_l(I-\gamma P^{\pi_l})^{-1} \\ &\leq \frac{K}{(1-\gamma)^2}K_\gamma^2(\rho)\big[K(\rho)\frac{K_\gamma(\rho)}{1-\gamma}(1+\gamma K(\rho))+K(\rho)\big]\rho\end{aligned}$$

Thus the first term in (12) is bounded by

$$\begin{aligned}&\sigma^p\rho\sum_{k=0}^{K-1}\sum_{m=0}^{M-1}\frac{\alpha_k^m}{\sigma}A_k^m \leq \Big[\frac{2\gamma}{(1-\gamma)^3}\Big]^p \\ &\quad KK_\gamma^2(\rho)K(\rho)\big[K_\gamma(\rho)(1+\gamma K(\rho))+1-\gamma\big]\rho/(2\gamma).\end{aligned}$$

This, together with the bound on r_K^M proves (10). ■

C. Proof of the Main Result

Proof: Given Lemma 6 and 7, the proof is a straightforward application of a union bounding argument (with respect to the error events of the individual iterations) and hence its details are omitted. ■

V. Discussion

A. Smoothness Conditions

One of the major conditions of the main result is that $C_i(\rho,\nu)$, $i=2,3$ are finite, or that $K(\rho), K_\gamma(\rho)$ are finite. The purpose of this section is to shed some light on the nature of these conditions.

In fact, notice first that the results continue to hold under the condition $C(\nu) < +\infty$. To see this just note that $C_i(\rho,\nu) < C(\nu)$, $i=2,3$ holds for any distribution ρ (this follows simply since $(\rho P^{\pi_1}P^{\pi_2}\dots P^{\pi_{m-1}})P^{\pi_m} \leq C(\nu)\nu$).

Note that $C(\nu) < \infty$ is satisfied whenever the transition density kernel is absolute continuous w.r.t. ν. This holds for compact state-space MDPs where the dynamics can be put in the state-space form $X_{t+1} = f(X_t,U_t) + W_t(X_t,U_t)$, where f is a measurable function and W_t is a zero-mean random variable that admits a density w.r.t. ν that is uniformly bounded and ν is lower bounded (such as the Lebesgue measure). The density may depend on X_t and U_t, but the bound on it should be independent of X_t, U_t. The "noisier" is the dynamics, the smallest is the smoothness constant $C(\nu)$. Whilst for $C(\nu) < \infty$, W_t must admit a density, ruling out deterministic systems or systems with jumps in the dynamics, $C_i(\rho,\nu) < \infty$ ($i=2,3$) may hold for such systems, see [15]. However even the class of MDPs that assume a bounded density and are Lipschitz continuous is large in a sense that their worst-case complexity is exponential in the dimension of $\mathcal{X}$ [16].

Now let us relate $C_i(\rho,\nu) < \infty$ ($i=2,3$) to the top-Lyapunov exponent of the system. As our starting point we take the definition of top-Lyapunov exponent associated with sequences of finite dimensional matrices: If $\{P_t\}_t$ is sequence of square matrices with non-negative entries and $\{y_t\}_t$ is a sequence of vectors that satisfy $y_{t+1} = P_t y_t$ then, by definition, the top-Lyapunov exponent is $\hat{\gamma}_{\rm top} = \limsup_{t\to\infty}(1/t)\log^+(\|y_t\|_\infty)$. If the top-Lyapunov exponent is positive then the associative system is sensitive to its initial conditions (unstable). A negative Lyapunov exponent, on the other hand, indicates that the system is stable; in case of certain stochastic systems the existence of strictly stationary non-anticipating realizations is equivalent to a negative Lyapunov exponent [17].[3]

[3]The lack of existence of such solutions would probably preclude any sample-based estimation of the system.

Now, one may think of y_t as a probability distribution over the state space and the matrices as the transition probabilities. One way to generalize the above definition to controlled systems and infinite state spaces is to identify y_t with the future state distribution when the policies are selected to maximize the growth rate of $\|y_t\|_\infty$. This gives rise to $\hat{\gamma}_{\rm top} = \limsup_{m\to\infty} \frac{1}{m} \log c(m)$, where $c(m)$ is defined by (4). Then, by elementary arguments, we get that if $\hat{\gamma}_{\rm top} < \log(1/\gamma)$ then $\sum_{m\geq 0} m^p \gamma^m c(m) < \infty$. In fact, if $\hat{\gamma}_{\rm top} \leq 0$ then $C_i(\rho,\nu) < \infty$, $i = 2, 3$. Hence, our conditions on $C_i(\rho,\nu)$ can be interpreted as some stability conditions.

Now let us to turn to the discussion of the assumption on the finiteness of $K(\rho)$ and $K_\gamma(\rho)$. By calling for the Neumann-series expansion of $(I - \gamma P)^{-1}$, it is easy to see that $K(\rho), K_\gamma(\rho) \leq C(\rho)$. We believe that the condition $K(\rho), K_\gamma(\rho) < \infty$ might be easier to verify than the finiteness of $C_i(\rho,\nu)$, $i = 2, 3$. When the Radon-Nikodym derivative $d\nu/d\rho$ exists and is bounded then $K(\rho)$ can be bounded in terms of $c(1)$ and $K_\gamma(\rho)$ can be bounded in terms of $\sum_{m\geq 0} \gamma^m c(m)$. However, in general, these two pairs of conditions need not be related.

B. Related Work

In terms of the tools and techniques used, the closest to the work presented here are our earlier papers [3], [7]. However, whilst in [3] we presented results for sample-based approximate value iteration where a generative model of the MDP was assumed to be available, in this paper we dealt with the significantly more complicated problem of analysing fitted policy iteration applied to a single trajectory. The paper [7] in this sense is closer to the present work. However, there a somewhat complicated Bellman-residual like criterion was considered (that requires the use of an auxiliary function) that might not look as appealing to some as the value-iteration based algorithm considered here, which is a "standard" algorithm. Thus, compared to this previous paper one of the main contribution is the analysis of an algorithm that is closest to the algorithms already in use in practice. However, it remains to be seen which of these algorithms performs better in applications. Also, the analysis presented here is substantially different because of the two nested loops of the algorithm (the price of which is a factor of $1/(1-\gamma)$ in the bound). Furthermore, as compared to [7] here a more detailed analysis of the smoothness constraints on the system is provided and new smoothness conditions are suggested. These might be easier to verify as they do not depend on the unknown distribution ν.

As compared to the conditions used by [3], our conditions on the function class $\mathcal{F}$ are (slightly) more restrictive, as far as the capacity constraints are concerned. The reason is that the previous iterate, Q_{k-1}, influences the next iterates through the greedy policy $\pi_k = \hat{\pi}(\cdot; Q_{k-1})$. Computing the greedy policy involves comparing the action-values and hence, in order to limit the complexity of the resulting policy space, we had to assume some more conditions on the function class $\mathcal{F}$. We believe that the constraints on the function class are satisfied by many popular function classes (e.g., regression trees, neural networks, etc.), but this remains to be proven. However, finite dimensional linear function spaces with truncation do satisfy the requirements of our results. It is also an open question of finiteness of the VC-crossing dimension is necessary for the stable behaviour of the algorithm.

VI. Conclusions

We have considered approximate policy iteration with trajectory based approximate value iteration. Our results show that the number of samples needed to achieve a small approximation error depends polynomially on the capacity of the function class used in the empirical loss minimization step and the smoothness of the dynamics of the system. One strength of our results is that they quantify the bias-variance tradeoff in RL. One of the most important open questions is how to resolve this tradeoff in an optimal fashion.

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

[3] Cs. Szepesvári and R. Munos. Finite time bounds for sampling based fitted value iteration. In *ICML'2005*, pages 881–886, 2005.

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

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

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

[7] A. Antos, Cs. Szepesvári, and R. Munos. Learning near-optimal policies with Bellman-residual minimization based fitted policy iteration and a single sample path. In *COLT-19*, pages 574–588, 2006.

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

[9] A. Antos, Cs. Szepesvári, and R. Munos. Approximate action-value iteration in continuous state spaces: learning with a single trajectory. (submitted), 2006.

[10] E.W. Cheney. *Introduction to approximation theory*. McGraw-Hill, London, New York, 1966.

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

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

[13] 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.

[14] R. Munos. Error bounds for approximate policy iteration. In *ICML'2003*, pages 560–567, 2003.

[15] R. Munos and Cs. Szepesvári. Finite time bounds for sampling based fitted value iteration. *Journal of Machine Learning Research*, 2005. submitted.

[16] C.S. Chow and J.N. Tsitsiklis. An optimal multigrid algorithm for continuous state discrete time stochastic control. *IEEE Transactions on Automatic Control*, 36(8):898–914, 1991.

[17] P. Bougerol and N. Picard. Strict stationarity of generalized autoregressive processes. *Annals of Probability*, 20:1714–1730, 1992.