# Finite Time Bounds for Sampling Based Fitted Value Iteration

**Csaba Szepesvári** SZCSABA@SZTAKI.HU
Computer and Automation Research Institute of the Hungarian Academy of Sciences, Kende u. 13-17, Budapest 1111, Hungary

**Rémi Munos** REMI.MUNOS@POLYTECHNIQUE.FR
Centre de Mathématiques Appliquées, Ecole Polytechnique, 91128 Palaiseau Cedex, France

## Abstract

In this paper we consider sampling based fitted value iteration for discounted, large (possibly infinite) state-space, finite action Markovian Decision Problems where only a generative model of the transition probabilities and rewards is available. At each step the image of the current estimate of the optimal value function under a Monte-Carlo approximation to the Bellman-operator is projected onto some function space. PAC-style bounds on the weighted $L^p$-norm approximation error are obtained as a function of the covering number, the approximation power of the function space, the iteration number and the sample size.

## 1. Introduction

In this paper we consider *fitted value iteration* (FVI) for solving expected total discounted reward, large state-space, finite action Markovian Decision Problems (MDP) under the assumption that the model is unknown, but a *generative* model of the MDP is available. Value iteration is the process of computing an approximation of the optimal value function by means of the iteration $V_{k+1} = TV_k$, where $T$ is the so-called Bellman-operator. FVI is an extension of value iteration that allows it to work in infinite or very large state spaces. FVI generates a series of value functions $V_0, V_1, \ldots, V_k, \ldots$ such that $V_{k+1}$ is obtained by projecting $TV_k$ (or an approximation of it) onto a space of functions, $\mathcal{F}$. FVI is a special form of *approximate value iteration* (AVI), which is a generic scheme where the iterates are given by $V_{k+1} = TV_k + \epsilon_k$. Typical results relate the asymptotic error of approximating the optimal value function in terms of the properties of error series $\{\epsilon_k\}$. When some bound on the error series impose a bound on the asymptotic approximation error then the iteration is called *stable*.

The origins of AVI date back to the early days of dynamic programming, e.g. (Samuel, 1959; Bellman & Dreyfus, 1959). Recent theoretical results concern supremum-norm approximation errors (Gordon, 1995; Tsitsiklis & Van Roy, 1996). The main underlying insight in these analysis is that if $\gamma \in (0, 1)$ is the discount factor of the MDP and if $A$ represents the operator that maps value functions to the space of functions of interest and if $A$ is $\gamma'$-Lipschitz w.r.t. the supremum-norm then the composite operator $AT$ is a contraction provided that $\gamma\gamma' < 1$ since in discounted problems $T$ is a contraction with contraction factor $\gamma$. Thus the iterates $V_{k+1} = ATV_k$ are guaranteed to converge.

Both of the above papers assume that the controlled system is known. In Section 7.2 of (Tsitsiklis & Van Roy, 1996) it is mentioned that the AVI can be extended to use Monte-Carlo approximation, but no theoretical analysis is presented there. In (Singh et al., 1995) the convergence of Q-learning is considered with "soft state-aggregation" (i.e. here the model is not assumed to be known), whilst in (Szepesvári & Smart, 2004) a more efficient, Rao-Blackwellised version of this algorithm was proposed and proven to yield convergent estimates.

The above results all concern approximations in the supremum-norm. However, it is both unrealistic and unnecessary to require good uniform approximation over the whole state space. In this article, we consider bounds on the approximation error in terms of weighted $L^p(\mu)$ norms, with $\|f\|_{p,\mu} = (\int |f(x)|^p d\mu)^{1/p}$, where $\mu$ is a probability distribution over the state space $\mathcal{X}$ and $p \geq 1$.

Appearing in *Proceedings of the 22$^{st}$ International Conference on Machine Learning*, Bonn, Germany, 2005. 

Previous work on weighted-norm stability analysis includes a stability result for linear function approximation and weighted Euclidean norms (defined over the finite dimensional parameter space) that is presented in Section 6.8 of (Bertsekas & Tsitsiklis, 1996b).

More recently, stability results in $L^2$ norm were derived for approximate policy iteration (Munos, 2003) and AVI (Munos, 2005), assuming the knowledge of the MDP. In this paper we extend these results to the case when the model of the MDP is not given explicitly, but only a simulation device is available. Further, we obtain bounds for weighted $L^p$-norms for any fixed $p \geq 1$. Results of Statistical Learning Theory are used to relate the complexity of the function space and the sample size required to obtain a randomized policy whose value function is within a specified tolerance $\epsilon$ of the optimal value function with high probability.

## 2. Outline of the Algorithm, Results

*Sampling based fitted value iteration* proceeds as follows: Let $V_k \in \mathcal{F}$ be the approximation of the optimal value function at stage $k$. A Monte-Carlo estimate of the image of $V_k$ under the Bellman-operator underlying the MDP is computed at selected random points:

$$\hat{V}(X_i) = \max_{a \in \mathcal{A}} \frac{1}{M} \sum_{j=1}^{M} \left[ R_j^{X_i,a} + \gamma V_k(Y_j^{X_i,a}) \right].$$

Here $i = 1, 2, \ldots, N$, $X_1, \ldots, X_N$ are sampled from some distribution $\mu$ defined over the state space $\mathcal{X}$, for each of these states $\{X_i\}_{1 \leq i \leq N}$ and for each possible action $a \in \mathcal{A}$, $Y_j^{X_i,a} \in \mathcal{X}$, and $R_j^{X_i,a} \in \mathbb{R}$, $j = 1, 2, \ldots, M$, are drawn using the generative model of the MDP. The next iterate $V_{k+1}$ is obtained as the best fit in $\mathcal{F}$ to the data $(X_i, \hat{V}(X_i)$, $i = 1, 2, \ldots, N$, in the sense of minimizing the empirical error

$$V_{k+1} = \operatorname*{argmin}_{f \in \mathcal{F}} \sum_{i=1}^{N} (f(X_i) - \hat{V}(X_i))^p. \tag{2.1}$$

This iteration is repeated $K$ times. Our main result is that under suitable conditions for large enough values of $N$, $M$, and $K$, the performance $V^{\pi_K}$ of the policy $\pi_K$ induced by the approximation $V_K$ is close to the optimal value function $V^*$ with high probability, i.e. for any $\delta > 0, \epsilon > 0$, there exists $K = O(p \log(\hat{R}_{\max}/(\epsilon(1-\gamma))))$, $N = O(\hat{R}_{\max}^2 (1/\epsilon)^p \log(\mathcal{N}/\delta))$, $M = O(\hat{R}_{\max}^2/\epsilon^2 \log(N|\mathcal{A}|/\delta))$ such that,

$$\mathbb{P}\left( \|V^* - V^{\pi_K}\|_{p,\mu} > \epsilon \right) \leq \delta,$$

where $\mathcal{N}$ is an appropriate covering number and $\hat{R}_{\max}$ is a bound on the (random) immediate rewards. We also derive a similar result on the performance of a randomized approximation of $\pi_K$.

We consider two variants of this algorithm: In the *single-sample* variant a single sample $X_i$, $Y_j^{X_i,a}$, $R_j^{X_i,a}$, $i = 1, \ldots, N$, $j = 1, \ldots, M$ is drawn during the initialization of the algorithm and is used in all the $K$ iterations. In the *multi-sample* variant a new sample is drawn in each of the $K$ iterations, independently of the previous samples.

## 3. Preliminaries

Due to the lack of space we cannot give a rigourous definition of MDPs, but introduce only the necessary notation. Readers not familiar with MDPs are referred to (Bertsekas & Tsitsiklis, 1996b). We consider MDPs with a measurable state-space $\mathcal{X}$ and a finite action space $\mathcal{A}$. The state-space can be finite, countable or uncountable. For simplicity we shall assume that $\mathcal{X}$ is a measurable subset of $\mathbb{R}^d$ for some $d > 0$. We assume that there exists a *sampling device* that can generate the MDP's transitions and rewards for any state-action pair $(x, a) \in \mathcal{X} \times \mathcal{A}$. We shall denote by $P(dy|x, a)$ $(S(dr|x, a))$ the transition probability kernel (resp., reward distribution).

Given an MDP, the goal is to find a (deterministic) stationary policy $\pi$ that maximizes the long-term total expected discounted reward given any initial state. The discount factor is denoted by $\gamma$ $(0 < \gamma < 1)$. The optimal long-term total expected discounted reward when the process is started from state $x$ shall be denoted by $V^*(x)$, and $V^*$ is called the optimal value function. A policy is called optimal, if it attains the optimal values $V^*(x)$ for *any* state $x \in \mathcal{X}$.

Let us denote the space of bounded measurable functions with domain $\mathcal{X}$ by $B(\mathcal{X})$. Further, the space of measurable functions bounded by $V_{\max} < +\infty$ shall be denoted by $B(\mathcal{X}; V_{\max})$.

A deterministic stationary policy $\pi : \mathcal{X} \to \mathcal{A}$ gives rise to the transition probability kernel $P^\pi(dy|x) = P(dy|x, \pi(x))$, from which two related operators are derived: a right linear operator $P^\pi \cdot : B(\mathcal{X}) \to B(\mathcal{X})$, defined for any bounded function $V \in B(\mathcal{X})$ by

$$(P^\pi V)(x) = \int V(y) P^\pi(dy|x),$$

and a left linear operator $\cdot P^\pi : M(\mathcal{X}) \to M(\mathcal{X})$, where $M(\mathcal{X})$ is the space of all probability distributions over $\mathcal{X}$, defined by

$$(\mu P^\pi)(dy) = \int P^\pi(dy|x) \mu(dx)$$

for $\mu \in M(\mathcal{X})$. The product of two kernels $P^{\pi_1}$ and $P^{\pi_2}$ is defined by

$$(P^{\pi_1} P^{\pi_2})(dz|x) = \int P^{\pi_1}(dy|x) P^{\pi_2}(dz|y).$$

We say that a (deterministic stationary) policy $\pi$ is *greedy* w.r.t. a function $V \in B(\mathcal{X})$ if, for all $x \in \mathcal{X}$,

$$\pi(x) \in \arg\max_{a\in\mathcal{A}}\{r(x,a) + \gamma \int V(y)P(dy|x,a)\}.$$

where $r(x,a) = \int zS(dz|x,a)$ is the expected reward of executing action $a$ in state $x$, and is assumed to be a bounded measurable function.

For any function $V$, such a greedy policy always exists because the maximum value in the previous equation is always reached, since $\mathcal{A}$ is finite.

## 4. Approximating the Bellman-operator

Define the Bellman-operator $T : B(\mathcal{X}) \to B(\mathcal{X})$ by

$$(TV)(x) = \max_{a\in\mathcal{A}}\{r(x,a) + \gamma \int V(y)dP(y|x,a)\}.$$

where $V : \mathcal{X} \to \mathbb{R}$ is an arbitrary, bounded measurable function. It is well known that the unique fixed point of $T$ is the optimal value function, $V^*$ (Bertsekas & Shreve, 1978). Further, the value iteration algorithm, $V_{k+1} = TV_k$, with an arbitrary $V_0 \in B(\mathcal{X})$ yields a sequence of functions $V_k$ that converges to $V^*$. Note that if $r(x,a)$ is bounded by $R_{\max} > 0$ then $V^*$ is bounded by $R_{\max}/(1-\gamma)$ and if $V_0 \in B(\mathcal{X}; R_{\max}/(1-\gamma))$ then also $V_k \in B(\mathcal{X}; R_{\max}/(1-\gamma))$.

Let us now fix an appropriate space of bounded measurable functions, $\mathcal{F}$. At this level of generality $\mathcal{F}$ could be any subset of $B(\mathcal{X})$. Typical choices would be e.g. a parameterized class of functions:

$$\mathcal{F} = \{\, f_\theta \in B(\mathcal{X}) \,|\, \theta \in \Theta \,\}$$

with either a linear ($f_\theta(x) = \theta^T\phi(x)$) or non-linear parameterization ($f_\theta(x) = f(x;\theta)$), such as in the case of neural-networks.

FVI works by projecting the iterates $V_k$ onto the space $\mathcal{F}$. In order to make this specific, let us pick any distribution $\mu$ over the state-space $\mathcal{X}$ and fix $p \ge 1$. Assuming for a moment that the *metric projection* of $TV$ onto $\mathcal{F}$ w.r.t. the $\mu$-weighted $L^p$-norm,

$$\Pi_{\mathcal{F}}TV = \operatorname*{argmin}_{f\in\mathcal{F}} \|f - TV\|_{p,\mu},$$

exists, FVI can be described as the algorithm that generates the iterates $V_{k+1} = \Pi_{\mathcal{F}}TV_k$.

FVI per se is not a practical algorithm when $\mathcal{X}$ is infinite. This is because $\Pi_{\mathcal{F}}TV_k$ cannot be computed analytically except in a few special cases. Besides, notice that computing $(TV)(x)$ itself involves an integration over the infinite state-space $\mathcal{X}$.

Let us now consider the approximate computation of $\Pi_{\mathcal{F}}TV$ by Monte-Carlo integration. As in the previous section, let $X_1, \ldots, X_N$ be an i.i.d. sample drawn from the distribution $\mu$. For each action $a \in \mathcal{A}$, $i$, $1 \le i \le N$ let $\{(R_j^{X_i,a}, Y_j^{X_i,a})\}$ be an i.i.d. sample of $M$ pairs, where $Y_j^{X_i,a} \sim P(dy|X_i,a)$, and $R_j^{X_i,a} \sim S(dr|X_i,a)$. Pick any function $V \in B(\mathcal{X}; V_{\max})$ and define the $N$ values

$$\hat{V}(X_i) = \max_{a\in\mathcal{A}} \frac{1}{M}\sum_{j=1}^{M}\left[R_j^{X_i,a} + \gamma V(Y_j^{X_i,a})\right],$$

$i = 1, 2, \ldots, N$. Let

$$V' = \operatorname*{argmin}_{f\in\mathcal{F}} \sum_{i=1}^{N}\left[f(X_i) - \hat{V}(X_i)\right]^p. \qquad (4.1)$$

For the sake of simplicity, we shall assume that the minimizer in Equation (4.1) exists.[1] This assumption simplifies the proofs, but it is by no means essential for our results. When the dependence of $V'$ on $V$ is important, we will write $\Pi_{\mathcal{F}}TV$ to denote $V'$ as defined above.

Central to our proofs will be Pollard's inequality that gives conditions under which, for i.i.d. samples $\{X_i\}_{i=1,\ldots,N}$, the sample averages $1/N\sum_{i=1}^N f(X_i)$ approximate the expectation $E[f(X_1)]$ *uniformly* over the space of functions $\mathcal{F}$. If $\mathcal{F}$ is finite, such a result follows by union bounds and exponential inequalities. When $\mathcal{F}$ is infinite, an appropriate and well-known measure of the 'size' of $\mathcal{F}$ is given by the *covering number* of $\mathcal{F}$: Let $x^{1:N} \stackrel{\text{def}}{=} (x_1, \ldots, x_N) \in \mathcal{X}^N$ be fixed. Fix $\epsilon > 0$, $q \ge 1$. The $(\epsilon, q)$-covering number (or covering number) of $\mathcal{F}(x^{1:N}) = \{(f(x_1), \ldots, f(x_N)) \,|\, f \in \mathcal{F}\}$ is the smallest integer $m$ such that $\mathcal{F}(x^{1:N})$ can be covered by $m$ balls of the normed-space $(\mathbb{R}^N, \|\cdot\|_q)$ with centers in $\mathcal{F}(x^{1:N})$ and radius $N^{1/q}\epsilon$. The $(\epsilon, q)$-covering number of $\mathcal{F}(x^{1:N})$ shall be denoted by $\mathcal{N}_q(\epsilon, \mathcal{F}(x^{1:N}))$. When $q = 1$, we use $\mathcal{N}$ instead of $\mathcal{N}_1$. When $X^{1:N}$ are i.i.d. with common underlying distribution $\mu$ then $\mathbb{E}\left[\mathcal{N}_q(\epsilon, \mathcal{F}(X^{1:N}))\right]$ shall be denoted by $\mathcal{N}_q(\epsilon, \mathcal{F}, N, \mu)$.

In what follows we shall make the following regularity assumptions regarding the MDP:

**Assumption A0** (Regularity) The MDP $(\mathcal{X}, \mathcal{A}, P, S, \gamma)$ satisfies the following conditions: $\mathcal{X}$ is a measurable subset of some Euclidean space, $\mathcal{A}$ is finite, the discount factor $\gamma$ satisfies $0 < \gamma < 1$. The reward kernel $S$ is such that the immediate reward function $r$ is a bounded, measurable function, with bound $R_{\max}$. Further, the support of

[1] This holds e.g. if $\mathcal{F}$ is closed. Note that for many popular choices, such e.g. for neural-networks with continuous transfer functions, $\mathcal{F}$ is not closed.

$S(\cdot|x,a)$ is included in $[-\hat{R}_{\max}, \hat{R}_{\max}]$ independently of $(x,a) \in \mathcal{X} \times \mathcal{A}$.

The following result holds true:

**Lemma 4.1.** *Fix $p \geq 1$, $\mu \in M(\mathcal{X})$. Let Assumption A0 hold and let $V_{\max} = R_{\max}/(1-\gamma)$, $V \in B(\mathcal{X}; V_{\max})$ and let $V'$ be defined as in Equation (4.1). Assume that $\mathcal{F} \subset B(\mathcal{X}; V_{\max})$. Fix $\epsilon, \delta > 0$ and assume that $\mathcal{F}$ is such that the error of approximating $TV$ by $\mathcal{F}$ is not larger than $\epsilon/5$:*

$$E_p(TV; \mathcal{F}) \stackrel{\text{def}}{=} \inf_{f \in \mathcal{F}} \|f - TV\|_{p,\mu} \leq \epsilon/5. \tag{4.2}$$

*If $N = O(V_{\max}^2 (1/\epsilon)^{2p} \log(\mathcal{N}(c\epsilon, \mathcal{F}, N, \mu)/\delta))$ and $M = O((\hat{R}_{\max} + \gamma V_{\max})^2/\epsilon^2 \log(N|\mathcal{A}|/\delta)$ then $\mathbb{P}\left(\|V' - TV\|_{p,\mu} > \epsilon\right) \leq \delta$. Here $c > 0$ is a constant independent of the parameters of the MDP and the function space.*

We remark that $N$ appears on both sides of its defining equation. For specific choices of $\mathcal{F}$ (such as e.g. when $\mathcal{F}$ is linearly parameterized with a bounded parameter space, or when $\mathcal{F}$ is the space of appropriately restricted neural networks (Anthony & Bartlett, 1999)) the covering number $\mathcal{N}$ will not depend on $N$, but $\log \mathcal{N}$ will depend on the dimensionality $d$ of the state space $\mathcal{X}$ (the dependence is typically of the order $O(\log d)$, $O(d)$ or $O(d \log d)$, cf. e.g. (Zhang, 2002; Anthony & Bartlett, 1999)). Hence follows the existence of a lower bound that is polynomial in $1/\epsilon$, $\log(1/\delta)$, $V_{\max}$, $\hat{R}_{\max}$, and $\log(|\mathcal{A}|)$.[2]

In the above result $V$ is assumed to be a deterministic function. However, the result continues to hold even if $V$ is random such that $V(\omega) \in \mathcal{F}$ holds for all $\omega \in \Omega$ (here $\Omega$ is the sample space). This extension is needed for the analysis of the single-sample variant of the algorithm. In order to appreciate the difference between these two variants consider bounding the probability of the error of the $k$th iteration of the multi-sample variant first. In this case, due to the independence of samples used in subsequent iterations, Lemma 4.1 can be used to bound $\mathbb{P}\left(\|V_{k+1} - TV_k\|_{p,\mu} > \epsilon | D_k\right)$, where $D_k$ denotes the samples used to obtain the $k$th iterate. Since $\mathbb{P}\left(\|V_{k+1} - TV_k\|_{p,\mu} > \epsilon\right) = \mathbb{E}\left[\mathbb{P}\left(\|V_{k+1} - TV_k\|_{p,\mu} > \epsilon | D_k\right)\right]$ this yields directly a bound on the probability of the error due to sampling, introduced in the $k$th iteration. In the single-sample variant the above argument does not work since the same single sample-set is used throughout all the iterations. Hence, for analyzing the behavior of the single-sample variant we need a new result. Define $\mathcal{F}_{T-} = \{\, f - Tg \,|\, f \in \mathcal{F}, g \in \mathcal{F} \,\}$. The result that we need is as follows:

**Lemma 4.2.** *The result of the previous lemma continues to hold if $V$ is a random function provided that $N = O(V_{\max}^2 (1/\epsilon)^{2p} \log(\mathcal{N}(c\epsilon, \mathcal{F}_{T-}, N, \mu)/\delta))$ and $M = O((\hat{R}_{\max} + \gamma V_{\max})^2/\epsilon^2 \log(8N|\mathcal{A}|\mathcal{N}(c'\epsilon, \mathcal{F}, M, \mu)/\delta)$, where $c, c' > 0$ are constants independent of the parameters of the MDP and the function space $\mathcal{F}$.*

An example when the covering numbers corresponding to the space $\mathcal{F}_{T-}$ can be bounded is when $\mathcal{X}$ is compact, $\mathcal{F} = \{f_\theta | \theta \in \Theta\}$, $\Theta$ is compact and the mapping $H : \theta \mapsto f_\theta$ is Lipschitz with coefficient $L$ when viewed as a mapping between the normed spaces $(\Theta, \|\cdot\|)$ and $(B(\mathcal{X}), L^\infty)$. In this case a standard argument allows us to bound the covering numbers of $\mathcal{F}_{T-}$ in terms of the covering numbers of $\mathcal{F}$.

Note that many popular choices of function approximators meet the requirement that $H$ is Lipschitz. Consider for example linear function approximators taking the form $f_\theta = \theta^T \phi$ with a suitable basis function $\phi : \mathcal{X} \to \mathbb{R}^d$. Straightforward calculations yield that $\left\|\theta_1^T \phi - \theta_2^T \phi\right\|_\infty = \sup_{x \in \mathcal{X}} |\langle \theta_1 - \theta_2, \phi(x) \rangle| \leq \|\theta_1 - \theta_2\|_2 \sup_{x \in \mathcal{X}} \|\phi(x)\|_2$. Hence, by choosing the $\ell^2$ norm in the space $\Theta$, we get that $\theta \mapsto \theta^T \phi$ is Lipschitz with coefficient $\|\|\phi(\cdot)\|_2\|_\infty$.

## 5. Main Results

Assume that $V_0 \in \mathcal{F}$ and let $\pi_k$ be a greedy policy w.r.t. the approximate $V_k$, where $V_k$ is generated by the sampled FVI algorithm. The next result makes use of one of the following assumptions:

**Assumption A1** [Uniformly stochastic transitions]. There exists a constant $C > 0$ such that, for all $x \in \mathcal{X}$, and all policy $\pi$,

$$P^\pi(\cdot|x) \leq C\mu(\cdot).$$

**Assumption A2** [Smooth future state distribution]. There exists a distribution $\rho$ and coefficients $\{c(m)\}_{m \geq 1}$ such that for all $m$ policies $\pi_1, \ldots, \pi_m$,

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

and the series $\sum_{m \geq 1} m\gamma^{m-1} c(m)$ converges. Under Assumption A2 the constant $C$ is redefined as follows:

$$C = (1-\gamma)^2 \sum_{m \geq 1} m\gamma^{m-1} c(m).$$

[2] There are cases when $\mathcal{N}$ can be bounded independently of $d$ and with a linear dependence on $N$, see e.g. (Zhang, 2002). The above conclusion still remains valid in this case.

Assumption A1 was introduced in (Munos, 2003) in finite state spaces for approximate policy iteration. In the Euclidean space considered here, assuming that the transition probability kernel $P(dy|x,a)$ admits a density representation $p(y|x,a)$ w.r.t. the Borel measure, Assumption A1 holds for example if the density $p(\cdot|\cdot,\cdot)$ is uniformly bounded by some constant $\beta$, under $\mu$ being the Borel measure. Such an example is illustrated in the numerical experiment below.

Assumption A2 is weaker than A1. A1 requires that the transitions be somehow uniformly stochastic. Thus, a pure deterministic MDP would not satisfy A1. However, such a deterministic MDP may satisfy A2 if the future state transition distribution is smooth compared to the initial distribution. Indeed, A2 implies that for any sequence of policies $\pi_1,\ldots,\pi_m$, the discounted future state distribution starting from $\rho$ is bounded by a constant $C$ times $\mu$: $(1-\gamma)^2\sum_{m\geq 1} m\gamma^{m-1}\mathbb{P}(X_m \in B|X_0\sim\rho,\, X_i \sim P(\cdot|X_{i-1},\pi_i(X_{i-1})),\ 1\leq i\leq m, X_0\sim\rho) \leq C\mu(B)$.

An example of MDP for which A2 holds but not A1 is the chain walk MDP described in (Munos, 2005).

These assumptions enable us to bound the performance error $V^*-V^{\pi_K}$ by a function of the $L^p(\mu)$ norm of the approximation errors $V_{k+1}-TV_k$ ($k\leq K$) due to sampling and approximation in the space $\mathcal{F}$:

**Theorem 5.1.** *Fix $p\geq 1$ and $\mu\in M(\mathcal{X})$, $V_0\in\mathcal{F}\subset B(\mathcal{X};V_{\max})$. Let the regularity Assumption A0 holds and also let Assumption A1 holds. Let $\epsilon>0,\delta>0$ be arbitrary. Assume that $\mathcal{F}$ is such that the worst-case approximation error of functions $\{TV\,|\,V\in\mathcal{F}\}$ satisfies*

$$\sup_{V\in\mathcal{F}} E_p(TV;\mathcal{F}) \leq \frac{(1-\gamma)^2\epsilon}{4C}. \tag{5.1}$$

*Then there exist integers $K,M$ and $N$ such that $K=O(\log(V_{\max}/(\epsilon(1-\gamma)^2)/\log(1/\gamma))$, $N$, $M$ are polynomial in $1/\epsilon$, $\log(1/\delta)$, $\log(1/(1-\gamma))$, $V_{\max}$, $\hat{R}_{\max}$, $\log(|\mathcal{A}|)$, $\log(\mathcal{N}(c\epsilon(1-\gamma)^2/C),\mathcal{F},\mu)))$ for some constant $c>0$, such that,*

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

*Now, let Assumptions A0, A2, and (5.1) hold with $C$ replaced by $C^{1/p}$. Then there exist integers $K,M$ and $N$ such that $K=O(\log(V_{\max}/(\epsilon(1-\gamma)^2)/\log(1/\gamma))$, $N$, $M$ are polynomials in $1/\epsilon$, $\log(1/\delta)$, $\log(1/(1-\gamma))$, $V_{\max}$, $\hat{R}_{\max}$, $\log(|\mathcal{A}|)$, $\log(\mathcal{N}(c\epsilon(1-\gamma)^2/C^{1/p}),\mathcal{F},\mu)))$ for some constant $c>0$, such that,*

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

*The results hold for both the single-sample and the multi-sample variants.*

### 5.1. Approximation Power

Let $\mathcal{X}$ be a compact subset of a $d$-dimensional Euclidean space and consider the problem of approximating functions bounded by some constant $V_{\max}>0$. Let $\{\mathcal{F}_n\}$ be a series of function spaces such that the 'complexity' of $\mathcal{F}_n$ increases with $n$. Typical examples would be classes where $n$ is proportional to the number of parameters in a parameterized class of functions. Consider the space of $V_{\max}$-bounded, Lipschitz-continuous functions $\cup_{L>0}\mathrm{Lip}(1;L) = \cup_{L>0}\{f\in B(\mathcal{X})|\,\|f\|_\infty\leq V_{\max}, |f(x)-f(y)|\leq L\|x-y\|\,\}$. For classical approximation classes, Jackson's theorem can be used to give lower bounds $n_0$ on the index $n$ such that $E_{n,\mu,q}(L) \stackrel{\text{def}}{=} \sup_{g\in\mathrm{Lip}(1;L)}\inf_{f\in\mathcal{F}_n}\|f-g\|_{p,\mu}\leq\epsilon$ holds whenever $n\geq n_0$.[3] We shall call an approximation class $\{\mathcal{F}_n\}$ universal if for any $L$ there exists an index $n_0$ such that for $n\geq n_0$ $E_{n,\mu,q}(L)\leq\epsilon$.

As an immediate corollary of Theorem 5.1 we have the following result:

**Corollary 5.2.** *Fix $p\geq 1$, $\mu\in M(\mathcal{X})$. Let $\mathcal{X}$ be a compact subset of a Euclidean space and consider an MDP satisfying Assumption A0. Further, let Assumption A1 hold. Fix $\epsilon>0$, $\delta>0$. Let $\{\mathcal{F}_n\}$ be a universal approximation class such that the covering numbers $\mathcal{N}(\epsilon,\mathcal{F}_n,N,\mu)$ are bounded and the dependency of these covering numbers on $N$ is $o(N)$. Then there exist an index $n_0$ such that if $n\geq n_0$ then there exist integers $K,N,M$ that are polynomial in $1/\epsilon$, $\log(1/\delta)$, $1/(1-\gamma)$, $V_{\max}$, $\hat{R}_{\max}$, $\log(|\mathcal{A}|)$, $\log(\mathcal{N}(c\epsilon(1-\gamma)^2/C^{1/p}),\mathcal{F}_n,\mu)))$ for some $c>0$, such that $\mathbb{P}\left(\|V^*-V^{\pi_K}\|>\epsilon\right)\leq\delta$. The results hold for both the single-sample and the multi-sample variants.*

### 5.2. Randomized Policies

Given $V_K$ and a state $x\in\mathcal{X}$ we can use sampling to determine an almost greedy action with high probability. In particular, let $\pi^K_{\alpha,\lambda}$ be the policy computed as follows: Draw $M'$ samples, $(R^{x,a}_j, Y^{x,a}_j)$ of reward-next-state pairs ($R^{x,a}_j\sim S(\cdot,x,a)$, $Y^{x,a}_j\sim P(\cdot|x,a)$) and compute

$$Q_{M'}(x,a) = \frac{1}{M'}\sum_{j=1}^{M'}\left\{R^{x,a}_j + \gamma V_K(Y^{x,a}_j)\right\}$$

Let $\pi^K_{\alpha,\lambda}$ select $\arg\max_{a\in\mathcal{A}} Q_{M'}(x,a)$. Call an action $a$ $\alpha$-greedy w.r.t. the function $V$, if $r(x,a)+\gamma\int V(y)P(dy|x,a)\geq(TV)(x)-\alpha$. The following result holds:

**Theorem 5.3.** *Let $M'$ be $O(|\mathcal{A}|R^2_{\max}/\alpha^2\log(|\mathcal{A}|/\lambda))$. If $p,\mu,\mathcal{F}$ are such as in Theorem 5.1 then under Assumptions A0 and A1, if $\alpha=(1-\gamma)2^{1/P-1}\epsilon/4$,*

[3] A typical results is that $n_0 = O(\epsilon^{-1})$.

*$\lambda = \epsilon(1-\gamma)/(4V_{\max})$, then there exist integers $K, N, M$ that are polynomial in $1/\epsilon$, $\log(1/\delta)$, $1/(1-\gamma)$, $V_{\max}$, $\hat{R}_{\max}$, $\log(|\mathcal{A}|)$, $\log(\mathcal{N}(c\epsilon(1-\gamma)^2/C^{1/p}), \mathcal{F}_n, \mu)))$ for some $c > 0$, such that*

$$\mathbb{P}\left(\left\|V^* - V^{\pi^K_{\alpha,\lambda}}\right\|_\infty > \epsilon\right) \leq \delta,$$

*where $\pi^K_{\alpha,\lambda}$ is a policy that selects $\alpha$-greedy actions w.r.t. the function $V_K$ with probability at least $\lambda$.*

*An analogous result in $L_{p,\mu}$ norm holds under Assumptions A0 and A2.*

Note that there exists better algorithms than the above described 'naive' algorithm for computing nearly greedy actions with high probability, e.g. the Median Elimination Algorithm of (E. Even-Dar & Mansour, 2002) would only need $M' = O(|\mathcal{A}|R^2_{\max}/\alpha^2 \log(1/\lambda))$ samples to achieve the same result.

We remark that using an entirely analogous reasoning, it is possible to extend Theorem 5.2 so that $\pi^K$ is replaced by $\pi^K_{\alpha,\lambda}$ with suitable chosen $\alpha$, $\lambda$. Due to space constraints we do not give this result here.

## 6. Numerical Experiment

We now illustrate the sampling based FVI algorithm on a simple one-dimensional optimal replacement problem, described e.g. in (Rust, 1996). The state variable $x_t \in \mathbb{R}_+$ measures the accumulated utilization (such as the odometer reading on a car) of a durable. By convention, $x_t = 0$ denotes a brand new durable. At each discrete time step $t$, there are two possible decisions: either keep ($a_t = K$) or replace ($a_t = R$) the durable. This latter action implies an additional cost $S$ of selling the existing durable and replacing it for a new one. The transition to a new state occurs with the following exponential densities: $p(y|x, a = K) = e^{-\beta(y-x)}\mathbb{I}(y \geq x)$ and $p(y|x, a = R) = e^{-\beta y}, \mathbb{I}(y \geq 0)$. The reward function (opposite of a cost) is $r(x, a = K) = -c(x)$ (we assume that $c(x)$ is an increasing function) and $r(x, a = R) = -C - c(0)$. The optimal value function solves the Bellman Optimality Equation. From this it is possible to derive an analytical expression of the optimal value function: $V^*(x) = \int_x^{\overline{x}} \frac{c'(y)}{1-\gamma}(1 - \gamma e^{-\beta(1-\gamma)(y-x)})dy - \frac{c(\overline{x})}{1-\gamma}, \mathbb{I}(x \leq \overline{x}) + \frac{-c(\overline{x})}{1-\gamma}\mathbb{I}(x > \overline{x})$, where $\overline{x}$ is the unique solution to

$$C = \int_0^{\overline{x}} \frac{c'(y)}{1-\gamma}(1 - \gamma e^{-\beta(1-\gamma)y})dy.$$

The optimal policy is $\pi^*(x) = K$ if $x \in [0, \overline{x}]$, and $\pi^*(x) = R$ if $x > \overline{x}$.

**Numerical Results**

We chose the numerical values $\gamma = 0.6$, $\beta = 0.5$, $C = 30$, $c(x) = 4x$. Thus, here $\overline{x} \simeq 4.8665$ and the optimal value function (see Figure 1) is

$$V^*(x) = \begin{cases} -10x + 30(e^{0.2(x-\overline{x})} - 1), & \text{if} \quad x \leq \overline{x}; \\ -10\overline{x}, & \text{if} \quad x > \overline{x}. \end{cases}$$

We consider linear approximation of the value function using polynomials of degree $L$. In order the theory to work we would need to put a bound on the weights (since otherwise the covering numbers become unbounded). However, for the sake of simplicity we did not impose any such bounds in these experiments.

In the numerical study, we modify the problem so that its state space becomes $[0, x_{\max}]$, with $x_{\max} = 10$: if the next state $y$ happens to be outside of the domain (i.e. $y > x_{\max}$) then the durable is replaced immediately, and a new state is drawn accordingly to the choice of action $R$. By the choice of $x_{\max}$, $\int_{x_{\max}}^\infty p(dy|x, R)$ is negligible and $x_{\max} > \overline{x}$. Hence the optimal value function of the altered problem closely matches that of the original problem over $[0, x_{\max}]$.

We chose the distribution $\mu$ to be uniform over the state-space $[0, x_{\max}]$. The transition density functions $p(\cdot|x, a)$ are bounded by $\beta$, thus Assumption A1 (hence also Assumption A2) holds with $C = \beta x_{\max} = 6$.

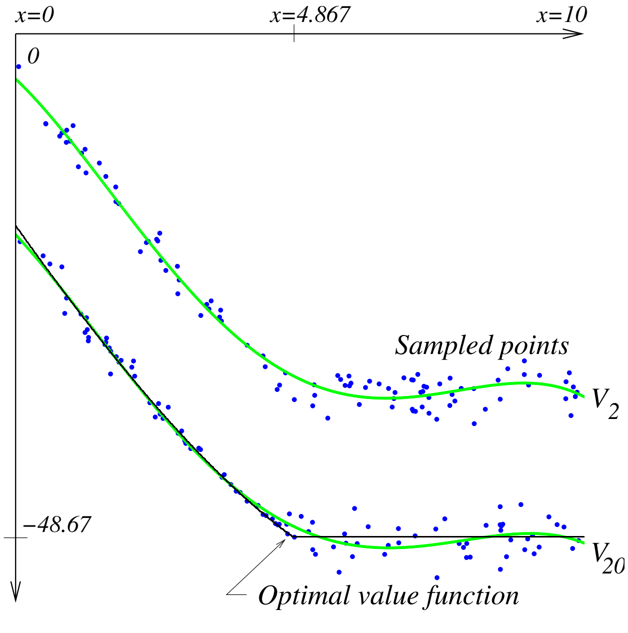


*Figure 1.* Illustration of Sampling based FVI at two iterations steps (up: $k = 2$, down: $k = 20$). The dots represent the $N = 100$ sampled points and their values (averaged over $M = 10$ samples), the grey curve is the best fit (among polynomials of degree $L = 4$) and the thin black curve is the optimal value function

Figure 1 illustrates two value iteration steps ($k = 2$ and $k = K = 20$) of the sampling based FVI algorithm: the dots represents the points $\{X_n\}_{1\le n\le N}$ for $N = 100$ samples drawn from distribution $\mu$ and the average values $\{\hat{V}(X_n)\}_{1\le n\le N}$ (over $M = 10$ samples). The black curve is the best fit (minimizing the least square error to the data) in $\mathcal{F}$ (for $L = 4$) and the grey curve is the optimal value function.

Table 1 shows the $L_\infty$ approximation errors $||V^* - V_K||_\infty$ for different values of the number of sampled states $N$, the number of sampled next states $M$, and the degree $L$ of the polynomials used for the approximations.[4]

From Table 1 we observe that for a specific value of $N$, when the degree of the polynomials increases, the approximation error decreases first (because of richer approximation spaces) but eventually increases because of overfitting. The overfitting effect decreases with the number of samples $N, M$, as expected.

*Table 1.* Approximation error of the optimal value function as a function of number of states $N$, the number of samples $M$, and the degree $L$ of the fitting polynomials

| N | M | L | $\|\|V^* - V_K\|\|_\infty$ |
|---|---|---|---|
| 100 | 10 | 2 | 3.08914 |
| 100 | 10 | 3 | 2.41143 |
| 100 | 10 | 4 | 1.22714 |
| 100 | 10 | 10 | 2.03977 |
| 1000 | 1000 | 4 | 0.783369 |
| 1000 | 1000 | 10 | 0.563451 |
| 1000 | 1000 | 20 | 0.346433 |
| 1000 | 1000 | 30 | 0.207297 |

## 7. Conclusions

We considered a sampling based FVI for discounted, large (possibly infinite) state-space, finite action Markovian Decision Problems where only a generative model of the problem is available. The algorithm computes an approximate metric projection of the image of the recent iterate under a Monte-Carlo approximation to the Bellman operator to some function space. PAC-style bounds on the weighted $p$-norm approximation error are obtained as a function of the covering number and approximation power of the function space, the iteration number and the sample size.

We combined results of approximation theory, learning theory and dynamic programming to show that sampling based FVI can be used to compute approximately optimal actions in time $O(|\mathcal{A}|R^2_{\max}/\alpha^2 \log(|\mathcal{A}|V_{\max}/(\epsilon(1-\gamma))))$ after an initialization phase that needs sample sizes that depend polynomially on $1/\epsilon$, $\log(1/\delta)$, $\log(1/(1-\gamma))$, $V_{\max}$, $\hat{R}_{\max}$, $\log(|\mathcal{A}|)$, and the logarithm of the covering numbers of the function space involved.

This is in contrast to earlier results e.g. those obtained in (Kearns et al., 1999) where the dependence on $1/(1-\gamma)$ is exponential. On the other hand, due to the inherent difficulty of regression, for specific choices of the function approximation method (e.g. linearly parameterized methods) our bounds will suffer from the curse of dimensionality. For such high dimensional spaces, dimension reduction techniques could be used along with our algorithm.

On the other hand, the performance of the algorithm could also be improved by other means. One such possibility is to change the number of samples by the iteration index, the intuition being that early in the iteration a cruder approximation to $T$ suffices. Another possibility is to speed up the calculations by adapting the number of samples $M$ at each sample $X_i$ along the lines of the algorithms in (E. Even-Dar & Mansour, 2002). In addition, our analysis could be extended to the study of approximate policy iteration. It seems to be possible to derive results similar to those obtained here but in the learning scenario when only a single sample path of a given policy is available. We believe that the results presented here can be used as the basis of building up a whole class of new results where tools of Dynamic Programming, Monte-Carlo integration/simulation, classification, regression are combined to prove the soundness of practical algorithms.

## 8. Appendix

In this Appendix we give the ideas of some of our results. The full proofs can be found in the extended version of this paper.

### 8.1. Proof of Lemma 4.1

*Proof.* The main idea of the proof is as follows: Let $\epsilon'' > 0$ be arbitrary and let $f^*$ be such that $\|f^* - TV\|_{p,\mu} \le \inf_{f\in\mathcal{F}} \|f - TV\|_{p,\mu} + \epsilon''$. Define the empirical norm $\|\cdot\|_{p,\hat{\mu}}$ by

$$\|f\|^p_{p,\hat{\mu}} = \frac{1}{N}\sum_{i=1}^{N} f(X_i)^p.$$

The following sequence of inequalities holds with probability at least $1 - \delta$:

[4] Although Table 1 shows the approximation errors $||V^* - V_K||_\infty$, a bound on the performance error $||V^* - V^{\pi_K}||_\infty$ of using the greedy policy $\pi_K$ w.r.t. $V_K$ may be obtained from the usual $L_\infty$ result $||V^* - V^{\pi_K}||_\infty \le \frac{2\gamma}{1-\gamma}||V^* - V_K||_\infty$ (cf. (Bertsekas & Tsitsiklis, 1996a)).

$$
\begin{aligned}
\|V' - TV\|_{p,\mu} &\leq \|V' - TV\|_{p,\hat{\mu}} + \epsilon' && (8.1)\\
&\leq \|V' - \hat{V}\|_{p,\hat{\mu}} + 2\epsilon' && (8.2)\\
&\leq \|f^* - \hat{V}\|_{p,\hat{\mu}} + 2\epsilon' && (8.3)\\
&\leq \|f^* - TV\|_{p,\hat{\mu}} + 3\epsilon' && (8.4)\\
&\leq \|f^* - TV\|_{p,\mu} + 4\epsilon' && (8.5)
\end{aligned}
$$

It follows then that $\|V' - TV\|_{p,\mu} \leq \inf_{f\in\mathcal{F}} \|f - TV\|_{p,\mu} + 4\epsilon' + \epsilon''$ with probability at least $1-\delta$. Since $\epsilon'' > 0$ was arbitrary, it also follows that $\|V' - TV\|_{p,\mu} \leq \inf_{f\in\mathcal{F}} \|f - TV\|_{p,\mu} + 4\epsilon'$ with probability at least $1-\delta$. Now, the desired result follows since by Equation (4.2), $\inf_{f\in\mathcal{F}} \|f - TV\|_{p,\mu} \leq \epsilon/5$, so choosing $\epsilon' = \epsilon/5$ can be used to finish the proof.

The proof of (8.3) follows due to the choice of $V'$ since $\|V' - \hat{V}\|_{p,\hat{\mu}} \leq \|f - \hat{V}\|_{p,\hat{\mu}}$ holds for all functions $f$ from $\mathcal{F}$ and thus the same inequality holds for $f^* \in \mathcal{F}$, too. Next one proves that each of (8.1),(8.2),(8.4) and (8.5) hold with probability at least $1-\delta'$ with $\delta' = \delta/4$. Pollard's inequality can be used to prove (8.1) and (8.5), whilst Hoeffding's inequality can be used to prove (8.2), (8.4): Hoeffding's inequality is used to obtain a bound on the probability of the error of approximating $(TV)(X_i)$ by $\hat{V}(X_i)$, whilst Pollard's inequality is used to obtain a similar result for the error of approximation of $\|V' - \hat{V}\|^p_{p,\mu}$ (resp. $\|f^* - TV\|^p_{p,\mu}$) by $\|V' - \hat{V}\|^p_{p,\hat{\mu}}$ (resp. $\|f^* - TV\|^p_{p,\hat{\mu}}$). □

### 8.2. Proof of Lemma 4.2

*Proof.* The proof is analogous to that of Lemma 4.1, the only difference is that at appropriate places in the proof where e.g. $\|f - TV\|_{p,\mu}$ appear, we use $\|f - Tg\|_{p,\mu}$ and take a supremum over $g \in \mathcal{F}$. As a result, instead of Hoeffding's inequality we use Pollard's inequality to derive bounds for (8.2) and (8.4). □

### 8.3. Proof of Theorem 5.1

The main idea is that iteration (2.1) may be written $V_{k+1} = TV_k + \varepsilon_k$ where $\varepsilon_k$ is the approximation error of the Bellman-operator applied to $V_k$ due to sampling. The proof of Theorem 5.1 follows directly from Lemma 4.1 and the next result (which makes use of Assumptions A1 and A2 like in (Munos, 2005)).

**Lemma 8.1.** *For any $\eta > 0$, there exists $K$ that is linear in $\log(1/\eta)$ (and $\log V_{\max}$) such that, if the $L_{p,\mu}$ norm of the approximation errors is bounded by some $\epsilon$, i.e. $\|\varepsilon_k\|_{p,\mu} \leq \epsilon$ for all $0 \leq k < K$, then given Assumption A1 (resp. Assumption A2), we have $\|V^* - V^{\pi_K}\|_\infty \leq \frac{2\gamma}{(1-\gamma)^2} C\epsilon + \eta$. (resp. $\|V^* - V^{\pi_K}\|_{p,\rho} \leq \frac{2\gamma}{(1-\gamma)^2} C^{1/p}\epsilon + \eta$.)*

**Acknowledgement**: Csaba Szepesvári was supported by OTKA Grant T047193 and by the Bolyai Fellowship of the Hungarian Academy of Sciences.

## References

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

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

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

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

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

E. Even-Dar, S. M., & Mansour, Y. (2002). PAC bounds for multi-armed bandit and Markov decision processes. *Fifteenth Annual Conference on Computational Learning Theory (COLT)* (pp. 255–270).

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

Kearns, M., Mansour, Y., & Ng, A. (1999). A sparse sampling algorithm for near-optimal planning in large Markovian decision processes. *Proceedings of IJCAI'99* (pp. 1324–1331).

Munos, R. (2003). Error bounds for approximate policy iteration. *19th International Conference on Machine Learning*, 560–567.

Munos, R. (2005). Error bounds for approximate value iteration. *American Conference on Artificial Intelligence.*

Rust, J. (1996). Numerical dyanmic programming in economics. In H. Amman, D. Kendrick and J. Rust (Eds.), *Handbook of computational economics.* Elsevier, North Holland.

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

Singh, S., Jaakkola, T., & Jordan, M. (1995). Reinforcement learning with soft state aggregation. *Proceedings of Neural Information Processing Systems 7* (pp. 361–368). MIT Press.

Szepesvári, C., & Smart, W. (2004). Interpolation-based Q-learning. *Proceedings of the International Conference on Machine Learning* (pp. 791–798).

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

Zhang, T. (2002). Covering number bounds of certain regularized linear function classes. *Journal of Machine Learning Research*, *2*, 527–550.