Finite Time Bounds for Sampling-Based Fitted Value Iteration

Rémi Munos, Csaba Szepesvari

Finite Time Bounds for Sampling-Based Fitted Value Iteration

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

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

Editor:

Abstract

In this paper we develop a theoretical analysis of the performance of sampling-based fitted value iteration (FVI) for solving large, or infinite state-space Markovian decision problems (MDP) with a generative model. Unlike most previous results, our theoretical guarantees apply to a large class of regressors (e.g. neural networks, adaptive regression trees, kernel machines, locally weighted learning). The bounds derived on the performance of sampling-based FVI make the dependence of loss explicit on the MDP's controllability and smoothness properties, the MDP's relation to the regressor employed and several other algorithmic choices. We discuss the relation of our results to previous results and illustrate some tradeoffs by means of a computer experiment.

Keywords: fitted value iteration, discounted Markovian decision problems, generative model, reinforcement learning, supervised learning, regression, Pollard's inequality, statistical learning theory, optimal control

1. Introduction

During the last decade, reinforcement learning (RL) algorithms have been successfully applied to a number of difficult problems, such as playing backgammon (Tesauro, 1995), job-shop scheduling (Zhang and Dietterich, 1995), elevator control (Crites and Barto, 1997), machine maintenance (Mahadevan et al., 1997), dynamic channel allocation (Singh and Bertsekas, 1997), and airline seat allocation (Gosavi, 2004), just to name a few. One common feature of these and similar practical problems is that their underlying state spaces is huge and given the available finite resources it is not possible to explore all parts of the state space. Algorithms designed to work for such large state space problems must be capable to generalize to unseen cases, which is often tackled by relying on a powerful function approximation technique. Yet, our understanding of the interaction of RL algorithms and function approximators is still very limited.

In this paper we study one of the simplest ways to combine RL and function approximators, namely, the combination of function approximators and value iteration. In value

iteration "value backups" generate a sequence of value functions (i.e., functions defined over the state space) in a recursive manner and it is guaranteed that after a sufficiently large number of iterations the resulting function can be used to compute a good policy in a simple manner. When the state space is large or infinite, neither exact computations, nor the exact representation of all possible iterates is possible. Hence, in any practical implementation approximate computations are called for and the iterates have to be restricted to lie in some function space (in other words, the iterates should be representable using some function approximator). The idea underlying *sampling-based fitted value iteration (FVI)* is to calculate the back-ups approximately at a finite number of points and then find the best fit to the computed values. The hope is that if the function approximator is sufficiently powerful, then the policy obtained ultimately will have a good performance. Indeed, this has been observed in practice in a number of cases (e.g. Wang and Dietterich, 1999; Dietterich and Wang, 2002; Lagoudakis and Parr, 2003; Jung and Uthmann, 2004; Ernst et al., 2005; Riedmiller, 2005), though almost all practitioner notes that the choice of the function approximator is critical for success.

Despite these good empirical results, there are still doubts about the extent to which sampling-based FVI can be expected to perform well. In particular, Baird (1995) and Tsitsiklis and Van Roy (1996) gave simple counterexamples showing that some instability may arise even in the simplest cases. In particular, in these example it is assumed that the environment is known, the back-ups of the iterates are calculated exactly, the optimal value function lies within the range of the function approximator (hence the function approximator looks powerful enough) and that the function approximator is linearly parameterized. Since value iteration without projection is stable, we must conclude that the instable behaviour is the result of the errors introduced when the iterates are projected into the function space. Our aim in this paper is to develop a better understanding of why, despite the conceivable difficulties, practitioners often observe good performance for sampling-based FVI.

The particular problem that we study here is to find a good policy given a generative model of the environment, i.e., planning using a generative model. A generative model allows the algorithm to sample any transitions for any particular state-action pair of the MDP (Kearns et al., 1999; Ng and Jordan, 2000; Kakade, 2003). Although the planning problem is considerably simpler than the learning problem, in the case of large state spaces it is still difficult enough to warrant its separate study.

The algorithm studied uses the generative model to sample transitions from randomly selected initial states. The resulting sample set is used effectively as an implicit, random representation of the MDP. We investigate two versions of the basic algorithm: In the multi-sample variant in each iteration a fresh sample set is generated, whilst in the single-sample variant the same sample set is used throughout all the iterations.

Our results come in the form of high-probability bounds on the performance as a function of the number of samples generated, some properties of the MDP and the function approximation class. The bounds resemble those available in supervised learning where alike bounds have two terms, one decaying with the increase of the *approximation power* of the function class, bounding the bias of the algorithm, whilst the other decaying polynomially with the number of samples, bounding the *estimation error*, sometimes also called the variance.

Whilst our bounds are similar to bounds of supervised learning (regression, in particular), there are some important differences. First, in our bounds the power of the function approximator is measured somewhat differently than it is done in e.g. regression where the power of a function space $\mathcal{F}$ is usually measured as the largest approximation error over some fixed reference class $\mathcal{G}$: $d(\mathcal{G}, \mathcal{F}) = \sup_{g \in \mathcal{G}} \inf_{f \in \mathcal{F}} \|f - g\|$. (The reference class is usually a classical smoothness class.) First, notice that this measure looks inadequate for our purposes since in the counterexamples of Baird (1995) and Tsitsiklis and Van Roy (1996) the target function can be approximated with zero error, but the algorithm still exhibits unbounded errors. Hence, our bounds depend on a different characterization of the approximation power of a function class. Namely, they rely on the *inherent Bellman error of the function space*:

$$d(T\mathcal{F}, \mathcal{F}) = \sup_{g \in \mathcal{F}} \inf_{f \in \mathcal{F}} \|f - Tg\|.$$

Here T is the Bellman operator underlying the MDP and $\|\cdot\|$ is an appropriate weighted p-norm that can be chosen by the user (the exact definitions will be given in Section 2). Observe that no fixed reference class is used in defining $d(T\mathcal{F}, \mathcal{F})$: the definition instead connects the Bellman operator (hence the MDP) and the function-space. Further, in the above-mentioned counterexamples the inherent Bellman error of the function space chosen is infinite (as expected).

Unlike the term bounding the approximation power, the term bounding the estimation error is very close to its counterpart from regression: our bounds also depend on the capacity of the function space employed and decay polynomially with the number of samples, though the rates are different. In particular, the bounds still indicate that the error of approximating the optimal value function can be made to approach the inherent Bellman error of the function space as the number of samples grows to infinity.

The combined bound also displays the so-called *bias-variance tradeoff*: In order to keep both error types small the number of samples have to be increased simultaneously with the increase of the power of the function class at an appropriate rate.

Our proof technique works only if the unknown MDP satisfies a number of conditions. The MDP is assumed to possess a continuous, compact state-space[1] and the number of actions is assumed to be finite. The main assumption on the dynamics is that the *discounted-average concentrability of future-state distributions* should be bounded. The *t-step concentrability* of the future-state distribution measures how much it can be concentrated in t steps as compared to some reference distribution. The discounted-average concentrability is the discounted-average of all the t-step concentrability factors. When the MDP's transition kernel possesses a bounded density, a case that is often considered in previous theoretical works (Chow and Tsitsiklis, 1989, 1991; Rust, 1996b; Szepesvári, 2000), then the t-step concentrability factors stay bounded and hence their discounted-average obeys the same bound.

As discussed previously our bounds depend on the inherent Bellman error of the chosen function space. When the MDP is unknown, in general it is difficult to select the function space so as to make its inherent Bellman error small. We will show that if one assumes that the transitions and rewards are smooth (most previous works such as the ones mentioned

1. In fact, for the sake of simplicity we work with MDPs whose state space is the d-dimensional unit-cube. However, the results can be extended to any compact-space MDPs without any difficulties.

above assume this condition, too) then it is possible to use the techniques of classical approximation theory (Cheney, 1966, e.g. Section 5.6) to construct such function spaces. In fact, this analysis also suggests that non-linear approximation methods that adapt to the local smoothness properties of the target function (such as wavelets employing shrinkage estimation) may yield better performance than linear methods.

Our results build on a recent proof technique, originally suggested by Munos (2003), to propagate weighted p-norm losses in value iteration. Unlike Munos (2003) and later (Munos, 2005), we do not assume that an exact model of the environment is available, but we consider in detail the influence that finite random samples can have on the final error. The main advantage of carrying out the analysis using weighted p-norms (especially important is the case when $p = 2$) is that these norms match the form of the empirical risk-term minimized in the projection step of the algorithm, hence they are expected to be stronger.

As a result of studying weighted p-norms, unlike the results of e.g. Gordon (1995) or Tsitsiklis and Van Roy (1996), our analysis technique allows us to bound the error even when a non-linear approximation method is used. It is also worth noting here that the technique developed here can be extended to more complicated learning scenarios. An example of such an extension has recently been worked out by Antos et al. (2006) who studied Bellman-residual minimization with a single trajectory of a fixed behaviour policy.

The paper is organized as follows: In the next section we formally introduce the concepts needed in the rest of the paper. The problem is defined and the algorithms are given in Section 3. Next, we develop finite-sample bounds for the error committed in a single step of the algorithms in Section 4. This bound is then used in proving our main result (Section 5). Next, we extend the result to the problem of obtaining a good policy (Section 6), followed by a construction that allows one to achieve an arbitrarily small final loss when the MDP is unknown but is assumed to be smooth (Section 7). Relationship to previous works is discussed in detail in Section 8. An experiment in a simulated environment, highlighting the main lessons of the analysis is given in Section 9. The full proofs of the statements are given in the Appendix.

2. Markovian Decision Problems

A discounted *Markovian Decision Problem* (discounted MDP) is a 5-tuple $(\mathcal{X}, \mathcal{A}, P, S, \gamma)$, where $\mathcal{X}$ is the *state space*, $\mathcal{A}$ is the *action space*, P is the *transition probability kernel*, S is *the reward kernel* and $0 < \gamma < 1$ is the *discount factor* (Bertsekas and Shreve, 1978; Puterman, 1994). In this paper we consider continuous state space, finite action MDPs (i.e., $|\mathcal{A}| < +\infty$). For the sake of simplicity we assume that $\mathcal{X}$ is a bounded, closed subset of a Euclidean space, $\mathbb{R}^d$.

The interpretation of an MDP as a control problem is as follows: Each initial state X_0 and action sequence $a_0, a_1, \ldots$ gives rise to a sequence of states $X_1, X_2, \ldots$ and rewards $R_1, R_2, \ldots$ satisfying, for any B and C Borel-measurable sets the equalities

$$\mathbb{P}\left(X_{t+1} \in B | X_t = x, a_t = a\right) = P(B|x, a),$$

and

$$\mathbb{P}\left(R_t \in C | X_t = x, a_t = a\right) = S(C|x, a).$$

Equivalently, we write $X_{t+1} \sim \mathbb{P}(\cdot|X_t, A_t)$, $R_t \sim \mathbb{P}(\cdot|X_t, A_t)$. In words, we say that when action a_t is executed from state $X_t = x$ the process makes a transition from x to the next state X_{t+1} and a reward, R_t, is incurred. The history of the process up to time t is $H_t = (X_0, a_0, R_0, \ldots, X_{t-1}, a_{t-1}, R_{t-1}, X_t)$. We assume that the random rewards $\{R_t\}$ are bounded by some positive number $\hat{R}_{\max}$, with probability one.

A *policy* is a sequence of functions that maps possible histories to probability distributions over the space of actions. Hence if the space of histories at time step t is denoted by $\mathcal{H}_t$ then a policy π is a sequence $\pi_0, \pi_1, \ldots$, where π_t maps $\mathcal{H}_t$ to $M(\mathcal{A})$, the space of all probability distributions over $\mathcal{A}$.[2] 'Following a policy' means that for any time step t given the history $x_0, a_0, \ldots, x_t$ the probability of selecting an action a equals $\pi_t(x_0, a_0, \ldots, x_t)(a)$. A policy is called *stationary* if π_t depends only on the last state visited. Equivalently, a policy $\pi = (\pi_0, \pi_1, \ldots)$ is called stationary if $\pi_t(x_0, a_0, \ldots, x_t) = \pi_0(x_t)$ holds for all $t \geq 0$. A policy is called *deterministic* if for any history $x_0, a_0, \ldots, x_t$ there exists some action a such that $\pi_t(x_0, a_0, \ldots, x_t)$ is concentrated on this action. Hence, any deterministic stationary policy can be identified by some mapping from the state space to the action space and so in the followings, at the price of abusing the notation and the terminology slightly, we will call such mappings policies, too.

The goal is to find a policy π that maximizes the expected total discounted reward given any initial state. Under this criterion the value of a policy π and a state $x \in \mathcal{X}$ is given by

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

where R_t^π is the reward incurred at time t when executing policy π. The optimal expected total discounted reward when the process is started from state x shall be denoted by $V^*(x)$; V^* is called the *optimal value function*. A policy π is called *optimal* if it attains the optimal values for *any* state $x \in \mathcal{X}$, i.e., if $V^\pi(x) = V^*(x)$ for all $x \in \mathcal{X}$. We also let $Q^*(x, a)$ denote the long-term total expected discounted reward when the process is started from state x, the first executed action is a and it is assumed that after the first step an optimal policy is followed. Under mild conditions, it can be shown that deterministic stationary optimal policies always exist (Bertsekas and Shreve, 1978).

Let us now introduce a few function spaces and operators that will be needed in the rest of the paper. 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 $0 < V_{\max} < +\infty$ shall be denoted by $B(\mathcal{X}; V_{\max})$. A deterministic stationary policy $\pi : \mathcal{X} \to \mathcal{A}$ defines the transition probability kernel P^π according to $P^\pi(dy|x) = P(dy|x, \pi(x))$. From this kernel, two related operators are derived: a right-linear operator, $P^\pi \cdot : B(\mathcal{X}) \to B(\mathcal{X})$, defined by

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

2. In fact, π_t must be a measurable mapping so that we are allowed to talk about e.g. the probability of executing an action. The reader interested in these issues is referred to Bertsekas and Shreve (1978), here we note only that for the class of discounted MDPs considered here there are no difficulties from this point of view.

and a left-linear operator, $\cdot P^\pi : M(\mathcal{X}) \to M(\mathcal{X})$, defined by

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

Here $\mu \in M(\mathcal{X})$ and $M(\mathcal{X})$ is the space of all probability distributions over $\mathcal{X}$.

In words, $(P^\pi V)(x)$ is the expected value of V after following π for a single time-step when starting from x, and μP^π is the distribution of states if the system is started from $X_0 \sim \mu$ and policy π is followed for a single time-step. The product of two kernels P^{π_1} and P^{π_2} is defined in the natural way:

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

Hence, $\mu P^{\pi_1}P^{\pi_2}$ is the distribution of states if the system is started from $X_0 \sim \mu$, policy π_1 is followed for the first step and then policy π_2 is followed for the second step. The interpretation of $(P^{\pi_1}P^{\pi_2}V)(x)$ is similar.

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

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

where $r(x,a) = \int zS(dz|x,a)$ is the expected reward of executing action a in state x. We assume that r is a bounded, measurable function. Actions maximizing $r(x,a) + \gamma \int V(y)P(dy|x,a)$ are said to be *greedy* w.r.t. V. Since $\mathcal{A}$ is finite the set of greedy actions is non-empty for any function V.

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

$$(TV)(x) = \max_{a\in\mathcal{A}} \left\{ r(x,a) + \gamma \int V(y)P(dy|x,a) \right\}, \quad V \in B(\mathcal{X}).$$

The operator T is called the *Bellman operator* underlying the MDP. Similarly, to any stationary deterministic policy π there corresponds an operator $T^\pi : B(\mathcal{X}) \to B(\mathcal{X})$ defined by

$$(T^\pi V)(x) = r(x,\pi(x)) + (P^\pi V)(x).$$

It is well known that T is a contraction mapping in supremum norm with contraction coefficient γ: $\|TV - TV'\|_\infty \le \gamma \|V - V'\|_\infty$. Hence, by Banach's fixed-point theorem, T possesses a unique fixed point. Moreover, under mild technical conditions this fixed point turns out to be equal to the optimal value function, V^*. Then a simple contraction argument shows that the so-called *value-iteration algorithm*,

$$V_{k+1} = TV_k,$$

with arbitrary $V_0 \in B(\mathcal{X})$ yields a sequence of iterates, V_k, that converge to V^* at a geometric rate. The contraction arguments also show 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 the same holds

for V_k, too. Proofs of these statements can be found in many textbooks such as (Bertsekas and Shreve, 1978).

Our initial set of assumptions on the class of MPDs considered is summarized as follows:

Assumption A0 [MDP Regularity] The MDP $(\mathcal{X}, \mathcal{A}, P, S, \gamma)$ satisfies the following conditions: $\mathcal{X}$ is a bounded, closed subset of some Euclidean space, $\mathcal{A}$ is finite, the discount factor γ 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 $S(\cdot|x,a)$ is included in $[-\hat{R}_{\max}, \hat{R}_{\max}]$ independently of $(x,a) \in \mathcal{X} \times \mathcal{A}$.

3. Sampling-based Fitted Value Iteration

We are now ready to introduce the sampling-based FVI algorithm:

The algorithm uses a function set $\mathcal{F} \subset B(\mathcal{X})$. The choice of $\mathcal{F}$ is governed by the need to be able to represent functions using a finite number of parameters and by some other requirements that we will discuss later. We shall also give some specific examples of how to select $\mathcal{F}$. The algorithm works by first selecting $V_0 \in \mathcal{F}$ and then computing a series of functions, $V_1, V_2, \ldots$ in a recursive manner. The $(k+1)$th function is obtained from the kth function as follows: First a Monte-Carlo estimate of TV_k is computed at a number of random states:

$$\hat{V}(X_i) = \max_{a \in \mathcal{A}} \frac{1}{M} \sum_{j=1}^{M} \left[R_j^{X_i,a} + \gamma V_k(Y_j^{X_i,a}) \right], \quad i = 1, 2, \ldots, N.$$

Here the *basepoints*, $X_1, \ldots, X_N$, are sampled from some distribution $\mu \in M(\mathcal{X})$, independently of each other. For each of these basepoints and for each possible action $a \in \mathcal{A}$ the next states, $Y_j^{X_i,a} \in \mathcal{X}$, and rewards, $R_j^{X_i,a} \in \mathbb{R}$, are drawn via the help of the generative model of the MDP:

$$\begin{aligned} Y_j^{X_i,a} &\sim P(\cdot|X_i,a), \\ R_j^{X_i,a} &\sim R(\cdot|X_i,a), \end{aligned}$$

where $j = 1, 2, \ldots, M$, $i = 1, \ldots, N$. We assume that $(Y_j^{X_i,a}, R_j^{X_i,a})$ and $(Y_{j'}^{X_{i'},a'}, R_{j'}^{X_{i'},a'})$ are independent of each other whenever $(i,j,a) \neq (i',j',a')$. Then, the next iterate V_{k+1} is obtained as the best fit in $\mathcal{F}$ to the data $(X_i, \hat{V}(X_i))_{i=1,2,\ldots,N}$ w.r.t. the p-norm based empirical loss

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

This computation is repeated K times for some K that is specified so as to ensure a small final estimation error of V^*.

We call the above version of the algorithm the *multi-sample variant* since a fresh sample is generated in each iteration step k. The *total number* of samples used by the multi-sample algorithm is $K \times N \times M$. Since in a single iteration only a fraction of these samples is used, one may wonder if it were more sample-efficient to use *all the samples* in all the iterations. We shall call the version of the algorithm that implements this idea the *single-sample variant*

of FVI (the equations for the iterates will be given in Section 5). A possible counterargument against the single-sample variant is that since the samples used in subsequent iterations are correlated, the bias due to sampling errors may get amplified in the iterations. One of our interesting theoretical findings is that the bias-amplification effect is not too severe and in fact, the single-sample variant of the algorithm is well behaving. In the experiments we will see a case when the single-sample variant is not only well behaving but can indeed outperform the multi-sample variant.

Let us now discuss the choice of the function space $\mathcal{F}$. Generally, $\mathcal{F}$ can be selected to be a parameterized class of functions:

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

In fact, we will see that our results apply to both linear ($f_\theta(x) = \theta^T \phi(x)$) and non-linear ($f_\theta(x) = f(x;\theta)$) parameterizations, such as neural networks. Another possibility is to use the kernel mapping idea underlying many of the recent top performing supervised-learning algorithms, such as support-vector machines, support-vector regression or Gaussian processes (Cristianini and Shawe-Taylor, 2000). In this case we let $\mathcal{F}$ be defined with the help of some (positive definite) kernel function $\mathcal{K}$: If $\mathcal{F}$ is a closed subset of $\mathcal{H}$, the Reproducing Kernel Hilbert Space (RKHS) associated with $\mathcal{K}$ then, thanks to the Representer Theorem of Kimeldorf and Wahba (1971), the optimization problem (1) admits a solution that can be expanded in terms of the training samples, i.e., written in the form of $\sum_{i=1}^n \alpha_i \mathcal{K}(x, X_i)$ with appropriate coefficients $\{\alpha_i\}$. Hence, although the space of functions $\mathcal{F}$ is not finitely parameterized, the optimization problem is still manageable.[3]

4. Approximating the Bellman Operator

The purpose of this section is to bound the error introduced in a single iteration of the algorithm. There are two components of this error: The approximation error caused by projecting the iterates into the function space $\mathcal{F}$ and the estimation error caused by using a finite, random sample.

The approximation error can be best explained by introducing the metric projection operator: Fix the sampling distribution $\mu \in M(\mathcal{X})$ and let $p \geq 1$. Let us define the *metric projection* of TV onto $\mathcal{F}$ w.r.t. the μ-weighted p-norm by

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

Here, for a real-valued measurable function g defined over $\mathcal{X}$, $\|g\|_{p,\mu}$ is defined by $\|g\|_{p,\mu}^p = \int |g(x)|^p \mu(dx)$. The space of functions with bounded $\|\cdot\|_{p,\mu}$-norm shall be denoted by $L^p(\mathcal{X};\mu)$. For $g \in L^p(\mathcal{X};\mu)$, $\Pi_{\mathcal{F}} g$ gives the *best approximation* to g in $\mathcal{F}$. The existence and uniqueness of best approximations is one of the fundamental problems of approximation theory. Existence can be guaranteed under fairly mild conditions, such as the compactness of $\mathcal{F}$ w.r.t. $\|\cdot\|_{p,\mu}$, or if $\mathcal{F}$ is finite dimensional (Cheney, 1966). Since the metric projection operator is needed for discussion purposes only, here we simply assume that $\Pi_{\mathcal{F}}$ is well-defined.

3. In the Representer Theorem of Kimeldorf and Wahba (1971), unlike in our case there is no restriction on where the solutions lie. Since the proof relies on the expansion of $f(X_i)$ in terms of $\{\mathcal{K}(\cdot, X_j)\}_{j=1}^N$ and an orthogonal component, it still goes through without any changes.

The approximation error in the kth step for $V = V_k$ is $d_{p,\mu}(TV, \mathcal{F}) = \|\Pi_{\mathcal{F}} TV - TV\|_{p,\mu}$, or more generally,

$$d_{p,\mu}(TV, \mathcal{F}) = \inf_{f \in \mathcal{F}} \|f - TV\|_{p,\mu} . \tag{2}$$

Hence, the approximation error can be made small by selecting $\mathcal{F}$ to be large enough.

Now, let us reiterate the algorithm in order to discuss the behaviour of the estimation error. As noted beforehand, sampling-based FVI can be thought to approximate $\Pi_{\mathcal{F}} TV$ by first employing Monte-Carlo integration to approximate TV:

$$\hat{V}(X_i) = \max_{a \in \mathcal{A}} \frac{1}{M} \sum_{j=1}^{M} \left[R_j^{X_i,a} + \gamma V(Y_j^{X_i,a}) \right], \quad i = 1, 2, \ldots, N, \tag{3}$$

followed by computing the best fit in $\mathcal{F}$ to the values obtained:

$$V' = \operatorname*{argmin}_{f \in \mathcal{F}} \sum_{i=1}^{N} \left| f(X_i) - \hat{V}(X_i) \right|^p . \tag{4}$$

(Here the random samples are assumed to satisfy the conditions of the previous section and for the sake of simplicity, we assume that the minimizer in Equation (4) exists.) The logics underlying Equations (3) and (4) is this: if M is big then $\hat{V}(X_i)$ will be a good approximation to $(TV)(X_i)$. Now, if N is big then for any $(f, g) \in \mathcal{F} \times \mathcal{F}$, the *empirical loss* $1/N \sum_{i=1}^{N} (f(X_i) - g(X_i))^p$ is a good approximation to the *true loss* $\|f - g\|_{p,\mu}^p$. Hence, we expect to find the minimizer of (4) to be close to the minimizer of $\|f - TV\|_{p,\mu}^p$. Since the function x^p is strictly increasing for $x > 0$, $p > 0$, the minimizer of $\|f - TV\|_{p,\mu}^p$ over $\mathcal{F}$ is just the metric projection of TV on $\mathcal{F}$, hence V' can be expected to be close to $\Pi_{\mathcal{F}} TV$.

Note that Equation (4) looks like an ordinary regression problem. However, unlike in regression, our target function, TV, is typically different from the regressor $g(x) = E[\hat{V}(X_i) | X_i = x]$. This is because the expectation of the maximum of some random variables is in general larger than the maximum of the expectation of the same random variables. In our case

$$\begin{aligned}
\mathbb{E}\left[\max_{a \in \mathcal{A}} \frac{1}{M} \sum_{j=1}^{M} \left[R_j^{X_i,a} + \gamma V(Y_j^{X_i,a}) \right] \Big| X_i \right] &\geq \max_{a \in \mathcal{A}} \mathbb{E}\left[\frac{1}{M} \sum_{j=1}^{M} \left[R_j^{X_i,a} + \gamma V(Y_j^{X_i,a}) \right] \Big| X_i \right] \\
&= \max_{a \in \mathcal{A}} \left[r(X_i, a) + \gamma \int V(y) P(dy | X_i, a) \right] \\
&= (TV)(X_i).
\end{aligned}$$

In fact, if the bias here were zero then we would have no reason to set $M > 1$. Indeed, in a pure regression setting the error does not improve if more than one sample is collected at the covariates, $\{X_i\}$; we must use $M > 1$ in order to decrease the bias only. As we will see, this results in an increase of the sample complexity (or decrease of convergence rate) as compared with the pure regression setting.

Above we argued that N should be increased to make the empirical loss approximate the true loss, i.e., to make the *estimation error* small. For any fixed function pair, this holds without any further conditions. However, we need this property to hold for V'. Since V' is

the minimizer of the empirical loss, it depends on the random sample and hence it is itself a random object. Hence the argument that ensures that the estimation error is small for any *fixed*, deterministic pair of functions does not apply to V'. The same problem arises in regression problems and so we use the techniques available for these problems. One widely used idea is to bound the estimation error at the random function by the worst estimation error over $\mathcal{F}$: For $g \in L^p(\mathcal{X};\mu)$,

$$\left| \frac{1}{N}\sum_{i=1}^{N}(V'(X_i)-g(X_i))^p - \left\|V'-g\right\|_{p,\mu}^p \right| \le \sup_{f\in\mathcal{F}} \left| \frac{1}{N}\sum_{i=1}^{N}(f(X_i)-g(X_i))^p - \|f-g\|_{p,\mu}^p \right|, \tag{5}$$

holds with probability one since $V' = V'(\omega)$, for any random event ω, is an element of $\mathcal{F}$. The right-hand-side here is the maximal deviation of certain empirical averages from their respective means. The behaviour of such quantities is the main focus of empirical process theory and we shall use the tools of this theory in developing our bounds.

When bounding the size of maximal deviations, clearly, the size of the function set becomes a major influential factor. When the function set has a finite number of elements, a bound follows by exponential inequalities and a union bounding argument. When $\mathcal{F}$ is infinite, the 'capacity' of $\mathcal{F}$ measured by the *covering number* of $\mathcal{F}$ is used to derive a bound. For completeness, let us now define covering numbers: Let $x^{1:N} \stackrel{\text{def}}{=} (x_1,\dots,x_N) \in \mathcal{X}^N$ be fixed. Fix $\epsilon > 0$, $q \ge 1$. The (ϵ,q)-covering number of the set $\mathcal{F}(x^{1:N}) = \{(f(x_1),\dots,f(x_N)) \,|\, f \in \mathcal{F}\}$ 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 (ϵ,q)-covering number of $\mathcal{F}(x^{1:N})$ is 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 μ then $\mathbb{E}\left[\mathcal{N}_q(\epsilon,\mathcal{F}(X^{1:N}))\right]$ shall be denoted by $\mathcal{N}_q(\epsilon,\mathcal{F},N,\mu)$. By Jensen's inequality $\mathcal{N}_p \le \mathcal{N}_q$ for $p \le q$. The logarithm of $\mathcal{N}_q$ is called the q-norm *metric entropy* of $\mathcal{F}$. For $q=1$, we shall call $\log\mathcal{N}_1$ the metric entropy of $\mathcal{F}$ (without any qualifiers).

The idea underlying covering numbers is that what really matters in bounding maximal deviations is how much the functions in the function space vary *at the actual samples*. Without imposing any conditions on the function space, covering numbers can grow as a function of the sample size. If this growth is too fast or some covering numbers are infinite then the tools presented here are insufficient to prove the convergence of the estimation error to zero with the sample-size going to infinity.

It turns out that for specific choices of $\mathcal{F}$ it is possible to bound the covering numbers *independently* of the number of samples. In fact, according to a well-known result due to Haussler (1995), covering numbers can be bounded as a function of the so-called *pseudo-dimension* of the function class if it is finite. The pseudo-dimension, or VC-subgraph dimension $V_{\mathcal{F}^+}$ of $\mathcal{F}$ is defined as the VC-dimension of the subgraphs of functions in $\mathcal{F}$.[4] The following statement gives the bound that is independent on the number of sample-points:

Proposition 1 (Haussler (1995), Corollary 3) *For any set $\mathcal{X}$, any points $x^{1:N} \in \mathcal{X}^N$, any class $\mathcal{F}$ of functions on $\mathcal{X}$ taking values in $[0,L]$ with pseudo-dimension $V_{\mathcal{F}^+} < \infty$, and*

4. The concept of VC-dimension was introduced by Sauer (1972); Vapnik and Chervonenkis (1971) and is defined as follows: Given a set system $\mathcal{C}$ with base set $\mathcal{X}$ we say that $\mathcal{C}$ shatters the points of $A \subset \mathcal{X}$ if all possible $2^{|A|}$ subsets of A can be obtained by intersecting A with elements of $\mathcal{C}$. The VC-dimension of $\mathcal{C}$ is the cardinality of the largest subset $A \subset \mathcal{X}$ that can be shattered.

any $\epsilon > 0$,

$$\mathcal{N}(\epsilon, \mathcal{F}(x^{1:N})) \le e(V_{\mathcal{F}^+} + 1)\left(\frac{2eL}{\epsilon}\right)^{V_{\mathcal{F}^+}}. \tag{6}$$

For a given set of functions $\mathcal{F}$ let $a+\mathcal{F}$ denote the set of functions shifted by the constant a: $a + \mathcal{F} = \{ f + a \,|\, f \in \mathcal{F} \}$. Clearly, covering numbers are not changed by shifts. This allows one to extend Proposition 1 to function sets with functions taking values in $[-L, +L]$. As advertised, the bound (6) does not depend on the sample size N. Bounds on the pseudo-dimension are known for many well-known function classes including linearly parameterized function classes, multi-layer perceptrons, radial basis function networks, several non- and semi-parametric function classes, cf. Niyogi and Girosi (1999); Anthony and Bartlett (1999); Györfi et al. (2002); Zhang (2002) and the references therein. In these bounds the metric entropy scales with the dimensionality q of the function space as $O(\log(q))$, $O(q)$ or $O(q \log q)$.[5] For these function spaces the generalization error is well behaving and can be bounded with high probability when N is polynomial in q, $1/\epsilon$, $\log(1/\delta)$, $V_{\max}$, $\hat{R}_{\max}$, and $\log(|\mathcal{A}|)$ (cf. Lemma 1 below).

Another route to get a useful bound on the number of samples is to derive an upper bound on the metric entropy that grows sublinearly with the number of samples. This way it is possible to obtain bounds for linear function classes that do not depend on the dimensionality of the function set. Consider the following class of bounded-weight linear functions:

$$\mathcal{F}_A = \{ f_\theta : \mathcal{X} \to \mathbb{R} \,|\, f_\theta(x) = \theta^T \phi(x),\ \|\theta\|_q \le A \}.$$

For finite-dimensional smooth parametric classes the metric entropy scales with $\dim(\phi) < +\infty$. On the other hand, if ϕ is the feature map associated with some positive definite kernel function $\mathcal{K}$ then ϕ can be infinite dimensional (one could argue that this class arises if one "kernelizes" FVI). In this case the bounds due to Zhang (2002) could be used. Since these scale proportionally with the logarithm of the number of samples, they will still give rise to roughly the same sample complexity even though the function space is infinite dimensional, as it follows from the result of the next section.[6]

4.1 Finite-sample Bounds

The following lemma bounds the number of samples needed to achieve a given accuracy relative to how well $\mathcal{F}$ can approximate the image of V under T:

Lemma 1 *Consider an MDP satisfying Assumption A0. Fix $p \ge 1$, $\mu \in M(\mathcal{X})$ and let $V_{\max} = R_{\max}/(1-\gamma)$. Pick any $V \in B(\mathcal{X}; V_{\max})$ and let V' be defined by Equation (4), where $X_i \sim \mu$. Assume that $\mathcal{F} \subset B(\mathcal{X}; V_{\max})$. Then for any $\epsilon, \delta > 0$,*

$$\left\| V' - TV \right\|_{p,\mu} \le d_{p,\mu}(TV, \mathcal{F}) + \epsilon \tag{7}$$

holds with probability at least $1 - \delta$, provided that

$$N > 512\, V_{\max}^2\, (4/\epsilon)^{2p} \left(\log(1/\delta) + \log(32 \mathcal{N}(\epsilon/32, \mathcal{F}, N, \mu)) \right) \tag{8}$$

5. Many bounds are known to hold for the supremum-norm metric entropy, as well.
6. The actual bounds have the form $\lceil A^2 B^2 / \epsilon^2 \rceil \log(2N+1)$, where B is an upper bound on $\sup_{x \in \mathcal{X}} \|\phi(x)\|_p$ with $p = 1/(1-1/q)$. Note that some of these bounds extend to supremum-norm metric entropies, as well.

and

$$M > \frac{8\,(\hat{R}_{\max} + \gamma V_{\max})^2}{\epsilon^2}\Big(\log(1/\delta) + \log(8N|\mathcal{A}|)\Big). \tag{9}$$

By our previous discussion, for a large number of choices of $\mathcal{F}$, the metric entropy of $\mathcal{F}$ is sublinear in N. In such cases Equation (8) yields an explicit bound on N. Note that the total number of samples, $N \times M$, scales with $\epsilon^{-(2p+2)}$ (apart from logarithmic terms). The comparable bound for the "pure" regression setting is ϵ^{-2p}. The additional quadratic factor represents the price we pay for the biasedness of the values $\hat{V}(X_i)$.

We use Figure 1 to illustrate the proof of this lemma which can be found in Appendix A. The figure has a left and a right sub-figure. The left-hand side sub-figure shows the space

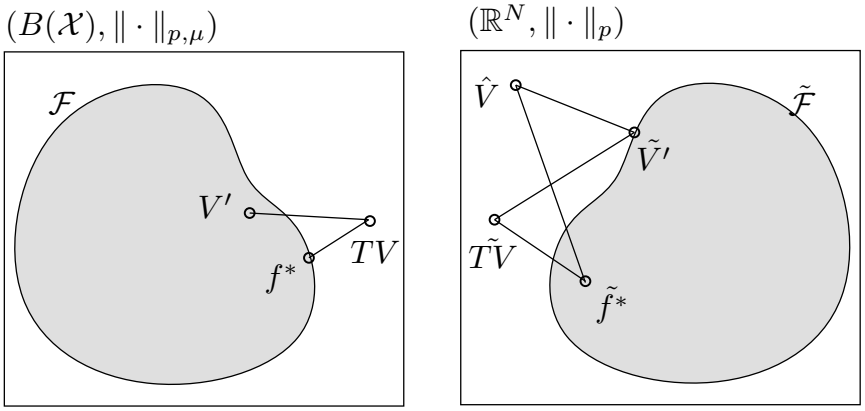

Figure 1: Illustration of the proof of Lemma 1 for bounding the distance of V' and TV in terms of the distance of f^* and TV, where f^* is the best fit to TV in $\mathcal{F}$. On the figure, V is a value function and V' is the result of minimizing the empirical loss, i.e., $V' = \operatorname{argmin}_{f\in\mathcal{F}} \|\widetilde{f} - \hat{V}\|_p^p$ (cf. Equations (3), (4)). For a function $f \in B(\mathcal{X})$, $\widetilde{f} = (f(X_1), \ldots, f(X_N))^T \in \mathbb{R}^N$. The set $\widetilde{\mathcal{F}}$ is defined by $\{\widetilde{f} \,|\, f \in \mathcal{F}\}$. Segments on the figure connect objects whose distances are compared in the proof.

of bounded functions over $\mathcal{X}$, whilst the right-hand side sub-figure shows a corresponding vector space. The connection between these two spaces is given by a mapping that maps functions $f \in B(\mathcal{X})$ to $\mathbb{R}^N$ using the mapping $f \mapsto (f(X_1), \ldots, f(X_N))^T$. Under this mapping, images of objects from the left-hand side are denoted by putting a tilde on the top of the corresponding symbol, i.e., $\widetilde{f} = (f(X_1), \ldots, f(X_N))^T$, $\widetilde{\mathcal{F}} = \{\widetilde{f} \,|\, f \in \mathcal{F}\}$, etc. The key observation is that by the law of large numbers, the distances of corresponding objects on the left-hand side and right-hand-side are expected to be close to each other when N is large, and if M is large then for any fixed i, $\hat{V}(X_i)$ will be close to $(TV)(X_i)$.

This observation is exploited as follows in the proof of the lemma: Our goal is to upper-bound the distance between V' and TV in terms of the distance between f^* and TV, where f^* is the best fit to TV in $\mathcal{F}$. (Remember that V' is the best fit in $\mathcal{F}$ to the data $(X_i, \hat{V}(X_i))_{i=1,\ldots,N}$ with respect to the p-norm $\|\cdot\|_p$). This is done by relating a series of distances to each other.

In particular, if N is large then $\|V' - TV\|_{p,\mu}^p$ and $\|\widetilde{V'} - \widetilde{TV}\|_p^p$ are expected to be close to each other. On the other hand, if M is large then $\hat{V}$ and $\widetilde{TV}$ are expected to be close to each other. Hence, $\|\widetilde{V'} - \widetilde{TV}\|_p^p$ and $\|\widetilde{V'} - \hat{V}\|_p^p$ are expected to be close to each other.

Since $\widetilde{V'}$ is the best fit to $\hat{V}$ in $\widetilde{\mathcal{F}}$, the distance between $\widetilde{V'}$ and $\hat{V}$ is not larger than the distance between the image $\widetilde{f}$ of an arbitrary function $f \in \mathcal{F}$ and $\hat{V}$. Choosing $f = f^*$ we conclude that the distance between $\widetilde{f^*}$ and $\hat{V}$ is not smaller than $\|\widetilde{V'} - \hat{V}\|_p^p$.

Exploiting again that M is large, we see that the distance between $\widetilde{f^*}$ and $\hat{V}$ must be close to that between $\widetilde{f^*}$ and $\widetilde{TV}$, which in turn must be close to the $L^p(\mathcal{X};\mu)$ distance between f^* and TV, provided that N is big enough. Hence, if $\|f^* - TV\|_{p,\mu}^p$ is small then so is $\|V' - TV\|_{p,\mu}^p$, finishing the outline of the proof.

4.2 Finite-sample Bounds for Random Functions

When an independent set of random samples is drawn in each iteration step (i.e., if we consider the multi-sample variant of sampling-based FVI) then due to the independence of samples used in subsequent iterations, Lemma 1 can be directly used to bound

$$\mathbb{P}\left(\|V_{k+1} - TV_k\|_{p,\mu} > d_{p,\mu}(TV_k, \mathcal{F}) + \epsilon | D_k\right)$$

thanks to

$$\mathbb{P}\left(\|V_{k+1} - TV_k\|_{p,\mu} > d_{p,\mu}(TV_k, \mathcal{F}) + \epsilon\right) = \mathbb{E}\left[\mathbb{P}\left(\|V_{k+1} - TV_k\|_{p,\mu} > d_{p,\mu}(TV_k, \mathcal{F}) + \epsilon | D_k\right)\right].$$

Here D_k denotes the samples drawn before the samples drawn to compute V_{k+1}.

In the single-sample variant, on the other hand, the same random sample is used in all the iterations. Then the above reasoning does not go through: in this case V_{k+1} is D_k-measurable and hence $\mathbb{P}\left(\|V_{k+1} - TV_k\|_{p,\mu} > \epsilon | D_k\right) = \mathbb{I}_{\{\|V_{k+1} - TV_k\|_{p,\mu} > \epsilon\}}$! Our next result extends Lemma 1 to this case by allowing the function V to be random. The approach that we will follow in the proof is similar to that used to derive a bound on the estimation error of data-dependent (random) functions. We need the following definition before the presentation of the result: Let

$$\mathcal{F}_{T-} = \{\, f - Tg \,|\, f \in \mathcal{F}, g \in \mathcal{F} \,\}$$

and denote by Ω the sample-space underlying the random variables $\{X_i\}$, $\{Y_j^{X_i,a}\}$, $\{R_j^{X_i,a}\}$, $i = 1, \ldots, N, j = 1, \ldots, M, a \in \mathcal{A}$. The analogue of Lemma 1 for random functions is as follows:

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

This lemma is proved in Appendix A.1. The bounds on the sample sizes are similar to those obtained in Lemma 1, except that N now depends on the metric entropy of $\mathcal{F}_{T-}$ and M depends on the metric entropy of $\mathcal{F}$. We now give two examples when explicit bounds on the covering number of $\mathcal{F}_{T-}$ can be given:

For the first example note that if $g : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ is Lipschitz with Lipschitz constant G then the ϵ-covering number of the space of functions of the form $h(x) = g(f_1(x), f_2(x))$, $f_1 \in \mathcal{F}_1$, $f_2 \in \mathcal{F}_2$ can be bounded by $\mathcal{N}(\epsilon/(2G), \mathcal{F}_1, n, \mu)\mathcal{N}(\epsilon/(2G), \mathcal{F}_2, n, \mu)$ (this follows directly from the definition of covering numbers). Since $g(x, y) = x - y$ is Lipschitz with $G = 1$, it suffices to bound the covering number of the space $\mathcal{F}_T = \{Tf | f \in \mathcal{F}\}$. One possibility to do this is as follows: Assume that $\mathcal{X}$ is compact, $\mathcal{F} = \{f_\theta | \theta \in \Theta\}$, Θ is compact and the mapping $H : (\Theta, \|\cdot\|) \to (B(\mathcal{X}), L^\infty)$ defined by $H(\theta) = f_\theta$ is Lipschitz with coefficient L. Fix $x^{1:n}$ and consider $\mathcal{N}(\epsilon, \mathcal{F}_T(x^{1:n}))$. Let θ_1, θ_2 be arbitrary. Then $|Tf_{\theta_1}(x) - Tf_{\theta_2}(x)| \le \|Tf_{\theta_1} - Tf_{\theta_2}\|_\infty \le \gamma \|f_{\theta_1} - f_{\theta_2}\|_\infty \le \gamma L \|\theta_1 - \theta_2\|$. Now assume that $C = \{\theta_1, \dots, \theta_m\}$ is an $\epsilon/(L\gamma)$-cover of the space Θ and consider any $n \ge 1$, $(x_1, \dots, x_n) \in \mathcal{X}^n$, $\theta \in \Theta$. Let θ_i be the nearest neighbour of θ in C. Then, $\left\|(Tf_\theta)(x^{1:n}) - (Tf_{\theta_i})(x^{1:n})\right\|_1 \le n \|Tf_\theta - Tf_{\theta_i}\|_\infty \le n\epsilon$. Hence, $\mathcal{N}(\epsilon, \mathcal{F}_T(x^{1:n})) \le \mathcal{N}(\epsilon/(L\gamma), \Theta)$.

Note that the mapping H can be shown to be Lipschitzian for many function spaces of interest. As an example let us consider the space of linearly parameterized functions taking the form $f_\theta = \theta^T \phi$ with a suitable basis function $\phi : \mathcal{X} \to \mathbb{R}^{d_\phi}$. By the Cauchy-Schwarz inequality, $\left\|\theta_1^T \phi - \theta_2^T \phi\right\|_\infty = \sup_{x \in \mathcal{X}} |\langle \theta_1 - \theta_2, \phi(x) \rangle| \le \|\theta_1 - \theta_2\|_2 \sup_{x \in \mathcal{X}} \|\phi(x)\|_2$. Hence, by choosing the ℓ^2 norm in the space Θ, we get that $\theta \mapsto \theta^T \phi$ is Lipschitz with coefficient $\left\|\|\phi(\cdot)\|_2\right\|_\infty$ (this gives a bound on the metric entropy that is linear in d_ϕ).

Another possibility to bound the 1-norm covering number of $\mathcal{F}_T$ is as follows: Let $x^{1:n}$ be fixed and consider an ϵ-cover of $\mathcal{F}(x^{1:n})$. Let the appropriate balls be $B_1, \dots, B_m$ with respective centers $c_1, \dots, c_m$. Clearly, $\{T(B_i)\}_{i=1,\dots,n}$ forms a cover of $\mathcal{F}_T(x^{1:n})$ and so does the system of minimum enclosing balls of TB_i. If we knew a bound on the radii of these balls, say $\alpha\epsilon$, then by a trivial rescaling argument we get $\mathcal{N}(\epsilon, \mathcal{F}_T(x^{1:n})) \le \mathcal{N}(\epsilon/\alpha, \mathcal{F}(x^{1:n}))$.

In order to derive a bound on the radii of the balls $\{TB_i\}_i$, pick any of them, say B_i and let $f \in B_i$, $c = c_i$. Then

$$\Delta = \sum_{i=1}^n |(Tf)(x_i) - (Tc)(x_i)| \le \gamma \sum_{i=1}^n \max_{a \in \mathcal{A}} \int |f(y) - c(y)| P(dy|x_i, a).$$

Assume that P satisfies the following locality condition

$$\int |g(y)| P(y|x, a) \le Lg(x),$$

for any $g \in \mathcal{F}$, $x \in \mathcal{X}$, $a \in \mathcal{A}$. Then $\Delta \le \gamma L \sum_{i=1}^n |f(x_i) - c(x_i)|$, hence $\alpha = \gamma L$. Clearly, for the locality condition to be satisfied it is critical for the functions in $\mathcal{F}$ to be smooth. If $\mathcal{F}$ contains only Λ-Lipschitz, uniformly bounded functions with uniform Lipschitz-coefficient $\Lambda > 0$ (this is satisfied automatically if $\mathcal{F}$ is a parametric class with a bounded weight-space) and a smooth parameterization then the locality condition is satisfied if for any x, a, the overwhelming mass of $P(\cdot|x, a)$ is concentrated in the $O(1/\Lambda)$ neighbourhood of x. If P, r are Lipschitz with coefficients L_p and L_r, respectively then the image space of the Bellman-operator T contains only $L_r + \gamma L_p$ Lipschitz functions. In this case it is reasonable to restrict $\mathcal{F}$ to $\Lambda = L_r + \gamma L_p$ Lipschitz functions. The locality assumption is reasonable when the MDP is obtained from an approximation to a continuous-time system through the discretization of time – a situation often found in finance applications.[7]

7. A time discretization of a (stochastic) differential equation with timestep Δt leads to a kernel $P(\cdot|x, a)$ whose overwhelming mass is concentrated in the $O((\Delta t)^{1/2})$ neighbourhood of x in the stochastic case

5. Main Results

For the sake of specificity, let us reiterate the algorithms. Let $V_0 \in \mathcal{F}$. Then the *single-sample variant* of sampling-based FVI produces a sequence of function $\{V_k\}_{0\le k\le K} \subset \mathcal{F}$ satisfying

$$V_{k+1} = \operatorname*{argmin}_{f\in\mathcal{F}} \sum_{i=1}^{N} \left| f(X_i) - \max_{a\in\mathcal{A}} \frac{1}{M} \sum_{j=1}^{M} \left[R_j^{X_i,a} + \gamma V_k(Y_j^{X_i,a}) \right] \right|^p. \tag{10}$$

The *multi-sample variant* is obtained by using a fresh set of samples in each iteration:

$$V_{k+1} = \operatorname*{argmin}_{f\in\mathcal{F}} \sum_{i=1}^{N} \left| f(X_i^k) - \max_{a\in\mathcal{A}} \frac{1}{M} \sum_{j=1}^{M} \left[R_j^{X_i^k,a,k} + \gamma V_k(Y_j^{X_i^k,a,k}) \right] \right|^p. \tag{11}$$

Let π_k be a greedy policy w.r.t. V_k. We are interested in bounding the loss due to using policy π_k and not an optimal one. In addition to the case when the loss is measured w.r.t. the supremum norm, we also consider the case when the loss is measured by a weighted p-norm:

$$L_k = \|V^* - V^{\pi_k}\|_{p,\rho}.$$

Here ρ is a distribution whose role is to put more weight to those parts of the state space where performance matters more. By using a weighted p-norm instead of the supremum norm we may avoid an overly pessimistic assessment of the performance of the algorithm. Note also that at the expense of posing more conditions on the MDP, we will still get some bounds on the supremum-norm loss, too. A particularly sensible choice is to set ρ to be the distribution over the states from which we start to use π_k. In this case if $p = 1$ then L_k measures the expected loss. For $p > 1$ the loss does not have a similarly simple interpretation, except that with $p \to \infty$ we recover the supremum norm loss, hence increasing p generally means that the evaluation becomes more pessimistic.

Let us now discuss how we arrive at a bound on the expected p-norm loss. By the results of the previous section we have a bound on the error introduced in any given iteration. Hence, all we need to show is that the errors stay controlled as they are propagated through the algorithm. Previously, such an analysis was primarily carried out by means of some contraction arguments. Contraction arguments, however, are not applicable since the previous section's bound is given in terms of weighted p-norms, whilst contraction arguments require supremum-norm bounds on the individual errors.

In our analysis we develop a different, though related argument. Firstly, we develop pointwise bounds. Pointwise bounds are advantageous since they naturally lead to norm bounds for any norm (i.e., also weighted p-norm bounds). One difficulty of working with the Bellman operator is that it is nonlinear and hence bounding $TV_1 - TV_2$ in terms of $V_1 - V_2$ looks difficult (max and substraction do not commute). Now, observe that for any k, TV_k

(this is usually the case e.g. in finance) and in the $O(\Delta t)$ neighbourhood of x in the deterministic case. In both cases, for a small enough Δt, the locality condition holds. TODO/cs/: Don't we need any conditions, e.g. smoothness?? /rm/ For a SDE $dX_t = f(X_t, u_t)dt + \sigma(X_t, u_t)dW_t$ with a uniform elliptic condition (ie. $\sigma\sigma' \ge \lambda I$, where $\lambda > 0$) on the noise term, then the $O((\Delta t)^{1/2})$ neighbourhood holds, otherwise, an $O(\Delta t)$ neighbourhood only (at least in some direction) holds. But I don't think we need to go in these details here.

can be equivalently written as $T^{\pi_k}V_k$, where π_k is the greedy policy w.r.t. V_k. Further, for an optimal policy π^*, we have $TV^* = T^{\pi^*}V^*$. Here the operators T^{π_k}, T^{π^*} are affine linear operators that are dependent on the linear operators P^{π_k} and P^{π^*}. Now, we can upper bound $V^* - V_{k+1}$ by $\gamma P^{\pi^*}(V^* - V_k) + \epsilon_k$ and lower bound it by $\gamma P^{\pi_k}(V^* - V_k) + \epsilon_k$, where $\epsilon_k = TV_k - V_{k+1}$ is the error introduced in the k-th step, where we exploited that linear operators commute with substraction. Ultimately since P^{π_k} commutes with multiplication by constants, too, we can collect exponents of γ and terms involving the stochastic kernels P^{π_k}. It is not hard then to foresee that as the errors are propagated through the iterations, future-state distributions of the form $P^{\pi_K}P^{\pi_{K-1}}\dots P^{\pi_k}$ enter our bounds.

However, since we have no non-trivial upper bound on $\|\epsilon_k\|_\infty$, instead of continuing this way, we take the pth power and integrate the pointwise bounds with respect to ρ. In order to derive a compact bound, we need a bound on the induced $L^p(\rho)$-norm of product operators of the above form, i.e., the maximum amplification such product operators can possibly achieve when "size" is measured with $L^p(\rho)$-norms. Note that if a supremum-norm analysis were followed ($p = \infty$), we would immediately find that the bound on the maximum amplification is 1 as $|\int V(y)P(dy|x,\pi(x))| \le \int |V(y)|P(dy|x,\pi(x)) \le \|V\|_\infty \int P(dy|x,\pi(x)) = \|V\|_\infty$, or $\|P^\pi V\|_\infty \le \|V\|_\infty$. In fact, at this point it is still possible to obtain the well-known supremum-norm bounds. As a similar inequality does not hold for weighted p-norms, we seek for conditions on the MDP that allow us to get bounds on the operator norm of these product operators. One simple assumption that allows us to arrive at such a conclusion is the following:

Assumption A1 [Uniformly stochastic transitions] For all $x \in \mathcal{X}$ and $a \in \mathcal{A}$, assume that $P(\cdot|x,a)$ is absolutely continuous w.r.t. μ and the Radon-Nikodym derivative of P w.r.t. μ is bounded uniformly with bound C_μ:

$$C_\mu \stackrel{\text{def}}{=} \sup_{x\in\mathcal{X}, a\in\mathcal{A}} \frac{dP(\cdot|x,a)}{d\mu} < +\infty.$$

Assumption A1 can be written in the form $P(\cdot|x,a) \le C_\mu \mu(\cdot)$, an assumption that was introduced by Munos (2003) in a finite MDP context for the analysis of approximate policy iteration. Clearly, if Assumption A1 holds then for $p \ge 1$, by Jensen's inequality, $|\int V(y)P(dy|x,\pi(x))|^p \le \int |V(y)|^p P(dy|x,\pi(x)) \le \int C_\mu |V(y)|^p d\mu(dy)$, hence $\|P^\pi V\|_{p,\rho} \le C_\mu^{1/p}\|V\|_{p,\mu}$. Note that when μ is the Lebesgue-measure over $\mathcal{X}$ then Assumption A1 becomes equivalent to assuming that the transition probability kernel $P(dy|x,a)$ admits a uniformly bounded density. Note that the "noisier" the dynamics is, the smaller is the constant C_μ. Although $C_\mu < +\infty$ looks like a strong restriction, the class of MDPs that admit this restriction is still very big in the sense that there are hard instances in it (this is discussed at some length in Section 8). However, this assumption certainly excludes completely or partially deterministic MDPs.

We now consider another assumption that allows for such systems, too. The idea is that for the analysis we only need to be able to reason about the operator norms of weighted sums of the product of arbitrary stochastic kernels. This motivates the following definition:

Assumption A2 [Bounded discounted-average concentrability of future-state distributions] Given ρ, μ, $m \ge 1$ and an arbitrary sequence of stationary policies $\{\pi_m\}_{m\ge 1}$, assume that

the future-state distribution $\rho P^{\pi_1} P^{\pi_2} \dots P^{\pi_m}$ is absolutely continuous w.r.t. μ. Moreover, with

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

the *discounted-average concentrability coefficient* of the future-state distributions,

$$C_{\rho,\mu} = (1-\gamma)^2 \sum_{m \geq 1} m \gamma^{m-1} c(m),$$

is finite.

Note that $c(m)$, called the m-step concentrability of a future-state distribution, measures how much ρ can get amplified in m steps as compared to the reference distribution μ. Thanks to discounting, $C_{\rho,\mu}$ is finite for a reasonably large class of systems. In fact $C_{\rho,\mu} < \infty$ holds when the so-called *top-Lyapunov exponent* of the system is non-positive, where, using our notation, the top-Lyapunov exponent of the MDP could be defined as $\hat{c} = \limsup_{m\to\infty} \frac{1}{m} \log c(m)$.[8] Top-Lyapunov exponents play a fundamental role in the modern stability analysis of stochastic systems. In fact, the existence of strictly stationary non-anticipating realizations to stochastic recursive equations is often associated with the negativity of the associated top-Lyapunov exponent (Bougerol and Picard, 1992; Whang and Linton, 1999).[9]

Let us now show that Assumption A2 is indeed weaker than Assumption A1. Clearly, since $C_{\rho,\mu}$ is just a weighted average of the m-step concentrability coefficients, $c(m)$, it suffices to show that $c(m) \leq C_\mu$ holds for any m. To see this observe that by definition for any distribution ν and policy π, $\nu P^\pi \leq C_\mu \mu$. Now take $\nu = \rho P^{\pi_1} \dots P^{\pi_{m-1}}$ and $\pi = \pi_m$ to conclude that $\rho P^{\pi_1} \dots P^{\pi_{m-1}} P^{\pi_m} \leq C_\mu \mu$ and so $c(m) \leq C_\mu$.

Since Assumption A1 is stronger than Assumption A2 in the proofs we will proceed by first proving the statement under Assumption A2 since then the result will follow automatically under Assumption A1. The reason Assumption A1 is still considered is that it actually allows one to arrive at supremum-norm performance bounds.

As a final preparatory step before the presentation of our main results, let us define the *inherent Bellman error* associated with the function space $\mathcal{F}$ by

$$d_{p,\mu}(T\mathcal{F}, \mathcal{F}) = \sup_{f \in \mathcal{F}} d_{p,\mu}(Tf, \mathcal{F}).$$

Note that $d_{p,\mu}(T\mathcal{F}, \mathcal{F})$ generalizes the notion of Bellman errors to function spaces in the natural way. As we have seen the error in iteration k depends on $d_{p,\mu}(T\hat{V}_k, \mathcal{F})$. Since $\hat{V}_k \in \mathcal{F}$, we see that the inherent Bellman error gives a uniform bound on the errors of the individual iterations.

The next theorem, the main result of the paper, states that with high probability the final performance of the policy found by the algorithm can be made as close to the inherent Bellman error associated with the function space $\mathcal{F}$ as desired by selecting a sufficiently high number of samples and hence AVI can be used to find near-optimal policies:

8. In fact, an elementary analysis shows that if $\hat{c} < \log(1/\gamma)$ then $\sum_{m\geq 0} m^p \gamma^m c(m) < \infty$. Hence if the top-Lyapunov exponent of the MDP is non-positive then $C_{\rho,\mu} < \infty$.
9. The lack of existence of such solutions would probably preclude any sample-based estimation of the system.

Theorem 2 *Consider an MDP satisfying Assumption A0 and A2. Fix $p \geq 1$, $\mu \in M(\mathcal{X})$ and let $V_0 \in \mathcal{F} \subset B(\mathcal{X}; V_{\max})$. Let $\{V_k\}_{k=1}^K$ denote the iterates generated by the multi-sample variant of sampling-based FVI with $X_i \sim \mu$. Then for any $\epsilon, \delta > 0$, there exist integers K, M and N such that K is linear in $\log(1/\epsilon)$, $\log V_{\max}$ and $\log(1/(1-\gamma))$, N, M are polynomial in $1/\epsilon$, $\log(1/\delta)$, $\log(1/(1-\gamma))$, $V_{\max}$, $\hat{R}_{\max}$, $\log(|\mathcal{A}|)$, $\log(\mathcal{N}(c\epsilon(1-\gamma)^2/(C_{\rho,\mu}^{1/p}\gamma), \mathcal{F}, N, \mu))$ for some constant $c > 0$, such that with probability at least $1-\delta$,*

$$\|V^* - V^{\pi_K}\|_{p,\rho} \leq \frac{2\gamma}{(1-\gamma)^2} C_{\rho,\mu}^{1/p} \, d_{p,\mu}(T\mathcal{F}, \mathcal{F}) + \epsilon.$$

If Assumption A1 holds as well then with probability at least $1-\delta$

$$\|V^* - V^{\pi_K}\|_{\infty} \leq \frac{2\gamma}{(1-\gamma)^2} C_{\mu}^{1/p} \, d_{p,\mu}(T\mathcal{F}, \mathcal{F}) + \epsilon$$

holds, too, with $C_{\rho,\mu}$ replaced by C_μ in the bounds on N, M.

Further, the results continue to hold for the single-sample variant of sampling-based FVI with the exception that N depends on $\log(\mathcal{N}(c\epsilon, \mathcal{F}_{T-}, N, \mu))$ and M depends on $\log(\mathcal{N}(c'\epsilon, \mathcal{F}, M, \mu))$ for appropriate $c, c' > 0$.

The proof is given in Appendix B. With some extra work, by a careful inspection of the proof and assuming that the pseudo-dimension of the function-space $\mathcal{F}$ is finite as in Proposition 1, it is possible to derive the following high-probability bound on the finite-sample performance of the multi-sample variant of FVI:

$$\begin{aligned}\|V^* - V^{\pi_K}\|_{p,\rho} \leq{}& \frac{2\gamma}{(1-\gamma)^2} C_{\rho,\mu}^{1/p} \, d_{p,\mu}(T\mathcal{F}, \mathcal{F}) + O(\gamma^K V_{\max}) \\ &+ \; O\left\{ \left(\frac{V_{\mathcal{F}^+}}{N} \left(\log(N) + \log(K/\delta)\right) \right)^{1/2p} + \left(\frac{1}{M} \left(\log(N|\mathcal{A}|) + \log(K/\delta)\right) \right)^{1/2} \right\}. \quad (12)\end{aligned}$$

Here N, M, K are arbitrary integers and the bound holds with probability $1-\delta$. The first term bounds the approximation error, the second arises due to the finite number of iterations, whilst the last two terms bound the estimation error.

This form of the bound is useful as it allows us to investigate problems like the optimal choice of N and M given a fixed budget of $n = K \times N \times M$ samples (or $\hat{n} = N \times M$ samples per iteration). Indeed, optimizing the bound yields that the best choice of N and M is given by $N = (V_{\mathcal{F}^+})^{1/(p+1)} \hat{n}^{p/(p+1)}$, $M = (\hat{n}/V_{\mathcal{F}^+})^{1/(p+1)}$, resulting in the bound $(n/(KV_{\mathcal{F}^+}))^{-1/(2p+2)}$ for the estimation error (disregarding logarithmic terms). (The choice of N, M does not influence the other error terms).

Now, let us consider the single-sample variant of FVI. Again, a careful inspection of the proof results in an inequality identical to (12) just with the pseudo-dimension of $\mathcal{F}$ replaced by the pseudo-dimension of $\mathcal{F}^-$ and $1/M$ replaced by $V_{\mathcal{F}^+}/M$. We may again ask the question of how to choose N, M, given a fixed-size budget of $n = N \times M$ samples. The formulae are similar to the previous ones. The resulting optimized bound on the estimation error is $(n/(V_{\mathcal{F}^-}V_{\mathcal{F}^+}))^{-1/(2p+2)}$. It follows that that if $K > V_{\mathcal{F}^-}$ and given a fixed budget of n samples the bound for the single-sample variant is better than for the multi-sample variant, where K could also be found by optimizing the bounds. When choosing K, we

have to balance the two terms where K appears in the bound. In fact in both cases $K \sim 1/\log(1/\gamma) \cong 1/(1-\gamma)$. Hence as γ approaches one, the single-sample variant of FVI can be expected to become more efficient, provided that everything else is kept the same. It is interesting to note that as γ becomes larger the number of times the samples are reused increases, too. One way to understand this result is that although the increased sample size might increase bias, it reduces the variance. We will get back to this observations in the section where we discuss the simulations.

Another way to use the above bound is to make comparisons with the rates available in non-parametric regression: The bounds in non-parametric regression have two terms, one that bounds the approximation error of the function class $\mathcal{F}$, and one that bounds the *estimation error*. Our bound, due to the specific properties of the problem, the approximation error of $\mathcal{F}$ is defined as the inherent Bellman error of $\mathcal{F}$. This seems reasonable since we are trying to find an approximate fixed point of T within $\mathcal{F}$. Let us turn our attention to the estimation error bounds. We know from the literature that for a total number of n samples the optimal rate for the estimation error in regression is $n^{-1/p}$ (Györfi et al., 2002). The rates obtained here have two terms. As before, optimizing the bound whilst keeping $n = N \times M$ fixed, we get the inferior rate $n^{-1/(2(p+1))}$. With considerably more work, using the techniques of Lee et al. (1996) (see also Chapter 11 of (Györfi et al., 2002)) it is possible to improve the exponent of N from $-1/2p$ to $-1/p$ in Equation 12. However, when doing so the factor multiplying the inherent Bellman error (playing the role of the approximation error) will be increased (the same happens in non-parametric regression) and thus the bound improves only when the estimation error dominates the first of the two terms. In this case, the optimized rate becomes $n^{-1/(p+1)}$. This is still slightly worse than the best possible rate for non-parameteric regression. The additional $1/n$ factor comes from the term in Equation 12 that depends on M. Thus, at least for FVI, the inferior rate as compared with regression seems unavoidable: we need $M \to \infty$ in order to control the bias of the target values used in the embedded regression problems, whilst in a pure regression setup, by definition one works with unbiased samples and M there would only control the variance of the dependent variable and not its bias. In such a case the optimal choice is $M = 1$: if we increase N whilst keeping $M = 1$ we gain more information than if we were to increase M, too. By switching from state-based value-iteration to value-iteration for action-value functions it seems possible to get rid of this second term. However, we leave the investigation of this for further work.

6. Randomized Policies

The previous result shows that by making the inherent Bellman error of the function space small enough, we can ensure a close-to-optimal performance if one uses the policy greedy with respect to the last value-function estimate, V_K. However, by definition this greedy policy depends on the unknown MDP. In this section we show that by computations analogous to that used in obtaining the iterates we can compute a randomized near-optimal policy.

Let us call an action a *α-greedy* w.r.t. the function V and state x, if

$$r(x,a) + \gamma \int V(y) P(dy|x,a) \geq (TV)(x) - \alpha.$$

Given V_K and a state $x \in \mathcal{X}$ we can use sampling to determine an α-greedy action with high probability: Let $\alpha, \lambda > 0$ be two parameters, to be selected later. Given any state $x \in \mathcal{X}$, for each action $a \in \mathcal{A}$, produce $M' = M'(\alpha, \lambda)$ reward-next-state pair samples, $(R_j^{x,a}, Y_j^{x,a})$, using the simulation model: $R_j^{x,a} \sim S(\cdot, x, a)$, $Y_j^{x,a} \sim P(\cdot|x, a)$. Based on these samples compute the approximate value of a at state x:

$$Q_{M'}(x, a) = \frac{1}{M'} \sum_{j=1}^{M'} \left[R_j^{x,a} + \gamma V_K(Y_j^{x,a}) \right].$$

Let the policy $\pi_{\alpha,\lambda}^K : \mathcal{X} \to \mathcal{A}$ select actions maximizing these approximate values:

$$\pi_{\alpha,\lambda}^K(x) = \arg\max_{a \in \mathcal{A}} Q_{M'}(x, a).$$

The following result is proved in Appendix C:

Theorem 3 *Consider an MDP satisfying Assumptions A0 and A2. Fix $p \geq 1$, $\mu \in M(\mathcal{X})$ and let $V_0 \in \mathcal{F} \subset B(\mathcal{X}; V_{\max})$ and let $\{V_k\}_{k=1}^K$ be the iterates generated by multi-sample FVI with $X_i \sim \mu$. Select $\alpha = (1-\gamma)\epsilon/8$, $\lambda = \epsilon/8(1-\gamma)/V_{\max}$ and let $M' = O(|\mathcal{A}|\hat{R}_{\max}^2 \log(|\mathcal{A}|/\lambda)/\alpha^2)$. Then, for any $\epsilon, \delta > 0$, there exist integers K, M and N such that K is linear in $\log(1/\epsilon)$, $\log V_{\max}$ and $\log(1/(1-\gamma))$, N, M are polynomial in $1/\epsilon$, $\log(1/\delta)$, $1/(1-\gamma)$, $V_{\max}$, $\hat{R}_{\max}$, $\log(|\mathcal{A}|)$, $\log(\mathcal{N}(c\epsilon(1-\gamma)^2/C_{\rho,\mu}^{1/p}), \mathcal{F}, \mu)))$ for some $c > 0$ and the policy $\pi_{\alpha,\lambda}^K$ defined above satisfies*

$$\left\| V^* - V^{\pi_{\alpha,\lambda}^K} \right\|_{p,\mu} \leq \frac{4\gamma}{(1-\gamma)^2} C_{\rho,\mu}^{1/p} d_{p,\mu}(T\mathcal{F}, \mathcal{F}) + \epsilon$$

with probability at least $1 - \delta$.

An entirely analogous result holds for the supremum-norm loss under Assumptions A0 and A1 with $C_{\rho,\mu}$ replaced by C_μ.

Similarly, the theorem can be stated for the single-sample variant with the obvious modifications. We note that in place of the above uniform sampling sampling model one could also use the Median Elimination Algorithm of Even-Dar et al. (2002), resulting in a reduction of M' by a factor of $\log(|\mathcal{A}|)$. However, for the sake of compactness we do not explore this option here.

7. Asymptotic Consistency

The first and weakest property for an algorithm designed to solve MDPs is that, as the sample grows, it should converge to achieve zero loss, i.e., the algorithm should be consistent. Sampling based FVI with a fixed function space $\mathcal{F}$ is not consistent: As it follows from our previous results, the loss converges to $2\frac{\gamma}{(1-\gamma)^2} C_{\rho,\mu}^{1/p} d_{p,\mu}(T\mathcal{F}, \mathcal{F})$. In order to make this residual term disappear we must increase $\mathcal{F}$ with the number of samples. The resulting method in regression is called the "method of sieves" and was proposed there by Grendander (1981). We chose this method since, despite its simplicity, it illustrates very well the fundamental tradeoffs involved in designing a consistent procedure.

In the result below we shall assume that both the reward and transition probabilities underlying the unknown MDP are "smooth" in a sense that if an initial state is perturbed a little then both the expected reward and the probability kernel change only by a little amount:

$$\begin{aligned}\max_{a\in\mathcal{A}} \sup_{B\subset\mathcal{X},x,x'\in\mathcal{X}} |P(B|x,a)-P(B|x',a)| &\le L_P \left\|x-x'\right\|^\alpha,\\ \max_{a\in\mathcal{A}} \sup_{x,x'\in\mathcal{X}} |r(x,a)-r(x',a)| &\le L_r \left\|x-x'\right\|^\alpha.\end{aligned}$$

Here $\alpha, L_p, L_r > 0$ are parameters that depend on the unknown MDP. These factors are typically unknown and are in general hard to estimate.[10] Hence, we seek a procedure that is asymptotically consistent and does not require the knowledge of L_p and L_r or α.

It is well known that if the MDP is smooth in the above sense and if $V \in B(\mathcal{X})$ is uniformly bounded by $V_{\max}$ then TV is $L = (L_r + \gamma V_{\max} L_P)$-Lipschitzian:

$$|(TV)(x) - (TV)(x')| \le (L_r + \gamma V_{\max} L_P)|x-x'|^\alpha, \quad \forall x, x' \in \mathcal{X}.$$

Now, if $\mathcal{F}_n$ is restricted to contain $V_{\max}$-bounded functions only then $T\mathcal{F}_n = \{TV \,|V \in \mathcal{F}_n\}$ will only contain L-Lipschitz functions:

$$T\mathcal{F}_n \subset \mathrm{Lip}(\alpha; L, V_{\max}) \stackrel{\mathrm{def}}{=} \{f \in B(\mathcal{X}) \,| \quad \|f\|_\infty \le V_{\max}, |f(x)-f(y)| \le L\,\|x-y\|^\alpha\,\}.$$

By the definition of $d_{p,\mu}$,

$$d_{p,\mu}(T\mathcal{F}_n, \mathcal{F}_n) \le d_{p,\mu}(\mathrm{Lip}(\alpha; L, V_{\max}), \mathcal{F}_n)$$

and hence if we make the right hand-side converge to zero then so will $d_{p,\mu}(T\mathcal{F}_n, \mathcal{F}_n)$. The quantity, $d_{p,\mu}(\mathrm{Lip}(\alpha; L, V_{\max}), \mathcal{F}_n)$, on the other hand, is just the approximation error of functions in the Lipschitz class $\mathrm{Lip}(\alpha; L, V_{\max})$ by elements of $\mathcal{F}_n$. The behaviour of the approximation error of Lipschitz functions is one of the most widely studied problems in classical approximation theory. In fact, for a large number of approximation classes, $\{\mathcal{F}_n\}$ (e.g. approximation by polynomials, wavelets, function dictionaries), variants of Jackson's theorem show that $d_{p,\mu}(\mathrm{Lip}(\alpha; L), \mathcal{F}_n) = O(Ln^{-\alpha})$, (e.g DeVore, 1997). Here $\mathrm{Lip}(\alpha; L)$ is the class of (L, α)-Lipschitz functions where no uniform bound is assumed on the functions in it. Thanks to the compactness of $\mathcal{X}$, $\mathrm{Lip}(\alpha; L) = \cup_{V_{\max}>0} \mathrm{Lip}(\alpha; L, V_{\max})$. In approximation theory an approximation class $\{\mathcal{F}_n\}$ is said to be *universal* if for any $\alpha, L > 0$, $\lim_{n\to\infty} d_{p,\mu}(\mathrm{Lip}(\alpha; L), \mathcal{F}_n) = 0$.

One remaining issue is that classical approximation spaces are not uniformly bounded (i.e., the functions in them do not assume a uniform bound), whilst our previous argument showing that the image space $T\mathcal{F}_n$ is a subset of Lipschitz functions critically relies on that $\mathcal{F}_n$ is uniformly bounded. A simple solution is to replace the function space $\mathcal{F}_n$ by $\mathcal{C}_{V_{\max}}\mathcal{F}_n = \{\,\mathcal{C}_{V_{\max}} V \,|\, V \in \mathcal{F}_n\,\}$, where for $\mathcal{C}_{V_{\max}}$ is the truncation operator which is defined by

$$\mathcal{C}_A r = \begin{cases} \mathrm{sign}(r)A, & \text{if } |r| > A, \\ r, & \text{otherwise.}\end{cases}$$

10. In regression, it is possible to prove consistency without imposing any smoothness assumptions. Whether this is possible in (off-line) planning in discounted MDPs is an open question.

Now, if $\mathcal{F}_n$ is a universally consistent sequence then we also have $\lim_{n\to\infty} d_{p,\mu}(\mathrm{Lip}(\alpha;L) \cap B(\mathcal{X};V_{\max}), \mathcal{C}_{V_{\max}}\mathcal{F}_n) = 0$. Now, if for a sample-size of n, the multi-sample FVI procedure is applied to the function space $\mathcal{F}_n(V_{\max}) \stackrel{\text{def}}{=} \mathcal{C}_{V_{\max}}\mathcal{F}_n$ then by our previous analysis, the estimation error will scale by $(n/(KV^+_{\mathcal{F}_n(V_{\max})}))^{1/(p+1)}$. Hence, in order to make the error bound converge to zero all we need to ensure is that the pseudo-dimension of $\mathcal{F}_n(V_{\max})$ grows sublinearly as a function of n. This way we get the following corollary to Theorem 2:

Corollary 4 *Consider an MDP satisfying Assumptions A0 and A2 and assume that both its immediate reward function and transition kernel are Lipschitzian. Fix $p \geq 1$, $\mu \in M(\mathcal{X})$ and let $\{\mathcal{F}_n\}$, $\mathcal{F}_n \subset B(\mathcal{X};V_{\max})$, be a universal approximation class such that the pseudo-dimension of $\mathcal{F}_n(V_{\max})$ grows sublinearly in n. Then, for each $\epsilon, \delta > 0$ there exist an index n_0 such that for any $n \geq n_0$ there exist integers K, N, M that are polynomial in $1/\epsilon$,$\log(1/\delta)$, $1/(1-\gamma)$, $V_{\max}$, $\hat{R}_{\max}$, $\log(|\mathcal{A}|)$, and $V_{\mathcal{F}_n(V_{\max})^+}$ such that if V_K is the output of multi-sample FVI when it uses the function set $\mathcal{F}_n(V_{\max})$ and $X_i \sim \mu$ then $\|V^* - V^{\pi_K}\|_{p,\rho} \leq \epsilon$ holds with probability at least $1-\delta$. An identical result holds for $\|V^* - V^{\pi_K}\|_\infty$ when Assumption A2 is replaced by Assumption A1.*

The result extends to single-sample FVI as before.

One specific aspect in which this corollary is not satisfactory is that solving the optimization problem over $\mathcal{C}_{V_{\max}}\mathcal{F}_n$ is computationally more challenging than to do the same over $\mathcal{F}_n$ (think of the case when $\mathcal{F}_n$ is a class of linearly parameterized functions and $p = 2$). An idea, borrowed from the regression literature is to first do the optimization and then truncate the obtained functions. The resulting procedure can be shown to be consistent using the methods of e.g. Chapter 10 of the book of Györfi et al. (2002).

It is important to emphasize that the construction used in this section is just one example of how our main results lead to consistent algorithms. An immediate extension of the present work would be to achieve the best possible convergence rates for a given MDP e.g. by using penalized estimation. However, we leave this for future work.

8. Discussion of Related Work

The origins of FVI date back to the early days of dynamic programming. One of the earliest example is the work of Samuel who used both linear and non-linear methods to approximate value functions in his programs that learned to play the game of checkers (Samuel, 1959, 1967). At the same time, Bellman and Dreyfus (1959) explored the use of polynomials for accelerating dynamic programming. Both in these works and in most later works (e.g. Reetz, 1977; Morin, 1978) FVI with representative states was considered. Of these authors, only Reetz (1977) presents theoretical results who, on the other hand, considered only one-dimensional feature spaces. Recent empirical studies include the works by Longstaff and Shwartz (2001); Haugh (2003) and Jung and Uthmann (2004), amongst others.

The theoretical analysis of AVI generating the iterates $\{V_t\}$, $V_t \in \mathcal{F}$, is straightforward for discounted MPDs as long as the approximation errors, $\epsilon_t = V_{t+1} - TV_t$, stay bounded in the supremum norm. Then the worst-case performance-loss for the policy greedy w.r.t. the most recent iterates can be bounded asymptotically by $\frac{2\gamma}{(1-\gamma)^2} \sup_{t\geq 1} \|\epsilon_t\|_\infty$ (e.g. Bertsekas and Tsitsiklis, 1996). When V_{t+1} is the best approximation of TV_t in $\mathcal{F}$ then $\sup_{t\geq 1} \|\epsilon_t\|_\infty$ can be upper bounded by the inherent Bellman error $d_\infty(T\mathcal{F}, \mathcal{F}) =$

$\sup_{f\in\mathcal{F}}\inf_{g\in\mathcal{F}}\|g-Tf\|_\infty$ and we get the loss-bound $\frac{2\gamma}{(1-\gamma)^2}d_\infty(T\mathcal{F},\mathcal{F})$. Apart from the smoothness factors ($C_{\rho,\mu}$, C_μ) and the estimation errors, our loss-bounds have the same form (cf. Equation (12)). In particular, we get back the previous bounds when $p\to\infty$, since then for μ being the Lebesgue measure, $C_\mu^{1/p}d_{p,\mu}(T\mathcal{F},\mathcal{F})\to d_\infty(T\mathcal{F},\mathcal{F})$. However, this does not mean that in our algorithm p should be taken as large as possible since, for finite p, $C_\mu^{1/p}d_{p,\mu}(T\mathcal{F},\mathcal{F})$ could be significantly smaller than $d_\infty(T\mathcal{F},\mathcal{F})$. Another reason not to take large values of p is that the estimation errors can be expected to grow with p.

Another way to approach the design and analysis of AVI is to consider algorithms of the form $V_{t+1}=\Pi TV_t$. Here Π is a mapping that maps functions to the function space $\mathcal{F}$. This form was originally proposed by (Gordon, 1995; Tsitsiklis and Van Roy, 1996). The results in these papers considered the planning scenario with known dynamics and making use of representative states. Results for more general settings, including learning and infinite state spaces, were considered in Singh et al. (1995) and lately by Szepesvári and Smart (2004). The algorithm of Ormoneit and Sen (2002) also belongs to this class.

The main insight underlying these algorithms is that if the composite operator ΠT is a supremum-norm contraction then the iterates converge to some limit V_∞. Then the worst-case loss of using the policy greedy w.r.t. V_∞ is bounded by $\frac{4\gamma}{(1-\gamma)^2}\epsilon_\Pi$, where ϵ_Π is the best approximation to V^* by fixed points of Π: $\epsilon_\Pi=\inf_{f\in\mathcal{F}:\Pi f=f}\|f-V^*\|_\infty$ (e.g. Tsitsiklis and Van Roy, 1996, Theorem 2). In practice, the requirement that Π has to be a non-expansion in the supremum norm is satisfied by using "averagers" (Gordon, 1995).

Averagers represent a large class of methods which include k-nearest neighbors with fixed centers, kernel-regression, linear interpolation with a fixed set of basis functions or splines with fixed knots. They were defined by Gordon (1995) for finite spaces. In order to facilitate a comparison with our results we restate the original definition:

Definition 5 ("Averagers" for finite $\mathcal{X}$; Gordon (1995)) *Let $\mathcal{X}$ be finite. $\Pi:B(\mathcal{X})\to B(\mathcal{X})$ is an averager if there exist functions α, $\beta\geq 0$ s.t. for all functions f, $(\Pi f)(x)=\alpha(x)+\sum_{x'}\beta(x,x')f(x')$ and if it holds that for all $x\in\mathcal{X}$, $\sum_{x'}\beta(x,x')\leq 1$.*

Note that $\Pi_0 f\stackrel{\text{def}}{=}\Pi f-\alpha$ is linear in f: for any $\lambda_1,\lambda_2\in\mathbb{R}$ and $f_1,f_2\in B(\mathcal{X})$, it holds that $\Pi_0(\lambda_1 f_1+\lambda_2 f_2)=\lambda_1\Pi_0 f_1+\lambda_2\Pi_0 f_2$. Further, the condition on the sum of the values $\beta(x,\cdot)$ is implied if the supremum norm of Π_0 is not larger than one. In fact, since all we want to ensure is that Π is a non-expansion, we can generalize the previous definition as follows without losing the previous results:

Definition 6 (Generalized Averagers) *$\Pi:B(\mathcal{X})\to B(\mathcal{X})$ is an averager if $\Pi f=\alpha+\Pi_0 f$, where $\alpha\in B(\mathcal{X})$ and $\Pi_0:B(\mathcal{X})\to B(\mathcal{X})$ is a linear operator with $\|\Pi_0\|_\infty\leq 1$.*

In practice, we must restrict ourselves to finitely parameterized averagers:

Definition 7 (Finitely Parameterized Generalized Averagers) *$\Pi:B(\mathcal{X})\to B(\mathcal{X})$ is a finitely parameterized averager if Πf can be written as $\Pi f=\alpha+\sum_i(L_i f)\phi_i$, where $\alpha\in B(\mathcal{X})$, $L_i:B(\mathcal{X})\to\mathbb{R}$ are linear non-expansions and $\sup_{x\in\mathcal{X}}\sum_i|\phi_i(x)|\leq 1$.*

One notable case is when $L_i f=f(x_i)$ for some points $\{x_i\}$, $\phi_0\geq 0$, $\sum_i\phi_i=1$ and $\alpha\equiv 0$. If, in addition, Π is an interpolator w.r.t. the set $\{x_i\}$ (i.e., $(\Pi f)(x_i)=f(x_i)$) and $(\phi_i(x_j))_{ij}$ has full rank then all the elements of the space spanned by the basis functions $\{\phi_i\}$ are fixed

points of Π. This is nice since then $\epsilon_\Pi = d_\infty(\mathcal{F}, V^*)$ and so the loss is directly controlled by the size of $\mathcal{F}$.[11] Note that since V^* is unknown it may be hard to guarantee *a priori* a small bound on ϵ_Π. Further, while for non-linear methods, such as n-term approximation or neural nets, a small number of parameters might be sufficient to approximate well functions which are smooth with the exception of some "small boundary set" or highly anisotropic, linear methods (such as averagers) are less flexible (e.g. DeVore, 1997). Hence, although the requirement that $\mathcal{F}$ has to approximate well V^* seems less stringent than the requirement that the inherent Bellman error of $\mathcal{F}$ is small, this advantage might be well offset by the better adaptation ability of non-linear methods, permitted in sampling based FVI.[12] In connection to this, note that if a function space approximates well the optimal value function then this often implies a small inherent Bellman error, too. This is because in the lack of the knowledge of V^*, one must resort to generic features of the MDP (such as smoothness) when the function space is designed. In fact, the easiest way to guarantee that V^* is well approximated if the image of all sufficiently regular functions (e.g. bounded and smooth functions) under T are well-approximated by $\mathcal{F}$ as it was done in Section 7. As we have seen there this way one automatically guarantees a small inherent Bellman error, too.

The work presented here builds on and extends our previous work. For finite state-space MDPs, working with a finite set of representative states Munos (2003, 2005) considered planning scenarios with known dynamics analyzing the stability of both approximate policy iteration and value iteration with weighted L^2 (resp., L^p) fitting. Preliminary versions of the results presented here were published in (Szepesvári and Munos, 2005). Using techniques similar to those developed here, recently we have proved results for the learning scenario when only a single trajectory of some fixed behaviour policy is known (Antos et al., 2006). We know of no other work that would have considered the analysis of approximate value iteration with weighted p-norms over the space of value-functions in discounted, infinite-horizon settings.

One closely related result is due to Tsitsiklis and Roy (1999) who studied sampling based approximate value iteration with linear function approximators. However, they only consider a restricted set of MDPs: finite horizon, optimal stopping with discounted total rewards (the motivation is to study pricing of American Options). In this setting the next-state distribution under the condition of not stopping is uncontrolled – the state of the market evolves independently of the decision maker. Tsitsiklis and Roy (1999) argue that in this case it is better to sample full trajectories than to generate samples in some other, arbitrary way. Their algorithm implements approximate backward propagation of the values (by L^2 fitting with linear function approximators), exploiting that the problem has a fixed, finite horizon. The main result of this paper shows that the estimation error converges to zero with probability one as the number of samples grows to infinity. Further, they

11. TODO: /cs/ Maybe we should get rid of this discussion. At the end, the point is that the definition of Gordon happens to require that averagers are affine linear operators. However, linearity is not needed and could be removed if we wanted. So it would probably be better to say that all practical examples of averagers are linear! And this is true, I guess. I leave it to you to do with this what you want!
12. TODO: /cs/ This is again an overstatement maybe. Also sampling based FVI allows nonlinear methods, e.g. shrinkage estimation currently lies outside of the scope of the analysis. Although it is true that neural nets share a lot of the nice properties of these nonlinear methods and they do lie within the scope of the results. (But then we mimic that we are able to solve the arising global optimization problems – though everyone is doing that in regression estimation, as well!)

bound the performance if the approximator is kept fixed when the sample size approaches infinity. Due to the special structure of the problem, the bound depends only on how well the optimal value function can be approximated by the chosen function space. Certainly, due to the known counterexamples (Baird, 1995; Tsitsiklis and Van Roy, 1996), we cannot hope that such a bound would hold in general.

Another relevant algorithm is conservative policy iteration (CPI) **?**Kakade (2003) designed for discounted infinite horizon problems. The algorithm searches in a fixed policy space Π, in each step picking a policy that maximizes the average of the empirical advantages of the previous policy at a number of states (basepoints) sampled from some distribution. The advantages are estimated by sampling sufficiently long trajectories from the basepoints. The policy picked this way is mixed into the previous policy to prevent performance drops due to drastic changes, hence the name of the algorithm.

There is a major difference concerning the goal of this algorithm and our goal: Whereas our goal is to study the dependence of the loss due to the lack of knowledge of the MDP and the use of a finite number of samples, theirs is to show that the problem of finding a near optimal policy can be reduced to that of finding the discounted future state distribution of a good policy. Indeed, their main result (e.g. Theorems 7.3.1 and 7.3.3 in Kakade, 2003) bounds the loss of using the policy returned instead of some other policy π in terms of the total variation distance between the distribution used to sample the basepoints and the discounted future state distribution underlying π. Thus, the result shows that if π is a good policy then in order to make the loss of CPI small, one has to select the sampling distribution to match closely the future state distribution of π. In other words, CPI allows one to reduce the problem of finding a good policy to that of finding the stationary distribution of a good policy. The result is stated as a high probability bound where the samples depend polynomially on $|\mathcal{A}|$, $1/(1-\gamma)$, $1/\epsilon$, the VC dimension of the policy class (assuming two actions) and $\log(1/\delta)$. Kakade emphasizes that this polynomial does not depend on the size (dimensionality) of the state-space. However, since in some sense the approximation error (bias) is not considered in this result and bounds on the estimation error (variance) typically scale with the VC dimension and are otherwise independent of the dimensionality of the space considered, this is not surprising at all (the bounds on the estimation error in Theorem 2 are also dimension independent).

Another relevant recent paper is by Murphy (2005) who, just like Tsitsiklis and Roy (1999), studied finite horizon problems with no discounting.[13] The algorithm studied is fitted value iteration for action-value functions, using backward propagation of the values and independent trajectories, similarly as was suggested by Tsitsiklis and Roy (1999). The error bounds come in the form of performance differences between pairs of greedy policies: One policy is greedy w.r.t. the value function returned by the algorithm, the other is greedy w.r.t. to any "test" function from the function set considered in the algorithm. Murphy (2005) concludes that the number of samples needed to achieve a small expected loss, ϵ, is exponential in the horizon and proportional to $1/\epsilon^4$. Dependence on the approximation error is avoided by making the assumption that the optimal action-value function could be

13. The results in (Murphy, 2005) are actually presented for partially observable problems. Since by using a limited set of value functions one implicitly introduces state aliasing anyway, value-function based methods can always be applied to partially observable problems: The loss due to partial observability is incorporated in the bounds through the limits of the approximation power of the function class.

added at virtually no cost to the function sets used by the algorithm. This is unfortunately not a real option unless the problem to be solved is known or is restricted in some other fundamental way. Murphy (2005) also emphasizes that the bounds do not depend on the dimensionality of the state space (observation space in her case). However, by considering that the trivial problem with a single step (horizon one) is equivalent to regression the known lower bounds for regression apply (see Stone (1980, 1982) and Chapter 3 of Györfi, Kohler, Krzyżak, and Walk (2002) for the exact statements and proofs) and show that it is not possible to avoid dependence on the dimensionality of the state space. In particular, the number of samples needed scales exponentially with the dimensionality of the state space if the class of MDPs is large enough. Generally it holds that when no special conditions are imposed on the MDP, FVI and its variants are subject to the curse-of-dimensionality: for a loss of the order ϵ one needs $N = \Omega((1/\epsilon)^d)$ samples. As we also argued above, this follows because regression already suffers from the curse-of-dimensionality.

It is important to note here that the curse-of-dimensionality is not specific to FVI variants but is actually inavoidable for large classes of continuous state-space MDPs. In fact, the result of Chow and Tsitsiklis (1989) states the following: Consider a class of MDPs with $\mathcal{X} = [0,1]^d$. Assume that the transition probabilities of all MDPs in the class assume a density and these densities have a common upper bound. Further, assume that the MDPs are uniformly smooth: Both the reward function and the transition densities are assumed to be Lipschitzian with fixed Lipschitz constants L_r, L_p. Fix a desired precision, $\epsilon > 0$. Then, any algorithm that is guaranteed to return an ϵ-optimal approximation to the optimal value function must query (sample) the reward function and the transition probabilities at least $\Omega(1/\epsilon)^d$-times, for some MDP within the class considered. Hence, even classes of smooth MDPs with uniformly bounded transition densities contain instances that are very hard to solve in this sense.[14]

The situation changes dramatically if one is allowed to draw samples in an on-line manner (in the off-line case we restrict the number of samples to a finite number; in the on-line model as time goes by, more and more samples are generated). In this case it is possible to show that in each step, using a sample-size that is polynomial in the important quantities (in particular, linear in the dimensionality of the state space) it is possible to achieve near-optimality. Following the idea of Rust (1996b), Szepesvári (2001) has actually shown that this statement holds for a class of MDPs that is actually larger than the one considered by Chow and Tsitsiklis (1989). Another on-line method is explored by Kearns et al. (1999) who proposed to build a sparse lookahead tree to compute an approximate evaluation of the actions. The size of this tree, however, scales exponentially in the ϵ-effective horizon $R_{\max}/(\epsilon(1-\gamma))$. However, it does not depend on the dimensionality of the state space at all. In fact, Kearns et al. (1999) argue that without imposing additional assumptions (i.e., smoothness) on the MDP the exponential dependence on the effective horizon time is unavoidable (a similar dependence on the horizon shows up in the bounds of Murphy (2005) and Kakade (2003)).[15]

14. TODO: /Cs/: This result talks about value function approximations. Can it be extended to the problem of finding a good policy? I asked John. He said "possibly", "likely". They actually wrote in the paper that this is not hard but he now thinks that maybe that statements was not well supported.

15. TODO: /Cs/ What are we after here? What is the conclusion? Ok, we were discussing sample complexity, curse-of-dimensionality. What is the problem of on-line algorithms? I mean based on these results we all

Although regression suffers from the curse-of-dimensionality in the worst-case (and so do algorithms that are based on regression estimation), there are special circumstances that make it possible for regression methods to perform well in high-dimensional problems. Of special interest are those *adaptive methods* that are capable of exploiting special features of the problem when such features are present, but which at the same time avoid sacrificing efficiency in the lack of the special circumstances. One strength of the techniques developed here is that they link the performance of regression and reinforcement learning in a direct way. We believe that this opens up the possibility to let advances in regression automatically give rise to similar advances in reinforcement learning. Hence, when flexible regression methods that adapt to unknown smoothness (e.g. Donoho and Johnstone, 1995), sparsity (e.g. Abramovich et al., 2006), or the number of relevant variables (e.g. Kohler and Krzyżak, 2005; Lin and Zhang, 2006) are combined with reinforcement learning (fitted value iteration), we expect to see the same performance gains that applies to these methods when they work in a pure regression setting. However, one should not forget that the situation for the procedures studied here is made more complex by the requirement that the approximation power of a function space has a self-referencing nature. This does not seem to cause much difficulties for adapting to the unknown smoothness or sparsity, but it looks less evident how the methods of automatic dimension estimation can be extended to RL. We think that the exploration of the available options for doing this is an important and interesting avenue for future research.

9. Simulation Study

As it was noted earlier, sampling based FVI has been used successfully in many prior empirical studies, such as the works by Longstaff and Shwartz (2001); Haugh (2003) or Jung and Uthmann (2004). Hence, our goal here is not to evaluate sampling-based FVI, but to illustrate the various tradeoffs on a simple example where the exact solution (actually, a close approximation to it) can be derived analytically.

9.1 An Optimal Replacement Problem

The problem used as a testbed is a simple one-dimensional optimal replacement problem, described e.g. by Rust (1996a). The system has a one-dimensional state, where state variable $x_t \in \mathbb{R}_+$ measures the accumulated utilization of a product, such as the odometer reading on a car. By convention, we let $x_t = 0$ denote a brand new product. At each discrete time step, t, there are two possible decisions: either keep ($a_t = \mathbf{K}$) or replace ($a_t = \mathbf{R}$)

should forget about off-line estimation, right? Where is the catch? We should say something, otherwise the whole paper becomes pointless. One point is the engineering viewpoint: You possibly does not have the resources to do the sampling on-line. Exponential number of samples in the horizon: That is bad! But the other method could work? What do you think??/Rm/ First I think the terminology on-line vs off-line used here is not conventional and rather confusing here. Or maybe I don't understand it. For example in Rust96 and Szepesvari01, the samples are drawn off-line, the same way we do here. So why is it more "online". I guess the difference is that in Rust and Sz., we assume we know the full model of the dynamics, whereas here we only assume we have access to a simulation device (generative model). So I guess the main argument here is that having a full model is sometimes unrealistic. Now, about the method of Kearns et al. which is indeed more "online", the main argument is as you said the exponential dependency wrt the horizon length.

the product. This latter action implies an additional cost C of selling the existing product

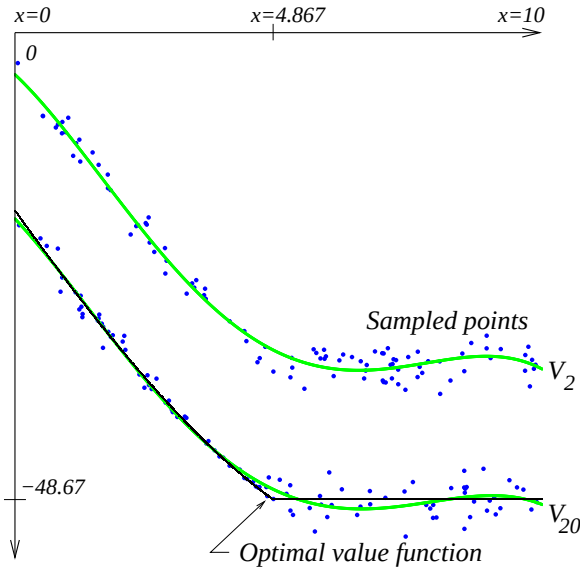

Figure 2: Illustration of Sampling based FVI at two iteration-steps (up: $k = 2$, down: $k = 20$). The dots represent the $N = 100$ sampled points and the corresponding target values that are obtained as the average of $M = 10$ samples. The grey curve is the best fit (among polynomials of degree $l = 4$). The thin black curve is the optimal value function.

and replacing it by a new one. The transition to a new state occurs with the following exponential densities:

$$p(y|x,\mathbf{K}) = \begin{cases} \beta e^{-\beta(y-x)}, & \text{if } y \geq x; \\ 0, & \text{if } y < x, \end{cases}$$

$$p(y|x,\mathbf{R}) = \begin{cases} \beta e^{-\beta y}, & \text{if } y \geq 0; \\ 0, & \text{if } y < 0. \end{cases}$$

The reward function is $r(x,\mathbf{K}) = -c(x)$, where $c(x)$, an increasing function of its argument, represents the cost of maintaining the product and $r(x,\mathbf{R}) = -C - c(0)$. The optimal value function solves the Bellman Optimality Equation:

$$V^*(x) = \max\Big[-c(x) + \gamma \int_x^\infty p(y|x,\mathbf{K}) V^*(y) dy, -C - c(0) + \gamma \int_0^\infty p(y|x,\mathbf{R}) V^*(y) dy \Big].$$

Here the first argument of max represents the total future reward given that the product is not replaced, whilst the second argument gives the total future reward provided that the product is replaced. From this equation it is possible to derive an analytical expression of the optimal value function:

$$V^*(x) = \begin{cases} \int_x^{\overline{x}} \frac{c'(y)}{1-\gamma}(1 - \gamma e^{-\beta(1-\gamma)(y-x)}) dy - \frac{c(\overline{x})}{1-\gamma}, & \text{if} \quad x \leq \overline{x}; \\ \frac{-c(\overline{x})}{1-\gamma}, & \text{if} \quad x > \overline{x}, \end{cases}$$

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

9.2 Results

We chose the numerical values $\gamma = 0.6$, $\beta = 0.5$, $C = 30$, $c(x) = 4x$. This gives $\overline{x} \simeq 4.8665$ and the optimal value function, plotted in Figure 2, is

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

We consider approximation of the value function using polynomials of degree l. Then, as suggested in Section 7, one should use truncation to keep the estimates bounded. In order to make the state space bounded, we introduce a problem that closely approximates the original one. For this we fix an upper bound for the states, $x_{\max} = 10 \gg \overline{x}$, and modify the problem definition such that if the next state y happens to be outside of the domain $[0, x_{\max}]$ then the product is replaced immediately, and a new state is drawn as if action $\mathbf{R}$ were chosen in the previous time-step.

By the choice of $x_{\max}$, $\int_{x_{\max}}^{\infty} p(dy|x, \mathbf{R})$ is negligible and hence the optimal value function of the altered problem closely matches that of the original problem when it is restricted to $[0, x_{\max}]$.

We chose the distribution μ to be uniform over the state space $[0, x_{\max}]$. The transition density functions $p(\cdot|x, a)$ are bounded by β, thus Assumption A1 holds with $C_\mu = \beta x_{\max} = 5$.

Figure 2 illustrates two iterates ($k = 2$ and $k = K = 20$) of sampling-based FVI: the dots represents the points $\{(X_n, \hat{V}(X_n))\}_{1 \le n \le N}$ for $N = 100$, where X_i is drawn from μ and $\{\hat{V}(X_n)\}_{1 \le n \le N}$ is computed using (3) with $V = V_k$ and $M = 10$ samples. The grey curve is the best fit (minimizing the least square error to the data) in $\mathcal{F}$ (for $l = 4$) and the thin black curve is the optimal value function.

Figure 9.2 shows the L_∞ approximation errors $||V^* - V_K||_\infty$ for different values of the degree l of the polynomial regression, and for two values of the number of basepoints ($N = 100$ and $N = 1000$). The number of iterations was set to $K = 20$.[16]

Note that in our previous analysis we derived bounds on the performance loss $||V^* - V^{\pi_K}||_\infty$ of using a policy π_K (instead of the optimal policy π^*) greedy w.r.t. the approximation V_K returned by sampling based FVI. However, in Figure 9.2, we show the approximation error $||V^* - V_K||_\infty$ instead. The reason is that in this numerical experiment, the performance loss is not very instructive, since even for a poorly approximated value function, the greedy policy is often close to the optimal policy. Thus, we found that the approximation error provides more information about the global performance of the method than the performance loss. Of course, the performance loss is always upper-bounded (in

16. TODO: /Cs/: How many repetitions are used, what is the variance of these values? Can we say that it is negligible? /Rm/ well... the variance is very small for good values of l and large for bad ls... I'm not sure this is very instructive...

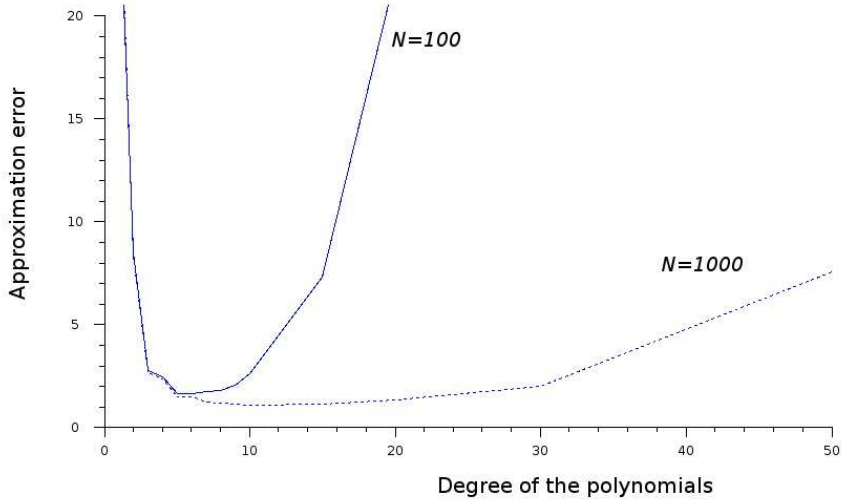

Figure 3: Approximation errors $||V^* - V_K||_\infty$ of the function V_K returned by sampling-based FVI after $K = 20$ iterations, for different values of the polynomials degree l, for $N = 100$ (plain curve) and $N = 1000$ (dash curve) samples. The number of sampled next states is $M = 10$. The plotted values are averaged over 100 independent runs.

L_∞-norm) by the approximation error, thanks to the well-known bound (Bertsekas and Tsitsiklis, 1996): $||V^* - V^{\pi_K}||_\infty \leq 2/(1-\gamma)||V^* - V_K||_\infty$.

From Figure 9.2 we observe that for $N = 100$, when the degree l of the polynomials increases, the error decreases first because the inherent approximation error, or bias decreases, but eventually increases because of overfitting (since the variance, or estimation error increases if we increase the function space). This results illustrate the different components of the bound (12) where the approximation error term $d_{p,\mu}(T\mathcal{F}, \mathcal{F})$ decreases with l (as discussed in Section 7) whereas the estimation term being a function of the pseudo-dimension of the $\mathcal{F}$ increases with l (in the linear approximation architecture considered here, the pseudo-dimension $V_{\mathcal{F}^+}$ equal the polynomials degree l plus two) with rate $O\big(\big(\frac{l+2}{N}\big)^{1/2p}\big) + O(1/M^{1/2})$, disregarding logarithmic factors. According to this bound, the estimation error decreases when the number of samples increases, which is corroborated by the experiments: As expected, overfitting decreases when the number of samples N, M increases.

In our second set of experiments we investigated whether the single-sample or the multi-sample variant of the algorithm is more advantageous in this particular problem. Figure 9.2 shows the distributional character of $V_K - V^*$ as a function of the state. The order of the

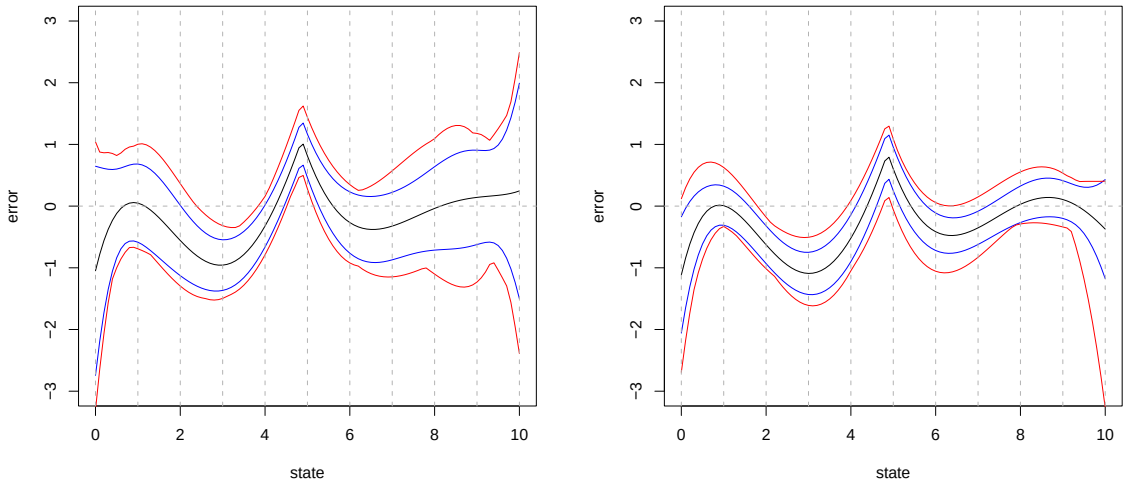

Figure 4: Approximation errors for the multi-sample (left figure) and the single-sample (right) variants of sampling based FVI. The figures show the distribution of errors for approximating the optimal value function as a function of state, as measured using 50 independent runs. For both version, $N = 100$, $K = 10$. However, for the single-sample variant $M = 100$, whilst for the multi-sample variant $M = 10$, making the total number of samples used in the two cases the same.

polynomial is 5. The black (middle) curve[17] shows the mean error (representing the bias) for 50 independent runs, the blue curves (just surrounding the middle curve) show the upper and lower confidence intervals at 1.5-times the observed standard deviation, whilst the red curves (outer curves, lightgrey) show the minimum/maximum approximation errors. Note the 'peak' at $\overline{x}$: The value function at this point is non-smooth, introducing a bias that converges to zero rather slowly (the same effect in Fourier analysis is known as the Gibbs phenomenon). It is also evident from the figure that the approximation error near the edges of the state space is larger. In polynomial interpolation for the uniform arrangements of the basepoints, the error actually blows up at the end of the intervals as the order of interpolation is increased (Runge's phenomenon). A general suggestion to avoid this is to increase the denseness of points near the edges or to introduce more flexible methods (e.g. splines). In FVI the edge effect is ultimately washed out, but still may slow down the procedure considerably in some cases when the behaviour of the value near the boundaries is critical.

Now, as to the comparison of the single- and multi-sample algorithms, it should be apparent from this figure that for this specific setup, the single-sample variant is actually preferable: The bias does not seem to increase much whilst the variance of the estimates

17. TODO: /Rm/ We cannot see the colors in the printed paper!!! /Cs/ This was plotted with R. Hmm, I guess I'd like to postpone plotting this again in greyscale when the paper is finalized...

decreases significantly. Thus, although the single-sample variant looks at a first sight looks to have a higher chance to overfit, we see that it can actually be more sample-efficient than the multi-sample variant, as predicted by the theory.

10. Conclusions

We considered sampling-based FVI for discounted, large (possibly infinite) state space, finite-action Markovian Decision Problems where only a generative model of the environment is available. In each iteration, a sampling-based approximation to the image of the previous iterate under Bellman operator is computed and a regression method is used to fit a function to the data obtained this way. The main contribution of the paper is a bound that shows how the final loss of sampling-based FVI depends on a number of important factors underlying the algorithm and the MDP. In particular, the bounds show that the loss can be bounded in terms of the Bellman residual of the function space used in the regression step, and the stochastic stability properties of the MDP (cf. Assumptions A1, A2). Our bounds and the counterexamples discussed in the introduction suggest that the Bellman residual of the function space plays a crucial role in the final performance of FVI. In particular, by increasing the number of samples and the richness of the function space at the same time, the resulting algorithm can be shown to be consistent for a wide class of MDPs. The derived rates show that in line with our expectations, FVI would typically suffer from the curse-of-dimensionality except when some specific conditions (extreme smoothness, only a few state variables are relevant, sparsity, etc.) are met. When such conditions are met, the resulting rates in regression are called 'fast'. Since these conditions could be difficult to verify a priori for any practical problem, methods that are capable of adapting to unknown properties that help to increase efficiency are of considerable interest. Although in the present paper we considered only the simplest risk minimization setting (with no such adaptation), we believe that since our results directly connect the regression literature and the RL literation, they could serve as a solid starting point for such further studies.

One immediate possibility along this line would be to extend our results to penalized empirical risk minimization when a penalty term penalizing the roughness of the candidate functions is added to the empirical risk. The advantage of this approach is well-known: Whilst in the method of sieves the user has to pick a function space tuned carefully to the specifics (sample size, assumed smoothness) of the problem, in penalized empirical risk minimization the same (rich) function space is used independently of the sample size and a multiplier of the penalty factor is used to control the complexity. Even better, it is then possible to design methods that select the multiplier in a close-to-optimal way (see Györfi et al., 2002, Section 21.2), enabling the algorithm to adapt to the actual (unknown) smoothness of the problem.

One disturbing feature of our bounds is that they scale slightly worse with the dimensionality than similar bounds available for regression estimation. We argued that it is possible to improve this to some extent, but the gap will never be eliminated unless action-value based FVI is considered. However, the analysis of this algorithm is left for future work.

There are other ways to improve the performance of our algorithm that are more directly related to specifics of RL. Both Tsitsiklis and Roy (1999) and Kakade (2003) argued that μ, the distribution used to sample the states should be selected to match the future state

distribution of a (near-)optimal policy. Since the only way to learn about the optimal policy is by running the algorithm, one idea is to change the sampling distribution by slowly moving it closer to the future state distribution of the most recent policy. The improvement presumably manifests itself by decreasing the term including $C_{\rho,\mu}$. It might be possible to improve the bound by a more careful analysis of how the errors propagate in the algorithm (the present analysis looks rather conservative). Another possibility is to adaptively choose M, the number of next states based on the available local information like in active learning, hoping that this way the sample-complexity of the algorithm could be further increased.

Acknowledgments

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).

Appendix A. Proof of Lemma 1

In order to prove Lemma 1 we will need the following inequality due to Pollard:

Theorem 8 (Pollard, 1984) *Let X_i, $i = 1, \ldots, N$, be an i.i.d. sequence taking values in the space $\mathcal{X}$. Let $\mathcal{F}$ be a set of measurable functions that are uniformly bounded and share the common bound K. Then*

$$\mathbb{P}\left(\sup_{f\in\mathcal{F}}\left|\frac{1}{N}\sum_{i=1}^{N} f(X_i) - \mathbb{E}\left[f(X_1)\right]\right| > \epsilon\right) \leq 8\mathbb{E}\left[\mathcal{N}(\epsilon/8, \mathcal{F}(X^{1:N}))\right] e^{-\frac{N\epsilon^2}{128K^2}}.$$

Here one should perhaps work with outer expectations because, in general, the supremum of an uncountably many random variables cannot be guaranteed to be measurable. However, since for specific examples of function space $\mathcal{F}$, measurability can typically be established by routine separability arguments, we will altogether ignore these measurability issues in this paper.

Now, let us prove Lemma 1 that stated the finite-sample bound for a single iterate.[18]

Proof Let $\epsilon'' > 0$ be arbitrary and let f^* be such that $\|f^* - TV\|_{p,\mu} \leq \inf_{f\in\mathcal{F}} \|f - TV\|_{p,\mu} + \epsilon''$. Define $\|\cdot\|_{p,\hat{\mu}}$ by

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

18. TODO: /Cs/: One reviewer suggested to repeat the text of the Lemmas. Should we do this? /rm/ I don't think so...

We will prove the lemma by proving that the following sequence of inequalities hold simultaneously on a set of events of measure not smaller than $1-\delta$:

$$\begin{aligned}
\left\|V' - TV\right\|_{p,\mu} &\leq \left\|V' - TV\right\|_{p,\hat{\mu}} + \epsilon' && (13)\\
&\leq \left\|V' - \hat{V}\right\|_{p,\hat{\mu}} + 2\epsilon' && (14)\\
&\leq \left\|f^* - \hat{V}\right\|_{p,\hat{\mu}} + 2\epsilon' && (15)\\
&\leq \left\|f^* - TV\right\|_{p,\hat{\mu}} + 3\epsilon' && (16)\\
&\leq \left\|f^* - TV\right\|_{p,\mu} + 4\epsilon' && (17)\\
&= d_{p,\mu}(TV, \mathcal{F}) + 4\epsilon' + \epsilon''. && (18)
\end{aligned}$$

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 will follow by choosing $\epsilon' = \epsilon/4$.

Now, let us prove (13)–(17). First, observe that (15) follows due to the choice of V' since $\left\|V' - \hat{V}\right\|_{p,\hat{\mu}} \leq \left\|f - \hat{V}\right\|_{p,\hat{\mu}}$ holds for all functions f from $\mathcal{F}$ and thus the same inequality holds for $f^* \in \mathcal{F}$, too.

Thus, (13)–(17) will be established if we prove that (13),(14),(16) and (17) all hold with probability at least $1-\delta'$ with $\delta' = \delta/4$. Let

$$Q = \max\Big(\Big|\left\|V' - TV\right\|_{p,\mu} - \left\|V' - TV\right\|_{p,\hat{\mu}}\Big|, \Big|\|f^* - TV\|_{p,\mu} - \|f^* - TV\|_{p,\hat{\mu}}\Big|\Big).$$

We claim that

$$\mathbb{P}\left(Q > \epsilon'\right) \leq \delta', \tag{19}$$

where $\delta' = \delta/4$. From this, (13) and (17) will follow.

In order to prove (19) note that for all $\omega \in \Omega$, $V' = V'(\omega) \in \mathcal{F}$. Hence,

$$\sup_{f\in\mathcal{F}} \Big|\|f - TV\|_{p,\mu} - \|f - TV\|_{p,\hat{\mu}}\Big| \geq \Big|\left\|V' - TV\right\|_{p,\mu} - \left\|V' - TV\right\|_{p,\hat{\mu}}\Big|$$

holds pointwise in Ω. Therefore the inequality

$$\sup_{f\in\mathcal{F}} \Big|\|f - TV\|_{p,\mu} - \|f - TV\|_{p,\hat{\mu}}\Big| > Q \tag{20}$$

holds pointwise in Ω, too and hence

$$\mathbb{P}\left(Q > \epsilon'\right) \leq \mathbb{P}\left(\sup_{f\in\mathcal{F}} \Big|\|f - TV\|_{p,\mu} - \|f - TV\|_{p,\hat{\mu}}\Big| > \epsilon'\right).$$

We claim that

$$\mathbb{P}\left(\sup_{f\in\mathcal{F}} \Big|\|f - TV\|_{p,\mu} - \|f - TV\|_{p,\hat{\mu}}\Big| > \epsilon'\right) \leq \mathbb{P}\left(\sup_{f\in\mathcal{F}} \Big|\|f - TV\|_{p,\mu}^p - \|f - TV\|_{p,\hat{\mu}}^p\Big| > (\epsilon')^p\right). \tag{21}$$

Consider any event ω such that

$$\sup_{f\in\mathcal{F}}\left|\|f-TV\|_{p,\mu}-\|f-TV\|_{p,\hat{\mu}}\right|>\epsilon' \tag{22}$$

holds. For any such event ω there exist a function $f'\in\mathcal{F}$ such that

$$\left|\left\|f'-TV\right\|_{p,\mu}-\left\|f'-TV\right\|_{p,\hat{\mu}}\right|>\epsilon'.$$

Pick such a function. Assume first that $\|f'-TV\|_{p,\hat{\mu}}\leq\|f'-TV\|_{p,\mu}$. Hence, $\|f'-TV\|_{p,\hat{\mu}}+\epsilon'<\|f'-TV\|_{p,\mu}$. Since $p\geq 1$, the elementary inequality $x^p+y^p\leq(x+y)^p$ holds for any non-negative numbers x,y. Hence we get $\|f'-TV\|_{p,\hat{\mu}}^p+\epsilon^p\leq(\|f'-TV\|_{p,\hat{\mu}}+\epsilon)^p<\|f'-TV\|_{p,\mu}^p$ and thus

$$\left|\left\|f'-TV\right\|_{p,\hat{\mu}}^p-\left\|f'-TV\right\|_{p,\mu}^p\right|>\epsilon^p.$$

This inequality can be shown to hold by an entirely analogous reasoning when $\|f'-TV\|_{p,\hat{\mu}}>\|f'-TV\|_{p,\mu}$.

Inequality (21) now follows since

$$\sup_{f\in\mathcal{F}}\left|\|f-TV\|_{p,\mu}^p-\|f-TV\|_{p,\hat{\mu}}^p\right|\geq\left|\left\|f'-TV\right\|_{p,\mu}^p-\left\|f'-TV\right\|_{p,\hat{\mu}}^p\right|.$$

Now, observe that $\|f-TV\|_{p,\mu}^p=\mathbb{E}\left[|(f(X_1)-(TV)(X_1))|^p\right]$, and $\|f-TV\|_{p,\hat{\mu}}^p$ is thus just the sample average approximation of $\|f-TV\|_{p,\mu}^p$. Hence, by noting that the covering number associated with $\{f-TV|f\in\mathcal{F}\}$ can be bounded in terms of the covering number of $\mathcal{F}$, calling for Theorem 8 results in

$$\mathbb{P}\left(\sup_{f\in\mathcal{F}}\left|\|f-TV\|_{p,\mu}^p-\|f-TV\|_{p,\hat{\mu}}^p\right|>(\epsilon')^p\right)\leq 8\mathbb{E}\left[\mathcal{N}(\epsilon'/8,\mathcal{F}(X^{1:N}))\right]\,e^{-\frac{N(\epsilon')^{2p}}{2(16V_{\max})^2}}.$$

By making the right-hand-side upper bounded by $\delta'=\delta/4$ we find a lower bound on N, displayed in turn in (8). This finishes the proof of (19).

Now, let us prove inequalities (14) and (16). Let f denote an arbitrary random function such that $f=f(x;\omega)$ is measurable for each $x\in X$ and assume that f is uniformly bounded by $V_{\max}$. Making use of the triangle inequality

$$\left|\|f-g\|_{p,\hat{\mu}}-\|f-h\|_{p,\hat{\mu}}\right|\leq\|g-h\|_{p,\hat{\mu}},$$

we get that

$$\left|\|f-TV\|_{p,\hat{\mu}}-\left\|f-\hat{V}\right\|_{p,\hat{\mu}}\right|\leq\left\|TV-\hat{V}\right\|_{p,\hat{\mu}}.$$

Hence, it suffices to bound $\left\|TV-\hat{V}\right\|_{p,\hat{\mu}}$.

For this purpose we shall use Hoeffding's inequality (Hoeffding, 1963) and union bound arguments. Fix any index i ($1 \le i \le N$). Let $K_1 = \hat{R}_{\max} + \gamma V_{\max}$. Then, by assumption $R_j^{X_i,a} + \gamma V(Y_j^{X_i,a}) \in [-K_1, K_1]$ holds with probability one and thus by Hoeffding's inequality,

$$\mathbb{P}\left(\left| \mathbb{E}\left[R_1^{X_i,a} + \gamma V(Y_1^{X_i,a}) \,|\, X^{1:N} \right] - \frac{1}{M}\sum_{j=1}^{M} R_j^{X_i,a} + \gamma V(Y_j^{X_i,a}) \right| > \epsilon' \,|\, X^{1:N} \right) \le 2\, e^{-\frac{2M(\epsilon')^2}{K_1^2}}, \tag{23}$$

where $X^{1:N} = (X_1, \dots, X_N)$. Making the right-hand-side upper bounded by $\delta'/(N|\mathcal{A}|)$ we find a lower bound on M (cf. (9)). Since

$$\left|(TV)(X_i) - \hat{V}(X_i)\right| \le \max_{a \in A} \left| \mathbb{E}\left[R_1^{X_i,a} + \gamma V(Y_1^{X_i,a}) \,|\, X^{1:N} \right] - \frac{1}{M}\sum_{j=1}^{M} \left[R_j^{X_i,a} + \gamma V(Y_j^{X_i,a}) \right] \right|$$

it follows by a union bounding argument that

$$\mathbb{P}\left(\left|(TV)(X_i) - \hat{V}(X_i)\right| > \epsilon' \,|\, X^{1:N} \right) \le \delta'/N,$$

and hence another union bounding argument yields

$$\mathbb{P}\left(\max_{i=1,\dots,N} \left|(TV)(X_i) - \hat{V}(X_i)\right|^p > (\epsilon')^p \,|\, X^{1:N} \right) \le \delta'.$$

Taking the expectation of both sides of this inequality gives

$$\mathbb{P}\left(\max_{i=1,\dots,N} \left|(TV)(X_i) - \hat{V}(X_i)\right|^p > (\epsilon')^p \right) \le \delta'.$$

Hence also

$$\mathbb{P}\left(\frac{1}{N}\sum_{i=1}^{N} \left|(TV)(X_i) - \hat{V}(X_i)\right|^p > (\epsilon')^p \right) \le \delta' \tag{24}$$

and therefore

$$\mathbb{P}\left(\left| \|f - TV\|_{p,\hat{\mu}} - \left\|f - \hat{V}\right\|_{p,\hat{\mu}} \right| > \epsilon' \right) \le \delta'$$

Using this with $f = V'$ and $f = f^*$ shows that inequalities (14) and (16) each hold with probability at least $1 - \delta'$. This finishes the proof of the lemma. ■

Now, let us turn to the proof of Lemma 2.

A.1 Proof of Lemma 2

Proof The proof is analogous to that of Lemma 1, hence we only give the differences. Up to (20) the two proofs proceed in an identical way, however, from (20) we continue by concluding that

$$\sup_{g \in \mathcal{F}} \sup_{f \in \mathcal{F}} \left| \|f - Tg\|_{p,\mu} - \|f - Tg\|_{p,\hat{\mu}} \right| > Q \tag{25}$$

holds pointwise in Ω. From this point onward, $\sup_{f\in\mathcal{F}}$ is replaced by $\sup_{g,f\in\mathcal{F}}$ throughout the proof of (19): The proof goes through as before until the point where Pollard's inequality is used. At this point, since we have two suprema, we need to consider covering numbers corresponding to the function set $\mathcal{F}_{T-} = \{ f - Tg \,|\, f \in \mathcal{F}, g \in \mathcal{F} \}$.

In the second part of the proof we must also use Pollard's inequality in place of Hoeffding's. In particular, (23) is replaced with

$$\mathbb{P}\left(\sup_{g\in\mathcal{F}} \left| \mathbb{E}\left[R_1^{X_i,a} + \gamma g(Y_1^{X_i,a}) \,|\, X^{1:N} \right] - \frac{1}{M}\sum_{j=1}^{M} R_j^{X_i,a} + \gamma g(Y_j^{X_i,a}) \right| > \epsilon' \,\middle|\, X^{1:N} \right) \leq \quad 8\mathbb{E}\left[\mathcal{N}(\epsilon'/8, \mathcal{F}_+(Z_{i,a}^{1:M})) \right] e^{-\frac{M(\epsilon')^2}{128K_1^2}},$$

where $Z_{i,a}^j = (R_j^{X_i,a}, Y_j^{X_i,a})$. Here $\mathcal{F}_+ = \{h : \mathbb{R}\times\mathcal{X} \to \mathbb{R} \,|\, h(s,x) = s\mathbb{I}_{\{|s|\leq V_{\max}\}} + f(x) \text{ for some } f \in \mathcal{F}\}$. The proof is concluded by noting that the covering numbers of $\mathcal{F}_+$ can be bounded in terms of the covering numbers of $\mathcal{F}$ using the arguments presented after the Lemma at the end of Section 4. ■

Appendix B. Proof of Theorem 2

First, note that iteration (11) (or (10)) may be written

$$V_{k+1} = TV_k - \varepsilon_k \tag{26}$$

where ε_k, defined by $\varepsilon_k = TV_k - V_{k+1}$, is the approximation error of the Bellman operator applied to V_k due to sampling. Before proving Theorem 2, we first state and prove two lemmas providing pointwise and L^p error bounds on the performance of using a policy π_k greedy w.r.t. V_k instead of the optimal policy π^*, as a function of the approximation errors ε_k, for $0 \leq k < K$.

B.1 Pointwise Error Bounds

Lemma 3 *We have*

$$V^* - V^{\pi_K} \leq (I - \gamma P^{\pi_K})^{-1}\Big\{ \textstyle\sum_{k=0}^{K-1} \gamma^{K-k}\big[(P^{\pi^*})^{K-k} + P^{\pi_K}P^{\pi_{K-1}}\dots P^{\pi_{k+1}}\big]|\varepsilon_k| + \gamma^{K+1}\big[(P^{\pi^*})^{K+1} + (P^{\pi_K}P^{\pi_K}P^{\pi_{K-1}}\dots P^{\pi_1})\big]|V^* - V_0|\Big\}. \tag{27}$$

Proof Since $TV_k \geq T^{\pi^*}V_k$, we have

$$V^* - V_{k+1} = T^{\pi^*}V^* - T^{\pi^*}V_k + T^{\pi^*}V_k - TV_k + \varepsilon_k \leq \gamma P^{\pi^*}(V^* - V_k) + \varepsilon_k,$$

from which we deduce by induction

$$V^* - V_K \leq \sum_{k=0}^{K-1} \gamma^{K-k-1}(P^{\pi^*})^{K-k-1}\varepsilon_k + \gamma^K(P^{\pi^*})^K(V^* - V_0). \tag{28}$$

Similarly, from the definition of π_k and since $TV^* \geq T^{\pi_k} V^*$, we have

$$V^* - V_{k+1} = TV^* - T^{\pi_k} V^* + T^{\pi_k} V^* - TV_k + \varepsilon_k \geq \gamma P^{\pi_k}(V^* - V_k) + \varepsilon_k.$$

Thus, by induction,

$$V^* - V_K \geq \sum_{k=0}^{K-1} \gamma^{K-k-1} (P^{\pi_{K-1}} P^{\pi_{K-2}} \dots P^{\pi_{k+1}}) \varepsilon_k + \gamma^K (P^{\pi_K} P^{\pi_{K-1}} \dots P^{\pi_1})(V^* - V_0). \quad (29)$$

Now, from the definition of π_K and since $TV_K \geq T^{\pi^*} V_K$, we have

$$\begin{aligned} V^* - V^{\pi_K} &= T^{\pi^*} V^* - T^{\pi^*} V_K + T^{\pi^*} V_K - TV_K + TV_K - T^{\pi_K} V^{\pi_K} \\ &\leq \gamma P^{\pi^*}(V^* - V_K) + \gamma P^{\pi_K}(V_K - V^* + V^* - V^{\pi_K}) \\ (I - \gamma P^{\pi_K})(V^* - V^{\pi_K}) &\leq \gamma (P^{\pi^*} - P^{\pi_K})(V^* - V_K), \end{aligned}$$

and since $(I - \gamma P^{\pi_K})$ is invertible and its inverse is positive (we may write $(I - \gamma P^{\pi_K})^{-1} = \sum_{m \geq 0} \gamma^m (P^{\pi_K})^m$), we deduce

$$V^* - V^{\pi_K} \leq \gamma (I - \gamma P^{\pi_K})^{-1} (P^{\pi^*} - P^{\pi_K})(V^* - V_K) \quad (30)$$

Now, using (28) and (29),

$$\begin{aligned} V^* - V^{\pi_K} \leq \; & (I - \gamma P^{\pi_K})^{-1} \Big\{ \textstyle\sum_{k=0}^{K-1} \gamma^{K-k} \big[(P^{\pi^*})^{K-k} - P^{\pi_K} P^{\pi_{K-1}} \dots P^{\pi_{k+1}} \big] \varepsilon_k \\ & + \gamma^{K+1} \big[(P^{\pi^*})^{K+1} - (P^{\pi_K} P^{\pi_K} P^{\pi_{K-1}} \dots P^{\pi_1}) \big] (V^* - V_0) \Big\}. \end{aligned}$$

We deduce (27) by taking the absolute value of both sides of the previous inequality. ■

B.2 L^p Error Bounds

We have the following approximation results.

Lemma 4 *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 ϵ, i.e. $\|\varepsilon_k\|_{p,\mu} \leq \epsilon$ for all $0 \leq k < K$, then*

- *Given Assumption A1 we have*

$$\|V^* - V^{\pi_K}\|_\infty \leq \frac{2\gamma}{(1-\gamma)^2} C_\mu^{1/p} \epsilon + \eta. \quad (31)$$

- *Given Assumption A2 we have*

$$\|V^* - V^{\pi_K}\|_{p,\rho} \leq \frac{2\gamma}{(1-\gamma)^2} C_{\rho,\mu}^{1/p} \epsilon + \eta. \quad (32)$$

Note that if $\|\varepsilon_k\|_\infty \leq \epsilon$ then letting $p \to \infty$ we get back the well-known, unimprovable supremum-norm error bounds

$$\limsup_{K\to\infty} \|V^* - V^{\pi_K}\|_\infty \leq \frac{2\gamma}{(1-\gamma)^2}\epsilon$$

for approximate value iteration (Bertsekas and Tsitsiklis, 1996). (In fact, by inspecting the proof below it turns out that for this the weaker condition, $\limsup_{k\to\infty} \|\varepsilon_k\|_\infty \leq \epsilon$ suffices, too.)

Proof We have seen that if A1 holds then A2 also holds, and for any distribution ρ, $C_{\rho,\mu} \leq C_\mu$. Thus, if the bound (32) holds for any ρ then choosing ρ to be a Dirac at each state proves (31). Thus we only need to prove (32).

We may rewrite (27) as

$$V^* - V^{\pi_K} \leq \frac{2\gamma(1-\gamma^{K+1})}{(1-\gamma)^2}\left[\sum_{k=0}^{K-1} \alpha_k A_k |\varepsilon_k| + \alpha_K A_K |V^* - V_0|\right],$$

with the positive coefficients

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

(defined such that they sum to 1) and the probability kernels:

$$\begin{aligned} A_k &= \frac{1-\gamma}{2}(I - \gamma P^{\pi_K})^{-1}\left[(P^{\pi^*})^{K-k} + P^{\pi_K} P^{\pi_{K-1}} \dots P^{\pi_{k+1}}\right], \text{ for } 0 \leq k < K, \\ A_K &= \frac{1-\gamma}{2}(I - \gamma P^{\pi_K})^{-1}\left[(P^{\pi^*})^{K+1} + P^{\pi_K} P^{\pi_K} P^{\pi_{K-1}} \dots P^{\pi_1}\right]. \end{aligned}$$

We have:

$$\begin{aligned} \|V^* - V^{\pi_K}\|_{p,\rho}^p &= \int \rho(dx)|V^*(x) - V^{\pi_K}(x)|^p \\ &\leq \left[\frac{2\gamma(1-\gamma^{K+1})}{(1-\gamma)^2}\right]^p \int \rho(dx)\left[\sum_{k=0}^{K-1} \alpha_k A_k |\varepsilon_k| + \alpha_K A_K |V^* - V_0|\right]^p (x) \\ &\leq \left[\frac{2\gamma(1-\gamma^{K+1})}{(1-\gamma)^2}\right]^p \int \rho(dx)\left[\sum_{k=0}^{K-1} \alpha_k A_k |\varepsilon_k|^p + \alpha_K A_K |V^* - V_0|^p\right](x), \end{aligned}$$

by using two times Jensen's inequality (since the sum of the coefficients α_k, for $k \in [0, K]$, is 1, and the A_k are stochastic operators) (i.e. convexity of $x \to |x|^p$).

The term $|V^* - V_0|$ may be bounded by $2V_{\max}$. Now, under Assumption A2, $\rho A_k \leq (1-\gamma)\sum_{m\geq 0} \gamma^m c(m+K-k)\mu$ and we deduce

$$\|V^* - V^{\pi_K}\|_{p,\rho}^p \leq \left[\frac{2\gamma(1-\gamma^{K+1})}{(1-\gamma)^2}\right]^p \left[\sum_{k=0}^{K-1} \alpha_k (1-\gamma) \sum_{m\geq 0} \gamma^m c(m+K-k) \|\varepsilon_k\|_{p,\mu}^p + \alpha_K (2V_{\max})^p\right].$$

Replace α_k by their values, and from the definition of $C_{\rho,\mu}$, and since $\|\varepsilon_k\|_{p,\mu} \leq \epsilon$, we have:

$$\begin{aligned}\|V^* - V^{\pi_K}\|_{p,\rho}^p &\leq \left[\frac{2\gamma(1-\gamma^{K+1})}{(1-\gamma)^2}\right]^p \Big[\frac{(1-\gamma)^2}{1-\gamma^{K+1}} \\ &\qquad \sum_{m\geq 0}\sum_{k=0}^{K-1} \gamma^{m+K-k-1} c(m+K-k)\epsilon^p + \frac{(1-\gamma)\gamma^K}{1-\gamma^{K+1}}(2V_{\max})^p\Big] \\ &\leq \left[\frac{2\gamma(1-\gamma^{K+1})}{(1-\gamma)^2}\right]^p \left[\frac{1}{1-\gamma^{K+1}} C_{\rho,\mu}\epsilon^p + \frac{(1-\gamma)\gamma^K}{1-\gamma^{K+1}}(2V_{\max})^p\right]\end{aligned}$$

Thus there is K linear in $\log(1/\eta)$ and $\log V_{\max}$, e.g. such that

$$\gamma^K < \left[\frac{(1-\gamma)^2}{4\gamma V_{\max}}\eta\right]^p$$

such that the second term is bounded by η^p, thus,

$$\|V^* - V^{\pi_K}\|_{p,\rho}^p \leq \left[\frac{2\gamma}{(1-\gamma)^2}\right]^p C_{\rho,\mu}\epsilon^p + \eta^p$$

thus

$$\|V^* - V^{\pi_K}\|_{p,\rho} \leq \frac{2\gamma}{(1-\gamma)^2} C_{\rho,\mu}^{1/p}\epsilon + \eta$$

■

B.3 Proof of Theorem 2

Proof Let us consider first the multi-sample variant of the algorithm under Assumption A2. Fix $\epsilon, \delta > 0$. Let the iterates produced by the algorithm be $V_1, \ldots, V_K$. Our aim is to show that by selecting the number of iterates, K and the number of samples, N, M sufficiently large, the bound

$$\|V^* - V^{\pi_K}\|_{p,\rho} \leq \frac{2\gamma}{(1-\gamma)^2}\, C_{\rho,\mu}^{1/p}\; d_{p,\mu}(T\mathcal{F}, \mathcal{F}) + \epsilon \tag{33}$$

holds with probability at least $1-\delta$. First, note that by construction the iterates V_k remain bounded by $V_{\max}$. By Lemma 4, under Assumption A2, for all those events, where the error $\epsilon_k = TV_k - V_{k+1}$ of the kth iterate is below (in $L^p(\mu)$-norm) some level ϵ_0, we have

$$\|V^* - V^{\pi_K}\|_{p,\rho} \leq \frac{2\gamma}{(1-\gamma)^2} C_{\rho,\mu}^{1/p}\epsilon_0 + \eta, \tag{34}$$

provided that $K = \Omega(\log(1/\eta))$. Now, choose $\epsilon' = (\epsilon/2)\,(1-\gamma)^2/(2\gamma C_{\rho,\mu}^{1/p})$ and $\eta = \epsilon/2$. Let $f(\epsilon,\delta)$ denote the function that gives lower bounds on N, M in Lemma 1 based on the value of the desired estimation error ϵ and confidence δ. Let $(N, M) \geq f(\epsilon', \delta/K)$. One difficulty is that V_k, the kth iterate is random itself, hence Lemma 1 (stated for deterministic functions) cannot be applied directly. However, thanks to the independence of samples between iterates, this is easy to fix. Let D_k denote the samples used up to

iteration k, i.e., V_k is measurable w.r.t. the sigma algebra generated by D_k, but V_{k+1} is not measurable w.r.t. D_k. Since V_k is D_k-measurable(i.e., 'non-random' conditionally on D_k), we can apply Lemma 1 to the conditional probability space defined by D_k to get $\mathbb{P}\left(\|\epsilon_k\|_{p,\mu} \le d_{p,\mu}(TV_k, \mathcal{F}) + \epsilon' | D_k\right) \ge 1 - \delta/K$. Taking expectation of both sides, we get that

$$\|\epsilon_k\|_{p,\mu} \le d_{p,\mu}(TV, \mathcal{F}) + \epsilon' \tag{35}$$

holds except for a set of bad events B_k of measure at most δ/K. Hence, inequality (35) holds simultaneously for $k = 1, \dots, K$, except for the events in $B = \cup_k B_k$. Note that $\mathbb{P}(B) \le \sum_{k=1}^{K} \mathbb{P}(B_k) \le \delta$. Now pick any event in the complementer of B. Thus, for such an event (34) holds with $\epsilon_0 = d_{p,\mu}(TV, \mathcal{F}) + \epsilon'$. Plugging in the definitions of ϵ' and η we obtain (33).

Now assume that the MDP satisfies Assumption A1. As before, we conclude that (35) holds except for the events in B_k and with the same choice of N and M, we still have $\mathbb{P}(B) = \mathbb{P}(\cup_k B_K) \le \delta$. Now, using (31) we conclude that except on the set B, $\|V^* - V^{\pi_K}\|_\infty \le \frac{2\gamma}{(1-\gamma)^2} C_{\rho,\mu}^{1/p} \, d_{p,\mu}(T\mathcal{F}, \mathcal{F}) + \epsilon$, concluding the first part of the proof.

For single-sample FVI the proof proceeds identically, except that now one uses Lemma 2 in place of Lemma 1. ■

Appendix C. Proof of Theorem 3

Proof We would like to prove that the policy defined in Section 6 gives close to optimal performance. Let us prove first the statement under Assumption A2.

By the choice of M', it follows using Hoeffding's inequality (see also Even-Dar et al., 2002, Theorem 1) that $\pi_{\alpha,\lambda}^K$ selects α-greedy actions with probability at least $1 - \lambda$.

Let π_α^K be a policy that selects α-greedy actions. A straightforward adaptation of the proof of Lemma 5.17 of Szepesvári (2001) yields that for all state $x \in \mathcal{X}$,

$$|V^{\pi_{\alpha,\lambda}^K}(x) - V^{\pi_\alpha^K}(x)| \le \frac{2V_{\max}\lambda}{1-\gamma}. \tag{36}$$

Now, use the triangle inequality to get

$$\left\|V^* - V^{\pi_{\alpha,\lambda}^K}\right\|_{p,\rho} \le \left\|V^* - V^{\pi_\alpha^K}\right\|_{p,\rho} + \left\|V^{\pi_\alpha^K} - V^{\pi_{\alpha,\lambda}^K}\right\|_{p,\rho}.$$

By (36), the second term can be bounded by $\frac{2V_{\max}\lambda}{1-\gamma}$, so let us consider the first term.

A modification of Lemmas 3 and 4 yields the following result, the proof of which will be given at the end of this section:

Lemma 5 *The following bound*

$$\left\|V^* - V^{\pi_\alpha^K}\right\|_{p,\rho} \le 2^{1-1/p}\left[\frac{2\gamma}{(1-\gamma)^2} C_{\rho,\mu}^{1/p} \max_{0\le k<K} \|\varepsilon_k\|_{p,\mu} + \eta + \frac{\alpha}{1-\gamma}\right] \tag{37}$$

holds for K such that $\gamma^K < \left[\frac{(1-\gamma)^2}{4\gamma V_{\max}}\eta\right]^p$.

Again, let $f(\epsilon,\delta)$ be the function that gives the bounds on N, M in Lemma 1 for given ϵ and δ and set $(N,M) \geq f(\epsilon',\delta/K)$ for ϵ' to be chosen later. Using the same argument as in the proof of Theorem 2 and Lemma 1 we may conclude that $\|\epsilon_k\|_{p,\mu} \leq d_{p,\mu}(TV_k,\mathcal{F}) + \epsilon' \leq d_{p,\mu}(T\mathcal{F},\mathcal{F}) + \epsilon'$ holds except for a set B_k with $\mathbb{P}(B_k) \leq \delta/K$.

Thus, except on the set $B = \cup_k B_k$ of measure not more than δ,

$$\begin{aligned}\left\|V^* - V^{\pi^K_{\alpha,\lambda}}\right\|_{p,\rho} &\leq 2^{1-1/p}\left[\frac{2\gamma}{(1-\gamma)^2}C^{1/p}_{\rho,\mu}\left(d_{p,\mu}(T\mathcal{F},\mathcal{F})+\epsilon'\right)+\eta+\frac{\alpha}{1-\gamma}\right]+\frac{2V_{\max}\lambda}{1-\gamma}\\ &\leq \left[\frac{4\gamma}{(1-\gamma)^2}C^{1/p}_{\rho,\mu}d_{p,\mu}(T\mathcal{F},\mathcal{F})+\frac{4\gamma}{(1-\gamma)^2}C^{1/p}_{\rho,\mu}\epsilon'+2\eta+\frac{2\alpha}{1-\gamma}\right]+\frac{2V_{\max}\lambda}{1-\gamma}.\end{aligned}$$

Now define $\alpha = \epsilon(1-\gamma)/8$, $\eta = \epsilon/8$, $\epsilon' = \epsilon/16(1-\gamma)^2/\gamma C^{-1/p}_{\rho,\mu}$ and $\lambda = \epsilon/8(1-\gamma)/V_{\max}$ to conclude that

$$\left\|V^* - V^{\pi^K_{\alpha,\lambda}}\right\|_{p,\rho} \leq \frac{4\gamma}{(1-\gamma)^2}C^{1/p}_{\rho,\mu}d_{p,\mu}(T\mathcal{F},\mathcal{F})+\epsilon \tag{38}$$

holds everywhere except on B. Also, just like in the proof of Theorem 2, we get that under Assumption A1 the statement for the supremum norm holds, as well.

It thus remained to prove Lemma 5:

Proof [Lemma 5] Write $\mathbf{1}$ for the constant function equals to 1. Since π^K_α is α-greedy w.r.t. V_K, we have $TV_K \geq T^{\pi^K_\alpha}V_K \geq TV_K - \alpha\mathbf{1}$. Thus, similarly to the proof of Lemma 3, we have

$$\begin{aligned}V^* - V^{\pi^K_\alpha} &= T^{\pi^*}V^* - T^{\pi^*}V_K + T^{\pi^*}V_K - TV_K + TV_K - T^{\pi^K_\alpha}V_K + T^{\pi^K_\alpha}V_K - T^{\pi^K_\alpha}V^{\pi^K_\alpha}\\ &\leq \gamma P^{\pi^*}(V^* - V_K) + \gamma P^{\pi^K_\alpha}(V_K - V^* + V^* - V^{\pi^K_\alpha}) + \alpha\mathbf{1}\\ &\leq (I-\gamma P^{\pi^K_\alpha})^{-1}\left[(\gamma P^{\pi^*} - P^{\pi^K_\alpha})(V^* - V_K)\right] + \frac{\alpha\mathbf{1}}{1-\gamma},\end{aligned}$$

and by using (28) and (29), we deduce

$$\begin{aligned}V^* - V^{\pi^K_\alpha} \leq\ & (I-\gamma P^{\pi^K_\alpha})^{-1}\Big\{\textstyle\sum_{k=0}^{K-1}\gamma^{K-k}\left[(P^{\pi^*})^{K-k} + P^{\pi^K_\alpha}P^{\pi_{K-1}}\dots P^{\pi_{k+1}}\right]|\varepsilon_k|\\ &+\gamma^{K+1}\left[(P^{\pi^*})^{K+1} + (P^{\pi^K_\alpha}P^{\pi_K}P^{\pi_{K-1}}\dots P^{\pi_1})\right]|V^* - V_0|\Big\} + \tfrac{\alpha\mathbf{1}}{1-\gamma}.\end{aligned}$$

Now, from the inequality $|a+b|^p \leq 2^{p-1}(|a|^p + |b|^p)$, we deduce, by following the same lines as in the proof of Lemma 4, that

$$\left\|V^* - V^{\pi^K_\alpha}\right\|^p_{p,\rho} \leq 2^{p-1}\left\{\left[\frac{2\gamma}{(1-\gamma)^2}\right]^p C_{\rho,\mu}(\max_{0\leq k<K}\|\varepsilon_k\|_{p,\mu})^p + \eta^p + \left[\frac{\alpha}{1-\gamma}\right]^p\right\},$$

and Lemma 5 follows. ■

■

References

F. Abramovich, Y. Benjamini, D.L. Donoho, and I.M. Johnstone. Adapting to unknown sparsity by controlling the false discovery rate. *Annals of Statistics*, 34(2), 2006. (to appear).

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

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-2006*, 2006. (to appear).

L. Baird. Residual algorithms: Reinforcement learning with function approximation. In Armand Prieditis and Stuart Russell, editors, *Proceedings of the Twelfth International Conference on Machine Learning*, pages 30–37, San Francisco, CA, 1995. Morgan Kaufmann.

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

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

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

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

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

C.S. Chow and J.N. Tsitsiklis. The complexity of dynamic programming. *Journal of Complexity*, 5:466–488, 1989.

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.

N. Cristianini and J. Shawe-Taylor. *An introduction to support vector machines (and other kernel-based learning methods)*. Cambridge University Press, 2000.

R.H. Crites and A.G. Barto. Improving elevator performance using reinforcement learning. In *Advances in Neural Information Processing Systems 9*, 1997.

R. DeVore. *Nonlinear Approximation*. Acta Numerica, 1997.

T. G. Dietterich and X. Wang. Batch value function approximation via support vectors. In T. G. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems 14*, Cambridge, MA, 2002. MIT Press.

D.L. Donoho and I.M. Johnstone. Adapting to unknown smoothness via wavelet shrinkage. *Journal of the American Statistical Association*, 90(432):1200–1224, 1995.

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

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

G.J. Gordon. Stable function approximation in dynamic programming. In Armand Prieditis and Stuart Russell, editors, *Proceedings of the Twelfth International Conference on Machine Learning*, pages 261–268, San Francisco, CA, 1995. Morgan Kaufmann.

A. Gosavi. A reinforcement learning algorithm based on policy iteration for average reward: Empirical results with yield management and convergence analysis. *Machine Learning*, 55:5–29, 2004.

U. Grendander. *Abstract Inference*. Wiley, New York, 1981.

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

M. Haugh. Duality theory and simulation in financial engineering. In *Proceedings of the Winter Simulation Conference*, pages 327–334, 2003.

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.

W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58:13–30, 1963.

T. Jung and T. Uthmann. Experiments in value function approximation with sparse support vector regression. In *ECML*, pages 180–191, 2004.

S.M. Kakade. *On the sample complexity of reinforcement learning*. PhD thesis, Gatsby Computational Neuroscience Unit, University College London, 2003.

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

G. Kimeldorf and G. Wahba. Some results on Tchebycheffian spline functions. *J. Math. Anal. Applic.*, 33:82–95, 1971.

M. Kohler and A. Krzyżak. Adaptive regression estimation with multilayer feedforward neural networks. *Journal of Nonparametric Statistics*, 17:891–913, 2005.

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

W.S. Lee, P.L. Bartlett, and R.C. Williamson. Efficient agnostic learning of neural networks with bounded fan-in. *IEEE Transactions on Information Theory*, 42(6):2118–2132, 1996.

Y. Lin and H.H. Zhang. Component selection and smoothing in smoothing spline analysis of variance models. *Annals of Statistics*, 34(5), 2006. (to appear).

F. A. Longstaff and E. S. Shwartz. Valuing american options by simulation: A simple least-squares approach. *Rev. Financial Studies*, 14(1):113–147, 2001.

S. Mahadevan, N. Marchalleck, T. Das, and A. Gosavi. Self-improving factory simulation using continuous-time average-reward reinforcement learning. In *Proceedings of the 14th International Conference on Machine Learning (IMLC '97)*, 1997.

T.L. Morin. Computational advances in dynamic programming. In *Dynamic Programming and its Applications*, pages 53–90. Academic Press, 1978.

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

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

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

A.Y. Ng and M. Jordan. PEGASUS: A policy search method for large MDPs and POMDPs. In *Proceedings of the 16th Conference in Uncertainty in Artificial Intelligence*, pages 406–415, 2000.

P. Niyogi and F. Girosi. Generalization bounds for function approximation from scattered noisy data. *Advances in Computational Mathematics*, 10:51–80, 1999.

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

M.L. Puterman. *Markov Decision Processes — Discrete Stochastic Dynamic Programming*. John Wiley & Sons, Inc., New York, NY, 1994.

D. Reetz. Approximate solutions of a discounted Markovian decision problem. *Bonner Mathematischer Schriften*, 98: Dynamische Optimierungen:77–92, 1977.

M. Riedmiller. Neural fitted Q iteration – first experiences with a data efficient neural reinforcement learning method. In *16th European Conference on Machine Learning*, pages 317–328, 2005.

J. Rust. Numerical dyanmic programming in economics. In H. Amman, D. Kendrick, and J. Rust, editors, *Handbook of Computational Economics*. Elsevier, North Holland, 1996a.

J. Rust. Using randomization to break the curse of dimensionality. *Econometrica*, 65: 487–516, 1996b.

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

A.L. Samuel. Some studies in machine learning using the game of checkers, II – recent progress. *IBM Journal on Research and Development*, pages 601–617, 1967.

N. Sauer. On the density of families of sets. *Journal of Combinatorial Theory Series A*, 13: 145–147, 1972.

S.P. Singh and D.P. Bertsekas. Reinforcement learning for dynamic channel allocation in cellular telephone systems. In *Advances in Neural Information Processing Systems 9*, 1997.

S.P. Singh, T. Jaakkola, and M.I. Jordan. Reinforcement learning with soft state aggregation. In *Proceedings of Neural Information Processing Systems 7*, pages 361–368. MIT Press, 1995.

C.J. Stone. Optimal rates of convergence for nonparametric estimators. *Annals of Statistics*, 8:1348–1360, 1980.

C.J. Stone. Optimal global rates of convergence for nonparametric regression. *Annals of Statistics*, 10:1040–1053, 1982.

Cs. Szepesvári. Efficient approximate planning in continuous space Markovian decision problems. *AI Communications*, 13:163–176, 2001.

Cs. Szepesvári. Efficient approximate planning in continuous space Markovian decision problems. *Journal of European Artificial Intelligence Research*, 2000. accepted.

Cs. Szepesvári and R. Munos. Finite time bounds for sampling based fitted value iteration. In *ICML'2005*, 2005.

Cs. Szepesvári and W.D. Smart. Interpolation-based Q-learning. In D. Schuurmans R. Greiner, editor, *Proceedings of the International Conference on Machine Learning*, pages 791–798, 2004.

G. Tesauro. Temporal difference learning and TD-Gammon. *Communications of the ACM*, 38:58–67, March 1995.

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

J.N. Tsitsiklis and B. Van Roy. Optimal stopping of Markov processes: Hilbert space theory, approximation algorithms, and an application to pricing financial derivatives. *IEEE Transactions on Automatic Control*, 44:1840–1851, 1999.

V. N. Vapnik and A. Ya. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, 16:264–280, 1971.

X. Wang and T.G. Dietterich. Efficient value function approximation using regression trees. In *Proceedings of the IJCAI Workshop on Statistical Machine Learning for Large-Scale Optimization*, Stockholm, Sweden, 1999.

Y. Whang and O. Linton. The asymptotic distribution of nonparametric estimates of the Lyapunov exponent for stochastic time series. *Journal of Econometrics*, 91:1–42, 1999.

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

W. Zhang and T. G. Dietterich. A reinforcement learning approach to job-shop scheduling. In *Proceedings of the International Joint Conference on Artificial Intellience*, 1995.