HAL
open science

Finite Time Bounds for Sampling-Based Fitted Value Iteration

Rémi Munos, Csaba Szepesvari

To cite this version:

Rémi Munos, Csaba Szepesvari. Finite Time Bounds for Sampling-Based Fitted Value Iteration. [Research Report] 2007, pp.47. inria-00120882v2

HAL Id: inria-00120882

https://inria.hal.science/inria-00120882v2

Submitted on 19 Jan 2007 (v2), last revised 5 Mar 2008 (v4)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

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) to solve infinite state-space, discounted-reward Markovian decision problems (MDPs) under the assumption that a generative model of the environment is available. Our main results come in the form of finite-time bounds on the performance of two versions of sampling-based FVI. The convergence rate results obtained allow us to show that both versions of FVI are well behaving in the sense that by using a sufficiently large number of samples for a large class of MDPs, arbitrary good performance can be achieved with high probability. An important feature of our proof technique is that it permits the study of weighted L^p-norm performance bounds. As a result, our technique applies to a large class of function-approximation methods (e.g., neural networks, adaptive regression trees, kernel machines, locally weighted learning), and our bounds scale well with the effective horizon of the MDP. The bounds show a dependence on the stochastic stability properties of the MDP: they scale with the discounted-average concentrability of the future-state distributions. They also depend on a new measure of the approximation capabilities of the function space, the inherent Bellman residual, which reflects how well the function space is "aligned" with the essential characteristics of the dynamics of the MDP. The conditions of the main result, as well as the results are extensively discussed and compared to previous theoretical results. Numerical experiments are used to substantiate the theoretical findings.

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 control problems, such as job-shop scheduling (Zhang and Dietterich, 1995), backgammon (Tesauro, 1995), elevator control (Crites and Barto, 1997), machine maintenance (Mahadevan et al., 1997), dynamic channel allocation (Singh and Bertsekas, 1997), or airline seat allocation (Gosavi, 2004). The set of possible states in

these problems is very large, and so only algorithms that can successfully generalize to unseen states are expected to achieve non-trivial performance. The approach in the above mentioned works is to learn an approximation to the optimal value function that assigns to each state the best possible expected long-term cumulative reward resulting from starting the control from the selected state. The knowledge of the optimal value function is sufficient to achieve optimal control (Bertsekas and Tsitsiklis, 1996), and in many cases an approximation to the optimal value function is already sufficient to achieve a good control performance. In large state-space problems, a function approximation method is used to represent the learnt value function. In all the above successful applications this is the approach used. Yet, the interaction of RL algorithms and function approximation methods is still not very well understood.

Our goal in this paper is to improve this situation by studying one of the simplest ways to combine RL and function approximation, namely, using function approximation in value iteration. The advantage of studying such a simple combination is that some technical difficulties are avoided, yet, as we will see, the problem studied is rich enough to make its study worthwhile.

Value iteration is a dynamic programming algorithm which uses 'value backups' to generate a sequence of value functions (i.e., functions defined over the state space) in a recursive manner. After a sufficiently large number of iterations the obtained function can be used to compute a good policy. Exact computations of the value backups require computing parametric integrals over the state space. Except in a few special cases, neither such exact computations, nor the exact representation of the resulting functions is possible. The idea underlying *sampling-based fitted value iteration* (FVI) is to calculate the back-ups approximately using Monte-Carlo integration at a finite number of points and then find a best fit within a user-chosen set of functions to the computed values. The hope is that if the function set is rich enough then the policy obtained will have a good performance, as observed experimentally by a number of researchers working with different variants of sampling-based FVI (e.g., Wang and Dietterich, 1999; Dietterich and Wang, 2002; Lagoudakis and Parr, 2003; Jung and Uthmann, 2004; Ernst et al., 2005; Riedmiller, 2005).

Despite these numerous successful empirical demonstrations, it was still unclear whether sampling-based FVI is indeed a "good" algorithm. In particular, Baird (1995) and Tsitsiklis and Van Roy (1996) gave simple counterexamples showing that FVI can be instable. These counterexamples are deceptively simple: the MDP is finite, exact backups can be and are calculated. the approximate value function is calculated using a linear combination of a number of fixed basis functions and the optimal value function can be represented exactly by such a linear combination. Hence, the function set seems sufficiently rich. Despite this, the iterates diverge. Since value iteration without projection is well-behaving, 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 find that sampling-based FVI is well behaving and, in particular, to work out a theory explaining when to expected good performance.

The setting studied in this paper is as follows: We assume that the state-space is a compact subset of a Euclidean space and that the MDP has a finite number of actions. The criterion of optimality is to maximize the expectation of the infinite-horizon, discounted

sum of rewards. We shall also assume that the algorithm can sample any transition from any initial state, i.e., that a *generative model* (or simulator) of the environment is available. This model has been used in a number of previous works (e.g., Kearns et al., 1999; Ng and Jordan, 2000; Kakade and Langford, 2002; Kakade, 2003). An extension of the present work to the case when only a single trajectory is available for learning is published elsewhere (Antos et al., 2006).

We investigate two versions of the basic algorithm: In the *multi-sample variant* a fresh sample set is generated in each iteration, while in the *single-sample variant* the same sample set is used throughout all the iterations. Interestingly, we find that no serious degradation of performance results from reusing the samples. In fact, we find that when the discount factor is close to one then the single-sample variant can be expected to be more efficient in the sense of yielding smaller errors using fewer samples.

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. We will compare our bounds to those available in supervised learning (regression), where alike bounds have two terms: one bounding the *bias* of the algorithm, while the other bounding the *variance*, or *estimation error*. The term bounding the bias decays when the *approximation power* of the function class is increased (hence this term is occasionally called the approximation error term). The term bounding the *variance* decays as the number of samples is increased, but increases when the richness of the function class is increased.

Although our bounds are similar to bounds of supervised learning, they are some notable differences. In regression estimation, the approximation power of the function set is usually measured with respect to 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 $\mathcal{G}$ is typically a classical smoothness class, such as a Lipschitz space. This measure is inadequate for our purposes since in the counterexamples of Baird (1995) and Tsitsiklis and Van Roy (1996) the target function (whatever function space it belongs to) can be approximated with zero error, but FVI still exhibits unbounded errors. In fact, our bounds use a different characterization of the approximation power of the function class $\mathcal{F}$, which we call its *inherent Bellman error*:

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

Here T is the Bellman operator underlying the MDP (capturing the essential properties of the MDP's dynamics) and $\|\cdot\|$ is an appropriate weighted p-norm that is chosen by the user (the exact definitions will be given in Section 2). Observe that no external reference class is used in the definition of $d(T\mathcal{F}, \mathcal{F})$: the inherent Bellman error reflects how well the function space $\mathcal{F}$ is 'aligned' to the Bellman operator, i.e., the dynamics of the MDP. In the above-mentioned counterexamples the inherent Bellman error of the function space chosen is infinite and so the bound (correctly) indicates the possibility of divergence.

The bounds on the variance are more similar to their regression counterpart: Just like in regression our variance bounds depend on the capacity (such as the VC-dimension) of the

function space used and decay polynomially with the number of samples. However, the rate of decay is slightly worse than the corresponding rate of (optimal) regression procedures. The difference comes from the inherent biasedness of the samples obtained in the Monte-Carlo estimation step. Nevertheless the bounds still indicate that the maximal error of the procedure in the limit when the number of samples grows to infinity converges to a finite positive number, which is in turn controlled by the inherent Bellman error of the function space.

Our bound also displays the *bias-variance tradeoff*: In order to keep both the approximation and estimation errors small the number of samples and the power of the function class has to be increased simultaneously. When this is done in a principled way, the resulting algorithm becomes *consistent*: It's error in the limit disappears for a large class of MDPs. Consistency is an important property of any MDP algorithms. If an algorithm fails to prove to be consistent, we would be suspicious about its use in any practical application.

Our bounds apply to MDPs satisfying the property that the *discounted-average concentrability of future-state distributions* is finite. The precise definition will be given in Section B.4, here we note in passing that this condition holds trivially in every finite MDP, and also, more generally, if the MDP's transition kernel possesses a bounded density. This latter condition has been considered in many previous theoretical works (e.g., Chow and Tsitsiklis, 1989, 1991; Rust, 1996b; Szepesvári, 2000) and this class of MDPs is quite large in the sense that they contain hard instances (see the discussion in Section 8). We also show a connection between Lyapunov exponents, well known from the stability analysis of dynamical systems, and our concentrability factor.

Our proofs build on a recent technique proposed by Munos (2003) that allows the propagation of weighted p-norm losses in approximate value iteration. In contrast, most previous analysis of FVI relied on propagating errors w.r.t. the maximum norm (Gordon, 1995; Tsitsiklis and Van Roy, 1996). The advantage of using p-norm loss bounds is that it allows the analysis of algorithms that use p-norm fitting (in particular, 2-norm fitting). Unlike Munos (2003) (and Munos 2005), we explicitly deal with infinite state-spaces, the effects of using a finite random sample, i.e., the bias-variance dilemma and the consistency of sampling-based FVI.

The paper is organized as follows: In the next section we introduce the concepts and notation used in the rest of the paper. The problem is formally defined and the algorithms are given in Section 3. Next, we develop finite-sample bounds for the error committed in a single iteration (Section 4). This bound is used in proving our main results in Section 5. We extend these results to the problem of obtaining a good policy in Section 6, followed by a construction that allows one to achieve asymptotic consistency when the unknown MDP is smooth with an unknown smoothness factor (Section 7). Relationship to previous works is discussed in detail in Section 8. An experiment in a simulated environment, highlighting the main points of the analysis is given in Section 9. The 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 system of Borel-measurable sets of $\mathcal{X}$ shall be denoted by $\mathcal{B}(\mathcal{X})$.

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 $X_{t+1} \sim \mathbb{P}\left(\cdot|X_t, A_t\right)$, $R_t \sim \mathbb{P}\left(\cdot|X_t, A_t\right)$. 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.[1]

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 $x \in \mathcal{X}$, that is 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

1. This condition could be replaced by a standard moment condition on the random rewards (Györfi et al., 2002) without changing the results.
2. In fact, π_t must be a measurable mapping so that we are allowed to talk about the probability of executing an action. Measurability issues by now are well understood and hence we shall not deal with them here. For a complete treatment the interested reader is referred to Bertsekas and Shreve (1978).

state x, the first executed action is a and it is assumed that after the first step an optimal policy is followed. Since by assumption the action space is finite and the rewards bounded and we assume discounting, the existence of deterministic stationary optimal policies is guaranteed (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

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

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 $r(x,a) = \int z S(dz|x,a)$ 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 $(x,a) \in \mathcal{X} \times \mathcal{A}$.

Let μ be a distribution over $\mathcal{X}$. 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 $L^p(\mathcal{X}; \mu)$.

3. Sampling-based Fitted Value Iteration

The parameters of sampling-based FVI are a distribution, $\mu \in M(\mathcal{X})$, a function set $\mathcal{F} \subset B(\mathcal{X})$, some initial value function, $V_0 \in \mathcal{F}$, and the integers N, M and K. The algorithm works by computing a series of functions, $V_1, V_2, \ldots \in \mathcal{F}$ in a recursive manner. The $(k+1)$th iterate 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 the distribution μ, 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:

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

$(j = 1, 2, \ldots, M,\ i = 1, \ldots, N)$. By assumption, $(Y_j^{X_i,a}, R_j^{X_i,a})$ and $(Y_{j'}^{X_{i'},a'}, R_{j'}^{X_{i'},a'})$ are independent of each other whenever $(i,j,a) \neq (i',j',a')$. 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

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

These steps are iterated $K > 0$ times, yielding the sequence $V_1, \ldots, V_K$.

We study two variants of this algorithm: When a fresh sample is generated in each iteration, we call the algorithm the *multi-sample variant*. The *total number* of samples used by the multi-sample algorithm is thus $K \times N \times M$. 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 the *single-sample variant* (details 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. 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 and can be expected to outperform the multi-sample variant. In the experiments we will see such a case.

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

$$\mathcal{F} = \{\, f_\theta \in B(\mathcal{X}) \,|\, \theta \in \Theta \,\}, \quad \dim(\Theta) < +\infty.$$

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 wavelet based approximations, multi-layer neural networks or kernel-based regression techniques. Another possibility is to use the kernel mapping idea underlying many recent state-of-the-art supervised-learning methods, such as support-vector machines, support-vector regression or Gaussian processes (Cristianini and Shawe-Taylor, 2000). Given a (positive definite) kernel function $\mathcal{K}$, $\mathcal{F}$ can be chosen as the closed convex subset of the reproducing-kernel Hilbert-space (RKHS) associated to $\mathcal{K}$. In this case the optimization problem (1) still admits a finite, closed-form solution despite that the function space $\mathcal{F}$ cannot be written in the above form for any finite dimensional parameters space (Kimeldorf and Wahba, 1971; Schölkopf and Smola, 2002).

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 $p \geq 1$. The *metric projection* of TV onto $\mathcal{F}$ w.r.t. the μ-weighted p-norm is defined by

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

Here $\Pi_{\mathcal{F}} : L^p(\mathcal{X};\mu) \to \mathcal{F}$ for $g \in L^p(\mathcal{X};\mu)$ gives the *best approximation* to g in $\mathcal{F}$.[3] The *approximation error* in the kth step for $V = V_k$ is $d_{p,\mu}(TV,\mathcal{F}) = \|\Pi_{\mathcal{F}}TV - TV\|_{p,\mu}$. More generally, we let

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

Hence, the approximation error can be made small by selecting $\mathcal{F}$ to be large enough. We shall discuss how this can accomplished for a large class of MDPs in Section 7.

In the kth iteration, given $V = V_k$, the function $V' = V_{k+1}$ is computed as follows:

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

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

where the random samples satisfy the conditions of the previous section and, for the sake of simplicity, we assume that the minimizer in Equation (4) exists.

Clearly, for X_i fixed, $1/M\sum_{j=1}^{M}(R_j^{X_i,a} + \gamma V(Y_j^{X_i,a})) \to (T_aV)(X_i)$ as $M \to \infty$ w.p.1. Hence, for M large enough, $\hat{V}(X_i)$ is a good approximation to $(TV)(X_i)$. On the other hand, if N is big then for any $(f,g) \in \mathcal{F}\times\mathcal{F}$, the *empirical p-norm 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 for N, M big, V' can be expected to be close to $\Pi_{\mathcal{F}}TV$.

Note that Equation (4) looks like an ordinary p-norm function fitting. One difference though is that in regression the target function equals the regressor $g(x) = E[\hat{V}(X_i)|X_i = x]$, whilst in our case the target function is TV and $TV \neq g$. This is because

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

In fact, if we had equality here then we would be no reason to set $M > 1$: in a pure regression setting it is always better to have a completely fresh pair of samples than to have a pair where the covariate is set to be equal to some previous sample. Due to $M > 1$ the rate of sampling-based FVI will be slightly worse than the rates available for regression.

Above we argued that for N large enough and for a fixed pair $(f,g) \in \mathcal{F}\times\mathcal{F}$, the empirical loss will approximate the true loss, i.e., the estimation error will be small. 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 a random object by itself and so the argument ensuring that the estimation error is small for any *fixed*, deterministic pair of functions cannot be used for V'. This is, however, the situation in supervised learning problems, too and so we

3. 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.

shall use techniques developed there. A simple idea is to bound the estimation error at the random function V' by the worst-case estimation error over $\mathcal{F}$, which can be done thanks to the inequality

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

which holds for any $g \in L^p(\mathcal{X};\mu)$, with probability one. The inequality holds because by construction, for any random event ω, $V' = V'(\omega)$ 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 developed there.

When bounding the size of maximal deviations, the size of the function set becomes a major factor. When the function set has a finite number of elements, a bound follows by exponential tail inequalities and a union bounding argument. When $\mathcal{F}$ is infinite, the 'capacity' of $\mathcal{F}$ measured by the *covering number* of $\mathcal{F}$ can be used to derive an appropriate bound: Let $\epsilon > 0$, $q \ge 1$, $x^{1:N} \stackrel{\text{def}}{=} (x_1,\dots,x_N) \in \mathcal{X}^N$ be fixed. 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 when bounding maximal deviations is how much the functions in the function space vary *at the actual samples.* Of course, without imposing any conditions on the function space, covering numbers can grow as a function of the sample size.

For specific choices of $\mathcal{F}$, however, it is possible to bound the covering numbers of $\mathcal{F}$ *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. 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 from 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 any $\epsilon>0$,*

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

4. The VC-dimension of a set system is defined as follows (Sauer, 1972; Vapnik and Chervonenkis, 1971): Given a set system $\mathcal{C}$ with base set $\mathcal{U}$ we say that $\mathcal{C}$ shatters the points of $A\subset\mathcal{U}$ 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 $A\subset\mathcal{U}$ that can be shattered.

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, neither the pseudo-dimension nor covering numbers are changed by shifts. This allows one to extend Proposition 1 to function sets with functions taking values in $[-L, +L]$.

Bounds on the pseudo-dimension are known for many popular 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). If q is the dimensionality of the function space, these bounds take the form $O(\log(q))$, $O(q)$ or $O(q \log q)$.[5]

Another route to get a useful bound on the number of samples is to derive an upper bound on the metric entropy that grows with the number of samples at a *sublinear* rate. As an example consider the following class of bounded-weight, linearly parameterized functions:

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

It is known that for finite-dimensional smooth parametric classes their metric entropy scales with $\dim(\phi) < +\infty$. If ϕ is the feature map associated with some positive definite kernel function $\mathcal{K}$ then ϕ can be infinite dimensional (this class arises if one 'kernelizes' FVI). In this case the bounds due to Zhang (2002) can be used. These bound the metric entropy by $\lceil A^2 B^2/\epsilon^2 \rceil \log(2N+1)$, where B is an upper bound on $\sup_{x\in\mathcal{X}} \|\phi(x)\|_p$ with $p = 1/(1-1/q)$.[6]

4.1 Finite-sample Bounds

The following lemma shows that with high probability, V' is a good approximation to TV when some element of $\mathcal{F}$ is close to TV and if the number of samples is high enough:

Lemma 1 *Consider an MDP satisfying Assumption A0. Let $V_{\max} = R_{\max}/(1-\gamma)$, fix a real number $p \ge 1$, integers $N, M \ge 1$, $\mu \in M(\mathcal{X})$ and $\mathcal{F} \subset B(\mathcal{X}; V_{\max})$. Pick any $V \in B(\mathcal{X}; V_{\max})$ and let $V' = V'(V, N, M, \mu, \mathcal{F})$ be defined by Equation (4). Let $\mathcal{N}_0 = \mathcal{N}\left(\frac{V_{\max}}{8p}\left(\frac{\epsilon}{4V_{\max}}\right)^p, \mathcal{F}, N, \mu\right)$. 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

$$N > 2\left(\frac{64 V_{\max}}{\epsilon}\right)^{2p}\Big(\log(1/\delta) + \log(32\mathcal{N}_0)\Big) \tag{8}$$

and

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

As we have seen before, for a large number of choices of $\mathcal{F}$, the metric entropy of $\mathcal{F}$ is independent of N. In such cases Equation (8) gives an explicit bound on N and M. In the

5. Again, similar bounds are known to hold for the supremum-norm metric entropy, as well.
6. Similar bounds exist for the supremum-norm metric entropy.

resulting bound, the total number of samples per iteration, $N \times M$, scales with $\epsilon^{-(2p+2)}$ (apart from logarithmic terms). The comparable bound for the pure regression setting is ϵ^{-2p}. The additional quadratic factor is the price to pay because of the biasedness of the values $\hat{V}(X_i)$.

The main ideas of the proof are illustrated using Figure 1 (the proof of the lemma can be found in Appendix A). The left-hand side of this figure depicts the space of bounded

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 $\mathcal{F}$ (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.

functions over $\mathcal{X}$, while the right-hand side figure shows a corresponding vector space. The spaces are connected by the mapping $f \mapsto \tilde{f} \stackrel{\text{def}}{=} (f(X_1), \ldots, f(X_N))^T$. In particular, this mapping sends the set $\mathcal{F}$ into the set $\tilde{F} = \{\tilde{f} \,|\, f \in \mathcal{F}\}$.

The proof goes by upper bounding the distance between V' and TV in terms of the distance between f^* and TV. Here $f^* = \Pi_{\mathcal{F}} TV$ is the best fit to TV in $\mathcal{F}$. The choice of f^* is motivated by the fact that V' can be seen as 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 $\|\cdot\|_p$. The bound is developed 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. Now, 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 of between $\widetilde{f^*}$ and $\widetilde{TV}$, which in turn must be close to the $L^p(\mathcal{X}; \mu)$ distance of f^* and TV if N is big enough. Hence, if $\|f^* - TV\|_{p,\mu}^p$ is small then so is $\|V' - TV\|_{p,\mu}^p$.

4.2 Bounds for the Single-Sample Variant

When analysing the error of sampling-based FVI, we would like to use Lemma 1 for bounding the error committed when approximating TV_k starting from V_k based on a new sample. When doing so, however, we have to take into account that V_k is random. Yet Lemma 1 requires that V, the function whose Bellman image is approximated, to be some fixed (non-random) function. The problem, however, is easily resolved by noting that the samples used in calculating V_{k+1} are independent of the samples used to calculate V_k. A formal argument is presented in Appendix B.3. The same argument, however, does not work for the single-sample variant of the algorithm when V_{k+1} and V_k are *both* computed using the *same set of random variables*. The purpose of this section is to extend Lemma 1 to cover this case.

In formulating this result we will need the following definition: For $\mathcal{F} \subset B(\mathcal{X})$ let us define

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

The following result holds:

Lemma 2 *Denote by Ω the sample-space underlying the random variables $\{X_i\}$, $\{Y_j^{X_i,a}\}$, $\{R_j^{X_i,a}\}$, $i = 1, \dots, N, j = 1, \dots, M, a \in \mathcal{A}$. Then the result of Lemma 1 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 $\mathcal{F}$.*

The proof can be found in Appendix A.1. Note that the sample-size bounds in this lemma are similar to those of Lemma 1, except that N now depends on the metric entropy of $\mathcal{F}_{T-}$ and M depends on the metric entropy of $\mathcal{F}$. Let us now give two examples when explicit bounds on the covering number of $\mathcal{F}_{T-}$ can be given using simple means:

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 $G = 1$, $\mathcal{N}(\epsilon, \mathcal{F}_{T-}, n, \mu) \le \mathcal{N}(\epsilon, \mathcal{F}, n, \mu)\mathcal{N}(\epsilon, \mathcal{F}_T, n, \mu)$. Hence it suffices to bound the covering numbers 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 $\|\|\phi(\cdot)\|_2\|_\infty$ (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 $B_1, \ldots, B_m$ with respective centers $c_1, \ldots, c_m$. Clearly, $\{T(B_i)\}_{i=1,\ldots,n}$ 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 would 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 one such ball, 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 enjoys 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 e.g. in financial applications.[7]

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

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

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

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

7. A time discretization of a (stochastic) differential equation $dX_t = f(X_t, u_t)dt + \sigma(X_t, u_t)dW_t$ satisfying a uniform elliptic condition on the noise term (ie. $\sigma\sigma' \ge \lambda I$, where $\lambda > 0$), 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 deterministic case a similar conclusion holds for the $O(\Delta t)$ neighborhood.

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 instead of an optimal one, where 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 on those parts of the state space where performance matters more. 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 small as they are propagated through the algorithm. Since the previous section's bounds are given in terms of weighted p-norms, it is natural to develop weighted p-norm bounds for the whole algorithm. Let us concentrate on the case when in all iterations the error committed is bounded. Since we use weighted p-norm bounds, the usual supremum-norm analysis does not work. However, a similar argument can be used.

The sketch of this argument is as fllows: Since we are interested in developing a bound on the performance of the greedy policy w.r.t. the final estimate of V^*, we first develop a pointwise analogue of supremum-norm Bellman-error bounds:

$$(I - \gamma P^\pi)(V^* - V^\pi) \le \gamma(P^{\pi^*} - P^\pi)(V^* - V).$$

Here V plays the role of the final value function estimate and π is a greedy policy w.r.t. V and V^π is its value-function. Hence, we see that it suffices to develop upper and lower bounds on $V^* - V$ with $V = V_K$. For the upper estimate, we use that $V^* - TV_k = TV^* - TV_k = T^{\pi^*}V^* - T^{\pi_k}V_k \le T^{\pi^*}V^* - T^{\pi^*}V_k = \gamma P^{\pi^*}(V^* - V_k)$. Hence, if $V_{k+1} = TV_k - \epsilon_k$ then $V^* - V_{k+1} \le \gamma P^{\pi^*}(V^* - V_k) + \epsilon_k$. An analogous reasoning results in the lower bound $V^* - V_{k+1} \ge \gamma P^{\pi_k}(V^* - V_k) + \epsilon_k$. Here π_k is a policy greedy w.r.t. V_k. Now, exploiting that the operator P^π is linear for any π, iterating these bounds yields upper and lower bounds on $V^* - V_K$. A crucial step of the argument is to replace T, the non-linear Bellman operator by linear operators (P^π, for suitable π) since propagating errors through linear operators is easy, while in general, it is impossible to do the same with non-linear operators. Actually, as we propagate the errors, it is not hard to foresee that operator products of the form $P^{\pi_K}P^{\pi_{K-1}}\dots P^{\pi_k}$ enter our bounds and that the amplification caused by these product operators is the major source of the possible increase of the error.

Note that if a supremum-norm analysis were followed ($p = \infty$), we would immediately find that the maximum amplification by these product operators bounded by one: Since, as it is well known, for any policy π, $|\int V(y)P(dy|x,\pi(x))| \le \int |V(y)|P(dy|x,\pi(x)) \le \|V\|_\infty \int P(dy|x,\pi(x)) = \|V\|_\infty$, i.e., $\|P^\pi\|_\infty \le 1$, and since

$$\|P^{\pi_K}\dots P^{\pi_k}\|_\infty \le \|P^{\pi_K}\|_\infty \dots \|P^{\pi_k}\|_\infty \le 1,$$

starting from the pointwise bounds, one recovers the well-known supremum-norm bounds by just taking the supremum of bound's two sides. Hence, the pointwise bounding technique

yields as tight bounds as the previous supremum-norm bounding technique. However, since in the algorithm only the weighted p-norm errors are controlled, instead of taking the pointwise supremum, we integrate the pointwise bounds with respect to the measure ρ to derive the desired p-norm bounds provided that the induced operator-norm of these operator products w.r.t. weighted p-norms can be bounded. One simple assumption that allows this is as follows:

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 $\|P^\pi V\|_{p,\rho} \le C_\mu^{1/p} \|V\|_{p,\mu}$ and thus $\|P^\pi\|_{p,\rho} \le C_\mu^{1/p}$. 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. 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 quite large in the sense that there are hard instances in it (this is discussed in detail in Section 8). However, the above assumption certainly excludes completely or partially deterministic MDPs, which might be important, for example, in financial applications.

Let us 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 [Discounted-average concentrability of future-state distributions] Given ρ, μ, $m \ge 1$ and an arbitrary sequence of stationary policies $\{\pi_m\}_{m\ge1}$, assume that the future-state distribution $\rho P^{\pi_1} P^{\pi_2} \dots P^{\pi_m}$ is absolutely continuous w.r.t. μ. Assume that

$$c(m) = \sup_{\pi_1,\dots,\pi_m} \left\| \frac{d(\rho P^{\pi_1} P^{\pi_2} \dots P^{\pi_m})}{d\mu} \right\|_\infty \tag{12}$$

satisfies that

$$C_{\rho,\mu} = (1-\gamma)^2 \sum_{m\ge1} m\gamma^{m-1} c(m) < +\infty.$$

We shall call $c(m)$ the m-step concentrability of a future-state distribution, while we call $C_{\rho,\mu}$ the *discounted-average concentrability coefficient* of the future-state distributions.
The number $c(m)$ measures how much ρ can get amplified in m steps as compared to the reference distribution μ. Hence, in general we expect $c(m)$ to grow with m. In fact, the condition that $C_{\rho,\mu}$ is finite is a growth rate condition on $c(m)$. Thanks to discounting,

$C_{\rho,\mu}$ is finite for a reasonably large class of systems: In fact, we will now argue that Assumption A2 is weaker than Assumption A1 and that $C_{\rho,\mu}$ is finite when the top-Lyapunov exponent of the MDP is finite.

To show the first statement it sufficies to see that $c(m) \le C_\mu$ holds for any m. However, this holds since by definition for any distribution ν and policy π, $\nu P^\pi \le C_\mu \mu$. Then take $\nu = \rho P^{\pi_1} \dots P^{\pi_{m-1}}$ and $\pi = \pi_m$ to conclude that $\rho P^{\pi_1} \dots P^{\pi_{m-1}} P^{\pi_m} \le C_\mu \mu$ and so $c(m) \le C_\mu$.

Let us now turn to the comparison with the top-Lyapunov exponent of the MDP. As our starting point we take the definition of top-Lyapunov exponent associated with sequences of finite dimensional matrices: If $\{P_t\}_t$ is sequence of square matrices with non-negative entries and $\{y_t\}_t$ is a sequence of vectors that satisfy $y_{t+1} = P_t y_t$ then, by definition, the top-Lyapunov exponent is $\hat{\gamma}_{\text{top}} = \limsup_{t\to\infty} (1/t) \log^+(\|y_t\|_\infty)$. If the top-Lyapunov exponent is positive then the associative system is sensitive to its initial conditions (unstable). A negative Lyapunov exponent, on the other hand, indicates that the system is stable; in case of certain stochastic systems the existence of strictly stationary non-anticipating realizations is equivalent to a negative Lyapunov exponent Bougerol and Picard (1992).[8]

Now, one may think of y_t as a probability distribution over the state space and the matrices as the transition probabilities. One way to generalize the above definition to controlled systems and infinite state spaces is to identify y_t with the future state distribution when the policies are selected to maximize the growth rate of $\|y_t\|_\infty$. This gives rise to $\hat{\gamma}_{\text{top}} = \limsup_{m\to\infty} \frac{1}{m} \log c(m)$, where $c(m)$ is defined by (12). Then, by elementary arguments, we get that if $\hat{\gamma}_{\text{top}} < \log(1/\gamma)$ then $\sum_{m\ge 0} m^p \gamma^m c(m) < \infty$. In fact, if $\hat{\gamma}_{\text{top}} \le 0$ then $C(\rho,\nu) < \infty$. Hence, $C(\rho,\nu) < +\infty$ can be interpreted as a weak stability condition.

Since Assumption A1 is stronger than Assumption A2 in the proofs we will proceed by first developing a proof under Assumption A2. The reason Assumption A1 is still considered is that it will allow us to derive supremum-norm performance bounds even if though in the algorithm we control only the weighted p-norm 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 a natural way: As we have seen the error in iteration k depends on $d_{p,\mu}(T\hat{V}_k,\mathcal{F})$. Since $\hat{V}_k \in \mathcal{F}$, the inherent Bellman error gives a uniform bound on the errors of the individual iterations.[9]

The next theorem is the main result of the paper. It 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 of the function space $\mathcal{F}$ as desired by selecting a sufficiently high number of samples. Hence, sampling-based FVI can be used to find near-optimal policies if $\mathcal{F}$ is sufficiently rich:

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

9. More generally, $d_{p,\mu}(\mathcal{G},\mathcal{F}) \stackrel{\text{def}}{=} \sup_{g\in\mathcal{G}} d_{p,\mu}(g,\mathcal{F}) \stackrel{\text{def}}{=} \sup_{g\in\mathcal{G}} \inf_{f\in\mathcal{F}} \|g-f\|_{p,\mu}$.

Theorem 2 *Consider an MDP satisfying Assumption A0 and A2. Fix $p \geq 1$, $\mu \in M(\mathcal{X})$ and let $V_0 \in \mathcal{F} \subset B(\mathcal{X}; V_{\max})$. 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 if the multi-sample variant of sampling-based FVI is run with parameters $(N, M, \mu, \mathcal{F})$ and π_K is a policy greedy with respect to the Kth iterate then 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 in addition Assumption A1 holds 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, when N, M selected as before just 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. Assuming that the pseudo-dimension of the function-space $\mathcal{F}$ is finite as in Proposition 1, a close examination of the proof gives the following high-probability bound for the multi-sample variant:

$$\begin{aligned} \|V^* - V^{\pi_K}\|_{p,\rho} &\leq \frac{2\gamma}{(1-\gamma)^2} C_{\rho,\mu}^{1/p} \; d_{p,\mu}(T\mathcal{F}, \mathcal{F}) + O(\gamma^K V_{\max}) \\ &+ \; O\left\{ \left(\frac{V_{\mathcal{F}^+}}{N} \left(\log(N) + \log(K/\delta)\right) \right)^{1/2p} + \left(\frac{1}{M} \left(\log(N|\mathcal{A}|) + \log(K/\delta)\right) \right)^{1/2} \right\}. \quad (13) \end{aligned}$$

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, while the last two terms bound the estimation error.

This form of the bound allows us to reason about the likely best 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 (apart from constants) 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. Note that the choice of N, M does not influence the other error terms.

Now, let us consider the single-sample variant of FVI. A careful inspection of the proof results in an inequality identical to (13) 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 given a fixed budget of n samples provided that $K > V_{\mathcal{F}^-}$ the bound for the single-sample variant is better than the one for the multi-sample

variant. In both cases a logical choice is to set K to minimize the respective bounds. In fact, the optimal choice turns out to be $K \sim 1/\log(1/\gamma) \approxeq 1/(1-\gamma)$ in both cases. 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. That the single-sample variant becomes more efficient is because the variance reduction effect of sample reuse is stronger than the increase in bias. Our computer simulations (Section 9) confirm this experimentally.

Another way to use the above bound is to make comparisons with the rates available in non-parametric regression: First, notice that the approximation error of $\mathcal{F}$ is defined as the inherent Bellman error of $\mathcal{F}$ instead of using an external reference class. This seems reasonable since we are trying to find an approximate fixed point of T within $\mathcal{F}$. The estimation error, for a sample size of n, can be seen to be bounded by $O(n^{-1/(2(p+1))})$, which for $p = 2$ gives $O(n^{-1/6})$. In regression, the comparable error (when using a bounding technique similars to ours) is bounded by $n^{-1/4}$ (Györfi et al., 2002). With considerably more work, using the techniques of Lee et al. (1996) (see also Chapter 11 of (Györfi et al., 2002)) in regression it is possible to get a rate of $n^{-1/2}$, at the price of multiplying the approximation error by a constant larger than one. It seems possible to use these techniques to improve the exponent of N from $-1/2p$ to $-1/p$ in Equation (13) (at the price of increasing the influence of the approximation error). Then the new rate would become $n^{-1/4}$. This is still worse than the best possible rate for non-parameteric regression. The additional factor comes from the need to use the samples to control the bias of the target values. Thus, in the case of FVI, the inferior rate as compared with regression seems unavoidable. By switching from state value functions to action-value functions it seems quite possible to eliminate this inefficiency. This improvement, however, must balance the capacity increase of the function space which depends on the size of the action-space.

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, 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 based on V_K.

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 draw an α-greedy action with probability at least $1-\lambda$ by executing the following procedure: Let $R_j^{x,a} \sim S(\cdot, x, a)$, $Y_j^{x,a} \sim P(\cdot|x,a)$, $j = 1, 2, \ldots, M'$ with $M' = M'(\alpha, \lambda)$ and compute the approximate value of a at state x using

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

Let the policy $\pi^K_{\alpha,\lambda} : \mathcal{X} \to \mathcal{A}$ be defined by

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

The following result holds:

Theorem 3 *Consider an MDP satisfying Assumptions A0 and A2. Fix $p \geq 1$, $\mu \in M(\mathcal{X})$ and let $V_0 \in \mathcal{F} \subset B(\mathcal{X}; V_{\max})$. Select $\alpha = (1-\gamma)\epsilon/8$, $\lambda = \epsilon/8(1-\gamma)/V_{\max}$ and let $M' = O(|\mathcal{A}|\hat{R}^2_{\max}\log(|\mathcal{A}|/\lambda)/\alpha^2)$. 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 $1/\epsilon$, $\log(1/\delta)$, $1/(1-\gamma)$, $V_{\max}$, $\hat{R}_{\max}$, $\log(|\mathcal{A}|)$, $\log(\mathcal{N}(c\epsilon(1-\gamma)^2/C^{1/p}_{\rho,\mu}), \mathcal{F}, \mu)))$ for some $c > 0$, such that if $\{V_k\}_{k=1}^K$ are the iterates generated by multi-sample FVI with parameters $(N, M, K, \mu, \mathcal{F})$ then the policy $\pi^K_{\alpha,\lambda}$ defined above, with probability at least $1-\delta$, we have*

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

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

The proof can be found in Appendix C.

A similar result holds for the single-sample variant of FVI. 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

A highly desirable property of any learning algorithm is that as the number of samples grows to infinity, the error of the algorithm should converge to zero; in other words, the algorithm should be *consistent*. Sampling based FVI with a *fixed* function space $\mathcal{F}$ is not consistent: Our previous results show that in such a case the loss converges to $\frac{2\gamma}{(1-\gamma)^2} C^{1/p}_{\rho,\mu} d_{p,\mu}(T\mathcal{F}, \mathcal{F})$. A simple idea to remedy this situation is to let the function space grow with the number of samples. In regression the corresponding method was proposed by Grendander (1981) and is called the *method of sieves*. The purpose of this section is to show that FVI combined with this method gives a consistent algorithm for a large class of MDPs, namely for those that have Lipschitzian rewards and transitions. It is important to emphasize that although the results in this section assume these smoothness conditions, the method itself does not require the knowledge of the Lipschitz factors. It is left for future work to determine whether similar results hold for larger classes of MDPs.

The smoothness of the transition probabilities and rewards is defined with respect to changes in the initial state:

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

Here $\alpha, L_p, L_r > 0$ are the unknown smoothness parameters of the MDP and $B \subset \mathcal{X}$ is Borel-measurable.

The method is built on the following observation: 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 then $T\mathcal{F}_n \stackrel{\text{def}}{=} \{TV \,|\, V \in \mathcal{F}_n\}$ contains L-Lipschitz $V_{\max}$-bound functions only:

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

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

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

Hence if we make the right-hand side converge to zero as $n \to \infty$ then so will do $d_{p,\mu}(T\mathcal{F}_n, \mathcal{F}_n)$. The quantity, $d_{p,\mu}(\mathrm{Lip}(\alpha; L, V_{\max}), \mathcal{F}_n)$ is nothing but the approximation error of functions in the Lipschitz class $\mathrm{Lip}(\alpha; L, V_{\max})$ by elements of $\mathcal{F}_n$. Now, $d_{p,\mu}(\mathrm{Lip}(\alpha; L, V_{\max}), \mathcal{F}_n) \le d_{p,\mu}(\mathrm{Lip}(\alpha; L), \mathcal{F}_n)$, where $\mathrm{Lip}(\alpha; L)$ is the set of Lipschitz-functions with Lipschitz constant L and we exploited that $\mathrm{Lip}(\alpha, L) = \cup_{V_{\max}>0}\mathrm{Lip}(\alpha; L, V_{\max})$ which holds thanks to the compactness of $\mathcal{X}$. In approximation theory an approximation class $\{\mathcal{F}_n\}$ is said to be *universal* if for any $\alpha, L > 0$, $\lim_{n\to\infty} d_{p,\mu}(\mathrm{Lip}(\alpha; L), \mathcal{F}_n) = 0$. For a large variety of approximation classes (e.g., approximation by polynomials, wavelets, function dictionaries) not only universality is established, but variants of Jackson's theorem give us rates of convergence of the approximation error: $d_{p,\mu}(\mathrm{Lip}(\alpha; L), \mathcal{F}_n) = O(Ln^{-\alpha})$ (e.g DeVore, 1997).

One remaining issue is that classical approximation spaces are not uniformly bounded (i.e., the functions in them do not assume a uniform bound), while 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. One solution is to use truncations: Let $\mathcal{T}_{V_{\max}}$ be the truncation operator,

$$\mathcal{T}_{V_{\max}} r = \begin{cases} \mathrm{sign}(r) V_{\max}, & \text{if } |r| > V_{\max}, \\ r, & \text{otherwise.} \end{cases}$$

Now, a simple calculation shows that

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

where $\mathcal{T}_{V_{\max}}\mathcal{F}_n = \{\mathcal{T}_{V_{\max}} f \,|\, f \in \mathcal{F}_n\}$. This, together with Theorem 2 gives rise to the following result:

Corollary 4 *Consider an MDP satisfying Assumptions A0 and A2 and assume that both its immediate reward function and transition kernel are Lipschitzian. Fix $p \ge 1$, $\mu \in M(\mathcal{X})$ and let $\{\mathcal{F}_n\}$, be a universal approximation class such that the pseudo-dimension of $\mathcal{T}_{V_{\max}}\mathcal{F}_n$ grows sublinearly in n. Then, for each $\epsilon, \delta > 0$ there exist an index n_0 such that for any $n \ge n_0$ 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 $V_{(\mathcal{T}_{V_{\max}}\mathcal{F}_n)^+}$ such that if V_K is the output of multi-sample FVI when it uses the function set $\mathcal{T}_{V_{\max}}\mathcal{F}_n$ and $X_i \sim \mu$ then $\|V^* - V^{\pi_K}\|_{p,\rho} \le \epsilon$ 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 defined by Equation (1) over $\mathcal{T}_{V_{\max}}\mathcal{F}_n$ is computationally challenging even when $\mathcal{F}_n$ is a class of linearly parameterized functions and $p = 2$. One idea is to do the optimization first over $\mathcal{F}_n$ and then truncate the obtained functions. The resulting procedure can be shown to be consistent (cf., Chapter 10 of Györfi et al. 2002 for an alike result in a regression setting).

It is important to emphasize that the construction used in this section is just one example of how our main result may lead to consistent algorithms. An immediate extension of the present work would be to achieve the best possible convergence rates for a given MDP by using penalized estimation. We leave the studying of such methods for future work.

8. Discussion of Related Work

Sampling based FVI has roots that date back to the early days of dynamic programming. One of the first examples of using value-function approximation methods 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 also 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.

FVI is a special case of *approximate value iteration* (AVI) which encompasses any algorithm of the form $V_{t+1} = TV_t + \epsilon_t$, where the errors ϵ_t are controlled in some way. If the error terms, ϵ_t, are bounded in supremum norm, then a straightforward analysis shows that asymptotically, the worst-case performance-loss for the policy greedy w.r.t. the most recent iterates can be bounded by $\frac{2\gamma}{(1-\gamma)^2}\sup_{t\ge 1}\|\epsilon_t\|_\infty$ (e.g. Bertsekas and Tsitsiklis, 1996). When V_{t+1} is the best approximation of TV_t in $\mathcal{F}$ then $\sup_{t\ge 1}\|\epsilon_t\|_\infty$ 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 error term, our loss-bounds have the same form (cf. Equation (13)). In particular, if μ is absolutely continuous w.r.t. the Lebesgue measure then letting $p \to \infty$ allows us to recover these previous bounds (since then $C_\mu^{1/p} d_{p,\mu}(T\mathcal{F},\mathcal{F}) \to d_\infty(T\mathcal{F},\mathcal{F})$). Further, we expect that the p-norm bounds would be tighter since the supremum norm is sensitive to outliers.

A different analysis, originally proposed by (Gordon, 1995; Tsitsiklis and Van Roy, 1996), goes by assuming that the iterates satisfy $V_{t+1} = \Pi TV_t$, where Π is an operator that maps bounded functions to the function space $\mathcal{F}$. While Gordon (1995) and Tsitsiklis and Van Roy (1996) considered the planning scenario with known dynamics and making use of

a set of representative states, subsequent results by Singh et al. (1995); Ormoneit and Sen (2002) and Szepesvári and Smart (2004) considered less restricted problem settings, though none of these authors presented finite-sample bounds. Further

The main idea in these analyses is that the above iterates must converge to some limit V_∞ if the composite operator ΠT is a supremum-norm contraction. Since T is a contraction, this holds wenever Π is a supremum-norm non-expansion. In this case, the loss of using the policy greedy w.r.t. V_∞ can be 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 a special class of approximation methods called *averagers* are used (Gordon, 1995). For these methods Π is guaranteed to be a non-expansion. Kernel regression methods, such as k-nearest neighbor smoothing with fixed centers, tree based smoothing (Ernst et al., 2005), or linear interpolation with a fixed set of basis functions such as spline-interpolation with fixed knots all belong to this class. In all these examples Π is a linear operator and takes the form $\Pi f = \alpha + \sum_{i=1}^n (L_i f)\phi_i$ with some function α, appropriate basis functions, ϕ_i, and linear functionals L_i $(i = 1, 2, \ldots, n)$. One particularly interesting case is when $L_i f = f(x_i)$ for some points $\{x_i\}$, $\phi_0 \geq 0$, $\sum_i \phi_i \equiv 1$, $\alpha \equiv 0$, $(\Pi f)(x_i) = f(x_i)$ and $(\phi_i(x_j))_{ij}$ has full rank. In this case all members of the space spanned by the basis functions $\{\phi_i\}$ are fixed points of Π. Hence $\epsilon_\Pi = d_\infty(\mathrm{span}(\phi_1, \ldots, \phi_n), V^*)$ and so the loss of the procedure is directly controlled by the size of $\mathcal{F}_n = \mathrm{span}(\phi_1, \ldots, \phi_n)$.

Let us now discuss the choice of the function spaces in averagers and sampling-based FVI. In the case of averagers, the class is restricted, but the approximation requirement, making ϵ_Π small, seems to be easier to satisfy than the corresponding requirement which asks for making the inherent Bellman residual of the function space $\mathcal{F}_n$ small. We think that in the lack of knowledge of V^* this advantage might be minor and well offset by the larger freedom to choose $\mathcal{F}_n$ (i.e., nonlinear, or kernel-based methods are allowed). In fact, when V^* is unknown one must resort to the generic properties of the class of MDPs considered (e.g., smoothness) in order to find the appropriate function space. Since also the optimal policy is unknown it does not seem as an advantage that only a single function (that depends on an unknown MDP) must be well approximated. Still, one may argue that the self-referential nature of the inherent Bellman-error makes the design for sampling-based FVI harder. As we have shown in Section 7, provided that the MDPs are smooth, designing these spaces is not necessarily harder than designing a function approximator for some regression task.

Let us now discuss some other related works where the authors consider the error resulting from some Monte-Carlo procedure. One set of results closely related to the ones presented here is due to Tsitsiklis and Roy (1999). These authors studied sampling-based fitted value iteration with linear function approximators. However, they considered a different class of MDPs: finite horizon, optimal stopping with discounted total rewards. 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. Their main result shows that the estimation error converges to

zero with probability one as the number of samples grows to infinity. Further, a bound on the asymptotic performance is given. Due to the special structure of the problem, this bound depends only on how well the optimal value function is approximated by the chosen function space. Certainly, because of the known counterexamples (Baird, 1995; Tsitsiklis and Van Roy, 1996), we cannot hope such a bound to hold in the general case.

The work presented here builds on our previous work. For finite state-space MDPs, Munos (2003, 2005) considered planning scenarios with known dynamics analysing the stability of both approximate policy iteration and value iteration with weighted L^2 (resp., L^p) fitting working with a finite set of representative states. 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 weighted p-norm error analysis of sampling-based FVI for continuous state-space MDPs and in a discounted, infinite-horizon settings.

One work where the author studies fitted value-iteration and which comes with a finite-sample analysis is by Murphy (2005), who, just like Tsitsiklis and Roy (1999), studied finite horizon problems with no discounting.[10][11] Because of the finite-horizon setting, the analysis is considerable simpler (the algorithm works backwards). The samples come from a number of independent trajectories just like in the case of Tsitsiklis and Roy (1999). The error bounds come in the form of performance differences between a pair of greedy policies: One of the policies from the pair is greedy w.r.t. the value function returned by the algorithm, while the other is greedy w.r.t. to some arbitrary 'test' function from the function set considered in the algorithm. The derived bound shows that the number of samples needed is exponential in the horizon of the problem and is proportional to ϵ^{-4}, where ϵ is the desired estimation error. The approximation error of the procedure, however, is not considered: Murphy suggests that the optimal action-value function could be added at virtually no cost to the function sets used by the algorithm. (Of course, when the optimal action-value function is unknown this option is not available to the user.) Murphy emphasizes the independence of her bounds of the dimensionality of the domain. This should not be surprising, though, since her bounds concern only the estimation error (the bounds on the estimation error in Theorem 2 are also dimension independent). We will come back to the discussion of worst-case sample-complexity after discussing the work by Kakade and Langford (Kakade and Langford, 2002; Kakade, 2003).

Kakade and Langford provide a finite-sample analysis of an algorithm introduced by them and called conservative policy iteration (CPI). This algorithm is designed for discounted infinite horizon problems. It 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. These advantages are esti-

10. We learnt of her results after submitting our paper.
11. The results in (Murphy, 2005) are actually presented for partially observable problems. It worth mentioning that since all value-function approximation methods introduce state aliasing anyway, results worked out for the fully observable case carry through to the limited feedback case without any changes except that the approximation power of the function approximation method is further limited by the information fed into the approximator. It is an interesting open problem if based on this observation it is possible to get consistent methods that avoid an explicit 'state estimation' component.

mated 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 between the goals of Kakade and Langford and that of the present paper: Whereas our goal is to study algorithms capable of solving MDPs, Kakade and Langford try to develop an algorithmic method (reduction) that transforms the problem of solving an MDP into another problem in an efficient way. In the case of CPI this problem is the problem of finding a discounted future state distribution of some good (near-optimal) policy: the relevant main results of (Kakade, 2003) (e.g., Theorems 7.3.1 and 7.3.3) bound the loss of using the policy returned by the procedure relative to using some other policy π (e.g., a near-optimal policy) as a function of the total variation distance between the distribution used to sample the basepoints (provided by the user) and the discounted future-state distribution underlying π. Unlike in this paper when the choice of the sampling distribution gets washed out if increasingly larger function sets are considered, the choice of the sampling distribution in the results of Kakade and Langford are not eliminated by considering more samples or a larger set of policies. The main result is stated as a high probability bound where the samples depend polynomially on the important parameters of the problem.

Now let us discuss the worst-case sample complexity of solving MDPs. A very simple observation is that it should be impossible to get bounds that scale polynomially with the dimension of the state-space unless special conditions are made on the problem. This is because the problem of estimating the value function of a policy in a trivial finite-horizon problem with a single time-step is equivalent to regression. Hence known lower bounds for regression must apply to RL, as well (see Stone (1980, 1982) and Chapter 3 of Györfi, Kohler, Krzyżak, and Walk (2002) for such bounds). In particular, from these bounds it follows that the minimax sample complexity of RL is exponential in the dimensionality of the state space provided the class of MDPs is large enough. Hence, it is not surprising that unless very special conditions are imposed on the class of MDPs considered, FVI and its variants are subject to the curse-of-dimensionality. One way to help with this exponential scaling is when the algorithm is capable of taking advantage of the possible advantageous properties of the MDP to be solved. In our opinion, one major open problem in RL is to design such methods (or to show that some existing method possess this property).

The curse-of-dimensionality is not specific to FVI variants but is inavoidable for large classes of continuous state-space MDPs. In fact, one result of Chow and Tsitsiklis (1989) states the following: Consider a class of MDPs with $\mathcal{X} = [0,1]^d$. Assume that for any MDP in the class, the transition probability kernel underlying the MDP has a density with respect to the Lebesgue measure and these densities have a common upper bound. Further, the MDPs within the class are assumed to be uniformly smooth: for any MDP in the class the Lipschitz constant of the reward function of the MDP is bounded by an appropriate constant and the same holds for the Lipchitz constant of the density function. Fix a desired precision, ϵ. 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 very hard instances.

The situation changes dramatically if one is allowed to interleave computations and control. In this case, building on the *random-discretization method* of Rust (1996b), it is possible to achieve near-optimal behaviour by using a number of samples per step that scales polynomially in the important quantities (Szepesvári, 2001). In particular, this result shows that it suffices to let the number of samples scale linearly in the dimensionality of the state space. Interestingly, this result holds for a class of MDPs that subsumes the one considered by Chow and Tsitsiklis (1989). The random-discretization method requires that the MDPs in the class satisfy Assumption A1 with a common constant and also the knowledge of the density underlying the transition probability kernel. The results presented here, on the other hand, require less stringent assumptions on the transition probability kernel (Assumption A2) and thus apply to a broader class of problems. Further, the advantage of the random-discretization method melts away when the density used by it is estimated from samples because density estimation itself is also subject to the curse of dimensionality.

Another method that interleaves computations and control is the sparse trajectory-tree method of Kearns et al. (1999). The sparse trajectory-tree method builds a random lookahead tree to compute sample based approximation to the values of each of the actions available at the current state. Although the size of this tree does not depend on the dimensionality of the state space at all, it scales exponentially in the 'ϵ-horizon', $\log_\gamma(R_{\max}/(\epsilon(1-\gamma)))$. When the the discount factor is close to one and the number of actions is moderate, the method cannot be expected to deliver reliable estimates. 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)).

9. Simulation Study

The purpose of this section is to illustrate the tradeoffs involved in using FVI. In particular, since identical or very similar algorithms have been used successfully in many prior empirical studies (e.g., Longstaff and Shwartz 2001; Haugh 2003; Jung and Uthmann 2004), we do not attempt a thorough empirical evaluation of the algorithm.

9.1 An Optimal Replacement Problem

The problem used as a testbed is a simple one-dimensional optimal replacement problem, described for example by Rust (1996a). The system has a one-dimensional state. The 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 time step, t, there are two possible decisions: either keep ($a_t = \mathbf{K}$) or replace ($a_t = \mathbf{R}$) the product. This latter action implies an additional cost C of selling the existing product and replacing it by a new one. The transition to a new state occurs with the following exponential densities:

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

The reward function is $r(x, \mathbf{K}) = -c(x)$, where $c(x)$ represents the cost of maintaining the product. By assumptions, c is monotonically increasing. The reward associated with the replacement of the product is independent of the state and is given by $r(x, \mathbf{R}) = -C - c(0)$.

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, while the second argument gives the total future reward provided that the product is replaced. This equation has a closed form solution:

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

Here $\overline{x}$ 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 $x > \overline{x}$.

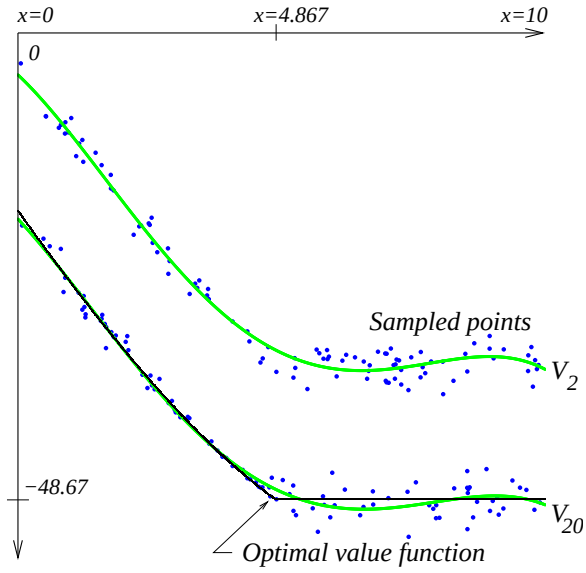

Figure 2: Illustration of two iteration steps of Sampling based FVI (up: $k = 2$, down: $k = 20$). The dots represent the pairs $(X_i, \hat{V}_{M,k}(X_i))$, $i = 1, \ldots, N$ based on $M = 10$ sampled transitions per basepoint and $N = 100$ basepoints. The grey curve is the best fit among polynomials of degree $l = 4$. The thin black curve is the optimal value function.

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

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

We consider approximation of the value function using polynomials of degree l. As suggested in Section 7, we used truncation. In order to make the state space bounded, we actually consider a problem that closely approximates the original one. **[CS: Remi: Can you confirm that the optimal value function for the new problem is unchanged? I believe this is the case. Then we should write this.]** 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 the multi-sample version of sampling-based FVI: the dots represents the points $\{(X_n, \hat{V}_{M,k+1}(X_n))\}_{1 \le n \le N}$ for $N = 100$, where X_i is drawn from μ and $\{\hat{V}_{M,k+1}(X_n)\}_{1 \le n \le N}$ 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, i.e., $p = 2$) in $\mathcal{F}$ (for $l = 4$) and the thin black curve is the optimal value function.

Figure 3 shows the L_∞ approximation errors $||V^* - V_K||_\infty$ for different values of the degree l of the polynomial regression, and different values of the number of basepoints N and the number of sampled next states M. The number of iterations was set to $K = 20$. The reason that the figure shows the error of approximating V^* by V_K (i.e., $\epsilon_K = \|V^* - V_K\|_\infty$) instead of $\|V^* - V^{\pi_K}\|_\infty$ is that in this problem this latter error converges very fast and thus is less interesting. (The technique developed in the paper can be readily used to derive a bound on the estimation error of V^*.) Of course, the performance loss is always upper-bounded (in L_∞-norm) by the approximation error, thanks to the well-known bound (e.g., Bertsekas and Tsitsiklis 1996): $||V^* - V^{\pi_K}||_\infty \le 2/(1-\gamma)||V^* - V_K||_\infty$.

From Figure 3 we observe that for $N = 100$, $M = 10$, when the degree l of the polynomials increases, the error decreases first because of the decrease of the inherent approximation error, but eventually increases because of overfitting. This graph thus illustrates the different components of the bound (13) where the approximation error term $d_{p,\mu}(T\mathcal{F}, \mathcal{F})$ decreases with l (as discussed in Section 7) whereas the estimation error bound, being a function of the pseudo-dimension of $\mathcal{F}$, increases with l (in this case, $V_{\mathcal{F}^+} = l + 2$) 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. Note that truncation never happens except in the when the degree of the polynomial is very large compared with the number of samples.

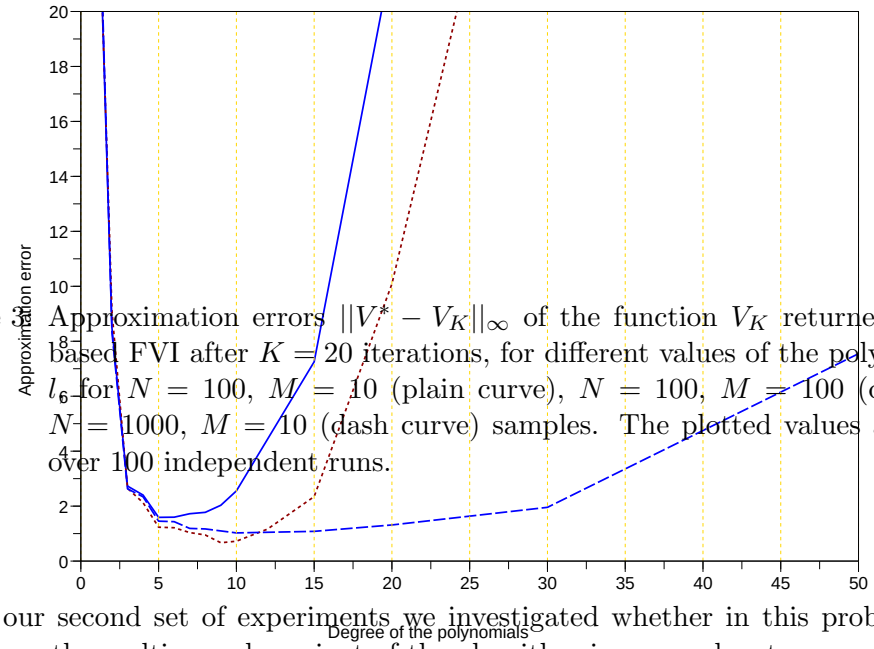

Figure 3: Approximation errors $||V^* - V_K||_\infty$ of the function V_K returned by sampling-based FVI after $K = 20$ iterations, for different values of the polynomials degree l for $N = 100$, $M = 10$ (plain curve), $N = 100$, $M = 100$ (dot curve), and $N = 1000$, $M = 10$ (dash curve) samples.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

In our second set of experiments we investigated whether in this problem the single-sample or the multi-sample variant of the algorithm is more advantageous.

Figure 4 shows the distributional character of $V_K - V^*$ as a function of the state. The order of the fitted polynomials is 5. The black (middle) curve 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, while 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 it may still

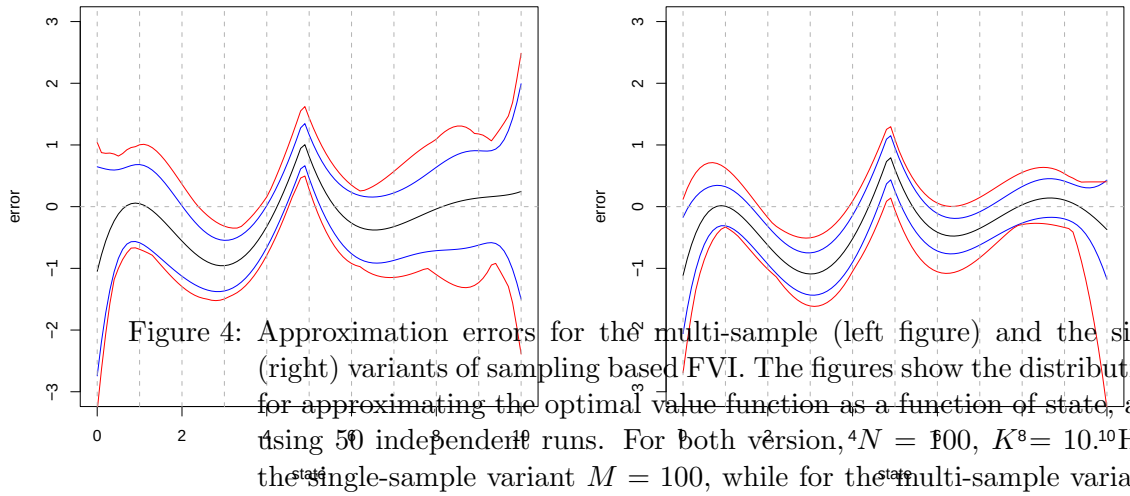

Figure 4: Approximation errors for the multi-sample (left figure) and the single-sample (right) variants of sampling based FVI. The figures show the distribution of errors for approximating the optimal value function as a function of state, as measured using 50 independent runs. For both version, $N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

cause a considerable slow-down of the procedure when the behaviour of the value-function 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 due to the reuse of samples, while the variance of the estimates decreases significantly.

10. Conclusions

We considered sampling-based FVI for discounted, large (possibly infinite) state space, finite-action Markovian Decision Problems when only a generative model of the environment is available. In each iteration, the image of the previous iterate under Bellman operator is approximated at a finite number of steps using a simple Monte-Carlo technique. A regression method is used then to fit a function to the data obtained. The main contribution of the paper are performance bounds for this procedure that holds with high probability. The bounds scale with the inherent Bellman error of the function space used in the regression step, and the stochastic stability properties of the MDP. It is an open question if the finiteness of the inherent Bellman error is necessary for the stability of FVI, but the counterexamples discussed in the introduction suggest that the inherent Bellman residual of the function space should indeed play a crucial role in the final performance of FVI. Even

less is known about whether the stochastic stability conditions are necessary or if they can be relaxed.

We argued that 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. Since these conditions could be difficult to verify a priori for any practical problem, adaptive methods are needed. We believe that the techniques developed in this paper may serve as a solid foundations for developing and studying such algorithms.

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 that without any a priori knowledge of the smoothness class, the method allows one to achieve the optimal rate of convergence (see Györfi et al., 2002, Section 21.2).

Another problem left for future work is to improve the scaling of our bounds. An important open question is to establish tight lower bounds for the rate of convergence for value-function based RL methods.

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 moving it closer to the future-state distribution of the most recent policy. The improvement presumably manifests itself by decreasing the term including $C_{\rho,\mu}$. Another possibility is to adaptively choose M, the number of next state sampled based on the available local information like in active learning, hoping that this way the sample-efficiency of the algorithm could be further improved.

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 use the following inequality due to Pollard:

Theorem 5 (Pollard, 1984) *Let X_i, $i = 1, \dots, N$, 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 (i.e., $0 \le |f(x)| \le K$ for any $(f, x) \in \mathcal{F}, \mathcal{X}$). Then*

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

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.

Proof Let Ω denote the sample-space underlying the random variables. Let $\epsilon'' > 0$ be arbitrary and let f^* be such that $\|f^* - TV\|_{p,\mu} \le \inf_{f\in\mathcal{F}} \|f - TV\|_{p,\mu} + \epsilon''$. Define $\|\cdot\|_{p,\hat{\mu}}$ by

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

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

$$\begin{aligned}
\left\|V' - TV\right\|_{p,\mu} &\le \left\|V' - TV\right\|_{p,\hat{\mu}} + \epsilon' && (14)\\
&\le \left\|V' - \hat{V}\right\|_{p,\hat{\mu}} + 2\epsilon' && (15)\\
&\le \left\|f^* - \hat{V}\right\|_{p,\hat{\mu}} + 2\epsilon' && (16)\\
&\le \|f^* - TV\|_{p,\hat{\mu}} + 3\epsilon' && (17)\\
&\le \|f^* - TV\|_{p,\mu} + 4\epsilon' && (18)\\
&= d_{p,\mu}(TV,\mathcal{F}) + 4\epsilon' + \epsilon''. && (19)
\end{aligned}$$

It follows then that $\|V' - TV\|_{p,\mu} \le \inf_{f\in\mathcal{F}} \|f - TV\|_{p,\mu} + 4\epsilon' + \epsilon''$ with probability at least $1-\delta$. Since $\epsilon'' > 0$ was arbitrary, it also follows that $\|V' - TV\|_{p,\mu} \le \inf_{f\in\mathcal{F}} \|f - TV\|_{p,\mu} + 4\epsilon'$ with probability at least $1-\delta$. Now, the Lemma follows by choosing $\epsilon' = \epsilon/4$.

Let us now turn to the proof of (14)–(18). First, observe that (16) holds due to the choice of V' since $\left\|V' - \hat{V}\right\|_{p,\hat{\mu}} \le \left\|f - \hat{V}\right\|_{p,\hat{\mu}}$ holds for all functions f from $\mathcal{F}$ and thus the same inequality holds for $f^* \in \mathcal{F}$, too.

Thus, (14)–(18) will be established if we prove that (14),(15),(17) and (18) all hold with probability at least $1-\delta'$ with $\delta' = \delta/4$. Let

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

We claim that

$$\mathbb{P}\left(Q > \epsilon'\right) \le \delta', \tag{20}$$

where $\delta' = \delta/4$. From this, (14) and (18) will follow.

In order to prove (20) note that for all $\omega \in \Omega$, $V' = V'(\omega) \in \mathcal{F}$. Hence,

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

holds pointwise in Ω. Therefore the inequality

$$\sup_{f\in\mathcal{F}} \left|\|f - TV\|_{p,\mu} - \|f - TV\|_{p,\hat{\mu}}\right| > Q \tag{21}$$

holds pointwise in Ω, too and hence

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

We claim that

$$\mathbb{P}\left(\sup_{f\in\mathcal{F}}\left|\|f - TV\|_{p,\mu} - \|f - TV\|_{p,\hat{\mu}}\right| > \epsilon'\right) \leq \mathbb{P}\left(\sup_{f\in\mathcal{F}}\left|\|f - TV\|_{p,\mu}^p - \|f - TV\|_{p,\hat{\mu}}^p\right| > (\epsilon')^p\right). \tag{22}$$

Consider any event ω such that

$$\sup_{f\in\mathcal{F}}\left|\|f - TV\|_{p,\mu} - \|f - TV\|_{p,\hat{\mu}}\right| > \epsilon'. \tag{23}$$

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 (22) 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 $\{f - TV | f \in \mathcal{F}\}$ is the same as the covering number of $\mathcal{F}$, calling for Theorem 5 results in

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

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 (20).

Now, let us prove inequalities (15) and (17). 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

$$\left| \|f - TV\|_{p,\hat{\mu}} - \left\| f - \hat{V} \right\|_{p,\hat{\mu}} \right| \le \left\| TV - \hat{V} \right\|_{p,\hat{\mu}} . \tag{24}$$

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,

$$\mathbb{P}\left(\left| \mathbb{E}\left[R_1^{X_i,a} + \gamma V(Y_1^{X_i,a}) \,|\, X^{1:N} \right] - \frac{1}{M} \sum_{j=1}^{M} R_j^{X_i,a} + \gamma V(Y_j^{X_i,a}) \right| > \epsilon' \,|\, X^{1:N} \right) \le 2\, e^{-\frac{2M(\epsilon')^2}{K_1^2}}, \tag{25}$$

where $X^{1:N} = (X_1, \ldots, X_N)$. 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

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

Taking the expectation of both sides of this inequality gives

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

Hence also

$$\mathbb{P}\left(\frac{1}{N} \sum_{i=1}^{N} \left| (TV)(X_i) - \hat{V}(X_i) \right|^p > (\epsilon')^p \right) \le \delta' \tag{26}$$

and therefore by (24),

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

Using this with $f = V'$ and $f = f^*$ shows that inequalities (15) and (17) 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, which stated a finite-sample bound for the single-sample variant of the algorithm.

A.1 Proof of Lemma 2

Proof The proof is analogous to that of Lemma 1, hence we only give the differences. Up to (21) the two proofs proceed in an identical way, however, from (21) we continue by concluding that

$$\sup_{g\in\mathcal{F}}\sup_{f\in\mathcal{F}}\left|\|f-Tg\|_{p,\mu}-\|f-Tg\|_{p,\hat{\mu}}\right|>Q \tag{27}$$

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 (20): 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, (25) is replaced with

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

where $Z_{i,a}^j=(R_j^{X_i,a},Y_j^{X_i,a})$. Here $\mathcal{F}_+=\{h:\mathbb{R}\times\mathcal{X}\to\mathbb{R}\,|\,h(s,x)=s\mathbb{I}_{\{|s|\leq V_{\max}\}}+f(x)$ 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

The theorem states states PAC-bounds on the sample size of sampling-based FVI. The idea of the proof is to show that *(i)* if the errors in each iteration are small then the final error will be small when K, the number of iterations is high enough and *(ii)* the previous results (Lemma 2,2) show that the errors stay small with high probability in each iteration provided that M,N is high enough. Putting these results together gives the main result. Hence, we need to show *(i)*.

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

$$V_{k+1}=TV_k-\varepsilon_k \tag{28}$$

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. The proof is done in two steps: we first prove a statement that gives pointwise bounds (i.e., the bounds hold for any state $x\in\mathcal{X}$) which is then used to prove the necessary L^p bounds. Parts *(i)* and *(ii)* are connected in Sections B.3, B.4.

B.1 Pointwise Error Bounds

Lemma 3 *We have*

$$\begin{aligned} V^* - V^{\pi_K} \leq \quad & (I-\gamma P^{\pi_K})^{-1}\Big\{ \textstyle\sum_{k=0}^{K-1} \gamma^{K-k}\big[(P^{\pi^*})^{K-k} + P^{\pi_K}P^{\pi_{K-1}}\dots P^{\pi_{k+1}}\big]|\varepsilon_k| \\ & +\gamma^{K+1}\big[(P^{\pi^*})^{K+1} + (P^{\pi_K}P^{\pi_K}P^{\pi_{K-1}}\dots P^{\pi_1})\big]|V^*-V_0|\Big\}. \end{aligned} \tag{29}$$

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

$$V^* - V_K \leq \sum_{k=0}^{K-1} \gamma^{K-k-1}(P^{\pi^*})^{K-k-1}\varepsilon_k + \gamma^K (P^{\pi^*})^K (V^*-V_0). \tag{30}$$

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,

$$V^* - V_K \geq \sum_{k=0}^{K-1} \gamma^{K-k-1}(P^{\pi_{K-1}}P^{\pi_{K-2}}\dots P^{\pi_{k+1}})\varepsilon_k + \gamma^K (P^{\pi_K}P^{\pi_{K-1}}\dots P^{\pi_1})(V^*-V_0). \tag{31}$$

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 a monotone operator[12] (we may write $(I-\gamma P^{\pi_K})^{-1} = \sum_{m\geq 0}\gamma^m (P^{\pi_K})^m$), we deduce

$$V^* - V^{\pi_K} \leq \gamma (I-\gamma P^{\pi_K})^{-1}(P^{\pi^*} - P^{\pi_K})(V^*-V_K) \tag{32}$$

Now, using (30) and (31),

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

from which (29) follows by taking the absolute value of both sides. ■

12. An operator T is monotone if for any $x \leq y$, $Tx \leq Ty$.

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 ($\|\varepsilon_k\|_{p,\mu} \le \epsilon$ for all $0 \le k < K$) then*

- *Given Assumption A1 we have*

$$\|V^* - V^{\pi_K}\|_\infty \le \frac{2\gamma}{(1-\gamma)^2} C_\mu^{1/p} \epsilon + \eta. \tag{33}$$

- *Given Assumption A2 we have*

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

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

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

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

Proof We have seen that if A1 holds then A2 also holds, and for any distribution ρ, $C_{\rho,\mu} \le C_\mu$. Thus, if the bound (34) holds for any ρ then choosing ρ to be a Dirac at each state proves (33). Thus we only need to prove (34).

We may rewrite (29) as

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

with the positive coefficients

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

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

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

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 $\log V_{\max}$ 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 From pointwise expectations to conditional expectations

We will need the following lemma in the proof of the theorem:

Lemma 5 *Assume that X, Y are independent random variables taking values in the respective measurable spaces, $\mathcal{X}$ and $\mathcal{Y}$. Let $f : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ be a Borel-measurable function such that $\mathbb{E}[f(X,Y)]$ exists. Assume that for all $y \in \mathcal{Y}$, $\mathbb{E}[f(X,y)] \geq 0$. Then $\mathbb{E}[f(X,Y)|Y] \geq 0$ holds, too, w.p.1.*

This lemma is an immediate consequence of the following result, whose proof is given for the sake of completeness:

Lemma 6 *Assume that X, Y are independent random variables taking values in the respective measurable spaces, $\mathcal{X}$ and $\mathcal{Y}$. Let $f : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ be a Borel-measurable function and assume that $\mathbb{E}[f(X,Y)]$ exists. Let $g(y) = \mathbb{E}[f(X,y)]$. Then $\mathbb{E}[f(X,Y)|Y] = g(Y)$ holds w.p.1.*

Proof Let us first consider the case when f has the form $f(x,y) = \mathbb{I}_{\{x \in A\}}\mathbb{I}_{\{y \in B\}}$, where $A \subset \mathcal{X}$, $B \subset \mathcal{Y}$ are measurable sets. Write $r(x) = \mathbb{I}_{\{x \in A\}}$ and $s(y) = \mathbb{I}_{\{y \in B\}}$. Then $\mathbb{E}[f(X,Y)|Y] = \mathbb{E}[r(X)s(Y)|Y] = r(Y)\mathbb{E}[s(X)|Y]$ since $s(Y)$ is Y-measurable. Since X and Y are independent, so are $s(X)$ and Y and thus $\mathbb{E}[s(X)|Y] = \mathbb{E}[s(X)]$. On the other hand, $g(y) = \mathbb{E}[r(X)s(y)] = s(y)\mathbb{E}[r(X)]$, and thus it indeed holds that $\mathbb{E}[f(X,Y)|Y] = g(Y)$ w.p.1. Now, by the additivity of expectations the same relation holds for sums of functions of the above form and hence, ultimately, for all simple functions. If f is nonnegative valued then we can find a sequence of increasing simple functions f_n with limit f. By Lebesgue's monotone convergence theorem, $g_n(y) \stackrel{\text{def}}{=} \mathbb{E}[f_n(X,y)] \to \mathbb{E}[f(X,y)] \, (= g(y))$. Further, since Lebesgue's monotone convergence theorem also holds for conditional expectations, we also have $\mathbb{E}[f_n(X,Y)|Y] \to \mathbb{E}[f(X,Y)|Y]$. Since $g_n(Y) = \mathbb{E}[f_n(X,Y)|Y] \to \mathbb{E}[f(X,Y)|Y]$ w.p.1., and $g_n(Y) \to g(Y)$ w.p.1., we get that $g(Y) = \mathbb{E}[f(X,Y)|Y]$ w.p.1. Extension to an arbitrary function follows by decomposing the function into its positive and negative parts. ■

B.4 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 large enough, the bound

$$\|V^* - V^{\pi_K}\|_{p,\rho} \leq \frac{2\gamma}{(1-\gamma)^2} C_{\rho,\mu}^{1/p} \, d_{p,\mu}(T\mathcal{F}, \mathcal{F}) + \epsilon \tag{35}$$

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

$$\|V^* - V^{\pi_K}\|_{p,\rho} \leq \frac{2\gamma}{(1-\gamma)^2} C_{\rho,\mu}^{1/p} \epsilon_0 + \eta, \tag{36}$$

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 via the application of Lemma 5.

To show this let us denote the collection of random variables used in the kth step by S_k. Hence, S_k consists of the N basepoints, as well as $|\mathcal{A}| \times N \times M$ next states and rewards. Further, introduce the notation $V'(V, S_k)$ to denote the result of solving the optimization problem (3)–(4) based on the sample S_k and starting from the value function $V \in B(\mathcal{X})$. By Lemma 1, $\mathbb{P}\left(\|V'(V,S_k) - TV\|_{p,\mu} \leq d_{p,\mu}(TV,\mathcal{F}) + \epsilon'\right) \geq 1 - \delta/K$. Now let us apply Lemma 5 with $X := S_k$, $Y := V_k$ and $f(S,V) = \mathbb{I}_{\{\|V'(V,S)-TV\|_{p,\mu} \leq d_{p,\mu}(TV,\mathcal{F})+\epsilon'\}} - (1-\delta/K)$. Since S_k is independent of V_k the lemma can indeed be applied. Hence, $\mathbb{P}\left(\|V'(V_k,S_k) - TV_k\|_{p,\mu} \leq d_{p,\mu}(TV_k,\mathcal{F}) + \epsilon' | V_k\right) \geq 1 - \delta/K$. Taking the expectation of both sides gives $\mathbb{P}\left(\|V'(V_k,S_k) - TV_k\|_{p,\mu} \leq d_{p,\mu}(TV_k,\mathcal{F}) + \epsilon'\right) \geq 1-\delta/K$. Since $V'(V_k,S_k) = V_{k+1}$, $\epsilon_k = TV_k - V_{k+1}$, we thus have that

$$\|\epsilon_k\|_{p,\mu} \leq d_{p,\mu}(TV,\mathcal{F}) + \epsilon' \tag{37}$$

holds except for a set of bad events B_k of measure at most δ/K.

Hence, inequality (37) holds simultaneously for $k = 1,\ldots,K$, except for the events in $B = \cup_k B_k$. Note that $\mathbb{P}(B) \leq \sum_{k=1}^K \mathbb{P}(B_k) \leq \delta$. Now pick any event in the complementer of B. Thus, for such an event (36) holds with $\epsilon_0 = d_{p,\mu}(TV,\mathcal{F}) + \epsilon'$. Plugging in the definitions of ϵ' and η we obtain (35).

Now assume that the MDP satisfies Assumption A1. As before, we conclude that (37) holds except for the events in B_k and with the same choice of N and M, we still have $\mathbb{P}(B) = \mathbb{P}(\cup_k B_K) \leq \delta$. Now, using (33) we conclude that except on the set B, $\|V^* - V^{\pi_K}\|_\infty \leq \frac{2\gamma}{(1-\gamma)^2}\, C_{\rho,\mu}^{1/p}\; d_{p,\mu}(T\mathcal{F},\mathcal{F}) + \epsilon$, 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 $x \in \mathcal{X}$,

$$|V^{\pi_{\alpha,\lambda}^K}(x) - V^{\pi_\alpha^K}(x)| \leq \frac{2V_{\max}\lambda}{1-\gamma}. \tag{38}$$

Now, use the triangle inequality to get

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

By (38), 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 7 *The following bound*

$$\left\|V^* - V^{\pi^K_\alpha}\right\|_{p,\rho} \le 2^{1-1/p}\left[\frac{2\gamma}{(1-\gamma)^2}C^{1/p}_{\rho,\mu}\max_{0\le k<K}\|\varepsilon_k\|_{p,\mu} + \eta + \frac{\alpha}{1-\gamma}\right] \tag{39}$$

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

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

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

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

$$\left\|V^* - V^{\pi^K_{\alpha,\lambda}}\right\|_{p,\rho} \le \frac{4\gamma}{(1-\gamma)^2}C^{1/p}_{\rho,\mu}d_{p,\mu}(T\mathcal{F},\mathcal{F}) + \epsilon \tag{40}$$

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 7:

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 $TV_K \ge T^{\pi^K_\alpha}V_K \ge TV_K - \alpha\mathbf{1}$. Thus, similarly to the proof of Lemma 3, we have

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

and by using (30) and (31), we deduce

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

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

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

and Lemma 7 follows. ■

■

References

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 G. Lugosi and H.U. Simon, editors, *The Nineteenth Annual Conference on Learning Theory, COLT 2006, Proceedings*, volume 4005 of *LNCS/LNAI*, pages 574–588, Berlin, Heidelberg, June 2006. Springer-Verlag. (Pittsburgh, PA, USA, June 22–25, 2006.).

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. 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. Kakade and J. Langford. Approximately optimal approximate reinforcement learning. In *Proceedings of the Nineteenth International Conference on Machine Learning*, pages 267–274, San Francisco, CA, USA, 2002. Morgan Kaufmann Publishers Inc.

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. 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.

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. In *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.

B. Schölkopf and A.J. Smola. *Learning with Kernels*. MIT Press, Cambridge, MA, 2002.

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*, pages 881–886, 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.

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.