

Efficient Approximate Planning in Continuous Space Markovian Decision Problems

Csaba Szepesvári^{a,*},

^a *Mindmaker Ltd.*

Budapest HUNGARY - 1121

Konkoly-Thege M. u. 29-33.

E-mail: szepes@mindmaker.hu

Monte-Carlo planning algorithms for planning in continuous state-space, discounted Markovian Decision Problems (MDPs) having a smooth transition law and a finite action space are considered. We prove various polynomial complexity results for the considered algorithms, improving upon several known bounds.

Keywords: Markovian Decision Problems, planning, value iteration, Monte-Carlo algorithms

1. Introduction

MDPs provide a clean and simple, yet fairly rich framework for studying various aspects of intelligence, such as planning. A well-known practical limitation of planning in MDPs is called the *curse of dimensionality* [1], referring to the exponential rise in the resources required to compute (even approximate) solutions to an MDP as the size of the MDP (the number of state variables) increases. For example, conventional dynamic programming (DP) algorithms, such as value- or policy-iteration scale exponentially with the size, even if they are combined with sophisticated multigrid algorithms [4]. Moreover, the curse of dimensionality is not specific to any algorithm, as shown by a result of Chow and Tsitsiklis [3].

Recently, Kearns et al. have shown that a certain on-line, tree building algorithm avoids the curse of dimensionality in discounted MDPs [9]. Recently, this result has been extended to partially observable MDPs (POMDPs) by the same authors [8]. The bounds in these two papers are independent of the size of the state space, but scale exponentially with $\frac{1}{1-\gamma}$, the *effective horizon-time*, where γ is the discount factor of the MDP.

In this paper we consider another on-line planning algorithm that will be shown to scale polynomially with the horizon-time, as well. The price of this is that we have to assume more regularity on

the MDPs we consider. In particular, we will restrict ourselves to stochastic MDPs with finite action spaces and state space $\mathcal{X} = [0, 1]^d$, and, more importantly, assume that the transition probability kernel of the MDPs are subject to the Lipschitz-condition $|p(x'|x_1, a) - p(x'|x_2, a)| \leq L_p \|x_1 - x_2\|_1$ for any states $x_1, x_2, x' \in [0, 1]^d$ and action $a \in \mathcal{A}$. Here $L_p > 0$ is a given fixed number and $\|\cdot\|_1$ denotes the ℓ^1 norm of vectors. Another restriction (quite common in the literature) that we will assume is the uniform boundedness of the transition probabilities (the bound shall be denoted by K_p) and of the immediate rewards (bound denoted by K_r). Further, our bounds will depend on the dimension of the state space, d .¹

The idea of the considered algorithms originates in the algorithm considered by Rust [13].² Rust studied a more restricted class of problems than considered in this paper and proved the following result. First, let us define the concept of ε -optimality in the mean. Fix an MDP with state space \mathcal{X} . A random, real-valued function \hat{V} with domain \mathcal{X} is called ε -optimal in the mean if $\mathbb{E} \left[\left\| \hat{V} - V^* \right\|_\infty \right] \leq \varepsilon$, where V^* is the optimal value function underlying the selected MDP and $\|\cdot\|_\infty$ is the maximum-norm and the expectation is

¹The bounds developed by Kearns et. al do not exhibit any dependence on the state space.

²The algorithm will be given in the next section.

taken for the random function \hat{V} . The input of the algorithm is a tolerance number, $\varepsilon > 0$. Given any $\varepsilon > 0$, the algorithm first builds up a (random) cache C_ε . Then, given a query state $x \in \mathcal{X}$ and the cache C_ε , the algorithm draws a sample of a random function $\hat{V}(x)$, \hat{V} being ε -optimal in the mean. Rust has shown that both phases of the algorithm are polynomial in $|\mathcal{A}|$, $K_r/(\varepsilon(1-\gamma))$, L_p , L_r , d , K_p . Here L_r is the Lipschitz factor of the immediate rewards. Note that Rust’s bound scales polynomially with the effective horizon-time, so our approach will be to extend his algorithm to planning.

The very first idea along this way is to make use of Markov’s inequality. The algorithm based on this idea would work as follows: Fix the random sample N and consider \hat{V} as given by Rust’s algorithm, and a state x . Using Markov’s inequality one gets that $P(\|\hat{V} - V^*\|_\infty \geq \delta) \leq \varepsilon/\delta$. Now, imagine that we can compute

$$\operatorname{argmax}_{a \in \mathcal{A}} \{r(x, a) + \gamma \int p(y|x, a) \hat{V}(y) \} dy.$$

A contraction argument would then show that drawing

$$N = \operatorname{poly}(K_r/(\varepsilon\delta), L_p, L_r, |\mathcal{A}|, d, K_r, K_p, 1/(1-\gamma))$$

samples is sufficient for ensuring the ε -optimality of π with probability at least $1-\delta$. Now, the $|\mathcal{A}|$ integrals can themselves be approximated by Monte-Carlo methods.³ The computational complexity of the resulting algorithm will depend polynomially on N and will thus scale polynomially with L_r and $1/\delta$.

There are a number of methods to boost the polynomial dependence on $1/\delta$ to $\log(1/\delta)$. Here, we are going to use maximal inequalities to arrive at such a result. This method will have the additional benefit that we can get rid of the Lipschitzian condition regarding the immediate rewards and boost the polynomial dependence of the complexity bounds on L_p to a poly-logarithmic one. Interestingly, our bound for the number of samples will be poly-logarithmic in the size of the action space, as well. Note, however, the the com-

³One might either want to reuse the samples drawn earlier or draw new samples. The second approach is easier to analyze, whilst the first one may appear more elegant for some.

plexity bounds will still scale polynomially with the size of the action space. We will also derive novel bounds for the complexity of calculating uniformly optimal policies.

The organization of the paper is as follows: In Section 2 we provide the necessary background. The algorithm is given in Section 3, the main result of the paper is formulated in Section 4. The proof of the main result is given in Section 5, and conclusions are drawn in Section 6.

2. Preliminaries

We assume that the reader is familiar with the basics of the theory of MDPs. Readers who lack the necessary background are referred to the book of Dynkin and Yuskovich [6] or the more recent books [2] and [11].

2.1. Notation

Let $p \in [1, +\infty]$. $\|\cdot\|_p$ refers to the ℓ^p norm of vectors and the L^p norm of functions, depending on the type of its argument. Lip_p denotes the set of mappings that are Lipschitz-continuous in the norm $\|\cdot\|_p$: $f \in \operatorname{Lip}_p$ means that there exists a positive constant $L > 0$ s.t. $\|f(x) - f(y)\|_p \leq L \|x - y\|_p$ (domains of the mappings are suppressed). L is called the $\|\cdot\|_p$ -Lipschitz constant of f . $\operatorname{Lip}_p(\gamma) \subset \operatorname{Lip}_p$ denotes the set of mappings whose $\|\cdot\|_p$ -Lipschitz constant is not larger than γ . A mapping T is called a contraction in the norm $\|\cdot\|_p$ if $T \in \operatorname{Lip}_p(\gamma)$ for some $0 \leq \gamma < 1$. Let \mathcal{V} be any set, $T : \mathcal{V} \rightarrow \mathcal{V}$ and $S : \mathcal{V} \rightarrow \mathcal{V}$. Then the mapping $TS : \mathcal{V} \rightarrow \mathcal{V}$ is defined by $(TS)v = T(Sv)$, $v \in \mathcal{V}$. The set of natural numbers will be denoted by \mathbb{N} , the set of reals by \mathbb{R} . If $t \in \mathbb{N}$ then T^t denotes the map that is the product of T with itself t -times. We say that $T = S$ iff $Tv = Sv$ holds for all $v \in \mathcal{V}$. ω will in general denote an elementary event of the probability space under consideration, lhs means “left-hand-side”, and rhs means “right-hand-side”. We define $B(\mathcal{X})$ to be the set of all bounded real-valued function over \mathcal{X} : $B(\mathcal{X}) = \{f : \mathcal{X} \rightarrow \mathbb{R} : \|f\|_\infty < +\infty, f \text{ is measurable}\}$. Further, for any $K > 0$, $B_K(\mathcal{X})$ shall denote the set of all bounded functions whose maximum-norm is below the constant K : $B_K(\mathcal{X}) = \{f \in B(\mathcal{X}) : \|f\|_\infty < K\}$.

Table 1
Pseudo-code of the algorithm

0.	Input: $x \in \mathcal{X}$ (query state), $\varepsilon > 0$ (tolerance), $p, r, \gamma, \mathcal{A}$ (model parameters).
1.	Compute t and N as defined in Theorem 5.18.
2.	Draw X_1, \dots, X_N independent samples uniformly distributed over \mathcal{X} .
3.	Compute $\hat{p}_{X^{1:N}}(X_i X_j, a)$ ($1 \leq i, j \leq N$) using $\hat{p}_{x^{1:N}}(x_i x, a) = p(x_i x, a) / \sum_{j=1}^N p(x_j x, a)$ if $\sum_{j=1}^N p(x_j x, a) > 0$, and let $\hat{p}_{x^{1:N}}(x_i x, a) = 0$ otherwise.
4.	Let $v_i = 0$, $1 \leq i \leq N$.
5.	Repeat t times: $v_i :=$ $\max_{a \in \mathcal{A}} \{r(X_i, a) + \gamma \sum_{j=1}^N \hat{p}_{X^{1:N}}(X_j X_i, a) v_j\}$, $1 \leq i \leq N$.
6.	Let $a^* =$ $\operatorname{argmax}_{a \in \mathcal{A}} \{r(x, a) + \gamma \sum_{j=1}^N \hat{p}_{X^{1:N}}(X_j x, a) v_j\}$.
7.	Return a^* .

2.2. The Model

Let us consider the continuous space discounted MDP given by $(\mathcal{X}, \mathcal{A}, p, r, \gamma)$, where $\mathcal{X} = [0, 1]^d$ ($d > 0$, $d \in \mathbb{N}$) is the state space, \mathcal{A} is the action space, p is a measurable transition density: $p(x'|x, a) \geq 0$ and $\int p(x'|x, a) dx' = 1 \forall (x, a) \in \mathcal{X} \times \mathcal{A}$, $r : \mathcal{X} \times \mathcal{A} \rightarrow \mathbb{R}$ is a measurable function, called the reward function and $0 \leq \gamma < 1$ is the discount factor. We further assume the followings:

Assumption 2.1. \mathcal{A} is finite.

Assumption 2.2. There exist constants $K_p, L_p > 0$ s.t. $\|p\|_\infty \leq K_p$ and $p(y|\cdot, a) \in \operatorname{Lip}_1(L_p)$ for all $(y, a) \in \mathcal{X} \times \mathcal{A}$.

Assumption 2.3. There exists some constant $K_r > 0$ s.t. $\|r\|_\infty < K_r$.

3. The Algorithm

The pseudo-code of the algorithm yielding uniformly approximately optimal policies can be seen in Table 1. Note that at the expense of increasing the computation time one may downscale the storage requirement of the algorithm from $O(N^2)$ to $O(N)$ if Step 3 of the algorithm is omitted. Then Equation (2) must be used in Steps 5 and 6. Note that one may still precompute the normalizing fac-

tor of (2) for speeding up the computations since the storage requirements for these normalizing factors depend only linearly on N .

Rust's original algorithm builds up the cache $C_\varepsilon = (v_1, \dots, v_N)$ using steps 1 – 5 with some N and t . Then, for any query state $x \in \mathcal{X}$ his algorithm returns the random value

$$\hat{V}(x) = \max_{a \in \mathcal{A}} \{r(x, a) + \gamma \sum_{j=1}^N \hat{p}_{X^{1:N}}(X_j|x, a) v_j\}.$$

It can be readily seen that our algorithm is just a straightforward extension of the one considered by Rust, the difficulty lies in deriving appropriate bounds for N and t .

Now, we introduce the notations needed to state the main results. Let $T_a : B(\mathcal{X}) \rightarrow B(\mathcal{X})$ be defined by

$$(T_a V)(x) = r(x, a) + \gamma \int p(y|x, a) V(y) dy.$$

Here $a \in \mathcal{A}$ is arbitrary and the integral should be understood here and in what follows to be over \mathcal{X} . For a stationary policy $\pi : \mathcal{X} \rightarrow \mathcal{A}$, let $T_\pi : B(\mathcal{X}) \rightarrow B(\mathcal{X})$ be defined by

$$(T_\pi V)(x) = (T_{\pi(x)} V)(x).$$

Finally, let the Bellman-operator $T : B(\mathcal{X}) \rightarrow B(\mathcal{X})$ be defined by

$$(TV)(x) = \max_{a \in \mathcal{A}} \{(T_a V)(x)\}.$$

Under our assumptions, T is known to have a unique fixed-point, V^* , called the optimal-value function. V^* is known to be uniformly bounded. It is also known that any (stationary) policy $\pi : \mathcal{X} \rightarrow \mathcal{A}$ satisfying $T_\pi V^* = TV^*$ is optimal in the sense that for any given initial state the total expected discounted return resulting from the execution of π is maximal. (The execution of a policy $\pi : \mathcal{X} \rightarrow \mathcal{A}$ means the execution of action $\pi(x)$ whenever the state is x .) A policy is called myopic or greedy w.r.t. the function $V \in B(\mathcal{X})$ if $T_\pi V = TV$. Since in our case the action set \mathcal{A} is finite, the existence of a myopic policy is guaranteed for any given uniformly bounded function V .

Now let $x_1, \dots, x_N \in \mathcal{X}$ be fixed elements of the state space. For brevity, let us denote the N -tuple

(x_1, \dots, x_N) by $x^{1:N}$. Let $\hat{T}_{x^{1:N} a} : B(\mathcal{X}) \rightarrow B(\mathcal{X})$ be defined by

$$(\hat{T}_{x^{1:N} a} V)(x) = r(x, a) + \gamma \sum_{i=1}^N \hat{p}_{x^{1:N}}(x_i | x, a) V(x_i), \quad (1)$$

where

$$\hat{p}_{x^{1:N}}(x_i | x, a) = \begin{cases} \frac{p(x_i | x, a)}{\sum_{j=1}^N p(x_j | x, a)}; & \text{if } \sum_{j=1}^N p(x_j | x, a) > 0, \\ 0; & \text{otherwise.} \end{cases} \quad (2)$$

The operator $\hat{T}_{x^{1:N} a}$ is obtained from T_a by approximating the integrals in T_a by finite sums. It should be clear that because of the Lipschitz conditions on p , $\hat{T}_{x^{1:N} a}$ does approximate T_a and the quality of approximation depends on the distribution of the points $x^{1:N}$. Using $\hat{T}_{x^{1:N} a}$ we introduce the operator $\hat{T}_{x^{1:N}}$ that is meant to approximate T . It is defined as follows: $\hat{T}_{x^{1:N}} : B(\mathcal{X}) \rightarrow B(\mathcal{X})$, and

$$(\hat{T}_{x^{1:N}} V)(x) = \max_{a \in \mathcal{A}} \{(\hat{T}_{x^{1:N} a} V)(x)\}. \quad (3)$$

Now, analogously with the previous definitions, $\hat{T}_{x^{1:N} \pi}$ is introduced by

$$(\hat{T}_{x^{1:N} \pi} V)(x) = (\hat{T}_{x^{1:N} \pi(x)} V)(x).$$

Throughout the paper we are going to work with independent random variables X_1, \dots, X_N , being uniformly distributed over \mathcal{X} .⁴ Similarly to the notation introduced for deterministic N -tuples of state space points, $X^{1:N}$ will be used to denote (X_1, \dots, X_N) . We define the random operators $\hat{T}_{N a}$, $\hat{T}_{N \pi}$ and \hat{T}_N by the respective equations

$$\hat{T}_{N a} = \hat{T}_{X^{1:N} a}, \quad \hat{T}_{N \pi} = \hat{T}_{X^{1:N} \pi}, \quad \text{and} \\ \hat{T}_N = \hat{T}_{X^{1:N}}.$$

⁴The uniform distribution is used for simplicity only. Any other sampling distribution with support covering \mathcal{X} could be used if the algorithm is modified appropriately (importance sampling) [7]. The form of the ideal sampling distribution is far from being clear since a single sample-set is used to estimate an infinite number of integrals. The form of the ideal distribution should be the subject of future research.

Here \hat{T}_N is called the *random Bellman-operator*. A great deal of effort in this paper will be devoted to show that \hat{T}_N and its powers approximate the true Bellman-operator T and its respective powers uniformly well, with high probability.

In order to connect the algorithm with the operators defined so far, let us introduce the projection operator $\hat{P}_{x^{1:N}} : B(\mathcal{X}) \rightarrow \mathbb{R}^N$ defined by

$$(\hat{P}_{x^{1:N}} V) = (V(x_1), \dots, V(x_N)),$$

and the ‘‘expansion’’ operators $\hat{E}_{x^{1:N} a}, \hat{E}_{x^{1:N}} : \mathbb{R}^N \rightarrow B(\mathcal{X})$ defined by the respective equations

$$(\hat{E}_{x^{1:N} a} v)(x) = r(x, a) + \gamma \sum_{j=1}^N \hat{p}_{x^{1:N}}(x_j | x, a) v(x_j), \quad a \in \mathcal{A},$$

and

$$(\hat{E}_{x^{1:N}} v)(x) = \max_{a \in \mathcal{A}} \{(\hat{E}_{x^{1:N} a} v)(x)\}.$$

Finally, let the finite state-space Bellman operator $\hat{L}_{x^{1:N}} : \mathbb{R}^N \rightarrow \mathbb{R}^N$ be defined by

$$(\hat{L}_{x^{1:N}} v)_i = \max_{a \in \mathcal{A}} \{r(x_i, a) + \gamma \sum_{j=1}^N \hat{p}_{x^{1:N}}(x_j | x_i, a) v_j\}.$$

The following proposition highlights the connection between the algorithm and these operators:

Proposition 3.1. *For any integer $t > 0$,*

$$\hat{T}_{x^{1:N}}^{t+1} = \hat{E}_{x^{1:N}} \hat{L}_{x^{1:N}}^t \hat{P}_{x^{1:N}}$$

and in particular,

$$\hat{T}_N^{t+1} = \hat{E}_{X^{1:N}} \hat{L}_{X^{1:N}}^t \hat{P}_{X^{1:N}}.$$

Proof. By inspection. \square

Remark 3.2. *According to Proposition 3.1, one can compute $(\hat{T}_N^{t+1} V)(x)$ in two phases, the first of which we could call the off-line phase and the second of which we could call the on-line phase. In the off-line phase one computes the N -dimensional vector $v_N^{(t)} = \hat{L}_{X^{1:N}}^t \hat{P}_{X^{1:N}} V$, which takes $O(tN^2|\mathcal{A}|)$ time, whilst in the second phase one computes the value of $(\hat{T}_N^{t+1} V)(x)$ by evaluating $(\hat{T}_N^{t+1} V)(x) = (\hat{E}_{X^{1:N}} v_N^{(t)})(x)$. This second*

step takes $O(N^2|\mathcal{A}|)$ time and thus the whole procedure takes $O(tN^2|\mathcal{A}|)$ time. Further, it is easy to see that the procedure takes $O(N + |\mathcal{A}|)$ space.⁵

Now, the algorithm whose pseudocode was given above can be formulated as follows:

Assume that we are given a fixed tolerance, $\varepsilon > 0$. On the basis of ε and $L_p, K_r, |\mathcal{A}|, \frac{1}{1-\gamma}$ we compute some integer $t > 0$ and another integer $N > 0$. Each time we need to compute an action of the randomized policy π for some state x , we draw a random sample $X^{1:N}$ and compute $v_N^{(t)} = \hat{L}_{X^{1:N}}^t \hat{P}_{X^{1:N}} V_0$ where $V_0(x) = 0$. Then a random action of $\pi(x)$ is computed by evaluating

$$\operatorname{argmax}_{a \in \mathcal{A}} (\hat{E}_{X^{1:N}} v_N^{(t)})(x). \quad (4)$$

The action of the argmax operator is returned. The resulting policy will be shown to be ε -optimal.

Another, computationally less expensive method is to hold the random sample $X^{1:N}$ fixed and compute $v_N^{(t)}$ only once. Then the computation of $\pi(x)$ using (4) costs only $O(|\mathcal{A}|N^2)$ steps.

4. Results

The first result that we will prove shows that the algorithm just described at the end of the previous section yields uniformly approximately optimal policies with high probability and has polynomial complexity:

Theorem 4.1. *Let $K = K_r/(1-\gamma)$ and let $\varepsilon > 0$, $\delta > 0$, $V_0 \in B_K(\mathcal{X})$ be fixed. Let $t = t(\varepsilon, \gamma, K)$, where*

$$t(\varepsilon, \gamma, K) = \left\lceil \frac{\log(8K) + \log(1/(\varepsilon(1-\gamma)))}{\log(1/\gamma)} \right\rceil$$

and let

$$N \geq 512K^2 K_p^2 \left(\frac{24(K+1)}{\varepsilon(1-\gamma)^2} \right)^2 \left(\log 8 + \log(t(\varepsilon, \gamma, K) + 1) + \log |\mathcal{A}| + d \log \left(\left\lceil \frac{384(K+1)^2 L_p d}{\varepsilon(1-\gamma)^2} \right\rceil + 1 \right) + \log \left(\frac{1}{\delta} \right) \right).$$

⁵Here we assume that the basic algebraic operations over reals take $O(1)$ time and that the storage of a real-number takes $O(1)$ space. We also assume that $\hat{p}_{X^{1:N}}$ is not stored.

Let $V = \hat{T}_N^t V_0$ and let the stationary policy π be defined by $\hat{T}_N \pi V = \hat{T}_N V$. Then

$$\mathbb{P}(\|V_\pi - V^*\|_\infty \leq \varepsilon) \geq 1 - \delta.$$

Further, the complexity of the algorithm is polynomial in $d, \varepsilon, K, K_p, \log(L_p), |\mathcal{A}|$ and $1/(1-\gamma)$.

Note that ideally the bound on N should depend only on K/ε , so that scaling the rewards would not change the complexity results. The bound given in the above theorem does not have this property, it has some “ K ”s without a corresponding ε . The cause of this will become clear during the course of the proof of this theorem and, more specifically, in the proof of Lemma 5.7. Note that if ε is sufficiently small, an upper bound on the above expression can always be derived by replacing K by K/ε at those occurrences of K that lack a corresponding ε term. This way one gets a less tight, but (in some sense) a better behaving bound.

The next result shows that the modified, fully on-line algorithm given in Table 1 yields a uniformly approximately optimal policy and has polynomial complexity. The above comments on scaling the rewards apply to this result, too.

Theorem 4.2. *Let $K = K_r/(1-\gamma)$ and let $\varepsilon > 0$. Fix some $V_0 \in B_K(\mathcal{X})$. Let $\varepsilon' = \varepsilon(1-\gamma)/(2(1+\gamma))$, $\delta = \varepsilon(1-\gamma)/(4K) (= \varepsilon(1-\gamma)^2/(4K_r))$. Further, let $t = t(\varepsilon', \gamma, K)$ and let N be the smallest integer larger than*

$$512K^2 K_p^2 \left(\frac{96(K+1)}{\varepsilon(1-\gamma)^3} \right)^2 \left(\log 8 + \log(t+1) + \log |\mathcal{A}| + d \log \left(\left\lceil \frac{768(K+1)^2 L_p d}{\varepsilon(1-\gamma)^3} \right\rceil + 1 \right) + \log \left(\frac{4K}{\varepsilon(1-\gamma)} \right) \right).$$

Let $V = \hat{T}_N^t V_0$ and let the stochastic stationary policy $\pi : \mathcal{X} \times \mathcal{A} \rightarrow [0, 1]$ be defined by $\pi(x, a) = \mathbb{P}(\pi_{X^{1:N}}(x) = a)$, where $\pi_{X^{1:N}}$ is the policy defined by $\hat{T}_N \pi_{X^{1:N}} V = \hat{T}_N V$. Then, π is ε -optimal and given a state x , a random action of π can be computed in time and space polynomial in $K_r/\varepsilon, d, K_r, K_p, \log L_p, |\mathcal{A}|$ and $1/(1-\gamma)$.

The rough outlines of the proofs of these theorems are as follows: Under our assumptions, Pollard’s maximal inequality (cf. [10]) ensures that for any given fixed function V_0 , $\left\|\hat{T}_N V_0 - T V_0\right\|_\infty$ is small with high probability.⁶ Using the triangle inequality one reduces the comparison of $\hat{T}_N^n V_0$ and $T^n V_0$ to those of $\hat{T}_N T^k V_0$ and $T T^k V_0$, where k varies from zero to $n - 1$. More precisely, one shows that if the differences between $\hat{T}_N T^k V_0$ and $T T^k V_0$ are small for all $k = 0, \dots, n - 1$ then $\left\|\hat{T}_N^n V_0 - T^n V_0\right\|_\infty$ will be small, too. Using this result, it is then easy to prove a maximal inequality for $\left\|\hat{T}_N^n V_0 - T^n V_0\right\|_\infty$.

Now, one can use standard contraction arguments to prove an inequality that bounds the difference of the value of a policy that is “approximately” greedy w.r.t. some function V in terms of the Bellman-residuals (see e.g. [14]). The plan is to use this inequality for $V = \hat{T}_N^n V_0$ and \hat{T}_N . Some more calculations yield Theorem 4.1.

Then, it is proven that if a policy selects only “good actions” (i.e., actions from $\mathcal{A}_\varepsilon(x) = \{a \in \mathcal{A} : (T_a V^*)(x) \geq (T V^*)(x) - \varepsilon\}$ for a suitable ε) then it is “good” itself (i.e., close to optimal). Next, we relax the condition of “selecting good actions” to “selecting good actions with high probability”. Such policies can be shown to be “good”, as well (cf. Lemma 5 of [9]). Finally, it is shown that if a policy is good with high probability then it selects good actions with high probability and thus, in turn, it must be “good”. This will finish the proof of Theorem 4.2.

One source of the complexity of the proof stems from the fact that Pollard’s inequality cannot be used in a simple way to bound $\left\|\hat{T}_N^n V_0 - T^n V_0\right\|_\infty$. This is because the usual induction argument that would bound $\left\|\hat{T}_N^n V_0 - T^n V_0\right\|_\infty$ based on a bound on $\left\|\hat{T}_N^{n-1} V_0 - T^{n-1} V_0\right\|_\infty$ does not quite work here. Typically, one argues that if \hat{T}_N approximates T uniformly well over the space of bounded functions (or some space of functions of interest) then $\left\|\hat{T}_N^n V_0 - T^n V_0\right\|_\infty$ will be small if

$\left\|\hat{T}_N^{n-1} V_0 - T^{n-1} V_0\right\|_\infty$ is small. Unfortunately, the space of all bounded functions is just too rich in our case: \hat{T}_N cannot approximate T uniformly well over this rather complex space. A smaller, but still appropriate space \mathcal{F} is needed - hence the complicated proof.

5. Proof

We prove the theorem in the next three sections. First, we prove some maximal inequalities for the random Bellman-operators $\hat{T}_{N a}$. Next we show how these can be extended to powers of \hat{T}_N and, finally, we apply all these to prove the main results.

5.1. Maximal Inequalities for Random Bellman Operators

We shall need some auxilliary operators which are easier to deal with using probability theory. Let

$$\tilde{T}_{N a} : B(\mathcal{X}) \rightarrow B(\mathcal{X})$$

$$(\tilde{T}_{N a} V)(x) = r(x, a) + \frac{\gamma}{N} \sum_{i=1}^N p(X_i | x, a) V(X_i);$$

$$\tilde{T}_N : B(\mathcal{X}) \rightarrow B(\mathcal{X})$$

$$(\tilde{T}_N V)(x) = \max_{a \in \mathcal{A}} \{(\tilde{T}_{N a} V)(x)\}.$$

Operator $\tilde{T}_{N a}$ is a simple Monte-Carlo estimate of operator T_a and will be shown to converge uniformly to T_a using standard methods. Unfortunately, $\tilde{T}_{N a}$ is not suitable for further analysis as it can be a non-contraction, and in order to analyze the iterations in our algorithms, the contraction property of the approximate Bellman operators will be needed. Hence the algorithms use $\hat{T}_{N a}$ and in a second step $\tilde{T}_{N a}$ will be related to $\hat{T}_{N a}$, and the approximation results will be extended to $\hat{T}_{N a}$.

We need some definitions and results from the theory of uniform deviations (cf. [10]).

Definition 5.1. *Let $A \subseteq \mathbb{R}^d$. The set $S \subseteq A$ is an ε -cover of A if for all $t \in A$ there exists an element s of S s.t. $\frac{1}{d} \|t - s\|_1 \leq \varepsilon$. The set of ε -covers of A will be denoted $\mathcal{C}(A; \varepsilon)$.*

⁶We must rely on Pollard’s maximal inequality instead of the simpler Chernoff-bounds because the state space is continuous and the sup-norm above involves a supremum over the state space. Further, this result is derived in two steps, using an idea of Rust [13].

Definition 5.2. The ε -covering number of a set A is defined by

$$\mathcal{N}(\varepsilon, A) = \min\{|S| : S \in \mathcal{C}(A; \varepsilon)\}.$$

The number $\log \mathcal{N}(\varepsilon, A)$ is called the ‘‘metric entropy’’ of A . Let $z^{1:n} = (z_1, \dots, z_n) \in (\mathbb{R}^d)^n$ and let $\mathcal{F} \subseteq \mathbb{R}^{\mathbb{R}^d}$. We define

$$\mathcal{F}(z^{1:n}) = \{(f(z_1), \dots, f(z_n)) : f \in \mathcal{F}\} \subseteq \mathbb{R}^n. \quad (5)$$

The following theorem is due to Pollard (see [10]):

Theorem 5.3 (Pollard, 1984). Let $n > 0$ be an integer, $\varepsilon > 0$, $M > 0$, $\mathcal{F} \subseteq [0, M]^{\mathbb{R}^d}$ be a set of measurable functions. Let $X_1, \dots, X_n \in \mathbb{R}^d$ be i.i.d. random variables. Then

$$\begin{aligned} \mathbb{P} \left(\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n f(X_i) - \mathbb{E}[f(X_1)] \right| > \varepsilon \right) &\leq \\ 8\mathbb{E} \left[\mathcal{N} \left(\frac{\varepsilon}{8}, \mathcal{F}(X^{(1:n)}) \right) \right] e^{-\frac{n\varepsilon^2}{128M^2}}. &\quad (6) \end{aligned}$$

An elegant proof of this theorem can be found in [5][pp. 492]. In general, some further assumptions are needed to make the result of the above sup measurable. Measurability problems, however, are now well understood so we shall not worry about this detail. Readers who keep worrying should take all the probability bounds except for the main result as outer/inner-probability bounds (whichever is appropriate). Note that in the final result we work with measurable sets and therefore there is no need to refer to outer/inner probability measures.

Firstly, we extend this theorem to functions mapping \mathbb{R}^d into $[-M, M]$.

Corollary 5.3.1. Let $n > 0$ be an integer, $\varepsilon > 0$, $M > 0$, $\mathcal{F} \subseteq [-M, M]^{\mathbb{R}^d}$ be a set of measurable functions. Let $X_1, \dots, X_n \in \mathbb{R}^d$ be i.i.d. random variables. Then

$$\begin{aligned} \mathbb{P} \left(\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n f(X_i) - \mathbb{E}[f(X_1)] \right| > \varepsilon \right) &\leq \\ 8\mathbb{E} \left[\mathcal{N} \left(\frac{\varepsilon}{8}, \mathcal{F}(X^{(1:n)}) \right) \right] e^{-\frac{n\varepsilon^2}{512M^2}}. &\quad (7) \end{aligned}$$

Proof. Apply Theorem 5.3 to $f^M = f + M$. \square

Definition 5.4. Let $d \in \mathbb{N}$, $d > 0$ and let $\sigma > 0$. Let

$$\begin{aligned} \text{Grid}(\sigma) = &\left\{ (2i_1\sigma, \dots, 2i_d\sigma) \in [0, 1]^d : \right. \\ &\left. 0 \leq i_k \leq \left\lfloor \frac{1}{2\sigma} \right\rfloor, 1 \leq k \leq d, i_k \in \mathbb{N} \right\} \end{aligned}$$

and let $P_\sigma : [0, 1]^d \rightarrow \text{Grid}(\sigma)$ be defined by

$$P_\sigma x = \operatorname{argmin}_y \{ \|x - y\|_1 : y \in \text{Grid}(\sigma) \}$$

where ties are broken in favor of points having smaller coordinates.

Remark 5.5. $\|x - P_\sigma x\|_1 \leq d\sigma$ and $|\text{Grid}(\sigma)| \leq (\lfloor \frac{1}{2\sigma} \rfloor + 1)^d$

Now we can prove our first result concerning the approximation of T_a by $\tilde{T}_{N,a}$.

Lemma 5.6. Let $K > 0$, $\varepsilon > 0$ and $\delta > 0$. Further, let $B_0 \subseteq B_K(\mathcal{X})$ be a finite set,

$$\begin{aligned} p_1(d, \varepsilon, \delta, K, K_p, L_p, |B_0|, |\mathcal{A}|, \gamma) = & \\ \frac{512K^2K_p^2}{\varepsilon^2} \left(\log 8 + \log |B_0| + \log |\mathcal{A}| \right) & \\ + d \log \left(\left\lfloor \frac{16KL_p d}{\varepsilon} \right\rfloor + 1 \right) + \log \left(\frac{1}{\delta} \right) &\quad (8) \end{aligned}$$

and $N \geq p_1(d, \varepsilon, \delta, K, K_p, L_p, |B_0|, |\mathcal{A}|, \gamma)$. Then

$$\mathbb{P} \left(\max_{V \in B_0} \max_{a \in \mathcal{A}} \left\| \tilde{T}_{N,a} V - T_a V \right\|_\infty > \varepsilon \right) \leq \delta. \quad (9)$$

Proof. We shall make use of Corollary 5.3.1. Let $\mathcal{F}(x^{1:N}) = \{z_N(V, x, a) : V \in B_0, a \in \mathcal{A}, x \in \mathcal{X}\}$, where

$$\begin{aligned} z_N(V, x, a) = & \\ (V(X_1)p(X_1|x, a), \dots, V(X_N)p(X_N|x, a)). & \end{aligned}$$

Easily, $z_N(V, x, a) \subseteq [-K K_p, K K_p]^N$. In order to bound $\mathcal{N}(\varepsilon, \mathcal{F}(X^{1:N}))$ from above, we construct an ε -cover of $\mathcal{F}(X^{1:N})$. We claim that

$$S_\sigma = \{z_N(V, x, a) : V \in B_0, a \in \mathcal{A}, x \in \text{Grid}(\sigma)\}$$

is an ε -cover of $\mathcal{F}(X^{1:N})$ if σ is chosen appropriately. In order to prove this let us pick up an arbitrary element $z_N(V, x, a)$ of $\mathcal{F}(X^{1:N})$. Then

$$\begin{aligned} & \frac{1}{N} \|z_N(V, x, a) - z_N(V, P_\sigma x, a)\|_1 = \\ & \frac{1}{N} \sum_{i=1}^N |V(X_i)p(X_i|x, a) - V(X_i)p(X_i|P_\sigma x, a)| \leq \\ & \frac{\|V\|_\infty}{N} \sum_{i=1}^N |p(X_i|x, a) - p(X_i|P_\sigma x, a)| \leq \\ & KL_p d \sigma. \end{aligned}$$

Therefore, if $\sigma = \varepsilon/(KL_p d)$ then S_σ is an ε -cover of $\mathcal{F}(X^{1:N})$. By Remark 5.5, $\mathcal{N}(\varepsilon, \mathcal{F}(X^{1:N})) \leq |B_0| |\mathcal{A}| \left(\left\lfloor \frac{2KL_p d}{\varepsilon} \right\rfloor + 1 \right)^d$. By Corollary 5.3.1, if

$$\begin{aligned} & \log 8 + \log |B_0| + \log |\mathcal{A}| \\ & + d \log \left(\left\lfloor \frac{16KL_p d}{\varepsilon} \right\rfloor + 1 \right) + \log \frac{1}{\delta} \leq \frac{N\varepsilon^2}{512K^2K_p^2} \end{aligned}$$

then (9) holds. \square

Now, we shall prove a similar result for \hat{T}_{Na} , using ideas from the proof of the Corollary to Theorem 3.4 of [13].

Lemma 5.7. *Let $K > 0$, $\varepsilon > 0$ and $\delta > 0$. Further, let $B_0 \subseteq B_K(\mathcal{X})$ be a finite set,*

$$\begin{aligned} & p_2(d, \varepsilon, \delta, K, K_p, L_p, |B_0|, |\mathcal{A}|, \gamma) = \\ & 512K^2K_p^2 \left(\frac{K+1}{\varepsilon} \right)^2 \left(\log 8 + \log(|B_0| + 1) \right. \\ & \left. + \log |\mathcal{A}| + d \log \left(\left\lfloor \frac{16(K+1)^2 L_p d}{\varepsilon} \right\rfloor + 1 \right) \right. \\ & \left. + \log \left(\frac{1}{\delta} \right) \right) \end{aligned} \quad (10)$$

If $N \geq p_2(d, \varepsilon, \delta, K, K_p, L_p, |B_0|, |\mathcal{A}|, \gamma)$ then

$$\mathbb{P} \left(\max_{V \in B_0} \max_{a \in \mathcal{A}} \left\| \hat{T}_{Na} V - T_a V \right\|_\infty > \varepsilon \right) \leq \delta. \quad (11)$$

Proof. Let us pick up some $V \in B_0$. By the triangle inequality

$$\begin{aligned} \left\| \hat{T}_{Na} V - T_a V \right\|_\infty & \leq \left\| \hat{T}_{Na} V - \tilde{T}_{Na} V \right\|_\infty \\ & + \left\| \tilde{T}_{Na} V - T_a V \right\|_\infty. \end{aligned} \quad (12)$$

Let

$$\bar{p}_N(x, a) = \frac{1}{N} \sum_{i=1}^N p(X_i|x, a).$$

If $\bar{p}_N(x, a) = 0$ then $\left| (\hat{T}_{Na} V)(x) - (\tilde{T}_{Na} V)(x) \right| = 0$. If $\bar{p}_N(x, a) \neq 0$ then by simple algebraic manipulations we get

$$\begin{aligned} & \left| (\hat{T}_{Na} V)(x) - (\tilde{T}_{Na} V)(x) \right| = \\ & \frac{\gamma |1 - \bar{p}_N(x, a)|}{\bar{p}_N(x, a)} \frac{1}{N} \left| \sum_{i=1}^N p(X_i|x, a) V(X_i) \right|. \end{aligned}$$

Since, by assumption $|V(X_i)| \leq K$, we have

$$\left| (\hat{T}_{Na} V)(x) - (\tilde{T}_{Na} V)(x) \right| \leq \gamma K |1 - \bar{p}_N(x, a)|. \quad (13)$$

Let $e : \mathcal{X} \rightarrow \mathbb{R}$ be defined by $e(x) = 1$ and observe that

$$\gamma(1 - \bar{p}_N(x, a)) = (T_a e)(x) - (\tilde{T}_{Na} e)(x)$$

and therefore by (13) we have

$$\begin{aligned} & \left| (\hat{T}_{Na} V)(x) - (\tilde{T}_{Na} V)(x) \right| \leq \\ & K \left| (T_a e)(x) - (\tilde{T}_{Na} e)(x) \right|. \end{aligned}$$

Note that this inequality holds also when $\bar{p}_N(x, a) = 0$. Taking the supremum over \mathcal{X} yields

$$\left\| \hat{T}_{Na} V - \tilde{T}_{Na} V \right\|_\infty \leq K \left\| T_a e - \tilde{T}_{Na} e \right\|_\infty.$$

By (12) we have

$$\begin{aligned} & \left\| \hat{T}_{Na} V - T_a V \right\|_\infty \leq \\ & K \left\| \tilde{T}_{Na} e - T_a e \right\|_\infty + \left\| \tilde{T}_{Na} V - T_a V \right\|_\infty \leq \\ & (K+1) \max_{V' \in B_0 \cup \{e\}} \left\| \tilde{T}_{Na} V' - T_a V' \right\|_\infty. \end{aligned}$$

Therefore

$$\begin{aligned} & \max_{V \in B_0} \max_{a \in \mathcal{A}} \left\| \hat{T}_{Na} V - T_a V \right\|_\infty \leq \\ & (K+1) \max_{a \in \mathcal{A}} \max_{V \in B_0 \cup \{e\}} \left\| \tilde{T}_{Na} V - T_a V \right\|_\infty. \end{aligned}$$

Now, the statement of the lemma follows using Lemma 5.6 with the choice $N \geq p_1(d, \varepsilon/(K+1), \delta, K+1, K_p, L_p, |B_0|+1, |\mathcal{A}|, \gamma)$. \square

5.2. Maximal Inequalities for Powers of Random Bellman Operators

First we need a proposition that relates the fixed point of a contraction operator and an operator that is "approximating" the contraction.

Proposition 5.8. *Let B be a space of bounded functions⁷, and fix some $V \in B$ and integer $t > 0$. Let $T_1, T_2 : B \rightarrow B$ be operators of B such that $T_1 \in \text{Lip}_\infty(\gamma)$ for some $0 \leq \gamma < 1$ and*

$$\|T_1 T_2^s V - T_2 T_2^s V\|_\infty \leq \alpha, \quad 0 \leq s \leq t-1 \quad (14)$$

for some $\alpha > 0$. Then

$$\|T_1^t V - T_2^t V\|_\infty \leq \frac{\alpha}{1-\gamma}. \quad (15)$$

Proof. We prove the statement by induction; namely, we prove that

$$\|T_1^s V - T_2^s V\|_\infty \leq \frac{\alpha}{1-\gamma} \quad (16)$$

holds for all $0 \leq s \leq t$. The statement is obvious for $s = 0$. Assume that we have already proven (16) for $s - 1$. By the triangle inequality, $\|T_1^s V - T_2^s V\|_\infty \leq \|T_1 T_1^{s-1} V - T_1 T_2^{s-1} V\|_\infty + \|T_1 T_2^{s-1} V - T_2 T_2^{s-1} V\|_\infty$. Since $T_1 \in \text{Lip}_\infty(\gamma)$, the first term of the rhs can be bounded by $\gamma \|T_1^{s-1} V - T_2^{s-1} V\|_\infty$, which in turn can be bounded by $\gamma\alpha/(1-\gamma)$, by the induction hypothesis. The second term, on the other hand, can be bounded by α , by (14). Since $\gamma\alpha/(1-\gamma) + \alpha = \alpha/(1-\gamma)$, inequality (16) holds for s as well, thus proving the proposition. \square

We cite the next proposition without proof, as the proof is both elementary and is well known.

Proposition 5.9. *Let $K = K_r/(1-\gamma)$. Then the Bellman-operator T maps $B_K(\mathcal{X})$ into $B_K(\mathcal{X})$.*

Now follows the main result of this section.

Lemma 5.10. *Let $t > 0$ be an integer, $\varepsilon > 0$, $\delta > 0$, $K = K_r/(1-\gamma)$, $V_0 \in B_K(\mathcal{X})$. Let*

$$\begin{aligned} p_3(d, \varepsilon, \delta, K, K_p, L_p, |B_0|, |\mathcal{A}|, \gamma) = & \\ & 512K^2 K_p^2 \left(\frac{K+1}{\varepsilon(1-\gamma)} \right)^2 \left(\log 8 + \log(|B_0| + 1) \right. \\ & + \log |\mathcal{A}| + d \log \left(\left\lfloor \frac{16(K+1)^2 L_p d}{\varepsilon(1-\gamma)} \right\rfloor + 1 \right) \\ & \left. + \log \left(\frac{1}{\delta} \right) \right). \end{aligned} \quad (17)$$

If $N \geq p_3(d, \varepsilon, \delta, K, K_p, L_p, |B_0|, |\mathcal{A}|, \gamma)$ then

$$\begin{aligned} \mathbb{P} \left(\max \left\{ \max_{a \in \mathcal{A}} \left\| \hat{T}_{N a} T^t V_0 - T_a T^t V_0 \right\|_\infty, \right. \right. \\ \left. \left. \left\| \hat{T}_N^t V_0 - T^t V_0 \right\|_\infty \right\} \leq \varepsilon \right) \geq 1 - \delta. \end{aligned}$$

Proof. Let $V_s = T^s V_0$, $0 \leq s \leq t$, $B_0 = \{V_0, V_1, \dots, V_t\}$. By Proposition 5.9, $B_0 \subseteq B_K(\mathcal{X})$. By Lemma 5.7, if

$$N \geq p_2(d, \varepsilon(1-\gamma), \delta, K, K_p, L_p, |B_0|, |\mathcal{A}|, \gamma)$$

then

$$\mathbb{P} \left(\max_{V \in B_0} \max_{a \in \mathcal{A}} \left\| \hat{T}_{N a} V - T_a V \right\|_\infty \leq \varepsilon(1-\gamma) \right) \geq 1 - \delta.$$

Let the elementary random event ω be such that

$$\max_{V \in B_0} \max_{a \in \mathcal{A}} \left\| \hat{T}_{N a}(\omega) V - T_a V \right\|_\infty \leq \varepsilon(1-\gamma).$$

If we show that

$$\begin{aligned} \max \left(\max_{a \in \mathcal{A}} \left\| \hat{T}_{N a}(\omega) T^t V_0 - T_a T^t V_0 \right\|_\infty, \right. \\ \left. \left\| \hat{T}_N(\omega)^t V_0 - T^t V_0 \right\|_\infty \right) \leq \varepsilon \end{aligned} \quad (18)$$

then the proof will be finished.

Obviously,

$$\max_{a \in \mathcal{A}} \left\| \hat{T}_{N a}(\omega) T^t V_0 - T_a T^t V_0 \right\|_\infty \leq \varepsilon \quad (19)$$

by the construction of B_0 and since $1-\gamma \leq 1$.

Now, note that

⁷More generally, B could be any Banach-space.

$$\left\| \hat{T}_N(\omega)V - TV \right\|_\infty \leq \max_{a \in \mathcal{A}} \left\| \hat{T}_{N,a}(\omega)V - T_a V \right\|_\infty$$

holds for all $V \in B(\mathcal{X})$. Since by the choice of ω and N ,

$$\max_{a \in \mathcal{A}} \left\| \hat{T}_{N,a}(\omega)T^s V_0 - T_a T^s V_0 \right\|_\infty \leq \varepsilon(1 - \gamma),$$

$$0 \leq s \leq t,$$

we also have

$$\left\| \hat{T}_N(\omega)T^s V_0 - T T^s V_0 \right\|_\infty \leq \varepsilon(1 - \gamma), \quad 0 \leq s \leq t.$$

Moreover, since $\hat{T}_N(\omega) \in \text{Lip}_\infty(\gamma)$, Proposition 5.8 can be applied with the choice $B = B(\mathcal{X})$, $T_1 = \hat{T}_N(\omega)$, $T_2 = T$ and $V = V_0$, yielding

$$\left\| \hat{T}_N(\omega)^t V_0 - T^t V_0 \right\|_\infty \leq \varepsilon.$$

This together with (19) yields (18), thus proving the theorem. \square

5.3. Proving the ε -optimality of the Algorithm

First, we prove an inequality similar to that of [14], but here we use both approximate value functions and approximate operators.

Lemma 5.11. *Let $V \in B(\mathcal{X})$, $x^{1:N} \in \mathcal{X}^N$ for some $N > 0$ and let $\pi : \mathcal{X} \rightarrow \mathcal{A}$ be such that*

$$\hat{T}_{x^{1:N}, \pi} V = \hat{T}_{x^{1:N}} V.$$

Then

$$\|V_\pi - V^*\|_\infty \leq \frac{2}{1 - \gamma} \left(\max_{a \in \mathcal{A}} \left\| T_a V - \hat{T}_{x^{1:N}, a} V \right\|_\infty + \gamma \|TV - V\|_\infty \right). \quad (20)$$

Note that since \mathcal{A} is finite, the policy defined in the lemma exists.

Proof. We compare $T_\pi^k V$ and $T^k V$ since these are known to converge to V_π and V^* , respectively. Firstly, we write the difference $T_\pi^k V - T^k V$ in the form of a telescoping sum:

$$T_\pi^k V - T^k V = \sum_{i=1}^{k-1} (T_\pi^{i+1} V - T_\pi^i V) + (T_\pi V - TV) - \sum_{i=1}^{k-1} (T^{i+1} V - T^i V).$$

Using the triangle inequality, the relations $T, T_\pi \in \text{Lip}_\infty(\gamma)$, and the inequality $\gamma^{k-1} + \gamma^{k-2} + \dots + \gamma \leq \gamma/(1 - \gamma)$, we get

$$\|T_\pi^k V - T^k V\|_\infty \leq \frac{\gamma}{1 - \gamma} \left(\|T_\pi V - V\|_\infty + \|TV - V\|_\infty \right) + \|T_\pi V - TV\|_\infty. \quad (21)$$

Using the identity $\hat{T}_{x^{1:N}, \pi} V = \hat{T}_{x^{1:N}} V$, we write

$$T_\pi V - TV = (T_\pi V - \hat{T}_{x^{1:N}, \pi} V) + (\hat{T}_{x^{1:N}} V - TV)$$

and thus

$$\|T_\pi V - TV\|_\infty \leq \left\| T_\pi V - \hat{T}_{x^{1:N}, \pi} V \right\|_\infty + \left\| \hat{T}_{x^{1:N}} V - TV \right\|_\infty \leq 2 \max_{a \in \mathcal{A}} \left\| T_a V - \hat{T}_{x^{1:N}, a} V \right\|_\infty. \quad (22)$$

On the other hand, $\|T_\pi V - V\|_\infty \leq \|T_\pi V - TV\|_\infty + \|TV - V\|_\infty$, and therefore by (21),

$$\|T_\pi^k V - T^k V\|_\infty \leq \frac{2\gamma}{1 - \gamma} \|TV - V\|_\infty + \left(\frac{\gamma}{1 - \gamma} + 1 \right) \|T_\pi V - TV\|_\infty$$

which combined with (22) yields

$$\|T_\pi^k V - T^k V\|_\infty \leq \frac{2}{1 - \gamma} \left(\gamma \|TV - V\|_\infty + \max_{a \in \mathcal{A}} \left\| T_a V - \hat{T}_{x^{1:N}, a} V \right\|_\infty \right).$$

Taking the limes superior of both sides when $k \rightarrow \infty$ yields the lemma. \square

Note that if $\hat{T}_{x^{1:N} a} = T_a$ then we get back the tight bounds of [14].⁸

The next lemma exploits that if $V_t = \hat{T}_{x^{1:N}}^t V_0$ for some $V_0 \in B_K(\mathcal{X})$ then the Bellman-error $\|TV_t - V_t\|_\infty$ can be related to the quality of approximation of T_a by $\hat{T}_{x^{1:N} a}$.

Lemma 5.12. *Let $K = K_r/(1 - \gamma)$, $\varepsilon > 0$ and let $V_0 \in B_K(\mathcal{X})$ fixed. Let*

$$t = t(\varepsilon, \gamma, K) = \left\lfloor \frac{\log(8K) + \log(1/(\varepsilon(1 - \gamma)))}{\log(1/\gamma)} \right\rfloor,$$

$x^{1:N} \in \mathcal{X}^N$, $V_t = \hat{T}_{x^{1:N}}^t V_0$ and assume that

$$\max_{a \in \mathcal{A}} \left\| T_a V_t - \hat{T}_{x^{1:N} a} V_t \right\|_\infty \leq \frac{\varepsilon(1 - \gamma)}{4(1 + \gamma)}. \quad (23)$$

Further, let $\pi : \mathcal{X} \rightarrow \mathcal{A}$ s.t. $\hat{T}_{x^{1:N} \pi} V_t = \hat{T}_{x^{1:N}} V_t$. Then π is ε -optimal, i.e., $\|V_\pi - V^*\|_\infty \leq \varepsilon$.

Proof. We use Lemma 5.11. Let $V = V_t$ and let us bound the Bellman-error $\|TV_t - V_t\|_\infty$ first:

$$\begin{aligned} \|TV_t - V_t\|_\infty &\leq \left\| TV_t - \hat{T}_{x^{1:N}} V_t \right\|_\infty \\ &+ \left\| \hat{T}_{x^{1:N}} V_t - V_t \right\|_\infty \leq \max_{a \in \mathcal{A}} \left\| T_a V_t - \hat{T}_{x^{1:N} a} V_t \right\|_\infty \\ &+ \left\| \hat{T}_{x^{1:N}}^{t+1} V_0 - \hat{T}_{x^{1:N}}^t V_0 \right\|_\infty. \end{aligned}$$

Since $\hat{T}_{x^{1:N}} \in \text{Lip}_\infty(\gamma)$, the second term is bounded by

$$\begin{aligned} \gamma^t \left\| \hat{T}_{x^{1:N}} V_0 - V_0 \right\|_\infty &\leq \\ \gamma^t \left(\left\| \hat{T}_{x^{1:N}} V_0 \right\|_\infty + \|V_0\|_\infty \right) &\leq 2K \gamma^t, \end{aligned}$$

where we have used that $\hat{T}_{x^{1:N}} : B_K(\mathcal{X}) \rightarrow B_K(\mathcal{X})$ and $V_0 \in B_K(\mathcal{X})$. Therefore, by Lemma 5.11 we have

$$\begin{aligned} \|V_\pi - V^*\|_\infty &\leq \frac{2(1 + \gamma)}{1 - \gamma} \max_{a \in \mathcal{A}} \left\| T_a V_t - \hat{T}_{x^{1:N} a} V_t \right\|_\infty \\ &+ \frac{4K \gamma^{t+1}}{1 - \gamma}. \end{aligned}$$

Using the definition of t and (23) we get

⁸Note that the lemma still holds if we replace the special operators $\hat{T}_{x^{1:N} a}$, $\hat{T}_{x^{1:N} \pi}$ and $\hat{T}_{x^{1:N}}$ by operators $\hat{T}_a, \hat{T}_\pi, \hat{T} \in \text{Lip}_\infty(\gamma)$ satisfying $(\hat{T}_\pi V)(x) = (\hat{T}_{\pi(x)} V)(x)$ and $(\hat{T} V)(x) = \max_{a \in \mathcal{A}} (\hat{T}_a V)(x)$.

$$\|V_\pi - V^*\|_\infty \leq \varepsilon,$$

proving the lemma. \square

Now, we are in the position to prove the first main result that was stated as Theorem 4.1 before:

Theorem 5.13. *Let $K = K_r/(1 - \gamma)$ and let $\varepsilon > 0$, $\delta > 0$, $V_0 \in B_K(\mathcal{X})$ be fixed. Let*

$$t = t(\varepsilon, \gamma, K)$$

and let

$$\begin{aligned} p_4(d, \varepsilon, \delta, K, K_p, L_p, |\mathcal{A}|, \gamma) &= \\ 512K^2 K_p^2 \left(\frac{24(K + 1)}{\varepsilon(1 - \gamma)^2} \right)^2 &\left(\log 8 + \log(t(\varepsilon, \gamma, K) + 1) \right) \\ + \log |\mathcal{A}| + d \log \left(\left\lfloor \frac{384(K + 1)^2 L_p d}{\varepsilon(1 - \gamma)^2} \right\rfloor + 1 \right) & \\ + \log \left(\frac{1}{\delta} \right). & \end{aligned} \quad (24)$$

Let $N \geq p_4(d, \varepsilon, \delta, K, K_p, L_p, |\mathcal{A}|, \gamma)$. Let $V = \hat{T}_N^t V_0$ and let the stationary policy π be defined by $\hat{T}_N \pi V = \hat{T}_N V$. Then

$$\mathbb{P}(\|V_\pi - V^*\|_\infty \leq \varepsilon) \geq 1 - \delta. \quad (25)$$

Proof. The proof combines Lemmas 5.12 and 5.10. Firstly, we bound

$$m = \max_{a \in \mathcal{A}} \left\| \hat{T}_N a \hat{T}_N^t V_0 - T_a \hat{T}_N^t V_0 \right\|_\infty.$$

Let $\tilde{\pi} : \mathcal{X} \rightarrow \mathcal{A}$ be defined by

$$\tilde{\pi}(x) = \operatorname{argmax}_{a \in \mathcal{A}} \left\| \hat{T}_N a \hat{T}_N^t V_0 - T_a \hat{T}_N^t V_0 \right\|_\infty$$

($\tilde{\pi}$ does not depend on x). Then

$$\begin{aligned}
m &= \left\| \hat{T}_{N\bar{\pi}} \hat{T}_N^t V_0 - T_{\bar{\pi}} \hat{T}_N^t V_0 \right\|_{\infty} \\
&\leq \left\| \hat{T}_{N\bar{\pi}} \hat{T}_N^t V_0 - \hat{T}_{N\bar{\pi}} T^t V_0 \right\|_{\infty} \\
&\quad + \left\| \hat{T}_{N\bar{\pi}} T^t V_0 - T_{\bar{\pi}} T^t V_0 \right\|_{\infty} \\
&\quad + \left\| T_{\bar{\pi}} T^t V_0 - T_{\bar{\pi}} \hat{T}_N^t V_0 \right\|_{\infty} \\
&\leq 2\gamma \left\| \hat{T}_N^t V_0 - T^t V_0 \right\|_{\infty} \\
&\quad + \max_{a \in \mathcal{A}} \left\| \hat{T}_{N a} T^t V_0 - T_a T^t V_0 \right\|_{\infty} \\
&\leq (2\gamma + 1) \\
&\quad \max \left\{ \max_{a \in \mathcal{A}} \left\| \hat{T}_{N a} T^t V_0 - T_a T^t V_0 \right\|_{\infty}, \right. \\
&\quad \left. \left\| \hat{T}_N^t V_0 - T^t V_0 \right\|_{\infty} \right\}.
\end{aligned}$$

Therefore if

$$N \geq p_3(d, \varepsilon(1 - \gamma)/(4(2\gamma + 1)(\gamma + 1)), \delta, K, K_p, L_p, t(\varepsilon, \gamma, K), |\mathcal{A}|, \gamma)$$

then by Lemma 5.10 and Lemma 5.12,

$$\|V_{\pi} - V^*\|_{\infty} \leq \varepsilon$$

with probability at least $1 - \delta$. \square

In order to finish the proof of the main theorem we will prove that in discounted problems stochastic policies that generate ε -optimal actions with high probability are uniformly good. This result appears in the context of finite models in [9]. For completeness, we present the proof here. We start with the definition of ε -optimal actions and then prove three simple lemmas.

Definition 5.14. Let $\varepsilon > 0$, and consider a discounted MDP $(\mathcal{X}, \mathcal{A}, p, r, \gamma)$. We call the set

$$\mathcal{A}_{\varepsilon}(x) = \{a \in \mathcal{A} : (T_a V^*)(x) \geq (TV^*)(x) - \varepsilon\}.$$

the set of ε -optimal actions. Elements of this set are called ε -optimal.

Lemma 5.15. Let $\pi : \mathcal{X} \times \mathcal{A} \rightarrow [0, 1]$ be a stationary stochastic policy that selects only ε -optimal actions: for all $x \in \mathcal{X}$ and $a \in \mathcal{A}$ from $\pi(x, a) > 0$ it follows that $a \in \mathcal{A}_{\varepsilon}(x)$. Then $\|V_{\pi} - V^*\|_{\infty} \leq \varepsilon/(1 - \gamma)$.

Proof. From the definition of π it is immediate that $\|T_{\pi} V^* - V^*\|_{\infty} \leq \varepsilon$. Clearly, $T_{\pi} V^* \leq V^*$ and

$$\begin{aligned}
(T_{\pi} V^*)(x) &= \sum_{a \in \mathcal{A}} \pi(x, a) (T_a V^*)(x) \\
&= \sum_{a \in \mathcal{A}_{\varepsilon}(x)} \pi(x, a) (T_a V^*)(x) \\
&\geq \sum_{a \in \mathcal{A}_{\varepsilon}(x)} \pi(x, a) ((TV^*)(x) - \varepsilon) \\
&= V^*(x) - \varepsilon.
\end{aligned}$$

Now, consider the telescoping sum

$$T_{\pi}^k V^* = T_{\pi} V^* + \sum_{i=1}^{k-1} (T_{\pi}^{i+1} V^* - T_{\pi}^i V^*).$$

Therefore,

$$\begin{aligned}
\|T_{\pi}^k V^* - V^*\|_{\infty} &\leq \\
&\leq \|T_{\pi} V^* - V^*\|_{\infty} + \sum_{i=1}^{k-1} \|T_{\pi}^{i+1} V^* - T_{\pi}^i V^*\|_{\infty} \\
&\leq \varepsilon + \frac{\gamma}{1 - \gamma} \varepsilon = \frac{\varepsilon}{1 - \gamma}.
\end{aligned}$$

\square

The next lemma will be applied to show if two policies are “close to each other” then so are their evaluation functions. Both the lemma and its proof are very similar to those of Proposition 5.8.

Lemma 5.16. Let B be a space of bounded functions⁹, $B_K \subseteq \{V \in B : \|V\|_{\infty} \leq K\}$. Assume that $T_1, T_2 : B_K \rightarrow B_K$ are such that for some $\alpha > 0$ $\|T_1 V - T_2 V\|_{\infty} \leq \alpha$ holds for all $V \in B_K$ and $T_1 \in \text{Lip}_{\infty}(\gamma)$ for some $0 \leq \gamma < 1$. Then $\|T_1^s V - T_2^s V\|_{\infty} \leq \alpha/(1 - \gamma)$. Further, let V_1^* be the fixed point of T_1 and V_2^* be the fixed point of T_2 . If $T_2 \in \text{Lip}_{\infty}(\gamma)$ then $\|V_1^* - V_2^*\|_{\infty} \leq \alpha/(1 - \gamma)$.

Proof. The proof is almost identical to that of Proposition 5.8. One proves by induction that $\|T_1^s V - T_2^s V\|_{\infty} \leq \alpha/(1 - \gamma)$ holds for all $s \geq 0$. Here $V \in B_K$ is fixed. Indeed, the inequality holds for $s = 0$. Assuming that it holds for $s - 1$ with $s \geq 1$ one gets

⁹Again, B could be any Banach-space.

$$\begin{aligned} \|T_1^s V - T_2^s V\|_\infty &\leq \|T_1 T_1^{s-1} V - T_1 T_2^{s-1} V\|_\infty \\ &+ \|T_1 T_2^{s-1} V - T_2 T_2^{s-1} V\|_\infty \leq \gamma\alpha/(1-\gamma) + \alpha \\ &= \alpha/(1-\gamma), \end{aligned}$$

showing the first part of the statement. The second part is proven by taking the limes superior of both sides when $s \rightarrow \infty$. \square

Now, we are ready to prove the lemma showing that policies that choose ε -optimal actions with high probability are uniformly good.

Lemma 5.17. *Let $\varepsilon > 0$, $1 > \delta > 0$ be given. Let $\pi : \mathcal{X} \times \mathcal{A} \rightarrow [0, 1]$ be a stochastic policy that selects ε -optimal actions with probability at least $1 - \delta$. Then $\|V_\pi - V^*\|_\infty \leq (\varepsilon + 2K\delta)/(1 - \gamma)$.*

Proof. Let $\delta(x) = \sum_{a \notin \mathcal{A}_\varepsilon(x)} \pi(x, a)$ denote the probability of selecting non- ε -optimal actions in state x ($x \in \mathcal{X}$). By assumption, $\delta(x) \leq \delta < 1$. Let $\pi' : \mathcal{X} \times \mathcal{A} \rightarrow [0, 1]$ be the policy defined by

$$\pi'(x, a) = \begin{cases} \frac{\pi(x, a)}{1 - \delta(x)}, & \text{if } a \in \mathcal{A}_\varepsilon(x), \\ 0, & \text{otherwise.} \end{cases}$$

We claim that T_π and $T_{\pi'}$ are ‘‘close’’ to each other. For, let $V \in B_K(\mathcal{X})$, where $K = K_r/(1 - \gamma)$.

$$\begin{aligned} (T_\pi V)(x) - (T_{\pi'} V)(x) &= \sum_{a \in \mathcal{A}} (\pi(x, a) - \pi'(x, a))(T_a V)(x) \end{aligned}$$

and since $\|T_a V\|_\infty \leq K$,

$$\|T_\pi V - T_{\pi'} V\|_\infty \leq K \sum_{a \in \mathcal{A}} |\pi(x, a) - \pi'(x, a)|.$$

Further,

$$\begin{aligned} &\sum_{a \in \mathcal{A}} |\pi(x, a) - \pi'(x, a)| \\ &= \sum_{a \in \mathcal{A}_\varepsilon(x)} |\pi(x, a) - \pi'(x, a)| \\ &\quad + \sum_{a \notin \mathcal{A}_\varepsilon(x)} |\pi(x, a) - \pi'(x, a)| \\ &= \sum_{a \in \mathcal{A}_\varepsilon(x)} \frac{\pi(x, a)}{1 - \delta(x)} - \pi(x, a) + \sum_{a \notin \mathcal{A}_\varepsilon(x)} \pi(x, a) \\ &= 2\delta(x) \leq 2\delta. \end{aligned}$$

Therefore, $\|T_\pi V - T_{\pi'} V\|_\infty \leq 2K\delta$. Since T_π and $T_{\pi'}$ and $B_K(\mathcal{X})$ satisfy the assumptions of

Lemma 5.16, and the fixed point of T_π and $T_{\pi'}$ are V_π and $V_{\pi'}$, respectively, we have

$$\|V_\pi - V_{\pi'}\|_\infty \leq 2K\delta/(1 - \gamma). \quad (26)$$

Further, by construction π' selects only ε -optimal actions and thus by Lemma 5.15,

$$\|V_{\pi'} - V^*\|_\infty \leq \varepsilon/(1 - \gamma).$$

Combining this with (26), we get that $\|V_\pi - V^*\|_\infty \leq (\varepsilon + 2K\delta)/(1 - \gamma)$, finishing the proof. \square

We are ready to prove the main result of the paper that was stated earlier as Theorem 4.2:

Theorem 5.18. *Let $K = K_r/(1 - \gamma)$ and let $\varepsilon > 0$. Fix some $V_0 \in B_K(\mathcal{X})$. Let $\varepsilon' = \varepsilon(1 - \gamma)/(2(1 + \gamma))$, $\delta = \varepsilon(1 - \gamma)/(4K)$ ($= \varepsilon(1 - \gamma)^2/(4K_r)$). Further, let*

$$t = t(\varepsilon', \gamma, K) = \left\lceil \frac{\log(32K/(\varepsilon(1 - \gamma)^2))}{\log(1/\gamma)} \right\rceil \quad (27)$$

and let

$$\begin{aligned} p_5(d, \varepsilon, K, K_p, L_p, |\mathcal{A}|, \gamma) &= \\ &512K^2 K_p^2 \left(\frac{96(K + 1)}{\varepsilon(1 - \gamma)^3} \right)^2 \left(\log 8 + \log(t + 1) \right. \\ &\quad \left. + \log |\mathcal{A}| + d \log \left(\left\lceil \frac{768(K + 1)^2 L_p d}{\varepsilon(1 - \gamma)^3} \right\rceil + 1 \right) \right. \\ &\quad \left. + \log \left(\frac{4K}{\varepsilon(1 - \gamma)} \right) \right). \end{aligned} \quad (28)$$

Choose

$$N \geq p_5(d, \varepsilon, K, K_p, L_p, |\mathcal{A}|, \gamma) \quad (29)$$

and let $V = \hat{T}_N^t V_0$. Further, let the stochastic stationary policy $\pi : \mathcal{X} \times \mathcal{A} \rightarrow [0, 1]$ be defined by

$$\pi(x, a) = \mathbb{P}(\pi_{X^{1:N}}(x) = a), \quad (30)$$

where $\pi_{X^{1:N}}$ is the policy defined by $\hat{T}_N^t \pi_{X^{1:N}} V = \hat{T}_N^t V$. Then π is ε -optimal and given a state x , a random action of π can be computed in time and space polynomial in $1/\varepsilon$, d , K , $\log L_p$, $|\mathcal{A}|$ and $1/(1 - \gamma)$.

Proof. The second part of the statement is immediate (cf. Remark 3.2). The bound on the time of computation is

$$O((t+1)N^2|\mathcal{A}|) \quad (31)$$

and the space requirement of the algorithm is¹⁰

$$O(N + |\mathcal{A}|) \quad (32)$$

For the first part, fix $X^{1:N}$. By Theorem 5.13, if $V = V_{\pi_{X^{1:N}}}$ then V satisfies $\mathbb{P}(\|V - V^*\|_\infty \leq \varepsilon') \geq 1 - \delta$. We claim that if ω is such that $\|V(\omega) - V^*\|_\infty \leq \varepsilon'$ then $\pi_{X^{1:N}}(\omega)(a) \in \mathcal{A}_{\varepsilon'(1+\gamma)}(x)$. Let us pick up such an ω , let $T_1 = T_{\pi_{X^{1:N}}(\omega)}$; note that $V = T_1 V$. Then $\|T_1 V^* - V^*\|_\infty \leq \|T_1 V^* - T_1 V\|_\infty + \|V - V^*\|_\infty \leq (1+\gamma)\varepsilon'$. Therefore, using the definition of $\mathcal{A}_\varepsilon(x)$, we get that $\pi_{X^{1:N}}(\omega)(a) \in \mathcal{A}_{\varepsilon'(1+\gamma)}(x)$. This shows that

$$\mathbb{P}(\pi_{X^{1:N}}(x) \in \mathcal{A}_{\varepsilon'(1+\gamma)}(x)) \geq 1 - \delta.$$

Now, by Lemma 5.17, the policy π defined by (30) is $((\varepsilon'(1+\gamma) + 2K\delta)/(1-\gamma))$ -optimal, i.e.,

$$\|V_\pi - V^*\|_\infty \leq (\varepsilon'(1+\gamma) + 2K\delta)/(1-\gamma).$$

Substituting the definitions of ε' and δ yields the result. \square

6. Conclusions and Further Work

In this article we have considered an on-line planning algorithm that was shown to avoid the curse of dimensionality. Bounds following from Rust's original result by Markov's inequality were improved on in several ways: our bounds depend poly-logarithmically on the Lipschitz constant of the transition probabilities, they do not depend on the Lipschitz constant of the immediate rewards (we dropped the assumption of having Lipschitz-continuous immediate reward functions), and the number of samples depends on the cardinality of the action set in a poly-logarithmic way, as well.

It is interesting to note that although our bounds depend poly-logarithmically on the Lipschitz constant of the transition probabilities (char-

acterizing how “fast” the dynamics is), they depend polynomially on the bound of the transition probabilities (characterizing the randomness of the MDP). Therefore, perhaps not surprisingly, for these kind of Monte-Carlo algorithms faster dynamics are easier to cope with than less random dynamics (with peaky transition probability functions).

As a consequence of our result, many interesting questions arise. For example, different variants of the proposed algorithm could be compared, such as multigrid versions, versions using quasi-random numbers, or versions that use importance sampling could be compared. In practice, one would probably choose not to recompute the cache C_ε for each query. Also, in practice, one would probably precompute the transition probability table $\hat{p}_{X^{1:N}}(X_i|X_j, a)$ and in order to speed-up the iterations one would probably eliminate the computation with those transition probability values that are very close to zero. This would considerably speed up the computations as one would expect that “distant” parts of the state space are “uncoupled”. However, the theoretical effect of these modifications needs to be explored.

Note that the Lipschitz condition on p can be replaced by an appropriate condition on the *metric-entropy* of $p(x|\cdot, a)$ and the proofs will still go through. Therefore the proofs can be extended to Hölder-classes of transition laws or local Lipschitz classes (e.g. $|p(x'|x_1, a) - p(x'|x_2, a)| \leq L(x', a) \|x_1 - x_2\|_1$) (in this case one would need to use bracketing numbers), smooth functions, Sobolev classes, etc.

One of the most interesting problems is to extend the results to infinite action spaces. For sure, such an extension needs some regularity assumptions on the dependence of the transition probability law and the reward function on the actions. It would also be interesting to prove analogous results for discrete MDPs having a factorized representation.

The presented algorithm may find applications in economic problems without any modifications [12]. We also work on applications on deterministic continuous state-space, finite-action space control problems and partially observable MDPs over discrete spaces. Also, a combination with look-a-head search can be interesting from the practical point of view.

The algorithm considered in the article was tried in practice on some standard problems (car-on-

¹⁰Assuming that only the normalization factors of the transition probabilities $\hat{p}_{X^{1:N}}$ are stored.

the-hill, acrobot) and it was observed to yield a reasonable performance even when the number of samples was kept quite small (in the range of a few hundred to few thousand samples). It was also observed that boundary effects can interfere negatively with the algorithm. Details of these experiments, however, will be described elsewhere.

References

- [1] R. Bellman. *Dynamic Programming*. Princeton University Press, Princeton, New Jersey, 1957.
- [2] D. P. Bertsekas. *Dynamic Programming: Deterministic and Stochastic Models*. Prentice-Hall, Englewood Cliffs, NJ, USA, 1989.
- [3] C.S. Chow and J.N. Tsitsiklis. The complexity of dynamic programming. *Journal of complexity*, 5:466–488, 1989.
- [4] 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.
- [5] L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Applications of Mathematics: Stochastic Modelling and Applied Probability. Springer-Verlag New York, 1996.
- [6] E.B. Dynkin and A.A. Yushkevich. *Controlled Markov Processes*. Springer-Verlag, Berlin, 1979.
- [7] G.S. Fishman. *Monte Carlo Concepts, Algorithms, and Applications*. Springer-Verlag, 1999.
- [8] M. Kearns, Y. Mansour, and A.Y. Ng. Approximate planning in large POMDPs via reusable trajectories. In S. A. Solla, T. K. Leen, and K. R. Müller, editors, *Advances in Neural Information Processing Systems 12*. MIT Press, Cambridge, MA, 1999. to appear.
- [9] 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*, 1999.
- [10] D. Pollard. *Convergence of Stochastic Processes*. Springer Verlag, New York, 1984.
- [11] M.L. Puterman. *Markov Decision Processes — Discrete Stochastic Dynamic Programming*. John Wiley & Sons, Inc., New York, NY, 1994.
- [12] J. Rust. Structural estimation of Markov decision processes. In *Handbook of Econometrics*, volume 4, chapter 51, pages 3082–3139. North Holland, 1994.
- [13] J. Rust. Using randomization to break the curse of dimensionality. *Econometrica*, 65:487–516, 1996.
- [14] R. J. Williams and L.C. Baird, III. Tight performance bounds on greedy policies based on imperfect value functions. In *Proceedings of the Tenth Yale Workshop on Adaptive and Learning Systems*, 1994.