ECOLE POLYTECHNIQUE

CENTRE DE MATHÉMATIQUES APPLIQUÉES
UMR CNRS 7641

91128 PALAISEAU CEDEX (FRANCE). Tél: 01 69 33 41 50. Fax: 01 69 33 30 11

http://www.cmap.polytechnique.fr/

Error Bounds for Approximate Value Iteration.

Rémi Munos

R.I. N^0 527 *April 2004*

Error Bounds for Approximate Value Iteration

Rémi Munos
Centre de Mathématiques Appliquées,
Ecole Polytechnique, 91128 Palaiseau, France.
Remi.Munos@polytechnique.fr

21st April 2004

Abstract

We study Approximate Value Iteration where value representations V_n are processed iteratively by $V_{n+1} = \mathcal{A}\mathcal{T}V_n$ where $\mathcal{T}$ is the Bellman operator and $\mathcal{A}$ an approximation operator. Bounds on the error between the performance of the policies induced by this algorithm and the optimal policy are given as a function of weighted L_1 or L_2-norms of the approximation errors. A Markov Decision Problem is thus reduced to successive resolutions of Supervised Learning problems.

1 Introduction

We study the resolution of *Markov Decision Problems* (MDPs) (Puterman, 1994) using approximate value function representations V_n. The **Approximate Value Iteration** (AVI) algorithm is defined by the iteration

$$V_{n+1} = \mathcal{A}\mathcal{T}V_n \tag{1}$$

where $\mathcal{T}$ is the *Bellman operator* and $\mathcal{A}$ an *approximation operator*, or equivalently a *supervised learning* (SL) algorithm. AVI is very popular and has been successfully implemented in many different settings in Dynamic Programming (DP) and Reinforcement Learning (RL) (Bertsekas & Tsitsiklis, 1996).

A simple version is: at stage n, select a sample of states $(x_k)_{k=1...K}$ from some distribution ρ_n, compute the backed-up values $\mathcal{T}V_n(x_k)$, then make a call to the SL algorithm. This returns a function V_{n+1} minimizing the average empirical loss $\frac{1}{K}\sum_k l(V_{n+1}(x_k) - \mathcal{T}V_n(x_k))$. Most SL algorithms use squared (L_2) or absolute (L_1) loss functions (or variants) thus perform a minimization problem in weighted semi-norm L_1 or L_2, where the weights are defined by ρ_n. It is therefore crucial to estimate the performance of AVI as a function of the weighted L_1 and L_2- norms of the SL approximation errors. The goal of this paper is to extend usual results in L_∞-norm to similar results in weighted L_1 and L_2- norms. The performance achieved by such a resolution of the MDP may then be directly related to the accuracy of the SL algorithm.

Alternative results in approximate DP with weighted norms include Linear Programming (de Farias & Roy, 2003) and Policy Iteration (Munos, 2003).

Let X be the state space (with $N < \infty$ states, although all results are extensible to the case $N = \infty$) and A the action space. Let $p(x, a, y)$ be the probability that the next state is y given that the current state is x and the action is a. Let $r(x, a, y)$ be the reward received when a transition $(x, a) \to y$ occurs.

A *policy* π is a mapping from X to A. We write P^π the $N \times N-$matrix with elements $P^\pi(x, y) = p(x, \pi(x), y)$ and r^π the vector with components $r^\pi(x) = \sum_y p(x, \pi(x), y) r(x, \pi(x), y)$.

For a policy π, we define the *value function* V^π which, in the discounted and infinite horizon case studied here, is the expected discounted sum of future rewards

$$V^\pi(x) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \, r(x_t, a_t, x_{t+1}) | x_0 = x, a_t = \pi(x_t) \right]$$

where $\gamma \in [0, 1)$ is a *discount factor*. V^π is the fixed point of the operator $\mathcal{T}^\pi : I\!R^N \to I\!R^N$ defined, for any vector $W \in I\!R^N$ by $\mathcal{T}^\pi W \stackrel{\text{def}}{=} r^\pi + \gamma P^\pi W$.

The *optimal value function* $V^* = \sup_\pi V^\pi$ is the fixed-point of the Bellman operator $\mathcal{T}$ defined, for any $W \in I\!R^N$, by

$$\mathcal{T} W(x) = \sup_{a \in A} \sum_y p(x, a, y)[r(x, a, y) + \gamma W(y)].$$

We say that a policy π is *greedy with respect to* $W \in I\!R^N$ if for all $x \in X$,

$$\pi(x) \in \arg\max_{a \in A} \sum_y p(x, a, y)[r(x, a, y) + \gamma W(y)].$$

An *optimal policy* π^* is a policy greedy w.r.t. V^*.

An exact resolution method for computing V^* is the *Value Iteration* (VI) algorithm defined by the iteration $V_{n+1} = \mathcal{T} V_n$. Due to the contraction property in $L_\infty-$norm of the operator $\mathcal{T}$, the iterates V_n converge to V^* as $n \to \infty$. However, problems with a large number of states prevent us from using such exact resolution methods; we need to represent the functions with a moderate number of coefficients and perform approximate iterations such as (1).

The paper is organized as follows. We first remind some approximation results in L_∞-norm, then give componentwise bounds and use them to derive error-bounds in L_1 and L_2-norms. Finally we detail some practical implementations. The main result of this paper is Theorem 2. All proofs are detailed in the Appendix.

We remind the definition of the norms: let $u \in I\!R^N$. Its L_∞-norm is $||u||_\infty \stackrel{\text{def}}{=} \sup_{x \in X} |u(x)|$. Let μ be a distribution on X. Its weighted L_1 and L_2-(semi)norms (noted $L_{1,\mu}$ and $L_{2,\mu}$) are $||u||_{1,\mu} \stackrel{\text{def}}{=} \sum_{x \in X} \mu(x)|u(x)|$ and $||u||_{2,\mu} \stackrel{\text{def}}{=} [\sum_{x \in X} \mu(x) u(x)^2]^{1/2}$. We note $||\cdot||_1$ or $||\cdot||_2$ the unweighted L_1 and L_2 norms (i.e. when μ is uniform).

2 Approximation results in L_∞-norm

The results in this section are given in (Bertsekas & Tsitsiklis, 1996). For consistency, the proofs are reminded in the Appendix. Consider the **AVI algorithm** defined by (1). This algorithm does not converge but its asymptotic behavior may be studied. If the approximation errors are bounded (in $L_\infty-$norm) $||V_{n+1} - \mathcal{A}\mathcal{T}V_n||_\infty \leq \varepsilon$ then, the limit superior of the error between the approximate and the optimal value functions satisfies

$$\overline{\lim_{n\to\infty}}||V^* - V_n||_\infty \leq \frac{1}{1-\gamma}\varepsilon. \tag{2}$$

A bound on the error between the asymptotic performance of a policy π_n greedy w.r.t. V_n and the optimal policy is

$$\overline{\lim_{n\to\infty}}||V^* - V^{\pi_n}||_\infty \leq \frac{2\gamma}{(1-\gamma)^2}\varepsilon. \tag{3}$$

This L_∞-bound requires a uniformly low approximation error over all states, which is difficult to get in practice (exceptions include (Gordon, 1995; Guestrin et al., 2001)), especially for large-scale problems. Most approximators such as those described in the next section perform a minimization problem using weighted L_1 and L_2 norms.

3 Approximation operators

A **supervised learning algorithm** returns a good fit g (within given classes of functions $\mathcal{F}$) of the data $(x_k, v_k) \in X \times I\!R$, $k = 1 \dots K$ (with the x_k sampled from some distribution μ and the values v_k being estimates of some function $f(x_k)$), by minimizing the average loss $\frac{1}{K}\sum_{k=1}^{K} l(v_k - g(x_k))$, mainly with L_1 or L_2 loss functions (or variants). Equivalently, $\mathcal{A}$ may be considered as an **approximation operator** that returns a compact representation $g \in \mathcal{F}$ of a general function f by minimizing some $L_{1,\mu}$ or $L_{2,\mu}$-norm. Approximation theory studies the approximation error as a function of the smoothness of f (DeVore, 1997).

The projection onto the span of a fixed family of functions (called *features*) is called *linear approximation* and include *Splines*, *Radial Basis, Fourier* or *Wavelet decomposition.* A better approximation is reached when choosing the features according to f (i.e. *feature selection*). This *non-linear approximation* is particularly efficient when f has piecewise regularities (e.g. in adaptive wavelet basis (Mallat, 1997) such functions are compactly represented with few non-zero coefficients). Greedy algorithms for selecting the best features among a given dictionary of functions include the *Matching Pursuit* and variants (Davies et al., 1997).

In Statistical Learning (Hastie et al., 2001), other SL algorithms are *Neural Network*, *Locally Weighted Learning* and *Kernel Regression* (Atkeson et al., 1997), *Support-Vectors* (SVs) and *Reproducing Kernels* (Vapnik et al., 1997).

We call $\mathcal{A}$ an $\varepsilon-$**approximation operator** if $\mathcal{A}$ returns an $\varepsilon-$approximation g of f: $||f-g|| \leq \varepsilon$.

4 Componentwise bounds

Here we provide several componentwise bounds that will be used in the next section. The next result provides two bounds on the error between the performance of a reference policy and a policy which is greedy w.r.t. a function V.

Lemma 1 *Let π_{ref} be a reference policy, $V \in \mathbb{R}^N$, and π a policy greedy w.r.t. V. Then*

$$V^{\pi_{\text{ref}}} - V^{\pi} \leq \gamma(I-\gamma P^{\pi_{\text{ref}}})^{-1}(P^{\pi_{\text{ref}}} - P^{\pi})(V^{\pi} - V), \quad (4)$$

$$V^{\pi_{\text{ref}}} - V^{\pi} \leq \gamma(I-\gamma P^{\pi})^{-1}(P^{\pi_{\text{ref}}} - P^{\pi})(V^{\pi_{\text{ref}}} - V). \quad (5)$$

Now consider the **AVI algorithm** defined by (1). Call $\varepsilon_n = \mathcal{T}V_n - V_{n+1}$ the **approximation error**. The error between the successive approximations V_n and the optimal value function V^* may be bounded by the approximation errors:

Lemma 2 *We have*

$$\gamma P^{\pi_n}(V^* - V_n) + \varepsilon_n \leq V^* - V_{n+1} \leq \gamma P^{\pi^*}(V^* - V_n) + \varepsilon_n.$$

Thus

$$V^* - V_n \leq \sum_{k=0}^{n-1} \gamma^{n-1-k}(P^{\pi^*})^{n-k-1}\varepsilon_k + \gamma^n (P^{\pi^*})^n (V^* - V_0),$$

$$V^* - V_n \geq \sum_{k=0}^{n-1} \gamma^{n-1-k}(\prod_{i=k+1}^{n-1} P^{\pi_i})\varepsilon_k + \gamma^n (\prod_{i=1}^{n} P^{\pi_i})(V^* - V_0).$$

As a consequence of Lemma 1 (inequality (5) applied to V_n and the choice of an optimal policy as the reference policy) we derive a bound on the error between the performance of a policy π_n greedy w.r.t. V_n and the optimal policy

$$V^* - V^{\pi_n} \leq \gamma(I-\gamma P^{\pi_n})^{-1}(P^{\pi^*} - P^{\pi_n})(V^* - V_n). \quad (6)$$

5 Approximation results in L_1 and L_2-norms

We use results of the previous section to extend the bound (3) in L_∞-norm to bounds in L_1 and L_2-norms. Let μ be a distribution on X (considered as a row vector). From Lemma 2 and (6) we deduce the result:

Theorem 1 *For all $0 \leq k \leq n$, define the stochastic matrices*

$$A_{n,k} = \frac{1-\gamma}{2}(I - \gamma P^{\pi_n})^{-1}\left[[(P^{\pi^*})^{n-k} + (\prod_{i=k+1}^{n} P^{\pi_i})\right].$$

Then $\mu_{n,k} \stackrel{\text{def}}{=} \mu A_{n,k}$ is a distribution on X, and for $i = 1$ or 2,

$$\overline{\lim_{n\to\infty}} ||V^* - V^{\pi_n}||^i_{i,\mu} \leq \frac{(2\gamma)^i}{(1-\gamma)^{2i-1}} \overline{\lim_{n\to\infty}} \sum_{k=0}^{n-1} \gamma^{n-1-k} ||\varepsilon_k||^i_{i,\mu_{n,k}} \tag{7}$$

This result extends the bound in $L_\infty-$norm (3). However, one might wonder how useful the bound (7) is, given that the distributions $\mu_{n,k}$ are unknown (since they make use of the a-priori unknown optimal policy). In the next section we detail some hypotheses on the structure of the MDP that will be used to prove Theorem 2 and which will fully justify the relevance of this bound.

5.1 Any practical use of these bounds?

At each iteration of the AVI algorithm the new function V_{n+1} is obtained by approximating $\mathcal{T}V_n$ via a call to a SL algorithm $\mathcal{A}$, which solves a minimization problem using some distribution ρ_n. The $\mu_{n,k}-$norm of ε_k in (7) may be bounded by its ρ_k-norm: $||\varepsilon_k||^i_{i,\mu_{n,k}} \leq \left\| \frac{\mu_{n,k}}{\rho_k} \right\|_\infty ||\varepsilon_k||^i_{i,\rho_k}$ where $\left\| \frac{\mu_{n,k}}{\rho_k} \right\|_\infty$ is the mismatch ratio between those distributions. In order to bound this ratio, we first require ρ_k to be lower bounded. This condition was already mentioned in (Koller & Parr, 2000; Kakade & Langford, 2002; Munos, 2003) to secure Policy Improvement steps in Approximate Policy Iteration. Here, we consider distributions of the form

$$\rho_n^\lambda = (1-\lambda)\mu + \lambda\rho_n \tag{8}$$

(for $0 \leq \lambda < 1$) where ρ_n is any distribution. A such example is the distribution $\rho_n^\lambda = (1-\lambda)\mu(I - \lambda P^{\pi_n})^{-1}$. This is the state visitation distribution of a Markov chain that starts with an initial state $x_0 \sim \mu$ and which at time t either follows policy π_n: $x_{t+1} \sim p(x_t, \pi_n(x_t), \cdot)$ with probability λ or restarts $x_{t+1} \sim \mu$ with probability $1-\lambda$. Note that when $\lambda \to 0$, ρ_n^λ tends to μ, and when $\lambda \to 1$, ρ_n^λ tends to the steady-state distribution for policy π_n.

A second requirement in order that the mismatch ratio be bounded is to upper bound $\mu_{n,k}$. In (Munos, 2003), we state a property (called *uniform stochasticity*) of the MDP under which we can prove that μ_R and $\mu_{n,k}$ are upper bounded. In practice though, this property does not cover a broad class of MDPs, and in particular never holds for deterministic MDPs. Here we investigate a much weaker assumption that holds for a large class of MDPs.

Hypothesis 1 *Given some distribution μ. For any $m \geq 0$ and any sequence of m policies $\pi_1, \pi_2, \ldots, \pi_m$, we have*

$$\mu P^{\pi_1} P^{\pi_2} \ldots P^{\pi_m} \leq h(m)\mu \tag{9}$$

where $h(m) \in \mathbb{R}$ is such that $\sum_{m\geq 0} \gamma^m h(m) \ll N$ for $N < \infty$. If $N = \infty$ this assumption reads that the series $\sum_{m\geq 0} \gamma^m h(m)$ converges (e.g. when h has at most an exponential growth (i.e. $h(m) = O(\beta^m)$) with some constant $\beta < 1/\gamma$).

Let us give some insight about this hypothesis. An example for which this hypothesis fails is an MDP where for a specific policy π all states jump to a given state -say state 1- with probability 1. Then if μ is an uniform distribution $\mu = (\frac{1}{N} \dots \frac{1}{N})$ we have $\mu(P^\pi)^m = (1\,0 \dots 0) \leq N\mu$. Thus $h(m) = N$ and $\sum_{m\geq 0} \gamma^m h(m) = \frac{N}{1-\gamma}$ which contradicts the hypothesis. This is really the worst case.

Intuitively this assumption means that for any sequence of policies $\pi_1, \dots, \pi_m$, the discounted state visitation distribution at any state y starting from $x_0 \sim \mu$ is upper bounded by a constant (much smaller than N) times $\mu(y)$:

$$\sum_{i=0}^{\infty} \gamma^m \Pr\left\{ x_m = y \middle| \begin{array}{l} x_0 \sim \mu, \\ x_i \sim p(x_{i-1}, \pi_i(x_{i-1}), \cdot) \end{array} \right\} \ll N\mu(y)$$

In short, this assumption prevents the mass of the future state visitation distribution to accumulate on few specific states. An important class of MDPs for which this assumption holds is defined in the following Lemma.

Lemma 3 *Assume that the MDP is embedded in $\mathbb{R}^d$, and that:*

1. *each state is represented by a point in $\mathbb{R}^d$,*
2. *the transitions are local (i.e. there exists $r > 0$ s.t. $p(x,a,y) > 0 \Rightarrow ||x - y|| \leq r$),*
3. *the density of points is bounded (i.e. there exists $\delta > 0$ s.t. for any volume $\mathcal{V} \in \mathbb{R}^d$, the number of points in $\mathcal{V}$ is at most $\delta|\mathcal{V}|$).*

Then Hypothesis 1 holds with the uniform distribution μ.

This Lemma encompasses a large class of MDPs among which those that result from a discretization of continuous (deterministic or stochastic) Markov processes (Kushner & Dupuis, 2001). Under this assumption, we now state the main result of this paper.

Theorem 2 *Let $i = 1$ or 2 and μ a distribution on X. Let $\mathcal{A}$ be an $\varepsilon-$approximation operator in L_{i,ρ_n^λ}-norm where the distribution ρ_n^λ is of the form (8) (i.e. for all $n \geq 0$, $||\varepsilon_n||_{i,\rho_n^\lambda} \leq \varepsilon$). Assume Hypothesis 1. Then*

$$\overline{\lim_{n\to\infty}} ||V^* - V^{\pi_n}||_{i,\mu} \leq \frac{2\gamma}{(1-\gamma)^2} \left(\frac{C}{1-\lambda} \right)^{1/i} \varepsilon \tag{10}$$

with the constant $C = h(1)[(1-\gamma)\sum_{m\geq 0} \gamma^m h(m)]^2$.

5.2 Illustration on the Chain Walk MDP

We illustrate the fact that the L_1 and L_2-norm bounds given in Theorem 2 may be much tighter than the $L_\infty-$norm (3) on the Chain Walk MDP defined in (Lagoudakis & Parr, 2003). This is a linear chain with N states with two dead-end states: states 1 and N. On each of the interior states $2 \leq x \leq N-1$ there are two possible actions: right or left, which changes the state in the intended direction with probability 0.9, and fails with probability 0.1 changing the state in the opposite direction. The reward simply depends on the current state and is 1 at boundary states and 0 elsewhere: $r = (1\,0 \ldots 0\,1)'$.

Consider approximation of the value function of the form $V_n(x) = \alpha_n + \beta_n x$ where $x \in \{1, \ldots, N\}$ is the state number. Assume that initial approximation is zero: $V_0 = (0 \ldots 0)'$. Then $\mathcal{T}V_0 = (1\,0 \ldots 0\,1)'$. The best fit in L_∞-norm is a constant function $V_1 = (\frac{1}{2} \ldots \frac{1}{2})'$ which produces an error $||V_1 - \mathcal{T}V_0||_\infty = \frac{1}{2}$.

Let us choose uniform distributions $\mu = \rho_n = (\frac{1}{N} \ldots \frac{1}{N})$ (i.e. $\lambda = 0$). In L_1-norm we find that the best fit is $V_1 = (0 \ldots 0)'$ (for $N > 4$) and the resulting error is $||V_1 - \mathcal{T}V_0||_1 = \frac{2}{N}$. In L_2-norm the best fit is also constant $V_1 = (\frac{2}{N} \ldots \frac{2}{N})'$ and the error is $||V_1 - \mathcal{T}V_0||_2 = \frac{\sqrt{2N-4}}{N}$.

In the three cases, by induction we see that the successive approximations V_n are constant, thus $\mathcal{T}V_n = r + \gamma V_n$ and the approximation errors remain the same as in the first iteration: $||V_{n+1} - \mathcal{T}V_n||_\infty = \frac{1}{2}$, $||V_{n+1} - \mathcal{T}V_n||_1 = \frac{2}{N}$, and $||V_{n+1} - \mathcal{T}V_n||_2 = \frac{\sqrt{2N-4}}{N}$. Since V_n is constant any policy π_n is greedy w.r.t. V_n. Hence for $\pi_n = \pi^*$ the l.h.s. of (3) and (10) are equal to zero. Now in order to compare the r.h.s. of these inequalities we need to estimate the constant C. The worst case in (9) is obtained when the mass of the state visitation distribution is mostly concentrated on one boundary state -say state 1- which corresponds to a policy π_{Left} that chooses everywhere action left. We see that $\mu(P^{\pi_{\text{Left}}})^m(x) \leq \mu(P^{\pi_{\text{Left}}})^m(1) \leq (1 + 0.9m)\mu(x)$ for all $x \geq 0$. Since $1 + 0.9m$ has sub-exponential growth, Hypothesis 1 is satisfied with $h(m) = 1 + 0.9m$. Note that this result is also a consequence of Lemma 3 since this Chain Walk is a typical example of an MDP embedded in $\mathbb{R}$. By noting that $\sum_{m \geq 0} m\gamma^m = \frac{\gamma}{(1-\gamma)^2}$, we derive the constant $C = 1.9[1 + 0.9\frac{\gamma}{(1-\gamma)}]^2$ which is independent from N. We deduce that when the number of states N is large, the L_1-norm bound gives an approximation of order $O(N^{-1})$, the L_2-norm bound is of order $O(N^{-1/2})$, whereas the $L_\infty-$norm bound (3) bound is only of order $O(1)$.

6 Practical algorithms

6.1 Model-based AVI

Let μ be a distribution on X. Given $\varepsilon > 0$ and an $\varepsilon-$approximation operator $\mathcal{A}$ in L_{i,ρ_n^λ}-norm (for $i = 1$ or 2). Assume Hypothesis 1. Successive iterations:

1. Select set of states $x_k \in X$, $k = 1 \ldots K$, sampled from the distribution $\rho_n^\lambda = (1 - \lambda)\mu + \lambda\rho_n$,

2. Compute the backed-up values $v_k = \mathcal{T}V_n(x_k)$,

3. Make a call to the supervised learning algorithm $\mathcal{A}$ with the data $\{x_k; v_k\}$, which returns an $\varepsilon-$approximation V_{n+1},

computes, after enough iterations, approximations V such that the error between the performance of a policy π greedy w.r.t. V and the optimal policy satisfies

$$||V^* - V^\pi||_{i,\mu} \leq \frac{2\gamma}{(1-\gamma)^2} \left(\frac{C}{1-\lambda}\right)^{1/i} \varepsilon. \tag{11}$$

6.2 Reinforcement Learning

Step 2 in the preceding algorithm requires the knowledge of a model of the transition probabilities (as well as a way to compute the expectations in operator $\mathcal{T}$). If this is not the case one may consider using a Reinforcement Learning (RL) algorithm (Sutton & Barto, 1998). Let us introduce the Q-values and the operator $\mathcal{R}$ defined on functions of $X \times A$,

$$\mathcal{R}Q(x,a) \stackrel{\text{def}}{=} \sum_{y \in X} p(x,a,y)\big[r(x,a,y) + \gamma \max_{b \in A} Q(y,b)\big].$$

The AVI algorithm is equivalent to defining successive approximations Q_n with iteration

$$Q_{n+1} = \mathcal{A}\mathcal{R}Q_n$$

where $\mathcal{A}$ is an approximation operator on $X \times A$. Thus, a model-free RL algorithm would be defined by the iteration:

1. Observe a set of transitions: $(x_k, a_k) \stackrel{r_k}{\to} y_k$, $k = 1 \ldots K$, where for current state x_k and action a_k, $y_k \sim p(x_k, a_k, \cdot)$ is the next observed state and r_k the received reward,

2. Compute the values $v_k = r_k + \gamma \max_b Q_n(y_k, b)$,

3. Make a call to the supervised learning algorithm $\mathcal{A}$ with the data $\{(x_k, a_k); v_k\}$, which returns an $\varepsilon-$approximation estimate $\widehat{Q_{n+1}}$,

An interesting case is when $\mathcal{A}$ is a linear operator *in the values* $\{v_k\}$ (which implies that the operators $\mathcal{A}$ and $\mathbb{E}$ commute) such as in Least Squares Regression, k-Nearest Neighbors, Locally Weighted Learning. Then the approximation $\widehat{Q_{n+1}}$ returned by $\mathcal{A}$ is an unbiased estimate of $\mathcal{A}\mathcal{R}Q_n$ (since the values $\{v_k\}$ are unbiased estimates of $\mathcal{R}Q_n(x_k, a_k)$). Thus when K is large such an iteration acts like a (model-based) AVI iteration, and bounds similar to those in Theorem 2 may be derived.

7 Conclusion

Theorem 2 provides a useful tool to relate the performance of AVI to the approximation power of the SL algorithm. Expressing the performance of AVI in the same norm as that used by the supervised learner to minimize the approximation error guarantees the tightness of this bound.

Extension to other loss functions l, such as ϵ-*insensitive* (used in SVs) or *Hubert's loss function* (for robust regression) is straightforward (Lemma 5 in the Appendix and thereby all other results extend to any semi-norm $L_{l,\mu}$ of the form $||u||_{l,\mu} = \sum_{x \in X} \mu(x) l(u(x))$ with an increasing an convex loss function l on $\mathbb{R}^+$).

Other possible extensions include Markov games and on-line RL.

A Appendix: proof of the results

Proof 1 (Proofs of the results in L_∞-norm) *Since π is greedy w.r.t. V (i.e. $\mathcal{T}^\pi V \geq \mathcal{T}^{\pi^*} V$), we have*

$$\begin{aligned} V^* - V^\pi &= V^* - \mathcal{T}^{\pi^*} V + \mathcal{T}^{\pi^*} V - \mathcal{T}^\pi V + \mathcal{T}^\pi V - \mathcal{T}^\pi V^\pi \\ &\leq V^* - \mathcal{T}^{\pi^*} V + \mathcal{T}^\pi V - \mathcal{T}^\pi V^\pi. \end{aligned}$$

Thus, in norm:

$$\begin{aligned} ||V^* - V^\pi||_\infty &\leq ||V^* - \mathcal{T}^{\pi^*} V||_\infty + ||\mathcal{T}^\pi V - \mathcal{T}^\pi V^\pi||_\infty \\ &\leq 2\gamma ||V^* - V||_\infty + \gamma ||V^* - V^\pi||_\infty \\ &\leq \frac{2\gamma}{1-\gamma} ||V^* - V||_\infty. \end{aligned} \tag{12}$$

From the fact that

$$\begin{aligned} ||V^* - V_{n+1}||_\infty &\leq ||\mathcal{T} V^* - \mathcal{T} V_n||_\infty + \varepsilon \\ &\leq \gamma ||V^* - V_n||_\infty + \varepsilon, \end{aligned}$$

we deduce (2) by taking the limit superior. Now, (12) applied to V_n and combined with (2) implies (3).

Now let us give two preliminary results.

Lemma 4 *Let A be an invertible matrix such that all the elements of its inverse are positive. Then the solutions to the inequality $Au \leq b$ are also solutions to $u \leq A^{-1}b$.*

Proof of Lemma 4. Let u be a solution to $Au \leq b$. This means that there exists a vector c with positive components s.t. $Au = b - c$, thus $u = A^{-1}b - A^{-1}c$. Since all components of $A^{-1}c$ are positive, we deduce that $u \leq A^{-1}b$. □

Lemma 5 *Let μ be a distribution on X, $\{S_k\}_{1\leq k\leq K}$ stochastic matrices, $\{\alpha_k\}_{1\leq k\leq K}$ positive numbers that sum to one, and u and $\{v_k\}_{1\leq k\leq K}$ (column) vectors that satisfy (componentwise) $0 \leq u \leq \sum_k \alpha_k S_k v_k$. Then $\mu_k \stackrel{\text{def}}{=} \mu S_k$ is a distribution on X and we have $||u||_{1,\mu} \leq \sum_k \alpha_k ||v_k||_{1,\mu_k}$ and $||u||_{2,\mu}^2 \leq \sum_k \alpha_k ||v_k||_{2,\mu_k}^2$.*

Proof of Lemma 5. We have

$$||u||_{1,\mu} = \mu u \leq \sum_k \alpha_k \mu S_k v_k = \sum_k \alpha_k ||v_k||_{1,\mu_k}.$$

Using two times Jensen's inequality (since the α_k sum to one and the S_k are stochastic matrices, respectively),

$$\begin{aligned} ||u||_{2,\mu}^2 &= \mu u^2 \leq \mu\big(\sum_k \alpha_k S_k v_k\big)^2 \leq \mu \sum_k \alpha_k \big(S_k v_k\big)^2 \\ &\leq \mu \sum_k \alpha_k S_k v_k^2 = \sum_k \alpha_k \mu_k v_k^2 = \sum_k \alpha_k ||v_k||_{2,\mu_k}^2 \,\square \end{aligned}$$

Proof of Lemma 1. Since $\mathcal{T}V \geq \mathcal{T}^{\pi_{\text{ref}}} V$, we have

$$\begin{aligned} V^{\pi_{\text{ref}}} - V^{\pi} &= \mathcal{T}^{\pi_{\text{ref}}} V^{\pi_{\text{ref}}} - \mathcal{T}^{\pi_{\text{ref}}} V + \mathcal{T}^{\pi_{\text{ref}}} V - \mathcal{T}V + \mathcal{T}V - \mathcal{T}^{\pi} V^{\pi} \\ &\leq \gamma P^{\pi_{\text{ref}}}(V^{\pi_{\text{ref}}} - V^{\pi} + V^{\pi} - V) + \gamma P^{\pi}(V - V^{\pi}). \end{aligned}$$

Thus

$$(I - \gamma P^{\pi_{\text{ref}}})(V^{\pi_{\text{ref}}} - V^{\pi}) \leq \gamma (P^{\pi_{\text{ref}}} - P^{\pi})(V^{\pi} - V)$$

and inequality (4) follows from Lemma 4. Similarly,

$$V^{\pi_{\text{ref}}} - V^{\pi} \leq \gamma P^{\pi_{\text{ref}}}(V^{\pi_{\text{ref}}} - V) + \gamma P^{\pi}(V - V^{\pi_{\text{ref}}} + V^{\pi_{\text{ref}}} - V^{\pi}),$$

thus

$$(I - \gamma P^{\pi})(V^{\pi_{\text{ref}}} - V^{\pi}) \leq \gamma (P^{\pi_{\text{ref}}} - P^{\pi})(V^{\pi_{\text{ref}}} - V),$$

which proves (5). $\square$

Proof of Lemma 2. Since $\mathcal{T}V_n \geq \mathcal{T}^{\pi^*} V_n$ and $\mathcal{T}V^* \geq \mathcal{T}^{\pi_n} V^*$, we have

$$\begin{aligned} V^* - V_{n+1} &= \mathcal{T}^{\pi^*} V^* - \mathcal{T}^{\pi^*} V_n + \mathcal{T}^{\pi^*} V_n - \mathcal{T}V_n + \varepsilon_n \\ &\leq \gamma P^{\pi^*}(V^* - V_n) + \varepsilon_n \\ V^* - V_{n+1} &= \mathcal{T}V^* - \mathcal{T}^{\pi_n} V^* + \mathcal{T}^{\pi_n} V^* - \mathcal{T}V_n + \varepsilon_n \\ &\geq \gamma P^{\pi_n}(V^* - V_n) + \varepsilon_n \end{aligned}$$

The other inequalities follow by induction. $\square$

Proof of Theorem 1. From Lemma 2 and by taking the absolute value in (6) we deduce that

$$0 \leq V^* - V^{\pi_n} \leq \sum_{k=0}^{n-1} \alpha_k A_{n,k} \frac{2\gamma(1-\gamma^n)}{(1-\gamma)^2} |\varepsilon_k| + O(\gamma^{n+1})$$

with $\alpha_k = \frac{1-\gamma}{1-\gamma^n}\gamma^{n-k-1}$ and we apply Lemma 5. Thus

$$||V^* - V^{\pi_n}||_{1,\mu} \leq \frac{2\gamma}{(1-\gamma)} \sum_{k=0}^{n-1} \gamma^{n-k-1} ||\varepsilon_k||_{1,\mu_{n,k}} + O(\gamma^{n+1})$$

and by taking the limit superior, (7) follows for $i = 1$. Similarly in L_2-norm,

$$\begin{aligned} ||V^* - V^{\pi_n}||_{2,\mu}^2 &\leq \left(\frac{2\gamma(1-\gamma^n)}{(1-\gamma)^2}\right)^2 \times \\ &\quad \sum_{k=0}^{n-1} \frac{1-\gamma}{1-\gamma^n} \gamma^{n-k-1} ||\varepsilon_k||_{1,\mu_{n,k}}^2 + O(\gamma^{n+1}) \end{aligned}$$

and (7) follows for $i = 2$. □

In order to prove Theorem 2, we first show that the distribution $\mu_{n,k}$ is upper bounded by a constant times μ.

Lemma 6 *If Hypothesis 1 holds for the distribution μ, then for all $0 \leq k \leq n$, $\mu_{n,k} \leq C_{n-k-1}\mu$, with the constant $C_{n-k-1} = h(1)h(n-k-1)(1-\gamma)\sum_{m=0}^{\infty}\gamma^m h(m)$.*

Proof of Lemma 6. From the definition of $\mu_{n,k}$,

$$\begin{aligned} \mu_{n,k} &= \mu A_{n,k} \leq \frac{1-\gamma}{2}\mu\big[\sum_{m=0}^{\infty}\gamma^m h(m)\big] \\ &\quad \big[P^{\pi^*}(P^{\pi^*})^{n-k-1} + P^{\pi_n}\prod_{i=k+1}^{n-1} P^{\pi_i}\big] \\ &\leq (1-\gamma)\big[\sum_{m=0}^{\infty}\gamma^m h(m)\big] h(1)h(n-k-1)\mu \ \square \end{aligned}$$

Proof of Theorem 2. Now, from the definition of ρ_n^λ we have $\left\|\frac{\mu_{n,k}}{\rho_k^\lambda}\right\|_\infty \leq \frac{C_{n-k-1}}{1-\lambda}$, thus $||\varepsilon_k||_{i,\mu_{n,k}}^i \leq \frac{C_{n-k-1}}{1-\lambda}\varepsilon^i$. For $i = 1$, we deduce from (7) that

$$\overline{\lim_{n\to\infty}}||V^* - V^{\pi_n}||_{1,\mu} \leq \frac{2\gamma}{(1-\gamma)} \sum_{m\geq 0} \gamma^m \frac{C_m}{1-\lambda}\varepsilon$$

and (10) follows. Similarly, for $i = 2$, from (7),

$$\begin{aligned} \overline{\lim_{n\to\infty}}||V^* - V^{\pi_n}||_{2,\mu}^2 &\leq \frac{(2\gamma)^2}{(1-\gamma)^3} \sum_{m\geq 0} \gamma^m \frac{C_m}{1-\lambda}\varepsilon^2 \\ &\leq \frac{(2\gamma)^2}{(1-\gamma)^4}\frac{C}{1-\lambda}\varepsilon^2 \square \end{aligned}$$

Proof of Lemma 3. Let $y \in X$. The number of states x that have a positive probability of reaching y in m steps, using any sequence of m policies

$\pi_1 \pi_2 \ldots \pi_m$, is less than the number of states in the ball of radius mr centered on y. Since a ball of radius mr has a volume $k(mr)^d$ where k is a constant that depend on the dimension of the space d, the sum over all states x of all $m-$steps probabilities $P^{\pi_1} P^{\pi_2} \ldots P^{\pi_m}(x, y)$ reaching y is bounded by

$$\sum_x P^{\pi_1} P^{\pi_2} \ldots P^{\pi_m}(x, y) \leq k(mr)^d \delta$$

Thus (9) holds with a uniform distribution μ and $h(m) = k(mr)^d \delta$. Since $h(m)$ is polynomial in m it has sub-exponential growth, and Hypothesis 1 holds. □

References

Atkeson, C. G., Schaal, S. A., & Moore, A. W. (1997). Locally weighted learning. *AI Review, 11.*

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

Davies, G., Mallat, S., & Avellaneda, M. (1997). Adaptive greedy approximations. *J. of Constr. Approx., 13,* 57–98.

de Farias, D., & Roy, B. V. (2003). The linear programming approach to approximate dynamic programming. *Operations Research, 51.*

DeVore, R. (1997). *Nonlinear approximation.* Acta Numerica.

Gordon, G. (1995). Stable function approximation in dynamic programming. *Proceedings of the International Conference on Machine Learning.*

Guestrin, C., Koller, D., & Parr, R. (2001). Max-norm projections for factored mdps. *Proceedings of the International Joint Conference on Artificial Intelligence.*

Hastie, T., Tibshirani, R., & Friedman, J. (2001). *The elements of statistical learning.* Springer Series in Statistics.

Kakade, S., & Langford, J. (2002). Approximately optimal approximate reinforcement learning. *Proceedings of the 19th International Conference on Machine Learning.*

Koller, D., & Parr, R. (2000). Policy iteration for factored mdps. *Proceedings of the 16th conference on Uncertainty in Artificial Intelligence.*

Kushner, H. J., & Dupuis, P. (2001). *Numerical methods for stochastic control problems in continuous time. second edition.* Applications of Mathematics. Springer.

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

Mallat, S. (1997). *A wavelet tour of signal processing.* Academic Press.

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

Puterman, M. L. (1994). *Markov decision processes, discrete stochastic dynamic programming.* A Wiley-Interscience Publication.

Sutton, R. S., & Barto, A. G. (1998). Reinforcement learning: An introduction. *Bradford Book.*

Vapnik, V., Golowich, S. E., & Smola, A. (1997). Support vector method for function approximation, regression estimation and signal processing. *In Advances in Neural Information Processing Systems*, 281–287.