On the Use of Non-Stationary Policies for Infinite-Horizon Discounted Markov Decision Processes

Bruno Scherrer

On the Use of Non-Stationary Policies for Infinite-Horizon Discounted Markov Decision Processes

Bruno Scherrer

March 25, 2012

Abstract

We consider infinite-horizon discounted Markov Decision Processes, for which it is known that there exists a *stationary* optimal policy. We consider the algorithm Value Iteration and the sequence of policies $\pi_1, \ldots, \pi_k$ it generates until some iteration k. We provide performance bounds for *non-stationary* policies involving the last m generated policies that reduce the state-of-the-art bound for the last *stationary* policy π_k by a factor $\frac{1-\gamma}{1-\gamma^m}$. In other words, and contrary to a common intuition, we show that it may be much easier to find a *non-stationary* approximately-optimal policy than a *stationary* one.

Suppose on runs an approximate version on Value Iteration, that is one builds a sequence of value-policy pairs as follows:

$$\begin{aligned} &\text{Pick any } \pi_{k+1} \text{ in } \mathcal{G}v_k \\ &v_{k+1} = T_{\pi_{k+1}} v_k + \epsilon_{k+1} \end{aligned}$$

where v_0 is arbitrary, $\mathcal{G}v_k$ is the set of policies that are greedy[1] with respect to v_k, and T_{π_k} is the linear Bellman operator associated to policy π_k. Let ϵ be a uniform bound on the norm of the errors $\|\epsilon_k\|_\infty$. A standard result (see for instance [1]) is the following performance guarantee:

Theorem 1. *The loss of policy π_k is bounded as follows:*

$$\|v_* - v_{\pi_k}\|_\infty \leq \frac{2(\gamma - \gamma^k)}{(1-\gamma)^2}\epsilon + \frac{2\gamma^k}{1-\gamma}\|v_* - v_0\|_\infty. \tag{1}$$

To our knowledge, there does not exist any example in the literature to support the tightness of this bound. It is, indeed, tight in the following sense:

Proposition 1. *For all k, there exists an MDP, an initial value v_0, a sequence of noise terms (ϵ_j) with $\|\epsilon_j\| \leq \epsilon$, such that running Value Iteration during k iterations with errors ϵ_k outputs a value function v_k of which a greedy policy satisfies Equation* (1) *with equality.*

Proof. Following Example 6.2 in [1], consider the deterministic MDP made of two states $\{s, s'\}$. s' is a terminal state (absorbing with 0 reward). The only choice is in s: either to stay (with reward $-\frac{2(\gamma-\gamma^k)}{1-\gamma}\epsilon$) or to switch to s' (with reward 0). There are two policies: the optimal policy π_* with value $v_* = (0\ ,\ 0)'$, and the non-optimal policy π_- with value $v_- = \left(-\frac{2(\gamma-\gamma^k)}{(1-\gamma)^2}\epsilon\ ,\ 0\right)'$. Consider the constant noise: $\epsilon_j = (\epsilon\ ,\ -\epsilon)'$. Initialize $v_0 = v_* = (0\ ,\ 0)$. By induction, it can be seen that for all $j \in \{1, ..., k-1\}$,

$$\begin{aligned} \mathcal{G}v_j &= \{\pi_*\} \\ \text{and} \qquad v_j &= \frac{(1-\gamma^j)}{1-\gamma}(\epsilon, -\epsilon)'. \end{aligned}$$

[1] There may be several greedy policies with respect to some value v, and what we write here holds whichever one is picked.

One can then observe that both policies are greedy with respect to v_{k-1}, so the bound of Equation (1) holds with equality for π_-. □

Remark 1. *The bound of Equation* (1) *tends to* $\frac{2\gamma}{(1-\gamma)^2}\epsilon$ *when* k *tends to* ∞. *This bound may be really bad when* γ *is close to 1. Moreover, compared to a value iteration algorithm for evaluating one single policy, and for which one can prove a dependency of the form* $\frac{1}{1-\gamma}\epsilon$, *there is an extra* $\frac{2\gamma}{1-\gamma}$ *that can significantly worsen the bound.*

Instead of running the last *stationary* policy π_k, one may consider running a periodic *non-stationary* policy, which is made of the last m policies. The following theorem shows that it is indeed a good idea.

Theorem 2. *Let* $\pi_{k,m}$ *be the following policy*

$$\pi_{k,m} = \pi_k \ \pi_{k-1} \ \cdots \ \pi_{k-m+1} \ \pi_k \ \pi_{k-1} \ \cdots .$$

Then its performance loss is bounded as follows:

$$\|v_* - v_{\pi_{k,m}}\|_\infty \leq \frac{2(\gamma - \gamma^k)}{(1-\gamma)(1-\gamma^m)}\epsilon + \frac{2\gamma^k}{1-\gamma^m}\|v_* - v_0\|_\infty.$$

Remark 2. *When* $m = 1$, *one recovers the standard result. For general* m, *this new bound is a factor* $\frac{1-\gamma}{1-\gamma^m}$ *better than the usual bound. Taking* $m = k$, *that is considering* all *the policies generated from the very start, one obtains the following bound:*

$$\|v_* - v_{\pi_{k,k}}\|_\infty \leq 2\left(\frac{\gamma}{1-\gamma} - \frac{\gamma^k}{1-\gamma^k}\right)\epsilon + \frac{2\gamma^k}{1-\gamma^k}\|v_* - v_0\|_\infty.$$

that tends to $\frac{2\gamma}{1-\gamma}\epsilon$ *when* k *tends to* ∞.

Remark 3. *From a bibliographical point of view, the idea of using non-stationary policies to improve error bounds already appears in [2]. However, in these works, the author considers* undiscounted finite-horizon *problems where the policy to be computed is naturally non-stationary. The fact that non-stationary policies (that loop over the last* m *computed policies) might also be useful in an infinite horizon context is to our knowledge new.*

Proof. The value of $\pi_{k,m}$ satisfies:

$$v_{\pi_{k,m}} = T_{\pi_k} T_{\pi_{k-1}} \cdots T_{\pi_{k-m+1}} v_{\pi_{k,m}}.$$

By induction, it can be shown that the sequence of values generated by the algorithm satisfies:

$$v_k = T_{\pi_k} T_{\pi_{k-1}} \cdots T_{\pi_{k-m+1}} v_{k-m} + \sum_{i=0}^{m-1} \Gamma_{k,i}\epsilon_{k-i}$$

where

$$\Gamma_{k,i} = P_{\pi_k} P_{\pi_{k-1}} \cdots P_{\pi_{k-i+1}}$$

in which, for all π, P_π denotes the stochastic matrix associated to policy π. By substracting the two equations, one obtains:

$$v_k - v_{\pi_{k,m}} = \Gamma_{k,m}(v_{k-m} - v_{\pi_{k,m}}) + \sum_{i=0}^{m-1} \Gamma_{k,i}\epsilon_{k-i}$$

and by taking the norm

$$\|v_k - v_{\pi_{k,m}}\|_\infty = \gamma^m \|v_{k-m} - v_{\pi_{k,m}}\|_\infty + \frac{1-\gamma^m}{1-\gamma}\epsilon. \tag{2}$$

Intuitively, Equation (2) shows that for sufficiently big m, v_k is a good approximation of the value of the non-stationary policy $\pi_{k,m}$ (whereas in general, it may be a poor approximation of the value of the stationary policy π_k).

By induction, it can also be proved that

$$\|v_* - v_k\|_\infty \leq \gamma^k \|v_* - v_0\|_\infty + \frac{1-\gamma^k}{1-\gamma}\epsilon. \tag{3}$$

Using Equations (2) and (3), we can conclude by observing that

$$\begin{aligned}
\|v_* - v_{\pi_{k,m}}\|_\infty &\leq \|v_* - v_k\|_\infty + \|v_k - v_{\pi_{k,m}}\|_\infty \\
&\leq \gamma^k \|v_* - v_0\|_\infty + \frac{1-\gamma^k}{1-\gamma}\epsilon + \gamma^m \|v_{k-m} - v_{\pi_{k,m}}\|_\infty + \frac{1-\gamma^m}{1-\gamma}\epsilon \\
&\leq \gamma^k \|v_* - v_0\|_\infty + \frac{1-\gamma^k}{1-\gamma}\epsilon + \gamma^m \left(\|v_{k-m} - v_*\|_\infty + \|v_* - v_{\pi_{k,m}}\|_\infty\right) + \frac{1-\gamma^m}{1-\gamma}\epsilon \\
&\leq \gamma^k \|v_* - v_0\|_\infty + \frac{1-\gamma^k}{1-\gamma}\epsilon + \gamma^m \left(\gamma^{k-m}\|v_* - v_0\|_\infty + \frac{1-\gamma^{k-m}}{1-\gamma}\epsilon + \|v_* - v_{\pi_{k,m}}\|_\infty\right) + \frac{1-\gamma^m}{1-\gamma}\epsilon \\
&= \gamma^m \|v_* - v_{\pi_{k,m}}\|_\infty + 2\gamma^k \|v_* - v_0\|_\infty + \frac{2(1-\gamma^k)}{1-\gamma}\epsilon. \qquad \square
\end{aligned}$$

References

[1] D.P. Bertsekas and J.N. Tsitsiklis. *Neuro-Dynamic Programming.* Athena Scientific, 1996.

[2] S.M. Kakade. *On the Sample Complexity of Reinforcement Learning.* PhD thesis, University College London, 2003.