HAL open science

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

Bruno Scherrer

To cite this version:

Bruno Scherrer. On the Use of Non-Stationary Policies for Infinite-Horizon Discounted Markov Decision Processes. [Research Report] 2012. hal-00682172v2

HAL Id: hal-00682172

https://inria.hal.science/hal-00682172v2

Submitted on 30 Mar 2012

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

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

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

Bruno Scherrer

March 30, 2012

Abstract

We consider infinite-horizon γ-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 implicitely 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 particular, the use of non-stationary policies allows to reduce the usual asymptotic performance bounds of Value Iteration with errors bounded by ϵ at each iteration from $\frac{\gamma}{(1-\gamma)^2}\epsilon$ to $\frac{\gamma}{1-\gamma}\epsilon$, which is significant in the usual situation when γ is close to 1. Given Bellman operators that can only be computed with some error ϵ, a surprising consequence of this result is that the problem of "computing an approximately optimal non-stationary policy" is much simpler than that of "computing an approximately optimal stationary policy", and even slightly simpler than that of "approximately computing the value of some fixed policy", since this last problem *only* has a guarantee of $\frac{1}{1-\gamma}\epsilon$.

Given a Markov Decision Process, suppose on runs an approximate version of 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. Though it does not appear exactly in this form in the literature, the following performance bound is somewhat standard.

Theorem 1. *Let $\epsilon = \max_{1\leq j<k} \|\epsilon_j\|_{sp}$ be a uniform upper bound on the span seminorm[2] of the errors before iteration k. The loss of policy π_k is bounded as follows:*

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

In Theorem 2, we will prove a generalization of this result, so we do not provide a proof here. Since for any f, $\|f\|_{sp} \leq 2\|f\|_\infty$, Theorem 1 constitutes a slight improvement and a (finite-iteration) generalization of the following well-known performance bound (see [1]):

$$\limsup_{k\to\infty} \|v_* - v_{\pi_k}\|_\infty \leq \frac{2\gamma}{(1-\gamma)^2} \max_k \|\epsilon_k\|_\infty.$$

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

[2] For any function f defined on the state space, the span seminorm of f is $\|f\|_{sp} = \max_s f(s) - \min_s f(s)$. The motivation for using the span seminorm instead of a more usual L_∞-norm is twofold: 1) it slightly improves on the state-of-the-art bounds and 2) it simplifies the construction of an example in the proof of the forthcoming Proposition 1.

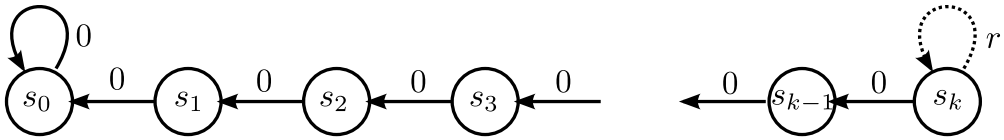

Figure 1: The deterministic MDP used in the proof of Proposition 1

Asymptotically, the above bounds involve a $\frac{\gamma}{(1-\gamma)^2}$ constant that may be really big when γ is close to 1. Compared to a value-iteration algorithm for approximately computing the value of some fixed policy, and for which one can prove a dependency of the form $\frac{1}{1-\gamma}\epsilon$, there is an extra term $\frac{\gamma}{1-\gamma}$ that suggests that the problem of "computing an approximately optimal policy" is significantly harder than that of "approximately computing the value of some fixed policy". To our knowledge, there does not exist any example in the literature that supports the tightness of the above mentionned bounds. The following proposition shows that the bound of Theorem 1 is in fact tight.

Proposition 1. *For all $\epsilon \geq 0$, $\Delta \geq 0$, and $k > 0$, there exists a $k+1$-state MDP, an initial value v_0 such that $\|v_* - v_0\|_{sp} = \Delta$, a sequence of noise terms (ϵ_j) with $\|\epsilon_j\|_{sp} \leq \epsilon$, such that running Value Iteration during iterations with errors (ϵ_j) outputs a value function v_{k-1} of which a greedy policy π_k satisfies Equation* (1) *with equality.*

Proof. Consider the deterministic MDP of Figure 1. The only decision is in state s_k, where one can stay with reward $r = -\frac{\gamma-\gamma^k}{1-\gamma}\epsilon - \gamma^k\Delta$ or move to s_{k-1} with 0 reward. All other transitions give 0 reward. Thus, there are only two policies, the optimal policy π_* with value equal to 0, and a policy $\bar{\pi}$ for which the value in s_k is $\frac{r}{1-\gamma}$. Take

$$v_0(s_l) = \begin{cases} -\Delta & \text{if } l = 0 \\ 0 & \text{else} \end{cases} \quad \text{and for all } j < k, \;\; \epsilon_j(s_l) = \begin{cases} -\epsilon & \text{if } j = l \\ 0 & \text{else.} \end{cases}$$

By induction, it can be seen that for all $j \in \{1, 2, \ldots, k-1\}$,

$$v_j(s_l) = \begin{cases} -\epsilon - \gamma\epsilon - \cdots - \gamma^{j-1}\epsilon - \gamma^j\Delta = -\frac{1-\gamma^j}{1-\gamma}\epsilon - \gamma^j\Delta & \text{if } j = l \\ 0 & \text{if } j < l \leq k \end{cases}$$

Since $\gamma v_{k-1}(s_{k-1}) = r$ and $v_{k-1}(s_k) = 0$, both policies are greedy with respect to v_{k-1}, and the bound of Equation (1) holds with equality for $\bar{\pi}$. □

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:

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

When $m = 1$, one exactly recovers the result of Theorem 1. 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:

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

that tends to $\frac{\gamma}{1-\gamma}\epsilon$ when k tends to ∞. In other words, we can see here that the problem of "computing a (non stationary) approximately-optimal policy" is not harder than that of "computing approximately the value of some fixed policy". Since the respective asymptotic errors are $\frac{\gamma}{1-\gamma}\epsilon$ and $\frac{1}{1-\gamma}\epsilon$, it seems even simpler !

Proof of Theorem 2. The value of $\pi_{k,m}$ satisfies:

$$v_{\pi_{k,m}} = T_{\pi_k} T_{\pi_{k-1}} \cdots T_{\pi_{k-m+1}} v_{\pi_{k,m}}. \tag{2}$$

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

$$T_{\pi_k} v_{k-1} = T_{\pi_k} T_{\pi_{k-1}} \cdots T_{\pi_{k-m+1}} v_{k-m} + \sum_{i=1}^{m-1} \Gamma_{k,i} \epsilon_{k-i} \tag{3}$$

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 Equations (3) and (2), one obtains:

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

and by taking the norm

$$\|T_{\pi_k} v_{k-1} - v_{\pi_{k,m}}\|_\infty = \gamma^m \|v_{k-m} - v_{\pi_{k,m}}\|_\infty + \frac{\gamma - \gamma^m}{1-\gamma} \epsilon_\infty \tag{4}$$

where $\epsilon_\infty = \max_{1 \le j \le k} \|\epsilon_j\|_\infty$. Essentially, Equation (4) shows that for sufficiently big m, $T_{\pi_k} v_{k-1}$ is an $\frac{\gamma}{1-\gamma}\epsilon$ approximation of the value of the non-stationary policy $\pi_{k,m}$ (whereas in general, it may be a much poorer approximation of the value of the stationary policy π_k.

By induction, it can also be proved that

$$\|v_* - v_k\|_\infty \le \gamma^k \|v_* - v_0\|_\infty + \frac{1-\gamma^k}{1-\gamma} \epsilon_\infty. \tag{5}$$

Using the fact that $\|T_{\pi_*} v_* - T_{\pi_k} v_{k-1}\|_\infty \le \gamma \|v_* - v_{k-1}\|_\infty$ since π_* (resp. π_k) is greedy with respect to v_* (resp. v_{k-1}), as well as Equations (4) and (5), we can conclude by observing that

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

Adding a constant to the value v_j at any step j of the algorithm does not affect the greedy policy set $\mathcal{G} v_j$ and only adds a constant to the next value v_{j+1}. As a consequence, we can assume witout loss of generality that $\|v_* - v_0\|_{sp} = 2\|v_* - v_0\|_\infty$, $\|\epsilon_j\|_{sp} = 2\|\epsilon_j\|_\infty$ and the result follows. □

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 *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) can also be useful in an infinite horizon context is to our knowledge new.

Acknowledgements I thank Boris Lesner for pointing out a flaw in a previous temptative proof of Proposition 1.

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.