Online Companion for "Multi-Objective Approaches to Markov Decision Processes with Uncertain Transition Parameters"

Dimitri Scheftelowitsch, Peter Buchholz, Vahid Hashemi, Holger Hermanns

October 18, 2017

Lemma 3.1. Let $\mathcal{P} = (S, A, T_{\updownarrow}, R_{\updownarrow}, Pr)$ be a SBMDP. Let furthermore π, π' be two policies where π' lies on the Pareto frontier. Then there exists a finite sequence of policies $\pi = \pi_0, \pi_1, \ldots, \pi_N = \pi'$ where $d(\pi_i, \pi_{i+1}) = 1, \mathbf{v}^{(\pi_i)} \not> \mathbf{v}^{(\pi_{i+1})}$ and, additionally, $N \leq |S|$.

Proof. We provide a proof by induction on $d(\pi, \pi')$. For $d(\pi, \pi') \in \{0, 1\}$, the statement holds obviously.

For $d(\pi, \pi') = c > 1$, the induction hypothesis is that the statement holds for $c - 1$. This means that for each policy π_1 with distance $d(\pi_1, \pi') = c - 1$ there exists a sequence of policies $\pi_1, \pi_2, \ldots, \pi_c = \pi'$ such that for any two adjacent policies π_i, π_{i+1} it is $\mathbf{v}^{(\pi_i)} \not> \mathbf{v}^{(\pi_{i+1})}$.

To show the induction step, we must infer the statement for $d(\pi, \pi') = c$. Suppose now for the sake of contradiction that it is not the case. We observe that under this assumption, for each state $s \in S$, the policy $\pi^{(s,\pi'(s))}$ that results from changing π in state s to choose action $\pi'(s)$ results in a value vector that is dominated by $\mathbf{v}^{(\pi)}$, i. e., $\mathbf{v}^{(\pi)} > \mathbf{v}^{(\pi^{(s,\pi'(s))})}$. Let us now consider a restricted SBMDP $\mathcal{P}^{[\pi,\pi']} = (S, A^{[\pi,\pi']}, T_{\updownarrow}^{[\pi,\pi']}, R_{\updownarrow}^{[\pi,\pi']})$ where the available actions are only those used in either π or π', that is, $A^{[\pi,\pi']} = \{a, b\}$ and the matrices P in $T_{\updownarrow}^{[\pi,\pi']}$ are constructed with $p_{s,s'}^{[\pi,\pi']a} = p_{s,s'}^{\pi(s)}$ and $p_{s,s'}^{[\pi,\pi']b} = p_{s,s'}^{\pi'(s)}$. The reward function is defined analogously by $R_{\updownarrow}^{[\pi,\pi']} = \left((\mathbf{r}_{\downarrow}^{[\pi,\pi']a}, \mathbf{r}_{\uparrow}^{[\pi,\pi']a}), (\mathbf{r}_{\downarrow}^{[\pi,\pi']b}, \mathbf{r}_{\uparrow}^{[\pi,\pi']b})\right)$ with

$$\mathbf{r}_{\downarrow s}^{[\pi,\pi']a} = \mathbf{r}_{\downarrow s}^{\pi(s)}, \mathbf{r}_{\uparrow s}^{[\pi,\pi']a} = \mathbf{r}_{\uparrow s}^{\pi(s)}$$
$$\mathbf{r}_{\downarrow s}^{[\pi,\pi']b} = \mathbf{r}_{\downarrow s}^{\pi'(s)}, \mathbf{r}_{\uparrow s}^{[\pi,\pi']b} = \mathbf{r}_{\uparrow s}^{\pi'(s)}.$$

It is easy to see that the policies π and π' can be executed in the new SBMDP $\mathcal{P}^{[\pi,\pi']}$. As all action changes from π lead to smaller value vectors in each component, we can see that π is locally optimal for each component, and thus, π is optimal for all components. Hence, π is an optimal policy in $\mathcal{P}^{[\pi,\pi']}$. Furthermore, π' is then dominated by π in all states and all components in $\mathcal{P}^{[\pi,\pi']}$ as well as in $\mathcal{P}$. Consequently, π' cannot lie on the Pareto frontier, which contradicts the initial assumption.

As we have arrived at a contradiction, we conclude that there must exist a state s where it is $\mathbf{v}^{(\pi)} \not> \mathbf{v}^{(\pi^{(s,\pi'(s))})}$, and, since $d(\pi^{(s,\pi'(s))}, \pi') = c-1$ and $d(\cdot, \cdot)$

can never exceed $|S|$, there exists, by induction hypothesis, a sequence of policies $\pi^{(s,\pi'(s))} = \pi_1, \pi_2, \ldots, \pi_c = \pi'$ for which $\mathbf{v}^{(\pi_i)} \not\succ \mathbf{v}^{(\pi_{i+1})}$. As $d(\pi, \pi^{(s,\pi'(s))}) = 1$, this concludes the proof.

Theorem 3.2. Algorithm 1 correctly computes $\mathcal{P}_{\text{Pareto}}$.

The correctness of the algorithm follows from Lemma 3.1. In detail, Algorithm 1 stores a set P of policies. In the i-th step, the set P is updated with policies that have distance 1 from already computed policies in P and distance i from π_0; a further constraint restricts the policies to be non-dominated by their "parent" in P. This way, after i steps P contains all policies with distance i from π_0 that follow a non-dominated path. By computing the non-dominated subset of currently found policies in line 6, we maintain a set of mutually non-dominated policies that are reachable on a non-dominated path from π_0. By Lemma 3.1., this captures all policies from $\mathcal{P}_{\text{Pareto}}$.