Policy and Value Transfer in Lifelong Reinforcement Learning (Appendix)

David Abel [† 1] **Yuu Jinnai** [† 1] **Yue Guo** [1] **George Konidaris** [1] **Michael L. Littman** [1]

We here include proofs and a visual of the Octogrid domain.

A. Proofs

Theorem 3.2. *For a distribution of MDPs with $R \sim D$,*

$$\mathbb{E}_{M\in\mathcal{M}}[V_M^{\pi^*_{avg}}(s)] \geq \max_{M\in\mathcal{M}} \Pr(M) V^*_M(s).$$

Proof. Ramachandran Amir (2007) also showed that the value function V^π_{avg} of an average MDP is the weighted average of the MDPs in the distribution,

$$V^\pi_{avg}(s) = \sum_{M\in\mathcal{M}} \Pr(M) V^\pi_M(s). \quad (1)$$

Thus,

$$\begin{aligned}
\mathbb{E}_{M\in\mathcal{M}}[V_M^{\pi^*_{avg}}(s)] &= \sum_{M\in\mathcal{M}} \Pr(M) V_M^{\pi^*_{avg}}(s)\\
&= V_{avg}^{\pi^*_{avg}}(s)\\
&= \max_\pi V^\pi_{avg}(s)\\
&= \max_\pi \sum_{M\in\mathcal{M}} \Pr(M) V^\pi_M(s)\\
&\geq \max_\pi \max_{M\in\mathcal{M}} \Pr(M) V^\pi_M(s)\\
&= \max_{M\in\mathcal{M}} \Pr(M) \max_\pi V^\pi_M(s)\\
&= \max_{M\in\mathcal{M}} \Pr(M) V^*_M(s).
\end{aligned}$$

Since we assume $\mathcal{R}(s,a) \geq 0$ for all s, a, we infer that $\sum_{M\in\mathcal{M}} \Pr(M) V^\pi_M(s) \geq \max_{M\in\mathcal{M}} \Pr(M) V^\pi_M(s)$, thus concluding the proof. □.

....

Corollary 3.2.1. *The bound in Theorem 3.2 is tight.*

Proof. Next we the bound is by an example MDP distribution shown in Figure 1.

In the MDP i the agent gets a reward if it executes a_i in MDP i:

$$R_M(s_0, a_i) = \begin{cases} 1 & M = i \\ 0 & \text{otherwise} \end{cases}$$

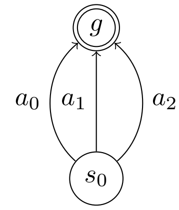

Figure 1: An example of a MDP which an average MDP solution returns a lower bound value.

In this distribution of MDPs, the optimal agent always gets reward of 1 where as the optimal average MDP agent gets $\max_{M\in\mathcal{M}} \Pr(M)$ reward on average. In this setting, $V^{\pi^*_{avg}}(s) = \max_{M\in\mathcal{M}} \Pr(M) V^*_M(s)$. Thus the bound is tight. □

....

Corollary 3.4. *For the $G \sim D$ setting,*

$$\begin{aligned}
&\mathbb{E}_{M\in\mathcal{M}}[V_M^{\pi^*_{avg}}(s)]\\
&\qquad \geq \min_{M\in\mathcal{M}} \Pr(M) \max_{M'\in\mathcal{M}} \Pr(M') V^*_{M'}(s).
\end{aligned}$$

Proof. We first leverage the following lemma:

Lemma 3.4.1.

$$\begin{aligned}
\max_{M\in\mathcal{M}} \Pr(M) V^\pi_M(s) &\leq V^\pi_{avg}(s)\\
&\leq \sum_{M\in\mathcal{M}} \Pr(M) V^\pi_M(s) / \min_{M'\in\mathcal{M}} \Pr(M')
\end{aligned}$$

(Proof sketch for lower bound): Let an MDP M' be the same MDP as M except it transits to a terminal state from goal nodes (and acquires a reward) by probability of $\Pr(M)$ instead of probability of 1. The value $V^\pi_{M'}(s)$ of state s in M' is at least as large as $\Pr(M) V^\pi_M(s)$. Thus, the value of state s in M' is lower than or equal to that in the average MDP as it reaches the goal less frequently. $V^\pi_{M'}(s)$ is smaller that or equal to $V^\pi_{avg}(s)$ as the average MDP has larger or equal probability of reaching the terminal state. Thus, for any $M \in \mathcal{M}$:

$$V^\pi_{avg}(s) \geq V^\pi_{M'}(s) \geq \Pr(M) V^\pi_M(s).$$

(Proof sketch for upper bound):

$$V^{\pi}_{avg}(s) \leq \sum_{M \in \mathcal{M}} V^{\pi}_{M}(s) \leq \sum_{M \in \mathcal{M}} \Pr(M) V^{\pi}_{M}(s) / \min_{M' \in \mathcal{M}} \Pr(M').$$

Now, we turn to the theorem.

$$\begin{aligned} \mathbb{E}_{M \in \mathcal{M}}[V^{\pi^*_{avg}}_{M}(s)] &= \sum_{M \in \mathcal{M}} \Pr(M) V^{\pi^*_{avg}}_{M}(s) \\ &\geq \min_{M \in \mathcal{M}} \Pr(M) V^{\pi^*_{avg}}_{avg}(s) \\ &= \min_{M \in \mathcal{M}} \Pr(M) \max_{\pi} V^{\pi}_{avg}(s) \\ &\geq \min_{M \in \mathcal{M}} \Pr(M) \max_{\pi} \max_{M' \in \mathcal{M}} \Pr(M') V^{\pi}_{M'}(s) \\ &= \min_{M \in \mathcal{M}} \Pr(M) \max_{M' \in \mathcal{M}} \Pr(M') V^{*}_{M'}(s). \qquad \square \end{aligned}$$

....

Theorem 3.8. *Suppose $\mathscr{A}$ is an algorithm that produces ε accurate Q functions for a subset of the state action space given an MDP M, an initial state s_0, and a horizon H. For a given $\delta \in (0, 1]$, after*

$$t \geq \frac{\ln(\delta)}{\ln(1 - p_{min})}, \tag{2}$$

sampled MDPs, for $p_{min} = \min_{M \in \mathcal{M}} \Pr(M)$, the updating-max shaping method will return a shaped Q-function $\hat{Q}_{max}$ such that for all state action pairs (s, a):

$$\hat{Q}_{max}(s, a) \geq \max_{M} Q^{*}_{M}(s, a), \tag{3}$$

with probability $1 - \delta$.

Proof. Consider an arbitrary state action pair (s,a).

After t samples, we choose:

$$\hat{Q}^{*}_{max}(s, a) \triangleq \max_{M} \hat{Q}^{*}_{M}(s, a). \tag{4}$$

After t samples, we let the following event define a mistake:

$$\hat{Q}^{*}_{max}(s, a) < \max_{M} Q^{*}_{M}(s, a). \tag{5}$$

First, we suppose that for each of sampled MDP M, our learning algorithm computes a *partial* but nearly *accurate* Q-function. That is, for some small ε:

$$\hat{Q}^{*}_{M}(s, a) = \begin{cases} Q^{*}_{M}(s, a) \pm \varepsilon & c(s, a) \geq m \\ \text{VMAX} & \text{otherwise} \end{cases} \tag{6}$$

That is, letting $c(s, a)$ denote the number of times a was executed in s: any state action pairs that were visited sufficiently often (more than m for some chosen $m << H$) result in an ε-accurate Q function. Otherwise, the algorithm returns VMAX.

Under these conditions, for a given state action pair, surely, for any MDP seen during the t samples M_i:

$$\hat{Q}^{*}_{max}(s, a) \geq \max_{M \in \mathcal{M}_{seen}} Q^{*}_{M}(s, a) \tag{7}$$

Therefore, the mistake event defined by Equation 5 only occurs when we *miss* an MDP in the distribution that has a higher $Q^*(s, a)$ than our estimate. We assume that the distribution has a lower bound on MDP probabilty:

$$p_{min} \triangleq \min_{M \in \mathcal{M}} \Pr(M). \tag{8}$$

Accordingly, we upper bound the mistake probability according to the probability that no such MDP was sampled over t samples, captured by the cumulative geometric distribution:

$$1 - (1 - p_{min})^{m} \geq 1 - \delta. \tag{9}$$

Simplifying:

$$\begin{aligned} 1 + \delta &\geq 1 + (1 - p_{min})^{t} \\ \ln(\delta) &\geq \ln(1 - p_{min}) \cdot t \\ \frac{\ln(\delta)}{\ln(1 - p_{min})} &\leq t \end{aligned}$$

Therefore, after

$$t \geq \frac{\ln(\delta)}{\ln(1 - p_{min})}, \tag{10}$$

sampled MDP we will have seen all MDPs in the distribution with high probability. $\square$

B. Octogrid

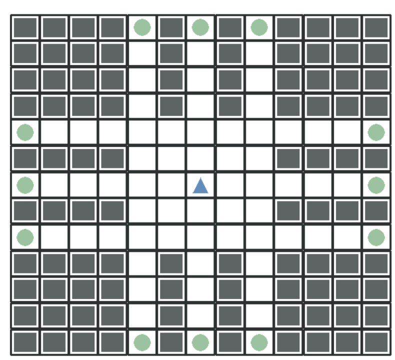

Figure 2: The Octogrid task distribution. The goal appears in exactly one of the 12 green circles chosen uniformly at random, with the agent starting in the center at the triangle.