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 \( V_{avg}^{\pi} \) of an average MDP is the weighted average of the MDPs in the distribution,
\[
V_{avg}^{\pi}(s) = \sum_{M \in \mathcal{M}} \Pr(M)V_M^{\pi}(s).
\]
Thus,
\[
\mathbb{E}_{M \in \mathcal{M}}[V_M^{\pi_{avg}}(s)] = \sum_{M \in \mathcal{M}} \Pr(M)V_M^{\pi_{avg}}(s)
= V_{avg}^{\pi_{avg}}(s)
= \max_{\pi} V_{avg}^{\pi}(s)
= \max_{\pi} \sum_{M \in \mathcal{M}} \Pr(M)V_M^{\pi}(s)
\geq \max_{\pi} \max_{M \in \mathcal{M}} \Pr(M)V_M^{\pi}(s)
= \max_{M \in \mathcal{M}} \Pr(M) \max_{\pi} V_M^{\pi}(s)
= \max_{M \in \mathcal{M}} \Pr(M)V_M^{*}(s).
\]
Since we assume \( \mathcal{R}(s,a) \geq 0 \) for all \( s,a \), we infer that
\[
\sum_{M \in \mathcal{M}} \Pr(M)V_M^{\pi}(s) \geq \max_{M \in \mathcal{M}} \Pr(M)V_M^{*}(s),
\]
thus concluding the proof. \( \square \)

**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 \( i \):
\[
R_M(s_0,a_i) = \begin{cases} 
1 & M = i \\
0 & \text{otherwise}
\end{cases}
\]
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, \( V_{avg}^{\pi_{avg}}(s) = \max_{M \in \mathcal{M}} \Pr(M)V_M^{*}(s) \). Thus the bound is tight. \( \square \)

**Corollary 3.4.** For the \( G \sim D \) setting,
\[
\mathbb{E}_{M \in \mathcal{M}}[V_M^{\pi_{avg}}(s)] \geq \min_{M \in \mathcal{M}} \Pr(M) \max_{M' \in \mathcal{M}} \Pr(M')V_{M'}^{*}(s).
\]

**Proof.** We first leverage the following lemma:

**Lemma 3.4.1.**
\[
\max_{M \in \mathcal{M}} \Pr(M)V_M^{\pi}(s) \leq V_{avg}^{\pi}(s)
\leq \sum_{M \in \mathcal{M}} \Pr(M)V_M^{\pi}(s)/\min_{M' \in \mathcal{M}} \Pr(M').
\]

(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 \( V_M^{\pi}(s) \) of state \( s \) in \( M' \) is at least as large as \( \Pr(M)V_M^{\pi}(s) \). 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. \( V_{avg}^{\pi}(s) \) is smaller that or equal to \( V_{avg}^{\pi}(s) \) as the average MDP has larger or equal probability of reaching the terminal state. Thus, for any \( M \in \mathcal{M} \):
\[
V_{avg}^{\pi}(s) \geq V_M^{\pi}(s) \geq \Pr(M)V_M^{*}(s).
\]
Proof sketch for upper bound:

\[ V^\pi_{\text{avg}}(s) \leq \sum_{M \in \mathcal{M}} V^\pi_M(s) \leq \sum_{M \in \mathcal{M}} \Pr(M) V^\pi_M(s) / \min_{M' \in \mathcal{M}} \Pr(M'). \]

Now, we turn to the theorem.

\[ \mathbb{E}_{M \in \mathcal{M}}[V^\pi_{\text{max}}^*(s)] = \sum_{M \in \mathcal{M}} \Pr(M) V^\pi_M(s) \geq \min_{M \in \mathcal{M}} \Pr(M) \max_{\pi} V^\pi_M(s) = \min_{M \in \mathcal{M}} \Pr(M) \max_{M' \in \mathcal{M}} \Pr(M') V^\pi_{M'}(s). \]

\[ \delta \leq \ln(1 - \min\{\mathbb{E}_{M \in \mathcal{M}}[V^\pi_{\text{max}}^*(s)], \mathbb{E}_{M \in \mathcal{M}}[V^\pi_{\text{max}}^*(s)] \}) \]

Theorem 3.8. Suppose \( \mathcal{A} \) is an algorithm that produces \( \varepsilon \) 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

\[ t \geq \frac{\ln(\delta)}{\ln(1 - \min\{\mathbb{E}_{M \in \mathcal{M}}[V^\pi_{\text{max}}^*(s)], \mathbb{E}_{M \in \mathcal{M}}[V^\pi_{\text{max}}^*(s)] \})}, \]

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

\[ \hat{Q}_{\text{max}}^*(s, a) \geq \max_M Q^*_M(s, a), \]

with probability \( 1 - \delta \).

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

After \( t \) samples, we choose:

\[ \hat{Q}_{\text{max}}^*(s, a) \triangleq \max_M Q^*_M(s, a). \]

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

\[ \hat{Q}_{\text{max}}^*(s, a) < \max_M Q^*_M(s, a). \]

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 \( \varepsilon \):

\[ \hat{Q}_M^*(s, a) = \begin{cases} Q^*_M(s, a) + \varepsilon & \text{if } c(s, a) \geq m \\ \text{VMAX otherwise} \end{cases} \]

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 \( \varepsilon \)-accurate \( Q \) function. Otherwise, the algorithm returns \text{VMAX}.

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

\[ \hat{Q}_{\text{max}}^*(s, a) \geq \max_{M \in \mathcal{M}} Q^*_M(s, a) \]

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 probability:

\[ p_{\min} \triangleq \min_{M \in \mathcal{M}} \Pr(M). \]

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:

\[ 1 - (1 - p_{\min})^t \geq 1 - \delta. \]

Simplifying:

\[ 1 + \delta \geq 1 + (1 - p_{\min})^t \]

\[ \frac{\ln(\delta)}{\ln(1 - p_{\min})} \leq t \]

Therefore, after

\[ t \geq \frac{\ln(\delta)}{\ln(1 - p_{\min})}, \]

sampled MDP we will have seen all MDPs in the distribution with high probability. \( \Box \)

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.