Supplementary Material for Monte Carlo Value Iteration with Macro-Actions

Lemma 1 *Given value functions U and V, $||HU - HV||_\infty \leq \gamma ||U - V||_\infty$.*

Proof.
Let b be an arbitrary belief and assume that $HV(b) \leq HU(b)$ holds. Let $\mathbf{a}^*$ be the optimal macro action for $HU(b)$. Then

$$\begin{aligned}
0 &\leq HU(b) - HV(b) \\
&\leq \mathbf{R}(b, \mathbf{a}^*) + \gamma \sum_{\mathbf{o} \in \mathcal{O}} p_\gamma(\mathbf{o}|\mathbf{a}^*, b) U(\tau(b, \mathbf{o}, \mathbf{a}^*)) - \mathbf{R}(b, \mathbf{a}^*) - \gamma \sum_{\mathbf{o} \in \mathcal{O}} p_\gamma(\mathbf{o}|\mathbf{a}^*, b) V(\tau(b, \mathbf{o}, \mathbf{a}^*)) \\
&= \gamma \sum_{\mathbf{o} \in \mathcal{O}} p_\gamma(\mathbf{o}|\mathbf{a}^*, b)[U(\tau(b, o, \mathbf{a}^*) - V(\tau(b, o, \mathbf{a}^*))] \\
&\leq \gamma \sum_{\mathbf{o} \in \mathcal{O}} p_\gamma(\mathbf{o}|\mathbf{a}^*, b) ||U - V||_\infty \\
&\leq \gamma ||U - V||_\infty.
\end{aligned}$$

Since $|| \cdot ||_\infty$ is symmetrical, the result is the same for the case of $HU(b) \leq HV(b)$. By taking $|| \cdot ||_\infty$ over all weighted belief, we get

$$||HU - HV||_\infty \leq \gamma ||U - V||_\infty.$$

Thus, H is a contractive mapping. □

Theorem 2 *The value function for an m-step policy is piecewise linear and convex and can be represented as*

$$V_m(b) = \max_{\alpha \in \Gamma_m} \sum_{s \in S} \alpha(s) b(s) \tag{1}$$

where Γ_m is a finite collection of α-vectors.

Proof.
We prove this property by induction. When $m = 1$, the intial value function V_1 is the best expected reward and can be written as

$$V_1(b) = \max_{\mathbf{a}} \mathbf{R}(b, \mathbf{a}) = \max_{\mathbf{a}} \sum_{s \in S} \mathbf{R}(s, \mathbf{a}) b(s).$$

This has the same form as $V_m(b) = \max_{\alpha_m \in \Gamma_m} \sum_{s \in S} \alpha_m(s) b(s)$ where there is one linear α-vector for each macro action. $V_1(b)$ can therefore be represented as a finite collection of α-vectors.

Assuming the optimal value function for any b_{i-1} is represented using a finite set of α-vector $\Gamma_{i-1} = \{\alpha^0_{i-1}, \alpha^1_{i-1}, ...\}$ and

$$V_{i-1}(b_{i-1}) = \max_{\alpha_{i-1} \in \Gamma_{i-1}} \sum_{s \in S} b_{i-1}(s) \alpha_{i-1}(s) \tag{2}$$

Substituting

$$b_{i-1}(s) = \sum_{j=1}^{\infty} \gamma^{j-1} \sum_{s'} p(s, \mathbf{o}, j | s', \mathbf{a}) b_i(s') / p_\gamma(\mathbf{o}|\mathbf{a}, b_i)$$

into (2), we get

$$V_{i-1}(b_{i-1}) = \max_{\alpha_{i-1} \in \Gamma_{i-1}} \sum_{s \in S} \frac{\sum_{j=1}^{\infty} \gamma^{j-1} \sum_{s'} p(s, \mathbf{o}, j | s', \mathbf{a}) b_i(s')}{p_\gamma(\mathbf{o}|\mathbf{a}, b_i)} \alpha_{i-1}(s).$$

Substituting it into the backup equation gives

$$\begin{aligned}
V_i(b_i) &= \max_{\mathbf{a}} \Big(\mathbf{R}(b_i, \mathbf{a}) + \gamma \sum_{\mathbf{o} \in \mathcal{O}} p_\gamma(\mathbf{o}|\mathbf{a}, b_i) \max_{\alpha_{i-1} \in \Gamma_{i-1}} \sum_{s \in S} \frac{\sum_{j=1}^{\infty} \gamma^{j-1} \sum_{s'} p(s, \mathbf{o}, j | s', \mathbf{a}) b_i(s')}{p_\gamma(\mathbf{o}|\mathbf{a}, b_i)} \alpha_{i-1}(s) \Big) \\
&= \max_{\mathbf{a}} \Big(\mathbf{R}(b_i, \mathbf{a}) + \gamma \sum_{\mathbf{o} \in \mathcal{O}} \max_{\alpha_{i-1} \in \Gamma_{i-1}} \sum_{s \in S} \sum_{j=1}^{\infty} \gamma^{j-1} \sum_{s'} p(s, \mathbf{o}, j | s', \mathbf{a}) b_i(s') \alpha_{i-1}(s) \Big) \\
&= \max_{\mathbf{a}} \max_{\alpha^1_{i-1} \in \Gamma_{i-1}, \ldots, \alpha^{|\mathcal{O}|}_{i-1}} \sum_{s' \in S} b_i(s') \left[\mathbf{R}(s', \mathbf{a}) + \gamma \sum_{\mathbf{o} \in \mathcal{O}} \sum_{s \in S} \sum_{j=1}^{\infty} \gamma^{j-1} p(s, \mathbf{o}, j | s', \mathbf{a}) \alpha^{\mathbf{o}}_{i-1}(s) \right]
\end{aligned}$$

The expression in the square bracket can evaluate to $|\mathcal{A}||\Gamma_{i-1}|^{|\mathcal{O}|}$ different vectors. We can rewrite $V_i(b_i)$ as:

$$V_i(b_i) = \max_{\alpha_i \in \Gamma_i} \sum_{s \in S} \alpha_i(s) b_i(s).$$

Hence $V_i(b_i)$ can be represented by a finite set of α-vector. $\square$