EVB Decomposition

Change in the value function due to learning after taking action a^*:

$$\begin{aligned} v(ba^*) - v(b) =& \sum_{b'} p(b' \mid b, a^*) \Big(v(b') - v(b) \Big) \\ =& \sum_{b'} p(b' \mid b, a^*) \Big(\sum_a \pi(a \mid b') q(b', a) - \sum_a \pi(a \mid b) q(b, a) \Big) \\ =& \sum_{b'} p(b' \mid b, a^*) \sum_a \Big(\big(\pi(a \mid b') - \pi(a \mid b) \big) q(b', a) \\ &+ \pi(a \mid b) \big(q(b', a) - q(b, a) \big) \Big) \end{aligned} \tag{1}$$

Expanding $q(b', a) - q(b, a)$:

$$\begin{aligned} q(b', a) - q(b, a) =& \sum_{b''} p(b'' \mid b', a) \big[r(b', a) + \gamma v(b'') \big] \\ &- \sum_{b'} p(g' \mid b, a) \big[r(b, a) + \gamma v(g') \big] \\ =& r(b', a) + \gamma \sum_{b''} p(b'' \mid b', a) v(b'') \\ &- r(b, a) + \gamma \sum_{g'} p(g' \mid b, a) v(g') \\ =& \underbrace{r(b', a) - r(b, a)}_{\substack{\text{Difference in the expected} \\ \text{immediate return}}} + \underbrace{\gamma \big[\sum_{b''} p(b'' \mid b', a) v(b'') - \sum_{g'} p(g' \mid b, a) v(g') \big]}_{\substack{\text{Difference in the expected} \\ \text{future return}}} \end{aligned} \tag{2}$$

So overall the EVB decomposes as:

$$\begin{aligned} v(ba^*) - v(b) =& \mathbb{E}_{b' \sim p(b' \mid b, a^*)} \Big[\sum_a \big(\pi(a \mid b') - \pi(a \mid b) \big) q(b', a) \\ &+ \mathbb{E}_{a \sim \pi(a \mid b)} \big[r(b', a) - r(b, a) \big] \\ &+ \mathbb{E}_{a \sim \pi(a \mid b)} \big[\gamma \sum_{b''} p(b'' \mid b', a) v(b'') - \gamma \sum_{g'} p(g' \mid b, a) v(g') \big] \Big] \end{aligned} \tag{3}$$

Simulations

blabla