

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

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

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

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

So overall the EVB decomposes as:

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