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:

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

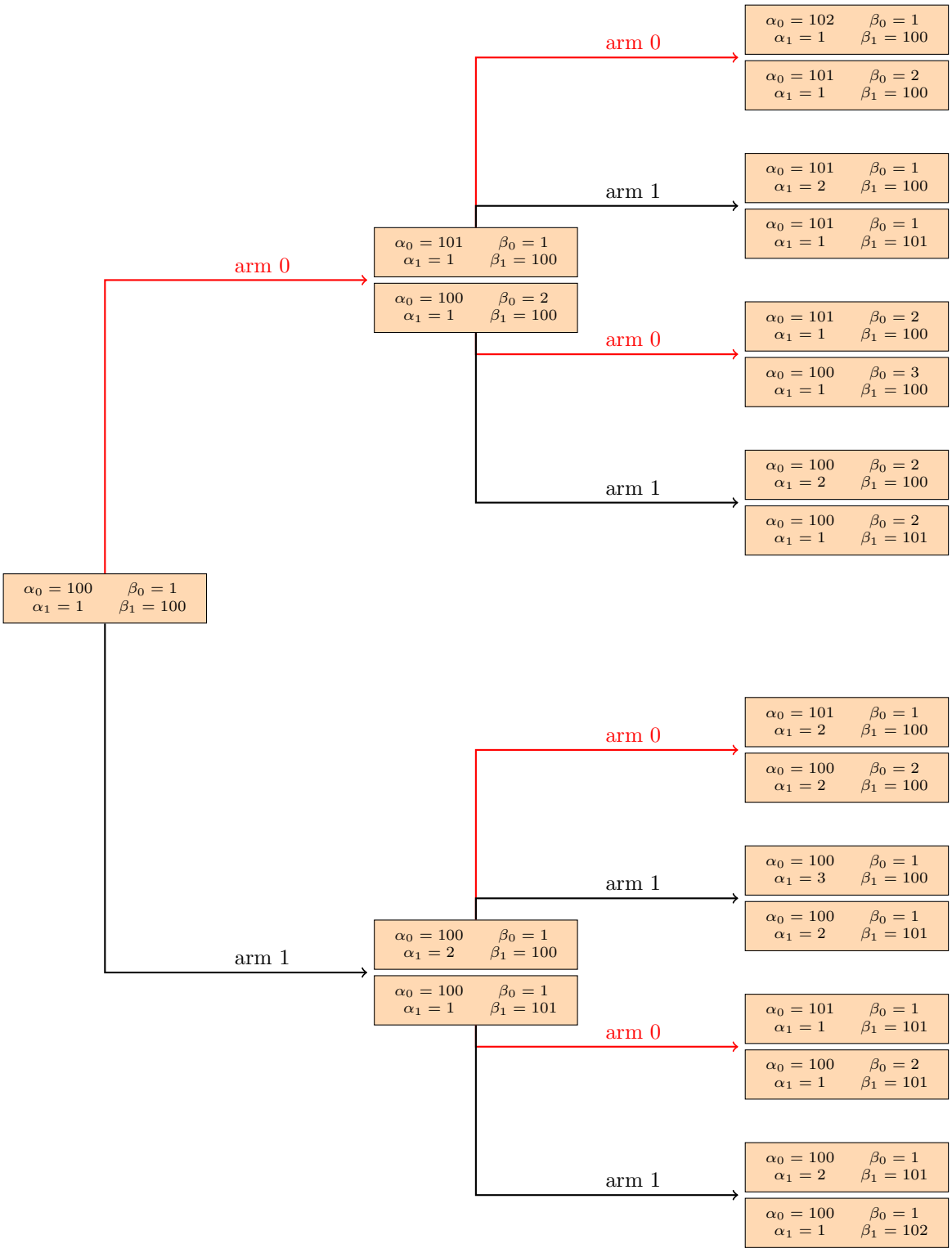
One last thing to consider is how the prioritisation of distal experiences should differ from those that are more immediate. This also has implications for how likely those experiences are to occur according to the current model.

For instance, if the agent considers updating $v(b)$ towards the value that would result from taking a particular action from that belief state – say, $v(ba^*)$ – the EVB associated with that update needs to be weighted by the probability of transitioning into belief state b in the first place (i.e., from the current root of the tree).

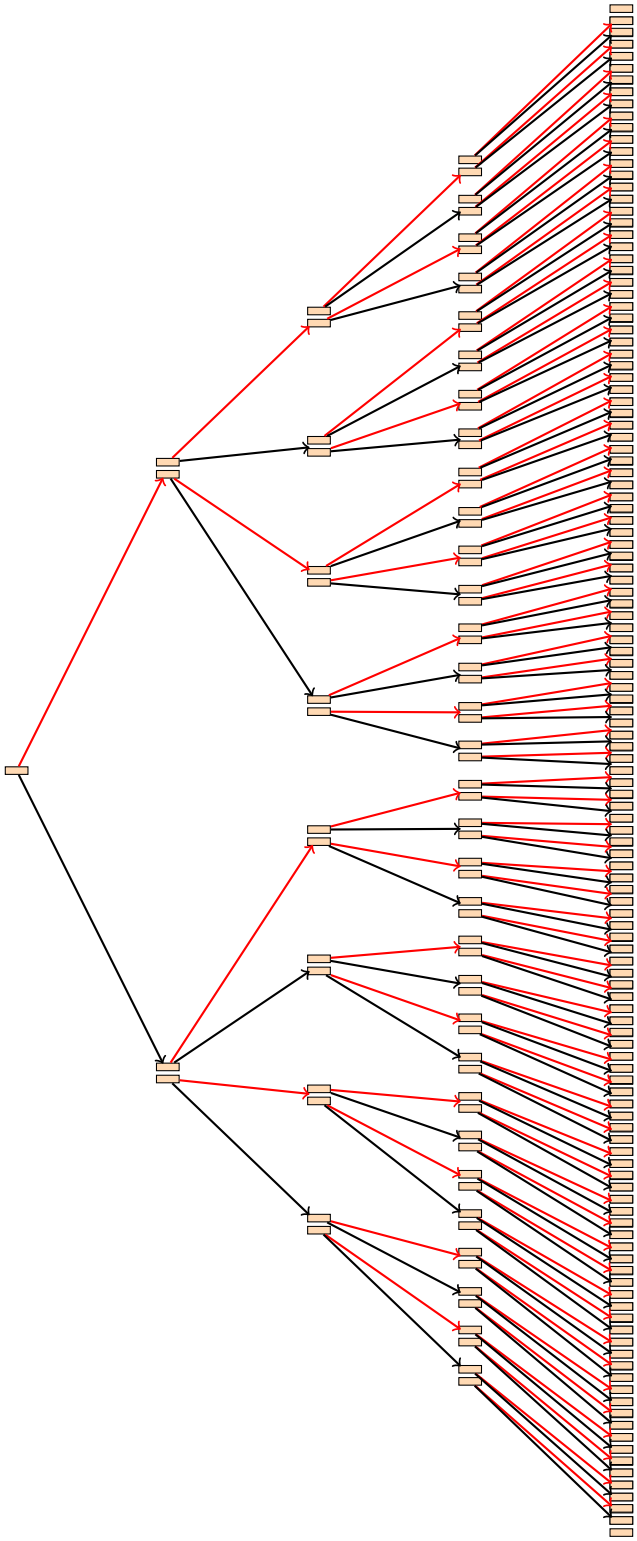
$$
\begin{aligned}
v(ba^*) - v(b) = p(b_{\text{root}} \to b) \times \Big(& \mathbb{E}_{b' \sim p(b' \mid b, a^*)} \Big[\sum_{a} \big(\pi(a \mid b') - \pi(a \mid b) \big) q(b', a) \\
&+ \mathbb{E}_{a \sim \pi(a \mid b)} \big[r(b', a) - r(b, a) \big] \\
&+ \mathbb{E}_{a \sim \pi(a \mid b)} \big[\gamma \sum_{b''} p(b'' \mid b', a) v(b'') - \gamma \sum_{g'} p(g' \mid b, a) v(g') \big] \Big] \Big)
\end{aligned}
\tag{4}
$$

Simulations

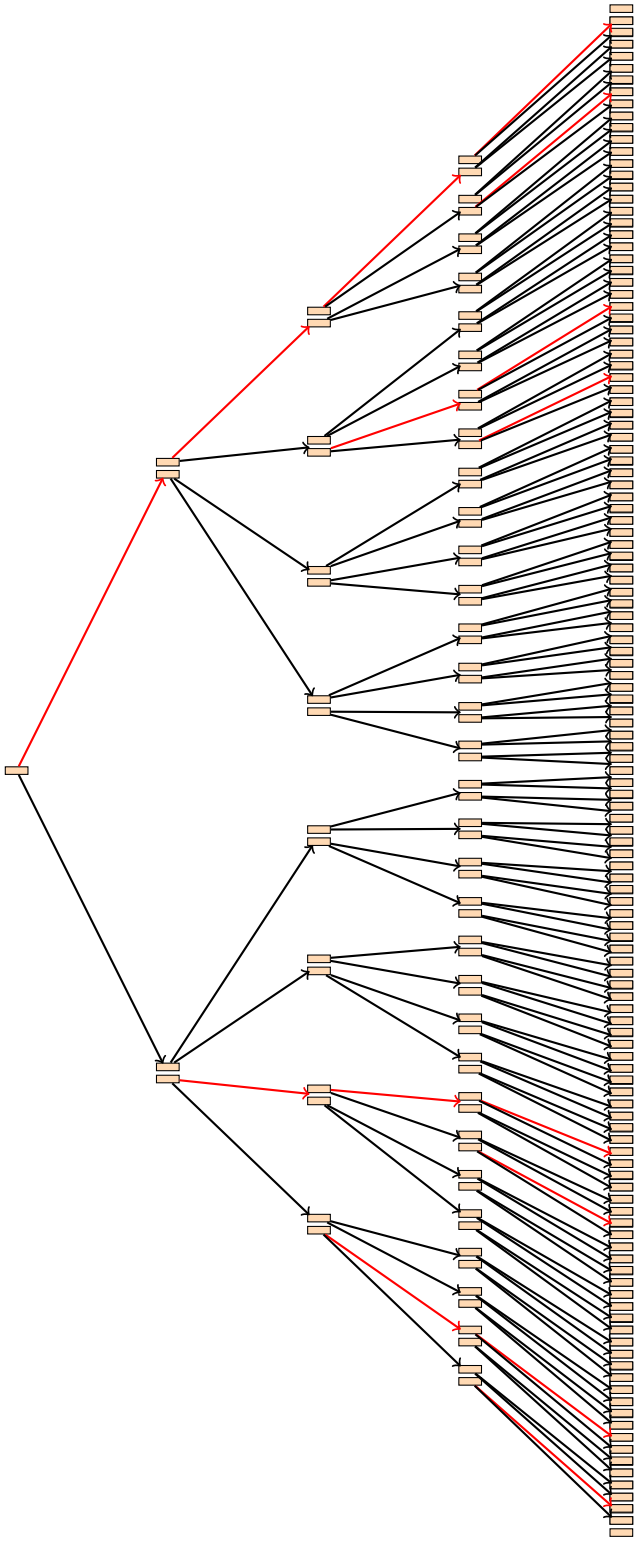
Prioritisation pattern with horizon $h = 2$. Each orange box is a belief state, and the parameters that correspond to that belief are written inside. Red arrows show which replay updates were executed.



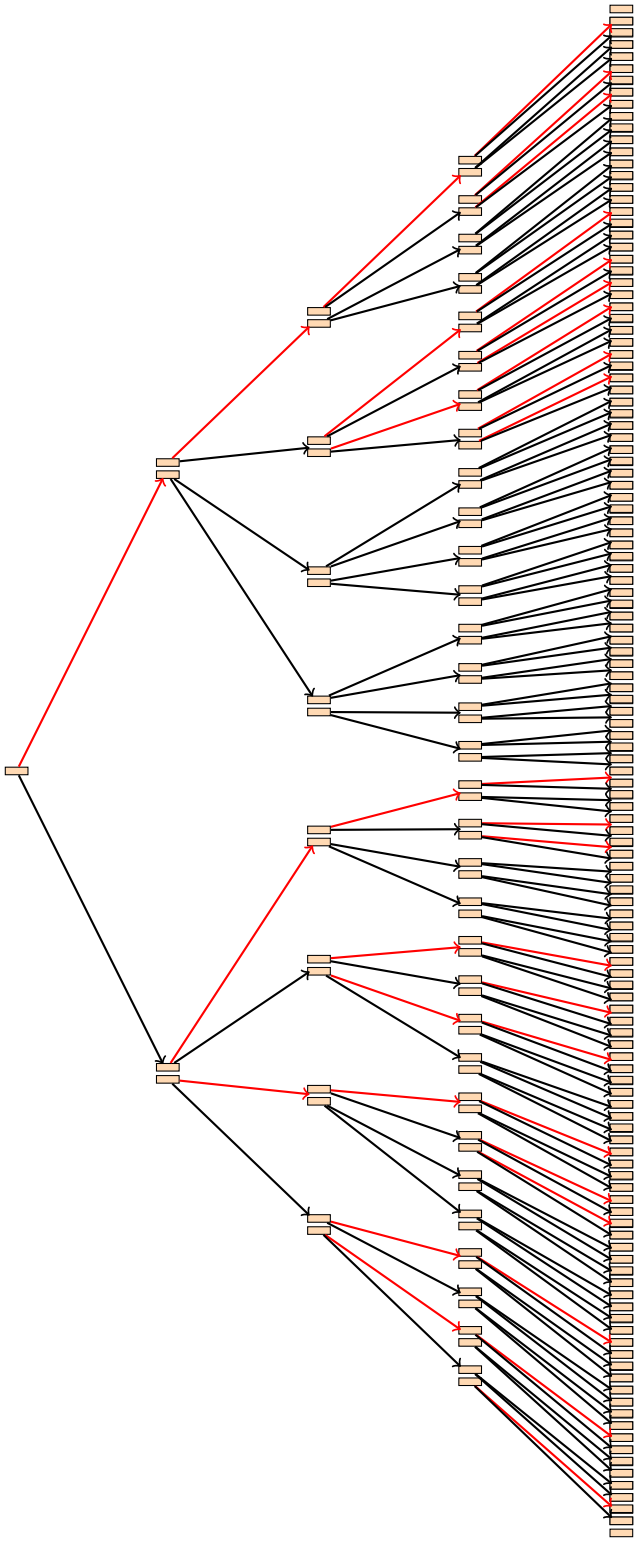
Prioritisation pattern with horizon $h = 4$. The prior belief at the root is set to the same values as in the above example with shorter horizon. Note that for this (and the previous) tree, EVB was not scaled by the probability of reaching any belief.



Prioritisation pattern with horizon $h = 4$. Same example as before, however, here, each EVB value $v(ba^*) - v(b)$ is scaled by the probability of reaching belief b from the root of the tree. Note how the less-likely branches are ignored in this case, as opposed to the one showed above.



Prioritisation pattern with horizon $h = 4$. In this case, the belief at the root was set to $(\alpha_0 = 100, \beta_0 = 1, \alpha_1 = 1, \beta_1 = 1)$ – i.e., complete uncertainty about the outcome of the second arm (branch that leads downwards from the root). Note how the 'exploration' of the lower sub-tree is much more extensive in this case, compared to the example shown above with the root prior $(\alpha_0 = 100, \beta_0 = 1, \alpha_1 = 1, \beta_1 = 100)$.



I omitted the model-free policy in the above examples – i.e., the policy was specified as the softmax of the immediate expected reward according to the model at any given belief state (including the root node).

Sequences

When each node/leaf is initialised to have a particular value (for instance, to the immediate reward according to the agent's model) – then the order of replay execution does not necessarily follow backward propagation. This is due to the fact that for higher values of the beta distribution parameters, each individual update has a smaller impact on the shape of the resulting posterior density.

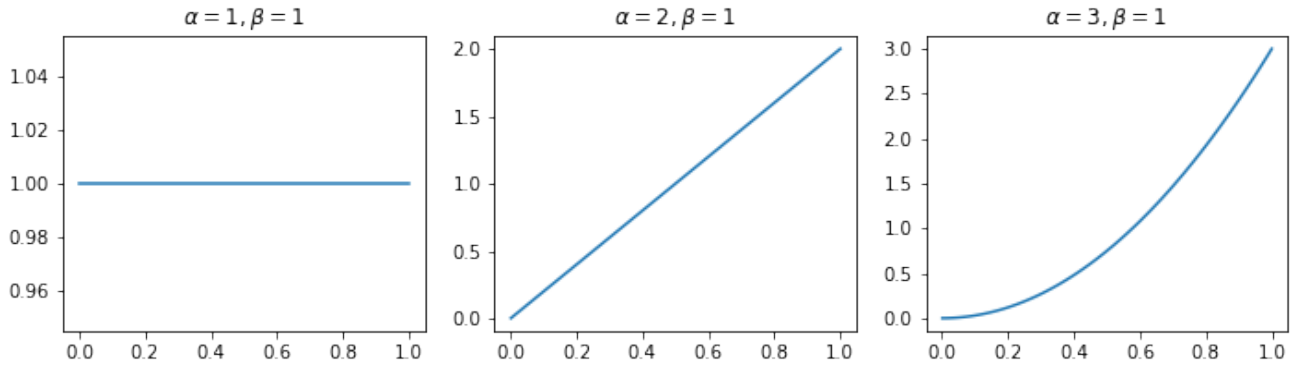


Figure 1: Consequitive Bayesian updates of the beta distribution. Left to right.

This is demonstrated in the example in Figure 1. KL divergence between the leftmost and middle pdfs is 33, whereas for the middle and rightmost pdfs it is 9.5. Therefore, the EVB associated with the update

$$v(\{\alpha = 1, \beta = 1\}) \leftarrow v(\{\alpha = 2, \beta = 1\})$$

will be greater than that of

$$v(\{\alpha = 2, \beta = 1\}) \leftarrow v(\{\alpha = 3, \beta = 1\})$$

The last example tree below shows how the sequence of replay events was executed where each node's value was set to 0, except for 1) the root node and 2) the leaf nodes. The prior at the root node was set to

$$\begin{array}{ll} \alpha_0 = 10 & \beta_0 = 5 \\ \alpha_1 = 3 & \beta_1 = 1 \end{array}$$