s_{agent}	current physical state of the agent
b_{agent}	current belief state of the agent
s_ρ	physical root state of a planning tree
N_{s_ρ}	number of steps it takes the agent to reach s_ρ from s_{agent}
s_k	physical state at which an update is considered
a_k	action for which an update is considered
b_k	belief state at which an update is considered

Table 1: Notation

EVB for information states

We are ultimately interested in how the value of the current information state $z_{\text{agent}} = \langle s_{\text{agent}}, b_{\text{agent}} \rangle$ changes as a result of a policy update at some future information state $z_k \in \mathcal{Z}$:

$$v_{\pi_{\text{new}}}(z_{\text{agent}}) - v_{\pi_{\text{old}}}(z_{\text{agent}}) = \sum_{z \in \mathcal{Z}} \underbrace{\sum_{i=0}^{\infty} \gamma^i P(z_{\text{agent}} \to z, i, \pi_{\text{old}})}_{\text{Need}} \times \underbrace{\sum_a \big[\pi_{\text{new}}(z,a) - \pi_{\text{old}}(z,a)\big] q_{\pi_{\text{new}}}(z,a)}_{\text{Gain}}$$

Need is currently estimated in the following way:

1. Simulate K forward trajectories from the agent's current state s_{agent} and belief b_{agent}, using the old policy π_{old}
2. Terminate each trajectory if $\gamma^d < \epsilon$ where d is the trajectory length
3. For all states, get the miminal number of steps (across those K trajectories) it takes the agent to get to those states. Denote this by N_{s_ρ} for each s

Then:

$$\begin{aligned}\widehat{\text{Need}}(\langle s_k, b_k \rangle) =& \gamma^{N_{s_\rho}} P(\langle s_{\text{agent}}, b_{\text{agent}} \rangle \to \langle s_\rho, b_{\text{agent}} \rangle, N_{s_\rho}, \pi_{\text{old}}) \times \gamma^h P(\langle s_\rho, b_{\text{agent}} \rangle \to \langle s_k, b_k \rangle, h, \pi_{\text{old}}) \\ &+ \sum_{i=N_{s_\rho}+H+1}^{\infty} \gamma^i P(\langle s_{\text{agent}}, b_k \rangle \to \langle s_k, b_k \rangle, i, \pi_{\text{old}})\end{aligned}$$

Do we need s_ρ? POCMP builds a partial tree (in the sense that not all histories are evaluated, but only the ones chosen by the tree policy)

One way of estimating Need is therefore to:

1. Simulate K forward trajectories from the agent's current information state $z_{\text{agent}} = \langle s_{\text{agent}}, b_{\text{agent}} \rangle$ using the old policy π_{old}. Update the belief along the way.
2. Terminate each trajectory if $\gamma^d < \epsilon$ where d is the trajectory length
3. Estimate the probability of reaching $z_k = \langle s_k, b_k \rangle$ and the average number of steps to reach it – denote this by $N(z_k)$
4. Estimate Need as:

$$\widehat{\text{Need}}(z_k) = \gamma^{N(z_k)} P(z_{\text{agent}} \to z_k, N(z_k), \pi_{\text{old}}) + \sum_{i=N(z_k)+1}^{\infty} \gamma^i P(\langle s_{\text{agent}}, b_k \rangle \to \langle s_k, b_k \rangle, i, \pi_{\text{old}})$$

Problems: i) this is not going to scale very well to larger problems; and ii) simulations are still done on-policy, for estimating Need under π_{old}.