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:

$$\text{EVB}(z_k, a_k) = \sum_{z \in \mathcal{Z}} \underbrace{\sum_{i=0}^{\infty} \gamma^i P_{0..i}(z_{\text{agent}} \to z, i, \pi_{\text{old}})}_{\text{Need}} \times \underbrace{\sum_a \left[\pi_{\text{new}}(z,a) - \pi_{\text{old}}(z,a)\right] q_{\pi_{\text{new}}}(z,a)}_{\text{Gain}} \tag{1}$$

where $P_{0..i}$ denotes transition models which are updated according to the successful transitions.

Note that M&D get rid of the first summation (in their case, it was over $\mathcal{S}$), since they assume that a single update changes the policy only at that update location (and hence Gain is 0 everywhere except that location). In our case, this is not quite true, because there are a number of information states that can potentially benefit from a policy update elsewhere – and thus ideally we want to keep that sum.

Our current model has a number of assumptions. For the updates of distal information states, we have:

$$\text{EVB}_{\text{dist}}(z_k = \langle s_k, b_k \rangle, a_k) = \gamma^i P_{0..i}(\langle s_\rho, b_{\text{agent}} \rangle \to \langle s_k, b_k \rangle, i, \pi_{\text{old}}) \times \text{Gain}(\langle s_k, b_k \rangle, a_k) \tag{2}$$

where $\langle s_\rho, b_{\text{agent}} \rangle$ is the information state at the root of the state-action tree which contains information state $z_k = \langle s_k, b_k \rangle$. The assumption is therefore that the agent can 'teleport' from its current physical state s_{agent} to the physical root state of that tree s_ρ without changing its belief b_{agent}.

As for the root updates, we have the following:

$$\text{EVB}_{\text{root}}(z_\rho = \langle s_\rho, b_\rho \rangle, a_k) = \underbrace{\sum_{i=0}^{\infty} \gamma^i P_0(\langle s_{\text{agent}}, b_{\text{agent}} \rangle \to \langle s_\rho, b_{\text{agent}} \rangle, i, \pi_{\text{old}})}_{\text{certainty-equivalent Need}} \times \underbrace{\text{Gain}_{\text{sweep}}(\langle s_\rho, b_{\text{agent}} \rangle, a_k)}_{\text{sweeped Gain}} \tag{3}$$

Where we partially account for the (fixed horizon) learnt information by our sweeping procedure when estimating Gain (by using equation 2). Partially because this isn't reflected in the Need term since we do not account for the changes in the transition model when calculating Need (hence P_0).

Another limitation is the lack of generalisaton across information states because i) we exclude the summation over $\mathcal{Z}$ in equation 1; and ii) the updates at distal information states are rooted in that tree – hence the potential information gain cannot be communicated to other information states rooted in other states.

The certainty-equivalence assumption right from the agent's curent physical location s_{agent} seems a bit odd, since we learn something about the transition model and then simply ignore it. My suggestion is to change how we compute Need to account for at least a fixed-horizon learning, and then do the certainty-equivalence approximation. This would be more consistent with how we calculate Gain, and in fact then we can have a single equation for both distal and root updates.

The idea is essentially to have two forms of certainty-equivalent Need. The first one would be a non-cumulative Need of reaching some root state s_ρ from the agent's current physical location s_{agent}. Then, from that root state we plan up to some fixed horizon, and finish off Need with certainty-equivalent Need based on the belief reached during that planning. Formally, this would look like this:

$$\begin{aligned}
\text{Need}(z_k = \langle s_k, b_k \rangle) = &\sum_{i=0}^{N} \gamma^i P_{0..i}(\langle s_{\text{agent}}, b_{\text{agent}} \rangle \to \langle s_k, b_{\text{agent}} \rangle, i, \pi_{\text{old}}) && \text{get to the root, fixed belief} \\
&+ \sum_{j=N}^{N+H} \gamma^j P_{N..j}(\langle s_{\text{agent}}, b_{\text{agent}} \rangle \to \langle s_k, b_k \rangle, j, \pi_{\text{old}}) && \text{consequitive belief updates} \\
&+ \sum_{k=N+H+1}^{\infty} \gamma^k P_{k..(N+H+1)}(\langle s_{\text{agent}}, b_k \rangle \to \langle s_k, b_k \rangle, k, \pi_{\text{old}}) && \text{fixed, updated belief}
\end{aligned}$$

where N is the expected number of steps needed to reach the root state s_ρ from the agent's current location s_{agent} (we need this to know how to discount the subsequent learning) and H is the planning horizon.

This has the advantage that we gain access to all information states and thus can account for the missing generalisaton.

The issue is of course estimating N. There is this recent 'First-occupancy representation' from Maneesh: https://arxiv.org/abs/2109.13863 – although it is not very useful since we need to know not the expected discount until N but the subsequent discount. That's why I've been thinking about the ensemble SR recently – although then it would be hard to justify the use of a particular γ.

- An interesting question is whether it's worth doing it – this would depend on the discount γ and/or the distance, since the contributions of future learning fade away. Perhaps not so important for our problem but can conceive of domains where it would be crucial?
- Makes me think – maybe we can derive the approximation error due to a limited horizon? E.g., some notion of 'information regret'