Partially Observable Markov Decision Process Modelling for Assessing Hierarchies: Supplemental Document

Weipeng Huang WEIPENG.HUANG@INSIGHT-CENTRE.ORG
Insight Centre for Data Analytics, University College Dublin, Ireland

Guangyuan Piao GUANGYUAN.PIAO@INSIGHT-CENTRE.ORG
Insight Centre for Data Analytics, National University of Ireland, Galway

Raul Moreno RAUL.MORENO.SALINAS@GMAIL.COM
Neil Hurley NEIL.HURLEY@INSIGHT-CENTRE.ORG
Insight Centre for Data Analytics, University College Dublin, Ireland

Editors: Sinno Jialin Pan and Masashi Sugiyama

Appendix A. Example of A POMDP for a Hierarchy

Consider a search over the three-node hierarchy with only three nodes, the root node c_0 and its two children c_1 and c_2. It contains eight possible states:

$$\langle c_0, 1\rangle, \langle c_0, 0\rangle, \langle c_1, 1\rangle, \langle c_1, 0\rangle, \langle c_2, 1\rangle, \langle c_2, 0\rangle, \langle \emptyset, 1\rangle, \langle \emptyset, 0\rangle \ .$$

The POMDP for this simple tree yields belief states b_{c_0}, b_{c_1}, b_{c_2} when the bot is at the corresponding node, and the trivial belief states at the fully observed terminal states $\langle \emptyset, 1\rangle$ and $\langle \emptyset, 0\rangle$. The bot moves using the guidance function values $\tilde{\eta} \triangleq \eta(c_0, c_1)$ and $\eta(c_0, c_2) = 1 - \tilde{\eta}$ to determine the next node when the action a_d is selectedguid The set of reachable belief states is represented in an AND-OR tree in Fig. 1 (see a similar figure in (Ross et al., 2008)). In this figure, an action must be chosen at an OR node, the choice of which leads to the set belief states, over all possible observations, that must be considered at the AND nodes. Expected rewards, $R(b, a)$ are represented on the arcs from OR- to AND-nodes, while the transition probabilities $p(o \mid b, a)$ are represented on the arcs from AND- to OR-nodes. Working from the leaf nodes back to the root, we can read from the tree that action a_s at node c_0, would lead to an expected reward of

$$Q(b_{c_0}, a_s) = b_{c_0} r(c_0) + (1 - b_{c_0})(-1) = b_{c_0}(r(c_0) + 1) - 1$$

while action a_d at c_0 would lead to an expected reward of

$$Q(b_{c_0}, a_d) = b_{c_0}\left[\tilde{\eta}^2(r(c_1) + 1) + (1 - \tilde{\eta})^2(r(c_2) + 1)\right] - 1$$

where the expression comes from $b_{c_1} = \tilde{\eta} b_{c_0}$ and $b_{c_2} = (1 - \tilde{\eta}) b_{c_0}$ via Eq. (1).

$$b'\left(\left\langle c', 1\right\rangle\right) = \eta(c, c') b\left(\langle c, 1\rangle\right) \qquad b'\left(\left\langle c', 0\right\rangle\right) = 1 - \eta(c, c') b\left(\langle c, 1\rangle\right) \ . \tag{1}$$

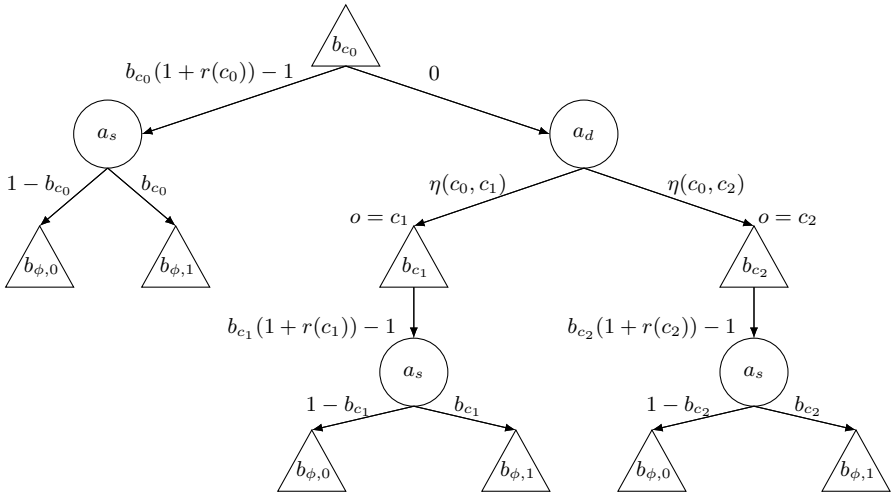

Figure 1: An AND-OR tree of reachable belief states from node c_0 of the three-node hierarchy

Appendix B. Insights into the Guidance Function

Recall that the guidance function is

$$\eta(c, c', x) = p(x \in c' \mid x \in c) \triangleq \frac{\exp\{\text{sim}(x, c')/\delta\}}{\sum_{c'' \in \mathcal{C}(c)} \exp\{\text{sim}(x, c'')/\delta\}}\,. \tag{2}$$

This can be any kind of similarity that is used in measuring clustering results. Leaving the choice of similarity function open adds flexibility to the measure by allowing the users to customise the comparison for various hierarchies of specific datasets. A simple choice is the inverse or the negative of Euclidean distance, assuming items can be mapped to points in $\mathbb{R}^n$; Bayesian models could use a similarity based on a distribution density; if items are represented as Term Frequency and Inverse Document Frequency (TF-IDF) vectors of textual data, then cosine similarity might be appropriate; and so on. The η function should return a higher probability score for the node c' that is "closest" to x among the siblings. The parameter δ is a *temperature* parameter in the Boltzmann function.

Suppose there are two child clusters A and B of parent C such that $\text{sim}(x, A) < s_\epsilon$ and $\text{sim}(x, B) < s_\epsilon$ for some $s_\epsilon \approx 0$, and yet $\text{sim}(x, A) \ll \text{sim}(x, B)$. For example, suppose the similarities of x to random cluster A and cluster B, respectively, return $1e-50$ and $1e-70$. It would be better to have $\eta(C, A) \approx \eta(C, B)$, rather than $\eta(C, A) \approx 1$, for two such clusters. Deep in the hierarchy, the bot should only choose to descend to a cluster when the evidence that it contains the target is strong. A δ parameter that increases with the depth of the tree ensures that the similarity values between two clusters must be increasingly more distinct deeper in the tree before one cluster is preferred over another.

Thus, δ_t can be defined as a function over the depth, t, of the hierarchy, s.t. $\delta_t \triangleq \delta\nu^t$ where $\nu \in [1, \infty)$. Setting $\nu = 1$ makes δ_t invariant with depth.

Appendix C. Policy

This section is devoted to two parts. The first is to show the complete version of the Real-Time Belief Space Search (RTBSS). Then, we analyse the complexity.

C.1. RTBSS

Algorithm 1 demonstrates the original RTBSS procedure as presented in (Ross et al., 2008). It heavily relies on the function Expand(b, a) in Algorithm 2 to explore the POMDP. The Boolean function IsLeaf(b) returns `true` if the only belief states reachable from b with non-zero probability are the terminal states. RTBSS is a greedy algorithm that explores a lower-bound on the optimal $V(b)$ using a diversity of actions within a limited number of look-aheads and selects the policy that maximises this lower-bound. Considering that the look-aheads are capped, this can also be described as a myopic policy and so follows the proposal in (Fern et al., 2007; Ie et al., 2019) to use myopic heuristics for approximating the Q value for each belief-action pair to alleviate the intractable computations in a POMDP.

Algorithm 1 RTBSS

Require: d, the maximum look-aheads which is fixed to 2 in our settings
Ensure: π, the policy function
Initialise b
repeat
  $L, a \leftarrow$ Expand(b, d)
  $\pi(b, a) \leftarrow 1$
  Execute a and perceive o
  $b \leftarrow \tau(b, a, o)$
until IsLeaf(b)

Algorithm 2 Expand

Require: b, the current belief state
Require: d, the number of levels to explore, must be ≥ 0
Ensure: L^*, optimal lower bound
Ensure: a^*, optimal action
$L^* \leftarrow -\infty$
if $d = 0$ or IsLeaf(b) **then**
  $L(a) \leftarrow R(b, a)$
else
  for $a \in A$ **do**
    $L(a) \leftarrow R(b, a) + \gamma \sum_{o \in O} p(o \mid b, a)$Expand$(\tau(b, a, o), d - 1)$
  end for
end if
$L^* \leftarrow \max_a L(a)$
$a^* \leftarrow \operatorname{argmax}_a L(a)$

C.2. Time Complexity of HQS

Proposition 1 *The worst case complexity of computing HQS for N items is $O(N^3\mathcal{F}(\text{sim}))$ where $\mathcal{F}(\text{sim})$ is the complexity of the similarity function.*

Proof [Sketch] Solving a finitely horizontal POMDP is PSPACE-complete (Papadimitriou and Tsitsiklis, 1987), while luckily it does not apply to our case. Consider the HQS calculation for a single target x. Let $\mathcal{F}(\text{sim})$ represent the complexity of calculating the similarity between x and a cluster. Let us reasonably assume that each non-root node has at least one sibling in the hierarchy. For the data with N entries and corresponding hierarchy with M^* nodes, the maximum M^* is $2N-1$. This holds when the tree splits one data point as a leaf and all others remain as one cluster, until all points become leaves.

Least optimally, the searcher needs to estimate the return at all nodes for a certain target, which will be in $O(M^*)$. However, the policy can still be pruned as reaching a wrong node finally receives the reward -1. Accordingly the reward computation can concentrate on the path wherein each node contains the target. Denote the number of children of the t^{th} parent in the right path by N_t. The complexity will then follow $O(\sum_t N_t) = O(M^*N)$ which is thus $O(N^2)$. Hereafter, the guidance function requires $O(N^2\mathcal{F}(\text{sim}))$ computations given that $O(\mathcal{F}(\text{sim}))$ is the complexity for calculating the similarity for a data point to a cluster—which will be polynomial for most commonly used similarity choices. ■

Nevertheless, the average case for the height of a tree is always logarithmic. We can write the average complexity of searching for a target as $O(a \log_a M) = O(a \log_a N)$ where a is a constant for the number of children. The average case for the HQS is therefore $O(a \log_a N \cdot N\mathcal{F}(\text{sim})) = O(N \log N\mathcal{F}(\text{sim}))$. The polynomial result concludes that HQS is practically applicable.

Appendix D. Experimental Setup

D.1. For the Implementation of HQS

We show some numerical details about the items in Table S1 and S2. We refer to items by their indices as indicated in the tables. The left column is the title of the item, and the second column contains the top 10 terms in the title and the description of the item, with the corresponding TF-IDF score in parentheses. As the examples all come from the large fashion category, the TF-IDF score is computed only on this subset of the `Amazon` data. However, when computing the similarities, given the small number of items, we still consider all the features without any feature selection techniques .

D.2. For Scaling the Approach

For the two experiments, we adopt the following similarity for use in the guidance function

$$\mathcal{S}(x, c) = \frac{1}{\|\boldsymbol{v}_x - \bar{\boldsymbol{v}}_c\|^2 + 0.0001}\,.$$

For `Amazon` after PCA applied, we set $\delta = \lceil \frac{1}{100} \rceil = 0.01$ where 100 is the number of dimensions that we keep.

index	short name	top 10 terms
0	Women Boots	Harness (0.2455), term (0.2186), long (0.1984), Womens (0.181), durability (0.1709), boots (0.1647), inch (0.1547), Crushed (0.1514), element (0.1465), tougher (0.1465),
1	Shoe Cream	Meltonian (0.2913), waxes (0.2768), cream (0.2169), cloth (0.1914), afterwards (0.1653), Misc (0.158), staining (0.1489), honored (0.1489), creamy (0.1489), terrific (0.1489),
2	Women Runner	Ascend (0.4741), Wave (0.3866), MIZUNO (0.237), EU (0.2266), SZ (0.2135), lends (0.2089), Mizuno (0.2049), UK (0.1822), Running (0.1822), China (0.1659),
3	Whitener	Whitener (0.5887), Sport (0.3376), chalky (0.2943), restores (0.2502), KIWI (0.2324), formula (0.218), scuffs (0.218), polish (0.1974), covers (0.1974), Kiwi (0.1925),
4	Sneaker	Coach (0.4753), signature (0.2487), leather (0.2173), preeminent (0.1759), Poppy (0.1759), Barrett (0.1759), emerged (0.1759), resulting (0.1682), Scribble (0.1682), coach (0.1682),
5	Biker Boot Straps	to (0.2226), 6in (0.2192), are (0.2111), clips (0.1902), in (0.1884), Straps (0.188), sold (0.1716), prevent (0.1564), Boot (0.15), SP6 (0.1402),

Table S1: Amazon item scores 1

References

Alan Fern, Sriraam Natarajan, Kshitij Judah, and Prasad Tadepalli. A Decision-Theoretic Model of Assistance. In *IJCAI*, pages 1879–1884, 2007.

Eugene Ie, Vihan Jain, Jing Wang, Sanmit Narvekar, Ritesh Agarwal, Rui Wu, Heng-Tze Cheng, Morgane Lustman, Vince Gatto, Paul Covington, Jim McFadden, Tushar Chandra, and Craig Boutilier. Reinforcement Learning for Slate-based Recommender Systems: A Tractable Decomposition and Practical Methodology. Technical report, 2019. arXiv preprint arXiv:1905.12767 [cs.LG].

Christos H Papadimitriou and John N Tsitsiklis. The Complexity of Markov Decision Processes. *Mathematics of operations research*, 12(3):441–450, 1987.

Stéphane Ross, Joelle Pineau, Sébastien Paquet, and Brahim Chaib-Draa. Online Planning Algorithms for POMDPs. *Journal of Artificial Intelligence Research*, 32:663–704, 2008.

index	short name	top 10 terms
6	Jewel Solution	cleaning (0.2535), components (0.2389), metals (0.2297), precious (0.1925), solution (0.1867), free (0.1512), accumulate (0.1482), Biodegradable (0.1482), titanium1 (0.1482), injectors (0.1482),
7	Dye Kit	TRG (0.4005), included (0.2872), Turquoise (0.2277), Detailed (0.2141), Everything (0.2117), coats (0.2105), dye (0.1958), instructions (0.1937), Self (0.1912), Dye (0.1852),
8	Silver Cloth	silver (0.3919), tarnishing (0.3315), tarnish (0.189), Silvershield (0.1713), Cadet (0.1713), shining (0.1414), drawer (0.1414), trade (0.1389), Tarnish (0.1389), will (0.1277),
9	Woman Trainer	Ryka (0.5414), Rythmic (0.406), Womens (0.1547), the (0.148), Athena (0.1353), cardio (0.1353), fittest (0.1353), Rhythmic (0.1353), kickboxing (0.1353), gain (0.1294),
10	Ultrasonic Cleaner	Professional (0.3063), ultrasonic (0.2914), grade (0.2771), NUMWPT (0.2082), Qt (0.2082), Joy4Less (0.2082), automotive (0.1925), transducer (0.1875), Heater (0.1875), controls (0.1834),
11	Boot Socks	dead (0.3354), Cuffs (0.3036), gorgeous (0.2605), drop (0.2396), lace (0.2396), season (0.2363), Add (0.2348), cuffs (0.234), put (0.2248), Socks (0.2194),

Table S2: Amazon item scores 2