Algorithm 1: New Anchor Beliefs (N = 50 in all experiments)

Data: S, G, \(\bar{V}(G)\), OCF, N, Q- in augmented space

Gnew = \{ \}

for all \(o \in \mathcal{O}\) do

\(b \leftarrow b(o)\)

end repeat

while \(b \notin G \cup Gnew\) do

add \(b\) into \(Gnew\)

return \(Gnew\)

Upper Bounds for POMDPs

- MDP: \(Q(s,a) = R(s) + \gamma \sum_{s'} T(s,a,s') \max_{a'} Q(s',a') \forall s,a\)
- QMDP: \(Q(s,a) = R(s) + \gamma \sum_{s'} T(s,a,s') \max_{a'} Q(s',a') \forall s,a\)
- FIB: \(Q(s,a) = R(s) + \gamma \sum_{s'} \max_{a'} \sum_{b \in \mathcal{B}} \bar{V}(b(s))Q(s,a') \forall s,a\)

Augmented POMDPs

- Add \(m\) interior beliefs to the set of \(n\) states of the original POMDP
- An initial belief \(P_{0}(b) = c\|b\) corresponds to interpolation of \(b_{0}\) by the convex combination \(c\) of anchor beliefs

Avoiding Lookahead

Observe that the convex combination of \(b\) can be seen as its embedding in the augmented space (\(c\) becomes a belief in the augmented space), and the policy can be queried directly.

AO-deterministic POMDPs

- Deterministic POMDPs in Littman’s thesis have deterministic \(T\) and \(Z\) (all probabilities are either zero or one)
- Quasi-deterministic POMDPs have deterministic \(T\) (Besse and Chaib-draa 2009)
- We introduce AO-deterministic POMDPs when all \(T_{a,b}\) matrices have at most one non-zero entry in every row—actions can be stochastic!
- All deterministic and quasi-deterministic POMDPs are AO-deterministic, but there exist POMDPs that are AO-deterministic but are neither deterministic nor quasi-deterministic (e.g. baseball)
- AO-deterministic POMDPs are AO-deterministic, e.g., rockSample-7\_8 and underwaterNav

Why AO-deterministic definition matters?

- The key conclusion: search for new beliefs going forward from corners as well (not only from \(b_{0}\) as it is the case in GapMin, HSVI, or SARSOP)

Our Algorithm

Theorem: Policies that are optimal for the underlying MDP of an AO-deterministic POMDP are also optimal at the corner beliefs of this POMDP.

Results—Execution Time and Quality of Upper Bound Policies

Table: The quality of upper bounds (UB) after 1000 seconds of planning (AO-deterministic POMDPs).

Price point algorithm

Table: The quality of upper bounds (UB) after 1000 seconds of planning (non AO-deterministic POMDPs).

Conclusion

- Efficient execution of upper bound policies (e.g. in an augmented space) was shown—useful for designing upper bound policies or giving them to guide branch-and-bound
- AO-deterministic POMDPs generalise existing definitions of deterministic and quasi-deterministic POMDPs, yet are specific enough to explain the process of refining upper bounds and to show where the augmented POMDP is converging to
- AO-deterministic POMDPs lead to a straightforward approach that can compute the tightest upper bounds without any use of lower bounds

Acknowledgements

This research was sponsored by NSERC and MITACS.