POMDP Planning and Execution in an Augmented Space

Marek Grześ and Pascal Poupart

David Cheriton School of Computer Science, University of Waterloo, Canada

WATERLOO CHERITON SCHOOL OF COMPUTER SCIENCE

AAMAS-14

Motivation for Investigating Upper Bounds

- Most point-based value iteration as well as branch-and-abound algorithms (including online planning) guide their optimisation by upper bounds
- There is growing interest in performance guarantees to estimate how far from optimal a policy can be; helps to check if a model fits a particular application
- An upper bound policy can be good and methods of fast execution are desirable
- Upper bounds are hard to improve; better understanding and methods are required

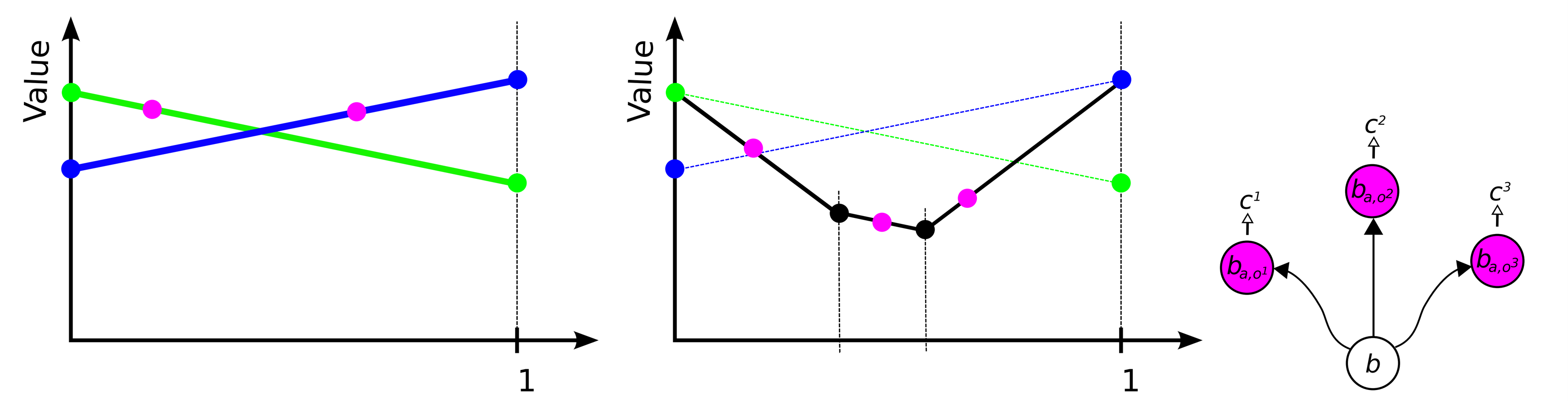
POMDPs and their $T_{a,o}$ Matrices

- $\langle S, A, O, T, Z, R, b_0, \gamma \rangle$
- $T_{a,o} = T_a Z_a(o) = \begin{array}{c|ccc} & s'_1 & \cdots & s'_n \\ s_1 & P(s'_1, o|a, s_1) & \cdots & P(s'_n, o|a, s_1) \\ \cdots & \cdots & \cdots & \cdots \\ s_n & P(s'_1, o|a, s_n) & \cdots & P(s'_n, o|a, s_n) \end{array}$
- $T_a = \begin{array}{c|ccc} & s'_1 & \cdots & s'_n \\ s_1 & t_{1,1} & \cdots & t_{1,n} \\ \cdots & \cdots & \cdots & \cdots \\ s_n & t_{n,1} & \cdots & t_{n,n} \end{array}$
- $Z_a = \begin{array}{c|ccc} & o_1 & \cdots & o_k \\ s'_1 & p_{1,1} & \cdots & p_{1,k} \\ \cdots & \cdots & \cdots & \cdots \\ s'_n & p_{n,1} & \cdots & p_{n,k} \end{array}$

Upper Bounds for POMDPs

- MDP: $Q(s,a) = R_a(s) + \gamma \sum_{s'} T_a(s,s') \max_{a'} Q(s',a') \; \forall s,a$
- QMDP: $Q(s,a) = R_a(s) + \gamma \sum_o \sum_{s'} T_{a,o}(s,s') \max_{a'} Q(s',a') \; \forall s,a$
- FIB: $Q(s,a) = R_a(s) + \gamma \sum_o \max_{a'} \sum_{s'} T_{a,o}(s,s') Q(s',a') \; \forall s,a$

Upper Bounds for Arbitrary Beliefs



$$\bar{V}(b) = \max_a \sum_s b(s)Q(s,a)$$

$$\pi(b) = \arg\max_a \sum_s b(s)Q(s,a)$$

$$\pi^G(b) = \arg\max_a \left\{ \sum_s b(s)R(s,a) + \gamma \sum_{o \in O} P(o|b,a)\bar{V}^G(b_{a,o}) \right\}$$

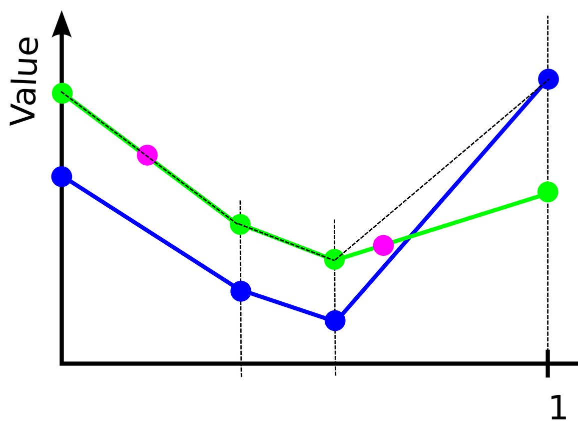
Augmented POMDPs

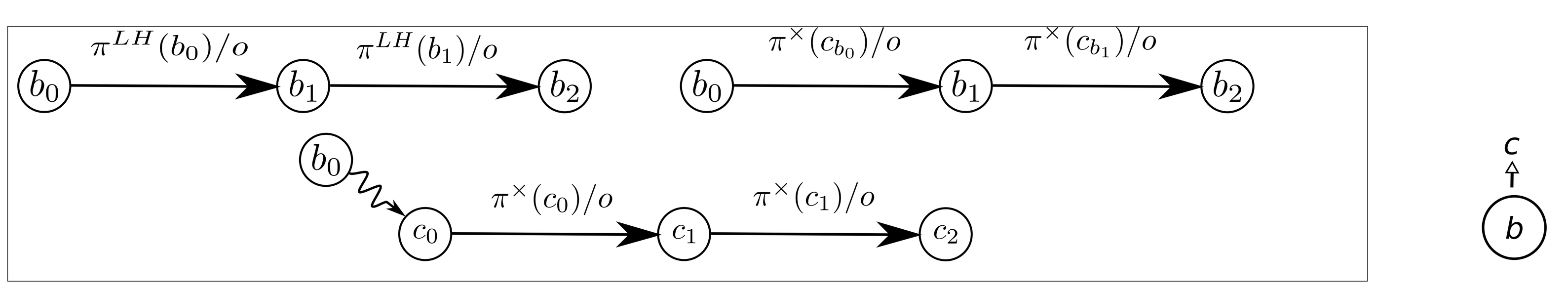
- Add m interior beliefs to the set of n states of the original POMDP
- An initial belief $Pr_0(b) = c(b)$ corresponds to interpolation of b_0 by the convex combination c of anchor beliefs
- $T_{a,o}(b, b') = P(b', o|a, b) = c(b')Z_a(o|b)$

$$T_{a,o} = \begin{array}{c|cccccc} & s'_1 & \cdots & s'_n & b'_{n+1} & \cdots & b'_{n+m} \\ s_1 & c_{1,1} & \cdots & c_{1,n} & c_{1,n+1} & \cdots & c_{1,n+m} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ s_n & c_{n,1} & \cdots & c_{n,n} & c_{n,n+1} & \cdots & c_{n,n+m} \\ b_{n+1} & c_{n+1,1} & \cdots & c_{n+1,n} & c_{n+1,n+1} & \cdots & c_{n+1,n+m} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ b_{n+m} & c_{n+m,1} & \cdots & c_{n+m,n} & c_{n+m,n+1} & \cdots & c_{n+m,n+m} \end{array}$$

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 $T_{a,o}$ 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)
- A few other benchmarks form ICAPS-IPPC are AO-deterministic, e.g., rockSample-7_8 and underwterNav

Why AO-deterministic definition matters?

- $T_{a,o} = \begin{array}{c|cccccc} & s'_1 & \cdots & s'_n & b'_{n+1} & \cdots & b'_{n+m} \\ s_1 & c_{1,1} & \cdots & c_{1,n} & c_{1,n+1} & \cdots & c_{1,n+m} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ s_n & c_{n,1} & \cdots & c_{n,n} & c_{n,n+1} & \cdots & c_{n,n+m} \\ b_{n+1} & c_{n+1,1} & \cdots & c_{n+1,n} & c_{n+1,n+1} & \cdots & c_{n+1,n+m} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ b_{n+m} & c_{n+m,1} & \cdots & c_{n+m,n} & c_{n+m,n+1} & \cdots & c_{n+m,n+m} \end{array}$
- If $b_{a,o}$ is a state of the augmented POMDP, then the row for (b, a, o) has at most one non-zero entry—$T_{a,o}$ is becoming "more deterministic" when upper bounds are improved

- 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

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

```
Data: S, G, V̄^G, OCF, N, Q- in augmented space
G_new ← ∅
if POMDP is AO-deterministic then
    for i=1 to N do
        if b_0 ∈ G then
            return G_new;                         /* nothing to improve */
        else
            b ← ForwardSearch or LAO*
            add b into G_new
else
    H ← SampleCorners(G, OCF, N) ;   /* sample among corners with non-deterministic transitions only */
    for all corner beliefs b ∈ H do
        repeat
            c ← embed b into augmented space
            a* ← action for c using augmented Q-values
            sample observation o according to P(o|b, a*)
            b ← b_{a,o}
        until b ∉ G ∪ G_new
        add b into G_new
return G_new
```

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

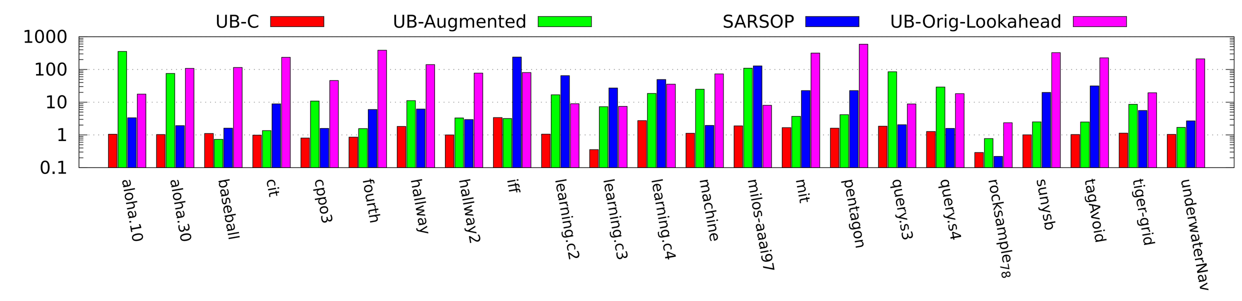


Figure: Ratio of the execution time to QMDP execution time

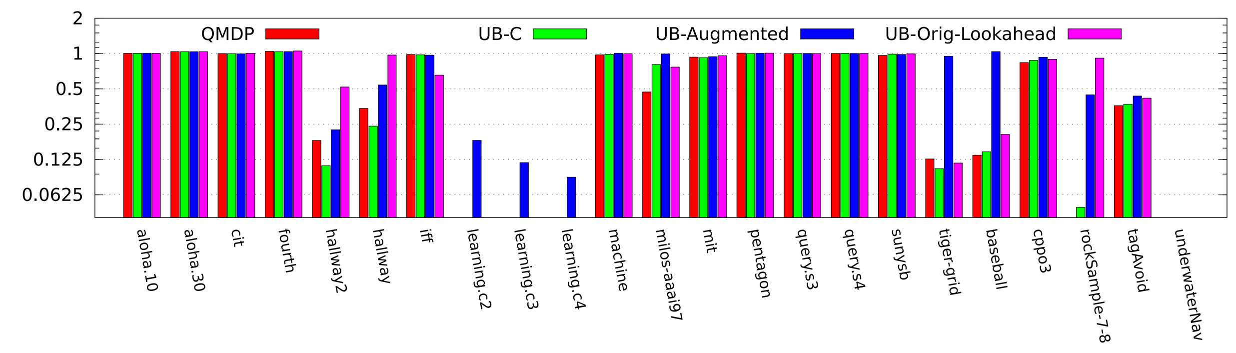


Figure: Ratio of simulated quality to SARSOP lower bound policies

Results—AO-deterministic and non AO-deterministic POMDPs

problem	algorithm	gap	LB	UB	$\lvert\Gamma\rvert$	$\lvert\bar{V}\rvert$	time	UB	$\lvert\bar{V}\rvert$	time
baseball	hsvi2	1e-3	0.6412	0.6412	991	n.a.	999			
$\lvert\mathcal{S}\rvert = 7681$	sarsop	7e-4	0.6412	0.6419	1453	1694	400	0.6412	3878	2346
$\lvert\mathcal{A}\rvert=6$ $\lvert\mathcal{O}\rvert=9$	GapMin	5.01	0.6346	5.6500	1	1	281	0.6434	52	15219
$\gamma = 0.999$	Aug-OCF			0.6413		3051	970			
rockSample-7_8	hsvi2	3.56	20.91	24.46	4752	n.a.	998			
$\lvert\mathcal{S}\rvert = 12545$	sarsop	4.12	20.91	25.02	3119	2473	999	24.46	8520	9806
$\lvert\mathcal{A}\rvert=13$ $\lvert\mathcal{O}\rvert=2$	GapMin	25.07	7.35	32.42	1	1	6.18	26.84	30	13855
$\gamma = 0.950$	Aug-OCF			24.81		3351	978			
underwaterNav	hsvi2	23.4	729.9	753.3	3545	n.a.	1000			
$\lvert\mathcal{S}\rvert = 2653$	sarsop	23.4	731.0	754.4	7918	2820	999	754.0	7947	10014
$\lvert\mathcal{A}\rvert=6$ $\lvert\mathcal{O}\rvert=103$	GapMin	80.2	675.06	755.3	1	1	742	754.8	115	10113
$\gamma = 0.950$	Aug-OCF			754.6		1830	471.0			

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

problem	algorithm	gap	LB	UB	$\lvert\Gamma\rvert$	$\lvert\bar{V}\rvert$	time	UB	$\lvert\bar{V}\rvert$	time
aloha.10	hsvi2	9.0	535.4	544.4	4729	n.a.	997	544.1	n.a.	10001.4
$\lvert\mathcal{S}\rvert = 30$	sarsop	9.5	535.2	544.7	48	2151	1000	544.3	8035	10000.5
$\lvert\mathcal{A}\rvert = 9$, $\lvert\mathcal{O}\rvert = 3$	GapMin	10.7	533.5	544.2	81	223	972	544.0	1140	10741.3
$\gamma = 0.999$	Aug-H			539.6 ± 0.01		1999.1 ± 21.7	981.9 ± 3.6			
	Aug-OCF			539.0 ± 0.01		3345 ± 22.8	984.5 ± 2.8			
hallway2	hsvi2	0.5250	0.3612	0.8862	2393	n.a.	997	0.8696	n.a.	10003.1
$\lvert\mathcal{S}\rvert = 92$	sarsop	0.5247	0.3737	0.8984	262	1519	992	0.8877	4029	10002.5
$\lvert\mathcal{A}\rvert = 5$, $\lvert\mathcal{O}\rvert = 17$	GapMin	0.4495	0.3497	0.7992	122	218	835.5			
$\gamma = 0.950$	Aug-H			0.897 ± 0.0		1349.6 ± 11.5	896.2 ± 17.6			
	Aug-OCF			0.805 ± 0.0		861.0 ± 6.3	944.1 ± 12.1			
hallway	hsvi2	0.250	0.945	1.195	1367	n.a.	996	1.185	n.a.	10026.6
$\lvert\mathcal{S}\rvert = 60$	sarsop	0.210	0.995	1.206	456	1713	998	1.196	5117	10002.6
$\lvert\mathcal{A}\rvert = 5$, $\lvert\mathcal{O}\rvert = 21$	GapMin	0.132	0.989	1.122	94	176	974	1.091	344	2035.3
$\gamma = 0.950$	Aug-H			1.186 ± 0.0		1189.7 ± 13.0	947.1 ± 13.3			
	Aug-OCF			1.095 ± 0.0		951.0 ± 7.0	946.1 ± 11.5			
machine	hsvi2	3.49	63.18	66.66	662	n.a.	982	66.34	n.a.	10003.5
$\lvert\mathcal{S}\rvert = 256$	sarsop	3.57	63.18	66.75	150	2742	998	66.4	9846	10004.6
$\lvert\mathcal{A}\rvert = 4$, $\lvert\mathcal{O}\rvert = 16$	GapMin	3.48	62.38	65.87	58	208	898	64.64	1174	12147.0
$\gamma = 0.990$	Aug-H			64.68 ± 0.0		972.0 ± 0.0	809.0 ± 4.1			
	Aug-OCF			63.84 ± 0.01		965.0 ± 12.3	918.7 ± 17.1			
query.s4	hsvi2	51.9	569.5	621.4	2846	n.a.	999	620.4	n.a.	10002.9
$\lvert\mathcal{S}\rvert = 81$	sarsop	54.3	569.1	623.4	166	6782	1000	622.8	23742	10014.1
$\lvert\mathcal{A}\rvert = 4$, $\lvert\mathcal{O}\rvert = 3$	GapMin	45.0	569.4	614.5	137	631	956	605.2	2945	13881.0
$\gamma = 0.990$	Aug-H			589.4 ± 0.06		1660.5 ± 12.8	892.0 ± 3.6			
	Aug-OCF			586.4 ± 0.03		1871 ± 10.4	949.6 ± 5.7			
tagAvoid	hsvi2	3.207	-6.150	-2.943	2896	n.a.	1000	-3.378	n.a.	10001.3
$\lvert\mathcal{S}\rvert = 870$	sarsop	3.455	-6.142	-2.686	9324	8049	989	-3.298	18099	10085.4
$\lvert\mathcal{A}\rvert = 5$, $\lvert\mathcal{O}\rvert = 30$	GapMin	12.70	-14.0	-1.291	77	310	773	-2.436	1800	10017.0
$\gamma = 0.950$	Aug-H			-0.672 ± 0.0		5840.3 ± 55.8	949.0 ± 5.5			
	Aug-OCF			-3.660 ± 0.0		6861.0 ± 50.8	990.5 ± 1.4			
cppo3	hsvi2	10.89	12.96	23.84	3773	n.a.	999	23.83	n.a.	10004.5
$\lvert\mathcal{S}\rvert = 180$	sarsop	9.69	14.69	24.38	242	3420	998	24.38	8879	10053.8
$\lvert\mathcal{A}\rvert = 6$, $\lvert\mathcal{O}\rvert = 6$	GapMin	6.87	15.43	22.30	497	1495	976	21.66	1624	14156.6
$\gamma = 0.900$	Aug-H			21.28 ± 0.01		2808.0 ± 27.8	920.6 ± 23.3			
	Aug-OCF			20.71 ± 0.03		1221 ± 34.7	937 ± 12			

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 deploying upper bound policies or using 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.

e-mail: {mgrzes|ppoupart}@cs.uwaterloo.ca

www: http://www.cs.uwaterloo.ca/~mgrzes