Closing the Gap: Improved Bounds on Optimal POMDP solutions

Pascal Poupart (U of Waterloo)

Kee-Eung Kim (KAIST)

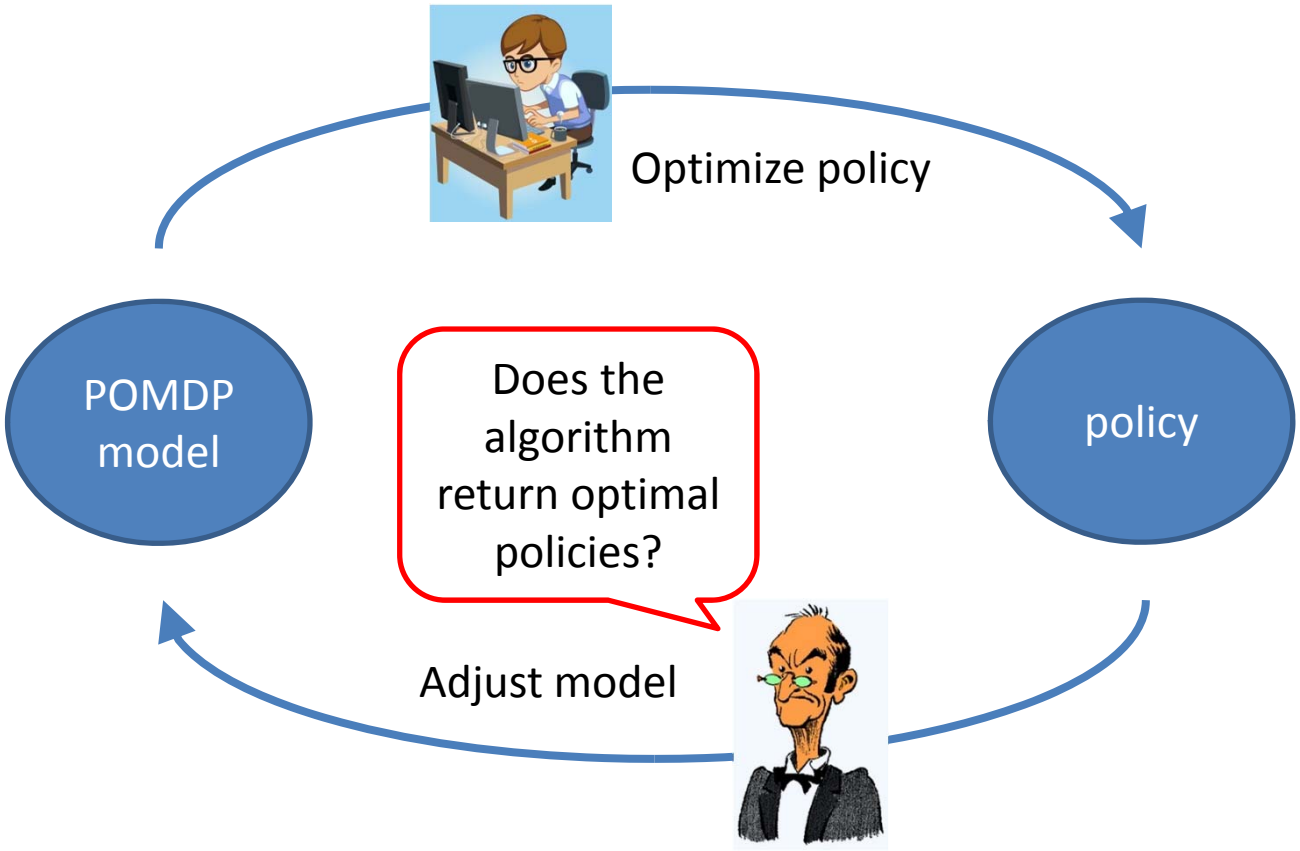
Dongho Kim (KAIST)

Outline

- Review
  - POMDPs
  - Bounds
- New algorithm: **GapMin**
  - Improved bounds
  - More compact representation
- Experimental results
- Conclusion

POMDP deployment

- Development cycle:



Motivation

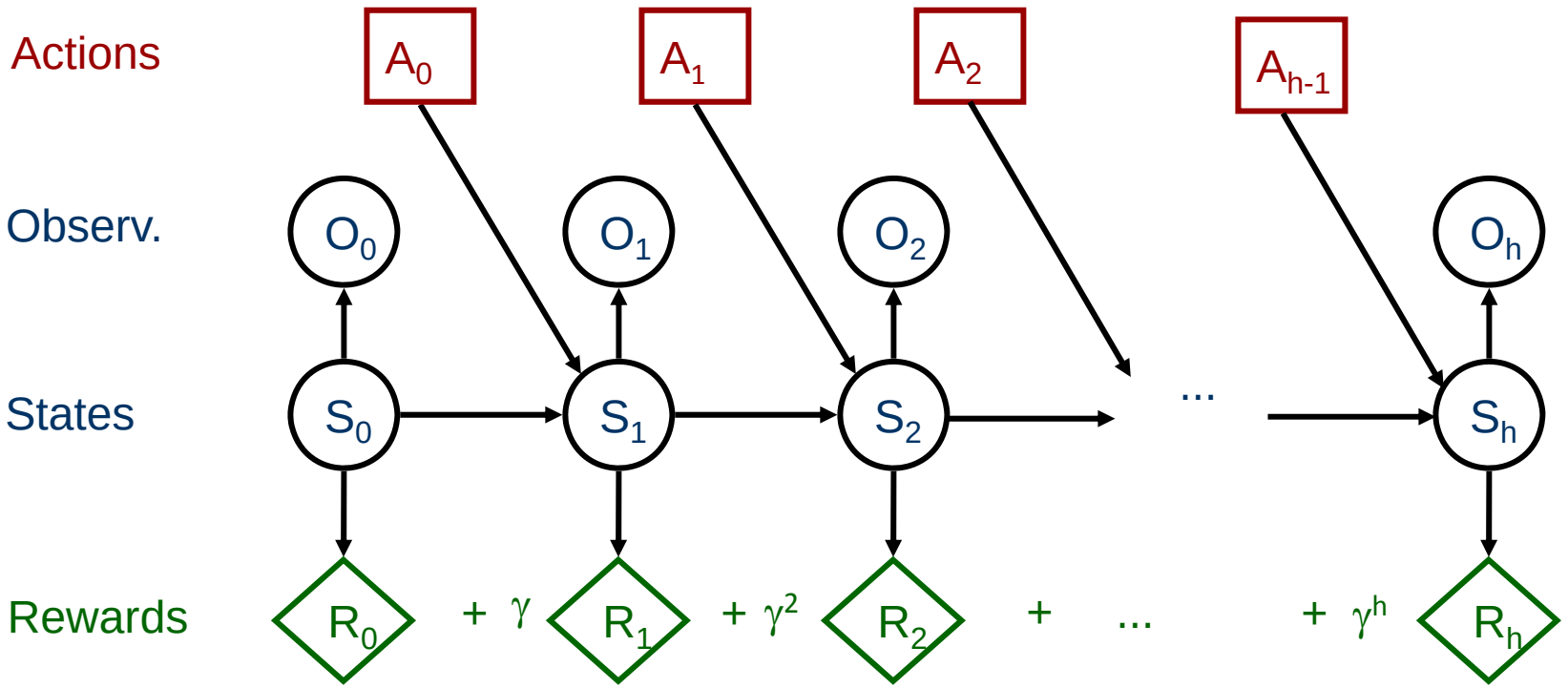
- Most POMDP algorithms
  - Output: value function or policy
  - No error bound
    - Exceptions: HSVI2, Symbolic HSVI, SARSOP
- Cassandra's repository (www.pomdp.org)
  - 68 POMDP benchmarks (10-15 years old)
  - Unknown optimal value for most benchmarks

Contributions

- New algorithm: **GapMin**
  - Tighter bounds on the optimal value
  - More compact bounds

- Cassandra's repository
  - (near) optimal value in < 1000 sec
    - SARSOP: 31 problems
    - HSVI2: 32 problems
    - **GapMin: 46 problems**

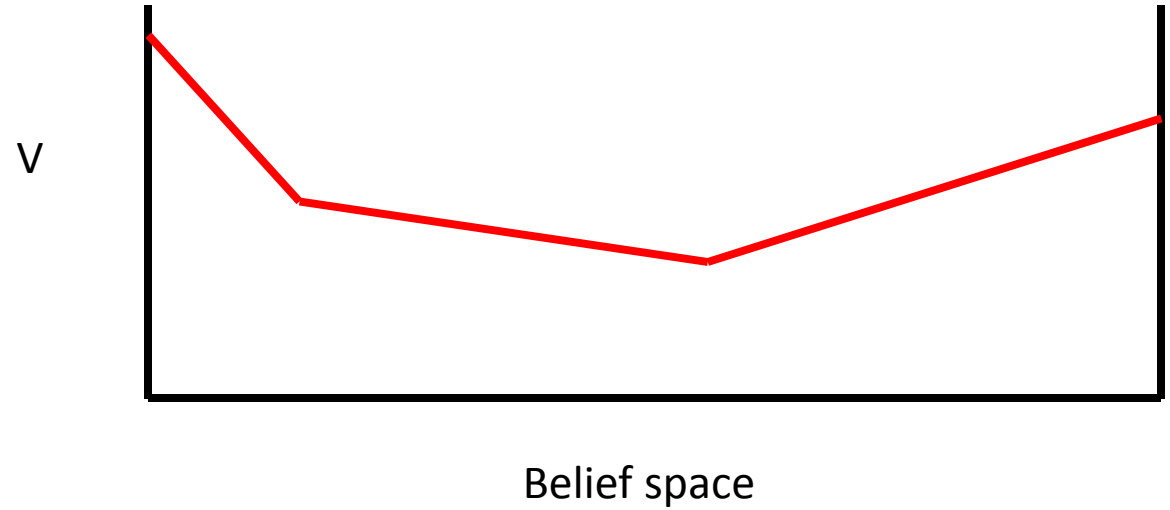
POMDP Graphical Representation



Solution: policy π maximizes expected total rewards

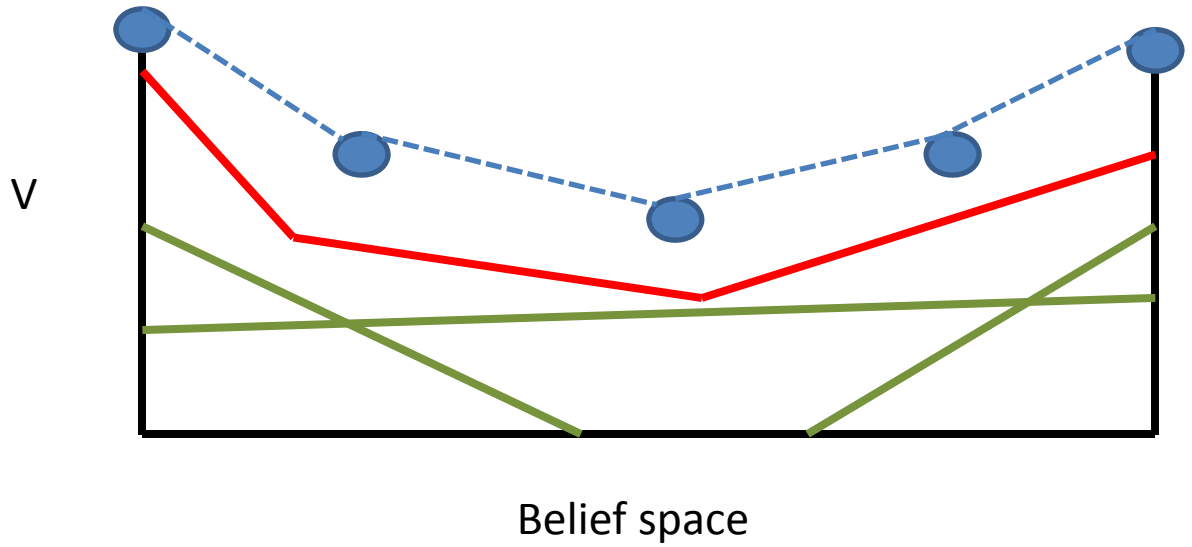
Optimal value function

- Piecewise linear and convex
  [Smallwood & Sondik, 73]



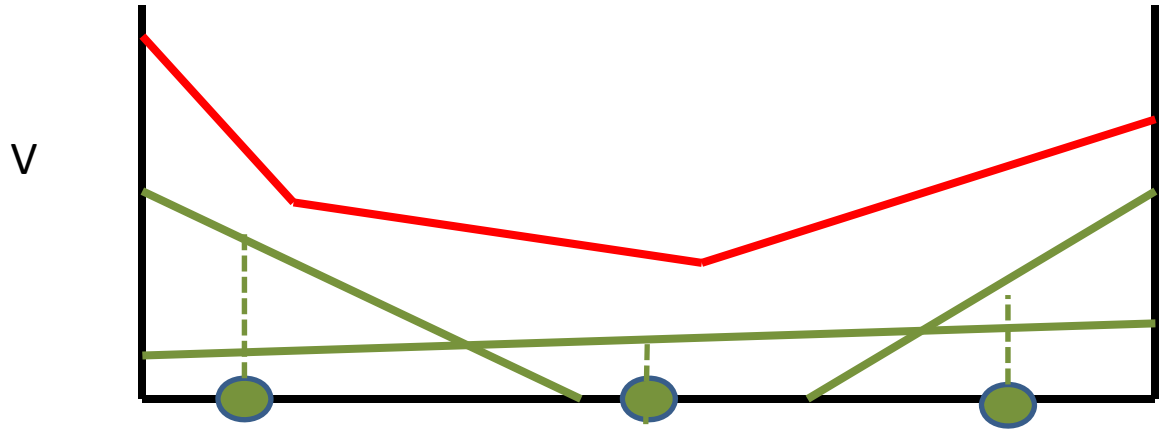
Bounds

- Lower bound: α-vectors
- Upper bound: belief-value pairs with interpolation



Lower Bound

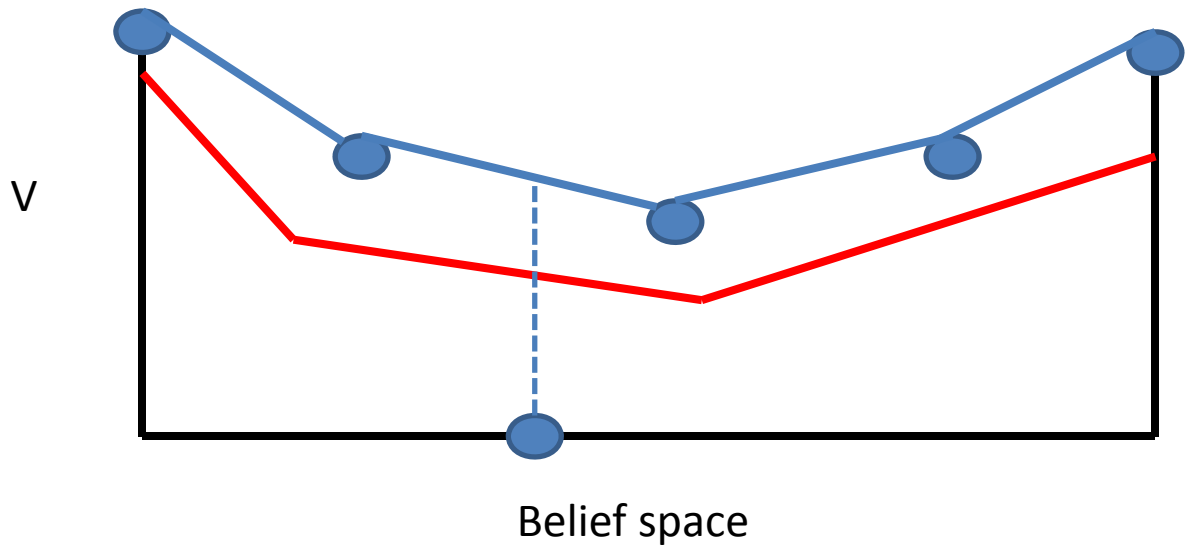
- Lower bound at reachable beliefs:



- Point-based value iteration
  - For each $b \in B$ do
    - $\alpha(b) \leftarrow \max_a R(b,a) + \gamma \sum_o \Pr(o|b,a) \max_{\alpha'} \alpha'(b_{ao})$

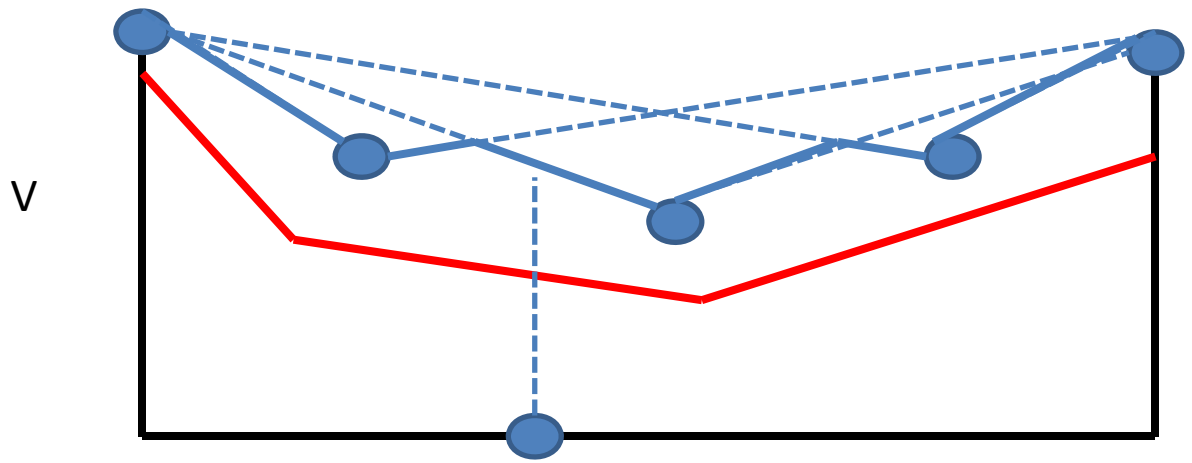
Upper Bound

- Belief-value pairs with LP interpolation
  - $\min_{c} \sum_{\bar{b}} c(\bar{b}) \bar{V}(\bar{b})$ s.t. $\sum_{\bar{b}} c(\bar{b}) \bar{b}(s) = b(s)\ \forall s$
  - Polynomial complexity (expensive)



Upper Bound

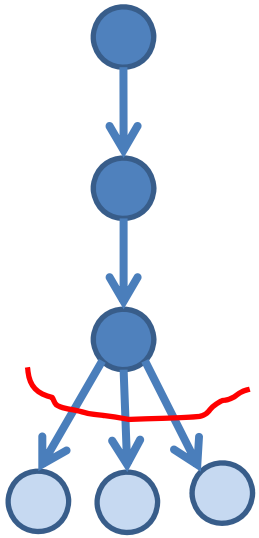
- Belief-value pairs with sawtooth interpolation
  - $\min_{\bar{b}} interpolate\ (b, \{\bar{b}\} \cup S, \bar{V})$
  - Linear complexity: $O(|S||\bar{B}|)$



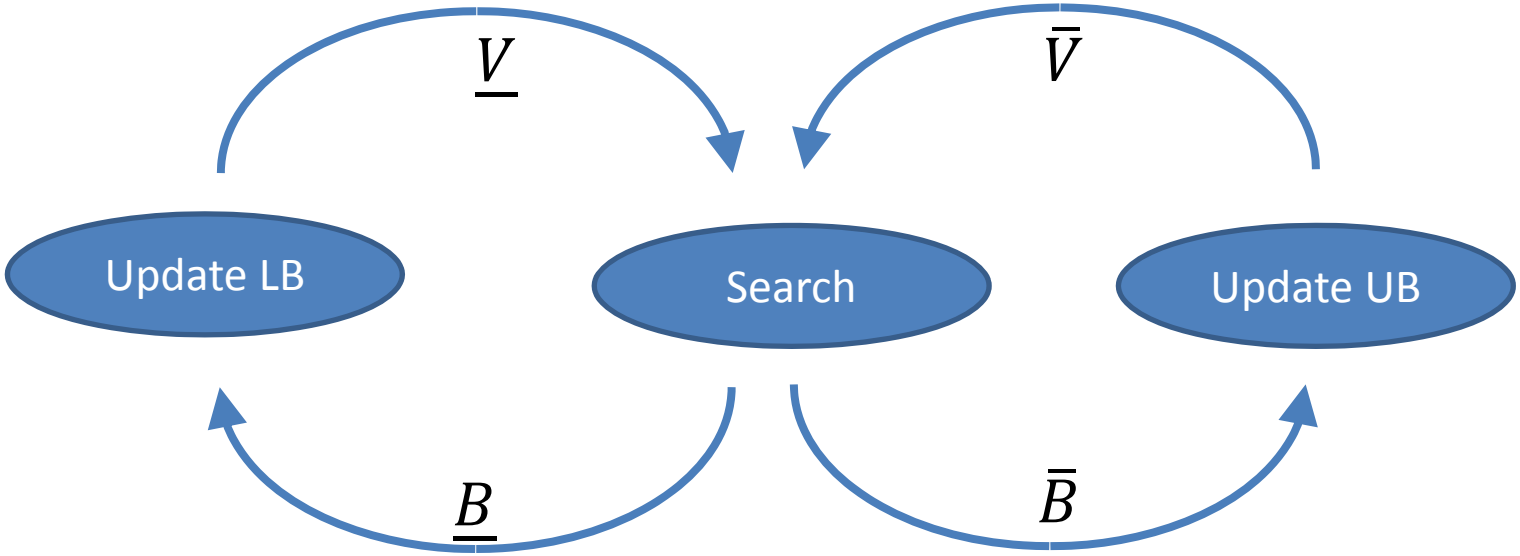
Belief search

- Find beliefs where the bounds are loose
  - Action selection:
    $a \leftarrow argmax_a \, R(b,a) + \gamma \sum_o \Pr(o|b,a) \, UB(b_{ao})$
  - Observation selection:
    $o \leftarrow argmax_o \, \Pr(o|b,a) \, gap(b_{ao})$

- Depth first search
- Guaranteed to find belief to be improved



Generic Bounding Algorithm

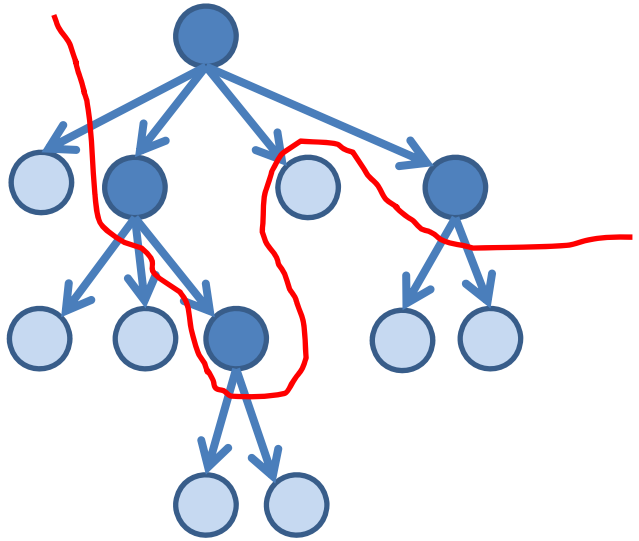


Comparison

	HSVI2 & SARSOP	GapMin
Search	Depth first	Priority queue Breadth first
LB	α-vectors	α-vectors
LB update	Path backup	PBVI
UB interpolation	sawtooth	LP
UB update	Path backup	FIB

GapMin belief search

- Priority queue
  - breadth first
  - fewer shallow beliefs
  - More compact bounds



  - Action selection: most promising action
    $a \leftarrow argmax_a \, R(b,a) + \gamma \sum_o \Pr(o|b,a) \, UB(b_{ao})$
  - Priority queue score: largest weighted gap
    $score(b) \leftarrow \gamma^{depth} \Pr(history) \, gap(b)$

UB update

- Point-based update:
  - For each $\bar{b} \in \bar{B} \cup S$ do
    - $\bar{V}(\bar{b}) \leftarrow \max_a R(\bar{b}, a) + \gamma \sum_o \Pr(o|\bar{b}, a)\, UB(\overline{b_{ao}})$
  - Costly: because of UB interpolation

Caching

- Lovejoy 91, Hauskrecht 00:
  - Cache interpolations
    - i.e., for each $\bar{b}_{ao}$ store convex combination c_{ao}
  - For each $\bar{b} \in \bar{B} \cup S$ do
    - $\bar{V}(\bar{b}) \leftarrow \max_a R(\bar{b}, a) + \gamma \sum_o \Pr(o|\bar{b}, a) \sum_{\bar{b}} c_{ao}(\bar{b}) \bar{V}(\bar{b})$

- Equivalent to value iteration in a belief MDP or augmented POMDP

UB update

- UB update with augmented POMDP:
  - QMDP
    - Policy: $S \rightarrow A$
    - $\bar{V}(s,a) = R(s,a) + \gamma \sum_{s'} \Pr(s'|s,a) \max_{a'} \bar{V}(s',a')$
  - Fast Informed Bound (FIB)
    - Policy: $S \times A \times O \rightarrow A$
    - $\bar{V}(s,a) = R(s,a) + \gamma \sum_{o} \max_{a'} \sum_{s'} \Pr(s'|s,a) \Pr(o|s',a) \bar{V}(s',a')$

- $\bar{V}_{QMDP} \geq \bar{V}_{FIB} \geq V^*$

Experiments

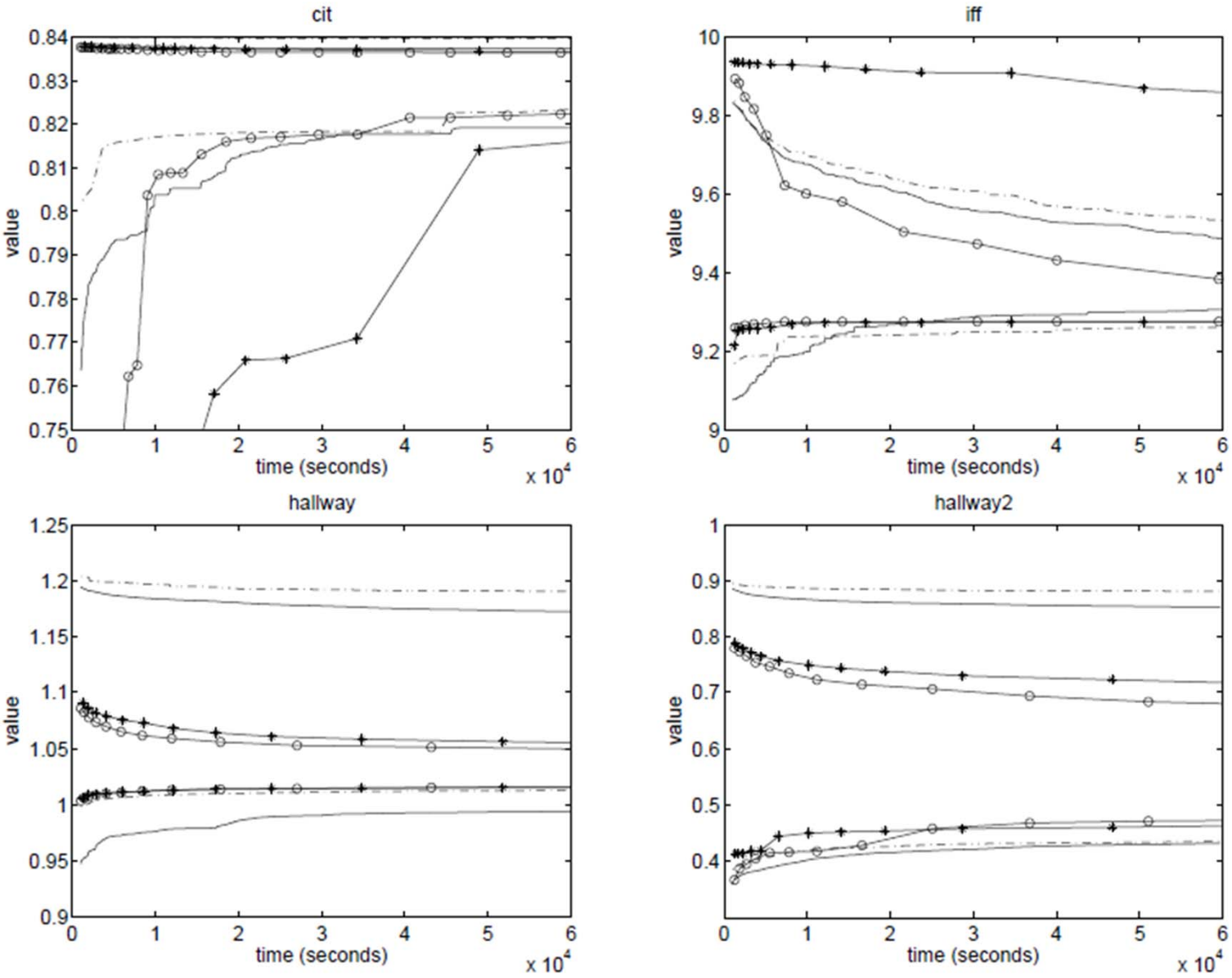
- Cassandra's repository (10-15 years old)
  - Tested 64 problems
  - Max time: 1000 sec

	HSVI2	SARSOP	GapMin
(near) optimal solutions (out of 64)	32	31	**46**
Smallest gap (out of 18)	2	4	**12**

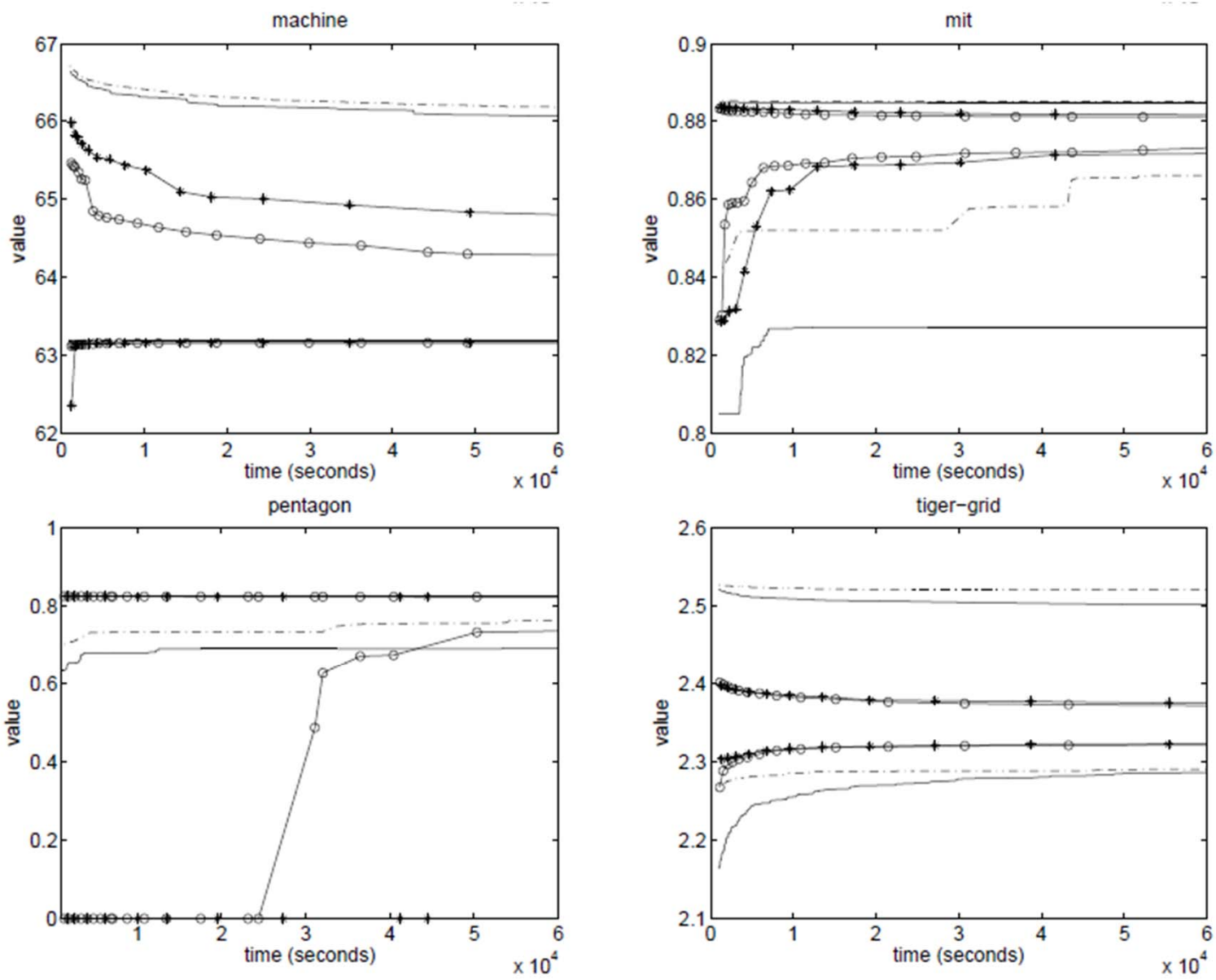
Experiments

- Max time: 50,000 sec

problem	algorithm	gap	LB	UB	$\|\Gamma\|$	$\|V\|$	time
cit	hsvi2	0.0182	0.8192	0.8373	29803	n.a.	49760
$\|\mathcal{S}\| = 284$	sarsop	0.0169	0.8228	0.8396	21168	9337	49916
$\|\mathcal{A}\| = 4, \|\mathcal{O}\| = 28$	gapMin ST	0.0226	0.8141	0.8367	739	681	48931
$\gamma = 0.990$	**gapMin LP**	**0.0149**	0.8215	0.8364	648	614	45473
hallway	hsvi2	0.179	0.994	1.173	15374	n.a.	49951
$\|\mathcal{S}\| = 60$	sarsop	0.178	1.013	1.191	3053	12869	49992
$\|\mathcal{A}\| = 5, \|\mathcal{O}\| = 21$	gapMin ST	0.043	1.015	1.058	947	2611	34828
$\gamma = 0.950$	**gapMin LP**	**0.036**	1.016	1.051	851	1904	43184
hallway2	hsvi2	0.4211	0.4319	0.8530	18505	n.a.	49983
$\|\mathcal{S}\| = 92$	sarsop	0.4482	0.4336	0.8818	1901	10908	49973
$\|\mathcal{A}\| = 5, \|\mathcal{O}\| = 17$	gapMin ST	0.2620	0.4605	0.7225	1647	2809	46687
$\gamma = 0.950$	**gapMin LP**	**0.2256**	0.4680	0.6936	1135	1798	36766
iff	hsvi2	0.199	9.302	9.501	40984	n.a.	50000
$\|\mathcal{S}\| = 104$	sarsop	0.290	9.259	9.549	54016	12237	49966
$\|\mathcal{A}\| = 4, \|\mathcal{O}\| = 22$	gapMin ST	0.634	9.273	9.908	1614	4502	34472
$\gamma = 0.999$	**gapMin LP**	**0.156**	9.275	9.431	1626	6231	40046
machine	hsvi2	2.89	63.18	66.07	7857	n.a.	49998
$\|\mathcal{S}\| = 256$	sarsop	3.02	63.18	66.20	996	22591	49963
$\|\mathcal{A}\| = 4, \|\mathcal{O}\| = 16$	gapMin ST	1.67	63.17	64.84	139	3807	49261
$\gamma = 0.990$	**gapMin LP**	**1.14**	63.17	64.30	173	1988	49036
mit	hsvi2	0.0575	0.8273	0.8848	34461	n.a.	49942
$\|\mathcal{S}\| = 204$	sarsop	0.0196	0.8655	0.8851	20662	12097	49616
$\|\mathcal{A}\| = 4, \|\mathcal{O}\| = 28$	gapMin ST	0.0105	0.8714	0.8819	861	984	41564
$\gamma = 0.990$	**gapMin LP**	**0.0091**	0.8721	0.8812	832	1051	43680
pentagon	hsvi2	0.1349	0.6910	0.8258	29033	n.a.	49924
$\|\mathcal{S}\| = 212$	**sarsop**	**0.0702**	0.7570	0.8271	21950	7534	49994
$\|\mathcal{A}\| = 4, \|\mathcal{O}\| = 28$	gapMin ST	0.8249	0.0000	0.8249	1	713	44437
$\gamma = 0.990$	gapMin LP	0.1497	0.6747	0.8244	425	846	40436
tiger-grid	hsvi2	0.217	2.286	2.502	28182	n.a.	49948
$\|\mathcal{S}\| = 36$	sarsop	0.231	2.290	2.522	5333	12504	49987
$\|\mathcal{A}\| = 5, \|\mathcal{O}\| = 17$	gapMin ST	0.055	2.322	2.377	2404	3752	38675
$\gamma = 0.950$	**gapMin LP**	**0.052**	2.321	2.373	2404	3778	43254



o: gapMinLP *: gapMinST ——: HSVI2 ---: SARSOP



o: gapMinLP *: gapMinST ——: HSVI2 ---: SARSOP

Conclusion

- New algorithm: **GapMin**
  - Tighter bounds
  - More compact representation
  - Online code: www.cs.uwaterloo.ca/~ppoupart/software.html
- Future work
  - Factored GapMin