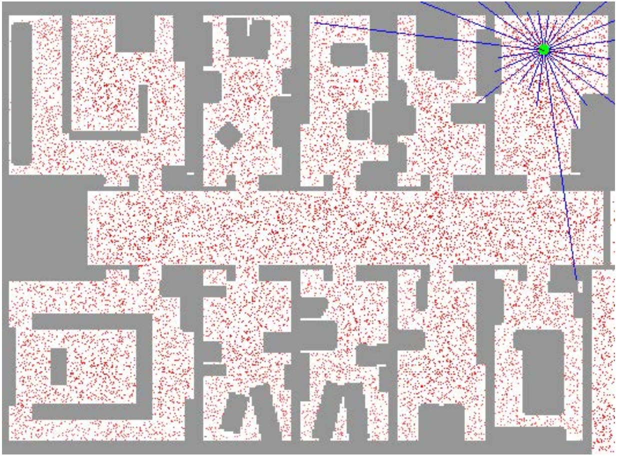
Information Gathering and Reward Exploitation of Subgoals for POMDPs

Hang Ma and Joelle Pineau
McGill University

AAAI January 27, 2015



http://www.cs.washington.edu/ai/Mobile_Robotics/mcl/animations/global-floor.gif

POMDPs

A partially observable Markov decision process (POMDP) is a tuple $\langle \mathcal{S}, \mathcal{A}, \Omega, T, O, R, b_0, \gamma \rangle$

- $\mathcal{S}$ is a finite set of states, $\mathcal{A}$ is a finite set of actions, Ω is a finite set of observations;
- $T(s, a, s') = p(s'|s, a)$ is the transition function that maps each state and action to a probability distribution over states;
- $O(s', a, o) = p(o|s', a)$ is the observation function that maps a state and an action to a probability distribution over possible observations;
- $R(s, a)$ is the reward function, $b_0(s)$ is the initial belief state, and γ is the discount factor.

Beliefs

A belief state $b \in \mathcal{B}$ is a sufficient statistic for the history, and is updated after taking action a and receiving observation o as follows:

$$b^{a,o}(s') = \frac{O(s', a, o) \sum_{s \in \mathcal{S}} T(s, a, s') b(s)}{p(o|a, b)}, \tag{1}$$

where $p(o|a, b) = \sum_{s \in \mathcal{S}} b(s) \sum_{s' \in \mathcal{S}} T(s, a, s') O(s', a, o)$ is a normalizing factor.

Policies

A policy is a mapping from the current belief state to an action.
A value function $\mathcal{V}_\pi(b)$ specifies the expected reward gained starting from b followed by policy π:

$$\mathcal{V}_\pi(b) = \sum_{s \in \mathcal{S}} b(s) R(s, \pi(b)) + \gamma \sum_{o \in \Omega} p(o|b, \pi(b)) \mathcal{V}_\pi(b^{\pi(b),o}).$$

POMDP Planning

Find an optimal policy that maximizes its value function.

Challenges

Computational complexity:

- We must reason in a $(n-1)$-dimensional continuous belief space (the curse of dimensionality).
- Complexity also grows fast with the length of planning horizon (the curse of history).

Challenges

Computational complexity:

- We must reason in a $(n-1)$-dimensional continuous belief space (the curse of dimensionality).
- Complexity also grows fast with the length of planning horizon (the curse of history).

Practical challenges:

- How to carry out intelligent information gathering in a large high dimensional belief space.
- How to scale up planning with long sequences of actions and delayed rewards.

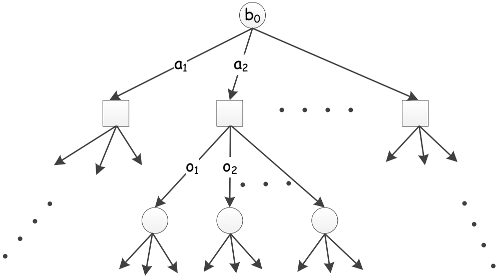


Figure: A belief tree rooted at b_0.

Existing Solvers

- PBVI (Pineau et al., 2003)
- HSVI/HSVI2 (Smith and Simmons, 2004, 2005)
- FSVI (Shani et al., 2007)
- SARSOP (Kurniawati et al., 2008)
- RTDP-Bel (Bonet and Geffner, 2009)
- MiGS (Kurniawati et al., 2011)
- Some online solvers, e.g. PUMA (He et al., 2010), POMCP (Silver and Veness, 2010)

Our Approach: Information Gathering and Reward Exploitation of Subgoals (IGRES)

Main ideas:

Our Approach: Information Gathering and Reward Exploitation of Subgoals (IGRES)

Main ideas:

1. Sample potentially important states as subgoals.

Our Approach: Information Gathering and Reward Exploitation of Subgoals (IGRES)

Main ideas:

1. Sample potentially important states as subgoals.
2. Generate macro-actions (sequences of actions) for transitions to subgoals.

Our Approach: Information Gathering and Reward Exploitation of Subgoals (IGRES)

Main ideas:

1. Sample potentially important states as subgoals.
2. Generate macro-actions (sequences of actions) for transitions to subgoals.
3. Generate macro-actions for information gathering and reward exploitation in the neighborhood of subgoals.

Capturing Important States

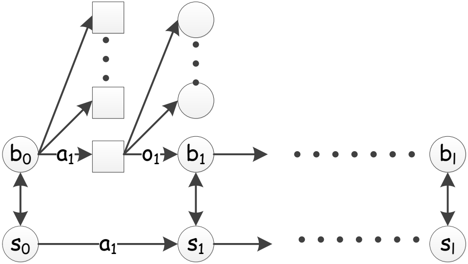
1. Identify importance heuristic functions with respect to:
   - Immediate reward;
   - Information gain.
2. Then a state is sampled as subgoal with probability of its importance.

Leverage the Structure

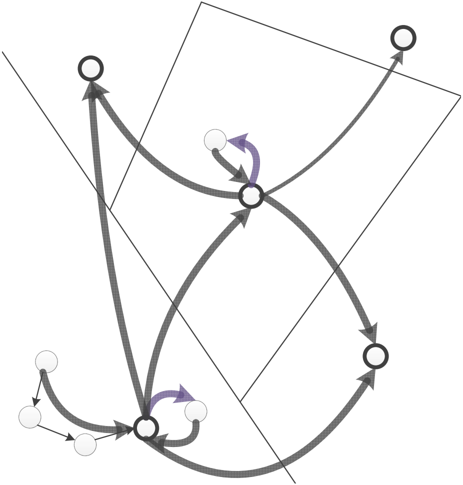
1. Specify distance between states with respect to the approximate similarity of their value.
2. Group all the states by the distance to the subgoals.

Sampling Belief States with Macro-actions

1. Associate each belief with an estimated current state.
2. Generate a macro-action towards the corresponding subgoal.
3. Gather information and exploit rewards around the subgoal.



Overview of IGRES



Conclusion

- Reduce planning complexity by using macro-actions.
- Sample potentially most useful beliefs (based on subgoal states).
- Capability of planning in large state space for a long horizon.

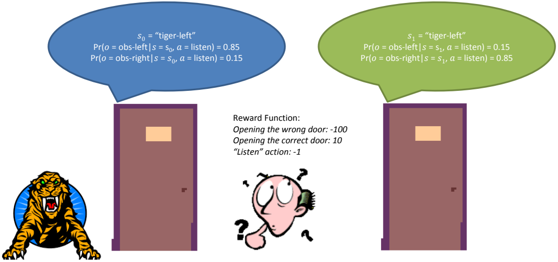


Figure: Tiger Domain

- Deterministic motions
- Noisy sensor for *Rock-goodness*
- +10 for sampling good
- -10 for sampling bad
- +10 for exiting
- No other cost/reward

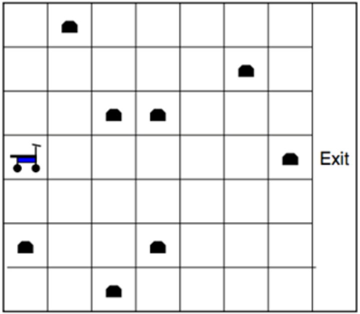


Figure: RockSample Domain

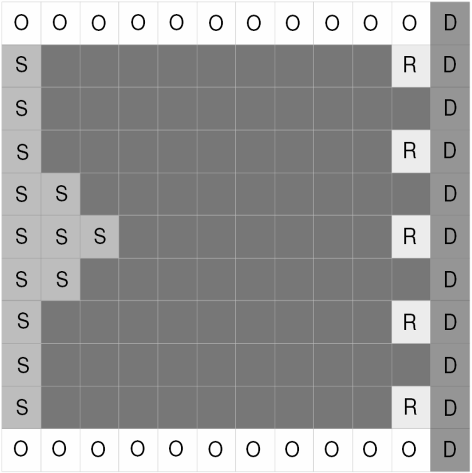


Figure: Underwater Domain

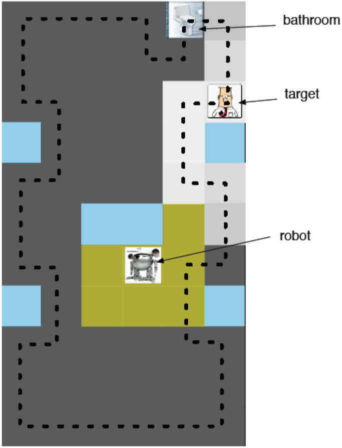


Figure: Homecare Domain

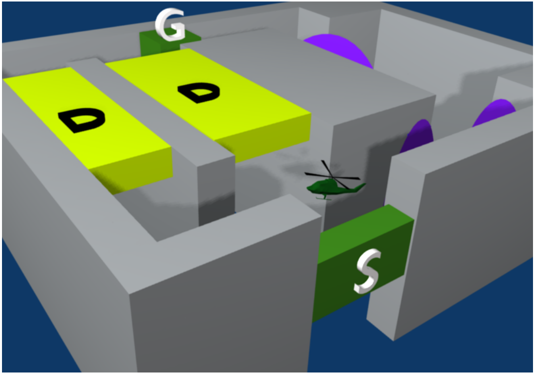


Figure: 3D-Navigation Domain

Results of the benchmark domains.

	Return	Time(s)
Tiger		
$\lvert\mathcal{S}\rvert = 2, \lvert\mathcal{A}\rvert = 3, \lvert\Omega\rvert = 2$		
RTDP-Bel	$\mathbf{19.42 \pm 0.59}$	0.30
HSVI2	$\mathbf{19.31 \pm 0.09}$	<1
FSVI*	N/A	
SARSOP	$\mathbf{18.59 \pm 0.61}$	0.09
MiGS	-19.88 ± 0	100
IGRES (# subgoals: 1)	$\mathbf{19.41 \pm 0.59}$	1
Noisy-tiger		
$\lvert\mathcal{S}\rvert = 2, \lvert\mathcal{A}\rvert = 3, \lvert\Omega\rvert = 2$		
RTDP-Bel	$\mathbf{-13.67 \pm 0.28}$	1.22
HSVI2	$\mathbf{-13.69 \pm 0.04}$	<1
FSVI*	N/A	
SARSOP	$\mathbf{-13.66 \pm 0.18}$	0.18
MiGS	-19.88 ± 0	100
IGRES (# subgoals: 1)	$\mathbf{-13.67 \pm 0.18}$	1
RockSample(4,4)		
$\lvert\mathcal{S}\rvert = 257, \lvert\mathcal{A}\rvert = 9, \lvert\Omega\rvert = 2$		
RTDP-Bel	$\mathbf{17.94 \pm 0.12}$	10.7
HSVI2	$\mathbf{17.92 \pm 0.01}$	<1
FSVI	$\mathbf{17.85 \pm 0.18}$	1
SARSOP	17.75 ± 0.12	0.7
MiGS	8.57 ± 0	100
IGRES (# subgoals: 4)	17.30 ± 0.12	10
RockSample(7,8)		
$\lvert\mathcal{S}\rvert = 12545, \lvert\mathcal{A}\rvert = 13, \lvert\Omega\rvert = 2$		
RTDP-Bel	20.55 ± 0.13	103
HSVI2	$\mathbf{21.09 \pm 0.10}$	100
FSVI	20.08 ± 0.20	102
SARSOP	$\mathbf{21.35 \pm 0.13}$	100
MiGS	7.35 ± 0	100
IGRES (# subgoals: 8)	19.54 ± 0.12	100
Hallway2		
$\lvert\mathcal{S}\rvert = 92, \lvert\mathcal{A}\rvert = 5, \lvert\Omega\rvert = 17$		
RTDP-Bel	0.237 ± 0.006	1004
HSVI2	0.507 ± 0.001	250
FSVI	0.494 ± 0.007	280
SARSOP	$\mathbf{0.530 \pm 0.008}$	200
MiGs	$\mathbf{0.522 \pm 0.008}$	200
IGRES (# subgoals: 20)	$\mathbf{0.530 \pm 0.008}$	200
Tag		
$\lvert\mathcal{S}\rvert = 870, \lvert\mathcal{A}\rvert = 5, \lvert\Omega\rvert = 30$		
RTDP-Bel	-6.32 ± 0.12	372
HSVI2	-6.46 ± 0.09	400
FSVI	$\mathbf{-6.11 \pm 0.11}$	35
SARSOP	$\mathbf{-6.08 \pm 0.12}$	30
MiGS	$\mathbf{-6.00 \pm 0.12}$	30
IGRES (# subgoals: 20)	$\mathbf{-6.12 \pm 0.12}$	30
Underwater Navigation		
$\lvert\mathcal{S}\rvert = 2653, \lvert\mathcal{A}\rvert = 6, \lvert\Omega\rvert = 103$		
RTDP-Bel	$\mathbf{750.07 \pm 0.28}$	338
HSVI2	718.37 ± 0.60	400
FSVI	725.88 ± 5.91	414
SARSOP	731.33 ± 1.14	150
MiGS	715.50 ± 1.37	400
IGRES (# subgoals: 20)	$\mathbf{749.94 \pm 0.30}$	50
Homecare		
$\lvert\mathcal{S}\rvert = 5408, \lvert\mathcal{A}\rvert = 9, \lvert\Omega\rvert = 928$		
RTDP-Bel**	N/A	
HSVI2	15.07 ± 0.37	2000
FSVI***	N/A	
SARSOP	$\mathbf{16.64 \pm 0.82}$	1000
MiGS	$\mathbf{16.70 \pm 0.85}$	1600
IGRES (# subgoals: 30)	$\mathbf{17.32 \pm 0.85}$	1000
3D-Navigation		
$\lvert\mathcal{S}\rvert = 16969, \lvert\mathcal{A}\rvert = 5, \lvert\Omega\rvert = 14$		
RTDP-Bel	-93.03 ± 0.01	2115
HSVI2	-91.98 ± 0	2000
FSVI**	N/A	
SARSOP	-99.97 ± 0	800
MiGS	$\mathbf{(2.977 \pm 0.512) \times 10^4}$	150
IGRES (# subgoals: 163)	$\mathbf{(3.272 \pm 0.193) \times 10^4}$	150

* ArrayIndexOutOfBoundsException is thrown.
** Solver is not able to compute a solution given large amount of computation time.
*** OutOfMemoryError is thrown.

Results of the adaptive management of migratory birds problem.

	Return	Time(s)
Lesser sand plover		
$\|\mathcal{S}\| = 108$, $\|\mathcal{A}\| = 3$, $\|\Omega\| = 36$		
symbolic Perseus*	4675	10
IGRES (# subgoals: 18)	5037.72 ± 8.82	10
Bar-tailed godwit b.		
$\|\mathcal{S}\| = 972$, $\|\mathcal{A}\| = 5$, $\|\Omega\| = 324$		
symbolic Perseus*	18217	48
IGRES (# subgoals: 36)	19572.41 ± 39.35	60
Terek sandpiper		
$\|\mathcal{S}\| = 2916$, $\|\mathcal{A}\| = 6$, $\|\Omega\| = 972$		
symbolic Perseus*	7263	48
IGRES (# subgoals: 72)	7867.95 ± 2.44	60
Bar-tailed godwit m.		
$\|\mathcal{S}\| = 2916$, $\|\mathcal{A}\| = 6$, $\|\Omega\| = 972$		
symbolic Perseus*	24583	58
IGRES (# subgoals: 72)	26654.06 ± 38.60	60
Grey-tailed tattler		
$\|\mathcal{S}\| = 2916$, $\|\mathcal{A}\| = 6$, $\|\Omega\| = 972$		
symbolic Perseus*	4520	378
IGRES (# subgoals: 72)	4860.91 ± 38.47	60
IGRES (# subgoals: 72)	4927.17 ± 38.14	300

* Results from Nicol et al. (2013).

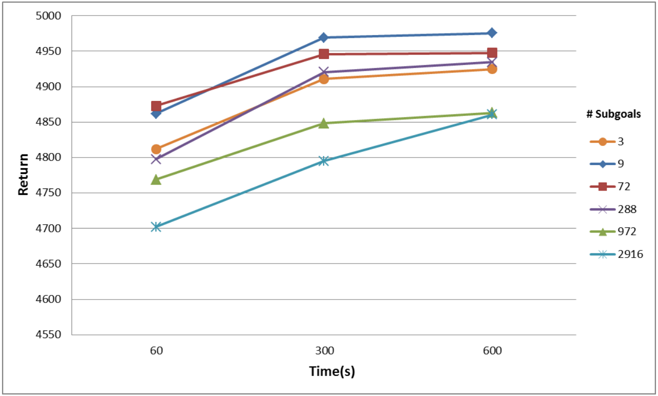


Figure: Performances on Grey-tailed tattler Domain

Thank You!

Reference I

B. Bonet and H. Geffner. Solving POMDPs: RTDP-bel vs. point-based algorithms. In *International Jont Conference on Artifical Intelligence*, 2009.

R. He, E. Brunskill, and N. Roy. Puma: Planning under uncertainty with macro-actions. In *National Conference on Artificial Intelligence*, 2010.

H. Kurniawati, D. Hsu, and W. S. Lee. SARSOP: Efficient point-based POMDP planning by approximating optimally reachable belief spaces. In *Robotics: Science and Systems*, 2008.

H. Kurniawati, Y. Du, D. Hsu, and W. S. Lee. Motion planning under uncertainty for robotic tasks with long time horizons. *International Journal of Robotics Research*, 30(3):308–323, 2011.

Reference II

S. Nicol, O. Buffet, T. Iwamura, and I. Chadès. Adaptive management of migratory birds under sea level rise. In *International Joint Conference on Artificial Intelligence*, 2013.

J. Pineau, G. Gordon, and S. Thrun. Point-based value iteration: An anytime algorithm for POMDPs. In *International Joint Conference on Artificial Intelligence*, 2003.

G. Shani, R. I. Brafman, and S. E. Shimony. Forward search value iteration for POMDPs. In *International Joint Conference on Artifical Intelligence*, 2007.

D. Silver and J. Veness. Monte-carlo planning in large POMDPs. In *Advances in Neural Information Processing Systems*, 2010.

T. Smith and R. G. Simmons. Heuristic search value iteration for POMDPs. In *Uncertainty in Artificial Intelligence*, 2004.

Reference III

T. Smith and R. G. Simmons. Point-based POMDP algorithms: Improved analysis and implementation. In *Uncertainty in Artificial Intelligence*, 2005.