Partially observable Markov decision processes

Matthijs Spaan

Institute for Systems and Robotics

Instituto Superior Técnico

Lisbon, Portugal

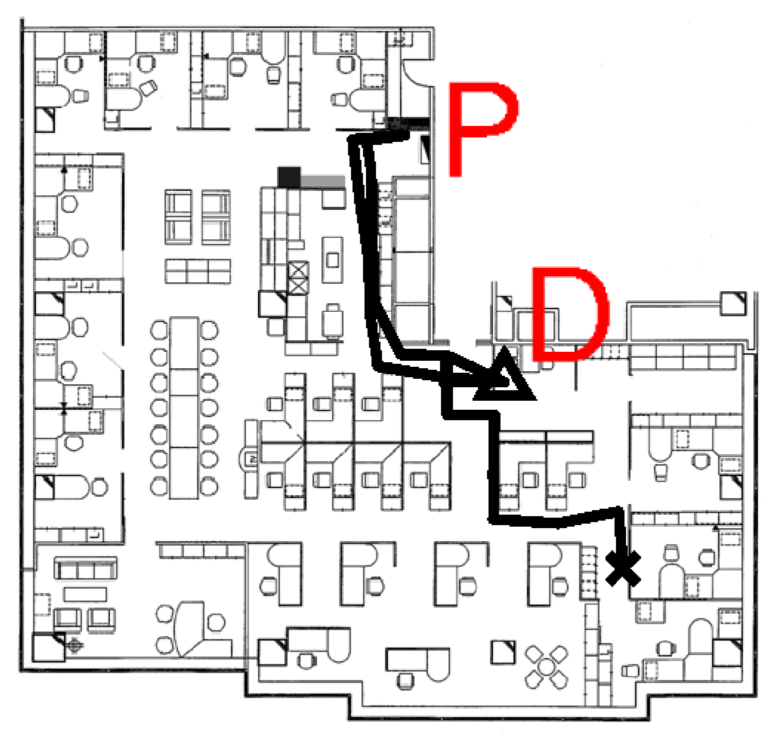
Reading group meeting, February 12, 2007

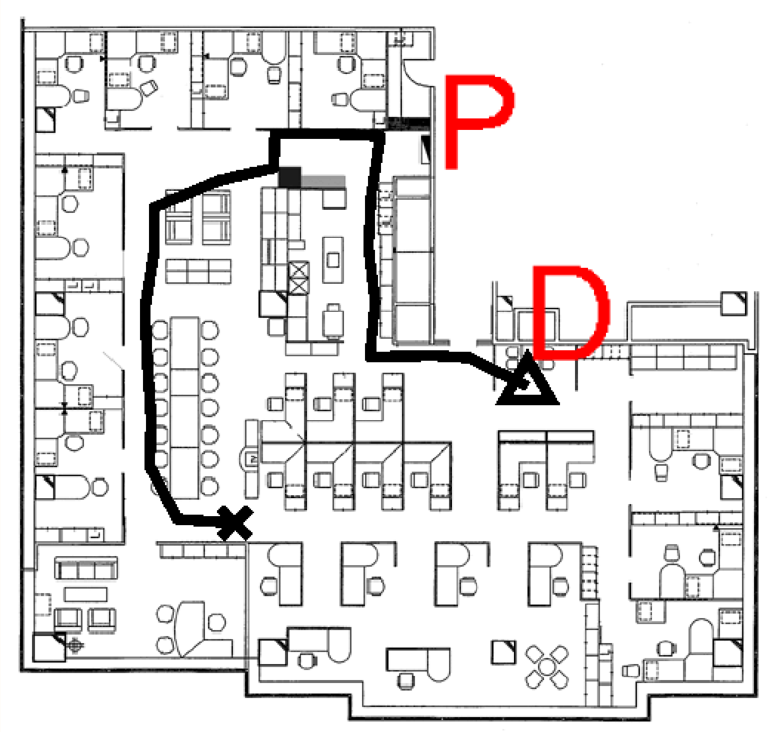
Overview

Partially observable Markov decision processes:

- Model.
- Belief states.
- MDP-based algorithms.
- Other sub-optimal algorithms.
- Optimal algorithms.
- Application to robotics.

A planning problem





Task: start at random position ($\times$) $\rightarrow$ pick up mail at P $\rightarrow$ deliver mail at D ($\triangle$).

Characteristics: motion noise, perceptual aliasing.

Planning under uncertainty

- Uncertainty is abundant in **real-world planning** domains.
- **Bayesian** approach $\Rightarrow$ probabilistic models.
- Common approach in robotics, e.g., robot localization.

POMDPs

Partially observable Markov decision processes (POMDPs) (Kaelbling et al., 1998):

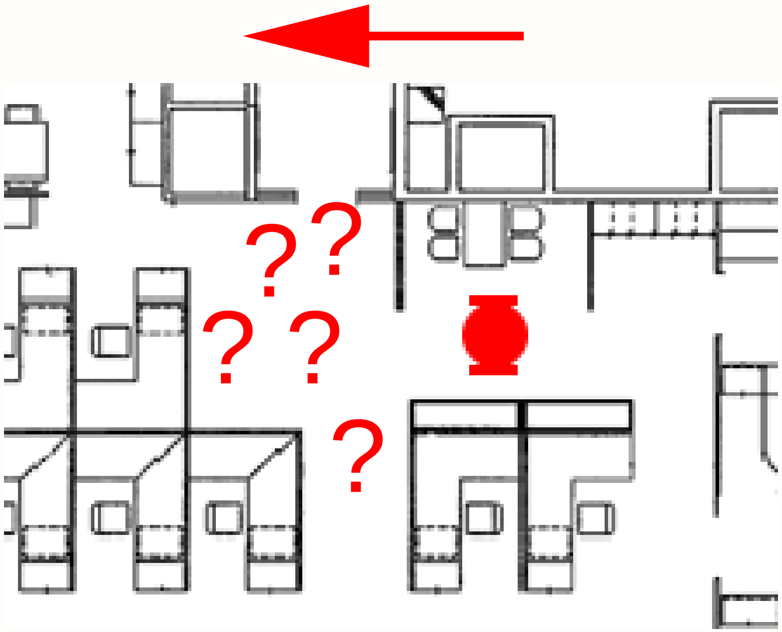
- Framework for agent planning under uncertainty.
- Typically assumes discrete sets of states S, actions A and observations O.
- Transition model $p(s'|s, a)$: models the effect of **actions**.
- Observation model $p(o|s, a)$: relates **observations** to states.
- Task is defined by a **reward** model $r(s, a)$.
- Goal is to compute plan, or **policy** π, that maximizes long-term reward.

POMDP applications

- Robot navigation (Simmons and Koenig, 1995; Theocharous and Mahadevan, 2002).
- Visual tracking (Darrell and Pentland, 1996).
- Dialogue management (Roy et al., 2000).
- Robot-assisted health care (Pineau et al., 2003b; Boger et al., 2005).
- Machine maintenance (Smallwood and Sondik, 1973), structural inspection (Ellis et al., 1995).
- Inventory control (Treharne and Sox, 2002), dynamic pricing strategies (Aviv and Pazgal, 2005), marketing campaigns (Rusmevichientong and Van Roy, 2001).
- Medical applications (Hauskrecht and Fraser, 2000; Hu et al., 1996).

Transition model

- For instance, robot motion is inaccurate.
- Transitions between states are **stochastic**.
- $p(s'|s,a)$ is the probability to jump from state s to state s' after taking action a.

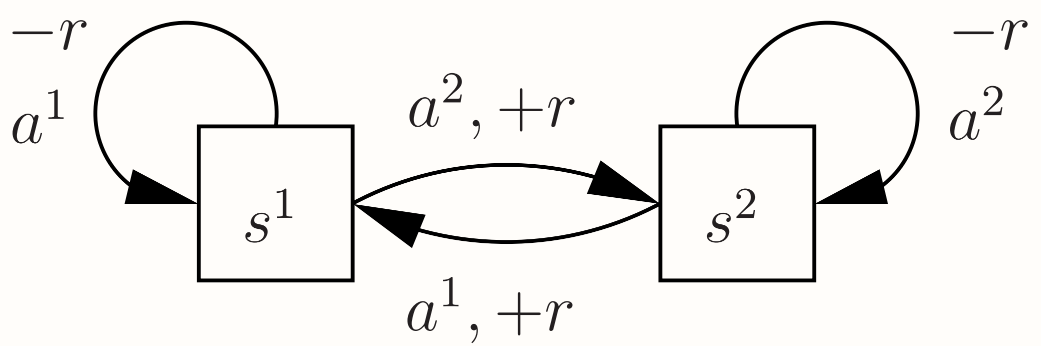


Observation model

- Imperfect sensors.
- Partially observable environment:
  - Sensors are **noisy**.
  - Sensors have a **limited view**.
- $p(o|s,a)$ is the probability the agent receives observation o in state s after taking action a.

Memory

A POMDP example that requires memory (Singh et al., 1994):



Method	Value
MDP policy	$V = \frac{r}{1-\gamma}$
Memoryless deterministic POMDP policy	$V_{\text{max}} = r - \frac{\gamma r}{1-\gamma}$
Memoryless stochastic POMDP policy	$V = 0$
Memory-based POMDP policy	$V_{\text{min}} = \frac{\gamma r}{1-\gamma} - r$

Beliefs

Beliefs:

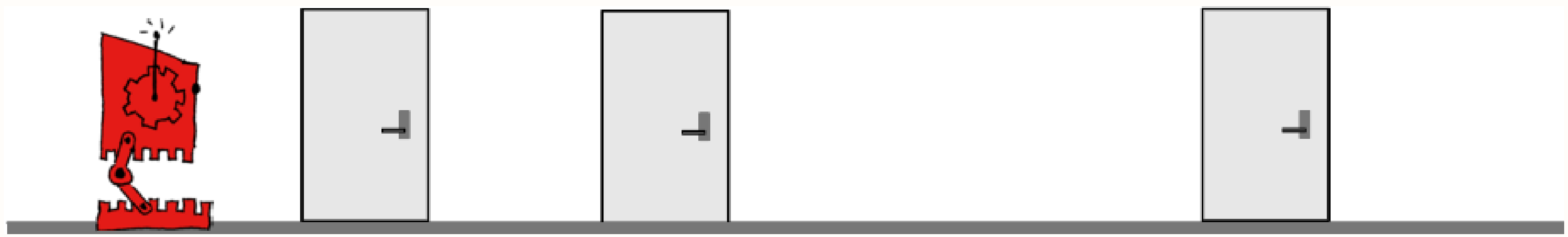
- The agent maintains a **belief** $b(s)$ of being at state s.
- After action $a \in A$ and observation $o \in O$ the belief $b(s)$ can be updated using Bayes' rule:

$$b'(s') \propto p(o|s') \sum_{s} p(s'|s,a) b(s)$$

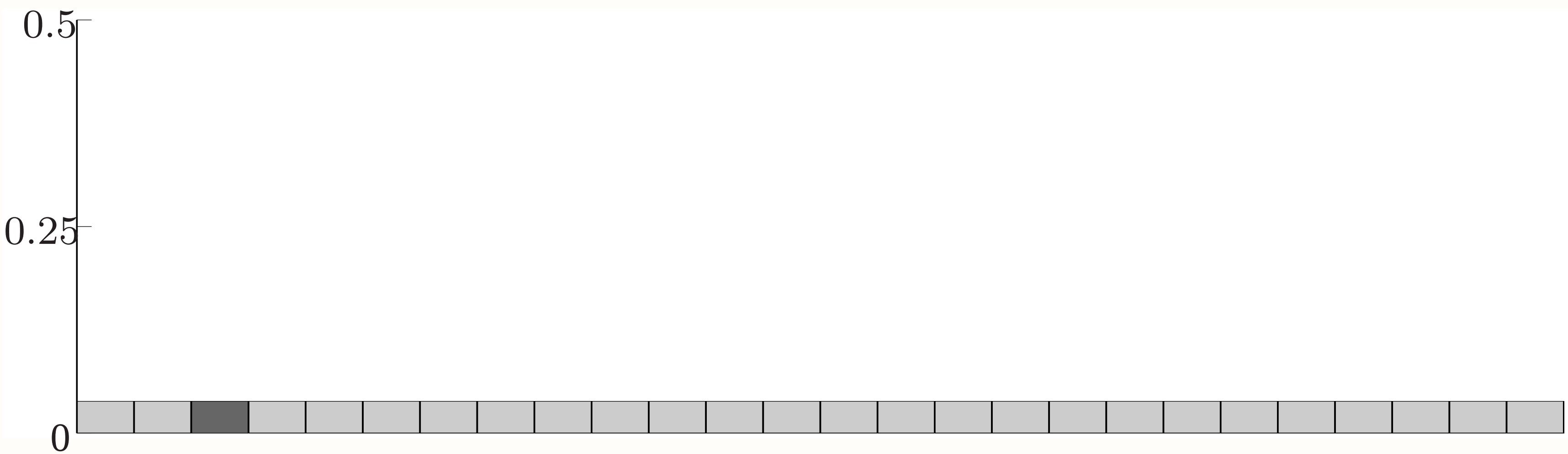
- The belief vector is a **Markov** signal for the planning task.

Belief update example

True situation:



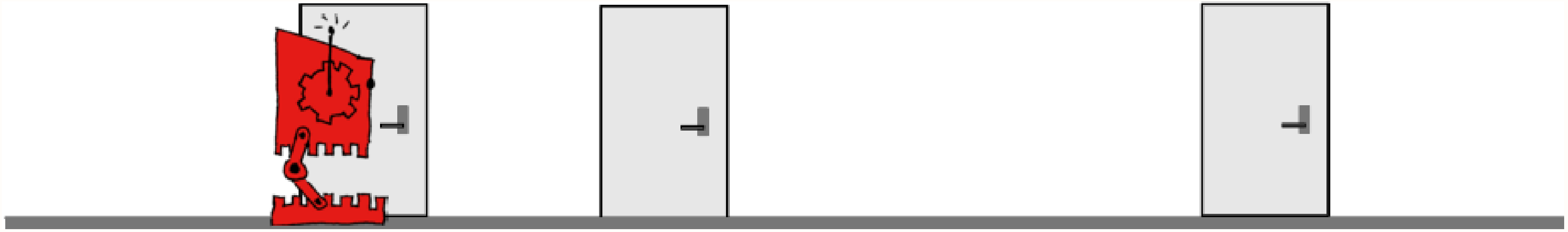
Robot's belief:



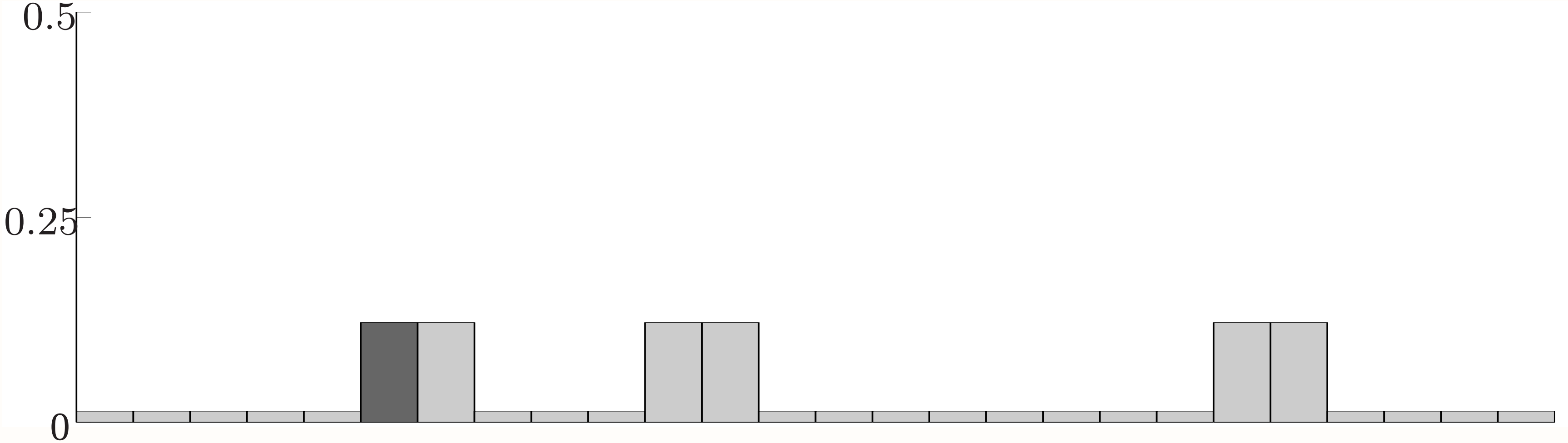
- Observations: *door* or *corridor*, 10% noise.
- Action: moves 3 (20%), 4 (60%), or 5 (20%) states.

Belief update example

True situation:



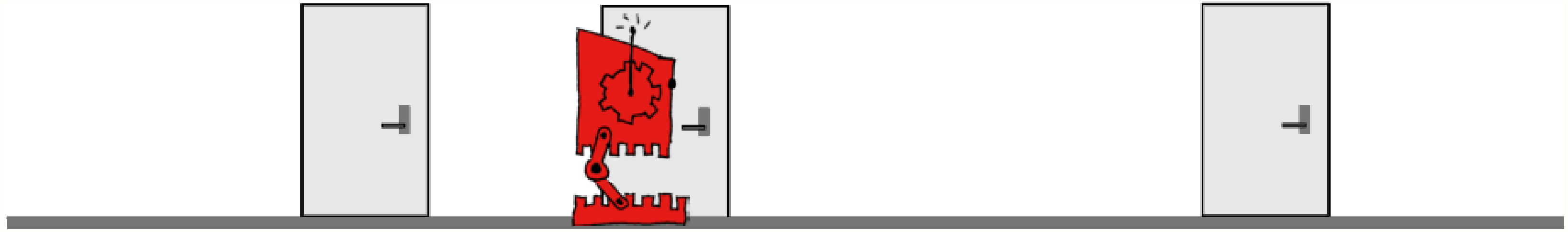
Robot's belief:



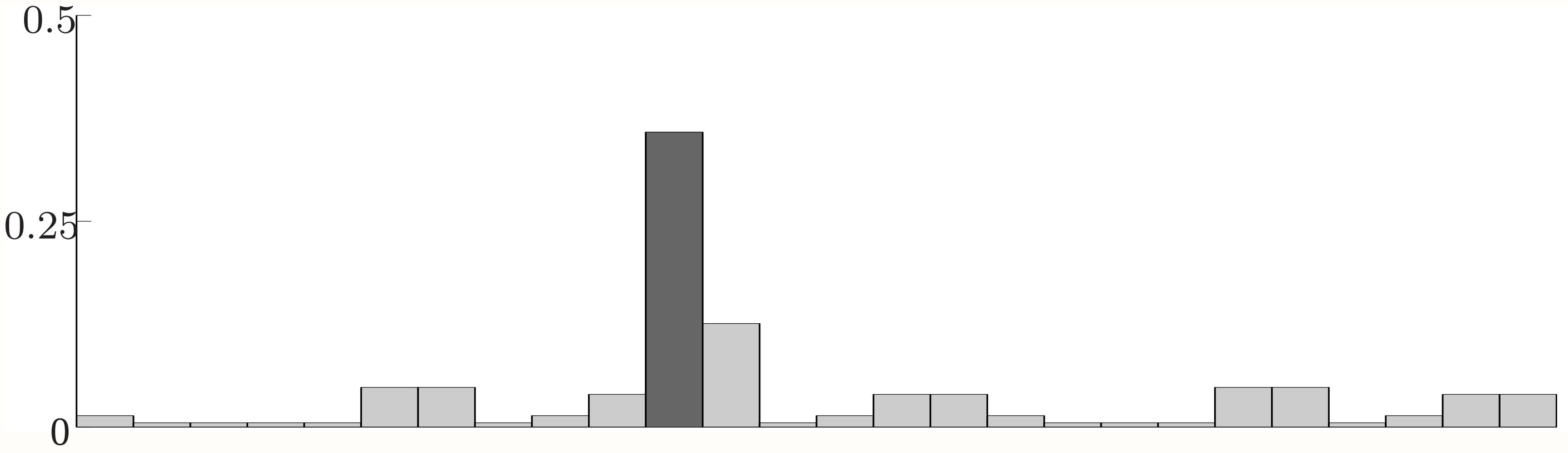
- Observations: **door** or *corridor*, 10% noise.
- Action: moves 3 (20%), 4 (60%), or 5 (20%) states.

Belief update example

True situation:



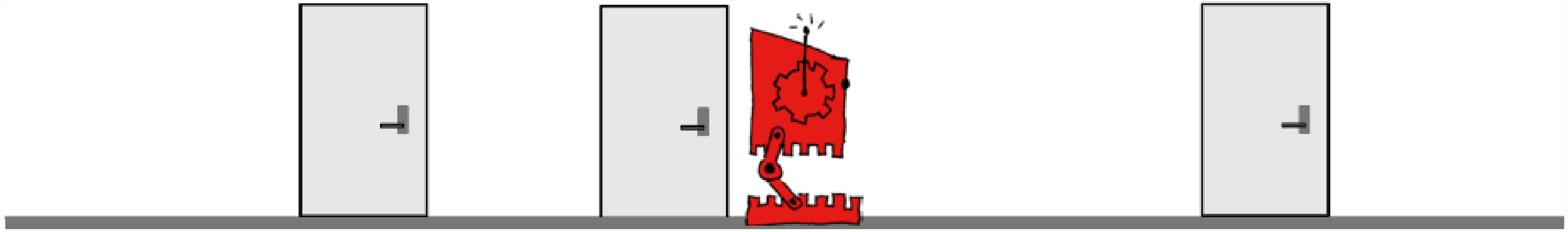
Robot's belief:



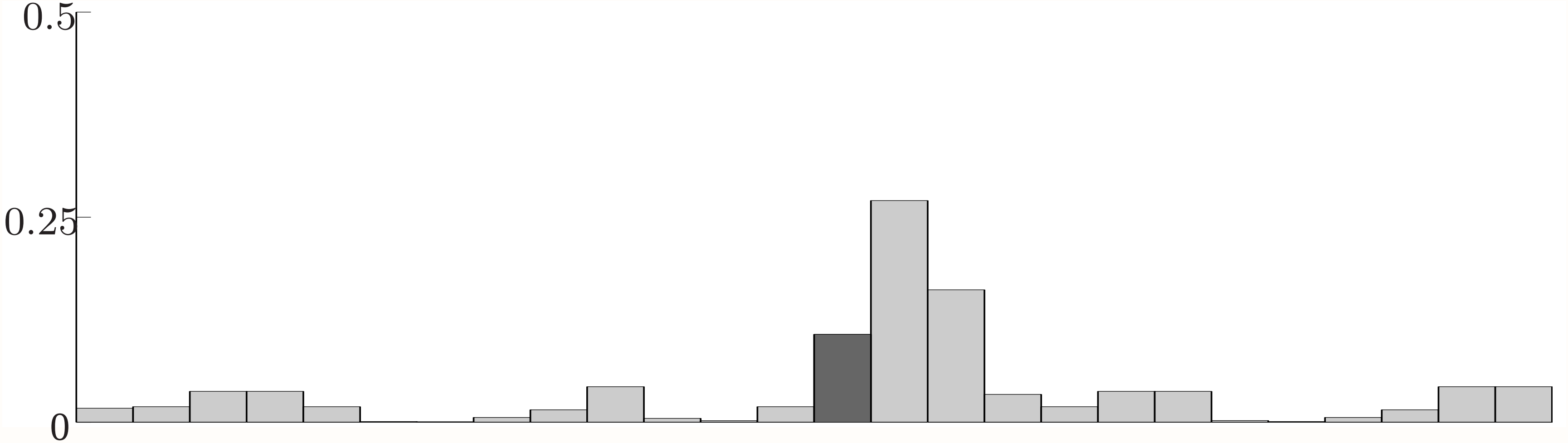
- Observations: **door** or *corridor*, 10% noise.
- Action: moves 3 (20%), 4 (60%), or 5 (20%) states.

Belief update example

True situation:



Robot's belief:



- Observations: *door* or **corridor**, 10% noise.
- Action: moves 3 (20%), 4 (60%), or 5 (20%) states.

Solving POMDPs

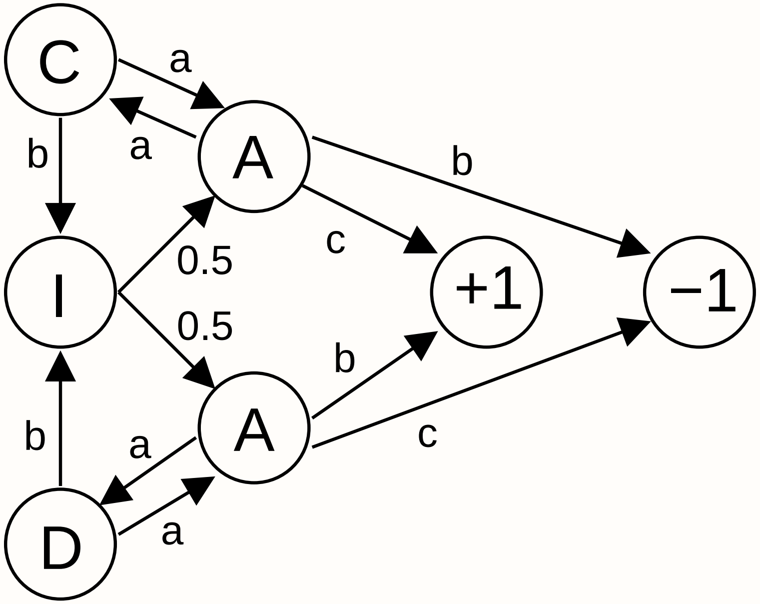
- A solution to a POMDP is a **policy**, i.e., a mapping $a = \pi(b)$ from beliefs to actions.
- An optimal policy is characterized by a **value function** that maximizes:

$$V_\pi(b_0) = E[\sum_{t=0}^{\infty} \gamma^t r(b_t, \pi(b_t))]$$

- Computing the optimal value function is a hard problem (PSPACE-complete for finite horizon).
- In robotics: a policy is often computed using simple MDP-based approximations.

MDP-based algorithms

- Use the solution to the MDP as an heuristic.
- Most likely state (Cassandra et al., 1996): $\pi_{MLS}(b) = \pi^*(\arg\max_s b(s))$.
- Q_{MDP} (Littman et al., 1995): $\pi_{Q_{\text{MDP}}}(b) = \arg\max_a \sum_s b(s)Q^*(s,a)$.



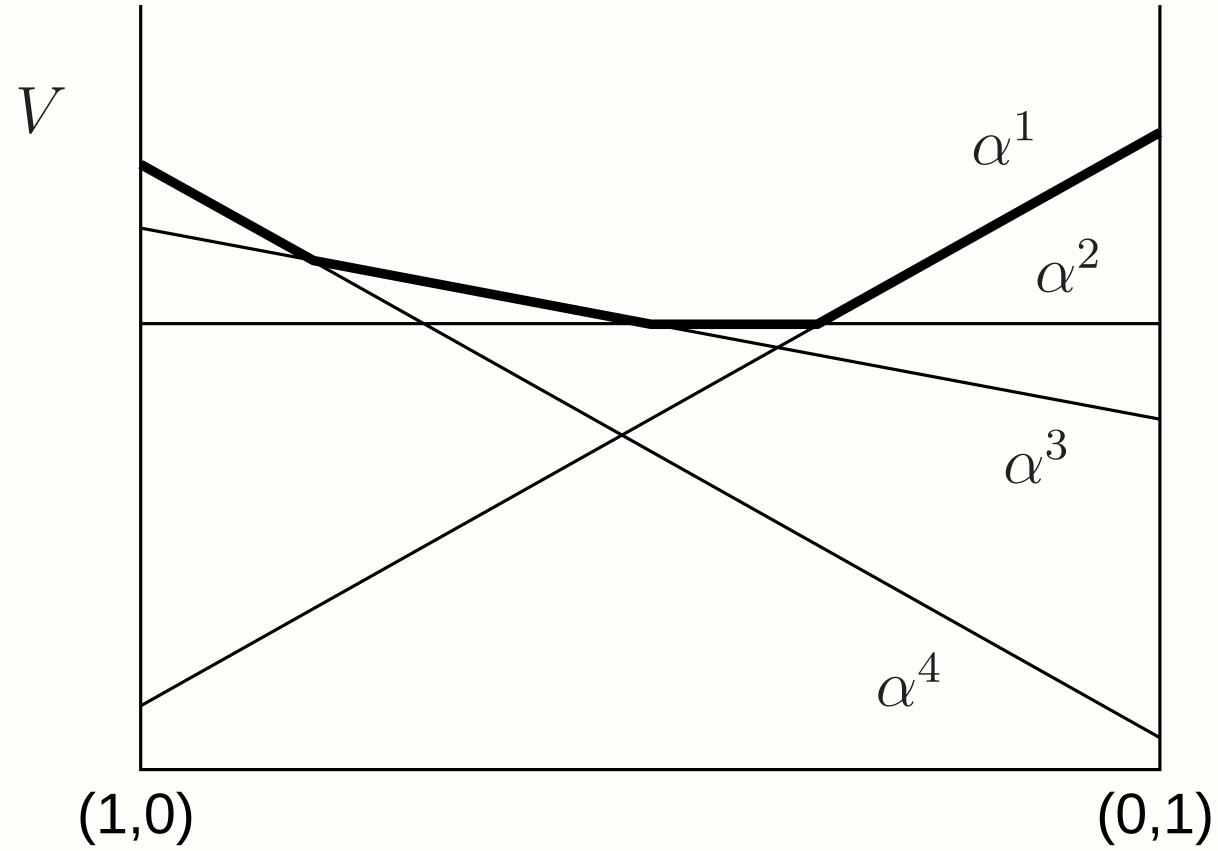
(Parr and Russell, 1995)

Other sub-optimal techniques

- Grid-based approximations (Drake, 1962; Lovejoy, 1991; Brafman, 1997; Zhou and Hansen, 2001; Bonet, 2002).
- Optimizing finite-state controllers (Platzman, 1981; Hansen, 1998b; Poupart and Boutilier, 2004).
- Gradient ascent (Ng and Jordan, 2000; Aberdeen and Baxter, 2002).
- Heuristic search in the belief tree (Satia and Lave, 1973; Hansen, 1998a; Smith and Simmons, 2004).
- Compressing the POMDP (Roy et al., 2005; Poupart and Boutilier, 2003).
- Point-based techniques (Pineau et al., 2003a; Spaan and Vlassis, 2005).

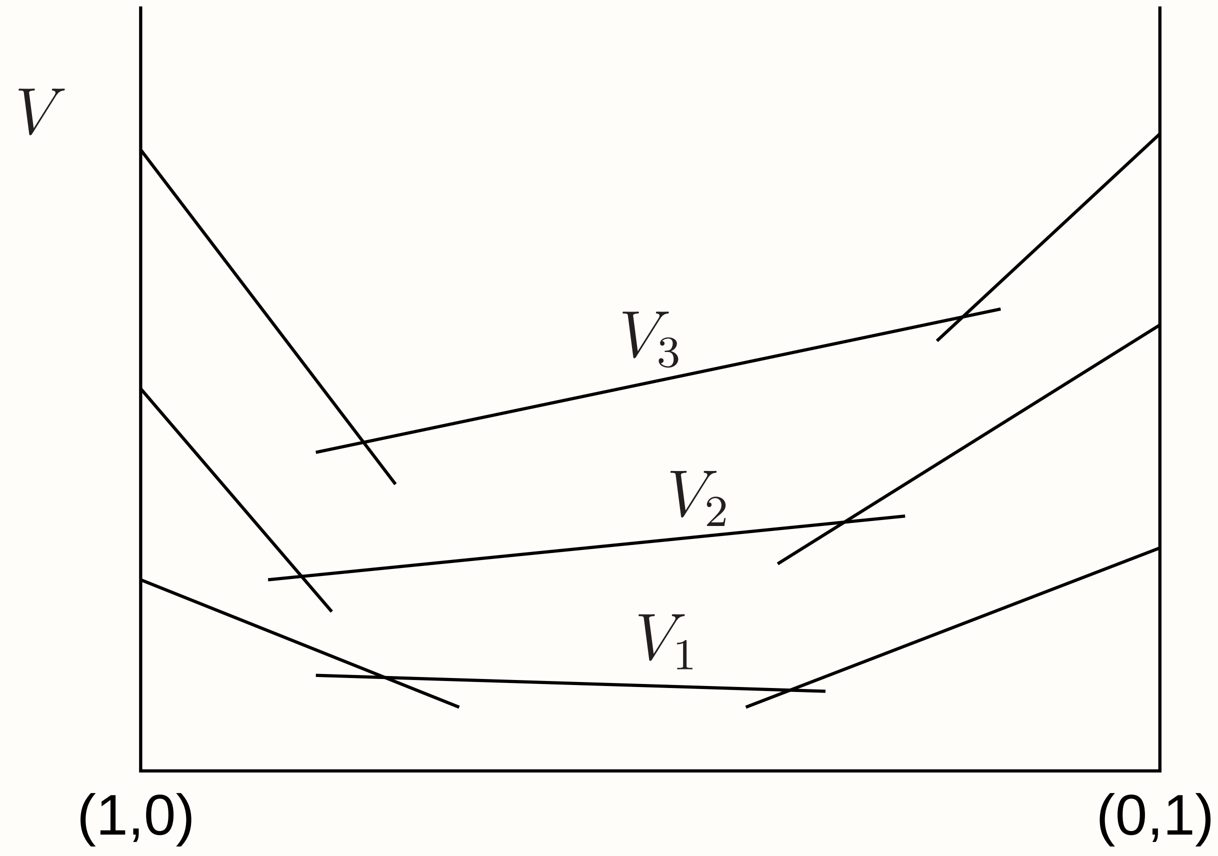
Optimal value functions

The optimal value function of a (finite horizon) POMDP is **piecewise linear** and **convex**: $V(b) = \max_\alpha b \cdot \alpha$.



Exact value iteration

Value iteration computes a sequence of value function estimates: $V_1, V_2, \ldots, V_n$.



Optimal POMDP methods

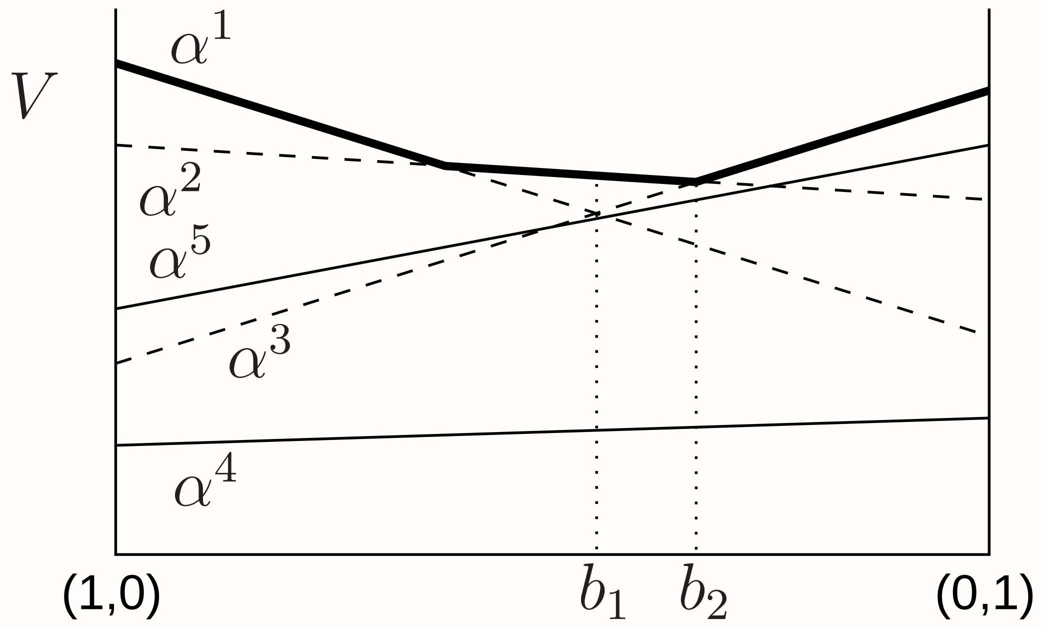
Enumerate and prune:

- Most straightforward: Monahan (1982)'s enumeration algorithm. Generates a maximum of $|A||V_n|^{|O|}$ vectors at each iteration, hence requires pruning.
- Incremental pruning (Zhang and Liu, 1996; Cassandra et al., 1997).

Search for witness points:

- One Pass (Sondik, 1971; Smallwood and Sondik, 1973).
- Relaxed Region, Linear Support (Cheng, 1988).
- Witness (Cassandra et al., 1994).

Vector pruning



Linear program for pruning:

variables: $\forall s \in S, b(s); x$

maximize: x

subject to:

$$b \cdot (\alpha - \alpha') \geq x, \forall \alpha' \in V, \alpha' \neq \alpha$$
$$b \in \Delta(S)$$

High dimensional sensor readings

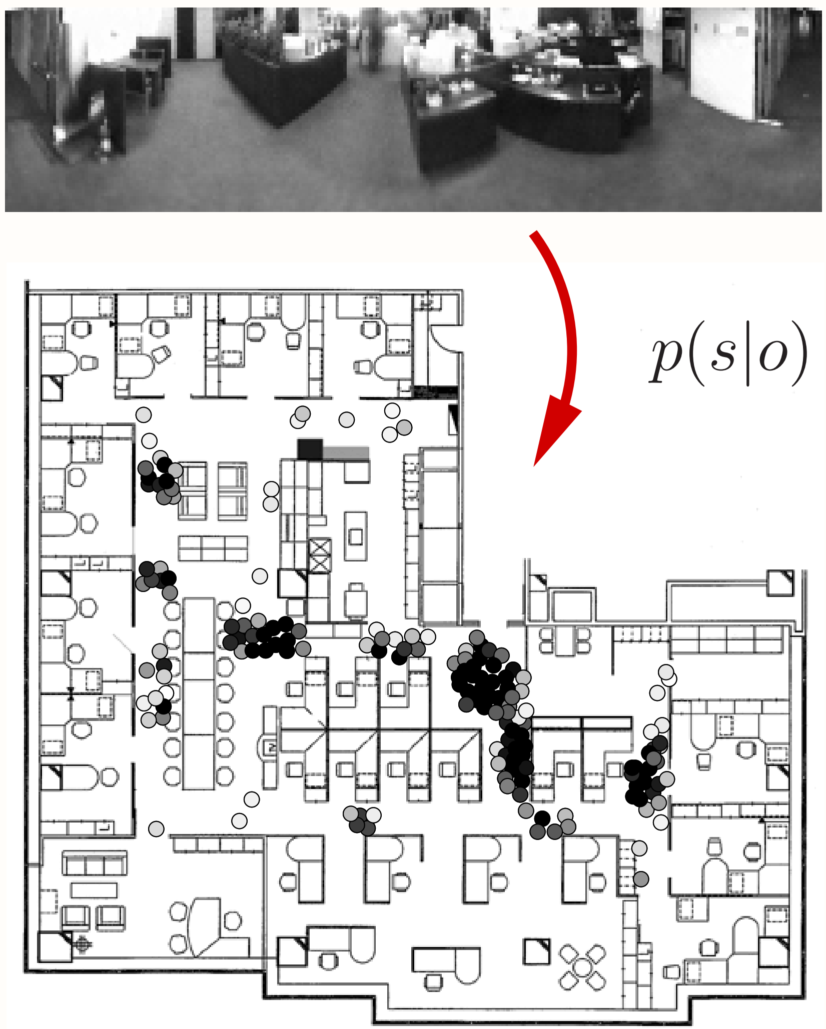
Omnidirectional camera images.

Example images ⇒

Dimension reduction:

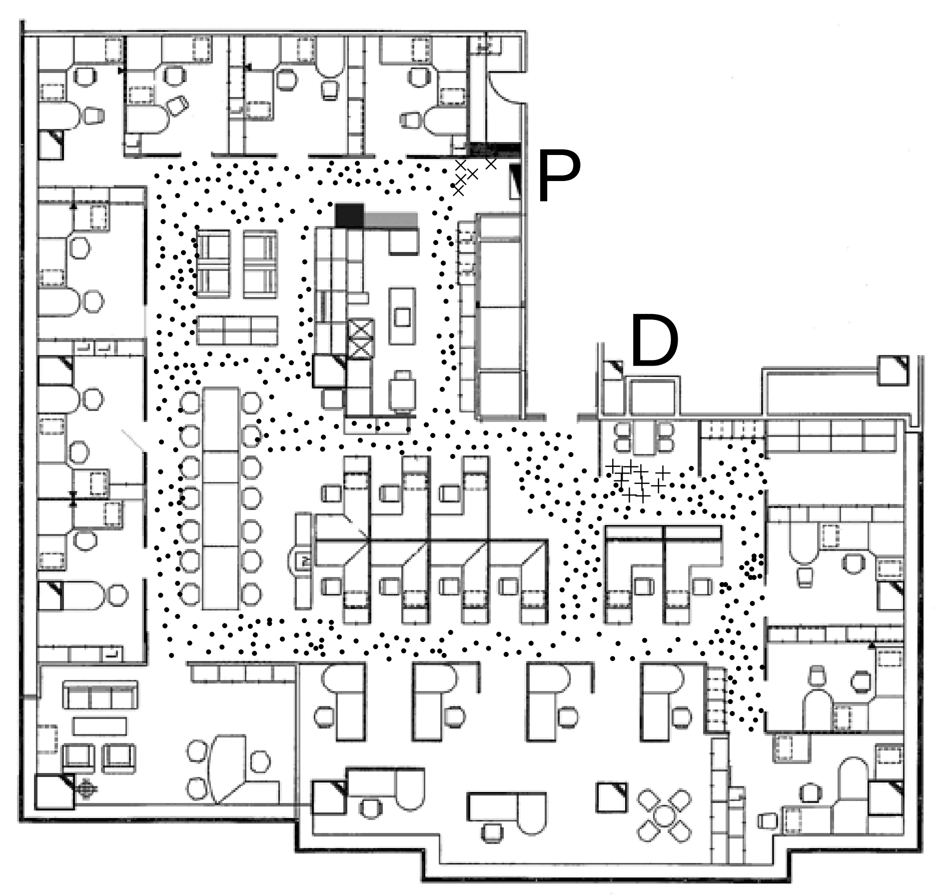
- Collect a database of images and record their location.
- Apply Principal Component Analysis on the image data.
- Project each image to the first 3 eigenvectors, resulting in a 3D feature vector for each image.

Observation model



- We cluster the feature vectors into 10 prototype observations.
- We compute a discrete observation model $p(o|s, a)$ by a histogram operation.

States, actions and rewards



- State: $s = (x, j)$ with x the robot's location and j the mail bit.
- Grid X into 500 locations.
- Actions: $\{\uparrow, \rightarrow, \downarrow, \leftarrow,$ *pickup*, *deliver*$\}$.
- Positive reward: only upon successful mail delivery.

References

D. Aberdeen and J. Baxter. Scaling internal-state policy-gradient methods for POMDPs. In *International Conference on Machine Learning*, 2002.

Y. Aviv and A. Pazgal. A partially observed Markov decision process for dynamic pricing. *Management Science*, 51(9):1400–1416, 2005.

J. Boger, P. Poupart, J. Hoey, C. Boutilier, G. Fernie, and A. Mihailidis. A decision-theoretic approach to task assistance for persons with dementia. In *Proc. Int. Joint Conf. on Artificial Intelligence*, 2005.

B. Bonet. An epsilon-optimal grid-based algorithm for partially observable Markov decision processes. In *International Conference on Machine Learning*, 2002.

R. I. Brafman. A heuristic variable grid solution method for POMDPs. In *Proc. of the National Conference on Artificial Intelligence*, 1997.

A. R. Cassandra, L. P. Kaelbling, and M. L. Littman. Acting optimally in partially observable stochastic domains. In *Proc. of the National Conference on Artificial Intelligence*, 1994.

A. R. Cassandra, L. P. Kaelbling, and J. A. Kurien. Acting under uncertainty: Discrete Bayesian models for mobile robot navigation. In *Proc. of International Conference on Intelligent Robots and Systems*, 1996.

A. R. Cassandra, M. L. Littman, and N. L. Zhang. Incremental pruning: A simple, fast, exact method for partially observable Markov decision processes. In *Proc. of Uncertainty in Artificial Intelligence*, 1997.

H. T. Cheng. *Algorithms for partially observable Markov decision processes*. PhD thesis, University of British Columbia, 1988.

T. Darrell and A. Pentland. Active gesture recognition using partially observable Markov decision processes. In *Proc. of the 13th Int. Conf. on Pattern Recognition*, 1996.

A. W. Drake. *Observation of a Markov process through a noisy channel*. Sc.D. thesis, Massachusetts Institute of Technology, 1962.

J. H. Ellis, M. Jiang, and R. Corotis. Inspection, maintenance, and repair with partial observability. *Journal of Infrastructure Systems*, 1(2):92–99, 1995.

E. A. Hansen. *Finite-memory control of partially observable systems*. PhD thesis, University of Massachusetts, Amherst, 1998a.

E. A. Hansen. Solving POMDPs by searching in policy space. In *Proc. of Uncertainty in Artificial Intelligence*, 1998b.

M. Hauskrecht and H. Fraser. Planning treatment of ischemic heart disease with partially observable Markov decision processes. *Artificial Intelligence in Medicine*, 18:221–244, 2000.

C. Hu, W. S. Lovejoy, and S. L. Shafer. Comparison of some suboptimal control policies in medical drug therapy. *Operations Research*, 44(5):696–709, 1996.

L. P. Kaelbling, M. L. Littman, and A. R. Cassandra. Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, 101:99–134, 1998.

M. L. Littman, A. R. Cassandra, and L. P. Kaelbling. Learning policies for partially observable environments: Scaling up. In *International Conference on Machine Learning*, 1995.

W. S. Lovejoy. Computationally feasible bounds for partially observed Markov decision processes. *Operations Research*, 39(1):162–175, 1991.

G. E. Monahan. A survey of partially observable Markov decision processes: theory, models and algorithms. *Management Science*, 28(1), Jan. 1982.

A. Y. Ng and M. Jordan. PEGASUS: A policy search method for large MDPs and POMDPs. In *Proc. of Uncertainty in Artificial Intelligence*, 2000.

R. Parr and S. Russell. Approximating optimal policies for partially observable stochastic domains. In *Proc. Int. Joint Conf. on Artificial Intelligence*, 1995.

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

J. Pineau, M. Montemerlo, M. Pollack, N. Roy, and S. Thrun. Towards robotic assistants in nursing homes: Challenges and results. *Robotics and Autonomous Systems*, 42(3–4):271–281, 2003b.

L. K. Platzman. A feasible computational approach to infinite-horizon partially-observed Markov decision problems. Technical Report J-81-2, School of Industrial and Systems Engineering, Georgia Institute of Technology, 1981. Reprinted in working notes AAAI 1998 Fall Symposium on Planning with POMDPs.

P. Poupart and C. Boutilier. Bounded finite state controllers. In *Advances in Neural Information Processing Systems 16*. MIT Press, 2004.

P. Poupart and C. Boutilier. Value-directed compression of POMDPs. In *Advances in Neural Information Processing Systems 15*. MIT Press, 2003.

N. Roy, J. Pineau, and S. Thrun. Spoken dialog management for robots. In *Proc. of the Association for Computational Linguistics*, 2000.

N. Roy, G. Gordon, and S. Thrun. Finding approximate POMDP solutions through belief compression. *Journal of Artificial Intelligence Research*, 23:1–40, 2005.

P. Rusmevichientong and B. Van Roy. A tractable POMDP for a class of sequencing problems. In *Proc. of Uncertainty in Artificial Intelligence*, 2001.

J. K. Satia and R. E. Lave. Markovian decision processes with probabilistic observation of states. *Management Science*, 20(1), 1973.

R. Simmons and S. Koenig. Probabilistic robot navigation in partially observable environments. In *Proc. Int. Joint Conf. on Artificial Intelligence*, 1995.

S. Singh, T. Jaakkola, and M. Jordan. Learning without state-estimation in partially observable Markovian decision processes. In *International Conference on Machine Learning*, 1994.

R. D. Smallwood and E. J. Sondik. The optimal control of partially observable Markov decision processes over a finite horizon. *Operations Research*, 21:1071–1088, 1973.

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

E. J. Sondik. *The optimal control of partially observable Markov processes*. PhD thesis, Stanford University, 1971.

M. T. J. Spaan and N. Vlassis. Perseus: Randomized point-based value iteration for POMDPs. *Journal of Artificial Intelligence Research*, 24:195–220, 2005.

G. Theocharous and S. Mahadevan. Approximate planning with hierarchical partially observable Markov decision processes for robot navigation. In *Proceedings of the IEEE International Conference on Robotics and Automation*, 2002.

J. T. Treharne and C. R. Sox. Adaptive inventory control for nonstationary demand and partial information. *Management Science*, 48(5):607–624, 2002.

N. L. Zhang and W. Liu. Planning in stochastic domains: problem characteristics and approximations. Technical Report HKUST-CS96-31, Department of Computer Science, The Hong Kong University of Science and Technology, 1996.

R. Zhou and E. A. Hansen. An improved grid-based approximation algorithm for POMDPs. In *Proc. Int. Joint Conf. on Artificial Intelligence*, 2001.