Published in the proceedings of the Workshop on "Structure & Priors in Reinforcement Learning" at ICLR 2019

PERCEPTION-AWARE POINT-BASED VALUE ITERATION FOR PARTIALLY OBSERVABLE MARKOV DECISION PROCESSES

Mahsa Ghasemi
Department of Mechanical Engineering
University of Texas at Austin
Austin, Texas 78712, USA
mahsa.ghasemi@utexas.edu

Ufuk Topcu
Department of Aerospace Engineering
and Engineering Mechanics
University of Texas at Austin
Austin, Texas 78712, USA
utopcu@utexas.edu

ABSTRACT

In conventional partially observable Markov decision processes, the observations that the agent receives originate from fixed known distributions. However, in a variety of real-world scenarios, the agent has an active role in its perception by selecting which observations to receive, leading to a combinatorial expansion of the action space. We use the structural decomposition of the action space to develop a novel and computationally efficient point-based value iteration algorithm. We prove that the proposed algorithm outputs a near-optimal value function and demonstrate its performance empirically.

1 INTRODUCTION

In the era of information explosion it is crucial to develop decision-making platforms that are able to judiciously extract useful information to accomplish a defined task. Such problems are composed of both an active perception element and a planning element, and appear in many applications including artificial intelligence, robotics, networked systems and internet of things.

We address joint perception and planning in partially observable Markov decision processes (POMDPs). Our main contribution is establishing near-optimal and tractable solutions for a class of problems where perception is defined as picking a constrained subset of information sources. We prove that it is possible to decouple the perception action space and the planning action space yet still achieve near-optimal strategies. To that end, we formulate the joint active perception and planning problem for POMDPs, develop a perception-aware point-based value iteration algorithm, and establish its theoretical guarantees.

The class of active perception considered in this paper, i.e., picking the most useful information sources, resembles the well-established problem of subset selection (Krause & Guestrin, 2007; Krause & Golovin, 2014; Qian et al., 2017). This type of active perception arises in various applications in control systems, robotics, and machine learning, where the constraints on sensing stem from power, processing capability, or communication limits.

1.1 RELATED WORK

Prior work such as Spaan (2008); Spaan & Lima (2009); Natarajan et al. (2015) model active perception as a POMDP. However, the most relevant work to ours is that of Araya et al. (2010); Spaan et al. (2015); Satsangi et al. (2018). Araya et al. (2010) proposed ρPOMDP framework where the reward depends on the entropy of the belief. Spaan et al. (2015) introduced POMDP-IR where the reward depends on an accurate prediction about the state. Satsangi et al. (2018) employed the submodularity of the underlying value function to use greedy scheme for sensor selection. The main difference of our work is that we consider active perception as a means to accomplishing the original task while in these work, active perception is the task itself and hence the POMDP rewards are metrics to capture perception quality.

2 PROBLEM FORMULATION

We introduce a new class of POMDP models, called AP^2-POMDP, that are suitable for problems with both elements of active perception and planning.

Definition 1. *An AP^2-POMDP is a tuple $\mathcal{P} = (S, A, k, T, \Omega, O, R, \gamma)$. S is the finite set of states. $A = A^{pl} \times A^{pr}$ denotes the finite set of paired actions with A^{pl} being the set of planning actions and A^{pr} being the set of perception actions. $A^{pr} = \{\delta \in \{0,1\}^n : \|\delta\|_0 \leq k\}$ constructs an n-dimensional lattice where k is the maximum number of information sources to be activated. Each component of an action $\delta \in A^{pr}$ determines whether to activate the corresponding information source, e.g. sensor. Let $\zeta(\delta) = \{i : \delta(i) = 1\}$ to denote the subset of information sources that are selected by δ. $T : S \times A^{pl} \times S \to [0,1]$ denotes the probabilistic transition function. $\Omega = \Omega^1 \times \Omega^2 \times \ldots \times \Omega^n$ is the partitioned set of observations, where each Ω_i corresponds to the set of measurements observable by information source i. $O : S \times A \times \Omega \to [0,1]$ denotes the probabilistic observation function. $R : S \times A^{pl} \to \mathbb{R}$ is the reward function, and $\gamma \in [0,1]$ is the discount factor.*

In many practical settings, the measurements from information sources only depend on the state and the previous action, as formally stated below.

Assumption 1. *We assume that given the current state and the previous action, the observations from information sources are mutually independent, i.e., $\forall I_1, I_2 \subseteq \{1, 2, \ldots, n\}$, $I_1 \cap I_2 = \emptyset$: $Pr(\bigcup_{i_1 \in I_1} \omega^{i_1}, \bigcup_{i_2 \in I_2} \omega^{i_2} | s, \beta) = Pr(\bigcup_{i_1 \in I_1} \omega^{i_1} | s, \beta) Pr(\bigcup_{i_2 \in I_2} \omega^{i_2} | s, \beta)$.*

Given the initial belief b_0, the following update equation holds between previous belief b and the belief $b_b^{'a,\omega}$ after taking action $a = (\beta, \delta)$ and receiving observation ω:

$$b_b^{'a,\omega}(s') = \frac{Pr(\omega|s', \beta, \delta) \sum_s Pr(s'|s, \beta) b(s)}{Pr(\omega|\beta, \delta)} = \frac{\prod_{i \in \zeta(\delta)} O_i(s', \beta, \omega^i) \sum_s T(s, \beta, s') b(s)}{\sum_{s''} \prod_{i \in \zeta(\delta)} O_i(s'', \beta, \omega^i) \sum_s T(s, \beta, s'') b(s)}. \tag{1}$$

The goal is to learn a pure policy to maximize $\mathbb{E}[\sum_{t=0}^{\infty} \gamma^t R(s_t, \beta_t) | b_0]$ where $\beta_t \in A^{pl}$. A pure policy is a mapping from beliefs to actions $\pi : B \to A$, where B is the set of beliefs that constructs a $(|S| - 1)$-dimensional probability simplex.

The POMDP solvers apply value iteration (Sondik, 1978), a dynamic programming technique, to find an optimal policy. Let V be a value function that maps beliefs to values in $\mathbb{R}$. The following recursive expression holds for V:

$$V_t(b) = \max_a \left(\sum_{s \in S} b(s) R(s, a) + \gamma \sum_{\omega \in \Omega} Pr(\omega|b, a) V_{t-1}(b_b^{'a,\omega}) \right). \tag{2}$$

The value iteration converges to the optimal value function V^* which satisfies the Bellman optimality equation (Bellman, 1957). Once the optimal value function is learned, an optimal policy can be derived. An important outcome of (2) is that at any horizon, the value function is piecewise-linear and convex (Smallwood & Sondik, 1973) and hence, can be represented by a finite set of hyperplanes. Each hyperplane is associated with an action. Let α denote the corresponding vector of a hyperplane and let Γ_t to be the set of α vectors at horizon t. Then,

$$V_t(b) = \max_{\alpha \in \Gamma_t} \alpha \cdot b. \tag{3}$$

This fact has motivated approximate point based solvers that try to approximate the value function by updating the hyperplanes over a finite set of belief points.

Next, we formulate the joint perception and planning problem.

Problem 1. *Let $\mathcal{P} = (S, A, k, T, \Omega, O, R, \gamma)$ to denote an AP^2-POMDP and b_0 to be an initial belief. The goal is to learn a pure belief-based policy $\pi(b) = (\beta, \delta)$ such that the expected discounted cumulative reward is maximized, i.e, $\pi^* = \text{argmax}_\pi \, \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t R(s_t, \pi(b_t)) | b_0\right]$.*

3 ACTIVE PERCEPTION WITH GREEDY SCHEME

For variety of performance metrics, finding an optimal subset of information sources poses a computationally challenging combinatorial optimization problem that is NP-hard (Williams & Young, 2007). Augmenting POMDP planning actions with $\binom{n}{k}$ active perception actions results in a combinatorial expansion of the action space. Thereupon, it is infeasible to directly apply existing POMDP solvers to Problem 1. Instead of concatenating both sets of actions and treating them similarly, we propose a greedy strategy for selecting perception actions that aims to pick the information sources that result in minimal uncertainty about the state. The key enabling factor is that the perception actions does not affect the transition, consequently, we can decompose the single-step belief update in (1) into two steps:

Algorithm 1 Greedy policy for perception action

Input: AP^2-POMDP $\mathcal{P} = (S, A, k, T, \Omega, O, R, \gamma)$, Current belief b, Planning action β.
Output: Perception action δ.
Initialize $\mathcal{X} = \{1, 2, \ldots, n\}$, $\zeta = \emptyset$.
for $l = 1, \ldots, k$ **do**
  $j^* = \text{argmax}_{j \in \mathcal{X} \setminus \zeta} -\mathcal{H}(\mathbf{s}|\tilde{b}_b^\beta, \bigcup_{i \in \zeta \cup \{j\}} \boldsymbol{\omega}^i)$
  $\zeta \leftarrow \zeta \cup \{j^*\}$
end for
return δ corresponding to ζ.

$$\tilde{b}_b^\beta(s') = \sum_s T(s, \beta, s')\tilde{b}(s), \qquad b'^{\delta,\omega}_{\tilde{b}}(s'') = \frac{\prod_{i \in \zeta(\delta)} O_i(s'', \beta, \omega^i)\tilde{b}(s'')}{\sum_{s'} \prod_{i \in \zeta(\delta)} O_i(s', \beta, \omega^i)\tilde{b}(s')}. \tag{4}$$

This in turn implies that after a transition is made, the agent should pick a subset of observations that lead to minimal uncertainty in $b'^{\delta,\omega}_{\tilde{b}}$.

Algorithm 2 BackUp step for AP^2-POMDP

Input: AP^2-POMDP $\mathcal{P} = (S, A, k, T, \Omega, O, R, \gamma)$, Current set of belief points B_t, Current set of α vectors Γ_{t-1}.
Output: Next set of α vectors Γ_t.
Initialize $\Gamma_t = \emptyset$, $\Gamma_t^{b,\beta} = \emptyset$ for all $b \in B_t$ and $\beta \in A^{pl}$.
for $\beta \in A^{pl}$ **do**
  $\Gamma_t^{\beta,*} \leftarrow \alpha^{\beta,*}(s) = R(s, \beta)$
  for $b \in B_t$ **do**
    $\bar{\delta} = \text{Greedy\_argmax}_{\delta \in A^{pr}} f(\zeta(\delta))$; $\Gamma_t^{b,\beta,\omega} = \emptyset$
    for $\omega \in \Omega_{i_1} \times \ldots \times \Omega_{i_k}, i_j \in \zeta(\bar{\delta})$ **do**
      for $\alpha \in \Gamma_{t-1}$ **do**
        $\alpha^{b,\beta,\omega}(s) = \gamma \sum_{s' \in S} \prod_{i_j \in \zeta(\bar{\delta})} O_i(s', \beta, \omega^{i_j}) T(s, \beta, s')\alpha(s')$; $\Gamma_t^{b,\beta,\omega} \leftarrow \Gamma_t^{b,\beta,\omega} \cup \alpha^{b,\beta,\omega}$
      end for
    end for
    $\alpha^{b,\beta} = \alpha^{\beta,*} + \sum_{\substack{\omega \in \Omega_{i_1} \times \ldots \times \Omega_{i_k} \\ i_j \in \zeta(\bar{\delta})}} \text{argmax}_{\alpha \in \Gamma_t^{b,\beta,\omega}} \alpha.b$; $\Gamma_t^{b,\beta} \leftarrow \Gamma_t^{b,\beta} \cup \alpha^{b,\beta}$
  end for
end for
for $b \in B_t$ **do**
  $\alpha^b = \text{argmax}_{\alpha \in \Gamma_t^{b,\beta}, \beta \in A^{pl}} \alpha.b$; $\Gamma_t = \Gamma_t \cup \alpha^b$
end for
return Γ_t.

To quantify state uncertainty, we use Shannon entropy of the belief. Since the observation values are unknown before selecting the sensors, we optimize conditional entropy that yields the expected value of entropy. With some algebraic manipulation, one obtains the conditional entropy of state

given current belief with respect to δ as:

$$\mathcal{H}(\mathbf{s}|b,\delta) = -\sum_{\omega^{i_1}\in\Omega^{i_1}}\cdots\sum_{\omega^{i_k}\in\Omega^{i_k}}\sum_{s\in S}\left(b(s)\prod_{i_j\in\zeta(\delta)}O_{i_j}(s,\beta,\omega^{i_j})\right.$$
$$\left.\log\left(\frac{b(s)\prod_{i_j\in\zeta(\delta)}O_{i_j}(s,\beta,\omega^{i_j})}{\sum_{s'\in S}b(s')\prod_{i_j\in\zeta(\delta)}O_{i_j}(s',\beta,\omega^{i_j})}\right)\right), \tag{5}$$

where $\zeta(\delta) = \{i_1, i_2, \ldots, i_k\}$. It is worth mentioning that b is the current distribution of $\mathbf{s}$ and is explicitly written only for the purpose of better clarity, otherwise, $\mathcal{H}(\mathbf{s}|b,\delta) = \mathcal{H}(\mathbf{s}|\delta)$. To minimize entropy, we define the objective function as the following set function:

$$f(\zeta) = \mathcal{H}(\mathbf{s}|\tilde{b}_b^\beta) - \mathcal{H}(\mathbf{s}|\tilde{b}_b^\beta, \bigcup_{i\in\zeta}\boldsymbol{\omega}^i) \tag{6}$$

and the optimization problem as:

$$\delta^* = \arg\max_{\delta\in A^{pr}} f(\zeta(\delta)). \tag{7}$$

We propose a greedy algorithm, outlined in Algorithm 1 to find an efficient solution to (7). The guarantee for the performance of the proposed greedy algorithm is stated in the next theorem.

Theorem 1. *Let ζ^* denote the optimal subset of observations with regard to objective function $f(\zeta)$, and ζ^g denote the output of the greedy algorithm in Algorithm 1. Then, the following performance guarantee holds:*

$$\mathcal{H}(\mathbf{s}|\tilde{b}_b^\beta, \bigcup_{i\in\zeta^g}\boldsymbol{\omega}^i) \leq \frac{1}{e}\mathcal{H}(\mathbf{s}|\tilde{b}_b^\beta) + \left(1-\frac{1}{e}\right)\mathcal{H}(\mathbf{s}|\tilde{b}_b^\beta, \bigcup_{i\in\zeta^*}\boldsymbol{\omega}^i). \tag{8}$$

Although Theorem 1 proves that the entropy of the belief point achieved by the greedy algorithm is close to the entropy of the belief point from the optimal solution, the key question is whether the value of these points are close. We prove that at each time step, in expectation, the value from greedy scheme is close to the value from optimal selection with regard to (7).

Theorem 2. *Let the agent's current belief to be b and its planning action to be β. Consider the optimization problem in* (7)*, and let δ^* and δ^g denote the optimal perception action and the perception action obtained by the greedy algorithm, respectively. It holds that $\mathbb{E}[\|b^g - b^*\|_1] \leq C_1$, where b^* and b^g are the updated beliefs according to* (4) *and C_1 is a constant value.*

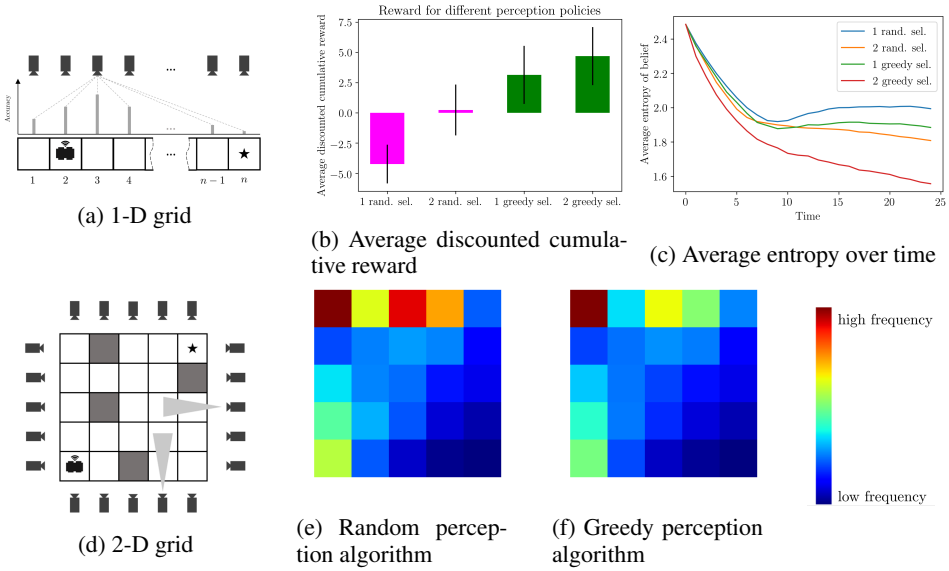
Theorem 3. *Instate the notation and hypothesis of Theorem 2. Additionally, let V to be the true value function for AP^2-POMDP. It holds that $\mathbb{E}[V(b^g) - V(b^*)] \leq C_2$, where C_2 depends on C_1 and parameters of the AP^2-POMDP.*

4 Perception-aware Point-based Value Iteration

We propose a novel point-based value iteration algorithm to approximate the value function for AP^2-POMDPs. The algorithm relies on the performance guarantee of the proposed greedy observation selection in previous section. The general procedure for a point-based solver consists of iterative belief point sampling, Bellman backup, and hyperplane pruning, until value function convergence (Araya et al., 2010). We develop a new BackUp step for AP^2-POMDPs that can be combined with any sampling and pruning method in other solvers, such as the ones developed by Spaan & Vlassis (2005), Kurniawati et al. (2008), and Smith & Simmons (2012).

In point-based solvers each witness belief point is associated with an α vector and an action. Nevertheless, for AP^2-POMDPs, each witness point is associated with two actions, β and δ. We compute δ based on greedy maximization of (7) so that given b and β, δ is uniquely determined. Henceforth, we can rewrite (2) using (3) to obtain:

$$V_t(b) = \max_\beta\left(\sum_{s\in S}b(s)R(s,\beta) + \gamma\sum_{\substack{\omega\in\Omega_{i_1}\times\ldots\times\Omega_{i_k}\\ i_j\in\zeta(\bar{\delta})}}\max_{\alpha\in\Gamma_{t-1}}\sum_{s\in S}\sum_{s'\in S}\alpha(s')\times\right.$$
$$\left.\prod_{i_j\in\zeta(\bar{\delta})}O_i(s',\beta,\omega^{i_j})T(s,\beta,s')b(s)\right). \tag{9}$$



(a) 1-D grid

(b) Average discounted cumulative reward

(c) Average entropy over time

(d) 2-D grid

(e) Random perception algorithm

(f) Greedy perception algorithm

Figure 1: Simulation settings and results for 1-D and 2-D robotic navigation tasks.

where $\bar{\delta} = \text{argmax}_{\delta \in A^{pr}} f(\zeta(\delta))$ and f is computed at $\tilde{b}_b^\beta$.

Based on the derivation in (9), we develop the BackUp step detailed in Algorithm 2 to compute the new set of α vectors from the previous ones using Bellman backup operation.

5 Simulation Results

We implement the proposed solver for AP^2-POMDPs. We initialize the belief set by uniform sampling from Δ_B (Devroye, 1986) and keep it fixed. However, one can integrate any sampling method such as the ones proposed by Smith & Simmons (2012), and Kurniawati et al. (2008). The α vectors are initialized by $\frac{1}{1-\gamma}\text{min}_{s,a}R(s,a).Ones(|S|)$ (Shani et al., 2013).

The first scenario models a robot that is moving in a 1-D discrete environment. The robot can move to adjacent cells by its navigation actions $A^{pl} = \{left, right, stop\}$ and has probabilistic transitions. The robot relies on a set of cameras for localization. To model the effect of robot's position on the accuracy of cameras' measurements, we use a binomial distribution with its mean at the cell that camera is on. The robot's objective is to reach a specific cell in the map. For that purpose, at each time step, the robot picks a navigation action and selects k cameras from the set of n cameras. We evaluate the computed policy by running 1000 Monte Carlo simulations. The robot starts at the origin and its initial belief is uniform. Figure 1-(b) demonstrates the discounted cumulative reward, averaged over 1000 runs, for random selection of 1 and 2 cameras, and greedy selection of 1 and 2 cameras. It shows that the greedy selection significantly outperforms the random selection. Figure 1-(c) depicts the belief entropy over the time. The lower entropy of greedy selection, compared to random selection, shows less uncertainty of the robot when taking planning actions.

The second scenario is a 2-D variant of the first scenario. The navigation actions of the robot are $A^{pl} = \{up, right, down, left, stop\}$. The rest of the setting is similar to 1-D case, except now the robot must avoid the obstacles in the map. We applied the proposed solver with both random perception and greedy perception on the 2-D example. Figure 1-(e&f) illustrates the normalized frequency of visiting each state for each perception algorithm. It can be seen that the policy learned by greedy active perception leads to better obstacle avoidance.

Acknowledgments

This work was supported in part by DARPA grant D19AP00004 and ONR grants N00014-18-1-2829 and N00014-19-1-2054.

REFERENCES

Mauricio Araya, Olivier Buffet, Vincent Thomas, and Françcois Charpillet. A pomdp extension with belief-dependent rewards. In *Advances in neural information processing systems*, pp. 64–72, 2010.

Karl J Åström. Optimal control of markov processes with incomplete state information. *Journal of Mathematical Analysis and Applications*, 10(1):174–205, 1965.

Richard Bellman. A markovian decision process. *Journal of Mathematics and Mechanics*, pp. 679–684, 1957.

Hsien-Te Cheng. *Algorithms for partially observable Markov decision processes*. PhD thesis, University of British Columbia, 1988.

Thomas M Cover and Joy A Thomas. *Elements of information theory*. John Wiley & Sons, 2012.

Luc Devroye. Sample-based non-uniform random variate generation. In *Proceedings of the 18th conference on Winter simulation*, pp. 260–265. ACM, 1986.

Abolfazl Hashemi, Mahsa Ghasemi, Haris Vikalo, and Ufuk Topcu. A randomized greedy algorithm for near-optimal sensor scheduling in large-scale sensor networks. In *2018 Annual American Control Conference (ACC)*, pp. 1027–1032. IEEE, 2018.

Chun-Wa Ko, Jon Lee, and Maurice Queyranne. An exact algorithm for maximum entropy sampling. *Operations Research*, 43(4):684–691, 1995.

Andreas Krause and Daniel Golovin. Submodular function maximization. In *Tractability: Practical Approaches to Hard Problems*, pp. 71–104. Cambridge University Press, 2014.

Andreas Krause and Carlos Guestrin. Near-optimal observation selection using submodular functions. In *AAAI*, volume 7, pp. 1650–1654, 2007.

Chris Kreucher, Keith Kastella, and Alfred O Hero Iii. Sensor management using an active sensing approach. *Signal Processing*, 85(3):607–624, 2005.

Hanna Kurniawati, David Hsu, and Wee Sun Lee. Sarsop: Efficient point-based pomdp planning by approximating optimally reachable belief spaces. In *Robotics: Science and systems*, volume 2008. Zurich, Switzerland., 2008.

William S Lovejoy. A survey of algorithmic methods for partially observed markov decision processes. *Annals of Operations Research*, 28(1):47–65, 1991.

Prabhu Natarajan, Pradeep K Atrey, and Mohan Kankanhalli. Multi-camera coordination and control in surveillance systems: A survey. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 11(4):57, 2015.

George L Nemhauser, Laurence A Wolsey, and Marshall L Fisher. An analysis of approximations for maximizing submodular set functionsi. *Mathematical programming*, 14(1):265–294, 1978.

Christos H Papadimitriou and John N Tsitsiklis. The complexity of markov decision processes. *Mathematics of operations research*, 12(3):441–450, 1987.

Joelle Pineau, Geoffrey Gordon, and Sebastian Thrun. Anytime point-based approximations for large pomdps. *Journal of Artificial Intelligence Research*, 27:335–380, 2006.

Chao Qian, Jing-Cheng Shi, Yang Yu, and Ke Tang. On subset selection with general cost constraints. In *IJCAI*, volume 17, pp. 2613–2619, 2017.

Yash Satsangi, Shimon Whiteson, Frans A Oliehoek, and Matthijs TJ Spaan. Exploiting submodular value functions for scaling up active perception. *Autonomous Robots*, 42(2):209–233, 2018.

Manohar Shamaiah, Siddhartha Banerjee, and Haris Vikalo. Greedy sensor selection: Leveraging submodularity. In *Decision and Control (CDC), 2010 49th IEEE Conference on*, pp. 2572–2577. IEEE, 2010.

Guy Shani, Joelle Pineau, and Robert Kaplow. A survey of point-based pomdp solvers. *Autonomous Agents and Multi-Agent Systems*, 27(1):1–51, 2013.

Richard D Smallwood and Edward J Sondik. The optimal control of partially observable markov processes over a finite horizon. *Operations research*, 21(5):1071–1088, 1973.

Trey Smith and Reid Simmons. Point-based pomdp algorithms: Improved analysis and implementation. *arXiv preprint arXiv:1207.1412*, 2012.

Edward J Sondik. The optimal control of partially observable markov processes over the infinite horizon: Discounted costs. *Operations research*, 26(2):282–304, 1978.

Matthijs TJ Spaan. Cooperative active perception using pomdps. In *AAAI 2008 workshop on advancements in POMDP solvers*, 2008.

Matthijs TJ Spaan and Pedro U Lima. A decision-theoretic approach to dynamic sensor selection in camera networks. In *ICAPS*, 2009.

Matthijs TJ Spaan and Nikos Vlassis. Perseus: Randomized point-based value iteration for pomdps. *Journal of artificial intelligence research*, 24:195–220, 2005.

Matthijs TJ Spaan, Tiago S Veiga, and Pedro U Lima. Decision-theoretic planning under uncertainty with information rewards for active cooperative perception. *Autonomous Agents and Multi-Agent Systems*, 29(6):1157–1185, 2015.

Jason D Williams and Steve Young. Partially observable markov decision processes for spoken dialog systems. *Computer Speech & Language*, 21(2):393–422, 2007.

Nevin Lianwen Zhang and Weihong Zhang. Speeding up the convergence of value iteration in partially observable markov decision processes. *Journal of Artificial Intelligence Research*, 14: 29–51, 2001.