On the Computational Complexity of Stochastic Controller Optimization in POMDPs*

Nikos Vlassis† Michael L. Littman‡ David Barber§

October 5, 2012

Abstract

We show that the problem of finding an optimal stochastic "blind" controller in a Markov decision process is an NP-hard problem. The corresponding decision problem is NP-hard, in PSPACE, and SQRT-SUM-hard, hence placing it in NP would imply breakthroughs in long-standing open problems in computer science. Our result establishes that the more general problem of stochastic controller optimization in POMDPs is also NP-hard. Nonetheless, we outline a special case that is convex and admits efficient global solutions.

Keywords: Partially observable Markov decision process, stochastic controller, bilinear program, computational complexity, Motzkin-Straus theorem, sum-of-square-roots problem, matrix fractional program, computations on polynomials, nonlinear optimization.

1 Introduction

Partially observable Markov decision processes (POMDPs) have proven to be a valuable conceptual tool for problems throughout AI, including reinforcement learning (Chrisman, 1992), planning under uncertainty (Kaelbling et al., 1998), and multiagent coordination (Bernstein et al., 2005). Briefly, a POMDP is a Markov decision process in which the decision maker is unable to perceive its current state directly, but has access to an observation function that relates states to observations. An important problem here is deciding how to select actions to minimize expected cost given the state uncertainty. Unfortunately, this problem is extremely challenging (Papadimitriou and Tsitsiklis, 1987; Mundhenk et al., 2000). In fact, the exact problem is unsolvable in the general case (Madani et al., 1999).

An alternative to finding optimal policies for POMDPs is to find low cost *controllers*—mappings from observation histories to actions (Sondik, 1971; Platzman, 1981). A restricted space of controllers can, in principle, be considerably easier to search than the space of all possible policies (Littman et al., 1998; Hansen, 1998; Meuleau et al., 1999). Various methods for controller optimization in POMDPs have been proposed in the literature, both for stochastic as

*To appear in ACM Transactions on Computation Theory, 4(4), 2012.
†Luxembourg Centre for Systems Biomedicine, Univ. of Luxembourg (nikos.vlassis@uni.lu)
‡Department of Computer Science, Brown University (mlittman@cs.brown.edu)
§Department of Computer Science, University College London (d.barber@cs.ucl.ac.uk)

well as for deterministic controllers: exhaustive search (Smith, 1971), branch and bound (Hastings and Sadjadi, 1979; Littman, 1994), local seach (Poupart and Boutilier, 2004; Serin and Kulkarni, 2005), constrained quadratic programming (Amato et al., 2007), and the EM algorithm (Toussaint et al., 2011).

A variety of complexity results are known for the problem of controller optimization in POMDPs. Most versions are known to be hard for classes that are believed to be above P (Papadimitriou and Tsitsiklis, 1987; Mundhenk et al., 2000). The computational decision problem asks, for a given controller class and a target cost, whether the target cost can be achieved by a controller in that class. Here, we consider several such controller classes.

Deterministic Time/History-Dependent Controller Such a controller chooses an action based on the current time period and/or the history of previous actions and observations. The problem is NP-complete or PSPACE-complete (Papadimitriou and Tsitsiklis, 1987; Mundhenk et al., 2000). In the remaining classes we assume stationary controllers.

Deterministic Controller of Polynomial Size Such a controller is represented by a graph in which nodes are labeled with actions and edges are labeled with observations. The problem is in NP in that we can guess a controller of the right size, then see if it incurs no more than the target cost by solving a system of linear equations. It is NP-hard even for the "easier" completely observable version (Littman et al., 1998).

Stochastic Controller of Polynomial Size This class extends deterministic controllers by allowing a probability distribution over actions at each node. There are POMDPs for which a stochastic controller of a given size can outperform any deterministic controller of the same size (Singh et al., 1994). In this article we show that this problem is NP-hard, in PSPACE, and SQRT-SUM-hard, hence showing it lies in NP would imply breakthroughs in long-standing open problems (Allender et al., 2009; Etessami and Yannakakis, 2010).

Deterministic Memoryless Controller A memoryless controller chooses an action based on the most recent observation only. These controllers are a special case of deterministic controllers with polynomial size as they can be represented as a graph with one node per observation. The problem is NP-complete (Littman, 1994; Papadimitriou and Tsitsiklis, 1987).

Stochastic Memoryless Controller These controllers are defined by a probability distribution over actions for each observation. They can be considerably more effective than the corresponding deterministic memoryless controllers. They are a generalization of the blind controllers we consider in this article, and it follows from our results that the problem is NP-hard, in PSPACE, and SQRT-SUM-hard.

Deterministic Blind Controller A blind controller for a POMDP is equivalent to a memoryless controller for an unobserved MDP. A deterministic blind controller consists of a single action that is applied (blindly) regardless of the observation history. It is straightforward to evaluate a deterministic blind controller—simply drop all actions but one from the POMDP and evaluate the resulting Markov chain. Thus, the decision problem for deterministic blind controllers is trivially in P as an algorithm can simply check each action to see which is best.

Stochastic Blind Controller Such a controller is a probability distribution over actions to be applied repeatedly at every timestep. This is the class of controllers we consider in this article. Again, the added power of stochasticity allows for much more effective policies to be constructed. However, as we show in the remainder of this article, the added power comes with a very high cost. The decision problem is NP-hard, in PSPACE, and SQRT-SUM-hard.

2 MDPs and blind controllers

We consider a discounted, with discount factor $\gamma < 1$, infinite-horizon Markov decision process (MDP) characterized by n states and k actions, state-action costs (negative rewards) c_{sa}, and starting distribution (μ_s) with $\mu_s \geq 0$ and $\sum_{s=1}^{n} \mu_s = 1$. Let $p(\bar{s}|s,a)$ denote the probability to transition to state $\bar{s}$ when action a is taken at state s. The following linear program can be used to find an optimal policy for the MDP:

$$\begin{aligned} \min_{x_{sa} \geq 0} \ & \sum_{sa} x_{sa} c_{sa}, \\ \text{s.t.} \ & \sum_{a} x_{\bar{s}a} = (1-\gamma)\mu_{\bar{s}} + \gamma \sum_{sa} p(\bar{s}|s,a) x_{sa} \quad \forall_{\bar{s}}, \end{aligned} \tag{1}$$

where x_{sa} denotes occupancy distribution over state-action pairs, and the constraints are the Bellman occupancy constraints. From an optimal occupancy x^*_{sa}, we can compute an optimal stationary and deterministic policy that maps states to actions (Puterman, 1994).

We consider now the case where we constrain the class of allowed policies to stochastic "blind" controllers, in which the controller cannot observe or remember anything (state, action, or time), but can only randomize over actions using the same distribution $\boldsymbol{\pi} = (\pi_a)$ at each time step, where $\boldsymbol{\pi} \in \Delta$ and $\Delta = \{\boldsymbol{\pi} : \boldsymbol{\pi} \ \geq \ 0, \sum_{a=1}^{k} \pi_a = 1\}$ is the standard probability simplex. Note that, unlike standard MDP policies, a blind controller $\boldsymbol{\pi}$ is not a function of state. (The related notion of a memoryless controller is a function of POMDP observations, but still not of state.) Explicitly encoding the controller parametrization

in (1) gives:

$$\begin{aligned} \min_{\mathbf{x}\geq 0, \boldsymbol{\pi}\in\Delta} \; & \sum_{sa} x_s \pi_a c_{sa}, \\ \text{s.t.} \quad & x_{\bar{s}} = (1-\gamma)\mu_{\bar{s}} + \gamma \sum_a \pi_a \sum_s p(\bar{s}|s,a) x_s \quad \forall_{\bar{s}}, \end{aligned} \tag{2}$$

where $\mathbf{x} = (x_s)$ is an occupancy distribution over states, with $\mathbf{x} \geq 0$. Note that the occupancy vector $\mathbf{x}$ satisfies $\sum_s x_s = 1$. When viewed as a function of both $\mathbf{x}$ and $\boldsymbol{\pi}$, the above program is a jointly constrained bilinear program. Such programs involve bilinear terms (like $x_s \pi_a$) in both the objective function as well as in the constraints, and are in general nonconvex in the joint vector $(\mathbf{x}, \boldsymbol{\pi})$ (Al-Khayyal and Falk, 1983).

Bilinear programs are known to be NP-hard to solve to global optimality in general, but could there be some special structure in (2) that renders that particular program tractable? In the next section, we answer this question in the negative, showing that finding an optimal stochastic blind controller is an NP-hard problem.

3 NP-hardness result

The decision problem we are addressing is the following.

Definition 1 (The STOCHASTIC-BLIND-POLICY problem)**.** *Given a discounted MDP and a target cost r, is there a stochastic blind controller $\boldsymbol{\pi}$ that incurs cost $J(\boldsymbol{\pi}) \leq r$?*

Here $J(\boldsymbol{\pi}) = \mathbf{x}^\top \mathbf{C} \boldsymbol{\pi}$ is the cost of controller $\boldsymbol{\pi}$ in (2), where $\mathbf{C} = (c_{sa})$ is an $n \times k$ matrix containing all state-action costs, and $\mathbf{x} = (x_s)$ is an $n \times 1$ occupancy vector defined via the Bellman occupancy constraints in (2). Let also $\boldsymbol{\mu} = (\mu_s)$ denote the $n \times 1$ starting distribution vector.

Theorem 1. *The* STOCHASTIC-BLIND-POLICY *problem is NP-hard.*

Proof. We reduce from the INDEPENDENT-SET problem. This problem asks, for a given (undirected and with no self-loops) graph $G = (V, E)$ and a positive integer $j \leq |V|$, whether G contains an independent set V' having $|V'| \geq j$. This problem is NP-complete, even when restricted to cubic graphs (a cubic graph is a graph in which every node has degree three) (Garey and Johnson, 1979).

Let $\mathbf{G}$ be the $n \times n$ (symmetric, 0-1) adjacency matrix of an input cubic graph G (hence each column of $\mathbf{G}$ sums to three). The reduction constructs an MDP with n states and n actions, uniform starting distribution $\boldsymbol{\mu}$, cost matrix $\mathbf{C} = \frac{1}{\gamma}(\mathbf{G} + \mathbf{I})$ where $\mathbf{I}$ is the identity matrix, and deterministic transitions $p(\bar{s}|s,a) = 1$ if $\bar{s} = a$ and 0 otherwise (the action variable a can be viewed as indexing the state space). Since the transitions $p(\bar{s}|s,a)$ are independent of s,

the occupancy vector in (2) reduces to $\mathbf{x} = (1-\gamma)\boldsymbol{\mu} + \gamma\boldsymbol{\pi}$, and the cost function becomes the quadratic

$$J(\boldsymbol{\pi}) = \frac{4(1-\gamma)}{n\gamma} + \boldsymbol{\pi}^\top(\mathbf{G}+\mathbf{I})\boldsymbol{\pi}, \tag{3}$$

where we used the fact that the input graph G is cubic and $\boldsymbol{\mu}$ is uniform. Moreover, for any graph G it holds (Motzkin and Straus, 1965)

$$\frac{1}{\alpha(G)} = \min_{\mathbf{y}\in\Delta} \mathbf{y}^\top(\mathbf{G}+\mathbf{I})\mathbf{y}, \tag{4}$$

where $\alpha(G)$ is the size of the maximum independent set (the stability number) of the graph. Let the target cost be $r = \frac{1}{j} + \frac{4(1-\gamma)}{n\gamma}$. Then, $J(\boldsymbol{\pi}) \leq r$ is equivalent to $\boldsymbol{\pi}^\top(\mathbf{G}+\mathbf{I})\boldsymbol{\pi} \leq \frac{1}{j}$, and hence from (4) follows that the existence of a vector $\boldsymbol{\pi}$ that satisfies $J(\boldsymbol{\pi}) \leq r$ would imply $\frac{1}{\alpha(G)} \leq \frac{1}{j}$, and hence $\alpha(G) \geq j$, or, in other words, $|V'| \geq j$ for some independent set $V' \subseteq V$. □

4 Connection to the SQRT-SUM problem

Our STOCHASTIC-BLIND-POLICY problem is contained in PSPACE, as it can be expressed as a system of polynomial inequalities—any such system is known to be solvable in PSPACE (Canny, 1988). But, is there a tighter upper bound?

We will attempt to address this question indirectly, by establishing a connection between the STOCHASTIC-BLIND-POLICY problem and the SQRT-SUM problem. The SQRT-SUM problem asks, for a given list of integers $c_1, \ldots, c_n$ and an integer d, whether $\sum_{i=1}^n \sqrt{c_i} \leq d$. The problem is conjectured to lie in P, however the best known complexity upper bound is the 4th level of the Counting Hierarchy (Allender et al., 2009). The difficulty of obtaining an exact complexity for this problem has been recognized for at least 35 years (Garey et al., 1976). Here we show that STOCHASTIC-BLIND-POLICY is at least as hard as SQRT-SUM. Hence a result that would for instance place STOCHASTIC-BLIND-POLICY in NP would resolve several open problems in computer science, as argued in a similar setting where SQRT-SUM is reduced to the 3-NASH problem (Etessami and Yannakakis, 2010).

Theorem 2. *The* STOCHASTIC-BLIND-POLICY *problem is* SQRT-SUM*-hard.*

Proof. Let $c_1, \ldots, c_n$ and d be the inputs of SQRT-SUM. The reduction constructs an MDP with $n+1$ states and n actions, where the $(n+1)$st state is absorbing (self-looping). The starting probabilities are $\mu_i = \frac{1}{n}$ for states $i = 1, \ldots, n$ and $\mu_{n+1} = 0$, and the costs depend only on state and are given by the inputs c_i for states $i = 1, \ldots, n$ and $c_{n+1} = 0$. From each state $i = 1, \ldots, n$, the ith action deterministically transitions to the absorbing state $n+1$, while all other actions deterministically transition back to state i.

For each state $i = 1, \ldots, n$, the Bellman occupancy constraint reads $x_i = \frac{1-\gamma}{n} + \gamma(1-\pi_i)x_i$. Let $\varepsilon = \frac{\gamma}{1-\gamma} > 0$. Then the cost function reads

$$J(\boldsymbol{\pi}) = \sum_{i=1}^{n} c_i x_i = \frac{1}{n} \sum_{i=1}^{n} \frac{c_i}{1+\varepsilon\pi_i}. \tag{5}$$

Multiplying and diving by $n+\varepsilon$, we can rewrite

$$J(\boldsymbol{\pi}) = \frac{n+\varepsilon}{n} \sum_{i=1}^{n} \frac{1+\varepsilon\pi_i}{n+\varepsilon} \Big(\frac{\sqrt{c_i}}{1+\varepsilon\pi_i} \Big)^2 \;\geq\; \frac{1}{n(n+\varepsilon)} \Big(\sum_{i=1}^{n} \sqrt{c_i} \Big)^2, \tag{6}$$

where we applied Jensen's inequality noting that $\sum_{i=1}^{n} \frac{1+\varepsilon\pi_i}{n+\varepsilon} = 1$. Since the last term in (6) is a constant independent of $\boldsymbol{\pi}$, we see that the cost function reaches its minimum when the above inequality is tight, which is achieved when all terms are equal. It follows therefore that the last term in (6) is the optimal cost J^*, and it is achieved when, for each i, holds:

$$\frac{1+\varepsilon\pi_i^*}{n+\varepsilon} = \frac{\sqrt{c_i}}{\sum_{j=1}^{n} \sqrt{c_j}}. \tag{7}$$

We define ε (and hence γ) so that $n+\varepsilon = n\sum_{i=1}^{n} c_i$. Note that ε is strictly positive if at least one of the c_i is larger than one (which we assume is true, otherwise the SQRT-SUM problem trivializes). Application of Jensen's bound gives

$$n+\varepsilon = n\sum_{i=1}^{n} c_i \geq \Big(\sum_{i=1}^{n} \sqrt{c_i} \Big)^2 \geq \sum_{i=1}^{n} \sqrt{c_i}, \tag{8}$$

which establishes that the optimal policy $\boldsymbol{\pi}^*$ in (7) is always positive.

The STOCHASTIC-BLIND-POLICY question of whether there exists a stochastic blind controller $\boldsymbol{\pi}$ with cost $J(\boldsymbol{\pi}) \leq r$ is clearly equivalent to the question whether $J^* \leq r$. By choosing $r = \frac{d^2}{n(n+\varepsilon)}$, we see from (6) that the condition $J^* \leq r$ is equivalent to $\sum_{i=1}^{n} \sqrt{c_i} \leq d$, and the reduction is complete. □

5 A tractable case

We describe here a special case that results in a cost function that is concave in $\boldsymbol{\pi}$, in which case an optimal controller can be trivially found in polynomial time.

For each action a, let $\mathbf{P}_a$ denote the corresponding transition matrix, with $\mathbf{P}_a(\bar{s}, s) = p(\bar{s}|s, a)$. The special case assumes that each matrix $\mathbf{P}_a$ is *symmetric* (and therefore doubly stochastic), and that the costs depend only on the state and are proportional to the starting distribution: $\mathbf{c} = -\kappa\boldsymbol{\mu}$, with $\kappa > 0$. (Note from (2) that shifting and scaling $\mathbf{c}$ by arbitrary constants does not affect the optimal policy.) The bilinear program (2) then reads:

$$\min_{\boldsymbol{\pi}\in\Delta} -\boldsymbol{\mu}^\top \Big(\mathbf{I} - \gamma\mathbf{M}_{\boldsymbol{\pi}} \Big)^{-1} \boldsymbol{\mu}, \qquad \text{where} \quad \mathbf{M}_{\boldsymbol{\pi}} = \sum_a \pi_a \mathbf{P}_a. \tag{9}$$

Lemma 1. *For any $\boldsymbol{\pi}$, the matrix $\mathbf{I} - \gamma \mathbf{M}_{\boldsymbol{\pi}}$ is symmetric positive definite.*

Proof. Since each matrix $\mathbf{P}_a$ is symmetric and stochastic, all its eigenvalues are real and satisfy $\lambda(\mathbf{P}_a) \leq 1$. Hence, the eigenvalues of $\mathbf{I} - \gamma \mathbf{P}_a$ are also real and satisfy $\lambda(\mathbf{I} - \gamma \mathbf{P}_a) = 1 - \gamma \lambda(\mathbf{P}_a) > 0$ because $\gamma < 1$. Therefore, $\mathbf{I} - \gamma \mathbf{P}_a$ is a symmetric positive definite matrix, and so must be the matrix $\mathbf{I} - \gamma \mathbf{M}_{\boldsymbol{\pi}}$ as it can be written as the convex combination (over $\boldsymbol{\pi}$) of positive definite matrices. □

Theorem 3. *The function $f(\boldsymbol{\pi}) = \boldsymbol{\mu}^\top \Big(\mathbf{I} - \gamma \mathbf{M}_{\boldsymbol{\pi}}\Big)^{-1} \boldsymbol{\mu}$ is convex in $\boldsymbol{\pi} \in \Delta$.*

Proof. The epigraph of f is (see also Boyd and Vandenberghe (2004, Section 3.1.7))

$$\textbf{epi } f = \Big\{ (\boldsymbol{\pi}, t) \mid \boldsymbol{\pi} \in \Delta,\ \boldsymbol{\mu}^\top \Big(\mathbf{I} - \gamma \mathbf{M}_{\boldsymbol{\pi}}\Big)^{-1} \boldsymbol{\mu} \leq t,\ \mathbf{M}_{\boldsymbol{\pi}} = \sum_a \pi_a \mathbf{P}_a \Big\} \tag{10}$$

$$= \Big\{ (\boldsymbol{\pi}, t) \ \Big|\ \boldsymbol{\pi} \in \Delta,\ \begin{bmatrix} \mathbf{I} - \gamma \sum_a \pi_a \mathbf{P}_a & \boldsymbol{\mu} \\ \boldsymbol{\mu}^\top & t \end{bmatrix} \succeq 0 \Big\}, \tag{11}$$

where we used Lemma 1 and the Schur complement condition for positive definite matrices (Boyd and Vandenberghe, 2004, Appendix A.5.5). The last condition in (11) is a linear matrix inequality in $(\boldsymbol{\pi}, t)$, hence **epi** f is a convex set and f is convex. □

The problem (9) becomes the minimization of the concave function $-f$ over the probability simplex, hence there must exist a globally optimal solution in a corner of the simplex. This means that there will always exist an optimal controller that is deterministic. Since there are only k deterministic controllers, evaluating each of them and selecting the optimal one takes $O(kn^3)$ operations.

6 Conclusions

In response to the computational intractability of searching for optimal policies in POMDPs, many researchers have turned to finite-state controllers as a more tractable alternative. We have provided here a computational characterization of exactly solving problems in the class of stochastic controllers, showing that (1) they are NP-hard, (2) they are in PSPACE, and (3) they are SQRT-SUM-hard, hence showing membership in NP would resolve long-standing open problems.

We note that our NP-hardness proof relies on the assumption that the costs c_{sa} are nondegenerate functions of both state and action. We have recently addressed the case of state-only-dependent costs, which can be shown to be NP-hard by a reduction from the general case. This work will be published elsewhere.

In this article we have only addressed the complexity of the decision problem for the discounted infinite-horizon case. There are several open questions, in particular the complexity of approximate optimization for this class of stochastic controllers. The related literature addresses only the case of deterministic controllers (Lusena et al., 2001).

Acknowledgments

We are grateful to Marek Petrik for his feedback and for pointing an error in an earlier version. The first author would like to thank Constantinos Daskalakis, Michael Tsatsomeros, John Tsitsiklis, and Steve Vavasis for helpful discussions.

References

Al-Khayyal, F. A. and Falk, J. E. (1983). Jointly constrained biconvex programming. *Mathematics of Operations Research*, 8(2):273–286.

Allender, E., Bürgisser, P., Kjeldgaard-Pedersen, J., and Miltersen, P. B. (2009). On the complexity of numerical analysis. *SIAM J. Comput.*, 38(5):1987–2006.

Amato, C., Bernstein, D. S., and Zilberstein, S. (2007). Solving POMDPs using quadratically constrained linear programs. In *Proc. 20th Int. Joint Conf. on Artificial Intelligence*, Hyderabad, India.

Bernstein, D. S., Hansen, E. A., and Zilberstein, S. (2005). Bounded policy iteration for decentralized POMDPs. In *Proc. 19th Int. Joint Conf. on Artificial Intelligence*, Edinburgh, Scotland.

Boyd, S. and Vandenberghe, L. (2004). *Convex Optimization*. Cambridge University Press, Cambridge, UK.

Canny, J. F. (1988). Some algebraic and geometric computations in PSPACE. In *ACM Symposium on Theory of Computing*, pages 460–467.

Chrisman, L. (1992). Reinforcement learning with perceptual aliasing: The perceptual distinctions approach. In *Proc. 10th National Conf. on Articial Intelligence*, San Jose, CA.

Etessami, K. and Yannakakis, M. (2010). On the complexity of Nash equilibria and other fixed points. *SIAM Journal on Computing*, 39(6):2531–2597.

Garey, M. R., Graham, R. L., and Johnson, D. S. (1976). Some NP-complete geometric problems. In *ACM Symposium on Theory of Computing*.

Garey, M. R. and Johnson, D. S. (1979). *Computers and Intractability: A Guide to the Theory of NP-Completeness*. W. H. Freeman & Co., New York, NY, USA.

Hansen, E. (1998). Solving POMDPs by searching in policy space. In *Proc. 14th Int. Conf. on Uncertainty in Artificial Intelligence*, Madison, Wisconsin, USA.

Hastings, N. A. J. and Sadjadi, D. (1979). Markov programming with policy constraints. *European Journal of Operations Research*, 3:253–255.

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

Littman, M. L. (1994). Memoryless policies: Theoretical limitations and practical results. In *Proc. 3rd Int. Conf. on Simulation of Adaptive Behavior*, Brighton, England.

Littman, M. L., Goldsmith, J., and Mundhenk, M. (1998). The computational complexity of probabilistic planning. *Journal of Artificial Intelligence Research*, 9:1–36.

Lusena, C., Goldsmith, J., and Mundhenk, M. (2001). Nonapproximability results for partially observable Markov decision processes. *Journal of Artificial Intelligence Research*, 14:2001.

Madani, O., Hanks, S., and Condon, A. (1999). On the undecidability of probabilistic planning and infinite-horizon partially observable Markov decision problems. In *Proc. 16th National Conf. on Artificial Intelligence*.

Meuleau, N., Kim, K., Kaelbling, L., and Cassandra, A. (1999). Solving POMDPs by searching the space of finite policies. In *Proc. 15th Conf. on Uncertainty in Artificial Intelligence*, Stockholm, Sweden.

Motzkin, T. S. and Straus, E. G. (1965). Maxima for graphs and a new proof of a theorem of Turán. *Canadian Journal of Mathematics*, 17:533–540.

Mundhenk, M., Goldsmith, J., Lusena, C., and Allender, E. (2000). Complexity of finite-horizon Markov decision process problems. *Journal of ACM*, 47:681–720.

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

Platzman, L. K. (1981). A feasible computational approach to infinite-horizon partially-observed Markov decision problems. Technical report, School of Industrial and Systems Engineering, Georgia Institute of Technology. J-81-2.

Poupart, P. and Boutilier, C. (2004). Bounded finite state controllers. In Thrun, S., Saul, L., and Schölkopf, B., editors, *Advances in Neural Information Processing Systems 16*, Cambridge, MA. MIT Press.

Puterman, M. (1994). *Markov decision processes : Discrete stochastic dynamic programming*. John Wiley & Sons, New York.

Serin, Y. and Kulkarni, V. G. (2005). Markov decision processes under observability constraints. *Mathematical Methods of Operations Research*, 61:311–328.

Singh, S. P., Jaakkola, T., and Jordan, M. I. (1994). Learning without state-estimation in partially observable Markovian decision processes. In *Proc. 11th Int. Conf. on Machine Learning*, San Francisco, CA.

Smith, J. L. (1971). Markov decisions on a partitioned state space. *IEEE Transactions on Systems, Man and Cybernetics, SMC-1*, pages 55–60.

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

Toussaint, M., Storkey, A., and Harmeling, S. (2011). Expectation-Maximization methods for solving (PO)MDPs and optimal control problems. In Barber, D., Cemgil, A. T., and Chiappa, S., editors, *Bayesian Time Series Models*. Cambridge University Press.