Open Problem: Approximate Planning of POMDPs in the class of Memoryless Policies

Kamyar Azizzadenesheli KAZIZZAD@UCI.EDU
University of California, Irvine

Alessandro Lazaric ALESSANDRO.LAZARIC@INRIA.FR
French Institute for Research in Computer Science and Automation (Inria)

Animashree Anandkumar A.ANANDKUMAR@UCI.EDU
University of California, Irvine

Abstract

Planning plays an important role in the broad class of decision theory. Planning has drawn much attention in recent work in the robotics and sequential decision making areas. Recently, Reinforcement Learning (RL), as an agent-environment interaction problem, has brought further attention to planning methods. Generally in RL, one can assume a generative model, e.g. graphical models, for the environment, and then the task for the RL agent is to learn the model parameters and find the optimal strategy based on these learnt parameters. Based on environment behavior, the agent can assume various types of generative models, e.g. Multi Armed Bandit for a static environment, or Markov Decision Process (MDP) for a dynamic environment. The advantage of these popular models is their simplicity, which results in tractable methods of learning the parameters and finding the optimal policy. The drawback of these models is again their simplicity: these models usually underfit and underestimate the actual environment behavior. For example, in robotics, the agent usually has noisy observations of the environment inner state and MDP is not a suitable model.

More complex models like Partially Observable Markov Decision Process (POMDP) can compensate for this drawback. Fitting this model to the environment, where the partial observation is given to the agent, generally gives dramatic performance improvement, sometimes unbounded improvement, compared to MDP. In general, finding the optimal policy for the POMDP model is computationally intractable and fully non convex, even for the class of memoryless policies. The open problem is to come up with a method to find an exact or an approximate optimal stochastic memoryless policy for POMDP models.

1. Introduction

The concept of planning, as a part of decision theory, in the AI literature has a long history. It is the bases for a variety of popular agent-environment interaction problems like Reinforcement Learning (RL). RL is an effective approach to solve the problem of sequential decision making under uncertainty. RL agents learn how to maximize long-term reward using a experience obtained by direct interaction with a stochastic environment (Bertsekas and Tsitsiklis, 1996; Sutton and Barto, 1998). Since the environment is initially unknown, the agent has to balance between *exploring* the environment to estimate its structure, and *exploiting* the estimates to compute a policy that maximizes the long-term reward. As a result, designing a RL algorithm requires three different elements: **1)** an estimator for the environment's structure, **2)** a planning algorithm to compute the optimal policy of the estimated environment (LaValle, 2006), and **3)** a strategy to make a trade off between exploration and exploitation to minimize the *regret*, i.e., the difference between the performance of the exact optimal policy and the rewards accumulated by the agent over time.

Most of RL literature assumes that the environment can be modeled as a Markov decision process (MDP), with a Markovian state evolution that is fully observed. A number of exploration–exploitation strategies have been shown to have strong performance guarantees for MDPs, either in terms of regret or sample complexity Auer et al. (2009). However, the assumption of full observability of the state evolution is often violated in practice, and the agent may have only noisy observations of the true state of the environment (e.g., noisy sensors in robotics). In this case, it is more appropriate to use the partially-observable MDP (POMDP) (Sondik, 1971) model.

Many challenges arise in designing RL algorithms for POMDPs. Unlike in MDPs, the estimation problem (element 1) involves identifying the parameters of a latent variable model (LVM). The planning problem (element 2), i.e., computing the optimal policy for a POMDP with known parameters, is PSPACE-complete (Papadimitriou and Tsitsiklis, 1987), and it requires solving an augmented MDP built on a continuous belief space (i.e., a distribution over the hidden state of the POMDP). Finally, integrating estimation and planning in an exploration–exploitation strategy (element 3) with guarantees is non-trivial and no no-regret strategies are currently known.

Previous works Ross et al. (2007) and Poupart and Vlassis (2008) present new active learning algorithms to estimate the belief state in a model-based Bayesian RL approach, where a distribution over possible MDPs is updated over time. The proposed algorithms are such that the Bayesian inference can be done at each step, a POMDP is sampled from the posterior and the corresponding optimal policy is executed. The regret bound and sample complexity are not provided.

Recently, the learning POMDP model parameter and imposing trade off between exploration and estimation (elements 1 and 3), are done at Azizzadenesheli et al. (2016). They propose the theoretical guaranty on regret bound given the oracle memoryless policy. Therefore to close the learning, planing, and exploration-exploitation loop, the missing part, planning (element 2), is the remaining part. Therefore, planing is a problem of finding the optimal memoryless policy, under uncertainty, in the class of stochastic memoryless polices. The overview complexity of planing in POMDP domain is discussed in Kaelbling et al. (1998).

2. Formal Definition

A POMDP M is a tuple $\langle \mathcal{X}, \mathcal{A}, \mathcal{Y}, \mathcal{R}, f_T, f_R, f_O \rangle$, where $\mathcal{X}$ is a finite state space with cardinality $|\mathcal{X}| = X$, $\mathcal{A}$ is a finite action space with cardinality $|\mathcal{A}| = A$, $\mathcal{Y}$ is a finite observation space with cardinality $|\mathcal{Y}| = Y$, and $\mathcal{R}$ is a finite reward space with cardinality $|\mathcal{R}| = R$ and largest reward $r_{\max}$. In addition f_T denotes the transition density, so that $f_T(x'|x, a)$ is the probability of transition to x' given the state-action pair (x, a), f_R is the reward density, so that $f_R(\boldsymbol{r}|x, a)$ is the probability of receiving the reward in $\mathcal{R}$ corresponding to the value

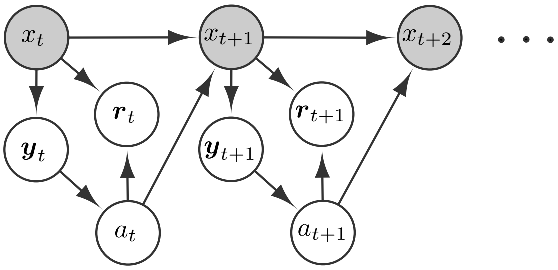

Figure 1: Graphical model of a POMDP under memoryless policies.

of the indicator vector $\boldsymbol{r}$ given the state-action pair (x, a), and f_O is the observation density, so that $f_O(\boldsymbol{y}|x)$ is the probability of receiving the observation in $\mathcal{Y}$ corresponding to the indicator vector $\boldsymbol{y}$ given the state x. Whenever convenient, we use tensor forms for the density functions such that $T_{i,j,l} = \mathbb{P}[x_{t+1} = j|x_t = i, a_t = l] = f_T(j|i, l), s.t.\ T \in \mathbb{R}^{X\times X\times A}$, $O_{n,i} = \mathbb{P}[\boldsymbol{y} = \boldsymbol{e}_n|x = i] = f_O(\boldsymbol{e}_n|i), s.t.\ O \in \mathbb{R}^{Y\times X}$, and $\Gamma_{i,l,m} = \mathbb{P}[\boldsymbol{r} = \boldsymbol{e}_m|x = i, a = l] = f_R(\boldsymbol{e}_m|i, l), s.t.\ \Gamma \in \mathbb{R}^{X\times A\times R}$. We also denote by $T_{:,j,l}$ the fiber (vector) in $\mathbb{R}^X$ obtained by fixing the arrival state j and action l and by $T_{:,:,l} \in \mathbb{R}^{X\times X}$ the transition matrix between states when using action l. The graphical model associated to the POMDP is illustrated in Fig. 1.

We focus on stochastic memoryless policies which map observations to actions and for any policy π we denote by $f_\pi(a|\boldsymbol{y})$ its density function. Acting according to a policy π in a POMDP M defines a Markov chain characterized by a transition density $f_{T,\pi}(x'|x) = \sum_a \sum_{\boldsymbol{y}} f_\pi(a|\boldsymbol{y}) f_O(\boldsymbol{y}|x) f_T(x'|x, a)$, and a stationary distribution ω_π over states such that $\omega_\pi(x) = \sum_{x'} f_{T,\pi}(x'|x)\omega_\pi(x')$. The expected average reward performance of a policy π is $\eta(\pi; M) = \sum_x \omega_\pi(x)\overline{r}_\pi(x)$, where $\overline{r}_\pi(x)$ is the expected reward of executing policy π in state x defined as $\overline{r}_\pi(x) = \sum_a \sum_{\boldsymbol{y}} f_O(\boldsymbol{y}|x) f_\pi(a|\boldsymbol{y})\overline{r}(x, a)$, and $\overline{r}(x, a) = \sum_r r f_R(r|x, a)$ is the expected reward for the state-action pair (x, a).

The best stochastic memoryless policy is $\pi^* = \arg\max_\pi \eta(\pi; M)$ and we denote by $\eta^* = \eta(\pi^*; M)$ its average reward. Finding the optimal policy π^* requires solving non-convex optimization and it is the desired open problem.

3. Related Work

Planning on uncertainty in a dynamic internal process is studied for infinite horizon Sondik (1978). It is shown that, finding the exact optimal policy for POMDP is followed by the curse of dimensionality and the curse of history. People uses point-based value iteration Pineau et al. (2006) to reduce the complexity of the planning. It is also common to use heuristic search value iteration Smith and Simmons (2004) and also policy tree with limited depth Kaelbling et al. (1998) to reduce the planning complexity. For a finite horizonBut the computation complexity of finding optimal policy grows exponentially by horizon. For an infinite horizon, each vector of state distribution can be any point in the continues space of the simplex subspace. This means the planning is over the continuous space which is PSPACE-complete. Sondik (1978) presented a method to partition the continuous space

of the state distribution and then the policy is just a mapping from these partitions to the action.
In general, planning in the space of memoryless policy has a lower level of complexity. Although, it seems to be easier than belief based planning, it is still an $NP-hard$ problem Vlassis et al. (2012). To breaking down this complexity, Li et al. (2011) presented a novel method for finding the optimal policy in the class of deterministic memoryless policies. Meanwhile, deterministic policies act poorly in the general case of POMDPs. The geometric representation of POMDP planning problem is shown in Montufar et al. (2015) and its geometric structure is well studied. Therefore, proposing a novel method to find the exact or approximated optimal memoryless policy (policy with performance $\epsilon-close$ to the performance of optimal policy) or limited history dependent policy under some mild conditions is the next step in the world of POMDP planning.

References

Peter Auer, Thomas Jaksch, and Ronald Ortner. Near-optimal regret bounds for reinforcement learning. In *Advances in neural information processing systems*, pages 89–96, 2009.

Kamyar Azizzadenesheli, Alessandro Lazaric, and Animashree Anandkumar. Reinforcement learning of pomdps using spectral methods. *arXiv preprint arXiv:1602.07764*, 2016.

D. Bertsekas and J. Tsitsiklis. *Neuro-Dynamic Programming*. Athena Scientific, 1996.

Leslie Pack Kaelbling, Michael L Littman, and Anthony R Cassandra. Planning and acting in partially observable stochastic domains. *Artificial intelligence*, 101(1):99–134, 1998.

Steven M LaValle. *Planning algorithms*. Cambridge university press, 2006.

Yanjie Li, Baoqun Yin, and Hongsheng Xi. Finding optimal memoryless policies of pomdps under the expected average reward criterion. *European Journal of Operational Research*, 211(3):556–567, 2011.

Guido Montufar, Keyan Ghazi-Zahedi, and Nihat Ay. Geometry and determinism of optimal stationary control in partially observable markov decision processes. *arXiv preprint arXiv:1503.07206*, 2015.

Christos Papadimitriou and John N. Tsitsiklis. The complexity of markov decision processes. *Math. Oper. Res.*, 12(3):441–450, August 1987. ISSN 0364-765X.

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

P. Poupart and N. Vlassis. Model-based bayesian reinforcement learning in partially observable domains. In *International Symposium on Artificial Intelligence and Mathematics (ISAIM)*, 2008.

Stephane Ross, Brahim Chaib-draa, and Joelle Pineau. Bayes-adaptive pomdps. In *Advances in neural information processing systems*, pages 1225–1232, 2007.

Trey Smith and Reid Simmons. Heuristic search value iteration for pomdps. In *Proceedings of the 20th conference on Uncertainty in artificial intelligence*, pages 520–527. AUAI Press, 2004.

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

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

Richard S Sutton and Andrew G Barto. *Introduction to reinforcement learning*. MIT Press, 1998.

Nikos Vlassis, Michael L Littman, and David Barber. On the computational complexity of stochastic controller optimization in pomdps. *ACM Transactions on Computation Theory (TOCT)*, 4(4):12, 2012.