Nonparametric Bayesian Learning of Other Agents' Policies in Interactive POMDPs

(Extended Abstract)

Alessandro Panella
apanel2@uic.edu

Piotr Gmytrasiewicz
piotr@uic.edu

Department of Computer Science
University of Illinois at Chicago
Chicago, IL 60607-7053

ABSTRACT

We consider an autonomous POMDP agent facing a multiagent environment with unknown opponents, that are modeled as finite state controllers. The agent first learns the models from (imperfectly) observed behavior, and subsequently exploits them in planning for its own optimal policy by constructing an interactive POMDP. In the learning phase, Bayesian nonparametric methods are used to sample from the posterior distribution over the infinite-dimensional space of all possible controllers, resulting in models whose size scales with the complexity of observed behavior. Experimental results show that learning improves the agent's performance, which increases with the amount of data collected during the learning phase.

Categories and Subject Descriptors

I.2.11 [**Distributed Artificial Intelligence**]: Multiagent Systems

Keywords

Multiagent Systems, Opponent Modeling, Probabilistic Inference, Bayesian Nonparametrics

1. INTRODUCTION

An autonomous POMDP agent operating in a multiagent environment must accurately predict the actions of other agents in order to achieve good performance. We consider an agent that maintains explicit models of other agents to generate such predictions.

Previous work [4] proposes maintaining a distribution over other agents' *intentional models*, that specify the other agents' own POMDP tuples, to be solved recursively. The probability space of all such models is very complex, especially if little is known about the other agents' preferences.

An alternative we pursue in this paper is to consider subintentional models, intended as stochastic processes underlying the other agents' behavior, without considering their beliefs or preferences. In particular, we consider a class of finite state controllers with deterministic transitions between nodes and stochastic action generations, named probabilistic deterministic finite state controllers (PDFCs), which provide a good trade-off between complexity and expressive power. Each node represents an internal state of the modeled agent that summarizes its past history, and contains a probabilistic mapping to the action space.

We assume realistically that our protagonist agent (i) has no a priori knowledge of the other agent's (j) model. This implies that the agent must learn a probability distribution over the set of all possible PDFCs. For this reason, we adopt Bayesian nonparametric methods (BNP) [5], which make it possible to specify a distribution over objects with an unbounded number of parameters. Intuitively, BNP methods allow the learned representations to grow with the observed complexity of the data. Since the problem we tackle is too complex to be amenable to conjugate analysis, we use the computational implementations of Bayesian inference based on the Dirichlet process and Gibbs sampling.

Although we view an on-line scenario in which the agent simultaneously learns about others and plans its own optimal policy as the most realistic, in this work we separate the learning and planning phases in order to assess the properties and merits of the learning methodology in isolation, and not subject to complications arising from a fully online approach, such as dealing with adaptive agents. During the learning phase our agent i accumulates its own observations which probabilistically depend on the state of the world and j's action. Using this training sequence, the agent computes a sample posterior distribution over j's models using Gibbs sampling. During the second phase the agents interact, and agent i exploits the learned models of j using a specialization of the interactive POMDP (I-POMDP) framework [4].

Our work is related to research in plan recognition [2] and goal-based POMDPs [7]. Moreover, similar BNP techniques have been applied to partially observable model-based reinforcement learning problems [3]. In game theory, finite automata have been used to represent agents in presence of *bounded rationality* with observable actions, such as the heuristic algorithm in [1]. In comparison, our framework tackles more general interactive settings, and does not rely on equilibrium-based solutions concepts, assuming instead a more procedural, behavioral perspective [8].

Appears in: *Proceedings of the 14th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2015), Bordini, Elkind, Weiss, Yolum (eds.), May 4–8, 2015, Istanbul, Turkey.*

2. OVERVIEW OF METHODOLOGY

The protagonist agent learns over a class C of PDFCs which represent j's policy, given a sequence of observations $\omega^i_{1:T}$ from the environment. The task is to infer the posterior distribution over all possible models $c \in C$:

$$p(c|\omega^i_{1:T}) \propto p(\omega^i_{1:T}|c)\ p(c). \quad (1)$$

To enable learning over controllers of unknown complexity, a nonparametric distribution $p(c)$ is used, based on the stick-breaking construction of the Dirichlet process. Theoretically, this distribution places an exponentially decreasing probability over an infinite set of objects (the nodes in the PDFC), each with its own parameters (the stochastic mapping to actions.) In practice, only a finite number of parameters will be instantiated from finite observation sequences. Dirichlet process priors lend themselves to efficient sampled-based inference via the "Chinese restaurant process" (CRP) conditional distribution, utilized in our Gibbs sampling learning algorithm. Additionally, a number of unobserved variables need to be inferred to enable sampling from the CRP, namely the sequences of world states $s_{1:T}$, and j's actions $a^j_{1:T}$ and own observations $\omega^j_{1:T}$.

The result of this inference is an ensemble of sampled models C_j that approximates the posterior distribution of Eq. 1. These models are used to augment the set of world states S in the subsequent planning phase, forming the *interactive state space* $\bar{S}_i = S \times M_j$. Unlike the more general, recursive I-POMDP framework, it is possible to construct a flat I-POMDP model on this augmented space, that can be solved using existing algorithm. Given that the interactive state space is much larger than the original S, but can be easily factorized, POMDP algorithms that work on factorized representations such as symbolic Perseus [6] are particularly useful.

3. RESULTS

We present results for the Multiagent Tiger Problem [4] in three instances, corresponding to different levels (0.96, 0.85, 0.7) of j's hearing accuracy, resulting in optimal controllers of size 3, 5, and 7 respectively, from which j's behavior is generated. In all three cases we evaluate i's performance for different lengths of observed sequence, averaged over 40 trials each. Fig. 1 reports the reward accumulated by i when interacting with j. Agent i operates accordingly to the solution to the I-POMDP augmented with j's learned models as described above, while j actually uses its real model. The vertical bars in the plots indicate the standard deviation over the different simulation trials, while the red line represents the reward that i would obtain if it were to know j's actual model. We observe that the increased length of the observation phase T_{learn} allows agent i to learn better quality models of j, resulting in increased performance. Moreover, we note that it takes a longer sequence to achieve good performance when the opponent's behavior is generated by a more complex policy (7-state controller in the right plot.)

4. CONCLUSIONS AND FUTURE WORK

We have presented a learning and planning framework for multiagent POMDP problems, where little is known about the other agents, using statistical models of their behavior in the form of finite controllers. Preliminary results validate the proposed methodology and motivate further investigation in this direction. Future efforts will concentrate on demonstrating the scalability of the methods presented here by using more complex examples, and enabling the modeling agent to interleave on-line learning and interaction.

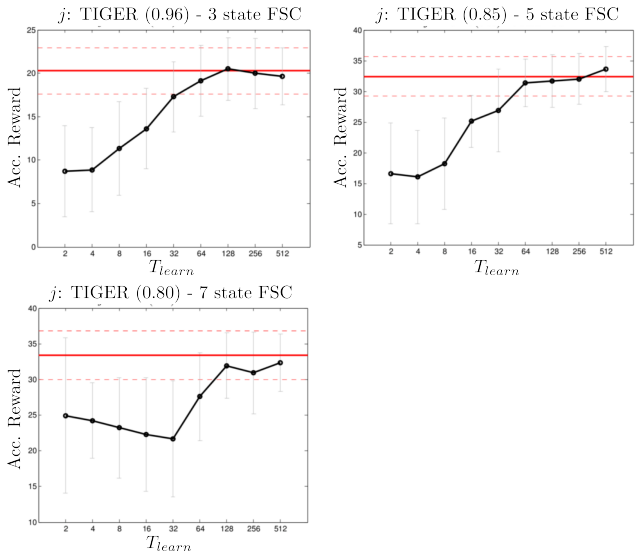

Figure 1: Agent's performance with respect to length of observed sequence.

REFERENCES

[1] D. Carmel and S. Markovitch. Learning models of intelligent agents. In *Proceedings of the 13th National Conference on Artificial intelligence*, pages 62–67, 1996.

[2] E. Charniak and R. P. Goldman. A Bayesian model of plan recognition. *Artificial Intelligence*, 64(1):53–79, Nov. 1993.

[3] F. Doshi-Velez, D. Pfau, F. Wood, and N. Roy. Bayesian nonparametric methods for partially-observable reinforcement learning. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 37(2):394–407, Feb 2015.

[4] P. J. Gmytrasiewicz and P. Doshi. A framework for sequential planning in multi-agent settings. *Journal of Artificial Intelligence Research*, 24(1):49–79, July 2005.

[5] N. L. Hjort, C. Holmes, P. Müller, and S. G. Walker, editors. *Bayesian Nonparametrics*. Cambridge University Press, Apr. 2010.

[6] P. Poupart. *Exploiting Structure to Efficiently Solve Large Scale Partially Observable Markov Decision Processes*. PhD thesis, University of Toronto, Toronto, Ont., Canada, 2005. AAINR02727.

[7] M. Ramirez and H. Geffner. Goal recognition over POMDPs: inferring the intention of a POMDP agent. In *International Joint Conference on Artificial Intelligence*, pages 2009–2014, 2011.

[8] J. R. Wright and K. Leyton-Brown. Behavioral game theoretic models: a bayesian framework for parameter analysis. In *International Conference on Autonomous Agents and Multiagent Systems, Valencia, Spain*, pages 921–930, 2012.