Learning Factored Markov Decision Processes with Unawareness

Extended Abstract

Craig Innes
University of Edinburgh
Edinburgh, UK
craig.innes@ed.ac.uk

Alex Lascarides
University of Edinburgh
Edinburgh, UK
alex@inf.ed.ac.uk

ABSTRACT

Methods for learning and planning in sequential decision problems often assume the learner is fully aware of all possible states and actions in advance. This assumption is sometimes untenable: evidence gathered via domain exploration or external advice may reveal not just information about which of the currently known states are probable, but that *entirely new* states or actions are possible. This paper provides a model-based method for learning factored markov decision problems from both domain exploration and contextually relevant expert corrections in a way which guarantees convergence to near-optimal behaviour, even when the agent is initially unaware of actions or belief variables that are critical to achieving success. Our experiments show that our agent converges quickly on the optimal policy for both large and small decision problems. We also explore how an expert's tolerance towards the agent's mistakes affects the agent's ability to achieve optimal behaviour.[1]

KEYWORDS

Reasoning in agent-based systems; Reward structures for learning; Reinforcement Learning

ACM Reference Format:
Craig Innes and Alex Lascarides. 2019. Learning Factored Markov Decision Processes with Unawareness. In *Proc. of the 18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019), Montreal, Canada, May 13–17, 2019,* IFAAMAS, 3 pages.

1 INTRODUCTION

In this paper, we show how an agent can learn optimal behaviour in a complex sequential decision problem under *unawareness*. By unawareness, we mean not just that the agent is uncertain about the transition probabilities between states, but that the agent may not initially know that certain states or actions even exist [4]. Such scenarios are common in the real world. In human discussion for example, answers to a person's inquiry may not only provide information about which of the questioner's existing hypotheses are likely, but may also reveal new unconsidered hypotheses [1]; in medicine, pharmacologists may discover a drug has a completely unforeseen side-effect in certain children, even after years of observing its effects on adults.

[1] A detailed account of this work can be found at [5]

Proc. of the 18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019), N. Agmon, M. E. Taylor, E. Elkind, M. Veloso (eds.), May 13–17, 2019, Montreal, Canada. © 2019 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org). All rights reserved.

Several methods jointly learn both a model of the environment and optimal behaviour of decision problems [2, 3, 8], but most assume the agent is aware of all relevant actions and states at the start of learning. Many works also leverage expert interventions to improve agent performance [6, 10–12] but again assume the expert's intended meaning can be understood without expanding the agent's awareness of the state and action space.

Recent work on *Markov Decision Processes with Unawareness* (MDPUS) [7] allow an agent to learn optimal behaviour even when it starts unaware of some states and actions. This work deliberately leaves abstract the mechanism by which agents discover unforeseen states and actions. Our paper builds on this work in two ways. First, we provide an agent who learns a structured model of the environment, and discovers *explicit belief variables* rather than atomic states. This makes learning tractable for larger, complex problems. Second, we provide a concrete mechanism by which an agent discovers unforeseen factors: it exploits the agent's reasoning and dialogue policies of both the learning agent and an expert.

2 UNAWARENESS LEARNING MODEL

We define *episodic, factored markov decision processes with unawareness* (FMDPU) with two tuples: $\langle \mathcal{X}^+, \mathcal{S}_s, \mathcal{S}_e, A^+, \mathcal{T}, \mathcal{R}^+ \rangle$ and $\langle \mathcal{X}^0, A^0, \mathcal{R}^0 \rangle$. The possible states $\mathcal{S}$ are represented as a joint assignment to the set of belief variables $\mathcal{X}^+$ (that is $\mathcal{S} = v(\mathcal{X}^+)$), while A^+ gives the set of possible actions; $S_s, S_e \subseteq \mathcal{S}$ are the possible start and end (terminal) states of an episode; $\mathcal{T} : \mathcal{S} \times A^+ \times \mathcal{S} \rightarrow [0, 1]$ is the *markovian transition function* $P(s'|s, a)$ and $\mathcal{R}^+ : v(scope_+(\mathcal{R})) \rightarrow \mathbb{R}$ is the immediate reward function (where $scope_+(\mathcal{R}) \subseteq \mathcal{X}^+$ is the subset of variables which determine the reward an agent receives). The sets $X^0 \subseteq \mathcal{X}^+$ and $A^0 \subseteq A^+$ define the agent's *initial awareness* of the belief variables and possible actions relevant to the problem. Likewise, for the agent's initial reward function - $\mathcal{R}^0 : scope_0(\mathcal{R}) \rightarrow \mathbb{R}$, the agent may be only aware of a subset of the variables which are relevant to determining the reward.

The agent's goal is to learn the optimal policy π_+ which chooses the action a which maximizes the expected discounted return (i.e., the *value function* $V_\pi(s) = \mathcal{R}^+(s) + \gamma \sum_{s' \in \mathcal{S}} P(s'|s, \pi(s)) V_\pi(s')$) across all states s. If the agent knows $\mathcal{T}$ and $\mathcal{R}^+$, we can calculate V_π via *value iteration* [9]. If not, we must learn the most likely transition structure Pa^a_X (where $Pa^a_X \subseteq \mathcal{X}^+$ is the set of variables which influence the value of X, given that the agent has performed action a) and associated parameters θ^a_X from the agent's interactions with the domain via sequential trials $D_{0:t} = [\langle s_0, a_0, r_1, s_1 \rangle, \ldots, \langle s_t, a_t, r_{t+1}, s_{t+1} \rangle]$, where s_t, a_t, and r_t are the current state, action taken, and reward received at time t. We can

then update V_π based on the agent's changing estimate of the environment using *incremental structured value iteration* (i-SVI) [2]. In our task however, the agent's unawareness may make converging on π_+ impossible. For example, the image of π_+ may contain some action $a \notin A^0$, or the optimal action may depend on the value of some variable $X_i \notin \mathcal{X}^0$.

Our model therefore also includes a cooperative, fully-informed expert who gives advice which might reveal members of $\mathcal{X}^+$, A^+ and $scope_+(\mathcal{R})$ that the agent is unaware of. Our aim was to mirror human teacher-apprentice interactions, so rather than assuming the expert can transfer all of its knowledge about the task at once (which is typically impossible due to cognitive bounds, ignorance, and other communicative constraints on human experts), we instead allow the expert to *interject* with a minimal amount of contextually relevant advice during learning. We identify three types of advice, whose combination guarantee the agent converges on near-optimal behaviour (see [5] for theorems and proofs):

(1) Advice on a better action (e.g., *"At time t, it would have been better to do action a' than a_t"*)
(2) Resolving misunderstandings between previous pieces of advice (e.g., *"I advised doing a' at time t, but not at time t' because the value of variable X_i at times t and t' was different"*)
(3) Explaining an observed reward the agent thinks is "impossible" given the size of $scope_0(\mathcal{R})$ (e.g., *"The value of X_j also affects your reward, which is why the last trial went poorly"*)

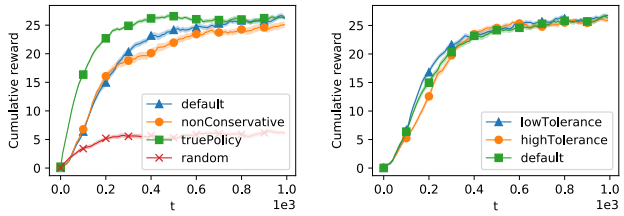
Advice of types (2) and (3) occur as answers to the learning agent's query, queries the agent asks when it infers that the hypothesis space of her current model is inconsistent with observed evidence. Advice of type (1) occurs when the expert observes that the agent has performed sufficiently poorly—the agent's performance drops some proportion β below π_+—and so this dialogue move depends on the expert's tolerance to the agent's mistakes.

As well as allowing the agent to overcome its unawareness, we also show how the agent can *conserve* its previous beliefs about $\mathcal{T}$ and V as its awareness expands (rather than resetting its beliefs upon each new discovery). On discovering a new variable Z at time t, the agent can conserve what it has learned about the likely parent structures by using the old *posterior* distribution $(Pa_X^a|D_{0:t-1})$ to inform a new *prior* over the new expanded set of possible parent structures (including those parent sets which might contain Z):

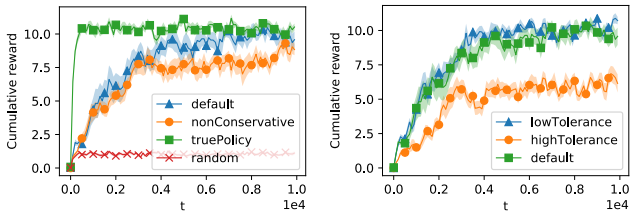
$$P'(Pa_X^a) = \begin{cases} (1-\rho)P(Pa_X^a|D_{0:t-1}) & \text{if } Z \notin Pa_X^a \\ \rho P((Pa_X^a \setminus Z)|D_{0:t-1}) & \text{otherwise} \end{cases}$$

Here, $\rho \leq 0.5$ controls the initial probability that the new variable Z is a parent to X. Intuitively, this prior assumes that the probabilistic structure in the expanded hypothesis space is close to the agent's estimates before discovering Z. Similarly, we can conserve the agent's estimate of each state's value by setting $V_\pi^t(s) = V_\pi^{t-1}(s[\mathcal{X}^{t-1}])$. This means our agent initially assumes Z has no effect on a state's value, until future trials show otherwise.

For a detailed description of our model, including a formal account of the messages between the agent and expert, and explicit algorithms for conserving information about $\mathcal{T}$ and V, see [5].



(a) *Coffee Robot* task (averaged over 50 experiments).



(b) *Factory* task (averaged over 20 experiments).

Figure 1: Cumulative Rewards. Shaded areas represent the standard error from the mean.

3 EXPERIMENTS

We tested variations of our agent and expert on two well-known sequential problems: *Coffee-Robot* (256 state/action pairs) and *Factory* (774144 state/action pairs).[2] In each experiment, our agent begins aware of only a subset of relevant to each problem. The **default** agent used the learning model described in section 2 and follows an ϵ-greedy policy throughout ($\epsilon = 0.1$, $\beta = 0.1$). The **truePolicy**/**random** agents give upper/lower bounds on performance by executing either an ϵ-greedy version of π_+, or a completely random action each time step, respectively. The **nonConservative** agent does not conserve information about V or $\mathcal{T}$ as described in section 2 but instead resets V and $\mathcal{T}$ to their initial values each time it discovers a new variable. The **lowTolerance**/**highTolerance** agents change the expert's tolerance to $\beta = 0.01$ and $\beta = 0.5$.

Figures 1a and 1b show the *cumulative reward*[3] gathered by each agent. Despite starting unaware of factors critical to success, the default agent quickly discovers the relevant actions and beliefs with the expert's aid, and converges on the optimal policy. The non-conservative agent also learns the optimal policy, but slower. This shows the value of conserving $\mathcal{T}$ and V on discovering new beliefs. We also see how expert tolerance affects performance. The agent paired with high tolerance expert discovers less actions and belief variables than its lower tolerance counter-parts, learns a (marginally) worse final policy, and (in the case of the larger *Factory* problem) earns a substantially lower cumulative reward. This is because, in the high tolerance case, the expert decides early on that the agent has learned a "good enough" policy, and so does not reveal additional actions which would yield only a minor increase in reward. For further analysis of these results, see [5].

[2]Full specifications at https://cs.uwaterloo.ca/~jhoey/research/spudd/index.php
[3]The cumulative rewards were discounted by 0.99 at each step to increase readability.

REFERENCES

[1] Anna Coenen, Jonathan D. Nelson, and Todd M. Gureckis. 2017. Asking the right questions about human inquiry. *OpenCoenen, Anna, Jonathan D Nelson, and Todd M Gureckis.âĂIJAsking the Right Questions About Human InquiryâĂİ. PsyArXiv* 13 (2017).

[2] Thomas Degris, Olivier Sigaud, and Pierre-Henri Wuillemin. 2006. Learning the structure of factored markov decision processes in reinforcement learning problems. In *Proceedings of the 23rd international conference on Machine learning*. ACM, 257–264.

[3] Carlos Diuk, Lihong Li, and Bethany R. Leffler. 2009. The adaptive k-meteorologists problem and its application to structure learning and feature selection in reinforcement learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*. ACM, 249–256.

[4] Joseph Y. Halpern, Nan Rong, and Ashutosh Saxena. 2010. MDPs with unawareness. *arXiv preprint arXiv:1006.2204* (2010).

[5] Craig Innes and Alex Lascarides. 2019. Learning Factored Markov Decision Processes with Unawareness. *arXiv e-prints* (Feb. 2019), arXiv:1902.10619.

[6] W. Bradley Knox and Peter Stone. 2009. Interactively shaping agents via human reinforcement: The TAMER framework. In *Proceedings of the fifth international conference on Knowledge capture*. ACM, 9–16.

[7] Nan Rong. 2016. *Learning in the Presence of Unawareness.* Ph.D. Dissertation. Cornell University.

[8] Alexander L. Strehl, Carlos Diuk, and Michael L. Littman. 2007. Efficient structure learning in factored-state MDPs. In *AAAI*, Vol. 7. 645–650.

[9] R. Sutton and A Barto. 1998. *Reinforcement Learning: An Introduction.* MIT Press.

[10] Lisa Torrey and Matthew Taylor. 2013. Teaching on a budget: Agents advising agents in reinforcement learning. In *Proceedings of the 2013 international conference on Autonomous agents and multi-agent systems*. International Foundation for Autonomous Agents and Multiagent Systems, 1053–1060.

[11] Fangkai Yang, Daoming Lyu, Bo Liu, and Steven Gustafson. 2018. PEORL: Integrating Symbolic Planning and Hierarchical Reinforcement Learning for Robust Decision-Making. *arXiv preprint arXiv:1804.07779* (2018).

[12] Matthieu Zimmer, Paolo Viappiani, and Paul Weng. 2014. Teacher-student framework: a reinforcement learning approach. In *AAMAS Workshop Autonomous Robots and Multirobot Systems*.