Maximizing Plan Legibility in Stochastic Environments

Extended Abstract

Shuwa Miura
University of Massachusetts Amherst
Amherst, Massachusetts
smiura@cs.umass.edu

Shlomo Zilberstein
University of Massachusetts Amherst
Amherst, Massachusetts
shlomo@cs.umass.edu

ABSTRACT

Legible behavior allows an observing agent to infer the intention of an observed agent. Producing legible behavior is crucial for successful multi-agent interaction in many domains. We introduce techniques for legible planning in stochastic environments. Maximizing legibility, however, presents a complex trade-off between maximizing the underlying rewards. Hence, we propose a method to balance the trade-off. In our experiments, we demonstrate that maximizing legibility results in unambiguous behaviors.

KEYWORDS

Planning; MDP; legibility; human-robot collaboration

ACM Reference Format:
Shuwa Miura and Shlomo Zilberstein. 2020. Maximizing Plan Legibility in Stochastic Environments. In *Proc. of the 19th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2020), Auckland, New Zealand, May 9–13, 2020*, IFAAMAS, 3 pages.

1 INTRODUCTION

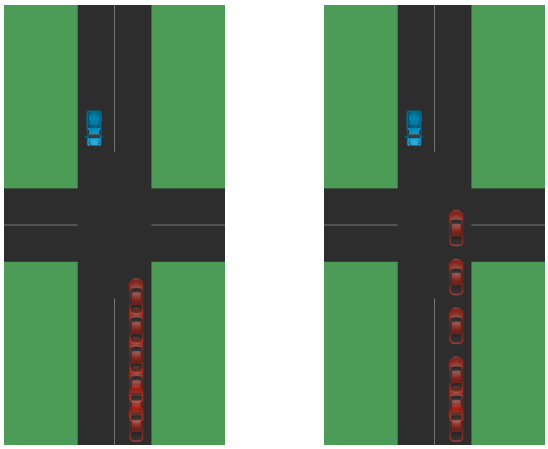
Reasoning about intentions – inferring what others are trying to do or making it clear what one is trying to do – is ubiquitous in our daily lives. For example, consider two cars approaching an intersection from opposite directions as shown in Fig. 1. Assume that the driver of the blue car is trying to make an unprotected left turn, but there is ambiguity whether they can complete the turn before the red car enters the intersection. If the red car is not yielding, the blue car should respect the traffic laws and wait for the red car to pass. To clearly convey its intention, the driver of the red car can accelerate before the intersection (Fig.1b). In this paper, we investigate *legible* behaviours for autonomous agents, which implicitly convey their intentions via the choice of actions.

The existing work on maximizing legibility [3–6] focuses on deterministic settings and cannot deal with stochastic environments. This is problematic for tasks such as autonomous driving, where the dynamics of the environment can be stochastic.

Another important challenge when maximizing legibility is how to balance the legibility score against the traditional plan execution cost. In our intersection example, while the driver of the non-yielding red car might want to make their intention clear, they probably would not want to risk getting into a crash.

In this paper, we introduce legible planning for a Markov decision process (MDP). We propose a constraint-based approach to balance the legibility and underlying rewards.

Proc. of the 19th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2020), B. An, N. Yorke-Smith, A. El Fallah Seghrouchni, G. Sukthankar (eds.), May 9–13, 2020, Auckland, New Zealand.



(a) Reward maximizing trace. **(b) Legibility maximizing trace.**

Figure 1: Example of traces produced by optimal policies that maximize reward (a) versus legibility (b).

2 LEGIBLE MDP

For an agent to make its intention clear, it is necessary to make an assumption on how its behavior reflects its intention. Hence, to tackle the problem of legible planning, we assume:

$$\pi_\theta(s,a) \propto exp(\beta Q^*_\theta(s,a)) \tag{1}$$

where a MDP is parameterized by θ . Intuitively, θ represents the goal/intention of the agent. At each time step, an agent takes an action with the probability exponentially proportionate to how good ($Q^*_\theta(s,a)$) the action is at the current state. β is a hyper-parameter representing how rational the agent is assumed to be. It has been demonstrated that the posterior probability of possible goals using Eqn. 1 and average human ratings over possible goals agree with each other [2].

We define a *legible MDP*, whose states contain the mental state of the observer according to the assumption Eqn. 1.

Definition 2.1. Given a MDP $M = \langle S, A, T, r, \gamma, \iota \rangle$, the set of possible parameters Θ, the true parameter θ^*, any distance measure between two beliefs ($dist : \Delta^\Theta \times \Delta^\Theta \to \mathcal{R}$), a new discount factor γ^L, a hyper-parameter β and the prior over the parameters P_Θ, a legible MDP $M^L = \langle S^L, A^L, T^L, r^L, \gamma^L, \iota^L \rangle$ is defined as follows:

- $S^L = S \times \Delta^{|\Theta|}$ where $\Delta^{|\Theta|}$ is a simplex over parameters.
- $A^L = A$.

- $T^L(\langle s, b\rangle, a, \langle s', b'\rangle) =$
  $\begin{cases} T_{\theta^*}(s,a,s') & b'(\theta) = \frac{T_\theta(s,a,s')\pi_\theta(s,a)b(\theta)}{\sum_{\theta'} T_{\theta'}(s,a,s')\pi_{\theta'}(s,a)b(\theta')} \\ 0 & otherwise \end{cases}$
  where $b'(\theta)$ is updated belief on the parameter θ and $\pi_\theta(s,a) \propto exp(\beta Q^*_\theta(s,a))$ using Eqn. 1.
- $r^L(\langle s, b\rangle, a, \langle s', b'\rangle) = -dist(b', b^*)$ where b^* is a one hot vector over the possible parameters (1 for the true parameter).
- $\iota^L = \langle \iota, [P_\Theta(\theta)|\theta \in \Theta]\rangle$.

For a legible MDP M^L, we call the parameterized MDP M *underlying MDP*. *dist* could be the L^2 norm, the Kullback-Leibler divergence or any distance measure. We use L^2 norm throughout the paper.

To justify the inclusion of belief over parameters in state factors, we make the following observation:

Remark. *An optimal policy of a legible MDP depends on the current belief of the observer.*

Of particular interest is a legible MDP where each parameter corresponds to a goal. We observe the following:

Remark. *For a legible MDP over goals, when*

- *A includes the special action* declare. *For each* M_g, declare *is only applicable at g, leading to the terminal absorbing state* s_∞ *with the probability 1.*
- *For the other actions,* T_g *is identical for every* $g \in \Theta$.
- $P_\Theta(g) \neq 0$ *for all* $g \in \Theta$,
- $V^*_g(s)$ *is bounded for all* $s \in S$ *and* $g \in \Theta$,
- $\iota \neq s_\infty$,

the belief has collapsed to the true goal iff $S_t = s_\infty$.

3 BALANCING LEGIBILITY AND COSTS

Balance legibility and the underlying rewards (or costs) presents a complex trade-off. We first observe that maximizing legibility and underlying rewards are two orthogonal problems.

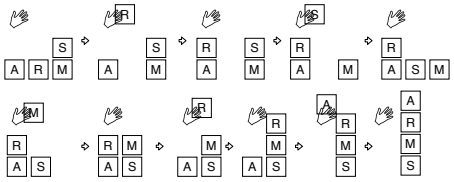
Remark. *A maximally legible policy* (π^*_L) *could be arbitrarily worse than an optimal policy* π^* *in terms of underlying rewards.*

In practice, it is often beneficial to strike a balance between legibility and underlying rewards. A widely used method for handling multiple objectives for an MDP is a Constrained MDP, or CMDP [1]. CMDP is a variant of MDP which maximizes the primary objective while having constraints on secondary objectives. In our case, legibility is the primary objective to maximize while underlying rewards are the secondary objective. Note that the constraints are only on expected returns. A widely used method for solving CMDP is based on linear programming [1]. This method, however, would require LP variables for each of the reachable states. For a legible MDP, this could be problematic as the number of reachable states could be large even for a finite-horizon version of the problem.

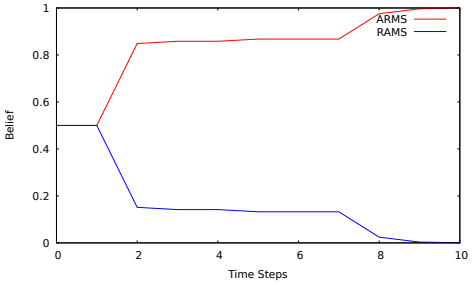
Instead, we propose to solve a more tractable version of the problem where we have bounds (δ) at each time step instead of the entire policy.

$$\pi^*_L(\delta) = \text{argmax}\ V^{\pi_L}(\iota^L), \tag{2}$$

$$\text{s.t. } |V^*(s) - Q^*(s, \pi_L(s))| \leq \delta \text{ for all } s^L \in S^L. \tag{3}$$



(a) Maximally legible policy with $\beta = 1.0$ and $H = 10$.



(b) Assumed belief changes over time.

Figure 2: Legible policy for Stochastic Blocks World.

4 ILLUSTRATIVE EXAMPLES

Stochastic Blocks World is a stochastic version of Blocks World. Picking up a block always succeeds with the probability 1. Putting down a block, however, fails with the probability 0.1 (in this case, the block falls on the table). Each action has the negative reward of -1. Starting from the initial state, there are two possible goals. One is to spell "ARMS" and the other is to spell "RAMS". Suppose the agent's true goal is to spell "ARMS". The optimal policy in terms of the underlying domain rewards is to first unstack the block "S". However, this is a part of the optimal policy to spell "RAMS" as well. Fig. 2a shows one trace the maximally legible policy can produce, where the agent first stacks the block "R" on top of "A". This makes harder or more costly to spell "RAMS", making the agent's intention to spell "ARMS" clearer. Fig. 2b shows the assumed belief about parameters over time according to our model.

For the intersection example earlier, when maximizing only for legibility, the red car does not necessarily stop even when the car ahead of it decides to decelerate. This is because, as we noted earlier, maximizing legibility and underlying rewards are two orthogonal problems. Having constraints on the underlying reward would prevent the red car from crashing into the car ahead of it.

5 CONCLUSION

We propose legible MDPs, where the agent's objective is to make its intention clear. We show that a maximally legible policy depends on the current belief of the observer, justifying our proposed algorithm to solve legible MDPs. A maximally legible policy, however, could be arbitrarily bad in terms of underlying rewards. Hence, we propose a constraint-based approach to balance the tradeoff between legibility and underlying rewards. Finally, the usefulness of legible MDPs is demonstrated through examples.

Acknowledgements This research was supported by the National Science Foundation grant number IIS-1724101.

REFERENCES

[1] *Eitan Altman. 1999.* Constrained Markov decision processes. *Vol. 7. CRC Press.*

[2] *Chris L. Baker, Rebecca Saxe, and Joshua B. Tenenbaum. 2009. Action understanding as inverse planning.* Cognition *113 (2009), 329–349.*

[3] *Anca Dragan, Kenton C. T. Lee, and Siddhartha Srinivasa. 2013. Legibility and predictability of robot motion. In* ACM/IEEE International Conference on Human-Robot Interaction. *301–308.*

[4] *Anca Dragan and Siddhartha Srinivasa. 2013. Generating Legible Motion. In* Proceedings of Robotics: Science and Systems. *Berlin, Germany.*

[5] *Anca D. Dragan, Shira Bauman, Jodi Forlizzi, and Siddhartha S. Srinivasa. 2015. Effects of Robot Motion on Human-Robot Collaboration. ACM, New York, USA, 51–58.*

[6] *Aleck M. MacNally, Nir Lipovetzky, Miquel Ramirez, and Adrian R. Pearce. 2018. Action Selection for Transparent Planning. In* International Conference on Autonomous Agents and MultiAgent Systems. *Stockholm, Sweden, 1327–1335. Stockholm, Sweden.*