Probabilistic Inference-based Reinforcement Learning

Quan Nguyen, Bert de Vries, Tjalling J. Tjalkens

Keywords: reinforcement learning, reward functions, probabilistic modeling, Bayesian inference.

Abstract
We introduce probabilistic inference-based reinforcement learning (PIReL), an approach to solve decision making problems by treating them as probabilistic inference tasks. Unlike classical reinforcement learning, which requires explicit reward functions, in PIReL they are implied by probabilistic assumptions of the model. This would enable a fundamental way to design the reward function by model selection as well as bring the potential to apply existing probabilistic modeling techniques to reinforcement learning problems.

1. Introduction
Reinforcement learning (RL) is a domain in machine learning concerning with how an agent makes decisions in an uncertain environment. In the traditional approach, the agent learns how to do a certain task by maximizing the expected total rewards. However, the reward functions are often handcrafted for specific problems than based on a general guideline.

In contrast to classical RL, probabilistic inference-based reinforcement learning (PIReL) treats the action as a hidden variable in a probabilistic model. Hence choosing actions that lead to the desired goal states can be treated in a straightforward manner as probabilistic inference.

This idea was in fact first proposed by (Attias, 2003). Our contribution is to extend the original framework so that it can take into account uncertainties about the goals. The extended framework shows its connection to classical RL. Particularly the reward function and discount factor in classical RL can be seen as certain probabilistic assumptions in the model. This interpretation provides us with a way to design appropriate reward function, by e.g., model selection.

2. Problem Modeling
The model is based on the Markov Decision Process. The interaction between the agent and the environment occurs in a time sequence, so subscription is used to indicate the time step. Under the Markov assumption, when the environment is in state $s_t$ (which is supposed to be fully observed by the agent), receives an action $a_t$ from the agent, changes to a new state $s_{t+1}$. The generative model is specified as:

$$p(s_1:T,a_1:T) = \pi \prod_{t=1}^{T-1} p(s_{t+1}|s_t,a_t), \quad (1)$$

where $\pi \triangleq \prod_{t=1}^{T-1} p(a_t)$ is the action prior (prior policy), and $p(s_{t+1}|s_t,a_t)$ is the transition probability. Unlike the standard MDP, there is no explicit reward here. Next we will explain how to infer actions.

2.1. Reinforcement Learning by Goal-based Probabilistic Inference
For the simplest decision making problem (Attias, 2003), at the initial state $s_1$, given a fixed horizon $T > 1$, and action prior $\pi$, the agent decides which actions $a_{1:T-1}$ should be done in order to archive the specified goal at the horizon, $s_T = g$. In the other words, we are interested in the posterior:

$$p(a_t|s_1,s_T = g), \forall t \in \{1,\ldots,T-1\}, \quad (2)$$

These probabilities have the form of a smoothing distribution, and the inference problem can be solved efficiently by a forward-backward-based algorithm.
3. Bayesian Policy and Relation to Classical Reinforcement Learning

In practice, it could be tricky to specify a desired goal precisely on $s_T$. Thus we introduce an abstract random binary variable $z$ that indicates whether $s_T$ is a good (rewarding) or bad state. The goal is instead set as $z = 1$ (good state).

In the special case when the given goal on $s_T$ is certain, we have $p(z = 1|s_T) \equiv \delta(s_T - g)$. And one could verify that

$$p(z = 1|s_1, T) = p(s_T = g|s_1, T)$$

while the updated policy (posterior) becomes

$$p(a_t|s_1, z = 1, T) = p(a_t|s_1, s_T = g), \forall t.$$  

For an uncertain goal on $s_T$, we have a generic form $\pi_T \triangleq p(z = 1|s_T)$, which is a probability function with input $s_T$ (since $z$ is always fixed at 1).

The current policy however still assumes that the horizon is known. Similarly, to accommodate the uncertainty about the horizon, we average over it. Without loss of generality, assume that horizon $T$ is upper bounded by $1 < T < \infty$, thus we have the full Bayesian policy

$$p(a_t|s_1, z = 1; \pi_T) = \sum_{T=2}^T \pi_T p(a_t|s_1, z = 1, T), \forall t,$$  

where $\pi_T \triangleq p(T)$ is the probability that the horizon is at time $T$. The marginal likelihood under $\pi_{ag} \triangleq \{\pi, \pi_T, \pi_z\}$ (policy, horizon, and goal distribution, respectively) is defined as:

$$p(z = 1|s_1; \pi_{ag}) = \sum_{T=2}^T \pi_T \int \pi_T \prod_{t=1}^{T-1} p(s_{t+1}|s_t; \pi) \text{d}s_{2:T}$$  

where $\gamma_T$ and $r_T$ denote the discount factor and instant reward at time $T$ respectively, while $R(s_T)$ is the reward function that returns a corresponding reward for state $s_T$.

It is clear that the horizon distribution $\pi_T$ behaves like the discount factor, while the goal distribution $\pi_{ag}$ acts like the reward function in classical reinforcement learning. In classical RL, both reward function and discount factor are often given. In contrast in our probabilistic framework, the optimal policy, horizon and goal distribution $\pi_{ag}$ that maximize the (log) marginal likelihood in eq. (4) can be estimated by e.g. EM algorithm (Dempster et al., 1977).

4. Related Work

The basic idea of PIReL originates from (Attias, 2003), where the agent infers actions in order to reach a certain goal at a fixed horizon. (Toussaint & Storkey, 2006) define the goal as to obtain the highest valued reward at the horizon; and propose an EM-based algorithm to derive the MAP estimation of action posterior with the horizon is marginalized out. By averaging over the horizons, the inferred policy also maximizes the expected return.

In the neuroscience and cognitive sciences literature, similar ideas to PIReL have been suggested, e.g., (Friston, 2010) and (Botvinick & Toussaint, 2012) discuss agents that infer actions that lead to a predefined goal. An alternative approach to improve the reward function is reward shaping, see e.g. (Ng et al., 1999), which however offers a limited alteration to the predefined rewards.

5. Conclusions

We discussed a framework where classical RL is recast as goal-based probabilistic inference. In this approach, there are no explicit reward functions as in classical RL, but instead the agent infers what actions to be do in order to reach a set of goals with different priorities. The reward function and discount factor can be interpreted as the goal and horizon distribution in this probabilistic framework. This potentially brings fundamental ways to improve or design an appropriate reward function and discount factor.

Acknowledgments

This work is part of the research programme HearScan with project number 13925, which is (partly) financed by the Netherlands Organisation for Scientific Research (NWO). We thank the anonymous reviewers.
for their thoughtful comments and suggestions.

References


