Mini-batch Bayesian Inverse Reinforcement Learning for Multiple Dynamics

Extended Abstract

Yusuke Nakata
Chiba University
a1019nakata@gmail.com

Sachiyo Arai
Chiba University
sachiyo@faculty.chiba-u.jp

ABSTRACT

Inverse reinforcement learning is a method that estates a reward function from experts demonstrations. Most existing inverse reinforcement learning methods assume that an expert gives demonstrations in a fixed environment, although the expert can provide demonstrations for a specific objective in multiple environments. In such cases, normal practice is to use demonstrations in multiple environments to estimate the expert's reward. Herein, we formulate this problem based on a Bayesian inverse reinforcement learning framework and propose a mini-batch Markov chain Monte Carlo method. An advantage of our method is scalability. Our proposed method is scalable with respect to a number of environments in which expert demonstrations are generated. Experimental results show quantitatively that the proposed method outperforms existing inverse reinforcement learning methods.

KEYWORDS

inverse reinforcement learning; bayesian inference

ACM Reference Format:
Yusuke Nakata and Sachiyo Arai. 2020. Mini-batch Bayesian Inverse Reinforcement Learning for Multiple Dynamics. In *Proc. of the 19th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2020), Auckland, New Zealand, May 9–13, 2020,* IFAAMAS, 2 pages.

1 INTRODUCTION

We formulate a problem of estimating rewards from expert demonstrations in multiple environments by using the Bayesian IRL (BIRL) framework [2]. This formulation enables rewards to be estimated from sub-optimal expert demonstrations with a stochastic policy with prior knowledge about expert rewards represented as a probability distribution. Furthermore, we propose a mini-batch Markov-chain Monte Carlo (MCMC) method for the formulated problem; this method uses part of expert demonstrations in each MCMC iteration to approximates the posterior distribution. The experimental results suggest that it is better to collect them in multiple environments than collecting it in a fixed environment, even if an available number of expert demonstrations in each environment are limited.

2 MARKOV DECISION PROCESS

A Markov decision process (MDP) is a classical formalization of the problem of sequential decision making. A finite MDP $\mathcal{M} = (E, R)$, comprises an environment $E = \langle\ \mathcal{S}, \mathcal{A}, T, \gamma\ \rangle$, and a reward R, where $\mathcal{S}$ is a finite set of states, $\mathcal{A}$ is a finite set of actions, $T(s'|s, a)$ is the probability of transition to $s' \in \mathcal{S}$ when the agent takes an action $a \in \mathcal{A}$ in state $s \in \mathcal{S}$, $\gamma \in [0, 1]$ is the discount factor, and the reward function $R : \mathcal{S} \rightarrow \mathbb{R}$ specifies the reward received in state $s \in \mathcal{S}$. An agent decides to take action a in state s with the probability specified by the policy $\pi : \mathcal{S} \times \mathcal{A} \rightarrow [0, 1]$.

A state value and an action value under the reward function R, and policy π are given respectively as,

$$V^{\pi}(s, R) = \mathbb{E}_{\pi,T}\left[\sum_{t=0}^{\infty} \gamma^t R(S_t) \middle| S_0 = s\right], \tag{1}$$

$$Q^{\pi}(s, a, R) = \mathbb{E}_{\pi,T}\left[\sum_{t=0}^{\infty} \gamma^t R(S_t) \middle| S_0 = s, A_0 = a\right]. \tag{2}$$

3 PROBLEM: BAYESIAN INVERSE REINFORCEMENT LEARNING FOR MULTIPLE DYNAMICS

BIRL-MD defines the problem of estimating the distribution of reward from the demonstrations of an expert under multiple dynamics. BIRL involves estimating the posterior distribution of reward $P(R|D) = P(R|D, E)$ given a fixed environment E and a dataset D, where D is generated by an expert in E. However, the proposed BIRL-MD involves estimating the posterior distribution of reward $P(R|\{(D_m, E_m)\}_{m=1}^{M})$ given environments with different dynamics $E_m = \langle \mathcal{S}, \mathcal{A}, T_m, \gamma \rangle$ and a set of datasets $\{D_m\}_{m=1}^{M}$ generated by an expert in each environment. In this paper, we model posterior distribution of the reward given a dataset of the expert $\{(E_m, D_m)\}_{m=1}^{M}$ can be expressed as

$$P(R|\{(D_m,E_m)\}_{m=1}^{M}) = \frac{1}{Z'} \exp\left(\kappa \sum_{m=1}^{M} \sum_{(s,a)\in D_m} Q^*(s, a, R, E_m)\right) P(R). \tag{3}$$

4 APPROACH

We propose an algorithm, mini-batch PolicyWalk for multiple dynamics (mini-batch PolicyWalk-MD), which samples reward from the posterior distribution $P(R|\{(D_m, E_m)\}_{m=1}^{M})$. The entire procedure of the mini-batch PolicyWalk-MD is detailed in Algorithm 1, which uses the technique explained in Section **??**. Our algorithm reduces the number of policy iterations for each MCMC iteration from M to a constant N. In the experiment, we show that our algorithm approximates the posterior distribution with a small constant N and reduces the computational time significantly.

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.

Algorithm 1 Mini-batch PolicyWalk for Multiple Dynamics

```
INPUT: Environments $\{E_m\}_{m=1}^M$, Demonstrations $\{D_m\}_{m=1}^M$, Prior $P(R)$, Step size $\delta$, Mini-batch size $N$
OUTPUT: Sampled Rewards $\{R_i\}_{i=1}^t$
Pick a random vector $R \in \mathbb{R}^{|S|}/\delta$
$\{\pi_m\}_{m=1}^M \leftarrow \{\text{Policy iteration}(E_m, R)\}_{m=1}^M$
for $i = 1$ do $t$
    Pick a reward vector $\tilde{R}$ uniformly at random from the neighbors of $R \in \mathbb{R}^{|S|}/\delta$
    $u \leftarrow$ Sample from uniform distribution $U(0,1)$
    $\tilde{\mathbb{N}} \leftarrow$ Sampled $N$ integers from $\{n \in \mathbb{N}| \; n \le M\}$ without repetition
    Compute $Q^\pi(s,a,R,E) \quad \forall\{s,a,(E_n,\pi_n)\} \in \mathcal{S} \times \mathcal{A} \times \{(E_n,\pi_n)\}_{n\in\tilde{\mathbb{N}}}$
    if $\exists\{s,a,(E,\pi)\} \in \mathcal{S} \times \mathcal{A} \times \{(E_n,\pi_n)\}_{n\in\tilde{\mathbb{N}}}, Q^\pi(s,\pi(s),\tilde{R},E) < Q^\pi(s,a,\tilde{R},E)$ then ▷ If any sampled policy is not optimal
        $\{\tilde{\pi}_n\}_{n\in\tilde{\mathbb{N}}} \leftarrow \{\text{Policy iteration}(E_n,\tilde{R})\}_{n\in\tilde{\mathbb{N}}}$
        if $\frac{1}{M}\log\left(u\frac{P(R)}{P(\tilde{R})}\right) < \frac{1}{N}\sum_{n\in\tilde{\mathbb{N}}}\log P\left(D_n,E_n\middle|\tilde{R}\right) - \log P\left(D_n,E_n\middle|R\right)$ then
            $R \leftarrow \tilde{R}$
            $\{\pi_n\}_{n\in\tilde{\mathbb{N}}} \leftarrow \{\tilde{\pi}_n\}_{n\in\tilde{\mathbb{N}}}$
    else if $\frac{1}{M}\log\left(u\frac{P(R)}{P(\tilde{R})}\right) < \frac{1}{N}\sum_{n\in\tilde{\mathbb{N}}}\log P\left(D_n,E_n\middle|\tilde{R}\right) - \log P\left(D_n,E_n\middle|R\right)$ then
        $R \leftarrow \tilde{R}$
    $R_i \leftarrow R$
```

5 EXPERIMENTS

The experimental environment is a windy grid world, in which each state has a wind direction, and the agent transitions to a wind direction with a certain probability regardless of the agent's action. Hence, we can create environments with different dynamics by varying the wind direction of each state. In this experiment, the probability of forced transition to the wind direction is set to 30%. The number of wind directions is five (i.e., up, down, left, right, and no wind), and the wind direction of each state is independent. The reward is 1.0 in a upper right-hand corner state (4, 4), and zero elsewhere.

We evaluate the estimated reward with a score that is known as the expected value difference (EVD) [1] which is a measure of how sub-optimal the learned policy is under the expert true reward. To calculate EVD, we used 100 environments with different wind directions (dynamics) generated from uniform distributions, and each experiment was conducted 10 times.

In Figure 1, the number of demonstrated environments is fixed to eight, and the number of environments used in each MCMC step (i.e., mini-batch size) are varied across {1, 2, 4, 8}. As Figure 1 shows, EVD does not change significantly with the mini-batch size N. Figure 2 evaluates our method with a mini-batch size of $N = 1$ by varying the number M of environments of expert's demonstrations. EVD decreases as the number of environments for reward estimation decreases.

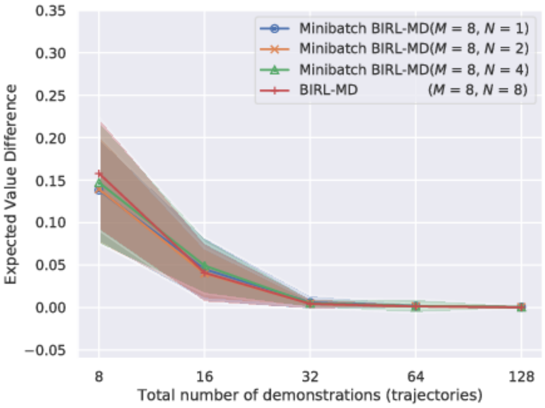

Figure 1: Evaluation of mini-batch BIRL-MD for different mini-batch size N.

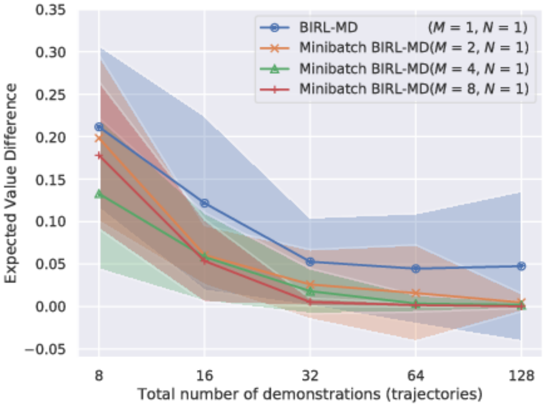

Figure 2: Evaluation of Minibatch BIRL-MD for Different Number of Environments of Expert's Demonstrations M with fixed mini-batch size N.

6 CONCLUSIONS AND FUTURE WORK

This paper formulates a Bayesian inverse reinforcement learning problem for expert demonstrations (i.e., sequences of both sensor inputs to expert and expert's actions) under multiple environments with different dynamics. An advantage of our method is its scalability with respect to the number of environments in which expert demonstrations are generated. Figure 1 and Table **??** show that our method can approximate the posterior distribution with a small mini-batch size in a computational time that is comparable with that of BIRL.

REFERENCES

[1] Sergey Levine, Zoran Popovic, and Vladlen Koltun. 2011. Nonlinear inverse reinforcement learning with gaussian processes. In *Advances in Neural Information Processing Systems.* 19–27.

[2] Deepak Ramachandran and Eyal Amir. 2007. Bayesian inverse reinforcement learning. *IJCAI International Joint Conference on Artificial Intelligence* (2007), 2586–2591.