Discount Factor Estimation in Inverse Reinforcement Learning

Babatunde H. Giwa, Chi-Guhn Lee

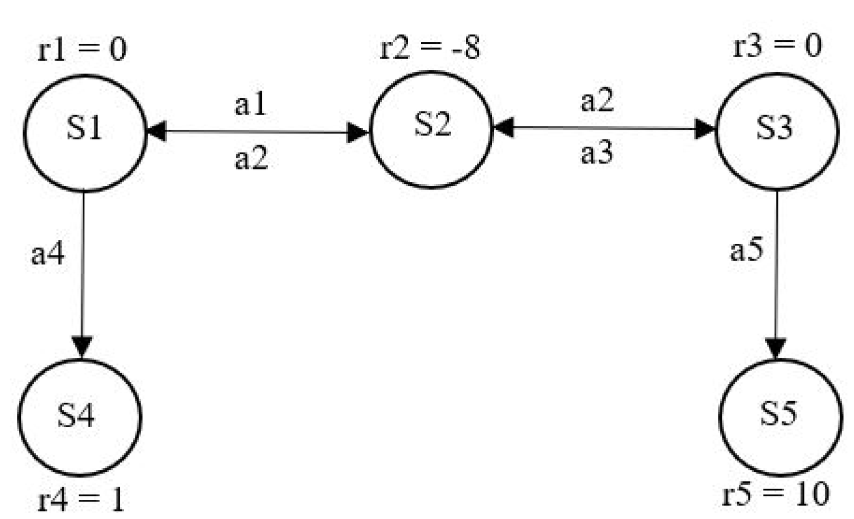
Introduction

- In a Markov decision process (MDP) environment, inverse reinforcement learning (IRL) primarily seeks to explain human behaviour, i.e., learn reward.
- In the IRL framework, experimental studies and theoretical intuitions show variability of learnt reward and optimal policy as discount factor (γ) changes.
- A feature-based gradient updates for simultaneous estimation of reward and discount factor.

Significance of Discount Factor

Given ground-truth policy (π^*) with a discount factor (γ^*). If γ^L is assumed in learning, i.e, $\gamma^* \neq \gamma^L$, learnt trajectories might differ from ground-truth.

Motivating Example



MDP with 5 states.

The optimal policies (originally π^*, learnt as $\tilde{\pi}^*$) differed when the discount factor varied as seen in state S_1.

	Original (($\gamma = 0.35$))		Learnt ($\tilde{\gamma} = 0.9$)	
State	Reward	π^*	Reward	$\tilde{\pi}^*$
S_1	0	a_4	0.1	a_2
S_2	-8	a_3	0.5	a_3
S_3	0	a_5	5.0	a_5
S_4	1	-	-0.1	-
S_5	10	-	9.9	-

Problem Definition

Given a finite-horizon MDP without discount factor (γ) and reward (R) with a set of trajectories, $\xi = \{(\tau_1, \tau_2, ..., \tau_{|\xi|})\}$, how do we jointly estimate discount factor and reward?

Mathematical Model for Solution Approach

Objective

$$\max \quad -\sum_{\tau\in\xi} p(\tau)\log p(\tau)$$

Constraints (subject to:)

$$\sum_{\tau\in\xi}\sum_{t=0}^{|\tau|}\gamma^t p(\tau)\phi(\tau;t) = \frac{1}{|\xi|}\sum_{\tau\in\xi}\sum_{t=0}^{|\tau|}\gamma^t\phi(\tau;t)$$

$$\sum_{\tau\in\xi} p(\tau) = 1$$

$$p(\tau) \geq 0$$

Following a Lagrangian and solution of first-order optimality equation, we evolve the following relation:

$$p(\tau) \propto e^{\sum_{t=0}^{|\tau|}\gamma^t\theta^\intercal\phi(\tau;t)} = e^{U(\tau)}$$

Goal: find θ^* and γ^* that maximizes the likelihood of demonstration set (ξ):

$$\theta^*, \gamma^* = \arg\max_{\theta,\gamma} \frac{1}{|\xi|}\prod_{\tau\in\xi}\frac{e^{\sum_{\tau\in\xi}\sum_{t=0}^{|\tau|}\gamma^t\theta^\intercal\phi(\tau;t)}}{Z}$$

Gradient Equations:

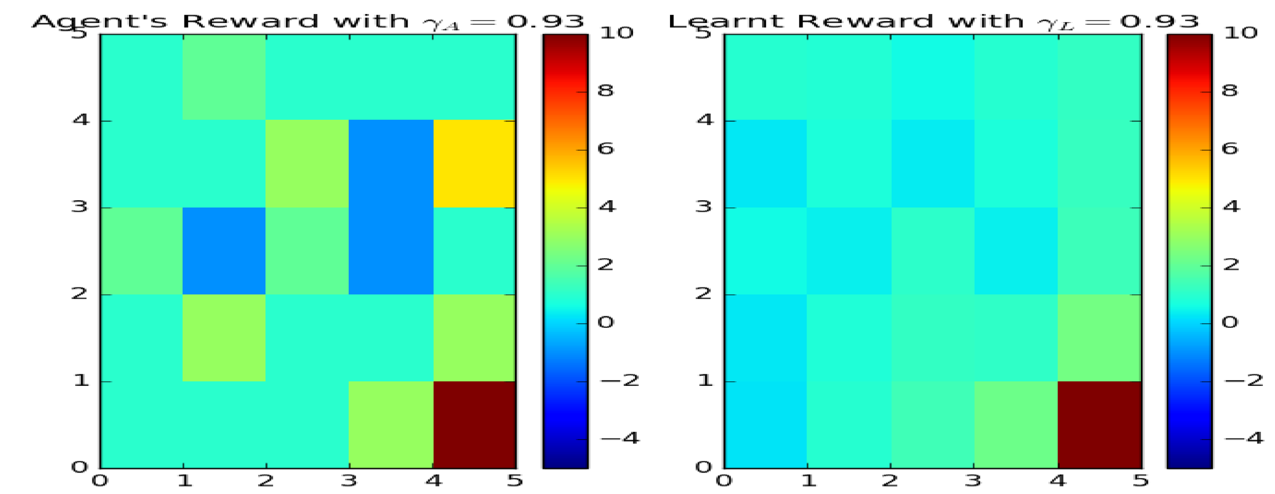
$$\nabla_\theta = \frac{1}{|\xi|}\sum_{\tau\in\xi}\sum_{t=0}^{|\tau|}\gamma^t\phi(\tau;t) - \sum_{s\in S}\sum_{t=0}^{|\tau|}\gamma^t P(s_t|\theta,\gamma)\phi(s_t)$$

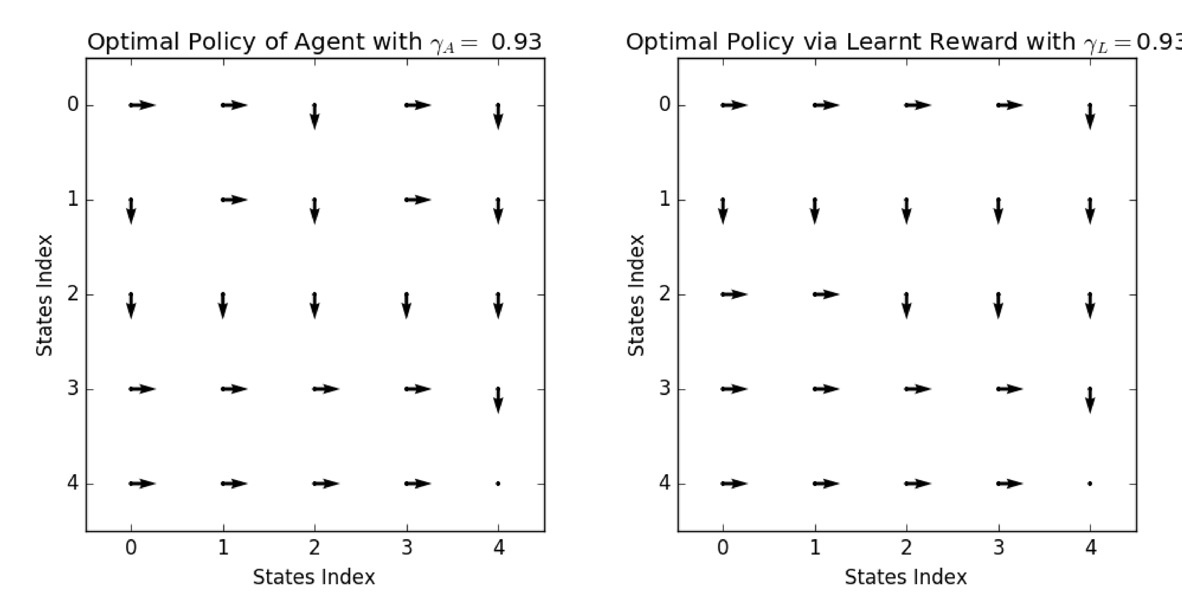
$$\nabla_\gamma = \frac{1}{|\xi|}\sum_{\tau\in\xi}\sum_{t=0}^{|\tau|}t\gamma^{t-1}\theta^\intercal\phi(\tau;t) - \sum_{s\in\xi}\sum_{t=0}^{|\tau|}t\gamma^{t-1}P(s_t|\theta,\gamma)\theta^\intercal\phi(s_t) + \lambda$$

where λ is a penalization to enforce boundary conditions on discount factor. With a learning rate (α), we define our update rules as follows:

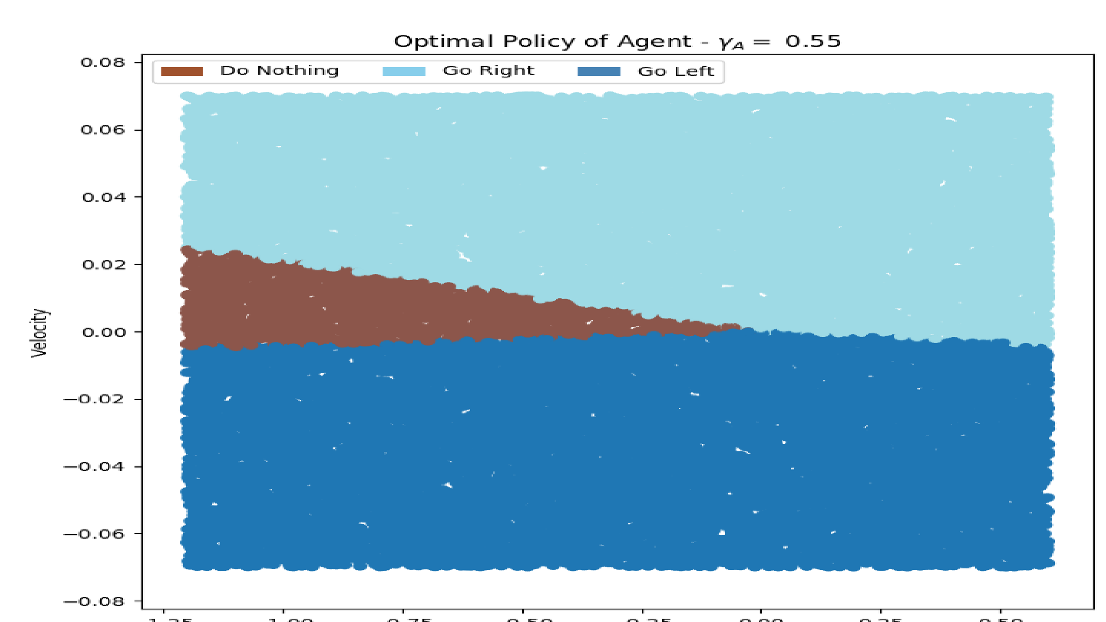
$$\theta \leftarrow \theta + \alpha\nabla_\theta$$
$$\gamma \leftarrow \gamma + \alpha\nabla_\gamma$$

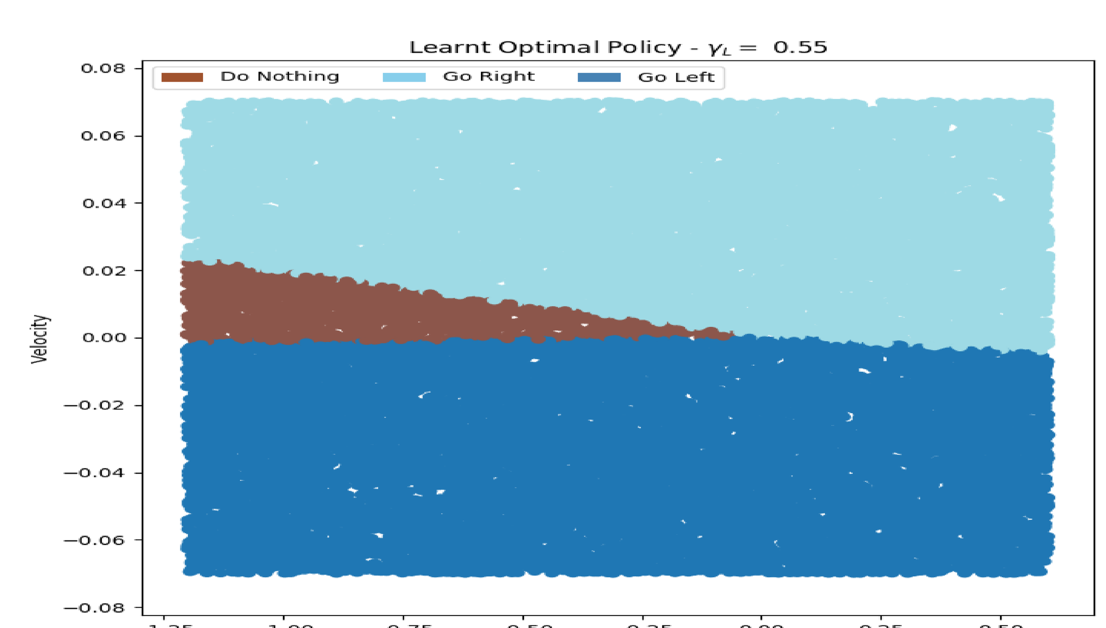
Experimental Results: Grid-World



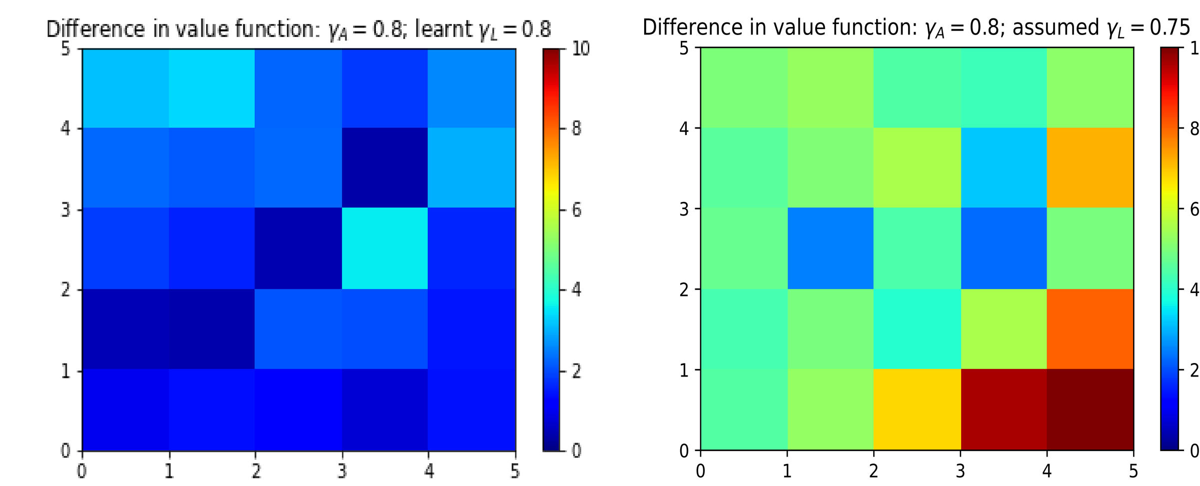


Mountain-Car Driving





Loss Metric: Value Function Difference (VFD) for a Grid-World Setup



(a) VFD (our approach) (b) VFD (baseline)

Concluding Remarks

- The utility-based approach facilitated the concurrent estimation of discount factor and reward in a model-based entropy framework.
- A suggestion for future work is sample complexity analysis of our approach given the VFD loss metric on real-world data.

Key References

- Ziebart, B.D., Bagnell, J.A. and Dey, A.K., 2010. Modeling interaction via the principle of maximum causal entropy. In ICML.
- Ng, A. Y., and Russell, S. J. 2000. Algorithms for inversereinforcement learning. InIcml, volume 1, 2.
- Sutton, R. S., and Barto, A. G. 2018. Reinforcement learn-ing: An introduction. MIT press.