Learning Guidance Rewards with Trajectory-space Smoothing
Tanmay Gangwani, Yuan Zhou, Jian Peng

Introduction and Motivation

- Reinforcement learning with end-of-episode feedback (episodic rewards)
- Representative of real-world decision making in robotics, finance, chemical-synthesis, healthcare. Below: a use-case in hardware design – learning a data-aware cache replacement policy

Guidance Rewards (Intuition and Definition)

Introduce a distribution over trajectories $M_\tau(\tau)$, parameterized by a reference trajectory $\tilde{\tau}$. Use this to define a modified RL objective:

$$\bar{\eta}(\pi_\theta) = \mathbb{E}_{\tilde{\tau} \sim \pi(\theta)} \left[ \mathbb{E}_{\tau \sim M_\tau} [R(\tau)] \right].$$

$M_\tau(\tau) \triangleq \delta(\tau = \tilde{\tau})$, recovers the standard RL objective

Let $p(\tilde{\tau})$ be some trajectory distribution. Define a reweighted $p(\tau)$ distribution to have mass only over trajectories that include the reference pair $(\tilde{s}, \tilde{a})$

$$p_{\tilde{S}}(\tau) \propto p(\tau) \mathbb{I}[(\tilde{s}, \tilde{a}) \in \tau]$$

Define $M_\tau(\tau)$ such that trajectories that overlap with the reference trajectory $\tilde{\tau}$ (in terms of states and actions) are preferred

$$M_\tau(\tau) = (1 - \gamma) \sum_{i=0}^{\infty} \gamma^i p_{\tilde{S}}(\tau_i) ; \tilde{\tau} = \{\tilde{s}_i, \tilde{a}_i\}_{i=0}^{\infty}$$

Insert in 1 and rearrange to get

$$\bar{\eta}(\pi_\theta) = \mathbb{E}_{(\tilde{s}, \tilde{a}) \sim \pi(\theta)} \left[ \sum_{i=0}^{\infty} \gamma^i r_\phi(\tilde{s}_i, \tilde{a}_i) \right] ; r_\phi(\tilde{s}, \tilde{a}) = \mathbb{E}_{\tau \sim p_{\tilde{S}}(\tau)} [R(\tau)]$$

Guidance Reward for a state-action pair can be computed as the expected return of the past trajectories which include that pair! (please see paper for further intuition as a uniform credit assignment mechanism)

Solution philosophy (reward decoupling) [1]

- Don’t use the episodic rewards directly for RL optimization
- Learn surrogate rewards to guide the agent towards maximizing the true (episodic) rewards

We refer to these as Guidance Rewards. Desired properties:

- affordable dense supervision
- efficient to compute (without any auxiliary learned networks)
- easy to incorporate into the state-of-the-art RL algorithms

Experiments

1) Tabular Q-learning in Grid-world with Episodic Rewards

Guidance rewards incorporated with SAC – referred to as SAC (IRCR). We contrast it with SAC (w/ env. rewards) and two recent approaches for dealing with sparse, delayed rewards – GASIL and Reward-Regression. In these environments, a reward is provided only at the last timestep of every episode

2) Soft Actor-Critic on MuJoCo Locomotion with Episodic Rewards

Guidance rewards used in a multi-agent particle domain where agents navigate to various points of interest in a 2D world with continuous state- and action-space. RL algorithms used are TD3 and distributional RL (CS1)