Paper & Project Proposal

Group 2

309505018 郭俊廷
309551026 黃柏愷

Paper Proposal

Soft Actor-Critic:
Off-Policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor

Issues to solve

Sample inefficency

Meticulous hyperparameter tuning

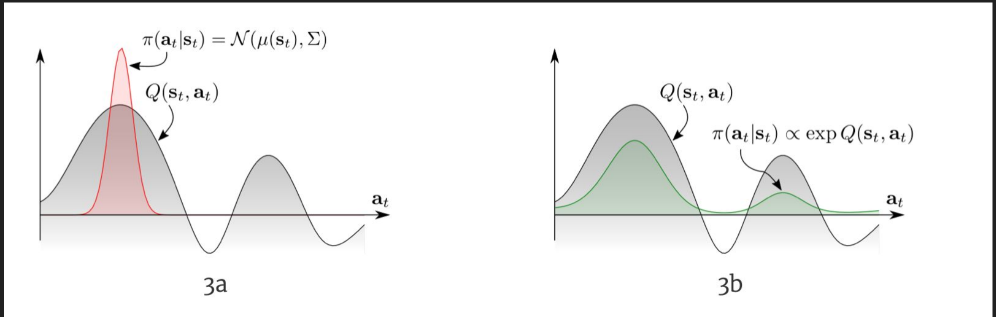
m

Maximum Entropy Reinforcement Learning

$$\pi^* = \arg\max_{\pi} \mathbb{E}_{(s_t, a_t) \sim \rho_\pi} \left[\sum_t R(s_t, a_t) \right]$$

$$\pi^* = \arg\max_{\pi} \mathbb{E}_{(s_t, a_t) \sim \rho_\pi} \left[\sum_t \underbrace{R(s_t, a_t)}_{reward} + \alpha \underbrace{H(\pi(\cdot|s_t))}_{entropy} \right]$$

Soft Q-Learning



- Learn Soft Q directly
- Policy is intractable in continuous domain

Soft Actor Critic

- Learn Soft Q of policy and the policy jointly
- like DDPG, but with stochastic policy

Soft Policy Iteration

- Soft Policy evaluation

$$\mathcal{T}^{\pi} Q(\mathbf{s}_t, \mathbf{a}_t) \triangleq r(\mathbf{s}_t, \mathbf{a}_t) + \gamma \, \mathbb{E}_{\mathbf{s}_{t+1} \sim p} \left[V(\mathbf{s}_{t+1}) \right],$$

where

$$V(\mathbf{s}_t) = \mathbb{E}_{\mathbf{a}_t \sim \pi} \left[Q(\mathbf{s}_t, \mathbf{a}_t) - \log \pi(\mathbf{a}_t | \mathbf{s}_t) \right]$$

Soft Policy Iteration

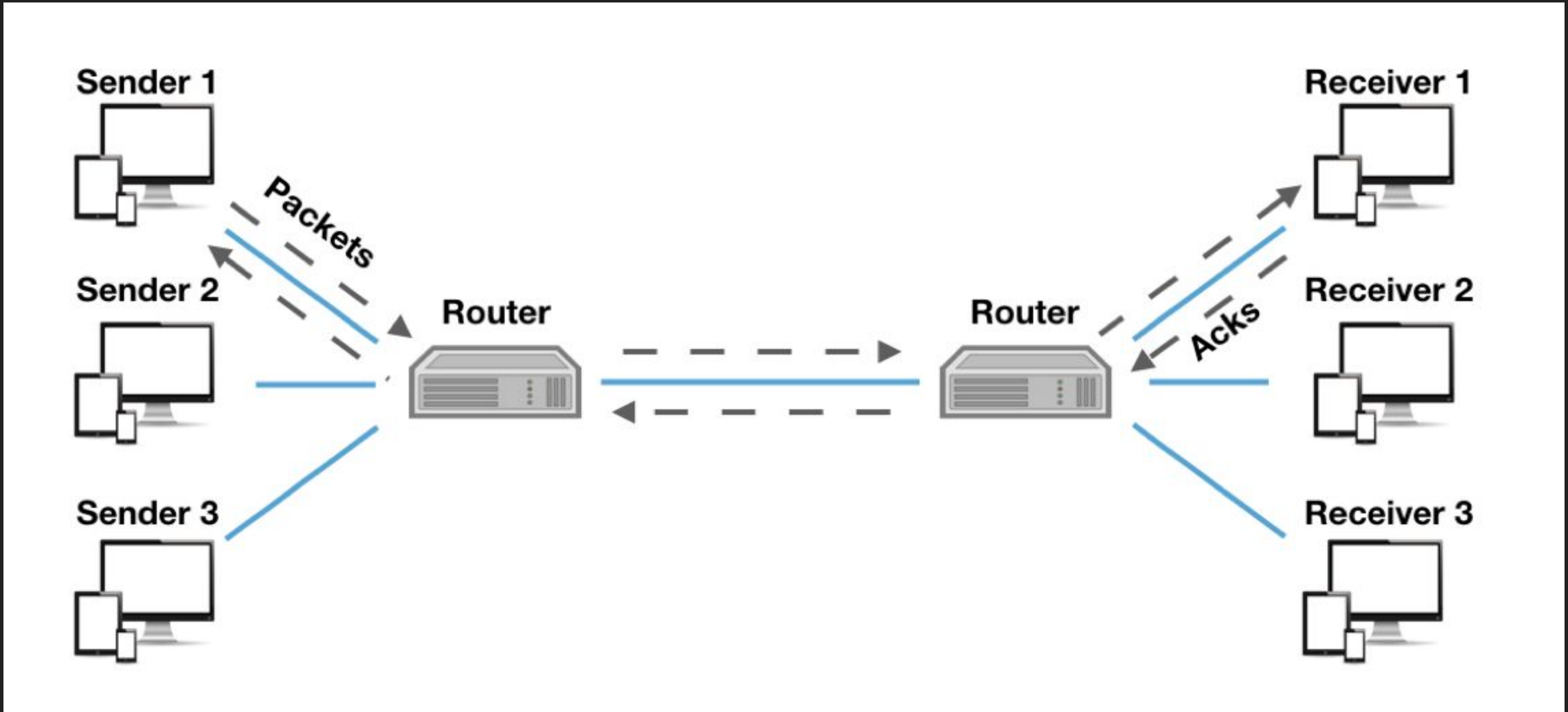
- Soft Policy Improvement

$$\pi_{\text{new}} = \arg\min_{\pi' \in \Pi} \mathrm{D}_{\mathrm{KL}}\left(\pi'(\,\cdot\,|\mathbf{s}_t) \,\middle\|\, \frac{\exp\left(Q^{\pi_{\text{old}}}(\mathbf{s}_t, \,\cdot\,)\right)}{Z^{\pi_{\text{old}}}(\mathbf{s}_t)}\right)$$

Project Proposal

Internet Congestion Control with extracted RL

Internet Congestion Control



A Deep Reinforcement Learning Perspective on Internet Congestion Control ICML 2019

State,Action, Reward Design

State

- Latency Gradient
- Latency Ratio
- Sending Ratio

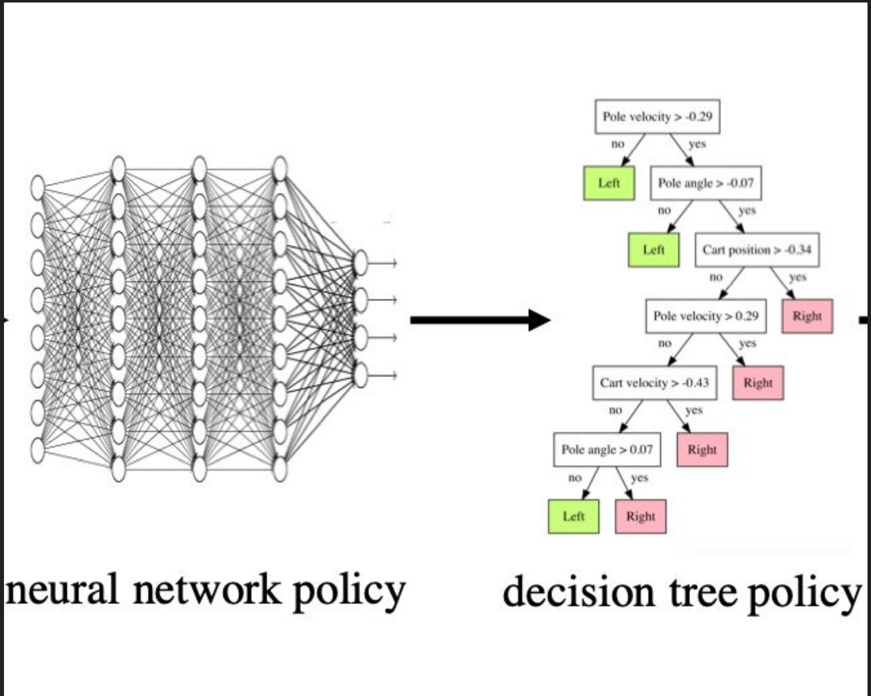
Action

- Sending rate

Reward

- $10 * \text{throughput} - 1000 * \text{latency} - 2000 * \text{loss}$

Policy Extraction via Q-Dagger



Verifiable Reinforcement Learning via Policy Extraction NIPS 2018