CS 594 Modern Reinforcement Learning

Lecture 3: Policy Gradients

Optimization

$$\max_{\theta} J(\theta)$$

$$\theta_{t+1} = \theta_t + \alpha \nabla \widehat{J(\theta_t)}$$

Parameterized Policies

- "Direct Tabular" Policy Parameterization:

$$\pi(s,a) = \theta_{s,a}$$

- "Softmax Tabular" Policy Parameterization:

$$\pi(s,a) = \frac{\exp(\theta_{s,a})}{\sum_a' \exp(\theta_{s,a'})}$$

- "Softmax Neural" Policy Parameterization:

$$\pi(s,a) = \frac{\exp(f_\theta(s,a))}{\sum_a' \exp(f_\theta(s,a'))}$$

Calculating J

$$v_\pi(s) = E_\pi[G_t \mid S_t = s] = E_\pi\left[\sum_{k=t}^{T} \gamma^{k-t} R_{k+1} \middle| S_t = s\right]$$

$$= \sum_{k=t}^{T} \sum_{s_k, a_k, r_{k+1}} \Pr_\pi[S_k = s_k | S_t = s_t]\, \pi(a_k|s_k) p(r_{k+1}|s_k, a_k) \gamma^{k-t} r_{k+1}$$

$$= \sum_{s,a,r} \eta(s)\, \pi(a|s) p(r|s,a)\, r$$

where we use the discounted occupancy measure

$$\eta(s) = \sum_{k=t}^{T} \Pr_\pi[S_k = s_k | S_t = s_t]\, \gamma^{k-t}$$

Discounted Occupancy Measure

$$\eta(s) = \sum_{k=t}^{T} \Pr_{\pi}[S_k = s_k | S_t = s_t]\, \gamma^{k-t}$$

Calculating ∇J

$$J(\theta) = \sum_{s,a,r} \eta(s)\, \pi(a|s,\theta) p(r|s,a) r$$

$$\nabla J(\theta) = \sum_{s,a,r} \nabla\eta(s)\pi(a|s,\theta)p(r|s,a)\, r + \eta(s)\nabla\pi(a|s,\theta)p(r|s,a)\, r$$

Uh oh!

Policy Gradient Theorem

$$\nabla J(\theta) = \nabla v_\pi(s) = \nabla \sum_a \pi(a|s,\theta) q_\pi(s,a)$$

$$= \sum_a \big(\nabla \pi(a|s,\theta)\big) q_\pi(s,a) + \pi(a|s,\theta) \nabla q_\pi(s,a)$$

$$= \sum_a \big(\nabla \pi(a|s,\theta)\big) q_\pi(s,a) + \pi(a|s,\theta) \nabla \sum_{s',r} p(s',r|s,a)(r + \gamma v_\pi(s'))$$

$$= \sum_a \big(\nabla \pi(a|s,\theta)\big) q_\pi(s,a) + \pi(a|s,\theta) \sum_{s'} p(s'|s,a) \gamma \nabla v_\pi(s')$$

Policy Gradient Theorem

$$\nabla v_\pi(s) = \sum_a \big(\nabla \pi(a|s,\theta)\big) q_\pi(s,a) + \pi(a|s,\theta) \sum_{s'} p(s'|s,a)\gamma \nabla v_\pi(s')$$

$$= \sum_a \big(\nabla \pi(a|s,\theta)\big) q_\pi(s,a) + \sum_{s'} \gamma \Pr_\pi[S_{t+1} = s' | S_t = s] \, \nabla v_\pi(s')$$

$$\nabla J(\theta) = \sum_s \eta(s) \sum_a \big(\nabla \pi(a|s,\theta)\big) q_\pi(s,a)$$

Policy Gradient Theorem

$$\nabla v_\pi(s) = \sum_a \big(\nabla\pi(a|s,\theta)\big) q_\pi(s,a) + \pi(a|s,\theta) \sum_{s'} p(s'|s,a)\gamma \nabla v_\pi(s')$$

$$= \sum_a \big(\nabla\pi(a|s,\theta)\big) q_\pi(s,a) + \sum_{s'} \gamma \Pr_\pi[S_{t+1} = s'|S_t = s]\, \nabla v_\pi(s')$$

$$\nabla J(\theta) = \sum_s \eta(s) \sum_a \big(\nabla\pi(a|s,\theta)\big) q_\pi(s,a)$$

$$\nabla J(\theta) = \sum_s \eta(s) \sum_a \pi(a|s,\theta) q_\pi(s,a) \frac{\nabla\pi(a|s,\theta)}{\pi(a|s,\theta)}$$

$$\nabla J(\theta) = \sum_s \eta(s) \sum_a \pi(a|s,\theta) q_\pi(s,a)\, \nabla \ln(\pi(a|s,\theta)$$

REINFORCE

- $\nabla J(\theta) = \sum_s \eta(s) \sum_a \pi(a|s,\theta) q_\pi(s,a) \nabla \ln(\pi(a|s,\theta))$

- Sample a trajectory
- For each time t calculate the Monte Carlo Estimate:
  - $g_t = \sum_{k=t}^{T-1} \gamma^{k-t} r_{t+1}$
- Calculate gradient estimate:
  - $\widehat{\nabla J(\theta)} = \sum_{t=0}^{T} \gamma^t \, \mathrm{g_t} \nabla \ln(\pi(a_t|s_t,\theta)$
- Update:
  - $\theta_{t+1} = \theta_t + \alpha \widehat{\nabla J(\theta_t)}$