Lecture11: Policy Gradients for Reinforcement Learning

Notes taken by squarezhong

Repo address: squarezhong/SDM5008-Lecture-Notes

Lecture11: Policy Gradients for Reinforcement Learning

Problem Formulation

Derivation of Policy Gradient

Compute $\nabla_{\theta} U(\theta)$

Summary of Policy Gradient Derivation

Problem Formulation

• RL Problem: find optimal policy π^*

$$V^*(s) = \max_{\pi}(E_{ au\sim\pi}(R(au)|S_0=s))$$

- Parameterize policy by a parameter vector θ , denote:
 - \circ $\pi_{ heta} := \pi_{ heta}(\cdot|s)$
 - $\circ \ P_{ heta}(au)$: the likelihood of trajectory au under policy $\pi_{ heta}$

$$P_{ heta}(au) = P(s_0) \prod_{t=0}^{T-1} \pi_{ heta}(a_t|s_t) P(s_{t+1}|s_t,a_t)$$

- Reformulate the MDP problem as an optimization problem
 - \circ assume s_0 distribution is included in trajectory likelihood $P_{ heta}(au)$
 - Utility function:

$$U(heta) = E_{ au \sim P_{ heta}(au)}[R(au)] = \sum_{ au} P_{ heta}(au) R(au)$$

 \circ RL problem reduces to finding the optimal policy parameter heta

$$\max_{ heta} U(heta) = \max_{ heta} \sum_{ au} P_{ heta}(au) R(au)$$

- \bullet dim is high
- $U(\theta)$: hard to compute (approximate)
- \circ Policy gradient use $\mathbf{1}^{st}$ order gradient ascend method

Derivation of Policy Gradient

Optimize $U(\theta)$ involves evaluating $U(\theta)$

- think of τ as a random variable (in trajectory space)
- By monte carlo: $U(\theta) = \sum_{ au} P_{ heta}(au) R(au) pprox rac{1}{N} \sum_{i} R(au^{(i)}), \quad au^{(i)} \stackrel{ ext{iid}}{\sim} P_{ heta}(au)$

Compute $\nabla_{\theta}U(\theta)$

$$abla_{ heta}U(heta) = rac{\partial}{\partial heta} \sum_{ au} P_{ heta}(au) R(au) = \sum_{ au} R(au) rac{\partial}{\partial heta} P_{ heta}(au)$$

Note the fact that $rac{
abla_{ heta}P_{ heta}(au)}{P_{ heta}(au)}=
abla_{ heta}\log(P_{ heta}(au))$

Then

$$\begin{split} \nabla_{\theta} U(\theta) &= \sum_{\tau} R(\tau) P_{\theta}(\tau) \frac{\nabla_{\theta} P_{\theta}(\tau)}{P_{\theta}(\tau)} = \sum_{\tau} R(\tau) P_{\theta}(\tau) \nabla_{\theta} \log(P_{\theta}(\tau)) \\ &= \sum_{\tau \sim P_{\theta}(\tau)} \left(R(\tau) \nabla_{\theta} \log(P_{\theta}(\tau)) \right) \\ &\approx \frac{1}{N} \sum_{\tau} R(\tau^{(i)}) \nabla_{\theta} \log(P_{\theta}(\tau^{(i)})) \quad \text{(monte carlo)} \end{split}$$

Substitute concrete expression

$$egin{aligned}
abla_{ heta} \log P_{ heta}(au) &=
abla_{ heta} \left(\log P(s_0) + \sum_{t=0}^{T-1} \log \pi_{ heta}(a_t|s_t) + \log P(s_{t+1}|s_t,a_t)
ight) \ &=
abla_{ heta} \sum_{t=0}^{T-1} \log \pi_{ heta}(a_t|s_t) \end{aligned}$$

Then

$$egin{aligned}
abla_{ heta}U(heta) &= \sum_{ au\sim P_{ heta}(au)}^{E} \left(R(au)
abla_{ heta} \sum_{t=0}^{T-1} \log \pi_{ heta}(a_{t}|s_{t})
ight) \ &= rac{1}{N} \sum_{i} R(au^{(i)})
abla_{ heta} \sum_{t=0}^{T-1} \log \pi_{ heta}(a_{t}^{(i)}|s_{t}^{(i)}) \quad ext{(monte carlo)} \end{aligned}$$

Summary of Policy Gradient Derivation

- ullet Roll out trajectories $au^{(i)} \sim P_{ heta}(\cdot), \ i=1,\cdots,N$
- Compute the empirical mean $\hat{g} = \frac{1}{N} \sum_i R(\tau^{(i)}) \nabla_{\theta} \sum_{t=0}^{T-1} \log \pi_{\theta}(a_t^{(i)}|s_t^{(i)})$
- By Monte Carlo, we know $E(\hat{g}) =
 abla_{ heta} U(heta)$
- In practice, the sample mean estimate \hat{g} has a high variance and many ways can be used to reduce the variance, leading to different algorithms.