

RL: Policy Gradient

Analytic Gradient

Marius Lindauer



Automated
Machine Learning
Hannover

Computing the gradient analytically

- We now compute the policy gradient analytically
- Assume policy π_θ is differentiable whenever it is non-zero
- and we know the gradient $\nabla_\theta \pi_\theta(s, a)$
- Denote a state-action trajectory as

$$\tau = (s_0, a_0, r_0, \dots, s_{T-1}, a_{T-1}, r_{T-1}, s_T)$$

- Use $R(\tau) = \sum_{t=0}^T R(s_t, a_t)$ to be the sum of rewards for a trajectory τ

↪ Focusing for now on $V(s_0, \theta) = \sum_{\tau} P(\tau; \theta) R(\tau)$

Computing the gradient analytically

- Denote a state-action trajectory as

$$\tau = (s_0, a_0, r_0, \dots, s_{T-1}, a_{T-1}, r_{T-1}, s_T)$$

- Use $R(\tau) = \sum_{t=0}^T R(s_t, a_t)$ to be the sum of rewards for τ
- Policy value is

$$V(\theta) = \mathbb{E} \left[\sum_{t=0}^T R(s_t, a_t); \pi_{\theta} \right] = \sum_{\tau} P(\tau; \theta) R(\tau)$$

where $P(\tau; \theta)$ is used to denote the probability over trajectories when executing policy π_{θ}

- In this new notation, our goal is to find the policy parameters θ^*

$$\theta^* \in \arg \max_{\theta} V(\theta) = \arg \max_{\theta} \sum_{\tau} P(\tau; \theta) R(\tau)$$

Computing the gradient analytically

- Our goal is to find the policy parameters θ^*

$$\theta^* \in \arg \max_{\theta} V(\theta) = \arg \max_{\theta} \sum_{\tau} P(\tau; \theta) R(\tau)$$

- Take the gradient with respect to θ :

$$\begin{aligned}\nabla_{\theta} V(\theta) &= \nabla_{\theta} \sum_{\tau} P(\tau; \theta) R(\tau) \\&= \sum_{\tau} \nabla_{\theta} P(\tau; \theta) R(\tau) \\&= \sum_{\tau} \frac{P(\tau; \theta)}{P(\tau; \theta)} \nabla_{\theta} P(\tau; \theta) R(\tau) \\&= \sum_{\tau} P(\tau; \theta) R(\tau) \underbrace{\frac{\nabla_{\theta} P(\tau; \theta)}{P(\tau; \theta)}}_{\text{likelihood ratio}} \\&= \sum_{\tau} P(\tau; \theta) R(\tau) \nabla_{\theta} \log P(\tau; \theta)\end{aligned}$$

Computing the gradient analytically

- Our goal is to find the policy parameters θ^*

$$\theta^* \in \arg \max_{\theta} V(\theta) = \arg \max_{\theta} \sum_{\tau} P(\tau; \theta) R(\tau)$$

- Take the gradient with respect to θ :

$$\nabla_{\theta} V(\theta) = \sum_{\tau} P(\tau; \theta) R(\tau) \nabla_{\theta} \log P(\tau; \theta)$$

- Approximate with empirical estimate for m sample trajectories under policy π_{θ} :

$$\nabla_{\theta} V(\theta) \approx \frac{1}{m} \sum_{i=1}^m R(\tau^{(i)}) \nabla_{\theta} \log P(\tau^{(i)}; \theta)$$

Decomposing the Trajectories Into States and Actions

- Approximate with empirical estimate for m sample trajectories under policy π_θ :

$$\nabla_\theta V(\theta) \approx \frac{1}{m} \sum_{i=1}^m R(\tau^{(i)}) \nabla_\theta \log P(\tau^{(i)}; \theta)$$

$$\begin{aligned} \nabla_\theta \log P(\tau^{(i)}; \theta) &= \nabla_\theta \log \left[\mu(s_0) \prod_{t=0}^{T-1} \pi_\theta(a_t | s_t) P(s_{t+1} | s_t, a_t) \right] \\ &= \nabla_\theta \left[\log \mu(s_0) + \sum_{t=0}^{T-1} \log \pi_\theta(a_t | s_t) + \log P(s_{t+1} | s_t, a_t) \right] \\ &= \sum_{t=0}^{T-1} \nabla_\theta \log \pi_\theta(a_t | s_t) \end{aligned}$$

⇒ No dynamics model required!