Policy Gradient

$$\max_{\theta} \mathbb{E}\left[\sum_{t=0}^{T-1} \gamma^t r(s_t, a_t) \,\middle|\, \pi_\theta \right] \rightarrow \text{objective fcn } J(\theta)$$

$$g = \nabla_\theta \mathbb{E}\left[\sum_{t=0}^{T-1} \gamma^t r(s_t, a_t) \,\middle|\, \pi_\theta \right]$$

estimate this gradient.

* Likelihood Ratio Policy

Define state-action trajectory, τ, as:

$$\tau = (s_0, a_0, r_0, s_1, a_1, r_1 \cdots s_{T-1}, a_{T-1}, r_{T-1}, s_T)$$

Write the objective function, $J(\theta)$

$$J(\theta) = \mathbb{E}_{\tau \sim \pi_\theta(\tau)}\left[\sum_{t=0}^{T-1} \gamma^t r(s_t, a_t) \,\middle|\, \pi_\theta \right]$$

$$= \sum_{\tau} \underbrace{P(\tau;\theta)}_{\text{prob}} \cdot \underbrace{R(\tau)}_{\text{R.V.}}$$

$P(\tau;\theta)$ is the probability of τ under π_θ:

$$P(\tau;\theta) = \mu(s_0) \prod_{t=0}^{T-1} \left[\pi(a_t | s_t; \theta) P(s_{t+1} | s_t, a_t) \right]$$

and the associated reward, $R(\tau)$, is defined as,

$$R(\tau) = \sum_{t=0}^{T-1} \gamma^t r(s_t, a_t)$$

The goal is to find

$$\nabla_\theta J(\theta) = \nabla_\theta \sum_\tau P(\tau;\theta) R(\tau)$$

$$= \sum_\tau \nabla_\theta P(\tau;\theta) R(\tau)$$

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

$$= \sum_\tau P(\tau;\theta) R(\tau) \underbrace{\frac{\nabla_\theta P(\tau;\theta)}{P(\tau;\theta)}}_{\text{likelihood ratio}}$$

$$\boxed{\nabla_\theta \log P(\tau;\theta) = \frac{\nabla_\theta P(\tau;\theta)}{P(\tau;\theta)}} \quad ①$$

①

$$= \sum_\tau \underbrace{P(\tau;\theta)}_{\text{Prob}} \underbrace{R(\tau) \nabla_\theta \log P(\tau;\theta)}_{\text{RV}}$$

②

$$\boxed{\mathbb{E} \to \text{PMF} \times \text{RV}}$$

$$= \mathbb{E}_\tau \left[R(\tau) \nabla_\theta \log P(\tau;\theta) \right]$$

$$\approx \frac{1}{N} \sum_{i=1}^{N} R(\tau_i) \nabla_\theta \log P(\tau_i;\theta)$$

The last term is called "Monte Carlo Sampling" to compute expectation.

Computing the gradient yields

$$\nabla_\theta J(\theta) \approx \frac{1}{N} \sum_{i=1}^{N} R(\tau_i) \nabla_\theta \log P(\tau_i;\theta)$$

Next page

$$\nabla_\theta \log P(\tau_i ; \theta) = \nabla_\theta \log \Big[\underbrace{\mu(s_0)}_{\text{initial state dist}} \prod_{t=0}^{T-1} \underbrace{\pi_\theta(a_t|s_t)}_{\text{policy}} \underbrace{P(s_{t+1}|s_t, a_t)}_{\text{trans. prob}} \Big]$$

$$= \nabla_\theta \Big[\log \mu(s_0) + \sum_{t=0}^{T-1} \log \pi_\theta(a_t|s_t) + \log P(s_{t+1}|s_t,a_t) \Big]$$

$$= \sum_{t=0}^{T-1} \underbrace{\nabla_\theta \log \pi_\theta(a_t|s_t)}_{\text{No dynamics required}}$$

we need a policy, $\pi_\theta(a_t|s_t)$ to be Stochastic

Summary

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

$$\approx \frac{1}{N} \sum_{i=1}^{N} \boxed{R(\tau_i)} \boxed{\nabla_\theta \log P(\tau_i ; \theta)}$$

$$= \frac{1}{N} \sum_{i=1}^{N} \boxed{\sum_{t=0}^{T-1} r(s_t, a_t)} \boxed{\sum_{t=0}^{T-1} \nabla_\theta \log \pi_\theta(a_t|s_t)}$$

$s_{i,t}, a_{i,t}$

Causality