Actor - Critic   Ch 13 - Sutton & Barto

policy $p(a \mid s; \theta)$

on-policy samples

$\sum_{t=0}^{T} r(s_t, a_t)$

payoff $J(\theta) = E_{\tau \sim p(\tau;\theta)} [r(\tau)]$

$(s_0, a_0, \ldots, s_T, a_T, s_{T+1})$

$= E_{\tau \sim T} [p(\tau;\theta)\, r(\tau)]$ ← importance sampling

$E_{x \sim p(x)} [f(x)] = E_{x \sim q(x)} \left[\frac{p(x)}{q(x)} f(x) \right]$

$\nabla_\theta J(\theta) = \nabla_\theta E_{\tau \sim T} [p(\tau;\theta)\, r(\tau)]$

$= E_{\tau \sim T} [\nabla_\theta p(\tau;\theta)\, r(\tau)]$ ← linearity of E, ∇_θ

$= E_{\tau \sim T} [p(\tau;\theta)\, \nabla_\theta \log p(\tau;\theta)\, r(\tau)]$ ← chain rule

$\nabla \log p = \frac{\nabla p}{p}$

$= E_{\tau \sim p(\tau;\theta)} [\nabla_\theta \log p(\tau;\theta)\, r(\tau)]$

$$\nabla_\theta \log p(\tau;\theta) = \nabla_\theta \log\left(p(s_0)\prod_{t=0}^{T} p(a_t \mid s_t;\theta)\, p(s_{t+1} \mid s_t, a_t)\right)$$

$$= \nabla_\theta \left(\log p(s_0) + \sum_{t=0}^{T} \log p(a_t \mid s_t;\theta) + \sum_{t=0}^{T} \log p(s_{t+1} \mid s_t, a_t)\right)$$

$$= \sum_{t=0}^{T} \nabla_\theta \log p(a_t \mid s_t;\theta)$$

$$\nabla_\theta J(\theta) = \mathbb{E}_{\tau\sim p(\tau;\theta)}\left[\left(\sum_{t=0}^{T} \nabla_\theta \log p(a_t \mid s_t;\theta)\right)\left(\sum_{t=0}^{T} r(s_t, a_t)\right)\right]$$

Causality $\longrightarrow$ $$= \mathbb{E}_{\tau\sim p(\tau;\theta)}\left[\sum_{t=0}^{T}\left(\nabla_\theta \log p(a_t \mid s_t;\theta) \sum_{k=t}^{T} r(s_k, a_k)\right)\right]$$

$$\nabla_\theta J(\theta) = E_{\tau \sim p(\tau;\theta)}\left[\sum_{t=0}^{T} \left(\nabla_\theta \log p(a_t \mid s_t;\theta) \sum_{k=t}^{T} r(s_k, a_k) \right) \right]$$

$$= \sum_{t=0}^{T} E_{\underbrace{(s_0, a_0, \dots, s_T, a_T, s_{T+1})}_{\sim p(\tau;\theta)}}\left[\nabla_\theta \log p(a_t \mid s_t;\theta) \sum_{k=t}^{T} r(s_k, a_k) \right]$$

$$= \sum_{t=0}^{T} E_{\underbrace{(s_0, a_0, \dots, s_t, a_t)}_{\sim p(\tau_{0:t};\theta)}}\left[\nabla_\theta \log p(a_t \mid s_t;\theta) \right]$$

$$\times \underbrace{E_{\underbrace{(s_{t+1}, a_{t+1}, \dots, s_T, a_T, s_{T+1})}_{\sim p(\tau_{t+1:T} \mid \tau_{0:t};\theta)}}\left[\sum_{k=t}^{T} r(s_k, a_k) \right]}_{Q(s_t, a_t)}$$

$$E[\dots \mid \tau_{0:t}] = E[\dots \mid s_t, a_t]$$

Markov property of MDPs

$$E_{(x,y)}\left[f(x)\, g(x,y) \right] = E_x\left[E_y\left[f(x)\, g(x,y) \mid x \right] \right]$$

$$= E_x\left[f(x)\, E_y\left[g(x,y) \mid x \right] \right]$$

$$\nabla_\theta J(\theta) = \sum_{t=0}^{T} E_{\tau_{0:t} \sim p(\tau_{0:t};\theta)} \left[\nabla_\theta \log p(a_t | s_t; \theta) \; Q_\theta^t(s_t, a_t) \right]$$

$$Q_\theta^t(s_t, a_t) = E_{\tau_{t:T}} \left[\sum_{k=t}^{T} r(s_t, a_t) \mid s_t, a_t \right]$$

$$E_x[f(x)] = E_{(x,y)}[f(x)] = E_x[f(x) \, E_y[1 | x]]$$

$$\nabla_\theta J(\theta) = E_{\tau \sim p(\tau;\theta)} \left[\sum_{t=0}^{T} \nabla_\theta \log \underbrace{p(a_t | s_t; \theta)}_{\text{actor}} \; \underbrace{Q_\theta^t}_{\text{critic}}(s_t, a_t) \right]$$

learn actor $p(a|s)$ and critic $Q(s,a)$ together using policy-gradient and Q-learning

this is good because Q is lower variance than $\sum_t r(s_t, a_t)$

Now consider infinite time, average reward case (no discounting).

$$J_{avg}(\theta) = \text{average reward per step}$$

$$= \lim_{T \to \infty} \frac{1}{T} \underbrace{E_{\tau_{0:T} \sim p(\tau_{0:T};\theta)} \left[\sum_{t=0}^{T} r(s_t, a_t) \right]}_{\text{previous payoff}}$$

$$= \sum_{s \in S} \sum_{a \in A} p(s, a; \theta) \, r(s, a)$$

ergodicity

probability of being in (a, s) as $t \to \infty$

invariant distribution of the MDP

$$= \sum_{s \in S} p(s; \theta) \sum_{a \in A} p(a \mid s; \theta) \, r(s, a)$$

Gradient of payoff $J_{avg}(\theta)$ in infinite-time, average-reward:

See B (13.5), p 326:

$$\nabla_\theta J(\theta) = \sum_s p(s;\theta) \sum_a Q_\theta(s,a) \nabla_\theta p(a|s;\theta)$$

$$= \sum_s p(s;\theta) \sum_a Q_\theta(s,a) \nabla_\theta \log p(a|s;\theta)\, p(a|s;\theta)$$

$$= \sum_s \sum_a p(s,a;\theta)\, \nabla_\theta \log p(a|s;\theta)\, Q_\theta(s,a)$$

$$= E_{(s,a)\sim p(s,a;\theta)}\left[\nabla_\theta \log p(a|s;\theta)\, Q_\theta(s,a)\right]$$

either samples over $S \times A$ or samples from on-policy trajectories

$E_\pi[\ldots]$

policy gradient theorem

$$\nabla_\theta J(\theta) = E_{(s,a)\sim p(s,a;\theta)}\left[\nabla_\theta \log p(a|s;\theta)\, Q_\theta(s,a)\right]$$

Actor-critic-Q method

$$\nabla_\theta J(\theta) = E_{(s,a)\sim p(s,a;\theta)}\left[\nabla_\theta \log p(a|s;\underset{\text{parameters of } p}{\theta})\, Q_\theta(s,a)\right]$$

Approximate $Q_\theta(s,a) \approx Q(s,a;\underset{\text{parameters of } Q}{w})$

loop over episodes:

$s \sim p(s)$

$a \sim p(a|s;\theta)$

loop over time:

$$\theta \leftarrow \theta + \underset{\text{learning rate for } p}{\alpha} \overbrace{\underbrace{\nabla_\theta \log p(a|s;\theta)\, Q(s,a;w)}_{\text{policy gradient}}}^{\approx \nabla_\theta J(\theta)}$$

$s' \sim p(s'|s,a)$

$a' \sim p(a'|s';\theta)$ ← on policy

$$\left.\begin{aligned} \delta &\leftarrow r(s,a) + Q(s',a';w) - Q(s,a;w) \\ w &\leftarrow w + \underset{\text{learning rate for } Q}{\beta}\, \delta\, \nabla_w Q(s,a;w) \end{aligned}\right\} \text{Q-learning}$$

$s \leftarrow s',\ a \leftarrow a'$