Reinforcement Learning Note

Chenggang Liu

1 short derivation of policy gradient method

$$J(\pi_\theta) = \mathbb{E}_{\tau\sim\pi_\theta}[R(\tau)]$$

$$\begin{aligned} eq:1\nabla_\theta J(\pi_\theta) &= \nabla_\theta \int_\tau p(\tau)R(\tau) && (1)\\ &= \int_\tau \nabla_\theta p(\tau) R(\tau) && (2)\\ &= \int_\tau p(\tau)\nabla_\theta \ln p(\tau) R(\tau) && (3)\\ &= \mathbb{E}_{\tau\sim\pi_\theta}\nabla_\theta \ln p(\tau)R(\tau) && (4)\end{aligned}$$

$$\ln p(\tau) = \ln p_0(s_0) + \sum_{t=0}^{T}[\ln P(s_{t+1}|s_t,a_t) + \ln \pi_\theta(a_t|s_t)] \quad (5)$$

the first and second term on the RHS is independent of π_θ, therefore,

$$\nabla_\theta \ln p(\tau) = \cancel{\nabla_\theta \ln p_0(s_0)} + \sum_{t=0}^{T}[\cancel{\nabla_\theta \ln P(s_{t+1}|s_t,a_t)} + \nabla_\theta \ln \pi_\theta(a_t|s_t)]$$

therefore, the policy gradient is

$$\nabla_\theta J(\pi_\theta) = E_{\tau\sim\pi_\theta}\textstyle\sum_{t=0}^{T}[\nabla_\theta \ln \pi_\theta(a_t|s_t)R(\tau)]$$

which can be improved by "causality trick":

$$\nabla_\theta J(\pi_\theta) = E_{\tau\sim\pi_\theta}\textstyle\sum_{t=0}^{T}[\nabla_\theta \ln \pi_\theta(a_t|s_t)r^t Q(s_t,a_t)]$$