π

$$y_t = r_t + \gamma \underbrace{V^{\pi}(s')}_{\theta^T \phi(s')}$$

$$\theta^* = \min_{\theta} L = \mathbb{E}\left[\left(V^{\pi}_{\theta} - y_t\right)^2\right]$$

? States $\rightarrow$ stationary state distribution after running policy π potentially for a long time

$$\lim_{t \to \infty} \theta_t \overset{?}{=} \theta^*$$

1997

$$J(\theta) = \mathbb{E}_{\tau \sim P_\theta(\tau)}\left[r(\tau) \right]$$

return یک اپیزود تا زمان τ

$P_\theta(\tau) \neq 0$

$$= \int r(\tau)\, P_\theta(\tau)\, d\tau$$

احتمال یک اپیزود

$$\nabla J(\theta) = \int \nabla_\theta P_\theta(\tau)\, r(\tau)\, d\tau$$

Under certain Continuity Cond.

$$= \int \frac{\nabla_\theta P_\theta(\tau)\, r(\tau)\, P_\theta(\tau)}{P_\theta(\tau)}\, d\tau$$

$\nabla_\theta \log P_\theta(\tau)$

$$= \int \underbrace{\nabla_\theta \log P_\theta(\tau) \; r(\tau)}_{g(\tau)} P_\theta(\tau)\, d\tau$$

$$= \mathbb{E}_{\tau \sim P_\theta(\tau)}\big(g(\tau)\big) \approx \frac{1}{N}\sum_{i=1}^{N} \nabla_\theta \log P_\theta(\tau_i)\, r(\tau_i)$$

$$P_\theta(\tau_i) = p(s_0^{(i)})\, p(s_1^{(i)} \mid s_0^{(i)}, a_0^{(i)})\, \pi_\theta(a_0^{(i)} \mid s_0^{(i)}) \cdots$$

$$\nabla_\theta \log P_\theta(\tau) = \nabla_\theta \log P(s_0) + \nabla_\theta \log P(s_1 \mid s_0, a_0) + \nabla_\theta \log \pi_\theta(a_0 \mid s_0) + \cdots$$

از شبکه میاد

$$\nabla_\theta J(\theta) \approx \frac{1}{N}\sum_{i=1}^{N}\left(\sum_{j=1}^{T}\nabla \log \pi\left(a_j^{(i)} \mid s_j^{(i)}\right)\right)\underbrace{\left(\sum_{j=1}^{T} r_j^{(i)}\right)}_{MC}$$

شبیه MLE وزن دار

هر چی return اون episode بیشتر باشه در شبکه بیشتر اثر داره

τ_1

واریانس بالا

τ_2

① بایستی تعداد زیادی نمونه بگیر

② ممکن است ایپیزودی به این صورت به احتمال تولید نمونه خوب رو ببره بالا

مفهوم سعی و خطا ⇐ احتمال return های خوب رو ببره بالا و برعکس احتمال return های بد رو بسیار پایین

τ_N

$$\theta^{(t+1)} \leftarrow \theta^{(t)} + \alpha \nabla_\theta J(\theta)$$