- Objective function

$$J(\theta) = \sum_s d^{\pi_\theta}(s) \cdot V_\pi(s) = \sum_s d^{\pi_\theta}(s) \cdot \sum_a \pi_\theta(a|s) \cdot q_\pi$$

$$\hookrightarrow d^{\pi_\theta}(s) = \lim_{t\to\infty} P[S_t = s \mid S_0, \pi_\theta]$$

- Policy Gradient Theorem

$$\nabla_\theta J(\theta) = \nabla_\theta \sum_s d^{\pi_\theta}(s) V_\pi(s) = \nabla_\theta \sum_s d^{\pi_\theta}(s) \cdot \sum_a \pi_\theta(a|s) \cdot q_{\pi_\theta}$$

$$\propto \sum_s d^{\pi_\theta}(s) \cdot \sum_a q_{\pi_\theta} \nabla_\theta \pi_\theta(a|s) \quad (\because d^{\pi_\theta}, q_{\pi_\theta} \to \theta\text{에 무관})$$

- Stationary state distribution

$$d^\pi(s) := \lim_{k\to\infty} P[S_k = s]$$

→ π를 따라 k step 움직일 때 $S_k = s$일 확률...

- Visitation probability

$$n(s) := \sum_{k=0}^{\infty} \rho^\pi(s_0 \to s, k)$$

→ π를 따라 갔을 때 State s를 방문할 확률

- Visitation probability → stationary state distribution

$$d^\pi(s) = \frac{n(s)}{\sum_s n(s)}$$

[Proof]

① $\nabla_\theta V_{\pi_\theta} = \nabla_\theta \sum_a \pi_\theta(a|s) q_{\pi_\theta}$

$= \sum_a (\nabla_\theta \pi_\theta(a|s) q_{\pi_\theta} + \pi_\theta(a|s) \cdot \nabla_\theta q_{\pi_\theta})$

$= \sum_a (\nabla_\theta \pi_\theta(a|s) q_{\pi_\theta} + \pi_\theta(a|s) \cdot \nabla_\theta [R_s^a + \sum_{s'} P_{ss'}^a \cdot V_{\pi_\theta}(s')])$

$= \sum_a (\nabla_\theta \pi_\theta(a|s) q_{\pi_\theta} + \pi_\theta(a|s) \cdot \nabla_\theta [\sum_{s'} P_{ss'}^a \cdot V_{\pi_\theta}(s')])$ ← R_s^a는 θ에 무관

$= \sum_a (\nabla_\theta \pi_\theta(a|s) q_{\pi_\theta} + \pi_\theta(a|s) \cdot \sum_{s'} P_{ss'}^a \cdot \nabla_\theta V_{\pi_\theta}(s'))$

↓ $q_\pi = R_s^a + \gamma \cdot \sum_{s'} P_{ss'}^a \cdot V_\pi(s')$ ⇒ $\gamma = 1$로... 실제 구현에서는 사용!

… $\nabla_\theta V_{\pi_\theta} = \sum_a (\nabla_\theta \pi_\theta(a|s) q_{\pi_\theta} + \pi_\theta(a|s) \cdot \sum_{s'} P_{ss'}^a \cdot \nabla_\theta V_{\pi_\theta}(s'))$

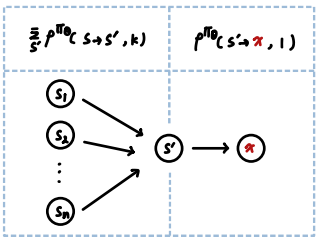
→ $\nabla_\theta V_{\pi_\theta}$가 Recursive하게 풀어질 수 있다.

② π_θ에 따라 state s에서 k번 움직여 state x에 도착하는 전이확률 : $\rho^\pi(s \to x, k)$

$s \longrightarrow s' \longrightarrow s'' \longrightarrow \cdots$

$a \sim \pi_\theta(\cdot|s)$ $a \sim \pi_\theta(\cdot|s')$ $a \sim \pi_\theta(\cdot|s'')$

- $k=0$: $\rho^{\pi_\theta}(s \to s, 0) = 1$
- $k=1$: $\rho^{\pi_\theta}(s \to s', 1) = \sum_a \pi_\theta(a|s) \cdot P_{ss'}^a$
- $k=k+1$: $\rho^{\pi_\theta}(s \to x, k+1) = \sum_{s'} \rho^{\pi_\theta}(s \to s', k) \times \rho^{\pi_\theta}(s' \to x, 1)$ ⟶ (All s'...)



③ $\nabla_\theta V_{\pi_\theta}(s) = \sum_a \left(\nabla_\theta \pi_\theta(a|s) \cdot q_{\pi_\theta} + \pi_\theta(a|s) \sum_{s'} p^a_{ss'} \cdot \nabla_\theta V_{\pi_\theta}(s') \right)$

↳ $\phi(s)$로 간략히...

$= \phi(s) + \sum_{s'} \sum_a \pi_\theta(a|s) p^a_{ss'} \cdot \nabla_\theta V_{\pi_\theta}(s')$

$= \phi(s) + \sum_{s'} p^{\pi_\theta}(s \to s', k=1) \cdot \nabla_\theta V_{\pi_\theta}(s')$ $\quad \left[\because \sum_a \pi_\theta(a|s) p^a_{ss'} = p^{\pi_\theta}(s \to s', k=1) \right]$

$= \phi(s) + \sum_{s'} p^{\pi_\theta}(s \to s', k=1) \cdot \left[\phi(s') + \sum_{s''} p^{\pi_\theta}(s' \to s'', k=1) \cdot \nabla_\theta V_{\pi_\theta}(s'') \right]$

$= \phi(s) + \sum_{s'} p^{\pi_\theta}(s \to s', k=1) \cdot \phi(s') + \sum_{s'} p^{\pi_\theta}(s \to s', k=1) \cdot \sum_{s''} p^{\pi_\theta}(s' \to s'', k=1) \cdot \nabla_\theta V_{\pi_\theta}(s'')$

$= \phi(s) + \sum_{s'} p^{\pi_\theta}(s \to s', k=1) \cdot \phi(s') + \sum_{s''} p^{\pi_\theta}(s \to s'', k=2) \cdot \nabla_\theta V_{\pi_\theta}(s'')$ $\quad \left[\because s \to s' \to s'' \right]$

$\vdots$

$= \sum_{x} \sum_{k=0}^{\infty} p^{\pi_\theta}(s \to x, k) \phi(x)$

$J(\theta)$ → ① episodic env : $J_1(\theta) = V_{\pi_\theta}(s_0)$ ← start value

→ ② Continuous ∞ : $J_{av}(\theta) = \sum_s d^\pi(s) \cdot V_\pi(s)$

↙ Sutton barto P325

④ $\nabla_\theta J(\theta) = \nabla_\theta V_{\pi_\theta}(s_0)$ ← random state!!

$= \sum_s \sum_{k=0}^{\infty} p^\pi(s_0 \to s, k) \phi(s)$

$= \sum_s n(s) \phi(s)$

$= \sum_s n(s) \cdot \sum_s \left(\frac{n(s)}{\sum_s n(s)} \right) \cdot \phi(s)$

$= \sum_s n(s) \cdot \sum_s d^\pi(s) \cdot \phi(s)$

$\sum_s n(s)$: 상수

$\propto \sum_s d^\pi(s) \cdot \phi(s)$

$= \sum_s d^\pi(s) \cdot \sum_a \nabla_\theta \pi(a|s) \cdot q_\pi$

$\longrightarrow \therefore \nabla_\theta J(\theta) \propto \sum_s d^\pi(s) \cdot \sum_a \nabla_\theta \pi_\theta(a|s) \cdot q_{\pi_\theta}$

② $\nabla_\theta J_{av}(\theta) = \sum_s \nabla_\theta d^\pi(s) \cdot V_\pi(s) + \sum_s d^\pi(s) \cdot \nabla_\theta V_\pi(s)$

$= \sum_s d^\pi(s) \cdot \nabla_\theta V_\pi(s)$

$= \sum_s d^\pi(s) \cdot \nabla_\theta \left(\sum_a \pi_\theta \cdot q_\pi \right)$

$= \sum_s d^\pi(s) \cdot \sum_a \nabla_\theta \pi_\theta \cdot q_\pi$

① ② 모두 동일 결과..

⑤ $\nabla_\theta J(\theta) \propto \sum_s d^\pi(s) \cdot \sum_a \nabla_\theta \pi_\theta(a|s) \cdot q_\pi$

$= \sum_s d^\pi(s) \cdot \sum_a \frac{\nabla_\theta \pi_\theta(a|s)}{\pi_\theta(a|s)} \pi_\theta(a|s) \cdot q_\pi$

$= \sum_s d^\pi(s) \cdot \sum_a \nabla_\theta \log \pi_\theta(a|s) \pi_\theta(a|s) \cdot q_\pi$

$= E \left[\nabla_\theta \log \pi_\theta(a|s) \cdot q_\pi \right]$ $\longrightarrow$ $\nabla_\theta \log \pi_\theta(a|s) \cdot q_\pi$ 이 d^π와 π_θ 분포를 따를 때 평균값

$(s \sim d^\pi, a \sim \pi_\theta)$