

عنین سوم

400104964

عنین سوم

۱) (الع) ۱) آنچه ترتیب ماتریس است احتمالات باید درسته و معنی داشت استیت هستم را از

ناتمام نموده هم سراسم به استیت های بعدی تغذیه بگیری این احتمالات ایجاد نمایم

۲) ماتریس انتقال احتمالات باید درسته و معنی داشتی برای این احتمالات ایجاد نمایم

۲) در ماتریس RL ماتریس موافقین عبارت روی در را جسته اوریم.

نحوی دلیل اینه اید را صفر مایه θ را بخواهیم اماده بگیرد در طبقه این θ به صرفت متناسب ظاهر شده است درسته

طریق ماتریس انتقال دهنده این استه را بخواهیم θ را از ترجیل بری اید برای انتقال دهنده سراسم در این را

$$\epsilon \sim p(\epsilon) \quad (S_t, a_t) = g_\theta(\epsilon, S_t, a_t)$$

$$\Rightarrow E_{\tau \sim P_\theta} \left[\sum_t r^t(S_t, a_t) \right] = E_{p(\epsilon)} \left[\sum_t r^t(S_t, a_t) \right]$$

درسته θ را برای این درسته از reparametrization trick استفاده کنید و در این طریق سراسم به صرفت متناسب

$$\nabla_\theta E_{\tau \sim P_\theta} \left[\sum_t r^t(S_t, a_t) \right] = \nabla_\theta E_{p(\epsilon)} \left[\sum_t r^t(g_\theta(\epsilon, S_t, a_t)) \right]$$

$$= E_{p(\epsilon)} \left[\sum_t r^t \nabla_\theta g_\theta(\epsilon, S_t, a_t) \right]$$

$$E[\hat{\nabla}_{\theta} j(\theta)] = E\left[\frac{1}{N} \sum_{i=1}^N \left(\sum_{t=0}^T \nabla_{\theta} \log \pi_{\theta}(a_{i,t} | s_{i,t}) \right) \sum_{t=0}^T r_t^+ r(s_{i,t}, a_{i,t}) \right] \quad (2)$$

$$= \frac{1}{N} \sum_{i=1}^N E\left[\sum_{t=0}^T \nabla_{\theta} \log \pi_{\theta}(a_{i,t} | s_{i,t}) \sum_{t=0}^T r_t^+ r(s_{i,t}, a_{i,t})\right] =$$

$$E\left[\left(\sum_{t=0}^T \nabla_{\theta} \log \pi_{\theta}(a_t, s_t)\right) \sum_{t=0}^T r_t^+ r(s_t, a_t)\right] \rightarrow \text{policy gradient} \quad \text{عوْجَهُ}$$

$$\nabla_{\theta} j(\theta) = \nabla_{\theta} E\left[\sum_{t=0}^T r_t^+ r(s_t, a_t)\right] = \nabla_{\theta} \int_{\mathcal{T}} \left(\sum_{t=0}^T r_t^+ r(s_t, a_t) \right) P_{\theta}(\mathcal{T}) d\mathcal{T}$$

$$\int_{\mathcal{T}} \nabla_{\theta} P_{\theta}(\mathcal{T}) \Gamma(\mathcal{T}) d\mathcal{T} = \int_{\mathcal{T}} \frac{\nabla_{\theta} P_{\theta}(\mathcal{T})}{P_{\theta}(\mathcal{T})} P_{\theta}(\mathcal{T}) \Gamma(\mathcal{T}) d\mathcal{T} = \int_{\mathcal{T}} \nabla_{\theta} \log P_{\theta}(\mathcal{T}) \Gamma(\mathcal{T}) P_{\theta}(\mathcal{T}) d\mathcal{T}$$

$$= \int_{\mathcal{T}} \nabla_{\theta} \log (P_{\theta}(s_0) \pi(s_0, a_0) p(s_1 | s_0, a_0) \dots) P_{\theta}(\mathcal{T}) \Gamma(\mathcal{T}) d\mathcal{T} = \int_{\mathcal{T}} \sum_{t=0}^T \nabla_{\theta} \log \pi_{\theta}(s_t, a_t) \Gamma(\mathcal{T}) P_{\theta}(\mathcal{T}) d\mathcal{T}$$

$$= E\left[\sum_{t=0}^T \nabla_{\theta} \log \pi_{\theta}(s_t, a_t) \sum_{t=0}^T r_t^+ r(s_t, a_t)\right]$$

در این فرض در یک تجربه واحد، حسین می‌تواند نتیجه ای را که می‌خواهد بگیرد.

تا است، این دو روش ممکن است.

$$E_{\mathcal{T}}\left[\sum_{t=0}^{\infty} \nabla_{\theta} \log \pi_{\theta}(s_t, a_t) \sum_{t=0}^{\infty} r_t^+ r(s_t, a_t)\right] = E_{\mathcal{T}}\left[\sum_{t=0}^{\infty} \nabla_{\theta} \log \pi_{\theta}(s_t, a_t) \sum_{t'=0}^{t-1} r_t^+ r(s_t, a_t) \right]$$

$$+ E_{\mathcal{T}}\left[\sum_{t=0}^{\infty} r_t^+ \nabla_{\theta} \log \pi_{\theta}(s_t, a_t) \sum_{t'=t}^{\infty} r_{t'}^+ r(s_{t'}, a_{t'})\right] \quad \text{***}$$

دلیل خاصیت طالسان می‌دانم که این دو روش ممکن است تفسیر شده باشند.

$$E_{\mathcal{T}}\left[\sum_{t=0}^{\infty} \nabla_{\theta} \log \pi_{\theta}(s_t, a_t) \sum_{t'=0}^{t-1} r_t^+ r(s_t, a_t)\right] = \sum_{t=0}^{\infty} E_{\mathcal{T}}\left[\nabla_{\theta} \log \pi_{\theta}(s_t, a_t) \sum_{t'=0}^{t-1} r_t^+ r(s_t, a_t)\right]$$

$$= \sum_{t=0}^{\infty} E_{\mathcal{T}}\left[\nabla_{\theta} \log \pi_{\theta}(s_t, a_t)\right] E_{\mathcal{T}}\left[\sum_{t'=0}^{t-1} r_t^+ r(s_t, a_t)\right]$$

$$\star \star \star E_{\mathcal{T}}\left[\nabla_{\theta} \log \pi_{\theta}(s_t, a_t)\right] = E_{\mathcal{T}}\left[\nabla_{\theta} \log P_{\theta}(\mathcal{T})\right] = \int_{\mathcal{T}} \frac{\nabla_{\theta} P_{\theta}(\mathcal{T})}{P_{\theta}(\mathcal{T})} P_{\theta}(\mathcal{T}) d\mathcal{T} = \nabla \int_{\mathcal{T}} P_{\theta}(\mathcal{T}) d\mathcal{T} = 0$$

$$\Rightarrow \sum_{t=0}^{\infty} E_{\mathcal{T}}\left[\nabla_{\theta} \log \pi_{\theta}(s_t, a_t)\right] E_{\mathcal{T}}\left[\sum_{t'=0}^{t-1} r_t^+ r(s_t, a_t)\right] = 0$$

$$\star \star \star \Rightarrow E_{\mathcal{T}}\left[\sum_{t=0}^{\infty} \nabla_{\theta} \log \pi_{\theta}(s_t, a_t) \sum_{t=0}^{\infty} r_t^+ r(s_t, a_t)\right] = \sum_{t=0}^{\infty} r_t^+ E\left[\nabla_{\theta} \log \pi_{\theta}(s_t, a_t) \hat{Q}(s_t, a_t)\right]$$

$$= \sum_{t=0}^{\infty} E_{s_0, a_0, \dots, s_{t-1}}\left[E_{s_t, a_t}\left[\underbrace{r_t^+ \nabla_{\theta} \log \pi_{\theta}(s_t, a_t) \hat{Q}(s_t, a_t)}_{\text{کار ایجاد}}\right]\right]$$

$$\begin{aligned}
 &= \sum_{t=0}^{\infty} E_{s_t, a_t} [r_t^+ \nabla_{\theta} \log \pi_{\theta}(s_t, a_t) \hat{Q}(s_t, a_t)] \\
 &= \sum_{t=0}^{\infty} \int_S \int_a r_t^+ \nabla_{\theta} \log \pi_{\theta}(s, a) \hat{Q}(s, a) p(s_t=s) \pi_{\theta}(s, a) da ds \\
 &= \int_S \sum_{t=0}^{\infty} r_t^+ p(s_t=s) \underbrace{\int_a \nabla_{\theta} \log \pi_{\theta}(s, a) \hat{Q}(s, a) \pi_{\theta}(s, a) da}_{\text{target}} ds \\
 &= \int_S E_{\pi} [\nabla_{\theta} \log \pi_{\theta}(s, a) \hat{Q}(s, a)] f_{\pi}(s) ds = E_{\pi, f_{\pi}} [\nabla_{\theta} \log \pi_{\theta}(s, a) \hat{Q}(s, a)]
 \end{aligned}$$

مقدمة في علم البرمجيات

$$g = \nabla_\theta \log \pi_\theta(s, a) (\hat{Q}(s, a) - b(s)) \quad (7)$$

$$Var(g) = E_{P_{\pi}, \pi} [g^T g] - E_{P_{\pi}, \pi} [g]^T E_{P_{\pi}, \pi} [g] = E_{P_{\pi}, \pi} [\nabla_{\theta} \log \pi_{\theta}(s, a)^T \nabla_{\theta} \log \pi_{\theta}(s, a) \hat{Q}(s, a)^2]$$

$$+ E_{\pi_\theta} [\nabla_\theta \log \pi_\theta(s, a)^T \nabla_\theta \log \pi_\theta(s, a)] b(s)^T - 2 E_{\pi_\theta} [\nabla_\theta \log \pi_\theta(s, a)^T \nabla_\theta \log \pi_\theta(s, a) \hat{Q}(s, a)] b(s)$$

$$-E_{\rho_{\pi, \pi}}[g]^T E_{\rho_{\pi, \pi}}[g] \rightarrow, \quad b(s) \text{ is decreasing.}$$

$$\frac{d \text{Var}(q)}{d b(s)} = 2 E_{P_\pi, \pi} [\nabla_\theta \log \pi_\theta(s, a)^\top \nabla_\theta \log \pi_\theta(s, a)] b(s) - 2 E_{P_\pi, \pi} [\nabla_\theta \log \pi_\theta(s, a)^\top \nabla_\theta \log \pi_\theta(s, a)]$$

$$\hat{Q}(s, a) = \mathbb{E}[\nabla_{\theta} \log \pi_{\theta}(s, a)^T \nabla_{\theta} \log \pi_{\theta}(s, a) \hat{Q}(s, a)]$$

(فرص مرسم الـ $\hat{Q}(s,a)$, $\nabla_{\theta} \log \pi_{\theta}(s,a)^T \nabla_{\theta} \log \pi_{\theta}(s,a)$ لابد من دفعه)

السلطنة صاحب الامارة من مملكة مملكة صاحب الامارة من مملكة نسيمة

$$b^*(s) = \frac{E[\nabla_\theta \log \pi_\theta(s, a)^T \nabla_\theta \log \pi_\theta(s, a) Q(s, a)]}{E[\nabla_\theta \log \pi_\theta(s, a)^T \nabla_\theta \log \pi_\theta(s, a)]} = E[\nabla_\theta \log \pi_\theta(s, a)^T \nabla_\theta \log \pi_\theta(s, a)] E[Q(s, a)]$$

$$= E_{P_1, \pi} [Q(S, a)] = V_\pi(S)$$

(الف) مقصود در توزيع احتمالات بivariate X باشد در اینجا μ, ν

$$\|\mu - \nu\|_{TV} = \inf \{ p(X \neq y) \mid (X, y) \text{ is a coupling of } \mu, \nu \}$$

$\forall A \subseteq X : \mu(A) - \nu(A) = p(X \in A) - p(Y \in A)$ ایجاب: میان هر دو (X, Y) coupling

$$\leq p(X \in A, Y \notin A) \leq p(X \neq Y)$$

$\forall A \subseteq X : \mu(A) - \nu(A) \leq \inf \{ p(X \neq y) \mid X, y \text{ is a coupling of } \mu, \nu \}$

$$\|\mu, \nu\|_{TV} = \max \{ \mu(A) - \nu(A) \} \leq \inf \{ p(X \neq y) \mid X, y \text{ is a coupling of } \mu, \nu \}$$

$$p = \sum_n f(n) \quad f(n) = \min \{ \mu(n), \nu(n) \} \quad \text{این تعاریف ایجاب است:}$$

$$\mu = \sum_n f(n) = \sum_{n: \mu(n) < \nu(n)} \mu(n) + \sum_{n: \nu(n) \leq \mu(n)} \nu(n) = \underbrace{\sum_{n: \mu(n) < \nu(n)} \mu(n)}_{1} + \sum_{n: \nu(n) \leq \mu(n)} \nu(n)$$

$$+ \underbrace{\sum_{n: \mu(n) \geq \nu(n)} \mu(n)}_{\mu(M)} - \underbrace{\sum_{n: \mu(n) \geq \nu(n)} \nu(n)}_{\nu(M)} = 1 - \sum_{n: \mu(n) \geq \nu(n)} \mu(n) - \nu(n)$$

$$\|\mu - \nu\|_{TV} = \sum_{n: \mu(n) \geq \nu(n)} \mu(n) - \nu(n)$$

$$M = \{ n \mid \mu(n) \geq \nu(n) \}$$

$$\forall A \subseteq X : \mu(n) - \nu(n) \leq \mu(A \cap M) - \nu(A \cap M)$$

$$\mu(n) - \nu(n) \leq 0 \quad n \in A \cap \bar{M} \rightarrow \mu(A \cap \bar{M}) - \nu(A \cap \bar{M}) \leq 0$$

$$\leq \mu(M) - \nu(M)$$

$$\mu(\bar{A} \cap M) - \nu(M \cap \bar{A}) \geq 0$$

این در این تعاریف ایجاب است: $0 \leq \mu(M) - \nu(M)$ زیرا $\mu(M) - \nu(A) = \mu(A) - \nu(M)$ ایجاب است که $A = M$ یا $A = \bar{M}$

$$\|\mu - \nu\|_{TV} = \mu(M) - \nu(M) = \sum_{n \in M} \mu(n) - \nu(n) \quad \text{درباره توزیع خود سوال نمایش داشت: } |\mu(A) - \nu(A)| = \mu(M) - \nu(M)$$

$$p = 1 - \|\mu - \nu\|_{TV}$$

حق در توزیع اینهاست:

خد ع را در تابع $f(n)$ این شکل نمایش دهیم: با احتمال p , Z را از توزیع احتمال μ و ν میگیریم و

$$f(n)$$

$$X = Y = Z$$

حل با اصل $\hat{d}_1(n), \hat{d}_2(n), \hat{d}_3(n)$ را به ترتیب از توزیع های زیر عده مرسیم.

$$\hat{d}_2(n) = \begin{cases} \frac{P(n) - V(n)}{\|P - V\|_{TV}} & P(n) > V(n) \\ 0 & \text{o.w.} \end{cases}$$

$$\hat{d}_3(n) = \begin{cases} \frac{V(n) - P(n)}{\|P - V\|_{TV}} & V(n) > P(n) \\ 0 & \text{o.w.} \end{cases}$$

$$X \sim p \hat{d}_1(n) + (1-p) \hat{d}_2(n) = p(n)$$

$$Y \sim p \hat{d}_1(n) + (1-p) \hat{d}_3(n) = V(n)$$

$$p(X \neq Y) = 1 - p(X = Y) = \|P - V\|_{TV}$$

درسته (ع) X, Y طبقی ایسته در مارکو

درسته نمایی بسیار سُلْطَنَه داشته ایست

دو توزیع $P_\theta(\cdot, S_+)$, $P_\theta(\cdot, S_-)$ را در نظر مرسیم بنابراین آنها ای وجود دارند که رابطه زیر را داشته باشند

$$X \sim P_\theta(\cdot, S_+) \quad Y \sim P_\theta(\cdot, S_-) \quad p(X \neq Y) = \|P_\theta(\cdot, S_+) - P_\theta(\cdot, S_-)\|_{TV} \leq \epsilon$$

$$\|P_\theta(\cdot, S_+) - P_\theta(\cdot, S_-)\|_{TV} = \alpha$$

$$P_\theta'(S_+) = ((1-\alpha)^+)^\top P_\theta(S_+) - (1-(1-\alpha)^+) P_{\text{mistake}}(S_+)$$

$$|P_\theta'(S_+) - P_\theta(S_+)| = (1-(1-\alpha)^+) |P_{\text{mistake}}(S_+) - P_\theta(S_+)| \stackrel{\alpha \leq \epsilon}{\leq} (1-(1-\epsilon)^+) |P_{\text{mistake}}(S_+) - P_\theta(S_+)|$$

$$\leq (1-(1-\epsilon)^+) (|P_{\text{mistake}}(S_+)| - |P_\theta(S_+)|) \leq (1-(1-\epsilon)^+) \times 2$$

$$\epsilon \geq 1 - (1-\epsilon)^+$$

حسن بخت شیر داریم

$\Rightarrow |P_\theta'(S_+) - P_\theta(S_+)| \leq 2\epsilon$ \rightarrow markov chain
برای حل این مسئله از ترتیب
و انتقالاتی در میان ایستگاه ها
و انتقالاتی در میان ایستگاه ها

(ب) منظمه مرسیم $|P_\theta'(\alpha, S_+) - P_\theta(\alpha, S_+)| \leq \epsilon$ \rightarrow مارکو

دانسته باشیم. باید دلخواه از ایستگاه S_+ برخوردی $P(S_+)$ در توزیع مرسیم

$$E_{P_\theta'}[\hat{d}(S_+)] = \sum_{S_+} P_\theta'(S_+) \hat{d}(S_+) = \sum_{S_+} (P_\theta'(S_+) - P_\theta(S_+) + P_\theta(S_+)) \hat{d}(S_+)$$

$$= \sum_{S_+} P_\theta(S_+) \hat{d}(S_+) - \sum_{S_+} (P_\theta(S_+) - P_\theta'(S_+)) \hat{d}(S_+) \geq E[\hat{d}(S_+)] - |P_\theta(S_+) - P_\theta'(S_+)| \max \hat{d}(S_+)$$

$$\geq E[\hat{d}(S_+)] - 2\epsilon + \max \hat{d}(S_+)$$

$$\text{حل متراس} \quad \text{لما} \quad E_{a_t \sim \pi(S_t, a_t)} \left[\frac{\pi_{\theta'}(S_t, a_t)}{\pi_{\theta}(S_t, a_t)} \gamma^t A^{\pi_{\theta}}(S_t, a_t) \right] \rightarrow \hat{A}(S_t)$$

$$\sum_{+} E_{P_{\theta}(S_t)} E_{a_t \sim \pi_{\theta}(S_t, a_t)} \left[\frac{\pi_{\theta'}(S_t, a_t)}{\pi_{\theta}(S_t, a_t)} \gamma^t A^{\pi_{\theta}}(S_t, a_t) \right]$$

$$\Rightarrow \sum_{+} E_{P_{\theta}(S_t)} E_{a_t \sim \pi_{\theta}(S_t, a_t)} \left[\frac{\pi_{\theta'}(S_t, a_t)}{\pi_{\theta}(S_t, a_t)} \gamma^t A^{\pi_{\theta}}(S_t, a_t) \right] - \underbrace{\sum_{t=0}^{\infty} 2\varepsilon t c}_{\Delta}$$

$$\Delta = \frac{T(T+1)}{2} \times 2 \times \varepsilon \times O(\Gamma_{\max}) = O(\varepsilon T^2 \Gamma_{\max})$$

$$c \leq \gamma^t \Gamma_{\max} \quad \Delta = \sum_{t=0}^{\infty} 2\varepsilon t c > 2\varepsilon \sum_{t=0}^{\infty} t \gamma^t \Gamma_{\max}$$

$$= 2\varepsilon \gamma \Gamma_{\max} \sum_{t=0}^{\infty} \frac{d}{dt} \gamma^t = 2\varepsilon \gamma \frac{d}{d\gamma} \sum_{t=0}^{\infty} \gamma^t = \frac{2\varepsilon \gamma \Gamma_{\max}}{(1-\gamma)^2} \quad \Delta = O\left(\frac{\varepsilon \gamma \Gamma_{\max}}{(1-\gamma)^2}\right)$$

(ج) من تراس از نهادی بین استانداری در سعیه طریق:

$$|\pi_{\theta'}(S_t, a_t) - \pi_{\theta}(S_t, a_t)| \leq \sqrt{\frac{1}{2} D_{KL}(\pi_{\theta'}(S_t, a_t) \| \pi_{\theta}(S_t, a_t))}$$

$$\Rightarrow D_{KL}(\pi_{\theta'}(S_t, a_t) \| \pi_{\theta}(S_t, a_t)) \geq 2 |\pi_{\theta'}(S_t, a_t) - \pi_{\theta}(S_t, a_t)| \geq 2\varepsilon^2$$

$$\bar{A}(\theta') = \sum_{+} E_{P_{\theta}} \left[E_{\pi} \left[\frac{\pi_{\theta'}(a_t | s_t)}{\pi_{\theta}(a_t | s_t)} \gamma^t A^{\pi_{\theta}}(s_t, a_t) \right] \right] \quad \text{ابه صورت در بر قریب مرسوم} \quad (\rightarrow)$$

$$\bar{A}(\theta') \approx \bar{A}(\theta) + \nabla_{\theta} \bar{A}(\theta)^T (\theta' - \theta) \quad \text{با توجه به تقریب بخط سری در طریق}$$

در اینجا \bar{A} ابه صورت در بر حساب مرسوم.

$$\nabla_{\theta} \bar{A}(\theta) = \nabla_{\theta} \sum_{+} E_{\substack{s_t \sim P_{\theta} \\ a_t \sim \pi_{\theta}}} \left[E_{\pi} \left[\frac{\pi_{\theta'}(a_t | s_t)}{\pi_{\theta}(a_t | s_t)} \gamma^t A^{\pi_{\theta}}(s_t, a_t) \right] \right] = \sum_{+} E_{P_{\theta}} \left[E_{\pi} \left[\frac{\nabla_{\theta} \pi_{\theta'}(s_t, a_t)}{\pi_{\theta}(s_t, a_t)} \gamma^t A^{\pi_{\theta}}(s_t, a_t) \right] \right]$$

$$= \sum_{+} E_{P_{\theta}} \left[E_{\pi_{\theta}} \left[\frac{\pi_{\theta}(s_t, a_t)}{\pi_{\theta}(s_t, a_t)} \gamma^t \nabla_{\theta} \log \pi_{\theta'}(a_t, s_t) A^{\pi_{\theta}}(s_t, a_t) \right] \right]$$

$$= \sum_{+} E_{P_{\theta}} \left[E_{\pi_{\theta}} \left[\gamma^t \nabla_{\theta} \log \pi_{\theta'}(a_t, s_t) A^{\pi_{\theta}}(s_t, a_t) \right] \right] \xrightarrow{\theta' = \theta} = \nabla_{\theta} J(\theta)$$

$$\Rightarrow \bar{A}(\theta') = \sum_{+} E_{\substack{s_t \sim P_{\theta} \\ a_t \sim \pi_{\theta}}} \left[E_{\pi_{\theta}} [\gamma^t A^{\pi_{\theta}}(s_t, a_t)] + \nabla_{\theta} J(\theta)^T (\theta' - \theta) \right]$$

(4)

$$\begin{aligned}
 \nabla_{\theta} V_{\theta}^{\pi}(s) &= \nabla_{\theta} Q_{\theta}^{\pi}(s, p_{\theta}(s)) = \nabla_{\theta} \left[\Gamma(s, p_{\theta}(s)) + \int_S \gamma p(s'|s, p_{\theta}(s)) V_{\theta}^{\pi}(s') ds' \right] \\
 &= \nabla_{\theta} p_{\theta}(s) \nabla_a \Gamma(s, a) \Big|_{a=p_{\theta}(s)} + \nabla_{\theta} \int_S \gamma p(s'|s, p_{\theta}(s)) V_{\theta}^{\pi}(s') ds \\
 &= \nabla_{\theta} p_{\theta}(s) \nabla_a \Gamma(s, a) \Big|_{a=p_{\theta}(s)} + \int_S \gamma (p(s'|s, p_{\theta}(s)) \nabla_{\theta} V_{\theta}^{\pi}(s') + \nabla_{\theta} p_{\theta}(s) \nabla_a p(s'|s, a) V_{\theta}^{\pi}(s')) ds \\
 &= \nabla_{\theta} p_{\theta}(s) \nabla_a (\Gamma(s, a) + \int_S \gamma p(s'|s, a) V_{\theta}^{\pi}(s') ds') \Big|_{a=p_{\theta}(s)} + \int_S \gamma p(s'|s, p_{\theta}(s)) \nabla_{\theta} V_{\theta}^{\pi}(s') ds' \\
 &= \nabla_{\theta} p_{\theta}(s) \nabla_a Q_{\theta}^{\pi}(s, a) \Big|_{a=p_{\theta}(s)} + \int_S \gamma p(s \rightarrow s', 1, p_{\theta}) \nabla_{\theta} V_{\theta}^{\pi}(s') ds' \\
 &\quad \text{جیب رابطہ ایسے سمجھوں (مرکزی ایسے جیسا کیا جائیں)۔} \\
 &= \nabla_{\theta} p_{\theta}(s) \nabla_a Q_{\theta}^{\pi}(s, a) \Big|_{a=p_{\theta}(s)} + \int_S \gamma p(s \rightarrow s', 1, p_{\theta}) \nabla_{\theta} p_{\theta}(s') \nabla_a Q_{\theta}^{\pi}(s', a) \Big|_{a=p_{\theta}(s')} ds' \\
 &\quad + \int_S \gamma p(s \rightarrow s', 1, p_{\theta}) \int_S \gamma p(s'' \rightarrow s'', 1, p_{\theta}) \nabla_{\theta} V_{\theta}^{\pi}(s'') ds'' ds' = \nabla_{\theta} p_{\theta}(s) \nabla_a Q_{\theta}^{\pi}(s, a) \Big|_{a=p_{\theta}(s)} \\
 &\quad + \int_S \gamma p(s \rightarrow s', 1, p_{\theta}) \nabla_{\theta} p_{\theta}(s') \nabla_a Q_{\theta}^{\pi}(s', a) \Big|_{a=p_{\theta}(s')} ds' + \int_S \gamma^2 p(s \rightarrow s, 2, p_{\theta}) \nabla_{\theta} V_{\theta}^{\pi}(s') ds' \\
 &= \dots = \int_S \sum_{t=0}^{\infty} \gamma^t p(s \rightarrow s', t, p_{\theta}) \nabla_{\theta} p_{\theta}(s') \nabla_a Q_{\theta}^{\pi}(s', a) \Big|_{a=p_{\theta}(s')} ds' \\
 \nabla_{\theta} J(p_{\theta}) &= \nabla_{\theta} \int_S p_{\theta}(s) V_{\theta}^{\pi}(s) ds = \int_S p_{\theta}(s) \nabla_{\theta} V_{\theta}^{\pi}(s) ds \quad \text{سب سے ایسا سمجھو، کہ } S \text{ میں } \\
 &= \int_S \int_S \sum_{t=0}^{\infty} \gamma^t p_{\theta}(s) p(s \rightarrow s', t, p_{\theta}) \nabla_{\theta} p_{\theta}(s') \nabla_a Q_{\theta}^{\pi}(s', a) \Big|_{a=p_{\theta}(s')} ds' ds \\
 &= \int_S p_{\theta}^{\pi}(s) \nabla_{\theta} p_{\theta}(s) \nabla_a Q_{\theta}^{\pi}(s, a) \Big|_{a=p_{\theta}(s)} ds
 \end{aligned}$$