Subject: Date: / /

سوال ①:

الف) با توجه به اینکه فضای state transition ما کاملاً probabilistic هستند، برای اینکه مقدوح پاداش را بخواهیم پیشینه کنیم، ممکن است که اگر در ایتِ S باشیم و یک action مانند a انجام بدهیم، ممکن است دو رویداد متفاوت بگیریم. ($P(s'|s,a)$ وجود دارد). همچنین می‌توان به این اشاره کرد که در یک state باشیم $\pi_\theta(\cdot|s)$ وجود دارد و رندومنس وجود دارد.

ب) $J(\theta) = \frac{1}{E_{\tau\sim p_\theta}} \{ \sum_t \gamma^t r(s_t, a_t) \}$

الگوریتم گرادیان کاهشی، از رابطهٔ بالا گرادیان نسبت به θ می‌گیرد.

با توجه به اینکه trajectory‌ها از یک توزیع p_θ می‌آیند و reward‌ها در حل هیچ نمی‌توان به صورت مستقیم از این عبارت گرادیان گرفت و باید این عبارت را بازکرده و پس گرادیان بگیریم. تا تابعیت θ محاسبه شود و قابل محاسبه شود.

برای رفع این مشکل می‌توان پاداش را به صورتی تعریف کنیم که تابعیتی از θ داشته باشد.

برای این کار می‌توان پارامتر جدیدی ϵ تعریف کنیم و طبق یک تابع مانند $\pi_\theta(s_t, a_t, \epsilon)$ اکشن را انتخاب کنیم و پس می‌توان τ را به صورت تابعی از ϵ و θ نوشت.

$$J(\theta) = \frac{1}{E_{\epsilon\sim\text{noise}}} \{ \sum_t \gamma^t r(s_t, \pi_\theta(s_t, a_t, \epsilon)) \}$$

s.a.m

سوال 2)

الف)

$$\nabla_\theta(J(\theta)) = \nabla_\theta\left\{\mathbb{E}_\tau\left\{\sum_t \gamma^t r(s_t,a_t)\right\}\right\} \rightarrow R_{(\tau)} = \sum_{t=0}^{\infty} \gamma^t r(s_t,a_t)$$

$$\tau = (s_0, a_0, s_1, a_1, \dots)$$

$$\nabla_\theta(J(\theta)) = \nabla_\theta\left\{\int P_\theta(\tau) R(\tau) d\tau\right\} = \int \nabla_\theta P_\theta(\tau) R(\tau) d\tau =$$

$$= \int \nabla_\theta \log P_\theta(\tau)\ P_\theta(\tau) R(\tau) d\tau =$$

$$= \mathbb{E}_\tau\left\{\nabla_\theta \log P_\theta(\tau) R(\tau)\right\} =$$

حال در نظر بگیرید که داریم:

$$\Rightarrow \nabla_\theta(J(\theta)) = \mathbb{E}_\tau\{\dots\}$$

$$P_\theta(\tau) = P(s_0)\, P(s_1|s_0,a_0)\, \pi_\theta(a_0|s_0) \dots \Rightarrow$$

$$\Rightarrow \nabla_\theta \log P_\theta(\tau) = \sum_{t=0}^{\infty} \nabla_\theta \log \pi_\theta(a_t|s_t)$$

پس داریم که: $\nabla_\theta(J_\theta) = \mathbb{E}_\tau\left\{\left(\sum \nabla_\theta \log \pi_\theta(a_t^{(i)}|s_t^{(i)})\right)\left(\sum \gamma^t r(s_t,a_t)\right)\right\}$

حال تخمینگر را در نظر می‌گیریم:

$$\mathbb{E}\left(\nabla_\theta \hat{J}(\theta)\right) = \mathbb{E}\left\{\frac{1}{N}\sum_{i=1}^{N}\left\{\left(\sum_{t=0} \nabla_\theta \log \pi_\theta(a_t|s_t)\right)\left(\sum \gamma^t r(s_t,a_t)\right)\right\}\right\} =$$

$$= \frac{1}{N}\sum_{i=1}^{N}\left\{\mathbb{E}\left\{\left(\sum \nabla_\theta \log \pi_\theta(a_t|s_t)\right)\left(\sum \gamma^t r(s_t,a_t)\right)\right\}\right\} =$$

$$= \frac{1}{N}\sum_{i=1}^{N} \nabla_\theta J_\theta = \nabla_\theta J_\theta \rightarrow \text{unbiased}$$

در حین اثبات فرض کردیم که

1. Trajectory ها کاملاً از هم مجزا هستند
2. در تمامی فضا π_θ هایی که داریم، مشتق پذیر هستند.

$$\rho_\pi(s) = \sum_{t=0}^{\infty} \gamma^t P(s_t = s)$$

$$B = \mathbb{E}_\tau \left\{ \left(\sum_{t=0}^{\infty} \nabla_\theta \log \pi_\theta(a_t|s_t) \right) \left(\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \right) \right\} =$$

$$= \sum_{t=0}^{\infty} \mathbb{E}_\tau \left\{ \nabla_\theta \log \pi_\theta(a_t|s_t) \sum_{t'=0}^{\infty} \gamma^{t'} r(s_{t'}, a_{t'}) \right\} =$$

$$= \sum_{t=0}^{\infty} \left\{ \mathbb{E}_\tau \left\{ \nabla_\theta \log \pi_\theta(a_t|s_t) \sum_{t'=0}^{t-1} \gamma^{t'} r(s_{t'}, a_{t'}) \right\} + \mathbb{E}_\tau \left\{ \nabla_\theta \log \pi_\theta(a_t|s_t) \sum_{t'=t}^{\infty} \gamma^{t'} r(s_{t'}, a_{t'}) \right\} \right\}$$

با توجه به Causality می دانیم که تصمیم که در مرحله t می گیریم، تأثیری در $\{0, \dots, t-1\}$ ها ندارد. پس می توان جمله اول را به این صورت نوشت که:

$$\mathbb{E}_\tau\left\{\nabla_\theta \log \pi_\theta(a_t|s_t) \sum_{t'=0}^{t-1} \gamma^{t'} r(s_{t'}, a_{t'})\right\} =$$

$$= \mathbb{E}_\tau\left\{\nabla_\theta \log \pi_\theta(a_t|s_t)\right\} \mathbb{E}_\tau\left\{\sum_{t'=0}^{t-1} \gamma^{t'} r(s_{t'}, a_{t'})\right\}$$

و همچنین داریم که:

$$\mathbb{E}\left\{\nabla_\theta \log \pi_\theta(a_t|s_t)\right\} = \int \pi_\theta \nabla_\theta \log \pi_\theta \, d\tau = \int \pi_\theta \frac{\nabla_\theta \pi_\theta}{\pi_\theta} d\tau = \nabla_\theta \int \pi_\theta \, d\tau = \nabla_\theta 1 = 0$$

پس می توان g را به این صورت نوشت که:

$$\sum_{t=1}^{\infty} \left\{\mathbb{E}_\tau\left\{\nabla_\theta \log \pi_\theta(a_t|s_t) \sum_{t'=t}^{\infty} \gamma^{t'} r(s_{t'}, a_{t'})\right\}\right\} =$$

$$= \sum_{t=1}^{\infty} \gamma^t \mathbb{E}_\tau\left\{\nabla_\theta \log \pi_\theta(a_t|s_t) \sum_{t'=t}^{\infty} \gamma^{t'-t} r(s_{t'}, a_{t'})\right\} =$$

$$= \sum_{t=1}^{\infty} \gamma^t \mathbb{E}_\tau\left\{\nabla_\theta \log \pi_\theta(a_t|s_t) \hat{Q}(s_t, a_t)\right\} =$$

متغیر به s_t و a_t وابسته است

$$= \sum_{t=1}^{\infty} \gamma^t \mathbb{E}_{s_t, a_t}\left\{\nabla_\theta \log \pi_\theta(a_t|s_t) \hat{Q}(s_t, a_t)\right\} =$$

$$= \sum_{t=1}^{\infty} \gamma^t \int_a \int_s \nabla_\theta \log \pi_\theta(a_t|s_t) \hat{Q}(s_t, a_t) P(s_t = s) \pi_\theta(s, a) \, da \, ds =$$

$$= \sum_{t=1}^{\infty} \int_s \gamma^t P(s_t = s) \int_a \nabla_\theta \log \pi_\theta(a_t|s_t) \hat{Q}(s_t, a_t) \pi_\theta(s, a) \, da \, ds =$$

$$=$$

s.a.m

$$\int_S \sum_{t=1}^{\infty} \gamma^t P(s_t = s) \; \mathop{E}_{\pi} \left\{ \nabla_\theta \log \pi_\theta(a_t/s_t) \hat{Q}(s_t, a_t) \right\} ds =$$

$$= \int_S \rho_\pi(s) \mathop{E}_{\pi} \left\{ \nabla_\theta \log \pi_\theta(a_t/s_t) \hat{Q}(s_t, a_t) \right\} ds =$$

$$= \mathop{E}_{\pi, \rho_\pi} \left\{ \nabla_\theta \log \pi_\theta(a_t/s_t) \hat{Q}(s_t, a_t) \right\}$$

← فرض کنید در نظر گرفتیم،

$$h = \nabla_\theta \log \pi_\theta(a, s) \left(\hat{Q}(s, a) - b(s) \right)$$

$$Var(h) = \mathop{E}_{\rho_\pi, \pi}(h^T h) - \mathop{E}_{\rho_\pi, \pi}(h)^T \mathop{E}(h)$$

$$= \mathop{E}_{\rho_\pi, \pi} \left((\nabla_\theta \log \pi_\theta(a, s))^T \nabla_\theta \log \pi_\theta(a, s) \right) b(s)^2$$

$$- 2 \mathop{E}_{\rho_\pi, \pi} \left((\nabla_\theta \log \pi_\theta(a, s))^T (\nabla_\theta \log \pi_\theta(a, s)) \hat{Q}(s, a) \right) b(s)$$

$$+ \mathop{E}_{\rho_\pi, \pi} \left((\nabla_\theta \log \pi_\theta(a, s))^T (\nabla_\theta \log \pi_\theta(a, s)) \hat{Q}(s, a)^2 \right)$$

$\neq \dots$

با توجه که تخمینگر ما نااریب است نتیجه حاصلات به $b(s)$ وابسته نیستند و در مشتق گرفتن تاثیری ندارند.

s.a.m

$$\frac{\partial(Var(h))}{\partial(b(s))} = 0 \Rightarrow 2b(s)\, \mathbb{E}_{\rho_{\pi_\theta},\pi}\left\{(\nabla_\theta \log\pi_\theta(a,s))^T \nabla_\theta \log\pi_\theta(a,s)\right\} - 2\mathbb{E}_{\rho_{\pi_\theta},\pi}\left\{(\nabla_\theta \log\pi_\theta(a,s))^T \nabla_\theta \log\pi_\theta(a,s)\hat{Q}(s,a)\right\} = 0$$

$$\Rightarrow b(s) = \frac{\mathbb{E}_{\rho_{\pi_\theta},\pi}\left\{(\nabla_\theta \log\pi_\theta(a,s))^T \nabla_\theta \log\pi_\theta(a,s)\hat{Q}(s,a)\right\}}{\mathbb{E}_{\rho_{\pi_\theta},\pi}\left\{(\nabla_\theta \log\pi_\theta(a,s))^T \nabla_\theta \log\pi_\theta(a,s)\right\}}$$

در نظر می‌گیریم با داشتن π_θ در عبارتی که نیاز داریم از هم مستقل باشند.

$\hat{Q}(s,a)$ و $\left(\nabla_\theta \log\pi_\theta(a,s)\right)^T\left(\nabla_\theta \log\pi_\theta(a,s)\right)$ از یکدیگر مستقل هستند.

$$b(s) = \frac{\mathbb{E}_{\rho_{\pi_\theta},\pi}\left\{(\nabla_\theta \log\pi_\theta(a,s))^T \nabla_\theta \log\pi_\theta(a,s)\right\}\mathbb{E}_{\rho_{\pi_\theta},\pi}\left\{\hat{Q}(s,a)\right\}}{\mathbb{E}_{\rho_{\pi_\theta},\pi}\left\{(\nabla_\theta \log\pi_\theta(a,s))^T \nabla_\theta \log\pi_\theta(a,s)\right\}}$$

$$\Rightarrow b(s) = \mathbb{E}_{\rho_{\pi_\theta},\pi}\left\{\hat{Q}(s,a)\right\} = V^\pi(s).$$

s.a.m

سوال ۶

الف) فرض کنید μ و ν دو توزیع احتمالاتی بر روی فضای $\mathcal{X}$ باشند. در این صورت داریم که:

$$\|\mu-\nu\|_{TV} = \inf\{P(X\neq Y) \mid X,Y \text{ یک Coupling از } \mu,\nu \text{ باشند}\}$$

اثبات قضیه ①:

$$\forall A\subseteq X: \mu(A)-\nu(A) = P(X\in A)-P(Y\in A) \leq P(X\in A, Y\notin A) \leq P(X\neq Y) \Rightarrow$$

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

$$\Rightarrow \|\mu-\nu\|_{TV} = \max_{A\subseteq X} |\mu(A)-\nu(A)| \leq \inf\{P(X\neq Y) \mid (X,Y) \text{ is a coupling of } \mu,\nu\}$$

کافی برای اثبات قضیه تنها کافی است که یک X,Y معرفی کنیم که در آن تساوی برقرار باشد. برای این کار به صورت زیر عمل می‌کنیم.

$$P = \sum \min\{\mu_{(x)}, \nu_{(x)}\} = \sum_{x:\mu_{(x)}\leq \nu_{(x)}} \mu_{(x)} + \sum_{x:\mu_{(x)}>\nu_{(x)}} \nu_{(x)}$$

s.a.m

$$\Rightarrow P_s = \sum_{x:\mu(x)\not\geq\nu(x)} \mu(x) + (1-1)\times \sum_{x:\mu(x)\geq\nu(x)} \mu(x) + \sum_{x:\mu(x)\geq\nu(x)} \nu(x) =$$

$$= 1 + \sum_{x:\mu(x)\not\geq\nu(x)} \{\nu(x) - \mu(x)\}$$

قضیهٔ دوم:

$$\|\mu - \nu\|_{TV} = \sum_{x:\mu(x)\not\geq\nu(x)} (\mu(x) - \nu(x))$$

اثبات قضیهٔ دوم: در نظر بگیرید یک مجموعه مانند B داریم به صورتی که:

$$B = \{x \mid \mu(x) \not\geq \nu(x)\}$$

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

~~$\forall x$~~

اثبات این دو نامساوی:

① $\forall x \in A\cap\bar{B} \Rightarrow \mu(x) - \nu(x) < 0$

$$\hookrightarrow \underbrace{(\mu(A) - \mu(A\cap B))}_{\mu(A\cap\bar{B})} - \underbrace{(\nu(A) - \nu(A\cap B))}_{\nu(A\cap\bar{B})} < 0$$

② ~~$\mu_{\cup B}$~~ $(\mu(B) - \mu(A\cap B)) - (\nu(B) - \nu(B\cap B)) \geq 0$

حال اگر در نامساوی بدست آمده $A = B$ قرار دهیم داریم

$$\left.\begin{array}{l} \mu(A) - \nu(A) = \mu(B) - \nu(B) \\ \mu(B) - \nu(B) \geq 0 \end{array}\right\} \Rightarrow |\mu(A) - \nu(A)| = \mu(B) - \nu(B)$$

s.a.m

با توجه به اثبات‌های گذشته، $\|\mu - \nu\|_{TV} = \mu(B) - \nu(B) = \sum_{x \in B} [\mu(x) - \nu(x)]$

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

حال دو متغیر X و Y را به صورت زیر می‌سازیم:

① با احتمال P Z را با احتمال $\frac{\min\{\mu(x), \nu(x)\}}{P}$ می‌سازیم و متغیر دوم $X = Y = Z$ ↳ $h_3(z)$

② با احتمال $1-P$ X و Y را از دو توزیع احتمال‌های زیر مستقل می‌گیریم:

$$h_1(x) = \begin{cases} \frac{\mu(x) - \nu(x)}{1-P} & \mu(x) > \nu(x) \\ 0 & \text{o.w.} \end{cases}$$

$$h_2(y) = \begin{cases} \frac{\nu(y) - \mu(y)}{1-P} & \nu(y) > \mu(y) \\ 0 & \text{o.w.} \end{cases}$$

در نتیجه داریم:

$$X \sim P h_3(x) + (1-P) h_1 = \mu(x)$$

$$Y \sim P h_3(y) + (1-P) h_2 = \nu(x)$$

در نتیجه (X, Y) یک Coupling است.

پس این X و Y که دقیقاً نامساوی داخل قضیه ② را به مساوی تبدیل می‌کند، کمترین ① ثابت می‌شود.

s.a.m

حال به اثبات صورت سوال می پردازیم، در نظر بگیرید در گام t ام هستیم و $\pi_Q(\cdot|s_t)$ و $\pi_{Q'}(\cdot|s_t)$ داریم. در نتیجه طبق قضیه‌ای که ثابت کردیم $X \sim R_Q(\cdot|s_t)$ و $Y \sim R_{Q'}(\cdot|s_t)$ ای داریم به صورتی که

$$\Rightarrow P(X \neq Y) = \underbrace{\|\pi_{Q'}(\cdot|s_t) - R_Q(\cdot|s_t)\|_{TV}}_{\leq \delta} \leq \varepsilon$$

$$P_{Q'}(s_t) = (1-\delta)^t P_Q(s_t) + \left(1-(1-\delta)^t\right) P_{mistake}(s_t) \Rightarrow$$

$$\Rightarrow |P_{Q'}(s_t) - P_Q(s_t)| = \left(1-(1-\delta)^t\right) |P_{mistake}(s_t) - P_Q(s_t)|$$

$$\leq 2\left(1-(1-\delta)^t\right)$$

$$\delta \leq \varepsilon \Rightarrow \left(1-(1-\delta)^t\right) \leq \left(1-(1-\epsilon)^t\right)$$

همچنین طبق بسط تیلور داریم که $(1-\varepsilon)^t \geq 1-\varepsilon t \Rightarrow 1-(1-\varepsilon)^t \leq \varepsilon t$

در نتیجه در نهایت داریم که

$$|P_{Q'}(s_t) - P_Q(s_t)| \leq 2\varepsilon t$$

s.a.m

Subject: Date: / /

ب) قضیه (3): فرض کنید یک تابع پاداش به صورت $h(s_t)$ داریم به صورت

$$\mathbb{E}_{P_{\theta'}(s_t)}\{h(s_t)\} \geq \mathbb{E}_{P_\theta(s_t)}\{h(s_t)\} - 2\varepsilon t \max_{s_t} |h(s_t)|$$

اثبات قضیه (3):

$$\mathbb{E}_{P_{\theta'}(s_t)}\{h(s_t)\} = \sum_{s_t} P_{\theta'}(s_t) h(s_t) =$$

$$= \sum_{s_t} P_\theta(s_t) h(s_t) + \sum_{s_t} (P_{\theta'}(s_t) - P_\theta(s_t)) h(s_t)$$

$$\geq \mathbb{E}_{P_\theta(s_t)}\{h(s_t)\} - |P_{\theta'}(s_t) - P_\theta(s_t)| \max_{s_t} h(s_t)$$

$$\geq \mathbb{E}_{P_\theta(s_t)}\{h(s_t)\} - 2\varepsilon \max_{s_t} |h(s_t)|$$

حال به اثبات صورت سؤال می پردازیم:

(1) حالت اول مجموعه:

$$h(s_t) = \mathbb{E}_{a_t \sim \pi_\theta(\cdot|s_t)}\left\{\frac{\pi_{\theta'}(a_t|s_t)}{\pi_\theta(a_t|s_t)} \gamma^t A^{\pi_\theta}(s_t, a_t)\right\}$$

$$\sum_t \mathbb{E}_{P_{\theta'}(s_t)}(|h(s_t)|) \geq \sum_t \mathbb{E}_{P_\theta(s_t)}\{h(s_t)\} - \sum_{t=0}^{T} 2\varepsilon t \max_{s_t} |h(s_t)|$$

s.a.m

Subject:

Date: / /

مرتبه اختلاف برابر است با:

$$\sum_{t=0}^{T} 2\varepsilon t \max_{s_t} |h(s_t)| = \varepsilon T(T+1) \times |h(s_t)|$$

با توجه به اینکه $h(s_t)$ تقریبا $O(r_{max})$ است، در نتیجه مرتبه اختلاف برابر است با $O(\varepsilon T^2 r_{max})$

maximum reward

حالت افق نامحدود.

اختلاف: $\sum_{t=0}^{\infty} 2\varepsilon t \max_{s_t} (h(s_t)) \leq 2\varepsilon \sum_{t=0}^{\infty} t\gamma^t r_{max}$

$$= 2\varepsilon r_{max} \times \gamma \sum_{t=0}^{\infty} \frac{d}{d\gamma}(\gamma^t)$$

$$= 2\varepsilon \gamma r_{max} \times \frac{d}{d\gamma}\left(\frac{1}{1-\gamma}\right) = \frac{2\varepsilon\gamma r_{max}}{(1-\gamma)^2}$$

مرتبه اختلاف: $O\left(\frac{2\varepsilon\gamma r_{max}}{(1-\gamma)^2}\right)$

Using pinsker inequality (2

$$\|\pi_{\theta'}(a_t|s_t) - \pi_\theta(a_t|s_t)\| \leq \sqrt{\frac{1}{2} D_{KL}(\pi_{\theta'}(\cdot|s_t) \| \pi_\theta(\cdot|s_t))} \Rightarrow$$

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

s.a.m

Subject: Date: / /

(2

$$D = \sum_t \mathbb{E}_{P_\theta(s_t)} \left\{ \mathbb{E}_{\pi_\theta(a_t|s_t)} \left\{ \frac{\pi_{\theta'}(a_t|s_t)}{\pi_\theta(a_t|s_t)} \gamma^t A^{\pi_\theta}(s_t, a_t) \right\} \right\}$$

$$\nabla_{\theta'} D = \sum_t \mathbb{E}_{P_\theta(s_t)} \left\{ \mathbb{E}_{\pi_\theta(a_t|s_t)} \left\{ \frac{\nabla_{\theta'} \pi_{\theta'}(a_t|s_t)}{\pi_\theta(a_t|s_t)} \gamma^t A^{\pi_\theta}(s_t, a_t) \right\} \right\}$$

$$= \sum_t \mathbb{E}_{P_\theta(s_t)} \left\{ \mathbb{E}_{\pi_\theta(a_t|s_t)} \left\{ \frac{\pi_{\theta'}(a_t|s_t)}{\pi_\theta(a_t|s_t)} \nabla_{\theta'} \log \pi_{\theta'}(a_t|s_t) \gamma^t A^{\pi_\theta}(s_t, a_t) \right\} \right\}$$

$$\Rightarrow \nabla_{\theta'} D \Big|_{\theta'=\theta} = \sum_t \mathbb{E}_{P_\theta(s_t)} \mathbb{E}_{\pi_\theta(a_t|s_t)}$$

$$\nabla_{\theta'} D \Big|_{\theta'=\theta} = \sum_t \mathbb{E}_{P_\theta(s_t)} \left\{ \mathbb{E}_{\pi_\theta(a_t|s_t)} \left\{ \gamma^t \nabla_{\theta'} \log \pi_{\theta'}(a_t|s_t) A^{\pi_\theta}(s_t, a_t) \right\} \right\}$$

$$D \geq \sum_t \mathbb{E}_{P_\theta(s_t)} \left\{ \mathbb{E}_{\pi_\theta(a_t|s_t)} \left\{ \gamma^t A^{\pi_\theta}(s_t, a_t) \right\} \right\} + \nabla_{\theta'} D \Big|_{\theta'=\theta}^{T} (\theta' - \theta)$$

قسمت اخر سوال

Markov Chains and Mixing Times

سوال ۴)

$$\nabla_\theta V^{\mu_\theta}(s) = \nabla_\theta Q^{\mu_\theta}(s, \mu_\theta(s)) = \nabla_\theta \left(r(s, \mu_\theta(s)) + \int_S \gamma p(s'|s, \mu_\theta(s)) V^{\mu_\theta}(s') ds' \right)$$

$$= \nabla_\theta \mu_\theta(s) \nabla_a r(s,a)\big|_{a=\mu_\theta(s)} + \nabla_\theta \int_S \gamma p(s'|s, \mu_\theta(s)) V^{\mu_\theta}(s') ds'$$

$$= \nabla_\theta \mu_\theta(s) \nabla_a r(s,a)\big|_{a=\mu_\theta(s)}$$

$$+ \int_S \gamma \left(p(s'|s, \mu_\theta(s)) \nabla_\theta V^{\mu_\theta}(s') + \nabla_\theta \mu_\theta(s) \nabla_a p(s'|s,a)\big|_{a=\mu_\theta(s)} V^{\mu_\theta}(s') \right) ds'$$

$$= \nabla_\theta \mu_\theta(s) \nabla_a \left(r(s,a) + \int_S \gamma p(s'|s,a) V^{\mu_\theta}(s') ds' \right)\Big|_{a=\mu_\theta(s)}$$

$$+ \int_S \gamma p(s'|s, \mu_\theta(s)) \nabla_\theta V^{\mu_\theta}(s') ds' =$$

$$= \nabla_\theta \mu_\theta(s) \nabla_a Q^{\mu_\theta}(s,a)\big|_{a=\mu_\theta(s)} + \int_S \gamma p(s \to s', 1, \mu_\theta) \nabla_\theta V^{\mu_\theta}(s') ds'$$

حال به صورت بازگشتی در این رابطه جایگذاری می کنیم

$$\nabla_\theta V^{\mu_\theta}(s) = \nabla_\theta \mu_\theta(s) \nabla_a Q^{\mu_\theta}(s,a)\big|_{a=\mu_\theta(s)}$$

$$+ \int_S \gamma p(s \to s', 1, \mu_\theta) \nabla_\theta \mu_\theta(s') \nabla_\theta \mu_\theta(s') \nabla_a Q^{\mu_\theta}(s',a)\big|_{a=\mu_\theta(s')} ds' +$$

$$+ \int_S \gamma p(s \to s', 1, \mu_\theta) \int_S \gamma p(s' \to s'', 1, \mu_\theta) \nabla_\theta V^{\mu_\theta}(s'') ds'' ds'$$

s.a.m

$$= \nabla_\theta \mu_\theta(s) \nabla_a Q^{\mu_\theta}(s,a)\Big|_{a=\mu_\theta(s)}$$
$$+ \int_S \gamma P(s \to s', 1, \mu_\theta) \nabla_\theta \mu_\theta(s') \nabla_a Q^{\mu_\theta}(s', a)\Big|_{a=\mu_\theta(s')} ds'$$
$$+ \int_S \gamma^2 P(s \to s', 2, \mu_\theta) \nabla_\theta V^{\mu_\theta}(s') ds'$$

به همین صورت اگر ادامه بدهیم به این می رسیم که

$$\nabla_\theta V^{\mu_\theta}(s) = \int_S \sum_{t=0}^{\infty} \gamma^t P(s \to s', t, \mu_\theta) \nabla_\theta \mu_\theta(s') \nabla_a Q^{\mu_\theta}(s', a)\Big|_{a=\mu_\theta(s')} ds'$$

حال با اضافه ریاض گرفتن از این عبارت می رسیم به

$$\nabla_\theta J(\mu_\theta) = \nabla_\theta \int_S P_1(s) V^{\mu_\theta}(s) ds$$
$$= \int_S P_1(s) \nabla_\theta V^{\mu_\theta}(s) ds$$
$$= \int_S \int_S \sum_{t=0}^{\infty} \gamma^t P_1(s) P(s \to s', t, \mu_\theta) \nabla_\theta \mu_\theta(s') \nabla_a Q^{\mu_\theta}(s', a)\Big|_{a=\mu_\theta(s')} ds' ds$$
$$= \int_S \rho^{\mu_\theta}(s) \nabla_\theta \mu_\theta(s) \nabla_a Q^{\mu_\theta}(s, a)\Big|_{a=\mu_\theta(s)} ds$$

s.a.m