دانیال پرنیان — یادگیری تقویتی

۴۰۱۱۱۰۳۰۷ — تمرین ۸

Subject :
Year . Month . Date . ( )

① از رابطه بلمن $V^{\pi_\theta}(s) = \sum_a \pi_\theta(a|s) Q^{\pi_\theta}(s,a)$ می دانیم. با مشتق گرفتن از این رابطه خواهیم داشت،

$$\nabla_\theta V^{\pi_\theta}(s) = \sum_a \left[(\nabla_\theta \pi_\theta(a|s)) Q^{\pi_\theta}(s,a) + \pi_\theta(a|s) \nabla_\theta Q^{\pi_\theta}(s,a) \right] \quad (I)$$

از رابطه بلمن برای Q می دانیم:

$$\nabla_\theta Q^{\pi_\theta}(s,a) = \nabla_\theta \left[r(s,a) + \gamma \sum_{s'} P(s'|s,a) V^{\pi_\theta}(s') \right] = \gamma \sum_{s'} P(s'|s,a) \nabla_\theta V^{\pi_\theta}(s')$$

همچنین می دانیم که $\nabla_\theta \pi_\theta(a|s) = \pi_\theta(a|s) \nabla_\theta \log \pi_\theta(a|s)$. اکنون با جایگذاری مشتق های $Q^{\pi_\theta}(s,a)$ و $\pi_\theta(a|s)$ در رابطه (I) خواهیم داشت:

$$\nabla_\theta V^{\pi_\theta}(s) = \sum_a \pi_\theta(a|s) \nabla_\theta \log \pi_\theta(a|s) Q^{\pi_\theta}(s,a) + \gamma \sum_a \pi_\theta(a|s) \sum_{s'} P(s'|s,a) \nabla_\theta V^{\pi_\theta}(s)$$

$$= E_{a \sim \pi_\theta(\cdot|s)} \left[\nabla_\theta \log \pi_\theta(a|s) Q^{\pi_\theta}(s,a) \right] + \gamma \sum_{s'} Pr^{\pi_\theta}(s_1 = s' | s_0 = s) \nabla_\theta V^{\pi_\theta}(s')$$

توجه کنید عبارت دوم با کمک $Pr^{\pi_\theta}(s_1 = s' | s_0 = s) = \sum_a \pi_\theta(a|s) P(s'|s,a)$ نتیجه شد، و درواقع احتمال رفتن از حالت s به s' را نشان می دهد (براساس سیاست π_θ).
عبارت اول را نیز از این به بعد با $f(s)$ نشان می دهیم.
بنابراین داریم:

$$\nabla_\theta J(\theta) = \nabla_\theta V^{\pi_\theta}(s_0) = f(s_0) + \gamma \sum_{s_1} Pr^{\pi_\theta}(s_1|s_0) \nabla_\theta V^{\pi_\theta}(s_1)$$

اکنون $\nabla_\theta V^{\pi_\theta}(s_1)$ را نیز با رابطه‌ای که داریم می نویسیم:

$$\nabla_\theta J(\theta) = f(s_0) + \gamma \sum_{s_1} Pr^{\pi_\theta}(s_1|s_0) \left[f(s_1) + \gamma \sum_{s_2} Pr^{\pi_\theta}(s_2|s_1) \nabla_\theta V^{\pi_\theta}(s_2) \right] \quad (II)$$

با توجه به ویژگی مارکفی، می دانیم $Pr^{\pi_\theta}(s_2|s_0) = \sum_{s_1} Pr^{\pi_\theta}(s_1|s_0) Pr^{\pi_\theta}(s_2|s_1)$. اثبات این نکته را در زیر آورده ام:

$$Pr^{\pi}(s_2 = s'' | s_0) = \sum_{a_0, s_1, a_1} P(s_1|s_0,a_0) \times \pi(a_0|s_0) \times P(s''|s_1,a_1) \times \pi(a_1|s_1)$$

$$= \sum_{s_1, a_1} Pr^{\pi}(s_1|s_0) \times P(s''|s_1,a_1) \times \pi(a_1|s_1)$$

$$= \sum_{s_1} Pr^{\pi}(s_1|s_0) \sum_{a_1} P(s''|s_1,a_1) \pi(a_1|s_1) = \sum_{s_1} Pr^{\pi}(s_1|s_0) \times Pr^{\pi}(s''|s_1)$$

P4PCO

Subject :
Year . Month . Date . ( )

بنابراین با کمک این قضیه و همچنین مجددا بازنویسی $\nabla_\theta V^{\pi_\theta}(s_0)$ در معادله (II) به کمک رابطه‌ای که داریم، می‌توان به نتیجه‌ی زیر رسید:

$$\nabla_\theta J(\theta) = f(s_0) + \sum_s \gamma P_r^{\pi_\theta}(s_1=s|s_0) f(s) + \sum_s \gamma^2 P_r^{\pi_\theta}(s_2=s|s_0) f(s) + \cdots$$
$$= \sum_{t=0}^{\infty} \sum_s \gamma^t P_r^{\pi_\theta}(s_t=s|s_0) f(s) = \sum_s \sum_{t=0}^{\infty} \gamma^t P_r^{\pi_\theta}(s_t=s|s_0) f(s) \quad \text{(III)}$$

از رابطه‌ای که در فرض‌ها داده شده می‌دانیم:

$$d_{s_0}^{\pi_\theta}(s) = (1-\gamma) \sum_{t=0}^{\infty} \gamma^t P_r^{\pi_\theta}(s_t=s|s_0)$$

بنابراین، رابطه‌ی بالا را در معادله (III) جایگذاری کنید:

$$\nabla_\theta J(\theta) = \sum_s \frac{d_{s_0}^{\pi_\theta}(s)}{1-\gamma} f(s) = \frac{1}{1-\gamma} \times \sum_s d_{s_0}^{\pi_\theta}(s) \, E_{a\sim\pi_\theta(\cdot|s)} \left[\nabla_\theta \log \pi_\theta(a|s) Q^{\pi_\theta}(s,a) \right]$$

که اگر براساس امید ریاضی از توزیع $s \sim d_{s_0}^{\pi_\theta}$ بازنویسی کنیم، به همان رابطه‌ی گفته شده در حکم سوال می‌رسیم:

$$\nabla_\theta J(\theta) = \frac{1}{1-\gamma} E_{s\sim d_{s_0}^{\pi_\theta}} \left[E_{a\sim\pi(\cdot|s)} \left[\nabla_\theta \log \pi_\theta(a|s) Q^{\pi_\theta}(s,a) \right] \right]$$

---

(۲) قضیه‌ی Compatible Function Approximation

برای اثبات حکم کافیست برابری مقداری که در سوال مثل برای $\nabla_\theta J(\theta)$ بدست آوردیم، و رابطه‌ی جدیدی که در سوال جدید گفته شده، ثابت کنیم. برابری این دو، معادل با حکم زیری می‌شود:

$$E_{s\sim d_{s_0}^{\pi_\theta}} E_{a\sim\pi_\theta(\cdot|s)} \left[\nabla_\theta \log \pi_\theta(a|s) \left(Q^{\pi_\theta}(s,a) - Q_\phi(s,a) \right) \right] = 0 \quad \text{(I)}$$

شرط دوم سوال می‌گوید که ϕ مقدار ϵ را کمینه می‌کند. بنابراین نتیجه می‌شود مشتق ϵ نسبت به ϕ برابر صفر است. اکنون نشان می‌دهیم $\nabla_\phi \epsilon = 0$ و رابطه‌ی (I) معادل هستند، و بنابراین حکم اثبات می‌شود.

با توجه به اینکه توزیع‌های $d_{s_0}^{\pi_\theta}$ و $\pi_\theta(\cdot|s)$ به ϕ وابسته نیستند، پس می‌توان ∇_ϕ را داخل امید ریاضی‌ها برد:

$$\nabla_\phi \epsilon = \nabla_\phi E_{s\sim d_{s_0}^{\pi_\theta}} E_{a\sim\pi_\theta(\cdot|s)} \left[\left(Q^{\pi_\theta}(s,a) - Q_\phi(s,a) \right)^2 \right]$$

نام‌گذاری مختصر توزیع‌ها:

$$= 0 \, E_A \, E_B \left[\nabla_\phi \left(Q^{\pi_\theta}(s,a) - Q_\phi(s,a) \right)^2 \right]$$
$$= E_A E_B \left[2 \left(Q^{\pi_\theta}(s,a) - Q_\phi(s,a) \right) \left(-\nabla_\phi Q_\phi(s,a) \right) \right]$$

استفاده از شرط اول:

$$= E_A E_B \left[2 \left(Q^{\pi_\theta}(s,a) - Q_\phi(s,a) \right) \left(-\nabla_\theta \log \pi_\theta(a|s) \right) \right]$$

Subject :
Year . Month . Date . ( )

اکنون اگر یک ۲- از داخل امید ریاضی فاکتور بگیریم خواهیم داشت :

۲- $E_{s\sim d_{s_0}}^{\pi_\theta} E_{a\sim\pi(\cdot|s)} \left[\nabla_\theta \log \pi_\theta(a|s) \left(Q^{\pi_\theta}(s,a) - Q_\phi(s,a) \right) \right] = 0$

که همان رابطهٔ (I) را نتیجه می‌دهد و حکم ثابت شد .

بخش 2.1

a) با توجه به تعاریف می‌توان نوشت :

$$\eta(\pi) = E_{s_0,a_0,s_1,\dots}\left[\sum_{t=0}^{\infty}\gamma^t r(s_t)\right] = E_{s_0} E_{a_0,s_1,\dots}\left[\sum_{t=0}^{\infty}\gamma^t r(s_t)\right] = E_{s_0\sim\rho(s_0)}\left[V_\pi(s_0)\right]$$

روابط را به نحو زیر ادامه می‌دهیم :

$$\begin{aligned}\eta(\tilde{\pi}) &= E_{s_0\sim\rho(s_0)}\left[V_{\tilde{\pi}}(s_0)\right] = E_{\tau\sim\tilde{\pi}}\left[V_{\tilde{\pi}}(s_0)\right] \\ &= E_{\tau\sim\tilde{\pi}}\left[\sum_{t=0}^{\infty}\gamma^t V_{\tilde{\pi}}(s_t) - \sum_{t=1}^{\infty}\gamma^t V_{\tilde{\pi}}(s_t)\right] \\ &= E_{\tau\sim\tilde{\pi}}\left[\sum_{t=0}^{\infty}\gamma^t\left(\gamma V_{\tilde{\pi}}(s_{t+1}) - V_{\tilde{\pi}}(s_t)\right)\right] \quad (I)\end{aligned}$$

$$\eta(\pi) = E_{\tau\sim\pi}\left[\sum_{t=0}^{\infty}\gamma^t r(s_t)\right] \quad (II)$$

همچنین می‌دانیم که :

$$(I),(II) \Rightarrow \eta(\tilde{\pi}) - \eta(\pi) = E_{\tau\sim\tilde{\pi}}\left[\sum_{t=0}^{\infty}\gamma^t\left(r(s_t) + \gamma V_\pi(s_{t+1}) - V_\pi(s_t)\right)\right] \quad (III)$$

توجه کنید بر اساس تعاریف به سادگی می‌توان دید

$$Q_\pi(s_t,a_t) = r(s_t) + \gamma V_\pi(s_{t+1})$$

و بنابراین :

$$A_\pi(s_t,a_t) = Q_\pi(s_t,a_t) - V_\pi(s_t) = r(s_t) + \gamma V_\pi(s_{t+1}) - V_\pi(s_t) \quad (IV)$$

$$(III),(IV) \Rightarrow \eta(\tilde{\pi}) = \eta(\pi) + E_{\tau\sim\tilde{\pi}}\left[\sum_{t=0}^{\infty}\gamma^t A_\pi(s_t,a_t)\right]$$

---

b) از نتیجهٔ بخش قبل استفاده می‌کنیم :

$$\begin{aligned}\eta(\tilde{\pi}) - \eta(\pi) &= E_{\tau\sim\tilde{\pi}}\left[\sum_{t=0}^{\infty}\gamma^t A_\pi(s_t,a_t)\right] \\ &= \sum_{t=0}^{\infty} E_{s_t\sim\rho_{\tilde{\pi}}(s_t)}\left[E_{a_t\sim\tilde{\pi}(\cdot|s_t)}\left[\gamma^t A_\pi(s_t,a_t)\right]\right] \\ &= \sum_{t=0}^{\infty}\sum_s P(s_t=s)\sum_a \tilde{\pi}(a|s)\,\gamma^t A_\pi(s_0,a_0) \\ &= \sum_s\left(\sum_{t=0}^{\infty}\gamma^t P(s_t=s)\right)\sum_a \tilde{\pi}(a|s) A_\pi(s_0,a_0)\end{aligned}$$

$$\Rightarrow \eta(\tilde{\pi}) = \eta(\pi) + \sum_s \rho_{\tilde{\pi}}(s)\sum_a \tilde{\pi}(a|s) A_\pi(s,a)$$

حکم ثابت شد .

Subject :
Year . Month . Date . ( )

c) از تعریف، می دانیم:

$$\bar{A}(s) = \sum_{a'} \pi'(a'|s) A_\pi(s,a') = \sum_{a'} \sum_{a} P(a,a'|s) A_\pi(s,a')$$

تساوی دوم از اینجا آمد که π'، روی a' از توزیع marginal dist. $a,a'|s$ است.

همچنین توجه کنید که $E_{a\sim\pi(\cdot|s)}[A_\pi(s,a)]=0$ زیرا:

$$E_{a\sim\pi(\cdot|s)}[A_\pi(s,a)] = E_{a\sim\pi(\cdot|s)}[Q_\pi(s,a)] - E_{a\sim\pi(\cdot|s)}[V_\pi(s)]$$
$$= V_\pi(s) - V_\pi(s) = 0$$

پس می توان نوشت:

$$0 = \sum_{a} \pi(a|s) A_\pi(s,a) = \sum_{a} \sum_{a'} P(a,a'|s) A_\pi(s,a)$$

با توجه به ۰ بودن این عبارت، می توان $\bar{A}(s)$ را بازنویسی کرد:

$$|\bar{A}(s)| = \left| \sum_{a',a} P(a,a'|s) A_\pi(s,a') - \sum_{a,a'} P(a,a'|s) A_\pi(s,a) \right|$$
$$= \left| \sum_{a,a'} P(a,a'|s) \left(A_\pi(s,a') - A_\pi(s,a) \right) \right|$$

نامساوی مثلث

$$\leq \sum_{a,a'} P(a,a'|s) \left| A_\pi(s,a') - A_\pi(s,a) \right|$$

حال توجه کنید اگر $a=a'$، آنگاه $A_\pi(s,a')-A_\pi(s,a)=0$.

در غیر این صورت نیز می توان کران بالای زیر را داد:

$$|A_\pi(s,a') - A_\pi(s,a)| \leq |A_\pi(s,a')| + |-A_\pi(s,a)| \leq 2 \max_{s,a} |A_\pi(s,a)|$$

در نتیجه می توان نوشت:

$$|\bar{A}(s)| \leq \sum_{a\neq a'} P(a,a'|s) \times 2 \max_{s,a} |A_\pi(s,a)|$$

$$\Rightarrow |\bar{A}(s)| \leq 2\alpha \max_{s,a} |A_\pi(s,a)|$$

حکم ثابت شد

طبق s: $\sum_{a\neq a'} P(a,a'|s) = P(a\neq a'|s) \leq \alpha$

d)

$$|E_{s_t\sim\pi'}[\bar{A}(s_t)] - E_{s_t\sim\pi}[\bar{A}(s_t)]| = \left| \sum_s P_{\pi'}(s_t=s) \bar{A}(s) - \sum_s P_\pi(s_t=s) \bar{A}(s) \right|$$

نامساوی مثلث

$$\leq \sum_s |P_{\pi'}(s_t=s) - P_\pi(s_t=s)| |\bar{A}(s)|$$

لم قبلی

$$\leq \sum_s |P_{\pi'}(s_t=s) - P_\pi(s_t=s)| \times 2\alpha \max_{s,a} |A_\pi(s,a)|$$

بنابراین کافیست ثابت کنیم $\sum_s |P_{\pi'}(s_t=s) - P_\pi(s_t=s)| \leq 2(1-(1-\alpha)^t)$ تا حکم نهایی اثبات شود. این موضوع نیز در کلاس ثابت شده بود و اثبات آن را در ادامه تکرار می کنم: (تقریبا)

Subject :
Year . Month . Date . ( )

فرض کنید S_n ، S'_n ، π' ، π به ترتیب حالت در زمان n براساس پالیسی π' و π باشد، با این فرض که هردو از حالت اولیه یکسان شروع شوند: $S'_0 = S_0 \sim P_0(s)$

در هر لحظه، اکشن های (a_n, a'_n) را از توزیع $P(a, a' | s_n)$ انتخاب می کنیم (اگر که $S_n = S'_n$ بود). اگر $a_n = a'_n$ باشد، آنگاه $S_{n+1} = S'_{n+1}$ خواهد بود (چون از random seed های transition یکسان استفاده می کنیم) که احتمال این اتفاق حداقل $1-\alpha$ است. اگر اینگونه نباشد، S_{n+1} و S'_{n+1} لزومی ندارد برابر شوند، هرچند بازهم ممکن است.

اکنون احتمال اینکه در لحظه n، $S_n \neq S'_n$ باشد به نحو زیر است:

$$P(S'_n \neq S_n) = P(S_n \neq S'_n | S_{n-1} = S'_{n-1}) \times P(S_{n-1} = S'_{n-1}) + P(S_n \neq S'_n | S_{n-1} \neq S'_{n-1}) P(S_{n-1} \neq S'_{n-1})$$

$$\begin{cases} P(S_n \neq S'_n | S_{n-1} = S'_{n-1}) \leq P(a_n \neq a'_n | S_{n-1} = S'_{n-1}) \leq \alpha \\ P(S_n \neq S'_n | S_{n-1} \neq S'_{n-1}) \leq 1 \end{cases}$$

$$\Rightarrow P(S_n \neq S'_n) \leq \alpha \times (1 - P(S_{n-1} \neq S'_{n-1})) + P(S_{n-1} \neq S'_{n-1}) = \alpha + (1-\alpha) P(S_{n-1} \neq S'_{n-1})$$

$$\Rightarrow P(S_n \neq S'_n) \leq \alpha + \alpha(1-\alpha) + \alpha(1-\alpha)^2 + \cdots + \alpha(1-\alpha)^{n-1} = \alpha \frac{1-(1-\alpha)^n}{1-(1-\alpha)} = 1 - (1-\alpha)^n \quad (*)$$

همچنین می دانیم: $\sum_s |P_{\pi'}(S_t = s) - P_{\pi}(S_t = s)| = 2 D_{TV}(P_{\pi'}(S_t = \cdot) \| P_{\pi}(S_t = \cdot))$

$$\leq 2 P(S'_t \neq S_t) \overset{(*)}{\leq} 2 (1 - (1-\alpha)^n)$$

حکم معادلی که می خواستیم ثابت شد و اثبات کامل شد.

---

e) از تعاریف موجود برای π و L_π می توان نوشت:

$$|\pi(\pi') - L_\pi(\pi')| = \left| E_{\tau \sim \pi'} \left[\sum_{t=0}^{\infty} \gamma^t \bar{A}(s_t) \right] - E_{\tau \sim \pi} \left[\sum_{t=0}^{\infty} \gamma^t \bar{A}(s_t) \right] \right|$$

خطی بودن امید ریاضی:

$$= \left| \sum_{t=0}^{\infty} \gamma^t \left(E_{\tau \sim \pi'} [\bar{A}(s_t)] - E_{\tau \sim \pi} [\bar{A}(s_t)] \right) \right|$$

نامساوی مثلث:

$$\leq \sum_{t=0}^{\infty} \gamma^t \left| E_{\tau \sim \pi'} [\bar{A}(s_t)] - E_{\tau \sim \pi} [\bar{A}(s_t)] \right|$$

لم سوال قبلی:

$$\leq \sum_{t=0}^{\infty} \gamma^t \left(\varepsilon \alpha (1 - (1-\alpha)^t) \epsilon \right) = \varepsilon \alpha \epsilon \sum_{t=0}^{\infty} \gamma^t (1 - (1-\alpha)^t)$$

اکنون کافیست ثابت کنیم $\sum_{t=0}^{\infty} \gamma^t (1-(1-\alpha)^t) \leq \frac{\alpha\gamma}{(1-\gamma)^2}$ تا حکم ثابت شود:

$$\sum_{t=0}^{\infty} \gamma^t (1-(1-\alpha)^t) = \sum_{t=0}^{\infty} \gamma^t - \sum_{t=0}^{\infty} (\gamma(1-\alpha))^t = \frac{1}{1-\gamma} - \frac{1}{1-\gamma(1-\alpha)}$$

$$= \frac{1-\gamma(1-\alpha) - (1-\gamma)}{(1-\gamma)(1-\gamma+\gamma\alpha)} = \frac{\gamma\alpha}{(1-\gamma)(1-\gamma+\gamma\alpha)} \leq \frac{\gamma\alpha}{(1-\gamma)^2}$$

حکم ثابت شد.

* برای نامساوی آخر، صرفا از این استفاده کردیم که $1-\gamma \leq 1-\gamma+\gamma\alpha$.

P4PCO

Subject :
Year. Month. Date. ( )

۴) لم 2.5 به ما می‌گوید اگر (π, π') یک α-coupled Policy Pair باشد (که طبق فرض هست) آنگاه داریم:

$$|\eta(\pi') - L_\pi(\pi')| \le \frac{4\alpha^2\gamma\epsilon}{(1-\gamma)^2} \Rightarrow \eta(\pi') \ge L_\pi(\pi') - \frac{4\alpha^2\gamma\epsilon}{(1-\gamma)^2}$$

با توجه به فرض‌ها، می‌توانیم در لم بالا قرار دهیم: $\alpha = D_{TV}^{max}(\pi, \pi')$:

$$\eta(\pi') \ge L_\pi(\pi') - \frac{4(D_{TV}^{max}(\pi,\pi'))^2\gamma\epsilon}{(1-\gamma)^2} \quad (*)$$

با توجه به نامساوی $D_{TV}(P\|Q)^2 \le D_{KL}(P\|Q)$ که در فرضیات داده شده، داریم:

$$D_{TV}(\pi(\cdot|s)\|\pi'(\cdot|s))^2 \le D_{KL}(\pi(\cdot|s)\|\pi'(\cdot|s)) \quad (I)$$

داریم:

$$(D_{TV}^{max}(\pi,\pi'))^2 = \left[\max_s D_{TV}(\pi(\cdot|s)\|\pi'(\cdot|s))\right]^2 = \max_s \left[D_{TV}(\pi(\cdot|s)\|\pi'(\cdot|s))\right]^2$$

$$\overset{(I)}{\le} \max_s D_{KL}(\pi(\cdot|s)\|\pi'(\cdot|s)) = D_{KL}^{max}(\pi,\pi')$$

$$\overset{(*)}{\Rightarrow} \eta(\pi') \ge L_\pi(\pi') - \frac{4\epsilon\gamma}{(1-\gamma)^2} \times D_{KL}^{max}(\pi,\pi')$$

حکم اثبات شد