تمرین ۷

دانیال پرنیان

۴۰۱۱۱۰۳۰۷

یادگیری تقویتی

Subject:
Year. Month. Date. ( )

۱.۱.۱ با استقرا روی k ثابت می کنیم $V_k^* \le R_{max}(1+\gamma+\dots+\gamma^{k-1})$ برای هر $k \ge 1$.

پایه استقرا: می دانیم $V_1^* = \max_a \sum_{s'} T(s,a,s')[R+\gamma V_0^*(s')]$ و $V_0^*(s')=0$ ، $R \le R_{max}$ بنابراین:

$$\forall s: V_1^*(s) \le \max_a \sum_{s'} T(s,a,s') \times R_{max} = \max_a R_{max} = R_{max}$$

پس برای $k=1$ حکم ثابت شد.

گام استقرا: فرض کنید برای $k-1$ برقرار باشد:

$$V_{k-1}^*(s') \le R_{max}(1+\gamma+\dots+\gamma^{k-2})$$

$$V_k^*(s) = \max_a \sum_{s'} T(s,a,s')\left[R(s,a,s') + \gamma V_{k-1}^*(s')\right] \le \max_a \sum_{s'} T(s,a,s')\left[R_{max} + \gamma V_{k-1}^*(s')\right]$$

$$\le \max_a \sum_{s'} T(s,a,s')\left[R_{max} + \gamma R_{max}(1+\gamma+\dots+\gamma^{k-2})\right] = R_{max}(1+\gamma+\dots+\gamma^{k-1})$$

پس ادعایمان ثابت شد و یک کران بالا برای V_k^* پیدا کردیم.

اثبات راحت تر: طبق تعریف، V_k^* مقدار Optimal Value است در حالتی که $H=k$ باشد یعنی کلا k گام حرکت کنیم. منطقا در هر گام حداکثر Reward ما R_{max} است پس به سادگی می توان نتیجه گرفت:

$$V_k^*(s) = \max_{a_0,a_1,\dots,a_{k-1}} E\left[\sum_{t=0}^{k-1} \gamma^t R(s_t,a_t,s_{t+1}) \mid s_0=s\right] \le \max E\left[\sum_{t=0}^{k-1} \gamma^t R_{max}\right]$$

$$\Rightarrow V_k^*(s) \le R_{max}(1+\gamma+\dots+\gamma^{k-1})$$

۲.۱.۱ فرض کنید π به گونه باشد که ابتدا k گام action بهینه را انجام دهد و سپس در گام $k+1$ رندم حرکت کند (مثلا اکشن a_k). در این صورت:

$$V_{k+1}^{\pi}(s) = V_k^*(s) + \gamma^k E[R(s_k,a_k)] \ge V_k^*(s)$$

خاصیتی را از اینجا نتیجه گرفتیم که تمام $R(s_k,a_k)$ ها نامنفی هستند. پس:

$$V_{k+1}^* \ge V_{k+1}^{\pi} \ge V_k^* \longrightarrow V_k^* \text{ is non-decreasing}$$

ما در Valve_iteration در هر مرحله قرار می دهیم $V^k(s) := \max_a \sum_{s'} P(s,a,s')[R+\gamma V^{k-1}(s')]$ که در واقع V^k همان V_k^* می شود. ثابت کردیم V^k ها صعودی هستند و همچنین طبق بخش قبل، یک کران بالا دارند. بنابراین تعداد متناهی گام می توانند اکیدا صعودی باشند و از جایی به بعد، ثابت می مانند و در واقع converge می کنند و در آنجا داریم $V^{n+1}(s) = V^n(s)$ پس $V^{n+1}(s) = \max_a \sum_{s'} P(s,a,s')[R+\gamma V^n(s')]$ یعنی V ها در معادله بلمن (optimality) صدق می کنند.

P4PCO

Subject :
Year. Month. Date. ( )

۳.۱.۱ فرض کنید $\lim_{k\to\infty} V^k = V^*$ و داریم:

$$\lim_{k\to\infty} V^{k+1}(s) = \lim_{k\to\infty} \max_a \sum_{s'} P(s,a,s')\cdot\left[R(s,a,s') + \gamma V^k(s')\right]$$

$$\Rightarrow V^*(s) = \max_a \{P(s,a,s')\left[R(s,a,s') + \gamma V^*(s')\right]$$

پس V^* در معادلهٔ بهینگی بلمن صدق می‌کند و بنابراین optimal value function است.

۲.۱ در هر حرکت یک r_0 به reward دریافت‌شده اضافه می‌شود بنابراین داریم:

$$\hat{V}_k^*(s) = V_k^*(s) + r_0(1+\gamma+\dots+\gamma^{k-1}) = V_k^*(s) + r_0\frac{1-\gamma^k}{1-\gamma} = V_k^*(s) + C$$

می‌توان دید به تمام value ها مقدار ثابت C مستقل از s اضافه شده. بنابراین نباید تفاوتی ایجاد شود و پالیسی بهینه برای $\hat{V}_k^*$ همان پالیسی بهینه برای V_k^* است.

اکنون اگر r_0 را طوری قرار دهیم که تمام $\hat{R}$ ها نامنفی باشد (مثلاً قرار دهیم $r_0 = -\min_{s,a} R(s,a)$) طبق آنالیزی که در بخش ۱.۱ انجام دادیم، نتیجه می‌شود $\hat{V}_k^*$ ها به مقدار بهینه همگرا می‌شوند و همانطور که گفتیم جواب سیاست بهینه، سیاست بهینهٔ مسئلهٔ اصلی هم خواهد بود. برای یافتن $\hat{V}^*$ توجه کنید:

$$\hat{V}^*(s) = \max_a E_{s'}\left[\hat{R}(s,a,s') + \gamma \hat{V}^*(s')\right]$$

اکنون قرار دهید $\hat{V}^* = V^* + C$. پس داریم:

$$V^*(s) + C = \max_a E_{s'}\left[R(s,a,s') + r_0 + \gamma(V^*(s')+C)\right] \quad (I)$$

از طرفی برای خود V^* داریم:

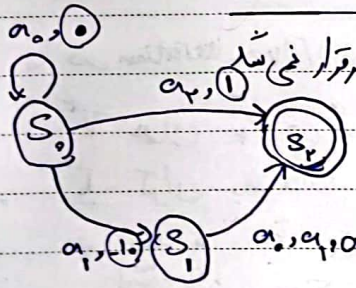
$$V^*(s) = \max_a E_{s'}\left[R(s,a,s') + \gamma V^*(s')\right] \quad (II)$$

با کم کردن (I) و (II) از هم نتیجه می‌شود:

$$C = r_0 + \gamma C \Rightarrow C = \frac{r_0}{1-\gamma} \Rightarrow \hat{V}^* = V^* + \frac{r_0}{1-\gamma}$$

همچنین می‌شد با استفاده از $\hat{V}_k^* = V_k^* + r_0\frac{1-\gamma^k}{1-\gamma}$ و قرار دادن $k\to\infty$ همین نتیجه را گرفت.

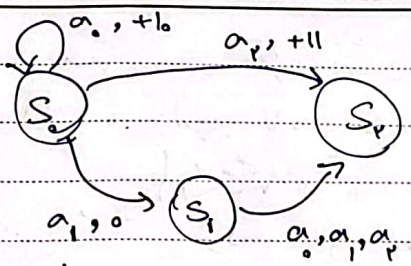
۵.۲.۱ اگر حالات terminate داشتیم دیگر رابطه $\hat{V}^* = V^* + \frac{r_0}{1-\gamma}$ لزوما برقرار نمی‌شد. به عنوان یک مثال نقض، به MDP روبرو توجه کنید. (s_2 حالت پایانی)



$$R(s_0,a_0)=1,\ R(s_0,a_1)=-10,\ R(s_0,a_2)=0$$
$$R(s_1,a_0)=-10$$

به وضوح در اینجا $\hat{V}^*(s_0) = 1$ و پالیسی بهینه در شروع انجام a_2 است.

Subject :
Year. Month. Date. ( )



حالا باید مقدار دهم $r_0 = 10$ و جوایز را زیاد کنیم:

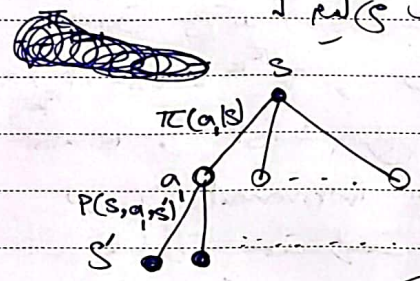
اکنون اینجا اگر a_2 را انجام دهیم جایزه ۱۱ را دریافت می کنیم و تمام می شود. اما اگر تا ابد a_0 را انجام دهیم، جایزه $10(1+\gamma+\gamma^2+\ldots) = 10 \times \frac{1}{1-\gamma}$ را دریافت می کنیم.

کافیست به جای ۱۰، عدد x را قرار می دادیم بطوریکه $x+1 < x \times \frac{1}{1-\gamma}$ یا معادلا $\gamma < \frac{1-\gamma}{\gamma}$ تا در حالت جدید اکشن بهینه a_0 باشد و $V^*(S_0) = x \times \frac{1}{1-\gamma}$.

۳.۱ (۶) توجه کنید در هرگام ما π_{k+1} را طوری قرار می دهیم که:

$$\pi_{k+1}(s) = \underset{a}{\operatorname{argmax}} \sum_{s'} P(s'|s,a)\left[R(s,a,s') + \gamma V^{\pi_k}(s')\right]$$

اکنون با کمک Backup diagram اثبات می کنیم که $V^{\pi_{k+1}}(s) \geq V^{\pi_k}(s)$.



حالا فرض کنید در لایه اول، به جای $\pi(\cdot|s)$، صرفا اکشن حریصانه را انتخاب کنیم. با اینکار $V(s)$ قطعا کم نمی شود (بیشتر مساوی می شود).

اکنون در لایه بعدی پالیسی را با حریصانه جایگزین کنید. $V(s')$ ها با اینکار بهتر می شود. بنابراین مقدار لایه بالاتر (که یک Convex Combination از $V(s')$هاست) هم کمتر نمی شود.

به همین شکل ادامه می دهیم و تا لایه ∞ جایگزین می کنیم. بنابراین پالیسی حریصانه مقدار Valueها را بهتر می کند و ثابت کردیم $V^{\pi_{k+1}}(s) \geq V^{\pi_k}(s)$.

در صورتی که برای تمام sها حالت مساوی رخ دهد، یعنی Converge کرده ایم و در غیر این صورت یعنی یک Value اکیدا بهتر شده.

(اگر $V^{\pi_{k+1}}(s) = V^{\pi_k}(s)$ یعنی از اینجا به بعد $V^{\pi_n}(s)$ ها ثابت خواهند ماند و می توانیم همان π_k را پالیسی پایانی بدانیم.)

۳.۱ (۷) با توجه به اینکه تعداد متناهی سیاست مختلف وجود دارد ($|A|^{|S|}$ حداکثر) و در هرگام سیاست بهتر می شود، قطعا همگرایی شود. اکنون توجه کنید در آن موقع داریم:

$$\pi_{k+1}(s) = \pi_k(s) \rightsquigarrow V^{\pi_{k+1}}(s) = \max_a \sum_{s'} [R(s,a) + \gamma V^{\pi_k}(s')] P(s,a,s')$$

و یعنی Vها در معادلات بهینگی بلمن صدق می کنند و طبق اثبات کلاس، بهینه هستند.

بنابراین π_k هم سیاست بهینه است.

Subject :
Year . Month . Date . ( )

۳.۱ ① در بخش های قبلی بیان دادیم که هردوی Value iteration و Policy iteration در هنگام همگرایی، دارای آ های هستند که در معادلهٔ بلمن صدق می کنند:

$$V^*(s) = \max_a \sum_{s'} P(s,a,s')\left[R(s,a) + \gamma V^*(s')\right]$$

در کلاس ثابت کردیم این معادلات یک جواب یکتا دارند (با کمک Banach fixed Point theorem) و بنابراین این دو، Value function هردو یک چیز هستند. سیاست های بهینه ممکن است متفاوت باشند، در صورتی که چند حالت با V یکسان داشته باشیم و همگی max باشند. اما در هر صورت مقدار Value فرقی نمی کند.

۳.۱ ② Policy iteration : فاز evaluation، $V^{\pi_k}_{i+1}(s) := \sum_{s'} P(s'|s,\pi_k(s))\left[R + \gamma V^{\pi_k}_i(s')\right]$ را تکرار می کنیم تا V^{π_k} همگرا شود. هر گام $O(|S|^2)$ طول می کشد و بطور متوسط به $O\left(\frac{\log(\frac{1}{\varepsilon})}{1-\gamma}\right)$ گام نیاز داریم. ($O(|S|^2 \frac{\log(\frac{1}{\varepsilon})}{1-\gamma})$ می تواند بسیار زیاد باشد.

راه دیگر حل با کمک Matrix inversion است. قرار است V^{π_k} را پیدا کنیم که در $V^{\pi_k} = T_{\pi_k} V^{\pi_k}$ صدق می کند (نقطه ثابت)، و $T_{\pi_k} V^{\pi_k} = r_{\pi_k} + \gamma P_{\pi_k} V^{\pi_k}$. بطور سراست داریم، $V^{\pi_k} = (I - \gamma P_{\pi_k})^{-1} r_{\pi_k}$، در $O(|S|^3)$ محاسبه می شود ($O(|S|^3)$ برای inverse گرفتن ماتریس $|S| \times |S|$، و $O(|S|^3)$ هم برای سایر عملیات)

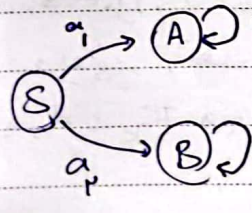
در فاز improvement نیز قرار می دهیم $\pi_{k+1}(s) := \arg\max_a \sum_{s'} P(s'|s,a)\left[R + \gamma V^{\pi_k}(s')\right]$ که محاسبهٔ عبارت سمت راست $O(|S||A|)$ طول می کشد و $|S|$ بار باید اینکار را انجام دهیم (تمام مقادیر s). پس $O(|S|^2|A|)$ طول می کشد.

پس هر iteration در کل $\{O(|S|^3) + O(|S|^2|A|)\}$ طول می کشد.

Value iteration : در هر گام قرار می دهیم $V^{k+1}(s) := \max_a \sum_{s'} P(s'|s,a)\left[R + \gamma V^k(s')\right]$ که یعنی برای تمام مقادیر s، $\{O(|S|^2|A|)\}$ طول می کشد.

پس هر iteration از روش Value iteration پیچیدگی محاسباتی کمتری دارد، اما در عوض نیاز به iteration های بیشتری برای همگرایی دارد، و در عمل روش Policy iteration سریعتر به پالیسی بهینه می رسد.

۳.۱ ⑩ در این حالت، دیگر Bellman Operator یک contraction mapping نیست، و اثبات همگرایی که در کلاس داشتیم صحیح نیست. در حقیقت اگر جمع ریوارد ها کران دار نباشد ($\sum_{i=0}^{\infty} R_i$) هیچ کدام از این دو الگوریتم تضمینی برای همگرایی ندارند.



حتی برای خود ماهم دشوار است در این حالت الگوریتم بهینه را پیدا کنیم. مثلاً در MDP روبرو، در هردو حالت return برابر ∞ خواهد بود و طبیعتاً اگر تمایلی به ریوارد کوتاه مدت نداشته باشیم ($\gamma = 1$) هیچ کدام را به دیگری ترجیح نمی دهیم.

P4PCO

Subject :
Year. Month. Date. ( )

۲. ۱. ① $\|B^\pi V - B^\pi V'\| = \max_s |B^\pi V(s) - B^\pi V(s)| = \gamma \max_s |\sum_{s'} P(s'|s,\pi(s))[V(s') - V'(s')]|$

نامساوی مثلث $\leq \gamma \max_s \sum_{s'} |P(s'|s,\pi(s))[V(s') - V'(s')]| = \gamma \max_s \sum_{s'} P(s'|s,\pi(s)) |V(s') - V'(s')|$

این عبارت یک Convex Combination از $|V(s') - V'(s')|$ هاست، بنابراین می‌توان نوشت:

$\gamma \max_s \sum_{s'} P(s'|s,\pi(s)) |V(s') - V'(s')| \leq \gamma \max_s \max_{s'} |V(s') - V'(s')|$

$= \gamma \max_s \|V - V'\| = \gamma \|V - V'\|$

$\Rightarrow \|B^\pi V - B^\pi V'\| \leq \gamma \|V - V'\|$ ■

---

۲. ۱. ② با توجه به سوال قبل که ثابت کردیم B^π یک Contraction mapping است، می‌توان از قضیه نقطه ثابت باناخ استفاده کرد و گفت B^π حتماً یک نقطه ثابت دارد.
همچنین می‌توان به سادگی با فرض خلف اثبات کرد.
فرض کنید V_1 و V_2 دو نقطه ثابت متفاوت باشند یعنی $B^\pi V_1 = V_1$ و $B^\pi V_2 = V_2$.
طبق نامساوی بخش قبل:

$\|V_1 - V_2\| = \|B^\pi V_1 - B^\pi V_2\| \leq \gamma \|V_1 - V_2\| \Rightarrow \gamma \geq 1$ ※

پس نقطه ثابت B^π یکتا است.

---

۲. ۱. ③ $B^\pi V'(s) - B^\pi V(s) = [r(s,\pi(s)) + \gamma \sum_{s'} P(s'|s,\pi(s)) V'(s')] - [r(s,\pi(s)) + \gamma \sum_{s'} P(s'|s,\pi(s)) \times V(s')]$

$= \underbrace{\gamma}_{>0} \sum_{s'} \underbrace{P(s'|s,\pi(s))}_{\geq 0} \times \underbrace{[V'(s') - V(s')]}_{\geq 0} \geq 0 \Rightarrow B^\pi V'(s) \geq B^\pi V(s) \quad \forall s \in S$

* توجه کنید اینکه $V'(s) - V(s) \geq 0$ است مستقیماً از فرض سوال نتیجه می‌شود.

---

۲. ۲. ④ فقط به ازای V^* (Optimal Value Function) می‌تواند ۰ شود:

$\|BV - V\| = 0 \iff BV - V = 0 \iff BV = V \iff V$ is fixed point of B

$\iff V = V^*$ (Since the fixed point is unique)

---

۲. ۲. ⑤ ابتدا طبق نامساوی مثلث داریم: (I) $\|V - V^\pi\| \leq \|V - B^\pi V\| + \|B^\pi V - V^\pi\|$

طبق نامساوی ثابت شده در سوال ۲، داریم: (II) $\|B^\pi V - B^\pi V^\pi\| \leq \gamma \|V - V^\pi\|$

با توجه به اینکه V^π نقطه ثابت B^π است داریم: $B^\pi V^\pi = V^\pi$ (III)

Subject :
Year . Month . Date . ( )

(II) , (III) $\Rightarrow \|B^\pi V - V^\pi\| \le \gamma \|V - V^\pi\|$

حالا اگر کران بدست آمده را در نامساوی (I) استفاده کنیم، نتیجه ی شود:

$$\|V - V^\pi\| \le \|V - B^\pi V\| + \gamma \|V - V^\pi\| \Rightarrow \|V - V^\pi\| \le \frac{\|V - B^\pi V\|}{1-\gamma}$$

اثبات نامساوی دوم نیز کاملا مشابه همین است! مراحل را بدون توضیحات می نویسم:

$$\|BV - BV^*\| \le \gamma \|V - V^*\| \xrightarrow{BV^* = V^*} \|BV - V^*\| \le \gamma \|V - V^*\| \quad (IV)$$

$$\|V - V^*\| \le \|V - BV\| + \|BV - V^*\| \overset{(IV)}{\le} \|V - BV\| + \gamma \|V - V^*\|$$

$$\Rightarrow \|V - V^*\| \le \frac{\|V - BV\|}{1-\gamma}$$

۲.۲.۶ توجه کنید اینکه π سیاست حریصانه استخراج شده از V است، به ما این نتیجه را می دهد که $BV = B^\pi V$. با نگاه کردن به روابط $BV(s)$ و $B^\pi V(s)$ این نتیجه به سادگی قابل دریافت است: $\pi(s) = \arg\max_a \left[r(s,a) + \gamma \sum_{s'} P(s'|s,a) V(s') \right]$

حالا به کمک این نامساوی های سوال ۵ را بازنویسی می کنیم:

$$\|V - V^\pi\| \le \frac{\|V - B^\pi V\|}{1-\gamma} = \frac{\|V - BV\|}{1-\gamma} \le \frac{\varepsilon}{1-\gamma}$$

$$\|V - V^*\| \le \frac{\|V - BV\|}{1-\gamma} \le \frac{\varepsilon}{1-\gamma}$$

$$\|V^* - V^\pi\| \le \|V - V^\pi\| + \|V^* - V\| \le \frac{2\varepsilon}{1-\gamma}$$ (نامساوی مثلث)

توجه کنید می دانیم $V^\pi \le V^*$ و بنابراین $|V^*(s) - V^\pi(s)| \le \|V^* - V^\pi\|$.

پس ثابت شد:

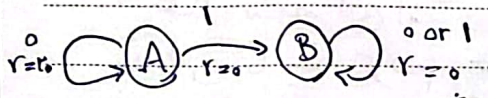
$$\forall s : V^*(s) - V^\pi(s) \le \frac{2\varepsilon}{1-\gamma} \Rightarrow V^\pi(s) \ge V^*(s) - \frac{2\varepsilon}{1-\gamma}$$

۲.۲.۷ یک کران پایین روی V^π می تواند تضمین کند که سیاست ما از یک حدی بدتر نیست، و در مسائلی مانند robotics (با فرض کنید یک هلیکوپتر پرنده!) که سیاست خیلی بد می تواند باعث آسیب جدی به تجهیزات ما شود (مثلا ربات سقوط کند و خراب شود!) داشتن کران پایین روی $V^\pi(s)$ می تواند یک راه برای تضمین آسیب ندیدن وسیله باشد. مثلا یک threshold تعیین می کنیم و اگر $V^\pi(s)$ از آن کران پایین بهتری داشت می دانیم اطمینان حاصل کنیم با سیاست فعلی رفتن به حالت s خطرناک نیست.

Subject :
Year. Month. Date. ( )

۲.۲ ۸) خیر، لزومی ندارد. فقط می توان نتیجه گرفت کران پایین هایی که برای $V^{\pi}(s)$ و $V^{*}(s)$ داریم برابر هستند، اما لزومی ندارد برابر باشند.

در ادامه یک مثال ساده می زنیم که ادعای سوال را نقض می کند:



همچنین در این مثال $\gamma = 0.9$ در نظر بگیرید.

فرض کنید $V(A)=20$, $V(B)=0$. در این حالت خواهیم داشت $\|BV-V\| = 1$

همچنین اگر $V'(A)=0$, $V'(B)=1$، بازهم خواهیم داشت $\|BV'-V'\| = 1$.

اما در حالت اول $\pi(A)=0$ خواهد بود، در حالت دوم $\pi'(A)=1$، و بنابراین خواهیم داشت $V^{\pi}(A)=10$ اما $V^{\pi'}(A)=0$ و برابر نیستند. این مثال هم نشان می دهد لزوما ادعا برقرار نیست.

۲.۲ ۹) در ابتدا همانطور که در سوال ۶ گفتیم می دانیم $BV = \hat{B}^{\pi}V$، و همچنین نامساوی اول سوال ۵ به شکل زیر در می آید:

$$\|V-V^{\pi}\| \leq \frac{\varepsilon}{1-\gamma} \quad (I)$$

حال ادعا می کنیم $\|V^{*}-V^{\pi}\| \leq \|V-V^{\pi}\|$.

توجه کنید در صورت اثبات این ادعا، نتیجه می شود (طبق (I)) که $\|V^{*}-V^{\pi}\| \leq \frac{\varepsilon}{1-\gamma}$

و بنابراین نامساوی خواسته شدهٔ سوال ثابت می شود: $V^{\pi}(s) \geq V^{*}(s) - \frac{\varepsilon}{1-\gamma}$

اکنون به اثبات ادعایمان می پردازیم. توجه کنید $V^{\pi} \leq V^{*}$ (چون $V^{*} = \max_{\pi} V^{\pi}$ است) و بنابراین اکنون داریم $V \geq V^{\pi} \leq V^{*}$. حالا توجه کنید:

$$\|V-V^{\pi}\| = \max_{s} |V(s)-V^{\pi}(s)| = \max_{s} V(s)-V^{\pi}(s)$$
$$\|V^{*}-V^{\pi}\| = \max_{s} |V^{*}(s)-V^{\pi}(s)| = \max_{s} V^{*}(s)-V^{\pi}(s)$$

با توجه به اینکه می دانیم $\forall s: V^{*}(s) \leq V(s)$، به وضوح نتیجه می شود:

$$\max_{s} V^{*}(s)-V^{\pi}(s) \leq \max_{s} V(s)-V^{\pi}(s) \Rightarrow \|V^{*}-V^{\pi}\| \leq \|V-V^{\pi}\|$$

بنابراین ادعایمان ثابت شد، و کران بهتر هم نتیجه می شود.

۲.۲ ۱۰) ابتدا به استقرا روی n ثابت می کنیم $\hat{B}^{n}V \leq V$. حالت $n=1$ به عنوان فرض داده شده. گام استقرا:

$$\left.\begin{array}{l} \hat{B}^{n}V \leq V \\ V \leq V' \Rightarrow BV \leq BV' \text{ : سوال ۳} \end{array}\right\} \Rightarrow B(\hat{B}^{n}V) \leq BV \Rightarrow \hat{B}^{n+1}V \leq BV \leq B$$

استقرا ثابت شد.

حالا توجه کنید $\lim_{n\to\infty} \hat{B}^{n}V = V^{*}$. از ادعایی که اثبات کردیم هم استفاده می کنیم:

$$V^{*} = \lim_{n\to\infty} \hat{B}^{n}V \leq V \Rightarrow V^{*} \leq V$$

Subject :
Year . Month . Date . ( )

(توجه کنید در سوال ۳ داشتیم اگر $V \le V'$، آنگاه $B^{\pi}V \le B^{\pi}V'$ اما اگر قرار دهیم $\pi = \pi^*_V$، آنگاه $BV = B^{\pi^*_V}V \le B^{\pi^*_{V'}}V' \le BV'$ که با لیسی بهینه با مقادیر V است، نتیجه‌ی نمی‌شود که همان نتیجه‌ای است که ما استفاده کردیم).

بنابراین ثابت کردیم $BV \le V$ یک شرط کافی برای $V^* \le V$ است و به وضوح چون دیگر نیاز به دانستن V^* نداریم، BV برآوردی قابل محاسبه است، چک کردن این شرط برایمان بسیار راحت‌تر است.

۲.۲.۱۱ در سوال ۵ نتیجه زیر را گرفتیم:

$$\|B^{\pi}V - V^{\pi}\| \le \gamma \|V - V^{\pi}\| \quad (I)$$

همچنین با توجه به اینکه $BV = B^{\pi}V$، در سوال ۹ گفتیم: (نتیجه سوال ۵)

$$\|V - V^{\pi}\| \le \frac{\varepsilon}{1-\gamma} \quad (II)$$

$$(I), (II) \Rightarrow \|B^{\pi}V - V^{\pi}\| \le \frac{\gamma\varepsilon}{1-\gamma} \quad (\star)$$

اکنون ادعا می‌کنیم $\|V^* - B^{\pi}V\| \le \frac{\gamma\varepsilon}{1-\gamma}$. توجه کنید درصورت اثبات این ادعا، با نامساوی مثلث سایر دلخواهمان را می‌گیریم:

$$\|V^* - V^{\pi}\| \le \|V^* - B^{\pi}V\| + \|B^{\pi}V - V^{\pi}\| \le \frac{2\gamma\varepsilon}{1-\gamma}$$

که مستقیما نامساوی اول را نتیجه می‌دهد. اکنون به اثبات ادعایمان می‌پردازیم.

در سوال ۵ اگر به‌جای V مقدار BV را قرار دهید نتیجه می‌گیریم: (نامساوی سوم)

$$\|BV - V^*\| \le \frac{\|BV - BBV\|}{1-\gamma} \quad (III)$$

$$\text{Contraction Property} \rightsquigarrow \|BV - B^2V\| \le \gamma\|V - BV\| = \gamma\varepsilon \quad (IV)$$

$$BV = B^{\pi}V \Rightarrow \|V^* - B^{\pi}V\| = \|BV - V^*\| \quad (V)$$

$$(III), (IV), (V) \Rightarrow \|V^* - B^{\pi}V\| \le \frac{\gamma\varepsilon}{1-\gamma} \rightarrow \text{ادعا ثابت شد}$$

نامساوی دوم نیز راحت است. توجه کنید $V \ge V^*$ نتیجه می‌دهد $BV \ge BV^* = V^*$.

پس:

$$V^* - V^{\pi} \le BV - V^{\pi} = B^{\pi}V - V^{\pi} \overset{(\star)}{\le} \frac{\gamma\varepsilon}{1-\gamma}$$

و نامساوی دوم هم اثبات شد. ■