مینی‌پروژه پنچم - تئوری

محمدجواد هزاره — شماره دانشجویی: ۹۸۱۰۱۰۷۴

سوال ۱

الف) برای یک استیت مشخص مانند s یکی از مقادیر $B_k v_1(s)$ یا $B_k v_2(s)$ بزرگ‌تر مساوی دیگری است. فرض کنیم این مقدار $B_k v_1(s)$ باشد، بنابراین می‌توان نوشت $B_k v_1(s) - B_k v_2(s) \geq 0$. حال فرض کنیم اکشن بهینه برای استیت s اکشن a^* باشد. بنابراین:

$$B_k v_1(s) = R(s, a^*) + \gamma_k \sum_{s' \in \mathcal{S}} p(s'|s, a^*) v_1(s')$$

بنابراین داریم:

$$B_k v_2(s) = \max_a \left[R(s, a) + \gamma_k \sum_{s' \in \mathcal{S}} p(s'|s, a) v_2(s') \right] \geq \left[R(s, a^*) + \gamma_k \sum_{s' \in \mathcal{S}} p(s'|s, a^*) v_2(s') \right]$$

چرا که بیشنه‌ی یک تابع حداقل به‌اندازه‌ی مقدار آن در نقطه‌ای دلخواه است. بنابراین:

$$\begin{array}{c} B_k v_1(s) = R(s, a^*) + \gamma_k \sum_{s' \in \mathcal{S}} p(s'|s, a^*) v_1(s') \\ -B_k v_2(s) \leq - \left[R(s, a^*) + \gamma_k \sum_{s' \in \mathcal{S}} p(s'|s, a^*) v_2(s') \right] \\ \hline B_k v_1(s) - B_k v_2(s) \leq \gamma_k \sum_{s' \in \mathcal{S}} p(s'|s, a^*) \left(v_1(s') - v_2(s') \right) \end{array}$$

سمت راست میانگین وزن‌دار تعدادی عبارت است که می‌دانیم این میانگین از قدرمطلق بیشنه‌ی آن‌ها کم‌تر خواهد بود. بنابراین:

$$B_k v_1(s) - B_k v_2(s) \leq \gamma_k \sum_{s' \in \mathcal{S}} p(s'|s, a^*) \left(v_1(s') - v_2(s') \right) \leq \gamma_k \|v_1 - v_2\|_\infty$$

از طرفی این رابطه برای تمام sها برقرار است، پس برای sای که سمت چپ را بیشینه کند نیز برقرار خواهد بود. هم‌چنین با توجه به آن‌چه در بند اول گفته شد این مقدار مثبت خواهد بود. (اگر مثبت نباشد ممکن بود عدد منفی بزرگی باشد که قدرمطلق آن از سمت راست بیش‌تر می‌شود.) بنابراین:

$$\left.\begin{aligned} &\max_s \big(B_k v_1(s) - B_k v_2(s)\big) \leq \gamma_k \|v_1 - v_2\|_\infty \\ &0 \leq \max_s B_k v_1(s) - B_k v_2(s) \end{aligned}\right\} \implies \|B_k v_1 - B_k v_2\|_\infty \leq \gamma_k \|v_1 - v_2\|_\infty$$

در انتخاب v_1 و v_2 فرض خاصی نکرده بودیم پس استدلال بالا برای تمام هر v_1 و v_2 برقرار است.

ب) با جایگذاری $\gamma_k = \frac{k}{k+1}$ خواهیم داشت:

$$\prod_{k=1}^{K} \gamma_k = \frac{1}{2} \times \frac{2}{3} \times \cdots \times \frac{K-1}{K} \times \frac{K}{K+1} = \frac{1}{K+1} \leq \frac{1}{K+1}$$

ج) اگر برای دو بردار v_1 و v_2 به ترتیب K مرحله عملگر B_k را اعمال کنیم، با توجه به انقباضی بودن تمام B_kها خواهیم داشت:

$$\begin{aligned} 0 \leq \|B_K \cdots B_1 v_1 - B_K \cdots B_1 v_2\|_\infty &\leq \gamma_K \|B_{K-1} \cdots B_1 v_1 - B_{K-1} \cdots B_1 v_2\|_\infty \\ &\leq \gamma_K \gamma_{K-1} \|B_{K-2} \cdots B_1 v_1 - B_{K-2} \cdots B_1 v_2\|_\infty \\ &\leq \qquad \vdots \\ &\leq \gamma_K \cdots \gamma_1 \|v_1 - v_2\|_\infty \\ &\leq \frac{1}{1+K} \|v_1 - v_2\|_\infty \end{aligned}$$

هم‌چنین برای بردار $\overrightarrow{0}$ داریم:

$$B_K \cdots B_1 \overrightarrow{0} = \gamma_K R_K + \gamma_K \gamma_{K-1} R_{K-1} + \cdots + \gamma_1 R_1 + R_0$$

که اگر مقدار پاداش‌ها متناهی باشد، با $K \to \infty$ عبارت بالا به مقداری متناهی هم‌گرا شده که آن را با v^* نشان می‌دهیم که این مقدار همان مطلوبیت استیت‌ها خواهد بود. حال اگر در رابطه‌ی بالا بردار v_2 را همان بردار $\overrightarrow{0}$ در

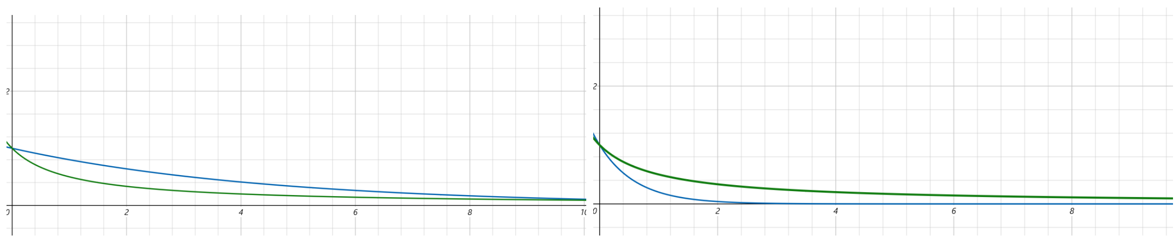
نظر بگیریم، در حد $K \to \infty$ خواهیم داشت:

$$\left.\begin{array}{r} 0 \le \lim\limits_{K\to\infty} \|B_K \cdots B_1 v_1 - v^*\|_\infty \\ \lim\limits_{K\to\infty} \|B_K \cdots B_1 v_1 - v^*\|_\infty \le \lim\limits_{K\to\infty} \dfrac{1}{1+K} \|v_1 - \overrightarrow{0}\|_\infty = 0 \end{array}\right\} \implies \lim_{K\to\infty} \|B_K \cdots B_1 v_1 - v^*\|_\infty = 0$$

$$\implies \lim_{K\to\infty} B_K \cdots B_1 v_1 = v^*$$

که در استدلال بالا فرض خاصی برای v_1 نکردیم، بنابراین استدلال برای هر v_1ای برقرار است.

د) به مقدار γ در روش ضریب ثابت بستگی دارد. اگر γ به‌اندازه‌ی کافی کوچک باشد روش ضریب ثابت بهتر عمل می‌کند و سریع‌تر به نزدیکی صفر می‌رسد. اما اگر γ به‌اندازه‌ی کافی بزرگ باشد استفاده کردن از روش جدید مناسب‌تر است چرا که سریع‌تر به صفر نزدیک می‌شود. به نمودارهای زیر می‌توان دقت کرد.



شکل ۱: نمودار آبی مربوط به روش ضریب ثابت و نمودار سبز مربوط به روش ضریب متغیر است. (در سمت راستی $\gamma = 0.2$ و در سمت چپی $\gamma = 0.8$)

سوال ۲

الف) با توجه به تصادفی بودن سیاست داده شده، برای محاسبه‌ی v_Π باید احتمال انجام شدن یک اکشن را در Q آن ضرب کرده و روی اکشن‌های مختلف جمع ببندیم. بنابراین:

$$v_\Pi(s) = \sum_a \Pi_{s,a} Q(s,a)$$

برای محاسبه‌ی Q نیز باید روی استیت‌های مختلف جمع بزنیم:

$$\begin{aligned} Q(s,a) &= \sum_{s'} p(s'|s,a) \left[W_{s,s'} + \gamma v_\Pi(s')\right] \\ &= \eta\big[W_{s,a} + \gamma v_\Pi(a)\big] + \frac{1-\eta}{d_s} \sum_{i \in N(s)} \big[W_{s,i} + \gamma v_\Pi(i)\big] \end{aligned}$$

که $N(s)$ همسایه‌های s هستند. بنابراین برای v_Π می‌توان نوشت:

$$v_\Pi(s) = \sum_a \Pi_{s,a} \left[\big[W_{s,a} + \gamma v_\Pi(a)\big] + \frac{1-\eta}{d_s} \sum_{i \in N(s)} \big[W_{s,i} + \gamma v_\Pi(i)\big] \right]$$

برای محاسبه‌ی v^* بهتر است از تصمیمات قطعی استفاده کنیم چرا که بیشنه‌ی Qها از هر میانگین وزن‌دار آن‌ها بیش‌تر خواهد بود. بنابراین در یک استیت s بهتر است به استیتی برویم که $Q(s,a)$ را بیشینه می‌کند. بنابراین روی اکشن‌های مختلفی که از s می‌توان انجام داد بیشینه می‌گیریم:

$$v^*(s) = \max_a \left[\big[W_{s,a} + \gamma v^*(a)\big] + \frac{1-\eta}{d_s} \sum_{i \in N(s)} \big[W_{s,i} + \gamma v^*(i)\big] \right]$$

ب) اگر $\eta = 1$ باشد داریم:

$$v_\Pi(s) = \sum_a \Pi_{s,a} \big[W_{s,a} + \gamma v_\Pi(a)\big] = \sum_a \Pi_{s,a} W_{s,a} + \gamma \sum_a \Pi_{s,a} v_\Pi(a)$$

جمله‌ی دوم به شکل برداری برابر $\gamma \Pi v_\Pi$ خواهد بود. جمله‌ی اول نیز به‌ازای یک s مشخص، جمع سطر sام از ماتریس $\Pi \odot W$ خواهد بود، بنابراین به شکل ماتریسی به فرم $(\Pi \odot W)\overrightarrow{1}$ خواهد بود. بنابراین برای رابطه‌ی

بلمن به شکل ماتریسی داریم:

$$v_\Pi = \gamma \Pi v_\Pi + (\Pi \odot W)\overrightarrow{1}$$
$$\implies \boxed{[I - \gamma\Pi]v_\Pi = (\Pi \odot W)\overrightarrow{1}}$$

ج) با توجه به قسمت قبل می‌توان نوشت:

$$v_\Pi = [I - \gamma\Pi]^{-1}(\Pi \odot W)\overrightarrow{1}$$

اما برای پیدا کردن بهترین سیاست باید Π را به‌گونه‌ای انتخاب کنیم که تک تک درایه‌های v_Π بیش‌ترین مقدار خود را داشته باشند. از طرفی مشخصا نمی‌خواهیم در سیاست خود برای رفتن از راس i به راس j که یالی به یکدیگر ندارند احتمالی در نظر بگیریم. به عبارتی در سیاست ما باید $\Pi_{i,j} = 0$ باشد اگر i به j یالی نداشته باشد. هم‌چنین به‌ازای هر راس مثل i، جمع احتمال تصمیم‌هایی که می‌گیریم باید برابر یک شود که این شرط نیز به شکل $\Pi\overrightarrow{1} = \overrightarrow{1}$ ظاهر می‌شود.