به نام خدا

# یادگیری تقویتی

## تمرین تئوری اول


یلدا شعبان زاده

۹۸۱۰۱۸۲۲


نیمسال دوم ۰۲-۰۱

# فهرست

سوال ۱ ........................................................................................................................................ ۳

سوال ۲ ........................................................................................................................................ ۸

سوال ۳ ........................................................................................................................................۱۱

سوال ۴ ........................................................................................................................................۱۵

## سوال ۱

(آ) می‌دانیم value iteration به فرم زیر است:

$$V_k^*(s) = \max_a \sum_{s'} P(s'|s,a)(R(s,a,s') + \gamma V_{k-1}^*(s'))$$

همچنین فرض کرده‌ایم که:

$$0 \leq R(s,a,s') \leq R_{max} \quad \forall s,a$$

و می‌دانیم:

$$0 \leq P(s'|s,a) \leq 1 \quad \forall s,a$$

حال می‌توان گفت:

$$\begin{aligned} V_k^*(s) = \max_a \sum_{s'} P(s'|s,a)\big(R(s,a,s') + \gamma V_{k-1}^*(s')\big) &\leq \sum_{s',a} P(s'|s,a)\big(R(s,a,s') + \gamma V_{k-1}^*(s')\big) \\ \leq \sum_{s',a} R(s,a,s') + \gamma V_{k-1}^*(s') &\leq R_{max} + \gamma \sum_{s'',a} R(s',a,s'') + \gamma V_{k-2}^*(s'') \\ \leq R_{max} + \gamma R_{max} + \gamma^2 R_{max} + \cdots + \gamma^{k-1} R_{max} &\leq \frac{R_{max} \times (1-\gamma^k)}{1-\gamma} \leq \frac{R_{max}}{1-\gamma} \end{aligned}$$

بنابراین به یک upper bound رسیدیم.

***

(ب) برای اینکه ثابت کنیم $V_k^*$ نسبت به k صعودی است؛ ابتدا $V_{k+1}^\pi$ را پیدا می‌کنیم به طوری که $V_{k+1}^\pi \geq V_k^*$ باشد.

برای این کار ابتدا از bellman optimality equation استفاده می‌کنیم.

$$V_{k+1}^*(s) = \max_a \sum_{s'} P(s'|s,a)(R(s,a,s') + \gamma V_k^*(s'))$$

از آنجا که policy بهینه همان $\pi^*$ است، می‌توان گفت که:

$$V_{k+1}^* \geq V_{k+1}^\pi \quad (1)$$

حال، فرض کنید $\pi$ یک policy دلخواه باشد که k تا step با policy $\pi^*$ برویم و در گام k+1ام policy $\pi$ را انتخاب کرده و با آن گام برداریم. از آنجا که rewardها نامنفی هستند داریم:

$$V_{k+1}^\pi \geq V_k^* \quad (2)$$

پس می‌توان گفت که:

$$(1,2) \Rightarrow V_{k+1}^* \geq V_k^*$$

حال، برای همگرایی می‌توان گفت از آنجایی که $V_k^*$ صعودی است و ما یک upper bound برایش بدست آوردیم، داریم:

$$V_0^* \leq \cdots \leq V_k^* \leq V_{k+1}^* \leq \cdots \leq \frac{R_{max}}{1-\gamma}$$
$$\Rightarrow \lim_{k\to\infty} V_k^* = \frac{R_{max}}{1-\gamma}$$

از آنجایی که این حد محدود است؛ الگوریتم ما همگرا خواهد بود.

***

(ج) طبق معادله بلمن داریم:

$$V_k^*(s) = \max_a \sum_{s'} P(s'|s,a)\big(R(s,a,s') + \gamma V_{k-1}^*(s')\big)$$
$$\Rightarrow \lim_{k\to\infty} V_k^*(s) = \lim_{k\to\infty} \max_a \sum_{s'} P(s'|s,a)\big(R(s,a,s') + \gamma V_{k-1}^*(s')\big)$$
$$= \max_a \sum_{s'} P(s'|s,a)\left(R(s,a,s') + \gamma \lim_{k\to\infty} V_{k-1}^*(s')\right) \qquad \text{پیوستگی}$$
$$\lim_{k\to\infty} V_k^*(s) = V^*(s) \xRightarrow{k=k-1=\infty} V^*(s) = \max_a \sum_{s'} P(s'|s,a)(R(s,a,s') + \gamma V^*(s))$$

پس در هر مرحله V افزایش پیدا می‌کند تا هنگامی که در بی‌نهایت convergence رخ می‌دهد و مقدارش برای آن state عوض نمی‌شود.
همچنین در حالت convergence بهینه است زیرا بنا بر تعریف در حالت convergence داریم $\pi_{k+1}(s) = \pi_k(s)$. یعنی:

$$V^{\pi_k}(s) = \max_a \sum_{s'} P(s'|s,a)\big(R(s,a,s') + \gamma V^{\pi_k}(s')\big)$$

از آنجایی که $V^{\pi_k}$ در معادله بلمن صدق می‌کند، می‌توان گفت $V^{\pi_k}$ برابر مقدار value function آپتیمال یا $V^*$ است.

***

(د) در MDP جدید فرض نامنفی بودن پاداش‌ها را نداریم. اما می‌توان با اضافه کردن اندازه کمترین پاداش به همه پاداش‌ها باند reward ها را نامنفی کرد. یعنی داریم:

$$0 \leq R'(s,a,s') \leq R_{max} + |R_{min}| \quad \forall s,a$$

حال طبق معادله بلمن داریم:

$$
\begin{aligned}
V'^*_k(s) &= \max_a \sum_{s'} P(s'|s,a)\big(R'(s,a,s') + \gamma V'^*_{k-1}(s')\big) \\
&= \max_a \sum_{s'} P(s'|s,a)\big(R(s,a,s') + |R_{min}| + \gamma V'^*_{k-1}(s')\big) \\
&= \max_a \left( \sum_{s'} P(s'|s,a)\left(R(s,a,s') + \gamma V'^*_{k-1}(s')\right) + \sum_{s'} P(s'|s,a)|R_{min}| \right) \\
&= \max_a \left( \left( \sum_{s'} P(s'|s,a)\left(R(s,a,s') + \gamma V'^*_{k-1}(s')\right)\right) + |R_{min}| \times 1 \right) \\
&= \max_a \left( \sum_{s'} P(s'|s,a)\left(R(s,a,s') + \gamma V'^*_{k-1}(s')\right)\right) + |R_{min}|
\end{aligned}
$$

از آنجایی که داریم $V'^*_0(s) = V^*_0(s)$ و $V^*_1(s) = \max_a \sum_{s'} P(s'|s,a)\big(R(s,a,s') + \gamma V^*_0(s')\big)$ می‌توان گفت:

$$
V'^*_1(s) = \max_a \left( \sum_{s'} P(s'|s,a)\big(R(s,a,s') + \gamma V^*_0(s')\big)\right) + |R_{min}| = V^*_1(s) + |R_{min}|
$$

$$
\begin{aligned}
V'^*_2(s) &= \max_a \left( \sum_{s'} P(s'|s,a)\big(R(s,a,s') + \gamma V'^*_1(s')\big)\right) + |R_{min}| \\
&= \max_a \left( \sum_{s'} P(s'|s,a)(R(s,a,s') + \gamma(V^*_1(s) + |R_{min}|))\right) + |R_{min}| \\
&= \max_a \left( \sum_{s'} P(s'|s,a)\big(R(s,a,s') + \gamma V^*_1(s)\big) + \gamma|R_{min}|\sum_{s'} P(s'|s,a)\right) + |R_{min}| \\
&= \max_a \left( \sum_{s'} P(s'|s,a)\big(R(s,a,s') + \gamma V^*_1(s)\big)\right) + \gamma|R_{min}| + |R_{min}|
\end{aligned}
$$

به همین صورت می‌توان به صورت افزایشی نشان داد که داریم:

$$
\begin{aligned}
V'^*_k(s) &= \max_a \sum_{s'} P(s'|s,a)\big(R'(s,a,s') + \gamma V'^*_{k-1}(s')\big) \\
&= \max_a \left( \sum_{s'} P(s'|s,a)\big(R(s,a,s') + \gamma V^*_{k-1}(s)\big)\right) + \sum_{i=0}^{k-1} \gamma^i |R_{min}| \\
&= V^*_{k-1}(s) + \sum_{i=0}^{k-1} \gamma^i |R_{min}| = V^*_{k-1}(s) + \sum_{i=0}^{k-1} \gamma^i r_0
\end{aligned}
$$

در حالت convergence نیز با شرط نداشتن terminating state داریم:

$$V'^{*}(s) = \lim_{k\to\infty} V'^{*}_{k}(s) = V^{*}(s) + \frac{r_0}{1-\gamma}$$

Policy بهینه قبلی در convergence برابر بود با: $\pi^{*}(s) = \arg\max_{\mathrm{a}} \sum_{s'} P(s'|s,a)\big(R(s,a,s') + \gamma V^{*}(s')\big)$

در این حالت نیز داریم:

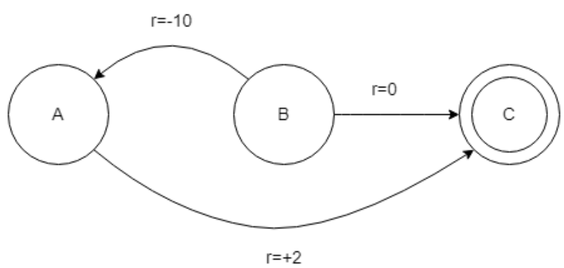
$$\begin{aligned}
\pi'^{*}(s) &= \arg\max_{\mathrm{a}} \sum_{s'} P(s'|s,a)\left(R(s,a,s') + \gamma V'^{*}(s')\right) \\
&= \arg\max_{\mathrm{a}} \sum_{s'} P(s'|s,a)\left(R(s,a,s') + \gamma (V^{*}(s) + \frac{r_0}{1-\gamma})\right) \\
&= \arg\max_{\mathrm{a}} \sum_{s'} P(s'|s,a)\left(R(s,a,s') + \gamma\big(V^{*}(s)\big)\right) + \frac{\gamma r_0}{1-\gamma} \\
&= \arg\max_{\mathrm{a}} \sum_{s'} P(s'|s,a)\left(R(s,a,s') + \gamma\big(V^{*}(s)\big)\right) = \pi^{*}(s)
\end{aligned}$$

***

ه) شرط نداشتن terminating state به این دلیل لازم است که در عبارت بالا بتوانیم k را به سمت بی‌نهایت میل دهیم و به عبارتی که به k وابسته نیست برسیم تا بتوانیم policy بهینه را بیابیم.

برای مثال نقض می‌توان گفت:

یک MDP را با سه حالت در نظر بگیرید: A و B و C. حالت C حالت پایانی است. حالت B نیز حالت شروع است.




در اینجا با شروع از B، دو اکشن چپ و راست دارد که احتمال هرکدام 0.5 است. در صورت انتخاب راست rewardهایش 0 است و در صورت حرکت به سمت چپ به A می‌رود که reward=-10 و سپس تنها اکشنش که حرکت به سمت C باشد را انجام می‌دهد و reward=+2 را می‌گیرد. پس rewardهایش برابر 8- است. پس policy بهینه با شروع از B حرکت به راست و دریافت reward ۰ است.

$$V^{*}(A) = \max_{\mathrm{a}}\Big(1 \times \big(2 + 1 \times V^{*}(C)\big)\Big) = 2 \Rightarrow \pi^{*}(A) = right$$

$$V^{*}(B) = \max_{\mathrm{a}}\left(\frac{1}{2} \times \big(-10 + 1 \times V^{*}(A)\big), \frac{1}{2} \times \big(0 + 1 \times V^{*}(C)\big)\right) = \max_{\mathrm{a}}(-4,0) = 0 \Rightarrow \pi^{*}(B) = right$$

$$V^{*}(C) = 0$$

اما اگر $r_0 = |R_{min}| = 10$ را به همه پاداش‌ها اضافه کنیم، سیاست بهینه تغییر می کند. اکنون، اگر از B به سمت چپ حرکت و سپس به C برویم، پاداش 12 میگیریم که با policy بهینه در مرحله قبل متفاوت است.

$$V^*(A) = \max_{a}\left(1 \times \left(12 + 1 \times V^*(C)\right)\right) = 12 \Rightarrow \pi^*(A) = right$$
$$V^*(B) = \max_{a}\left(\frac{1}{2} \times \left(0 + 1 \times V^*(A)\right), \frac{1}{2} \times \left(10 + 1 \times V^*(C)\right)\right) = \max_{a}(6,5) = 6 \Rightarrow \pi^*(B) = left$$
$$V^*(C) = 0$$

بنابراین policy بهینه در B با policy بهینه در مرحله قبل تفاوت دارد.

## سوال ۲

(آ) فرض کنید $V^{\pi_t}(s)$ و $V^{\pi_{t+1}}(s)$ به ترتیب توابع مقدار حالت برای سیاست های $\pi_t$ و $\pi_{t+1}$ باشند و فرض کنید $\pi^*$ پالیسی بهینه باشد. فرض شده که $V_\infty^{\pi_{t+1}}(s) = V^{\pi_{t+1}}(s) = V_\infty^{\pi_t}(s) = V^{\pi_t}(s)$ حال اثبات می‌کنیم $\pi_{t+1} = \pi^*$

برای تناقض فرض کنید $\pi_{t+1}$ برابر $\pi^*$ نیست. سپس یک حالت $s$ وجود دارد که $\pi_{t+1}(s) \neq \pi^*(s)$

همچنین داریم:

$$Q^{\pi_t}\big(s,\pi^*(s)\big) = R\big(s,\pi^*(s)\big) + \gamma \sum_{s'} P(s'|s,\pi^t(s))V^{\pi_t}(s')$$
$$Q^{(\pi_{t+1})(s,\pi^{(t+1)(s)})} = R\big(s,\pi^{(t+1)}(s)\big) + \gamma \sum_{s'} P\left(s'\middle|s,\pi^{(t+1)}(s)\right) V^{(\pi_{t+1})}(s')$$

با فرض، $V_\infty^{\pi_{t+1}}(s) = V_\infty^{\pi_t}(s)$ بنابراین دو مقدار Q برابر هستند. از آنجایی که $\pi_t$ و $\pi_{t+1}$ برابر نیستند، داریم:

$$R(s,\pi^*(s)) < R(s,\pi^{t+1}(s))$$

بنابراین، می‌توانیم معادله فوق را به صورت زیر بازنویسی کنیم:

$$V^{\pi_t}(s) + \varepsilon < V^{\pi_{t+1}}(s)$$

بنابراین به تناقض رسیدیم و نتیجه می‌گیریم اگر تابع ارزش-حالت برای یک policy بین دو تکرار تغییر نکند، به پالیسی بهینه رسیده‌ایم.

***

(ب) می‌توان تعداد مراحل policy evaluation را حداکثر برابر با $|A|^{|S|}$ در نظر گرفت که در آن $|S|$ تعداد حالت ها و $|A|$ تعداد actionها است. زیرا در نظر بگیرید که هر مرحله ارزیابی تابع مقدار را برای هر حالت با استفاده از معادله بلمن محاسبه می کند:

$$V_{k+1}^{\pi}(s) = \sum_{s'} P(s'|s,\pi(s))(R(s,\pi(s),s') + \gamma V_k^{\pi}(s'))$$

از آنجا که $\pi$ از فضای stateها به actionها است و صعودی است می‌توان گفت که $|A|^{|S|}$ policy داریم پس این الگوریتم تمام می‌شود. زیرا با توجه به قسمت بعد نیز درصورت برابر شدن دو V به پالیسی بهینه رسیده‌ایم و در لوپ نمی‌افتیم.

برای اثبات اینکه policy iteration در نهایت خاتمه می یابد و به مقدار بهینه همگرا می شود، می توانیم از استدلال زیر استفاده کنیم:

policy iteration شامل دو مرحله است: policy evaluation و policy improvement. در مرحله policy evaluation، value function را تا زمانی که به تابع ارزش واقعی policy فعلی همگرا شود، به روز می کنیم. در مرحله policy improvement، policy جدیدی را انتخاب می کنیم که با توجه به value function فعلی انتخاب شده است.

از آنجایی که value function تضمین شده است که در تعداد محدودی از تکرارها به تابع ارزش واقعی برای policy فعلی همگرا شود، مرحله ارزیابی policy در نهایت خاتمه می یابد. علاوه بر این، سیاست جدید انتخاب شده در مرحله بهبود policy تضمین شده است که بهتر یا برابر با policy فعلی باشد. اگر policy جدید کاملاً بهتر از policy فعلی باشد، مرحله ارزیابی policy را با policy جدید تکرار می کنیم تا زمانی که تابع ارزش به تابع ارزش واقعی policy جدید همگرا شود.

از آنجایی که فقط تعداد محدودی از policy ها وجود دارد، و هر policy را می توان در تعداد محدودی از مراحل بهبود بخشید، الگوریتم policy iteration تضمین شده است که خاتمه یابد و به policy بهینه همگرا شود.

***

(ج) دو الگوریتم پالیسی ایتریشن (Policy Iteration) و ولیو ایتریشن (Value Iteration) جزو الگوریتم‌های متداول برای حل مسئله فرایند تصمیم‌گیری مارکوف (MDP) به منظور یافتن سیاست بهینه هستند. این الگوریتم‌ها دارای مزایای مختلفی هستند. در ادامه به بررسی مزایای الگوریتم پالیسی ایتریشن نسبت به الگوریتم ولیو ایتریشن پرداخته می‌شود:

۱- همگرایی سریعتر: الگوریتم پالیسی ایتریشن معمولاً سریعتر همگرا می‌شود. زیرا در هر مرحله بین مراحل ارزیابی سیاست و بهبود سیاست جابجا شده می‌شود. در مقابل، الگوریتم ولیو ایتریشن در هر مرحله ارزیابی سیاست را انجام می‌دهد که ممکن است زمان‌بر باشد.

۲- همگرایی تضمین‌شده: الگوریتم پالیسی ایتریشن تضمین می‌کند که در تعداد محدودی از مراحل به سیاست بهینه دست پیدا می‌کند، در حالی که الگوریتم ولیو ایتریشن تنها همگرایی بهینه را در حالت نامحدود تضمین می‌کند. این به این معناست که الگوریتم پالیسی ایتریشن ممکن است به‌صورت دقیق‌تر و سریع‌تر به راه‌حل MDP برسد.

۳- عملکرد بهتر درMDP های بزرگ: الگوریتم پالیسی ایتریشن برایMDP های بزرگ کارایی محاسباتی بیشتری دارد. چرا که به تعداد کمتری از مراحل نسبت به الگوریتم ولیو ایتریشن نیاز دارد. این به این معناست که الگوریتم پالیسی ایتریشن به‌صورت مستقیم سیاست را به به‌جای به‌روزرسانی مقادیر همه حالت‌ها و اقدامات در هر تکرار، پالیسی را مستقیماً به‌روزرسانی می‌کند.

۴- اجرای ساده تر: اجرای پالیسی ایتریشن از نظر مفهومی ساده تر از value iteration است، زیرا شامل دو مرحله اصلی است: policy evaluation و بهبود policy. در مقابل، ولیو ایتریشن فقط یک مرحله را شامل می شود، اما می تواند پیچیده تر باشد زیرا به محاسبه تابع مقدار بهینه نیاز دارد.

به طور کلی، policy iteration می‌تواند انتخاب بهتری نسبت به value iteration از نظر سرعت همگرایی، کارایی محاسباتی و سهولت اجرا، به ویژه برای MDP های بزرگتر باشد. با این حال، انتخاب بین دو الگوریتم در نهایت به ویژگی‌های خاص مسئله MDP در حال حل بستگی دارد.

***

(د) اگر داشته باشیم:

$$V_0^{\pi_{t+1}}(s) = \sum_{s'} P(s' \mid \pi_{t+1}(s), s)[R(s', \pi_{t+1}(s), s) + \gamma V_\infty^{\pi_t}(s')]$$

ابتدا ثابت می‌کنیم که:

$$\forall s \;\; V_0^{\pi_{t+1}}(s) \geq V_\infty^{\pi_t}(s)$$

برای اینکار می‌توان گفت که با توجه به policy evaluation داریم:

$$V_\infty^{\pi_t}(s) = \sum_{s'} P(s' \mid \pi_t(s), s)\big[R(s', \pi_t(s), s) + \gamma V_\infty^{\pi_t}(s')\big]$$

همچنین می‌توان گفت که بهترین policy ما در این لحظه $\pi_{t+1}$ است. پس می‌توان گفت که:

$$\sum_{s'} P(s'|\pi_{t+1}(s),s)\left[R(s',\pi_{t+1}(s),s)+\gamma V_{\infty}^{\pi_t}(s')\right] \geq \sum_{s'} P(s'|\pi_t(s),s)\left[R(s',\pi_t(s),s)+\gamma V_{\infty}^{\pi_t}(s')\right]$$
$$\Rightarrow V_0^{\pi_{t+1}}(s) \geq V_{\infty}^{\pi_t}(s) \qquad \forall s \quad (1)$$

حال فرض می‌کنیم مقادیر در policy evaluation صعودی است. یعنی: $\forall s \;\; V_{k+1}^{\pi_{t+1}}(s) \geq \; V_k^{\pi_{t+1}}(s)$

برای اثبات $V_{\infty}^{\pi_{t+1}}(s) \geq V_{\infty}^{\pi_t}(s)$ از آنجا که policy evaluation صعودی است می‌توان گفت:

$$V_0^{\pi_{t+1}}(s) \leq \cdots \leq V_k^{\pi_{t+1}}(s) \leq \; V_{k+1}^{\pi_{t+1}}(s) \leq \cdots \leq V_{\infty}^{\pi_{t+1}}(s) \qquad (2)$$

پس داریم:

$$(1,2) \Rightarrow V_{\infty}^{\pi_{t+1}}(s) \geq V_0^{\pi_{t+1}}(s) \geq V_{\infty}^{\pi_t}(s) \qquad \forall s$$

پس درحالت تغییر یافته این عبارت برقرار است.

برای حالت تغییر نیافته نیز داریم:

$$V_0^{\pi_{t+1}}(s) = V_0^{\pi_t}(s) = 0$$
$$V_{\infty}^{\pi_t}(s) = \sum_{s'} P(s'|\pi_t(s),s)\left[R(s',\pi_t(s),s)+\gamma V_{\infty}^{\pi_t}(s')\right]$$
$$V_{\infty}^{\pi_{t+1}}(s) = \sum_{s'} P(s'|\pi_{t+1}(s),s)\left[R(s',\pi_{t+1}(s),s)+\gamma V_{\infty}^{\pi_{t+1}}(s')\right]$$

اکنون، برای اثبات اینکه در هر تکرار از policy iteration، مقدار همگرایی با مقدار اولیه صفر صعودی است، باید نشان دهیم که تابع مقدار برای هر حالت با هر تکرار ارزیابی پالیسی در حال افزایش است.

با فرض اینکه تابع مقدار برای هر حالت به صفر مقداردهی شود، می‌توانیم نشان دهیم که تابع مقدار برای هر حالت با هر تکرار ارزیابی سیاست افزایش می‌یابد. زیرا معادله بلمن یک معادله بازگشتی است که به تابع مقدار حالت بعدی بستگی دارد. از آنجایی که تابع مقدار برای هر حالت به صفر مقداردهی می شود، تابع مقدار برای هر حالت تنها با هر تکرار ارزیابی سیاست افزایش می یابد.

هنگامی که تابع مقدار همگرا شد، مرحله بهبود پالیسی، پالیسی را به روز می کند تا عملی را انتخاب کند که تابع مقدار همگرا را برای هر حالت به حداکثر می رساند. این به این معنی است که سیاست جدید بهتر از سیاست قبلی خواهد بود، که تضمین می کند که ارزش همگرایی صعودی است.

بنابراین، می‌توان نتیجه گرفت که در هر تکرار از تکرار سیاست، مقدار همگرایی با مقدار اولیه صفر صعودی است.

## سوال ۳

(آ) بر اساس سوال داریم:

$$R = E\left[\sum_{t=0}^{H} r_t\right]$$
$$Gini(P) = \sum_k p_k(1 - p_k)$$

حال اگر در نظر بگیریم ۲ action داریم خواهیم داشت:

$$\begin{aligned}
&Gini(P) = p_1(1 - p_1) + p_2(1 - p_2)\\
&s.t.\ \ p_1 + p_2 = 1,\ \ p_1, p_2 \geq 0\\
&\Rightarrow Gini(P) = p_1 p_2 + p_2 p_1 = 2p_1 p_2 = 2p_1(1 - p_1)
\end{aligned}$$

حال داریم:

$$\begin{aligned}
&\arg\max Gini(P) = \arg\max 2p_1(1 - p_1) = \arg\max p_1 - {p_1}^2\\
&\Rightarrow 1 - 2p_1^* = 0 \Rightarrow p_1^* = \frac{1}{2} \Rightarrow p_2^* = \frac{1}{2}
\end{aligned}$$

برای کمینه کردن اختلاف می‌توان گفت:

$$\begin{aligned}
&\arg\min|p_1 - p_2| = \arg\min|p_1 - (1 - p_1)| = \arg\min|1 - 2p_1|\\
&\Rightarrow 1 - 2p_1^* = 0 \Rightarrow p_1^* = \frac{1}{2} \Rightarrow p_2^* = \frac{1}{2}
\end{aligned}$$

از دو عبارت بالا می‌توان نتیجه گرفت که:

$$\arg\max Gini(P) = \arg\min|p_1 - p_2|$$

بنابراین این توزیع به کم شدن تفاوت احتمال انتخاب شدن action ها کمک می‌کند.

***

(ب) می‌خواهیم با استفاده از KKT تابع لاگرانژ مربوط به بهینه سازی عبارت زیر را بنویسیم: (با توجه به فرم کلی قیدها را به این صورت تغییر می‌دهیم: منبع) همچنین از آنجایی که Gini به فرم concave است و اکسپکتد یک عدد است؛ می‌توان در نظر گرفت تابع f به صورت convex است. (در این سوال i و a نشان‌دهنده‌ی ایندکس iام و iaction ام هستند و A همان مجموعه G است.)

$$\begin{aligned}
&\min_{\pi_A}\ f(\pi_A) = -E_{\pi_A}[r(a)] - \beta Gini(\pi_A)\\
&subject\ to \qquad l(\pi_A) = \sum_a \pi_A(a) - 1 = 0\\
&\qquad\qquad\qquad h_i(\pi_A) = -\pi_A(a) \leq 0 \quad \text{for all i} \ = \ 0, 1, \ldots, |\mathrm{A}| - 1
\end{aligned}$$

حال مسئله زیر را به عنوان فرم dual مسئله بالا در نظر می‌گیریم:

$$\max_{\mathrm{u,v}} \min_{\pi_A} \; f(\pi_A) + \sum_i u_i h_i(\pi_A) + vl(\pi_A)$$
$$subject\ to \qquad u \geq 0$$

تابع لاگرانژ به صورت زیر است که در اینجا u,v ضرایب لاگرانژ هستند:

$$\mathcal{L}(\pi_A, u, v) = f(\pi_A) + \sum_i u_i h_i(\pi_A) + vl(\pi_A)$$
$$= -E_{\pi_A}[r(a)] - \beta Gini(\pi_A) + \sum_{i,a} -u_i \pi_A(a) + v(\sum_a \pi_A(a) - 1)$$

***

ج) سپس می‌توان شرایط KKT را برای این مسئله بیان کنیم:

Stationarity Condition:

$$0 \in \partial\left( f(\pi_A) + \sum_i u_i h_i(\pi_A) + vl(\pi_A) \right)$$

Primal feasibility:

$$h_i(\pi_A) \leq 0 \quad \text{for all i } = \ 0, 1, \dots, |\mathrm{A}| - 1$$
$$l(\pi_A) = \sum_a \pi_A(a) - 1 = 0$$

Dual feasibility:

$$u_i \geq 0 \quad \text{for all i } = \ 0, 1, \dots, |\mathrm{A}| - 1$$

Complementary slackness:

$$u_i h_i(\pi_A) = 0, \text{for all i } = \ 0, 1, \dots, |\mathrm{A}| - 1$$

حال می‌توان گفت:

$$\frac{\partial \mathcal{L}(\pi_A, u, v)}{\partial \pi_A} = -r(a) - \beta\big(1 - 2\pi_A^*(a)\big) - u_i^* + v^* = 0$$
$$\Rightarrow \pi_A^*(a) = \frac{r(a) + \beta + u_i^* - v^*}{2\beta}$$

همچنین داریم:

$$l(\pi_A) = \sum_a \pi_A(a) - 1 = 0 \Rightarrow \sum_a \pi_A^*(a) = 1 \Rightarrow \sum_a \frac{r(a) + \beta + u_i^* - v^*}{2\beta} = 1 \Rightarrow \sum_{a,i} r(a) + \beta + u_i^* - v^*$$
$$= 2\beta \Rightarrow |A|\beta - |A|v^* + \sum_{a,i} r(a) + u_i^* = 2\beta \Rightarrow |A|v^* - \sum_i u_i^*$$
$$= \sum_a r(a) + \beta(|A| - 2)$$

از آنجایی که احتمالات غیر صفر هستند، می‌توان گفت که $h_i(\pi_A) < 0$ . همچنین چون در شروط KKT داریم که $u_i h_i(\pi_A) = 0$ می‌توان گفت که:

$$u_i = 0 \quad \text{for all i} = 0, 1, \dots, |\text{A}| - 1$$
$$\Rightarrow v^* = \frac{(\sum_a r(a)) + \beta(|A| - 2)}{|A|}$$

پس داریم:

$$\pi_A^*(a) = \frac{r(a) + \beta + u_i^* - v^*}{2\beta} = \frac{r(a)|A| + \beta|A| - (\sum_a r(a)) - \beta(|A| - 2)}{2\beta|A|}$$
$$= \frac{r(a)|A| - (\sum_a r(a))}{2\beta|A|} + \frac{1}{|A|}$$

***

(د) ب رای تبدیل مسئله داده شده به یک مسئله فرم درجه دوم بدون داشتن A، می توانیم از روش زیر استفاده کنیم:

ابتدا Gini را بر حسب مقدار مورد انتظار توزیع احتمال مجذور، به صورت زیر بیان می‌کنیم:

$$Gini(\pi) = \left(\pi(a) - \frac{1}{|A|}\right)^2$$

این یک نتیجه استاندارد در تئوری احتمال است که نشان می دهد معیار ناخالصی جینی معادل واریانس توزیع احتمال است، زمانی که میانگین $\frac{1}{|A|}$ ثابت شود. حال مجموعه جدیدی از متغیرها را مشخص می‌کنیم که با *y_a* نشان داده می شوند، به طوری که $\pi(a) = \frac{exp(y_a)}{Z}$، که در آن Z یک normalization constant است به طوری که $\sum_a exp(y_a) = 1$. این تبدیل تضمین می کند که جمع احتمالات ۱ شوند، یعنی $\sum_a \pi(a) = 1$.

مسئله را بر حسب *y_a* با استفاده از عبارت بالا برای *π(a)* بازنویسی می‌کنیم:

$$\min_y f(y) = -E[r(a)] - \beta \sum_a \left(\frac{exp(y_a)}{Z} - \frac{1}{|A|}\right)^2$$
$$subject\ to \quad l(y) = \sum_a exp(y_a) - Z = 0$$
$$h_{i(y)} = -y_a \leq 0 \quad for\ all\ i = 0,1,\dots,|A| - 1$$

مسئله به دست آمده یک مسئله فرم درجه دوم با شکل زیر است:

$$minimize: Q(y) = -E[r(a)] - \beta \sum_a \left(\frac{exp(y_a)}{Z} - \frac{1}{|A|}\right)^2$$
$$subject\ to: \sum_a exp(y_a) - Z = 0$$

برای یافتن راه‌حل بهینه برای مجموعه اقدام A، می‌توانیم با استفاده از تکنیک‌های بهینه‌سازی استاندارد، مانند gradient descent یا روش نیوتن، مسئله شکل درجه دوم بالا را حل کنیم. هنگامی که جواب بهینه $y$ را به دست آوردیم، می‌توانیم توزیع احتمال بهینه $\pi$ را با استفاده از فرمول بازیابی کنیم:

$$\pi(a) = \frac{exp(y_a)}{Z}$$

در نهایت، ما می‌توانیم مجموعه اقدامات بهینه A را با انتخاب اقدامات با احتمال غیر صفر بازیابی کنیم، یعنی $A = \{a: \pi(a) > 0\}$.

## سوال ۴

(آ) طبق روابط داده شده می‌دانیم:

$$E_{-1}(s) = 0$$
$$E_t(s) = \gamma\lambda E_{t-1}(s) + I_{ss_t}$$

حال می‌خواهیم ثابت کنیم:

$$E_t(s) = \sum_{k=0}^{t} (\gamma\lambda)^{t-k} I_{ss_k}$$

برای اینکار از استقرا استفاده می‌کنیم:

فرض استقرا:

$$E_{-1}(s) = 0$$
$$E_0(s) = \gamma\lambda E_{-1}(s) + I_{ss_0} = I_{ss_0} = \sum_{k=0}^{0} (\gamma\lambda)^{0-k} I_{ss_k}$$

و همینطور برای t=1 داریم:

$$E_1(s) = \gamma\lambda E_0(s) + I_{ss_1} = \gamma\lambda I_{ss_0} + I_{ss_1} = \sum_{k=0}^{1} (\gamma\lambda)^{1-k} I_{ss_k}$$

پس برای فرض اثبات شد.

گام استقرا: فرض می‌کنیم برای t-1 فرض برقرار است. حال نشان می‌دهیم برای t نیز برقرار است.

$$E_t(s) = \gamma\lambda E_{t-1}(s) + I_{ss_t} = \gamma\lambda \sum_{k=0}^{t-1} (\gamma\lambda)^{t-1-k} I_{ss_k} + I_{ss_t} = \sum_{k=0}^{t-1} (\gamma\lambda)^{(t-1-k)+1} I_{ss_k} + I_{ss_t}$$
$$= \sum_{k=0}^{t} (\gamma\lambda)^{t-k} I_{ss_k}$$

بنابراین فرض اثبات شد.

***

(ب) می‌دانیم:

$$\delta_t = R_{t+1} + \gamma V(s_{t+1}) - V(s_t)$$
$$V(s) \leftarrow V(s) + \alpha \delta_t E_t(s)$$

و همچنین:

$$\sum_{t=0}^{T-1} \Delta V_t^{TD}(s) = \sum_{t=0}^{T-1} \alpha \delta_t E_t(s) = \sum_{t=0}^{T-1} \alpha \delta_t \sum_{k=0}^{t} (\gamma\lambda)^{t-k} I_{ss_k}$$
$$= \alpha \delta_0 I_{ss_0}$$
$$+\alpha \delta_1 (\gamma\lambda) I_{ss_0} + \alpha \delta_1 I_{ss_1}$$
$$+\alpha \delta_2 (\gamma\lambda)^2 I_{ss_0} + \alpha \delta_2 (\gamma\lambda) I_{ss_1} + \alpha \delta_2 I_{ss_2}$$
$$+ \cdots$$
$$\alpha \delta_{T-1} (\gamma\lambda)^{T-1} I_{ss_0} + \alpha \delta_{T-1} (\gamma\lambda)^{t-2} I_{ss_1} + \cdots + \alpha \delta_{T-1} I_{ss_{T-1}}$$
$$= \sum_{t=0}^{T-1} \alpha I_{ss_t} \sum_{i=0}^{T-1-t} (\gamma\lambda)^i \delta_{i+t} = \sum_{i=0}^{T-1} \alpha I_{ss_i} \sum_{k=t}^{T-1} (\gamma\lambda)^{k-t} \delta_k$$

تساوی مورد نظر اثبات شد.

***

ج) بر اساس نگاه forward داریم:

$$\frac{1}{\alpha} \Delta V_t^{\lambda}(s) = \frac{1}{\alpha}\left[\alpha\left(G_t^{\lambda} - V_t(s_t)\right)\right] = G_t^{\lambda} - V_t(s_t)$$

$$G_t^{\lambda} = (1-\lambda) \sum_{n=1}^{\infty} \lambda^{n-1} G_t^{(n)} = (1-\lambda) \sum_{n=1}^{\infty} \lambda^{n-1} \left[\left(\sum_{k=0}^{n-1} r_{t+1+k} \gamma^k\right) + \gamma^n V_t(s_{t+n})\right]$$
$$= (1-\lambda) \sum_{n=1}^{\infty} \lambda^{n-1} \left[\left(\sum_{k=1}^{n} r_{t+k} \gamma^{k-1}\right)\right] + (1-\lambda) \sum_{n=1}^{\infty} \lambda^{n-1} \gamma^n V_t(s_{t+n})$$
$$= (1-\lambda) \sum_{k=1}^{\infty} \gamma^{k-1} r_{t+k} \left[\left(\sum_{n=k-1}^{\infty} \lambda^n\right)\right] + (1-\lambda) \sum_{n=1}^{\infty} \lambda^{n-1} \gamma^n V_t(s_{t+n})$$
$$= (1-\lambda) \sum_{k=1}^{\infty} \gamma^{k-1} r_{t+k} \frac{\lambda^{k-1}}{1-\lambda} + (1-\lambda) \sum_{n=1}^{\infty} \lambda^{n-1} \gamma^n V_t(s_{t+n})$$
$$= \sum_{k=1}^{\infty} (\gamma\lambda)^{k-1} r_{t+k} + (1-\lambda) \sum_{n=1}^{\infty} \lambda^{n-1} \gamma^n V_t(s_{t+n})$$
$$= \sum_{n=1}^{\infty} (\gamma\lambda)^{n-1} r_{t+n} + \sum_{n=1}^{\infty} \lambda^{n-1} \gamma^n V(s_{t+n}) - \sum_{n=1}^{\infty} (\gamma\lambda)^n V_t(s_{t+n})$$

بنابراین می‌توان گفت:

$$\begin{aligned}
\frac{1}{\alpha}\Delta V_t^{\lambda}(s) &= \sum_{n=1}^{\infty}(\gamma\lambda)^{n-1}r_{t+n} + \sum_{n=1}^{\infty}\lambda^{n-1}\gamma^{n}V_t(s_{t+n}) - \sum_{n=1}^{\infty}(\gamma\lambda)^{n}V_t(s_{t+n}) - V_t(s_t) \\
&= \sum_{n=1}^{\infty}(\gamma\lambda)^{n-1}r_{t+n} + \sum_{n=1}^{\infty}\lambda^{n-1}\gamma^{n}V_t(s_{t+n}) - \sum_{n=0}^{\infty}(\gamma\lambda)^{n}V_t(s_{t+n}) = \\
&= \sum_{n=1}^{\infty}(\gamma\lambda)^{n-1}r_{t+n} + \sum_{n=1}^{\infty}\lambda^{n-1}\gamma^{n}V_t(s_{t+n}) - \sum_{n=0}^{\infty}(\gamma\lambda)^{n}V_t(s_{t+n}) \\
&= \sum_{n=1}^{\infty}(\gamma\lambda)^{n-1}\big(r_{t+n} + \gamma V_t(s_{t+n})\big) - \sum_{n=1}^{\infty}(\gamma\lambda)^{n-1}V_t(s_{t+n-1})
\end{aligned}$$

پس داریم:

$$\begin{aligned}
\frac{1}{\alpha}\Delta V_t^{\lambda}(s) &= \sum_{n=1}^{\infty}(\gamma\lambda)^{n-1}\big(r_{t+n} + \gamma V_t(s_{t+n}) - V_t(s_{t+n-1})\big) \\
&= \sum_{n=0}^{\infty}(\gamma\lambda)^{n}\big(r_{t+n+1} + \gamma V_t(s_{t+n+1}) - V_t(s_{t+n})\big) \\
&= \sum_{k=t}^{\infty}(\gamma\lambda)^{k-t}\big(r_{k+1} + \gamma V_t(s_{k+1}) - V_t(s_k)\big)
\end{aligned}$$

تساوی مورد نظر اثبات شد.

***

(د)

$$\begin{aligned}
\frac{1}{\alpha}\Delta V_t^{\lambda}(s) &= \sum_{k=t}^{\infty}(\gamma\lambda)^{k-t}\big(r_{k+1} + \gamma V_t(s_{k+1}) - V_t(s_k)\big) \Rightarrow \Delta V_t^{\lambda}(s) \\
&= \alpha\sum_{k=t}^{\infty}(\gamma\lambda)^{k-t}\big(r_{k+1} + \gamma V_t(s_{k+1}) - V_t(s_k)\big) \xRightarrow{offline} \sum_{t=0}^{T-1}\Delta V_t^{\lambda}(s) \\
&= \sum_{t=0}^{T-1}\alpha\sum_{k=t}^{\infty}(\gamma\lambda)^{k-t}\delta_t \overset{2}{\Rightarrow} \sum_{t=0}^{T-1}\Delta V_t^{TD}(s) = \sum_{t=0}^{T-1}\alpha I_{ss_t}\sum_{k=t}^{T-1}(\gamma\lambda)^{k-t}\delta_t = \sum_{t=0}^{T-1}\Delta V_t^{\lambda}(s) I_{ss_t}
\end{aligned}$$

در حالت offline به‌روزرسانی‌ها روی هم انباشته و در انتهای episode اعمال می‌شوند اما در حالت آنلاین پس از هر step در هر episode اعمال می‌شود. حالت تساوی در حالت offline دقیق است، زیرا در این حالت $V_t$ تغییر نمی‌کند ولی در حالت آنلاین تغییر می‌کند. همه مراحل پس از terminal state دارای پاداش صفر و ارزش صفر هستند. بنابراین همه آنها نیز صفر هستند. (۲)