به نام خدا

یادگیری تقویتی

تمرین تئوری چهارم

یلدا شعبان زاده

۹۸۱۰۱۸۲۲

نیمسال دوم ۰۲-۰۱

فهرست

سوال ۱- کشف مهارت

(آ)

.۱

$$I(S;Z) = H(Z) - H(Z|S)$$

در این عبارت ترم اول یا $H(Z)$ با استفاده از uniform prior p(Z) ماکزیمم می‌شود و ترم دوم یا $H(Z|S)$ با ماکزیمم کردن log p(Z|S) مینیمم می‌شود و هرچه log p(Z|S) بیشتر باشد یعنی مهارت‌ها بهتر متمایز شده‌اند. این عبارت اطلاعات متقابل بین حالت های محیط (S) و متغیر پنهان نشان دهنده مهارت (Z) را نشان می‌دهد. این مقدار اطلاعاتی را که متغیر مهارت در مورد وضعیت های محیط ارائه می‌دهد اندازه گیری می‌کند. مقدار بالاتر I(S;Z) نشان می‌دهد که مهارت‌ها آموزنده هستند و مقدار قابل توجهی از اطلاعات را در مورد حالت ها ارائه می دهند، به این معنی که مهارت های مختلف منجر به حالت های متمایز و قابل تشخیص محیط می شود.

.۲

$$I(A;Z|S) = H(Z|S) - H(Z|A,S) = H(A|S) - H(A|S,Z)$$

این عبارت اطلاعات متقابل مشروط بین اقدامات انجام شده توسط عامل (A) و متغیر مهارت پنهان (Z) را با توجه به حالات مشاهده شده محیط (S) نشان می‌دهد. مقدار اطلاعاتی را که متغیر مهارت در مورد اقدامات ارائه می‌دهد، با در نظر گرفتن حالات مشاهده شده اندازه گیری می‌کند. مقدار بالاتر I(A;Z|S) نشان می‌دهد که مهارت‌ها آموزنده هستند و با توجه به حالت‌های مشاهده‌شده، مقدار قابل توجهی از اطلاعات را در مورد اقدامات ارائه می‌دهند. این نشان می‌دهد که مهارت‌های مختلف با در نظر گرفتن زمینه خاص ارائه شده توسط stateها منجر به الگوهای اقدام متمایز و قابل تشخیص می‌شود.

.۳

$$H(A|S)$$

این عبارت بیانگر آنتروپی مشروط اقدامات انجام شده توسط عامل (A) با توجه به حالات مشاهده شده محیط (S) است. عدم قطعیت یا تصادفی بودن اعمال عامل را با در نظر گرفتن حالات مشاهده شده اندازه‌گیری می‌کند. این مقدار اطلاعاتی را که با توجه به حالات مشاهده شده هنوز در مورد اقدامات نامشخص یا غیرقابل پیش بینی هستند، اندازه‌گیری می‌کند. مقدار کمتر H(A|S) نشان می‌دهد که اقدامات عامل با توجه به حالات مشاهده شده قطعی‌تر یا قابل پیش‌بینی‌تر هستند. این نشان می‌دهد که اقدامات عامل بیشتر محدود شده یا تحت تأثیر حالات مشاهده شده است و تنوع یا تصادفی کمتری در فرآیند انتخاب اکشن وجود دارد.

***

(ب)

$$F(\theta) \triangleq I(S;Z) + H(A|S) - I(A;Z|S)$$

I(S;Z): این عبارت شرایطی را برآورده می‌کند که مهارت‌های مختلف باید منجر به مشاهده حالات مختلف محیط شود. مقدار بالاتر I(S;Z) نشان می‌دهد که مهارت‌ها اطلاعات بیشتری در مورد حالت‌ها ارائه می‌دهند، به این معنی که مهارت‌های مختلف منجر به حالت‌های متمایز و متنوع می‌شوند. (شرط اول) همچنین نشان می‌دهد برای تمییز بین مهارت‌های مختلف ما تنها به حالت های محیط نیاز داریم نه عمل هایی که توسط

عامل انجام می‌شود. درواقع $H(Z|S)$ هرچقدر کمتر باشد یعنی با دانستن S می‌توان راحت‌تر Z را پیش‌بینی کرد و اینکه برای تمییز بین مهارت‌های مختلف ما تنها به حالت های محیط نیاز داریم نه عمل هایی که توسط عامل انجام می‌شود. (شرط دوم)

H(A|S): با به حداکثر رساندن H(A|S)، این معیار عامل را تشویق می‌کند تا سطح بالاتری از تصادفی یا تغییرپذیری را در اعمال خود، با توجه به حالات مشاهده‌شده، نشان دهد. این کار باعث می‌شود کشف مهارت‌های متنوع برای عامل با اجازه دادن برای کشف گزینه‌های عمل مختلف در پاسخ به حالت‌های یکسان، بیشتر شود.

I(A;Z|S): این عبارت با در نظر گرفتن تأثیر حالت‌های مشاهده شده، با جریمه کردن وابستگی یا همبستگی بین اقدامات و مهارت ها، معیار را تکمیل می کند. با به حداقل رساندن I(A;Z|S)، این معیار با توجه به حالات مشاهده شده، اقدامات را تشویق می‌کند تا کمتر به مهارت‌ها وابسته باشند. این امر کشف مهارت‌هایی را که از یکدیگر متمایز هستند، ترویج می‌کند، زیرا اقدامات تشویق می‌شوند تا اتکای کمتری به مهارت خاص مرتبط با حالت‌های مشاهده‌شده داشته باشند. (شرط سوم)

با ترکیب این عبارات در معادله (۱)، این معیار شرایط مشاهده حالات قابل تمایز، ترویج اقدامات متنوع با توجه به حالت‌ها و به حداقل رساندن وابستگی بین اقدامات و مهارت‌ها را در نظر می‌گیرد.

***

(ج)

$$
\begin{aligned}
F(\theta) &\triangleq I(S;Z) + H(A|S) - I(A;Z|S) = I(S;Z) + H(A|S) - H(A|S) + H(A|S,Z) \\
&= H(Z) - H(Z|S) + H(A|S,Z)
\end{aligned}
$$

عبارت اول توزیع prior را روی p(z) تشویق می کند تا آنتروپی بالایی داشته باشد. p(z) می‌تواند در این رویکردمان یکنواخت شود تا تضمین شود که حداکثر آنتروپی را دارد. (در DIAYN به این صورت است)

عبارت دوم نشان می‌دهد که استنتاج مهارت z از وضعیت فعلی باید آسان باشد. درواقع $H(Z|S)$ هرچقدر کمتر باشد یعنی با دانستن S می‌توان راحت‌تر Z را پیش‌بینی کرد و اینکه برای تمییز بین مهارت‌های مختلف ما تنها به حالت های محیط نیاز داریم نه عمل هایی که توسط عامل انجام می‌شود.

عبارت سوم نشان می‌دهد که هر مهارت باید تا حد امکان تصادفی عمل کند، که ما با استفاده از حداکثر سیاست آنتروپی برای نشان دادن هر مهارت به آن دست می‌یابیم. این کار باعث می‌شود کشف مهارت‌های متنوع برای عامل با اجازه دادن برای کشف گزینه‌های عمل مختلف در پاسخ به حالت‌های یکسان، بیشتر شود.

***

(د)

$$
\begin{aligned}
F(\theta) \triangleq H(Z) - H(Z|S) + H(A|S,Z) &= H(A|S,Z) + E_{z\sim p(z),s\sim\pi(z)}[\log p(z|s)] - E_{z\sim p(z)}[\log p(z)] \\
&= H(A|S,Z) + E_{z\sim p(z),s\sim\pi(z)}[\log p(z|s)] - E_{z\sim p(z),s\sim\pi(z)}[\log p(z)] \\
&= H(A|S,Z) + E_{z\sim p(z),s\sim\pi(z)}[\log p(z|s) - \log p(z)] \\
&= H(A|S,Z) + E_{z\sim p(z),s\sim\pi(z)}\left[\log \frac{p(z|s)}{p(z)}\right] \\
&= H(A|S,Z) + E_{z\sim p(z),s\sim\pi(z)}\left[\log\left(\frac{p(z|s)}{q_\phi(z|s)} \times \frac{q_\phi(z|s)}{p(z)}\right)\right] \\
&= H(A|S,Z) + E_{z\sim p(z),s\sim\pi(z)}\left[\log \frac{p(z|s)}{q_\phi(z|s)} + \log \frac{q_\phi(z|s)}{p(z)}\right] \\
&= H(A|S,Z) + E_{z\sim p(z),s\sim\pi(z)}\left[D_{KL}\left(p(z|s)||q_\phi(z|s)\right)\right] + E_{z\sim p(z),s\sim\pi(z)}\left[\log \frac{q_\phi(z|s)}{p(z)}\right] \\
&\geq H(A|S,Z) + E_{z\sim p(z),s\sim\pi(z)}\left[\log q_\phi(z|s) - \log p(z)\right] = G(\theta,\phi)
\end{aligned}
$$

***

(ه) اولین ترم در $G(\theta,\phi)$ نشان‌دهنده‌ی این است که هر مهارت باید تا حد امکان تصادفی عمل کند، یعنی آنتروپی بالایی از اکشن نسبت به دانستن استیت و مهارت می‌خواهیم. یعنی همان exploration؛ در دومین ترم در بخش $E_{z\sim p(z),s\sim\pi(z)}[-\log p(z)]$ می‌خواهیم آنتروپی مهارت‌ها را زیاد کنیم. یعنی به احتمال یکسانی برای هر مهارت نزدیک شویم. در بخش $E_{z\sim p(z),s\sim\pi(z)}[\log q_\phi(z|s)]$ هم می‌خواهیم احتمال discriminate کردن مهارت‌ها را به شرط دانستن استیت افزایش دهیم. به همین دلیل با افزایش ترم دوم؛ امکان تصادفی عمل کردن کاهش می‌یابد و برعکس. پس ما یک trade-off بین exploration و discrimination داریم. به همین دلیل در DIAYN با اضافه کردن ضریب α به ترم اول این مصالحه را کنترل می‌کنند.

***

(و) SAC آنتروپی سیاست را بر اقدامات به حداکثر می‌رساند، که از عبارت آنتروپی در هدف G ما مراقبت می‌کند. می‌توان ریوارد را با توجه به تابع هدف و discriminative بودن برابر زیر قرار داد:

$$r_z(s,a) = \log q_\phi(z|s) - \log p(z)$$

با این روش عامل برای بازدید از حالاتی که به راحتی قابل تشخیص است پاداش دریافت می کند، در حالی که discriminator آپدیت می‌شود تا مهارت z را از حالت های بازدید شده بهتر استنتاج کند. تنظیم آنتروپی به عنوان بخشی از به روز رسانی SAC رخ می دهد.

***

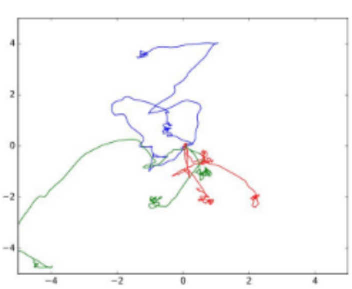
(ز)

$$\begin{aligned}
I(s';z|s) &= H(z|s) - H(z|s',s) = H(s'|s) - H(s'|s,z) \\
&= -E_{s,s'\sim p}[\log p(s'|s)] + E_{z,s,s'\sim p}[\log p(s'|s,z)] = E_{z,s,s'\sim p}\left[\log\frac{p(s'|s,z)}{p(s'|s)}\right] \\
&= E_{z,s,s'\sim p}\left[\log\left(\frac{p(s'|s,z)}{q_\phi(s'|s,z)}\times\frac{q_\phi(s'|s,z)}{p(s'|s)}\right)\right] \\
&= E_{z,s,s'\sim p}\left[\log\frac{p(s'|s,z)}{q_\phi(s'|s,z)} + \log\frac{q_\phi(s'|s,z)}{p(s'|s)}\right] \\
&= E_{s,z\sim p}\left[D_{KL}\left(p(s'|s,z)||q_\phi(s'|s,z)\right)\right] + E_{z,s,s'\sim p}\left[\log\frac{q_\phi(s'|s,z)}{p(s'|s)}\right]
\end{aligned}$$

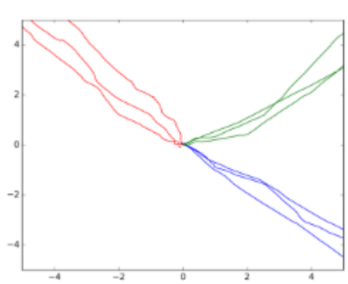
اطلاعات متقابل در معادله اول می‌گوید که چقدر می توان در مورد s′ با توجه به s,z یا به طور متقارن، z را با توجه به انتقال از s → s′ دانست. از معادله دوم، حداکثر کردن این هدف مربوط به به حداکثر رساندن تنوع انتقال‌های تولید شده در محیط است، که با آنتروپی $H(s'|s)$ نشان داده می شود، در حالی که z را با به حداقل رساندن آنتروپی $H(s'|s,z)$ در مورد وضعیت بعدی s′ اطلاع رسانی می‌کند. به طور شهودی، مهارت‌های z را می‌توان به‌عنوان توالی‌های عمل انتزاعی تفسیر کرد که با انتقال‌های ایجاد شده در محیط (و نه فقط با وضعیت فعلی) قابل شناسایی هستند.

بنابراین، بهینه‌سازی این اطلاعات متقابل را می‌توان به عنوان رمزگذاری مجموعه‌ای از مهارت‌ها در فضای پنهان Z درک کرد، در حالی که انتقال‌ها را برای z∈ Z معین قابل پیش‌بینی می‌کند.

تفاوت این با تابع هدف در قسمت ب هم در این است که در این قسمت ترنزیشن‌ها یا به عبارتی استیت‌های بعدی نیز در نظر گرفته شده‌اند تا از حرکات رندم در یک محدوده توسط عامل جلوگیری شود زیرا در اینجا $H(s'|s)$ قرار است ماکزیمم شود و یعنی استیت‌های آینده احتمالات یکسانی بهتر است داشته باشند. اما با مینیمم کردن $H(s'|s,z)$ یعنی شرطی شدن روی مهارت باعث افزایش احتمال استیت s′ نسبت به s و z می‌شود. اما در ب وجود اکشن های متفاوت ممکن است باعث تمایز استیت نهایی در آن مهارت نشود. همچنین DIAYN بر این ایده استوار است که مهارت‌های متنوع را می‌توان برای حل وظایف مختلف مورد استفاده قرار داد و می‌تواند منجر به عملکرد بهتر از یادگیری یک پالیسی واحد برای همه کارها شود. این موجب افزایش پیچیدگی مهارت‌های یادگرفته شده می‌شود. اما DADS بر این ایده استوار است که مهارت ها را می توان با کشف دینامیک‌های زیربنایی محیط و سپس استفاده از این دانش برای یادگیری سیاست‌هایی که می تواند وظایف را حل کند، یاد گرفت.



DIAYN



DADS

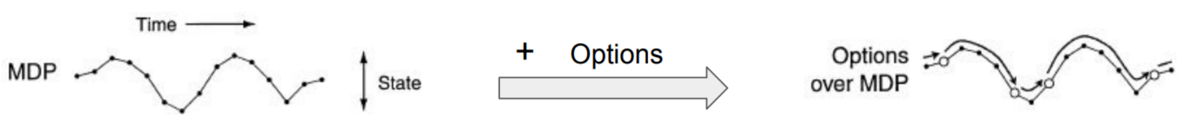
سوال ۲- یادگیری سلسله مراتبی

(آ)

دلیل استفاده از semi-MDP زمانی که می‌خواهیم temporal abstraction داشته باشیم این است که اقدامات و برنامه‌ریزی‌های موقتی را با temporally abstract knowledge امکان‌پذیر می‌کند. علاوه بر این، انتقال دانش با استفاده از دانش دامنه برای تعریف گزینه‌ها امکان‌پذیر است و راه‌حل‌های اهداف فرعی را می‌توان مجدداً مورد استفاده قرار داد.

به طور خلاصه این فواید عبارتند از:

- Knowledge Transfer: استفاده از دانش دامنه برای تعریف optionها؛ راه حل های اهداف فرعی را می توان مجددا استفاده کرد.
- semi-MDP = option + MDP: یک تئوری معرفی شده است که در آن می‌گوید برای هر MDP و هر مجموعه‌ای از optionهای تعریف شده بر روی آن MDP، فرآیند تصمیم گیری که فقط از بین آن optionها انتخاب می شود و هر کدام را تا خاتمه اجرا می‌کند، یک نیمه MDP است.



- یادگیری و planning به صورت efficientتر

منبع

***

(ب)

همانطور که می‌دانیم مقدار value حالت s زیر پالیسی π برابر است با: (π state-value function)

$$V^{\pi}(s) = E[r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \cdots | s_t = s, \pi] = E[r_{t+1} + \gamma V^{\pi}(s_{t+1}) | s_t = s, \pi]$$
$$= \sum_{a \in A_s} \pi(a|s) \left[r_s^a + \gamma \sum_{s'} p_{ss'}^a V^{\pi}(s') \right]$$

حال optimal state-value function یا همان معادله بلمن برای s تحت π برابر است با:

$$V^*(s) = \max_{\pi} V^{\pi}(s) = \max_{a \in A_s} E[r_{t+1} + \gamma V^*(s_{t+1}) | s_t = s, a_t = a] = \max_{a \in A_s} \left[r_s^a + \gamma \sum_{s'} p_{ss'}^a V^*(s') \right]$$

حال اگر option داشته باشیم برای هر مارکف پالیسی μ state-value function را می‌توان به صورت زیر نوشت:

$$V^{\mu}(s) = E[r_{t+1} + \gamma r_{t+2} + \cdots + \gamma^{k-1} r_{t+k} + \gamma^k V^{\mu}(s_{t+k}) | \varepsilon(\mu, s, t)] = \sum_{o \in O_s} \mu(s, o) \left[r_s^o + \sum_{s'} p_{ss'}^o V^{\mu}(s') \right]$$

همچنین معادله بلمن برای s,a برابر است با:

$$\begin{aligned}
Q^{\mu}(s,o) &= E[r_{t+1} + \gamma r_{t+2} + \cdots + \gamma^{k-1} r_{t+k} + \gamma^{k} V^{\mu}(s_{t+k}) | \varepsilon(o,s,t)] \\
&= E\left[r_{t+1} + \gamma r_{t+2} + \cdots + \gamma^{k-1} r_{t+k} + \gamma^{k} \sum_{o \in O_s} \mu(s_{t+k}, o) Q^{\mu}(s_{t+k}, o') \,|\varepsilon(o,s,t)\right] \\
&= r_s^o + \sum_{s'} p_{ss'}^{o} \sum_{o \in O_{s'}} \mu(s', o') Q^{\mu}(s', o')
\end{aligned}$$

معادله دوم بر اساس تعریف مقدار حالت بعدی به عنوان expectation مقادیر Q در وضعیت بعدی نسبت به سیاست فعلی به دست آمده است.

***

ج) ابتدا برای $V_O^*(s)$ داریم:

$$\begin{aligned}
V_O^*(s) &\coloneqq \max_{\mu \in \Pi(O)} V^{\mu}(s) = \max_{o \in O_s} E[r_{t+1} + \gamma r_{t+2} + \cdots + \gamma^{k-1} r_{t+k} + \gamma^{k} V_O^*(s_{t+k}) | \varepsilon(o,s,t)] \\
&= \max_{o \in O_s} \left[r_s^o + \sum_{s'} p_{ss'}^{o} V_O^*(s')\right] = \max_{o \in O_s} E[r + \gamma^k V_O^*(s) | \varepsilon(o,s)]
\end{aligned}$$

حال می‌دانیم که: $V_O^*(s) = \max_{o \in O_s} Q_O^*(s, o')$

$$\begin{aligned}
Q_O^*(s,o) &\coloneqq \max_{\mu \in \Pi(O)} Q^{\mu}(s,o) = E[r_{t+1} + \gamma r_{t+2} + \cdots + \gamma^{k-1} r_{t+k} + \gamma^{k} V_O^*(s_{t+k}) | \varepsilon(o,s,t)] \\
&= E\left[r_{t+1} + \gamma r_{t+2} + \cdots + \gamma^{k-1} r_{t+k} + \gamma^{k} \max_{o \in O_{s_{t+k}}} Q_O^*(s_{t+k}, o') \,|\varepsilon(o,s,t)\right] \\
&= r_s^o + \sum_{s'} p_{ss'}^{o} \max_{o \in O_{s'}} Q_O^*(s', o') = E\left[r + \gamma^k \max_{o \in O_{s'}} Q_O^*(s', o') \,|\varepsilon(o,s)\right]
\end{aligned}$$

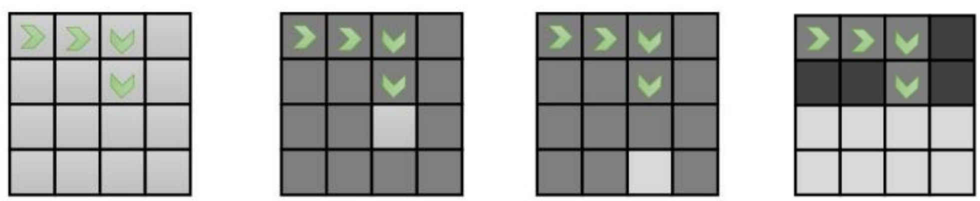
سوال ۳- مروری بر یادگیری تقویتی معکوس

(آ) یادگیری تقویتی معکوس (IRL) تکنیکی است که به جای یادگیری مستقیم یک پالیسی، هدف آن یادگیری تابع پاداش از رفتار اکسپرت است. دلایل مختلفی وجود دارد که چرا ما ممکن است به جای یادگیری مستقیم یک پالیسی از رفتار اکسپرت در این زمینه، استفاده از یادگیری تابع پاداش از طریق IRL را انتخاب کنیم:

تابع پاداش آموخته شده می تواند بینشی در مورد ترجیحات و اهداف اساسی متخصص ارائه دهد که می تواند در درک کار و طراحی پالیسی‌های جدید مفید باشد. تابع پاداش آموخته شده می تواند وظیفه اساسی را به روشی قوی‌تر از یادگیری مستقیم یک پالیسی از رفتار اکسپرت نشان دهد، زیرا تابع پاداش می‌تواند رفتار متخصص را در طیف وسیعی از موقعیت ها توضیح دهد. درواقع می‌تواند generalization بهتری داشته باشد و در حالات OOD موجب خسارات برگشت ناپذیر نشود. یادگیری تابع پاداش، انعطاف‌پذیری بیشتری را در تعریف کار امکان‌پذیر می‌کند، زیرا می‌توانیم رفتار مورد نظر را از طریق تابع پاداش به جای تکیه بر مجموعه ثابتی از نمایش‌های اکسپرت مشخص کنیم.

پس با این کار می‌توان امید داشت که پالیسی یادگرفته‌ی ما بهتر از پالیسی اکسپرت بشود. درواقع ما intent را در اینجا می‌خواهیم یاد بگیریم.

اما یک مشکل در IRL وجود دارد و آن هم under specification است. یعنی به ازای یک رفتار مشخص و ثابت از اکسپرت، ریوارد فانکشن‌های متفاوتی می‌توانند آن رفتار را نشان دهند. برای مثال اگر ما یک demonstrationای مثل این جهت‌های سبز رنگ در شکل زیر داشته باشیم، ریواردهای مختلفی می‌توانند با این رفتار متناسب باشند.



درواقع مسئله ill-posed است و ما باید بتوانیم بین این حالات یکی را به عنوان هدف انتخاب کنیم زیرا نمی‌توان مشخص کرد هدف اصلی کدام بوده است.

***

(ب)

۱. روش‌های inverse RL یک مشکلی که دارند under specification است. پس باید یک سری constraintهایی داشته باشیم که این مشکل را حل کند. به همین دلیل فرض کردیم ریوارد فانکشن ما حالت پارامتریک دارد و از فیچرهای استیت-اکشن استفاده می‌کند و به شکل یک تابع خطی با پارامترهای ناشناخته φ ریوارد تعریف می‌شود.

$$r_\psi(s,a) = \sum_{i=1}^{n} \psi_i f_i(s,a) = \psi^T f(s,a)$$

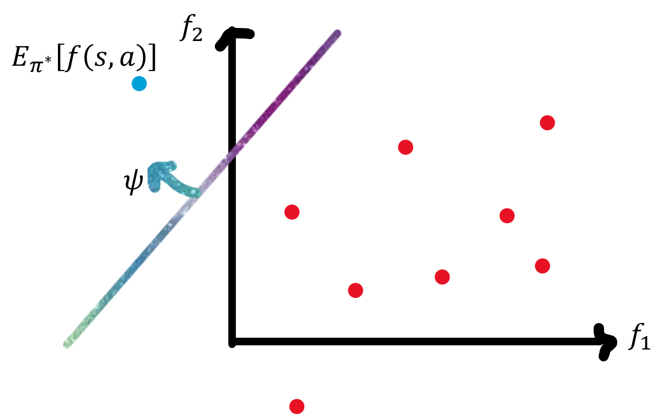
حال فرض کنید که π^{r_ψ} پالیسی آپتیمال برای r_ψ باشد. آنگاه داریم:

$$E_{\pi^{r_\psi}}[f(s,a)] = E_{\pi^*}[f(s,a)]$$

این مفهوم feature matching است و درواقع می‌گوید این فیچرهایی که تعریف کردیم و ریوارد از آنها ساخته می‌شود می‌خواهیم با پالیسی‌ای که از اکسپرت می‌آید اکسپکتد یکسانی داشته باشد تا trajectoryهای بین اکسپرت با فیچرهایی که ما یاد گرفتیم مچ بشوند.

۲. این کار مشکل under specification را حل نمی‌کند. زیرا این تساوی می‌تواند با ψهای مختلف برقرار شود. مثلا اگر در یک grid world فیچرها را به شکل باینری در نظر بگیریم که آیا در خانه i,j قرار گرفته‌ایم یا نه. یعنی در جدول m*n، m*n تا فیچر باینری تعریف به این صورت کنیم، در هر استیتی باشیم یکی ۱ و بقیه ۰ هستند. در این حالت پس باز هم مشکل under specification وجود دارد.

۳. برای حل این مشکل یک کار استفاده از ایده‌ی SVM و maximum margin است. به این صورت که در فضای فیچرها هرکدام از $E[f(s,a)]$ یک نقطه هستند. اگر فیچر اسپیس ما به اندازه‌ی کافی قوی باشد نقطه‌ی متناظر با $E_{\pi^*}[f(s,a)]$ از سایر نقاط که $E_{\pi^{r_\psi}}[f(s,a)]$ هستند (به ازای ψهای مختلف)، به خوبی تفکیک می‌شود.



این مسئله می‌تواند ill-posed باشد زیرا یک ψ واحد وجود ندارد. بنابراین از maximum margin استفاده می‌کنیم. پس می‌توانیم مسئله را به شکل زیر تعریف کنیم:

$$\max_{\psi, m} \quad m$$
$$s.t. \quad \psi^T \mathbb{E}_{\pi^*}[f(s,a)] \geq \max_{\pi} \psi^T \mathbb{E}_{\pi}[f(s,a)] + m$$

که در اینجا m متغیر کمکی است. مشکلات این روش این است که ما در فضای همه‌ی پالیسی‌ها می‌خواهیم این کلسیفیکیشن را انجام دهیم. بنابراین ممکن است یکی از آن پالیسی‌ها همان π^* باشد. پس این m صفر می‌شود و نمی‌خواهیم این اتفاق بیفتد. کاری که برای حل این مشکل می‌کنند استفاده از KL-divergence است.

$$\min_{\psi} \frac{1}{2}\|\psi\|^2 \quad \text{such that } \psi^T E_{\pi^\star}[\mathbf{f}(\mathbf{s},\mathbf{a})] \geq \max_{\pi \in \Pi} \psi^T E_{\pi}[\mathbf{f}(\mathbf{s},\mathbf{a})] + D(\pi, \pi^\star)$$

e.g., difference in feature expectations!

در اینجا ما فاصله را برای سیاست‌هایی که متمایزتر از سیاست‌های اکسپرت هستند، بیشتر می‌کنیم، در حالی که سیاست‌های مشابه فاصله کمتری دارند و حاشیه کمتری دریافت می‌کنند.

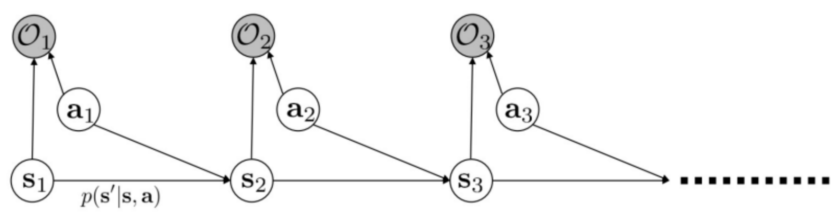
بنابراین مسئله به این آپتیمیزیشن می‌تواند تبدیل شود. مشکلاتی که باز هم دارد این است که فرض کردیم پالیسی اکسپرت آپتیمال است اما ممکن است sub optimal باشد. می‌توان slack variableها را اضافه کرد اما کلا مسئله‌ی constraint optimization زیاد در دیپ لرنینگ خوش تعریف نیست.

***

(ج)

۱.

$$p(O_t|s_t,a_t) = \exp\left(r_\psi(s_t,a_t)\right)$$
$$\Rightarrow p(\tau|O_{1:T}) = \frac{p(\tau, O_{1:T})}{p(O_{1:T})} \propto p(\tau)p(O_{1:T}|\tau) = p(\tau)\Pi_{t=1}^{T} p(O_t|\tau) = p(\tau)\Pi_{t=1}^{T} \exp\left(r_\psi(s_t,a_t)\right)$$
$$= p(\tau)\exp\left(\sum\nolimits_t r_\psi(s_t,a_t)\right)$$



به طور دقیق‌تر می‌توان گفت که مساوی در حالت زیر است که در اینجا Z همان partition function است. درواقع:

$$p(\tau|O_{1:T}) = \frac{1}{Z}p(\tau)\exp\left(\sum\nolimits_t r_\psi(s_t,a_t)\right), \qquad Z = \int p(\tau)\exp\left(\sum\nolimits_t r_\psi(s_t,a_t)\right)d\tau$$

در maximum likelihood learning داریم:

$$L_\psi \coloneqq \frac{1}{N}\sum_{i=1}^{N} \log p(\tau_i|O_{1:T},\psi)$$

همچنین:

$$p(\tau|O_{1:T}) \propto \underbrace{p(\tau)}_{\text{can ignore (independant of } \psi)} \exp\left(\sum\nolimits_t r_\psi(s_t,a_t)\right)$$

بنابراین داریم:

$$
\begin{aligned}
L_\psi &= \frac{1}{N}\sum_{i=1}^{N}\log p(\tau_i|O_{1:T},\psi) = \frac{1}{N}\sum_{i=1}^{N}\log\frac{1}{Z}p(\tau_i)\exp\left(\sum\nolimits_t r_\psi(s_t,a_t)\right) \\
&= \frac{1}{N}\sum_{i=1}^{N}\log p(\tau_i) + \log\exp\left(\sum\nolimits_t r_\psi(s_t,a_t)\right) - \log Z \\
&= \frac{1}{N}\sum_{i=1}^{N}\log p(\tau_i) + \log\exp\left(\sum\nolimits_t r_\psi(s_t,a_t)\right) - \log Z \\
\Rightarrow \max_\psi L_\psi &= \max_\psi \frac{1}{N}\sum_{i=1}^{N}\log p(\tau_i|O_{1:T},\psi) = \max_\psi \frac{1}{N}\sum_{i=1}^{N}\log\exp\left(\sum\nolimits_t r_\psi(s_t,a_t)\right) - \log Z \\
&= \max_\psi \frac{1}{N}\sum_{i=1}^{N}\sum\nolimits_t r_\psi(s_{t,i},a_{t,i}) - \log Z = \max_\psi \frac{1}{N}\sum_{i=1}^{N} r_\psi(\tau_i) - \log Z
\end{aligned}
$$

بنابراین:

$$
\max_\psi \frac{1}{N}\sum_{i=1}^{N}\log p(\tau_i|O_{1:T},\psi) = \max_\psi \frac{1}{N}\sum_{i=1}^{N} r_\psi(\tau_i) - \log Z
$$

نشان دادیم L_ψ در معادله‌ی بالا صدق می‌کند.

۲. اگر $\log Z$ را نداشته باشیم؛ آنگاه انگار می‌خواهیم میانگین $r_\psi(\tau_i)$ ها را حساب کنیم و این باعث می‌شود trajectoryهایی که اکسپرت دیده است نسبت به سایر trajectoryها محتمل‌تر بشوند و دیگر سایر آنها حتی اگر ریوارد بالایی داشته باشند محتمل نشوند.

۳.

$$
\begin{aligned}
\nabla_\psi L_\psi &= \nabla_\psi \frac{1}{N}\sum_{i=1}^{N} r_\psi(\tau_i) - \log Z = \frac{1}{N}\sum_{i=1}^{N}\nabla_\psi r_\psi(\tau_i) - \nabla_\psi \log Z = \frac{1}{N}\sum_{i=1}^{N}\nabla_\psi r_\psi(\tau_i) - \frac{1}{Z} \\
&= E_{\tau\sim\pi^*(\tau)}\left[\nabla_\psi r_\psi(\tau)\right] - \underbrace{\frac{1}{Z}\int p(\tau)\exp\left(r_\psi(\tau)\right)}_{p(\tau|O_{1:T},\psi)}\nabla_\psi r_\psi(\tau)d\tau \\
&= E_{\tau\sim\pi^*(\tau)}\left[\nabla_\psi r_\psi(\tau)\right] - E_{p(\tau|O_{1:T},\psi)}\left[\nabla_\psi r_\psi(\tau)\right]
\end{aligned}
$$

حال برای ترم دوم می‌توان گفت:

$$
E_{p(\tau|O_{1:T},\psi)}\left[\nabla_\psi r_\psi(\tau)\right] = E_{p(\tau|O_{1:T},\psi)}\left[\nabla_\psi \sum\nolimits_t r_\psi(s_t,a_t)\right] = \sum_{t=1}^{T} E_{(s_t,a_t)\sim p(s_t,a_t|O_{1:T},\psi)}\left[\nabla_\psi r_\psi(s_t,a_t)\right]
$$

همچنین می‌توان گفت:

$$p(s_t, a_t | O_{1:T}, \psi) = p(a_t | s_t, O_{1:T}, \psi) p(s_t | O_{1:T}, \psi)$$

که بنا بر تعریف بیزین از اسلایدها می‌دانیم:

$$p(a_t | s_t, O_{1:T}, \psi) = \frac{\beta(s_t, a_t)}{\beta(s_t)}, \qquad p(s_t | O_{1:T}, \psi) \propto \beta(s_t)\alpha(s_t)$$
$$\Rightarrow p(s_t, a_t | O_{1:T}, \psi) \propto \underbrace{\beta(s_t, a_t)}_{backward\ message} \underbrace{\alpha(s_t)}_{forward\ message} = \mu_t(s_t, a_t)$$

بنابراین داریم:

$$E_{p(\tau | O_{1:T}, \psi)}[\nabla_\psi r_\psi(\tau)] = \sum_{t=1}^{T} E_{(s_t, a_t) \sim p(s_t, a_t | O_{1:T}, \psi)}[\nabla_\psi r_\psi(s_t, a_t)]$$
$$= \sum_{t=1}^{T} \int \int \mu_t(s_t, a_t) \nabla_\psi r_\psi(s_t, a_t) ds_t\, da_t$$

۴. در این عبارت جدید که بدست آوردیم ما نیاز داریم که روی فضای استیت-اکشن انتگرال بگیریم درحالی که در عبارت قبلی نیاز بود تا روی همه‌ی trajectoryها انتگرال بگیریم که این کار غیرممکن و بسیار سخت بود. بنابراین این روش یک روش عملی برای حساب کردن وزن‌های ψ است.

نکته‌ای که وجود دارد این است که در این روش نیز باید فضای استیت و اکشن محدود و کوچک باشد؛ اما در فضای محدود و کوچک نیز trajectoryها بزرگ می‌شوند.

۵. پس از هر آپدیت ψ باید هرسری بیزین RL بزنیم و به همین دلیل خیلی scalable نیست و ایده‌اش قابل استفاده است. برای عملی شدن باید بتوانیم فضای بزرگ state action نیز هندل کنیم و اینکه stateها فقط از sampling بیایند و دینامیک ناشناخته باشد نیز مشکلات اپلای کردن این روش در فضاهای بزرگ است. حال فرض می‌کنیم ψ را در هر گام می‌دانیم. پس با این یک ریوارد تعریف می‌کنیم و یک MaxEnt-RL ران می‌کنیم و حال می‌توان از این RL سمپل گرفت. زیرا یک پالیسی رندم بدست می‌آید و با هر ران یک trajectory می‌گیریم.

MaxEnt RL به فرم زیر است:

$$J(\theta) = \sum_t E_{\pi(s_t, a_t)}[r_\psi(s_t, a_t)] + E_{\pi(s_t)}[H(\pi(a|s_t))]$$

حال داریم:

$$\nabla_\psi L_\psi \approx \frac{1}{N} \sum_{i=1}^{N} \nabla_\psi r_\psi(\tau_i) - \frac{1}{M} \sum_{i=1}^{M} \nabla_\psi r_\psi(\tau_i)$$

که در آن عبارت اول از مسیرهای اکسپرت و عبارت دوم از ران کردن سیاست فعلی آمده است. این کار باعث می‌شود استیمیتور جمله‌ی دوم بایاس شود زیرا trajectoryها را از پالیسی‌ای گرفته است که الزاما بهینه نیست (یا حتی soft optimal) و ما صرفا یک گام $p(s_t, a_t|O_{1:T}, \psi)$ را بهتر کردیم. بنابراین برای حل این مشکل باید از Importance sampling استفاده کنیم.

در صورت مقدار آپتیمالیتی را می‌گذاریم که exp جمع ریواردها را دارد و در مخرج نیز پالیسی ای که تا الان improve کردیم می‌گذاریم:

$$\nabla_\psi L_\psi \approx \frac{1}{N}\sum_{i=1}^{N} \nabla_\psi r_\psi(\tau_i) - \frac{1}{\sum_j w_j}\sum_{i=1}^{M} w_j \nabla_\psi r_\psi(\tau_i)\,, \qquad w_j = \frac{p(\tau)\exp\left(r_\psi(\tau_j)\right)}{\pi(\tau_j)}$$

پس از ساده کردن وزن‌ها داریم:

$$w_j = \frac{p(\tau)\exp\left(r_\psi(\tau_j)\right)}{\pi(\tau_j)} = \frac{p(s_1)\Pi_t p(s_{t+1}|s_t, a_t)\exp\left(r_\psi(s_t, a_t)\right)}{p(s_1)\Pi_t p(s_{t+1}|s_t, a_t)\pi(a_t|s_t)} = \frac{\Pi_t \exp\left(r_\psi(s_t, a_t)\right)}{\Pi_t \pi(a_t|s_t)}$$
$$= \frac{\exp\left(\sum_t r_\psi(s_t, a_t)\right)}{\Pi_t \pi(a_t|s_t)} = \frac{\exp\left(r_\psi(\tau_j)\right)}{\Pi_t \pi(a_t|s_t)}$$

سوال ۴- یادگیری برون‌خط

(آ)

$$\mathcal{L}(\pi) = E_{\rho_\pi(\tau)}\left[\sum_{t=0}^{H} \delta(a_t, a_t^*)\right],\ \ \pi(a_t \neq a_t^*|s) \le \epsilon$$

فرض کنید ϵ_i برابر اکسپکتد 0-1 loss در زمان i از π باشد. یعنی داریم:

$$\epsilon_i = E_{s\sim d_{\pi^*}^i}[C_\pi(s)] = E_{s\sim d_{\pi^*}^i}\left[E_{a\sim\pi(.|s)}\left[\delta\left(a, \pi^*(s)\right)\right]\right],\ \ \text{for i} = 1,2,\dots,\text{H}$$

آنگاه داریم:

$$\Rightarrow \epsilon \ge \frac{1}{H}\sum_{i=1}^{H}\epsilon_i$$

درواقع ϵ_i احتمالی را نشان می‌دهد که π تحت توزیع $d_{\pi^*}^i$ اشتباه عمل کند. حال فرض کنید p_t برابر احتمالی باشد که π در آن در t قدم اول اشتباه عمل نکرده باشد. (با در نظر گرفتن π^*)؛ و d_t توزیع حالات π روی زمان t به شرط اینکه تا الان اشتباهی عمل نکرده باشد، باشد. همچنین $d_t{}'$ برابر توزیع حالات در زمان t که با پیروی از π^* به دست می‌آید، اما مشروط به این واقعیت است که π حداقل یک اشتباه در اولین t-1 حالت‌های بازدید شده مرتکب شده است. پس داریم:

$$d_{\pi^*}^t = p_{t-1}d_t + (1-p_{t-1})d_t'$$

اکنون در زمان t، هزینه مورد انتظار π اگر تا کنون اشتباه کرده است حداکثر ۱ است، یا اگر هنوز اشتباه نکرده است $E_{s\sim d_t}[C_\pi(s)]$ است. پس داریم:

$$J(\pi) \le \sum\nolimits_{t=1}^{H} p_{t-1}E_{s\sim d_t}[C_\pi(s)] + (1-p_{t-1}) \times 1$$

فرض کنید e_t و e_t' برابر احتمال اشتباه π در توزیع d_t و d_t' باشد. آنگاه:

$$E_{s\sim d_t}[C_\pi(s)] \le E_{s\sim d_t}[C_{\pi^*}(s)] + e_t$$

همچنین داریم:

$$\epsilon_t = p_{t-1}e_t + (1-p_{t-1})e_t' \Rightarrow p_{t-1}e_t \le \epsilon_t$$
$$p_t = (1-e_t)p_{t-1} \Rightarrow p_t \ge p_{t-1} - \epsilon_t \ge 1 - \sum_{i=1}^{t}\epsilon_i \Rightarrow 1 - p_{t-1} + \epsilon_t \le \sum_{i=1}^{t}\epsilon_i$$

و نیز داریم:

$$J(\pi^*) = \sum\nolimits_{t=1}^{H} p_{t-1}E_{s\sim d_t}[C_{\pi^*}(s)] + (1-p_{t-1})E_{s\sim d_t'}[C_{\pi^*}(s)]$$

حال می‌توان گفت:

$$J(\pi) \leq \sum_{t=1}^{H} p_{t-1} E_{s\sim d_t}[C_\pi(s)] + (1-p_{t-1}) \times 1 \leq \sum_{t=1}^{H} p_{t-1}\big(E_{s\sim d_t}[C_{\pi^*}(s)] + e_t\big) + (1-p_{t-1})$$
$$= \sum_{t=1}^{H} p_{t-1}\big(E_{s\sim d_t}[C_{\pi^*}(s)]\big) + \underbrace{p_{t-1}e_t}_{\leq \epsilon_t} + (1-p_{t-1}) \leq J(\pi^*) + \sum_{t=1}^{H} \epsilon_t + (1-p_{t-1})$$
$$\leq J(\pi^*) + \sum_{t=1}^{H} \sum_{i=1}^{t} \epsilon_i \leq J(\pi^*) + H \sum_{t=1}^{H} \epsilon_t \leq J(\pi^*) + H^2\epsilon$$
$$\Rightarrow J(\pi) - J(\pi^*) \leq H^2\epsilon$$

بنابراین خطای یادگیری دارای حد بالای $O(H^2\epsilon)$ است.

***

(ب) در حالت on policy از آنجا که عامل تعامل با محیط دارد می‌توان گفت:

$$\epsilon_i = E_{s\sim d_\pi^i}[C_\pi(s)] = E_{s\sim d_\pi^i}\left[E_{a\sim\pi(.|s)}\left[\delta\big(a, \pi^*(s)\big)\right]\right], \text{ for i} = 1,2, \dots, \text{H}$$

و بنابراین به طور واضح داریم:

$$J(\pi) \leq J(\pi^*) + \sum_{i=1}^{H} \epsilon_i \leq J(\pi^*) + \sum_{i=1}^{H} \epsilon = J(\pi^*) + H\epsilon$$

بنابراین خطای یادگیری دارای حد بالای $O(H\epsilon)$ است.