سؤال ۱:

توضیح کامل الگوریتم SAC:

الگوریتم SAC یک الگوریتم model free (با استفاده از نمونه‌گیری) و off-policy است.

الگوریتم‌های RL معمولی تلاش می‌کنند که راه‌حلی را بهینه کردن پاداش می‌اضافه رادنبال می‌کنند. اما در این الگوریتم

رادنبال راه‌حلی (سیاستی هستم) که هم سعی کند بیشترین پاداش ممکن را به دست آورد اما هزمان تا حدی هم بتواند رفتار تصادفی

پیش‌بینی و اکتشافی خود را حفظ کند. سه ویژگی اساسی: پایداری بیشتر / اکتشاف بیشتر

هدف SAC بیشینه کردن مجموع پاداش‌ها به همراه آنتروپی است:

$$J(\pi) = \sum_{t=0}^{T} E_{(s_t,a_t)\sim \rho_\pi}\left[r(s_t,a_t) + \alpha\, H(\pi(\cdot|s_t))\right]$$

(پاداش، ضریب اهمیت، آنتروپی سیاست)

حال به اجزای SAC می‌پردازیم: (البته sac بدون value):

۱- شبکه Actor (سیاست گذار): این شبکه state را گرفته و پارامترهای یک توزیع (گاوسی) شامل mean و std را برای

انتخاب عمل خروجی می‌دهد.

۲- شبکه critic (منتقد): این شبکه می‌خواهد $Q(s,a)$ را البته بصورت soft-max تخمین بزند ($H(X)$)،

توجه داریم استفاده از دو شبکه و انتخاب min بین آنها با جلوگیری از overfit شود.

۳- دو شبکه target-critic:

این شبکه‌ها کپی با تأخیر از critic هستند برای یادگیری خود critic و actor، این کپی باعث جلوگیری از

overfit می‌شود.

توابع های loss و روند یادگیری:

critic: تلاش می‌کند خطای MSE بین پیش‌بینی خود و یک مقدار هدف را کمینه کند:

$$y = r(s,a) + \gamma\left(\min_{i=1,2} Q_i(s',a') - \alpha \log \pi(a'|s')\right)$$

(y: هدف)

$$SSE = \sum (y - Q(s,a))^2$$

تابع هزینه Actor: براساس کمینه کردن واگرایی KL بین سیاست و توزیع ایده آل (Q-function) soft-max است:

$$L_{actor}(\phi) = E_{a\sim\pi_\phi}\left[\alpha \log \pi_\phi(a|s) - \min_{i=1,2} Q_{\theta_i}(s,a)\right]$$

به ست اعمالی رود که $Q(s,a)$ بیشترین یا آستروپی بیشتر شود.

در نسخه‌ی ریاضیات ده کردیم خود پارامتر α بیشتر پارامتر قابل یادگیری بود.

سوال ۲. نقش آنتروپی در GAIL با objective

سبب می‌شود که هدف پیدا کردن صرفاً بهترین مسیر به پیدا کردن مجموعه‌ای مسیرهای خوب و انعطاف‌پذیر تغییر کند.

a. آنتروپی $-\sum \pi(a|s) \log \pi(a|s)$ در پاداش باعث می‌شود که یک پاداش درونی برای کنجکاوی اعمال شود. نتیجه این کار می‌شود اینکه اکتشاف به مقدار کافی انجام می‌شود و احتمال گیر افتادن در optimal های local کاهش یابد.

b. افزایش پایداری و جلوگیری از overfit: یک سیاست با آنتروپی بالا انعطاف‌پذیر است، به‌جای یادگیری کامل و دقیق برای یک موقعیت خاص یاد می‌گیرد مجموعه‌ای از اعمال خوب وجود دارند.

بنابراین تضمین می‌شود که سیاست شکننده نباشد اگر محیطی تغییر کند یا اتفاق غیرمنتظره‌ای رخ دهد عامل گزینه‌های دیگری برای واکنش دارد.

سوال ۳ :

معادله بلمن با soft-max به صورت زیر است :

$$Q^*_{soft}(s,a) = r(s,a) + \gamma E_{s' \sim p, a' \sim \pi^*} [Q^*_{soft}(s',a') - \alpha \log \pi^*(a'|s')]$$

سیاست بهینه

با اعمال مکرر این اپراتور به این خاطر که Q-function یک تابع نقطه ثابت است (fined-point) به نقطه ثابت خود همگرا می شود. حالا این Q^* چه اطلاعاتی می دهد؟

ارزشی که با ورود از s و اجرای action=a و سپس دنبال کردن بهترین سیاست تعادلی که تعادل بین پاداش و آنتروپی را برقرار کند به دست می آید.

$$T^{\pi} Q(s,a) \triangleq r(s,a) + \gamma E_{s' \sim p}[V_{soft}(s')]$$

$$V_{soft}(s') = E_{a' \sim \pi}[Q(s',a') - \alpha \log \pi(a'|s')]$$

$T^{\pi} Q = Q$ این تابع زمانی به نقطه ثابت خود می رسد که اعمال اپراتور دیگر باعث تغییر نشود یعنی

و طبق قضیه نقطه ثابت اثبات می شود این مقدار، مقدار همگرایی است.

توجه داریم که یک گذشت انقباضی است ($\gamma < 1$) اثبات می شود به این دلیل همگرا به نقطه همگرایی می شود.

سوال ۴:

هدف به دست آوردن مقدار رو به رو است:

$$\text{Ideal policy}(a|s) = \frac{\exp(\frac{1}{\alpha} Q_\theta(s,a))}{Z_\theta}$$

این Z_θ جمع زدن روی exp ها است

و برای آن باید در فضای پیوسته انتگرال را حساب کنیم که خب این انتگرال به دست آوردنش سخت است

بنابراین از ابزار فاصله و کمینه کردن آن استفاده می‌کنیم

هدف ما این خواهد شد:

$$\min_\phi D_{KL}\left(\pi_\phi(a|s) \,\|\, \frac{\exp(\frac{1}{\alpha} Q(s,a))}{Z_\theta}\right)$$

فاصله

یا پارامترهای به دست آمده از ϕ را باید طوری انتخاب کرد که توزیع خروجی آن کمترین فاصله را با مقدار ایده‌آل داشته باشد.

در ریاضیات داریم:

$$D_{KL}(p\|Q) = E_{x\sim p}[\log p - \log Q(x)]$$

پس برای مسئله ما خواهیم داشت:

$$D_{KL} = E_{a\sim\pi_\phi}\left[\log \pi_\phi(a|s) - \log\left(\frac{e^{\frac{1}{\alpha}Q(s,a)}}{Z_\theta}\right)\right]$$

$$= E_{a\sim\pi_\phi}\left[\log \pi_\phi(a|s) - \log(e^{\frac{1}{\alpha}Q_\phi(s,a)}) + \log Z_\theta\right]$$

↓

ربطی به ϕ ندارد پس از حساب کردن حذف می‌شود و داریم:

$$= E_{a\sim\pi_\phi}\left[\log \pi_\phi(a|s) - \frac{1}{\alpha} Q_\phi(s,a)\right]$$

$$= E_{a\sim\pi_\phi}\left[\alpha \log \pi_\phi(a|s) - Q_\phi(s,a)\right]$$

و این همان تابع خطایی است که در Actor داریم یعنی باید مینیمم کنیم:

$$E_{a\sim\pi_\phi}\left[\alpha \log \pi_\phi(a|s) - Q_\phi(s,a)\right]$$