تمرین سوم – یادگیری تقویتی

دانشجو: سایه جارالهی

شماره دانشجویی: ۹۸۱۰۱۳۳۹

استاد: دکتر رهبان- آقای حسنی

بهار ۱۴۰۲

سوال یک-

بخش الف –

همانطور که میدانیم LQR یک روش کنترل است که در آن cost function به صورت یک تابع quadratic مدل میشود و دینامیک محیط نیز به صورت یک تابع linear مدل میشود. پس از آن optimization انجام میشود تا سری اکشن ها به دست آیند (درواقع ماتریس های K و k به دست آیند که بتوان اکشن را به دست آورد) باتوجه به تابع هایی که ارائه دادیم، فرض کردمی که دینامیک مدل با تابع خطی مدل می شود، در نتیجه اگر در واقعیت دینامیک مدل بسیار پیچیده تر باشد LQR به طور مستقیم همگرا نمیشود و باید ابتدا خطی سازی داینامیک را انجام دهیم. همچنین تابع درجه دو برای cost function مشخص شده و در صورتی که cost function بسیار پیچیده تر از این تابع باشد لزوما همگرا به جواب بهینه نمیشویم.

یکی دیگر از موارد fully observable بودن محیط است. در صورتی که محیط کاملا شناخته شده نباشد، مدل کردن دینامیک با تابع خطی کار نادرستی است و باعث میشود به جواب بهینه همگرا نشویم زیرا برخی اکشن ها و استیت ها ناشناخته هستند و در نتیجه با انجام یک اکشن ممکن است به جای کاملا متفاوتی حرکت کنیم.

در روش LQR ساده، در نظر میگیریم که محیط stable است. درصورتی که رندومنس داشته باشیم و محیط stable نباشد در ابتدا باید محیط را stable کنیم.

همچنین برای ایتکه LQR پاسخ داشته باشد، در برخی از مراحل باید واردن ماتریس C به دست آید. در نتیجه برای آنکه جواب داشته باشد باید ماتریس cost ها مثبت معین باشد.

بخش ب-

طبق مقاله ای که قید شده است، یکی از راه های استفاده از LQR در محیط های partially observable آن است که با استفاده ازKalman Filter بخش هایی را که unobserved هستند تخمین بزنیم. در واقع Kalman Filter با استفاده از سری داده هایی که در طول زمان مشاهده کرده است، استیت یک پراسس را تخمین میزند، به صورت یک تابع بازگشتی پیاده سازی شده است و توسط داینامیکی که مشاهده میکند و مقادیری که دارد استیت را تخمین میزند.

درواقع میتوان گفت از LQG(Linear quadratic-Gaussian) استفاده میشود. در این روش از Kalman filter برای تخمین استفاده میشود. همچنین از یک noise گاوسی مشابه آنچه در کلاس درس دیدیم استفاده میشود تا در صورتی که Kalman filter تخمین بدون نویز را در نظر میگیرد، یک نویز نیز به آن اضافه کنیم. با این دو کار و به کمک روش LQR (پس از تخمین استیت) به خروجی نهایی میرسیم. منبع

بخش ج-

همانطور که میدانیم LQR یک روش model-based است. میتوان مطابق مثالی که در اسلایدها برای ترکیب روش های model-free و model-based برای Half-cheetah زده شد(در آن مثال در ابتدا به روش بر پایه مدل ترین انجام میشد و سپس با روش model-free) ، عمل کرد.

به این صورت که در ابتدا LQR به عنوان یک پایه در نظر گرفته شود و از خروجی آن خروجی بهینه به دست آید. در گام بعدی توسط یک روش model-free در RL یک correction term به دست آید تا در نقاطی که LQR ضعف دارد، خروجی و اکشن ها را اصلاح کند. یکی از مقالاتی که در این رابطه استفاده میشود در لینک آورده شده است.

همچنین یک روش دیگر آن است که خروجی LQR را با model-free با نح خاصی ترکیب کنیم. در واقع این دو جداگانه آموزش داده شوند و خروجی آن ها به دست آید و اکشن نهایی به صورت ضریبی از این دو باشد و در واقع weighted sum گرفته شود و یا با روش های دیگر ترکیب شوند.

بخش د-

مشابه با آنچه در کلاس درس دیدیم، در روش iLQR با استفاده از اعمال یک توزیع گاوسی برای استیت بعدی میتوان exploration را هدایت کرد و یا uncertainty درسیستم را مدل کرد. برای اینکار میتوان یک نویز گاوسی را به dynamic مسئله اضافه کرد که

میانگین آن با استفاده از تابع های داینامیک تعریف شده باشد و ماتریس کواریانس نیز به اعمال نویز کمک میکند. با اینکار و با تغییر ماتریس کواریانس به مقادیر مناسب، میتوان exploration را تقویت کرد زیرا استیت های بعدی که برمیگردند دارای کواریانس بیشتری هستند و به exploration کمک میکند. همچنین برای مدل کردن uncertainty محیط نیز مشابه با LQR میتواند مفید باشد زیرا به ازای استیت و اکشن، استیت بعدی از یک توزیع می آید و همیشه مقدار مشخصی ندارد.

الف) $r \sim Gamma(\alpha, \beta) \rightarrow p(r|\alpha,\beta) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} r^{\alpha-1} e^{-\beta r}$

$t \sim exp(\lambda) \longrightarrow p(t|\lambda) = \lambda e^{-\lambda t}$

روبرو : $(t=5, r=200)$ $(t=15, r=100)$

$$\begin{cases} \beta \sim Gamma(\varepsilon, \omega) \\ \lambda \sim Gamma(\delta, \eta) \end{cases}$$

سرویس (t_1, r_1)

همان طور که گفته شد، با بیشینه درست نمایی، توزیع پیشین را به دست می‌آوریم

با توجه به استقلال تجربه ها می‌توان گفت:

$$P(R|\alpha,\beta) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} \times (100)^{\alpha-1} \times e^{-100\beta} \times \frac{\beta^{\alpha}}{\Gamma(\alpha)} \times (200)^{\alpha-1} e^{-200\beta}$$

$$= \left(\frac{\beta^{\alpha}}{\Gamma(\alpha)}\right)^2 \times (100 \times 200)^{\alpha-1} \times e^{-300\beta} = A$$

فقط نسبت به β بیشینه می‌کنیم ⇐ MLE

$$\frac{dA}{d\beta} = \frac{(20000)^{\alpha-1}}{\Gamma(\alpha)^2} \left[2\alpha \beta^{2\alpha-1} e^{-300\beta} - 300 \beta^{2\alpha} e^{-300\beta} \right] = 0$$

$$\Rightarrow 2\alpha \beta^{2\alpha-1} e^{-300\beta} = 300 \beta^{2\alpha} e^{-300\beta}$$

$$\boxed{\frac{\alpha}{150} = \beta}$$

$$P(T|\lambda) = \lambda e^{-10\lambda} \times \lambda e^{-0\lambda} = \lambda^2 e^{-20\lambda} = B$$

MLE: $\frac{dB}{d\lambda} = 2\lambda e^{-20\lambda} - 20\lambda^2 e^{-20\lambda} = 0$

$2\lambda e^{-20\lambda} = 20\lambda^2 e^{-20\lambda}$

$$\lambda = \frac{1}{T_c}$$

حال محاسبه می‌کنیم

$$p(\beta \mid r_1, \alpha, \varepsilon, w)$$

مسأله آمده در کلاس درس حل کنیم. اما r_1 را دست می‌زنیم و سپس ساده‌سازی می‌کنیم

$$= \frac{p(\beta, r_1 \mid \alpha, \varepsilon, w)}{p(r_1 \mid \alpha, \varepsilon, w)} = \frac{p(r_1 \mid \alpha, \beta, \varepsilon, w)\, p(\beta \mid \alpha, \varepsilon, w)}{p(r_1 \mid \alpha, \varepsilon, w)}$$

به دلیل استقلال می‌توان ساده‌سازی کرد

$$= \frac{p(r_1 \mid \alpha, \beta)\, p(\beta \mid \varepsilon, w)}{p(r_1 \mid \alpha, \varepsilon, w)} = \frac{\beta^{\alpha}}{\Gamma(\alpha)} r_1^{\alpha-1} e^{-\beta r} \times \frac{w^{\varepsilon}}{\Gamma(\varepsilon)} \frac{\beta^{\varepsilon-1} e^{-w\beta}}{p(r_1 \mid \alpha, \varepsilon, w)}$$

صرف‌نظر از مخرج

$$= \frac{\beta^{\alpha+\varepsilon-1} \times r_1^{\alpha-1} \times e^{-(w+r_1)\beta} \times w^{\varepsilon}}{\Gamma(\alpha)\Gamma(\beta) \times p(r_1 \mid \alpha, \varepsilon, w)} \quad (\text{I})$$

$$(*)\; p(r_1 \mid \alpha, \varepsilon, w) = \int_{\beta} p(\beta \mid \alpha, \varepsilon, w)\, p(r_1 \mid \alpha, \beta, \varepsilon, w)\, d\beta$$

$$= \int_{\beta} \frac{w^{\varepsilon}}{\Gamma(\varepsilon)} \beta^{\varepsilon-1} e^{-w\beta}\, \frac{\beta^{\alpha} r_1^{\alpha-1} e^{-\beta r_1}}{\Gamma(\alpha)}$$

$$= \boxed{\frac{w^{\varepsilon} r_1^{\alpha-1}}{\Gamma(\varepsilon)\Gamma(\alpha)} \int_{\beta} \beta^{\alpha+\varepsilon-1} e^{-(w+r_1)\beta}}$$

در نتیجه برای ⓛ داریم:

$$\frac{\beta^{\alpha+\varepsilon-1} \times e^{-(w+r_1)\beta}}{\int_\beta \beta^{\alpha+\varepsilon-1} e^{-(w+r_1)\beta}}$$

(با کمک حالت بالا، سعیمان را ...) از آن جا که عبارت بالا شبیه توزیع گاما است، صورت و مخرج را $\frac{(r_1+w)^{\alpha+\varepsilon}}{\Gamma(\alpha+\varepsilon)}$ ضرب می‌کنیم. مخرج دقیقا توزیع $\int_\beta Gamma(\alpha+\varepsilon, r_1+w)$ می‌شود که چون انتگرال روی تمام مقادیر توزیع است یک می‌شود. در نتیجه فقط صورت به صورت زیر باقی می‌ماند:

$$\frac{(r_1+w)^{\alpha+\varepsilon}}{\Gamma(\alpha+\varepsilon)} \beta^{\alpha+\varepsilon-1} e^{-(w+r_1)\beta}$$

این توزیع همان $Gamma(\alpha+\varepsilon, r_1+w)$ است

$$\Rightarrow \begin{cases} \varepsilon' = \alpha+\varepsilon \\ w' = w+r_1 \end{cases}$$

✓

Subject:
Date

$t \sim \exp(\lambda) \rightarrow p(t|\lambda) = \lambda e^{-\lambda t}$

$\lambda \sim \text{Gamma}(\alpha, \eta)$

$$p(\lambda | t_1, \alpha, \eta) = \frac{p(\lambda, t_1 | \alpha, \eta)}{p(t_1 | \alpha, \eta)} = \frac{p(\lambda|\alpha,\eta)\, p(t_1|\alpha,\eta,\lambda)}{p(t_1|\alpha,\eta)}$$

$$= \frac{\frac{\eta^\alpha \lambda^{\alpha-1} e^{-\lambda\eta}}{\Gamma(\alpha)} \times \lambda e^{-\lambda t_1}}{p(t_1|\alpha,\eta)} = \frac{\eta^\alpha \lambda^\alpha e^{-\lambda(\eta+t_1)}}{\Gamma(\alpha)\, p(t_1|\alpha,\eta)} = (\text{II})$$

$$(**)\quad p(t_1|\alpha,\eta) = \int_\lambda p(t_1, \lambda|\alpha,\eta) = \int_\lambda p(\lambda|\alpha,\eta)\, p(t_1|\alpha,\eta,\lambda)\, d\lambda$$

$$= \int_\lambda \frac{\eta^\alpha \lambda^\alpha e^{-\lambda(\eta+t_1)}}{\Gamma(\alpha)} d\lambda = \frac{\eta^\alpha}{\Gamma(\alpha)} \int_\lambda e^{-\lambda(\eta+t_1)} \lambda^\alpha d\lambda$$

$$(**) \Rightarrow (\text{II}) = \frac{\lambda^\alpha e^{-\lambda(\eta+t_1)}}{\int_\lambda e^{-\lambda(\eta+t_1)} \lambda^\alpha d\lambda}$$

طبق بخش قبل روابط مقابل صورت و مخرج ساده $\frac{(\eta+t_1)^{\alpha+1}}{\Gamma(\alpha+1)}$ ضرب کرد تا مخرج یک شود و این

صورت را شبیه آن کنیم

$$\lambda^\alpha e^{-\lambda(\eta+t_1)} \frac{(\eta+t_1)^{\alpha+1}}{\Gamma(\alpha+1)} \sim \text{Gamma}(\alpha+1, t_1+\eta)$$

$$\Rightarrow \begin{cases} \alpha' = \alpha+1 \\ \eta' = \eta + t_1 \end{cases}$$

$$p(t_2 \mid t_1) = \int_\lambda p(t_2 \mid t_1, \lambda)\, \underbrace{p(\lambda \mid t_1)}_{\text{همان posterior است}}\, d\lambda \qquad (\sim$$

$$= \int \overbrace{p(t_2 \mid t_1, \lambda)}^{\text{prior ما}} \times \frac{\eta'^{\alpha'} \lambda^{\alpha'-1} e^{-\eta'\lambda}}{\Gamma(\alpha')}$$

$$= \int_\lambda \lambda e^{-\lambda t_2} \frac{\eta'^{\alpha'} \lambda^{\alpha'-1} e^{-\eta'\lambda}}{\Gamma(\alpha')}$$

$$= \frac{\eta'^{\alpha'}}{\Gamma(\alpha')} \int_0^\infty \lambda^{\alpha'} e^{-\lambda(t_2+\eta')}\, d\lambda \qquad \text{(I)}$$

$$= \frac{\eta'^{\alpha'}}{\Gamma(\alpha')} \left[\cancel{\frac{\lambda^{\alpha'} e^{-\lambda(t_2+\eta')}}{-(t_2+\eta')}}\Big|_0^\infty - \frac{(\alpha')}{-(t_2+\eta')} \int_0^\infty \lambda^{\alpha'-1} e^{-\lambda(t_2+\eta')}\, d\lambda \right]$$

$$= \frac{\eta'^{\alpha'}}{(\alpha'-1)!} \times \frac{\alpha'}{(t_2+\eta')} \underbrace{\int_0^\infty \lambda^{\alpha'-1} e^{-\lambda(t_2+\eta')}\, d\lambda}_{X}$$

از مقایسه (I) و عبارت بالا در می‌یابیم که عبارت داخل انتگرال رابطه بازگشتی را برای X نوشت و تنها
ضریب λ یکی کم شد. در نتیجه می‌توان گفت با مقدار از تکرار نوشتن عبارت‌ها و انتگرال جزء به جزء داریم

$$= \frac{\eta'^{\alpha'}}{\cancel{(\alpha'-1)!}} \times \frac{\cancel{\alpha'!}}{(t_2+\eta')^{\alpha'+1}} = \frac{\eta'^{\alpha'} \times \alpha'}{(t_2+\eta')^{\alpha'+1}} = \frac{\eta'^{\alpha'} \times \alpha'}{(\eta')^{\alpha'+1} \left(\frac{t_2}{\eta'}+1\right)^{\alpha'+1}}$$

$$= \frac{\alpha'}{\eta'} \times \left(\frac{t_2}{\eta'}+1\right)^{-(\alpha'+1)}$$

Subject:
Date

ج) اگر بدترین رفتار را معادل این بگیریم که همیشه آن را مانند همیشه پیش‌بینی کند، ترجیح می‌دهد که هیچ وقت بازی نکند و کارش را ادامه دهد.

اما وقتی ریسک‌پذیر باشد، چون می‌داند که با احتمال نزدیک به صفر ممکن است جایزه بالایی ببرد، همین احتمال (هر چند کم) برای او کافی است که وارد بازی شود

اما وقتی ریسک خنثی است، باید ببینید چطور میانگین جایزه تقسیم می‌شود.

وقتی ریسک خنثی است، با توجه به باورش به پیدا کردن (posterior)، باید ببینیم در یک ساعت چطور میانگین جایزه reward ممکن است تغییر کند. درنتیجه یک راه این است که $E(t)$ را محاسبه کنیم، سپس $E(r)$ را نیز با توجه به posterior محاسبه کنیم. حال اگر $E(r) < K \times E(t)$ شده باشد ترک کند، در غیر این صورت باید به بازی ادامه دهد. توجه داریم که فرض کردیم t برحسب ساعت بررسی شده است

PAPCO