یادگیری تقویتی عمیق
دکتر رهبان

دانشگاه صنعتی شریف
دانشکده مهندسی کامپیوتر

امیر کوشان فتاح حصاری ۴۰۱۱۰۲۱۹۱

تمرین سری ۹
Bandit ها و یادگیری برخط

۱۶ خرداد ۱۴۰۴

یادگیری تقویتی عمیق

تمرین سری ۹

امیر کوشان فتاح حصاری ۴۰۱۱۰۲۱۹۱

---

۱ توزیع های **light-tailed**

نابرابری هوفدینگ

بخش اول : لم هوفدینگ

اگر یک متغیر تصادفی X با $\mathbf{E}[X] = 0$ داشته باشیم و بدانیم که $a \leq X \leq b$ ، با استفاده از تعریف زیر رابطه خواسته شده را اثبات میکنیم.
ابتدا تابع $\phi(s)$ را به شکل زیر تعریف میکنیم :

$$
\begin{aligned}
&\phi(s) = \log \mathbf{E}\left[e^{sX}\right] \\
&\phi'(s) = \frac{\partial \log \mathbf{E}\left[e^{sX}\right]}{\partial s} = \frac{\frac{\partial \mathbf{E}\left[e^{sX}\right]}{\partial s}}{\mathbf{E}\left[e^{sX}\right]} = \frac{\mathbf{E}\left[X \cdot e^{sX}\right]}{\mathbf{E}\left[e^{sX}\right]} = \int X \cdot \underbrace{\frac{e^{sX} d\mathbf{P}(X)}{\mathbf{E}\left[e^{sX}\right]}}_{d\mathbf{P_s}(X)} \\
&\qquad = \mathbf{E}_{x\sim \mathbf{P}_s(X)}\left[X\right] \\
&\phi''(s) = \frac{\partial}{\partial s}\mathbf{E}_{x\sim \mathbf{P_s}(X)}\left[X\right] = \frac{\mathbf{E}\left[X^2 \cdot e^{sX}\right]\mathbf{E}\left[e^{sX}\right] - \left(\mathbf{E}\left[X \cdot e^{sX}\right]\right)^2}{(\mathbf{E}\left[e^{sX}\right])^2} \\
&\qquad = \frac{\mathbf{E}\left[X^2 \cdot e^{sX}\right]}{\mathbf{E}\left[e^{sX}\right]} - \left(\frac{\mathbf{E}\left[X \cdot e^{sX}\right]}{\mathbf{E}\left[e^{sX}\right]}\right)^2 \\
&\phi''(s) = \mathbf{E}_{x\sim \mathbf{P_s}(X)}\left[X^2\right] - \left(\mathbf{E}_{x\sim \mathbf{P_s}(X)}\left[X\right]\right)^2 = \mathbb{V}_{x\sim \mathbf{P_s}(X)}\left[X\right]
\end{aligned}
$$

همچنین میدانیم که :

$$
\begin{aligned}
\mathbb{V}\left[X\right] = \mathbb{V}\left[X - (\frac{a+b}{2})\right] &= \mathbf{E}\left[(X - (\frac{a+b}{2}))^2\right] - \left(\mathbf{E}\left[X - (\frac{a+b}{2})\right]\right)^2 \\
&\leq \mathbf{E}\left[\left(X - (\frac{a+b}{2})\right)^2\right]
\end{aligned}
$$

امید ریاضی نهایی در رابطه بالا زمانی ماکسیموم میشود که متغیر تصادفی X یکی از مقادیر اکستریم (یعنی a یا b) را اختیار کند. در این صورت داریم که :

$$
\begin{aligned}
&X = a \Rightarrow (a - \frac{a+b}{2})^2 = (b - \frac{a+b}{2})^2 = (\frac{b-a}{2})^2 \\
&\mathbf{E}\left[\left(X - \frac{a+b}{2}\right)^2\right] \leq \left(\frac{b-a}{2}\right)^2 = \frac{(b-a)^2}{4} \Rightarrow \mathbb{V}\left[X\right] \leq \frac{(b-a)^2}{4}
\end{aligned}
$$

و در نهایت داریم : [۱]

لم هوفدینگ

$$\phi(s) = \int\int \phi''(s) = \int_0^s \int_0^\mu \mathbb{V}_{x\sim\mathbf{P_q}(X)}\left[X\right] dq\, d\mu \leq \int_0^s \int_0^\mu \frac{(b-a)^2}{4} dq\, d\mu$$
$$= \int_0^s \frac{\mu(b-a)^2}{4} d\mu$$
$$= \frac{s^2(b-a)^2}{8}$$

$$\boxed{\phi(s) = \log \mathbf{E}\left[e^{sX}\right] \leq \frac{s(b-a)^2}{8} \Rightarrow \mathbf{E}\left[e^{sX}\right] \leq \exp(\frac{s^2(b-a)^2}{8})}$$

بخش دوم : نابرابری هوفدینگ

از تعریف توابع زیر گاوسی با پارامتر σ [۲] میدانیکه که در رابطه زیر صدق میکنند :

$$\mathbf{E}\left[e^{\lambda x}\right] \leq \exp(\frac{\lambda^2\sigma^2}{2})$$

حال با استفاده از نابرابری مارکف و روش کرامر–چرنف [۳] رابطه زیر را اثبات میکنیم که در روند اثبات نابرابری هوفدینگ استفاده خواهد شد :

$$\mathbf{P}(X \geq \epsilon) \leq \exp(\frac{-\epsilon^2}{2\sigma^2})$$

$$\mathbf{P}(X \geq \epsilon) = \mathbf{P}(\lambda X \geq \lambda\epsilon) = \mathbf{P}(\exp(\lambda X) \geq \exp(\lambda\epsilon)) \leq \frac{\mathbf{E}\left[\exp(\lambda X)\right]}{\exp(\lambda\epsilon)} \quad \text{Markov}$$
$$\leq \exp(\frac{\lambda^2\sigma^2}{2} - \lambda\epsilon) \quad \text{Def. of subgaussianity}$$

حال اگر λ را برابر با $\frac{\epsilon}{\sigma^2}$ در نظر بگیریم داریم که :

$$\exp(\frac{\lambda^2\sigma^2}{2} - \lambda\epsilon) = \exp(\frac{\epsilon^2\sigma^2}{2\sigma^4} - \frac{\epsilon^2}{\sigma^2}) = \exp(-\frac{\epsilon^2}{2\sigma^2})$$
$$\Rightarrow \mathbf{P}(X \geq \epsilon) \leq \exp(-\frac{\epsilon^2}{2\sigma^2}) \qquad \blacksquare$$

همچنین اگر متغیر های تصادفی X_i را داشته باشیم که مستقل و زیر گاوسی با پارامتر σ_i باشند آنگاه داریم که :

$$q = \sum_i X_i$$

$$\mathbf{E}\left[e^{\lambda q}\right] \overset{\text{independent}}{=\!=\!=} \prod_i \mathbf{E}\left[e^{\lambda X_i}\right] \leq \exp(\frac{\lambda^2 \sum \sigma_i^2}{2}) \Rightarrow q \sim \mathcal{SG}(\sqrt{\sum_i \sigma_i^2})$$

[۱] رابطه ای که بالاتر برای پیدا کردن کران بالایی برای واریانس استفاده کردیم به نام نابرابری Popoviciu's بر روی واریانس ها شناخته میشود.

[۲] σ-subgaussian

[۳] Cramer–Chernoff method

همچنین cX توزیع زیر گاوسی با پارامتر $|c|\sigma$ خواهد بود.
حال اگر متغیر های تصادفی $Z_i \; \forall i \in \{1, 2, ..., n\}$ کران دار باشند و در بازه $[a, b]$ قرار بگیرند آنگاه داریم که :

$$\mathbf{P}(Z_i \geq \epsilon) \leq \exp(-\frac{\epsilon^2}{2\sigma^2})$$

از بخش اول میدانیم که چنین متغیر های تصادفی زیر گاوسی با پارامتر $\frac{(b-a)}{2}$ می باشند. پس داریم که :

نابرابری هوفدینگ (طرف راست)

$$\mathbf{P}(Z_i \geq \epsilon) \leq \exp(-\frac{\epsilon^2}{2\sigma^2}) = \exp(-\frac{\epsilon^2 \cdot 2}{(b-a)^2})$$

$$\mathbf{P}(\sum_i \overbrace{(Z_i - \mathbf{E}\left[Z_i\right])}^{q_i} \geq \epsilon) \; = \mathbf{P}(\sum_i q_i \geq \epsilon) \leq \exp(-\frac{\epsilon^2}{2 \cdot (\sqrt{\sum_i \sigma_i^2})^2}) = \exp(-\frac{4 \cdot \epsilon^2}{2 \cdot n(b-a)^2})$$

$$\mathbf{P}(\frac{1}{n}\sum_i q_i \geq \epsilon) \leq \exp(-\frac{n^2 \cdot 4 \cdot \epsilon^2}{2 \cdot n \cdot (b-a)^2}) = \exp(-\frac{n \cdot 2 \cdot \epsilon^2}{(b-a)^2})$$

$$\boxed{\mathbf{P}(\frac{1}{n}\sum_i (Z_i - \mathbf{E}\left[Z_i\right]) \geq \epsilon) \leq \exp(-\frac{2 \cdot n \cdot \epsilon^2}{(b-a)^2})}$$

برای طرف چپ معادله نیز با تغییر متغیر زیر جلو میرویم :

$$Y_i = -Z_i\,; \quad -b \leq Y_i \leq -a\,; \quad \mathbf{E}\left[Y_i\right] = -\mathbf{E}\left[Z_i\right]$$

$$\frac{1}{n}\sum_i Z_i - \mathbf{E}\left[Z_i\right] \leq -\epsilon \Rightarrow \frac{1}{n}\sum_i Y_i - \mathbf{E}\left[Y_i\right] \geq \epsilon$$

$$\mathbf{P}(\frac{1}{n}\sum_i Z_i - \mathbf{E}\left[Z_i\right] \leq -\epsilon) = \mathbf{P}(\frac{1}{n}\sum_i Y_i - \mathbf{E}\left[Y_i\right] \geq \epsilon) \leq \exp(-\frac{2 \cdot n \cdot \epsilon^2}{(b-a)^2})$$

نابرابری هوفدینگ (طرف چپ)

$$\mathbf{P}(\frac{1}{n}\sum_i Z_i - \mathbf{E}\left[Z_i\right] \leq -\epsilon) \leq \exp(-\frac{2 \cdot n \cdot \epsilon^2}{(b-a)^2})$$

زیر گاوسی

بخش اول

سه رابطه زیر برای متغیر های زیر گاوسی را اثبات میکنیم :
برای رابطه اول از روش کرامر_چرنف که در بخش قبل نیز بیان شد استفاده میکنیم :

$$\mathbf{P}(\underbrace{X-\mathbf{E}[X]}_{Z} \geq \epsilon) \leq \exp(\frac{-\epsilon^2}{2\sigma^2})$$

$$\mathbf{P}(Z \geq \epsilon) = \mathbf{P}(\lambda Z \geq \lambda\epsilon) = \mathbf{P}(\exp(\lambda Z) \geq \exp(\lambda\epsilon)) \leq \frac{\mathbf{E}[\exp(\lambda Z)]}{\exp(\lambda\epsilon)} \quad \text{Markov}$$

$$\leq \exp(\frac{\lambda^2\sigma^2}{2} - \lambda\epsilon) \quad \text{Def. of subgaussianity}$$

سمت راست

$$\lambda = \frac{\epsilon}{\sigma^2}; \forall\, \epsilon > 0$$

$$\exp(\frac{\lambda^2\sigma^2}{2} - \lambda\epsilon) = \exp(\frac{\epsilon^2\sigma^2}{2\sigma^4} - \frac{\epsilon^2}{\sigma^2}) = \exp(-\frac{\epsilon^2}{2\sigma^2})$$

$$\Rightarrow \mathbf{P}(X-\mathbf{E}[X] \geq \epsilon) \leq \exp(-\frac{\epsilon^2}{2\sigma^2}) \quad \blacksquare$$

برای رابطه دوم که برای دم سمت چپ توزیع میباشد داریم که :

$$\mathbf{P}(X < \mathbf{E}[X]-\epsilon) = \mathbf{P}(\underbrace{X-\mathbf{E}[X]}_{Z} < -\epsilon) = \mathbf{P}(Z < -\epsilon)$$

$$\mathbf{P}(Z < -\epsilon) \xrightarrow{\lambda<0} \mathbf{P}(\lambda Z > -\lambda\epsilon) = \mathbf{P}(\exp(\lambda Z) > \exp(-\lambda\epsilon)) \leq \frac{\mathbf{E}[\lambda Z]}{\exp(-\lambda\epsilon)} \quad (\text{Markov})$$

$$(\text{Def. of Subgaussianity}) \quad \leq \frac{\exp(\frac{\lambda^2\sigma^2}{2})}{\exp(-\lambda\epsilon)} = \exp(\frac{\lambda^2\sigma^2}{2} + \lambda\epsilon)$$

سمت چپ

$$\lambda = \frac{-\epsilon}{\sigma^2}; \forall\, \epsilon > 0$$

$$\exp(\frac{\epsilon^2\sigma^2}{2\sigma^2} - \frac{\epsilon^2}{\sigma^2}) = \exp(-\frac{\epsilon^2}{2\sigma^2})$$

$$\boxed{\Rightarrow \mathbf{P}(Z < -\epsilon) = \mathbf{P}(X < \mathbf{E}[X]-\epsilon) \leq \exp(-\frac{\epsilon^2}{2\sigma^2}) \quad \blacksquare}$$

برای رابطه سوم از قضیه کران اجتماع[۴] استفاده میکنیم. داریم که :

$$\mathbf{P}\left(\bigcup_i A_i\right) \leq \sum_i \mathbf{P}\left(A_i\right)$$

$$\begin{aligned}\mathbf{P}\left(\left|X-\mathbf{E}\left[X\right]\right| \geq \epsilon\right) &= \mathbf{P}\left(X-\mathbf{E}\left[X\right] \geq \epsilon \bigcup -X+\mathbf{E}\left[X\right] \geq \epsilon\right)\\ &= \mathbf{P}\left(X-\mathbf{E}\left[X\right] \geq \epsilon \bigcup X-\mathbf{E}\left[X\right] \leq -\epsilon\right)\end{aligned}$$

اجتماع دو رابطه

$$\mathbf{P}\left(X-\mathbf{E}\left[X\right] \geq \epsilon \bigcup X-\mathbf{E}\left[X\right] \leq -\epsilon\right) \leq \mathbf{P}\left(X-\mathbf{E}\left[X\right] \geq \epsilon\right)+\mathbf{P}\left(X-\mathbf{E}\left[X\right] \leq -\epsilon\right)$$

$$\mathbf{P}\left(\left|X-\mathbf{E}\left[X\right]\right| \geq \epsilon\right) \leq 2\exp(-\frac{\epsilon^2}{2\sigma^2})$$

بخش دوم : نابرابر هوفدینگ

همانطور که در اثبات نابرابر هوفدینگ در سوال اول دیدیم ، و با استفاده از رابطه سوم بخش اول همین سوال داریم که :

$$\mathbf{P}\left(\left|X_i-\mathbf{E}\left[X_i\right]\right| \geq \epsilon\right) \leq 2\exp(-\frac{\epsilon^2}{2\sigma_i^2})$$

و میدانیم که مجموع n تا متغیر زیر گاوسی با پارامتر های σ_i در نهایت خود یک توزیع زیر گاوسی با پارامتر $\sqrt{\sum_i \sigma_i^2}$ می باشد. با استفاده از این نکته داریم :

نابرابری هوفدینگ سوال ۲

$$\mathbf{P}(\sum_i |X_i-\mu_i| \geq \epsilon) \leq 2\exp(-\frac{\epsilon^2}{2\sum_i \sigma_i^2})$$

[۴] Union bound

بخش سوم

اگر متغیر های X_i از توزیع زیر گاوسی با پارامتر σ^2 پیروی بکنند داریم که :

$$q = \sum_i X_i;\ \mathbf{E}\left[e^{\lambda q}\right] \overset{\text{independent}}{=\!=\!=} \prod_i \mathbf{E}\left[e^{\lambda X_i}\right] \leq \exp(\frac{\lambda^2 \sum \sigma_i^2}{2}) \Rightarrow q \sim \mathcal{SG}(\sqrt{\sum_i \sigma_i^2})$$

$$\mathbf{P}(X_i \geq \epsilon) \leq \exp(-\frac{\epsilon^2}{2\sigma^2})$$

$$\mathbf{P}(\sum_i \overbrace{(X_i - \mathbf{E}\left[X_i\right])}^{q_i} \geq \epsilon) = \mathbf{P}(\sum_i q_i \geq \epsilon) \leq \exp(-\frac{\epsilon^2}{2\cdot(\sqrt{\sum_i \sigma_i^2})^2}) = \exp(-\frac{\epsilon^2}{2\cdot\sigma^2\cdot n})$$

$$\mathbf{P}(\frac{1}{n}\sum_i (X_i - \mathbf{E}\left[X_i\right]) \geq \epsilon) \xrightarrow{cX\sim\mathcal{SG}(|c|\sigma)} \mathbf{P}(\frac{1}{n}\sum_i (X_i - \mathbf{E}\left[X_i\right]) \geq \epsilon) \leq \exp(-\frac{\epsilon^2}{2\cdot\frac{\sigma^2}{n^2}\cdot n})$$

اثبات بخش سوم

$$\mathbf{P}(\frac{1}{n}\sum_i (X_i - \mathbf{E}\left[X_i\right]) \geq \epsilon) \leq \exp(-\frac{n\epsilon^2}{2\cdot\sigma^2})$$

حال اگر این کران بالا را برابر با δ بگیریم داریم که :

$$\delta = \exp(-\frac{n\epsilon^2}{2\sigma^2})$$

$$\log\delta = -\frac{n\epsilon^2}{2\sigma^2} \Rightarrow 2\sigma^2\log(\frac{1}{\delta}) = n\epsilon^2$$

$$\epsilon = \sqrt{\frac{2\sigma^2\log(\frac{1}{\delta})}{n}}$$

$$\Rightarrow \mathbf{P}(\frac{1}{n}\sum_i (X_i - \mathbf{E}\left[X_i\right]) \geq \sqrt{\frac{2\sigma^2\log(\frac{1}{\delta})}{n}}) \leq \delta$$

رابطه دوم بخش سوم

احتمال اینکه عبارت $\frac{1}{n}\sum_i (X_i - \mathbf{E}\left[X_i\right])$ از $\sqrt{\frac{2\sigma^2\log(\frac{1}{\delta})}{n}}$ بیشتر باشد از δ کمتر هست. پس احتمال اینکه $\frac{1}{n}\sum_i (X_i - \mathbf{E}\left[X_i\right])$ از $\sqrt{\frac{2\sigma^2\log(\frac{1}{\delta})}{n}}$ کمتر باشد از $1-\delta$ بیشتر هست.

$$\mathbf{P}(\frac{1}{n}\sum_i (X_i - \mathbf{E}\left[X_i\right]) < \sqrt{\frac{2\sigma^2\log(\frac{1}{\delta})}{n}}) > 1-\delta$$

۲ UCB

الگوریتم Upper Condfidence Bound

بخش اول : اثبات لم تجزیه regret

ابتدا چند تعریف را بیان میکنیم . در هر مرحله یک عمل از بین n عمل موجود را میتوانیم انتخاب کنیم. پس در هر قدم زمان t حاصل جمع زیر برابر با یک می باشد:

$$\sum_{a\in\mathcal{A}} \mathbb{I}\{A_t = a\} = 1$$

و به ازای انجام هر عمل در هر مرحله (یا قدم) یک جایزه به مقدار X_t دریافت میکنیم. مجموع این جایزه ها را بعد از n قدم با S_n نشان میدهیم :

$$S_n = \sum_t X_t = \sum_t \sum_{a\in\mathcal{A}} X_t \cdot \mathbb{I}\{A_t = a\}$$

همچنین تفاوت بین جایزه دریافتی و جایزه بهینه را در هر مرحله به شکل زیر بیان میکنیم :

$$\Delta_a = \mu^* - \mu_a(V)$$

که μ_a اینجا میانگین جایزه دریافتی از arm_a هست که در محیط v قرار دارد. همچنین μ^* بهترین جایزه دریافتی از بین تمامی arm ها میباشد.
با این تعاریف ، میتوانیم regret کلی بعد از n قدم را به شکل زیر بیان کنیم :

$$R_n = n\mu^* - \mathbf{E}[S_n] = \sum_t \mu^* - \sum_t \mathbf{E}[X_t] = \sum_t \mathbf{E}[\mu^* - X_t]$$
$$= \sum_t \sum_{a\in\mathcal{A}} \mathbf{E}[(\mu^* - X_t)\mathbb{I}\{A_t = a\}]$$

اگر در قدم t ام ، عمل $A_t = a$ را انجام دهیم ، جایزه X_t برابر با $\mu_a(v)$ میشود. پس داریم که :

لم تجزیه

$$\sum_t \sum_{a\in\mathcal{A}} \mathbf{E}[(\mu^* - \mu_a(v))\mathbb{I}\{A_t = a\}] = \sum_{a\in\mathcal{A}} \underbrace{(\mu^* - \mu_a(v))}_{\Delta_a} \mathbf{E}\left[\overbrace{\sum_t \mathbb{I}\{A_t = a\}}^{T_a(n)}\right]$$
$$\Rightarrow R_n = \sum_a \Delta_a \mathbf{E}[T_a(n)]$$

بخش دوم : عواقب انتخاب نادرست δ

اگر به ازای انتخاب δ اشتباه بازه اطمینان ما که همان $\sqrt{\frac{2\log(\frac{1}{\delta})}{T_i(t-1)}}$ می باشد دیگر فایده ای نداشته باشد و ایندکس الگوریتم ما (که همان $UCB(T_i(t), \delta)$ هست) به کمتر از میانگین حقیقت آن arm برسد ، الگوریتم دیگر arm بهینه را نتخاب نمیکند و دچار regret خطی میشود.
برای جلوگیری از این مشکل میتوان مقدار δ را جوری انتخاب کرد که با گذر زمان کاهش یابد که بازه اطمینان مان نیز کاهش یابد و احتمال اینکه ایندکس ما به کمتر از میانگین حقیقی برسد را کاهش میدهیم. انتخاب هایی مانند $\delta = \frac{1}{n^2}, \frac{1}{n}, \ldots$ میتوانند گزینه های خوبی باشند.

بخش سوم : تعداد دفعات کشیده شدن دست i ام

فرض کنید فقط با اتفاق های خوب سر و کله میزنیم. (G_i) اگر داشته باشیم که $u_i > T_i(n)$ یعنی دست شماره i بیشتر از u_i بار کشیده شده است ، پس در این n بار بازی کردن ، یک t ای وجود دارد که به ازای آن داریم :

$T_i(t-1) = u_i \Rightarrow A_t = i$

تناقض

با استفاده از تعریف G_i داریم که :

$$UCB_i(t-1,\delta) = \hat{\mu}_i(t-1) + \sqrt{\frac{2\log(\frac{1}{\delta})}{T_i(t-1)}} = \hat{\mu}_{iu_i} + \sqrt{\frac{2\log(\frac{1}{\delta})}{u_i}} < \mu_1 < UCB_1(t-1,\delta)$$

در هر مرحله عملی که در مرحله بعدی انجام میشود طبق رابطه زیر به دست میاید :

$A_t = \text{argmax}_j \, \text{UCB}_j(t-1,\delta)$

و چون دیدیم که UCB دست i ام در قدم $t-1$ از UCB دست 1 کمتر هست قطعا انتخاب نمیشود و امکان ندارد که بیشتر از u_i بار بازی شود.
پس با تناقض نشان دادیم که در نهایت arm شماره i ماکسیموم u_i بار بازی میشود.

بخش چهارم : کران بالا برای امید ریاضی

با نوشتن عبارت $T_i(n)$ به شکل زیر داریم که :

$\mathbf{E}\left[T_i(n)\right] = \mathbf{E}\left[\mathbb{I}\{G_i\}T_i(n)\right] + \mathbf{E}\left[\mathbb{I}\{G_i^c\}T_i(n)\right]$

به صورت جمع امید ریاضی مواقعی که اتفاق خوب رخ داده و امید ریاضی مواقعی که اتفاق بد رخ داده

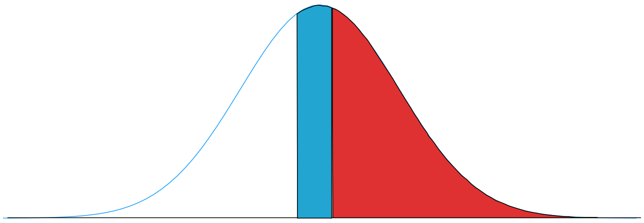
$\mathbf{E}\left[\mathbb{I}\{G_i\}T_i(n)\right] < u_i \quad$ (Maximum number of T_i in Good events)
$\mathbf{E}\left[\mathbb{I}\{G_i^c\}T_i(n)\right] < n\mathbb{P}(G_i^c) \quad$ (At worst all n times the Bad events happen)
$*\mathbf{E}\left[\mathbb{I}\{G_i^c\}\right] = \int_w \mathbb{I}\{G_i^c\}d\mathbb{P}(w) = \mathbb{P}(G_i^c)$
$\mathbf{E}\left[T_i(n)\right] \leq u_i + n\mathbb{P}(G_i^c)$

بخش پنجم :کران برای احتمال $\hat{\mu}_{iu_i}$

رابطه داده شده را به شکل زیر بازنویسی میکنیم :

$$\mathbf{P}\left(\hat{\mu}_{iu_i} + \sqrt{\frac{2\log(\frac{1}{\delta})}{u_i}} \geq \mu_1\right) = \mathbf{P}\left(\hat{\mu}_{iu_i} - \mu_i \geq (\mu_1 - \mu_i) - \sqrt{\frac{2\log(\frac{1}{\delta})}{u_i}}\right)$$

$$\underbrace{\mathbf{P}\left(\hat{\mu}_{iu_i} - \mu_i \geq \Delta_i - \sqrt{\frac{2\log(\frac{1}{\delta})}{u_i}}\right)}_{\text{the area in red}} \leq \underbrace{\mathbf{P}\left(\hat{\mu}_{iu_i} - \mu_i \geq c\Delta_i\right)}_{\text{the area in red + blue}}$$



شکل ۱: احتمال رخ دادن $(\hat{\mu}_{iu_i} - \mu_i \geq c\Delta_i)$ بیشتر هست.

حال با استفاده از روش کرامر–چرنف داریم که :

$$\mathbf{P}(\hat{\mu}_{iu_i} - \mu_i \geq c\Delta_i) \leq \exp(-\frac{(c\Delta_i)^2 n}{2\sigma^2})$$
$$\xrightarrow{\sigma=1, n=u_i} \mathbf{P}(\hat{\mu}_{iu_i} - \mu_i \geq c\Delta_i) \leq \exp(-\frac{c^2\Delta_i^2 u_i}{2})$$

بخش ششم : کران بالا برای احتمال رخداد بد

برای G_i^c که مکمل G_i هست داریم که :

$$G_i = \{\mu_1 \leq \min_{t\in[n]} UCB_1(t,\delta)\} \bigcap \{\hat{\mu}_{iu_i} + \sqrt{\frac{2\log(\frac{1}{\delta})}{n}} \leq \mu_1\}$$
$$G_i^c = \{\mu_1 > \min_{t\in[n]} UCB_1(t,\delta)\} \bigcup \{\hat{\mu}_{iu_i} + \sqrt{\frac{2\log(\frac{1}{\delta})}{n}} > \mu_1\}$$

همچنین میدانیم وقتی مجموعه ای داریم که عضو های آن از مینیموم مجموعه ای دیگر بیشتر هستند ، پس باید از تک تک اعضای آن مجموعه نیز بیشتر باشند. پس مجموعه اول رابطه بالا را میتوان به شکل زیر نوشت :

$$\{\mu_1 > \min_{t\in[n]} UCB_1(t,\delta)\} \subseteq \bigcup_t \{\mu_1 > UCB_1(t,\delta)\} \quad (**)$$

$$\Rightarrow \mathbf{P}(\{\mu_1 > \min_{t\in[n]} UCB_1(t,\delta)\} \bigcup \{\hat{\mu}_{iu_i} + \sqrt{\frac{2\log(\frac{1}{\delta})}{n}} > \mu_1\}) \leq$$
$$\mathbf{P}(\{\mu_1 > \min_{t\in[n]} UCB_1(t,\delta)\}) + \mathbf{P}(\{\hat{\mu}_{iu_i} + \sqrt{\frac{2\log(\frac{1}{\delta})}{n}} > \mu_1\}) \quad \text{(Union Bound)}$$

احتمال مجموعه اول را حساب میکنیم :

$$\mathbf{P}(\{\mu_1 > \min_{t\in[n]} UCB_1(t,\delta)\}) \leq \mathbf{P}(\bigcup_t \{\mu_1 > UCB_1(t,\delta)\}) \quad (**)$$
$$\leq \sum_t \mathbf{P}(\mu_1 > UCB_1(t,\delta)) = \sum_{t=0}^{n} \mathbf{P}(\mu_1 > \hat{\mu}_1 + \sqrt{\frac{2\log(\frac{1}{\delta})}{t}})$$

- :

$$\mathbf{P}(X - \hat{X} \geq \epsilon) \leq \exp(-\frac{n\epsilon^2}{2\sigma^2}) \Rightarrow \epsilon = \sqrt{\frac{2\log(\frac{1}{\delta})}{t}} \Rightarrow \exp(-\frac{n\epsilon^2}{2\sigma^2}) = \delta\,; \sigma = 1 \text{ (assumption)}$$

$$\sum_t \mathbf{P}(\mu_1 > \hat{\mu}_{1t} + \sqrt{\frac{2\log(\frac{1}{\delta})}{t}}) \leq n\delta$$

احتمال دوم ، یعنی $\mathbf{P}(\{\hat{\mu}_{iu_i} + \sqrt{\frac{2\log(\frac{1}{\delta})}{n}} > \mu_1\})$ نیز در بخش قبلی (بخش پنجم) محاسبه شده است. پس در آخر داریم :

احتمال وقوع رخداد بد

$$\begin{aligned}\mathbf{P}(G_i^c) &= \mathbf{P}(\{\mu_1 > \min_{t\in[n]} UCB_1(t,\delta)\}) + \mathbf{P}(\{\hat{\mu}_{iu_i} + \sqrt{\frac{2\log(\frac{1}{\delta})}{n}} > \mu_1\}) \\ &\leq n\delta \qquad\qquad + \exp(-\frac{c^2\Delta_i^2 u_i}{2})\end{aligned}$$

$$\boxed{\mathbf{P}(G_i^c) \leq n\delta + \exp(-\frac{c^2\Delta_i^2 u_i}{2})}$$

بخش هفتم : کران بالا برای امید ریاضی

از بخش شش که احتمال $\mathbf{P}(G_i^c)$ را محاسبه کردیم در رابطه ای که برای امیدریاضی $T_i(n)$ به دست آورده بودیم قرار میدهیم و داریم که :

$$\mathbf{E}\left[T_i(n)\right] \leq u_i + n\mathbb{P}(G_i^c)$$

$$\mathbf{E}\left[T_i(n)\right] \leq u_i + n(n\delta + \exp(-\frac{c^2\Delta_i^2 u_i}{2}))$$

مقدار u_i که انتخاب میکنیم باید در رابطه بخش پنجم $(\Delta_i - \sqrt{\frac{2\log(\frac{1}{\delta})}{u_i}} \geq c\Delta_i)$ نیز صدق کند. یک انتخاب معمول ، انتخاب کوچکترین مقدار u_i برحسب این نامعادله می باشد :

$$\Delta_i - \sqrt{\frac{2\log(\frac{1}{\delta})}{u_i}} = c\Delta_i$$

$$\Delta_i(1-c) = \sqrt{\frac{2\log(\frac{1}{\delta})}{u_i}}$$

$$(\Delta_i(1-c))^2 = \frac{2\log(\frac{1}{\delta})}{u_i} \Rightarrow u_i = \frac{2\log(\frac{1}{\delta})}{(\Delta_i(1-c))^2}$$

$$u_i = \left\lceil \frac{2\log(\frac{1}{\delta})}{(\Delta_i(1-c))^2} \right\rceil \quad \text{(Because } u_i \text{ is integer)}$$

با قرار دادن $\delta = \frac{1}{n^2}$ و قرار دادن مقدار u_i در معادله خواهیم داشت :

$$\mathbf{E}[T_i(n)] \leq u_i + 1 + n^{1-\frac{2c^2}{(1-c)^2}} = \left\lceil \frac{2\log(n^2)}{(1-c)^2\Delta_i^2} \right\rceil + 1 + n^{1-\frac{2c^2}{(1-c)^2}}$$

اگر c را برابر با $\frac{1}{2}$ در نظر بگیریم داریم که :

$$\left\lceil \frac{2\log(n^2)}{(1-c)^2\Delta_i^2} \right\rceil + 1 + n^{1-\frac{2c^2}{(1-c)^2}} = \left\lceil \frac{16\log(n)}{\Delta_i^2} \right\rceil + 1 + n^{-1}$$

کران بالایی برای $n^{-1} \leq 1$ داریم و برای $\left\lceil \frac{16\log(n)}{\Delta_i^2} \right\rceil \leq \frac{16\log(n)}{\Delta_i^2} + 1$. با قرار دادن این دو کران در رابطه بالا داریم :

$$\left\lceil \frac{16\log(n)}{\Delta_i^2} \right\rceil + 1 + n^{-1} \leq \frac{16\log(n)}{\Delta_i^2} + 1 + 1 + 1 = \frac{16\log(n)}{\Delta_i^2} + 3$$

$$\mathbb{E}[T_i(n)] \leq \frac{16\log(n)}{\Delta_i^2} + 3$$

بخش هشتم

با جایگذاری رابطه نهایی بخش هفتم در رابطه تجزیه بخش اول داریم :

$$R_n = \sum_a \Delta_a \mathbf{E}\left[T_a(n)\right] \leq \sum_a \Delta_a\left(\frac{16\log(n)}{\Delta_a^2} + 3\right) = \sum_{a:\Delta_a \neq 0} \frac{16\log(n)}{\Delta_a} + 3\sum_a \Delta_a$$

$$R_n \leq \sum_{a:\Delta_a \neq 0} \frac{16\log(n)}{\Delta_a} + 3\sum_a \Delta_a$$

بخش نهم

با انتخاب $\Delta = \sqrt{\frac{16klog(n)}{n}}$ خواهیم داشت :

$$\begin{aligned} R_n &= \sum_{i=1}^{k} \Delta_i \mathbb{E}[T_i(n)] = \sum_{i:\Delta_i < \Delta} \Delta_i \mathbb{E}[T_i(n)] + \sum_{i:\Delta_i \geq \Delta} \Delta_i \mathbb{E}[T_i(n)] \\ &\leq n\Delta + \sum_{i:\Delta_i \geq \Delta} \left(3\Delta_i + \frac{16\log(n)}{\Delta_i}\right) \leq n\Delta + \frac{16k\log(n)}{\Delta} + 3\sum_i \Delta_i \\ &\leq 8\sqrt{nk\log(n)} + 3\sum_{i=1}^{k} \Delta_i, \end{aligned}$$

۳ یادگیری برخط

الگوریتم RWM

بخش اول : بروز کردن وزن ها

$$\mathbf{P}(X=i)=\frac{w_i(t)}{S_t}$$
$$w_i(t+1)=w_i(t)(1-\epsilon\cdot\mathbb{I}(\hat{m}_t=1))$$
$$S_{t+1}=\sum_i w_i(t+1)=\sum_i w_i(t)(1-\epsilon\cdot\mathbb{I}(\hat{m}_t=1))$$

$$\mathbf{E}\left[S_{t+1}\right]=\mathbf{E}\left[\sum_i w_i(t+1)\right]=\mathbf{E}\left[\sum_i w_i(t)\cdot(1-\epsilon\cdot\mathbb{I}(\hat{m}_t=1))\right]$$
$$\mathbf{E}\left[S_{t+1}\right]=\mathbf{E}\left[S_t\right](1-\epsilon\cdot\mathbf{E}\left[\mathbb{I}(\hat{m}_t=1)\right])=\mathbf{E}\left[S_t\right](1-\epsilon\cdot\mathbf{P}(\hat{m}_t=1))$$

بخش دوم

$$\begin{aligned}\mathbf{E}\left[S_{T+1}\right]&=\mathbf{E}\left[S_T\right](1-\epsilon\cdot\mathbf{P}(\hat{m}_T=1))\\&=\mathbf{E}\left[S_{T-1}\right](1-\epsilon\cdot\mathbf{P}(\hat{m}_T=1))(1-\epsilon\cdot\mathbf{P}(\hat{m}_{T-1}=1))\\&.\\&.\\&.\\&=\mathbf{E}\left[S_0\right]\prod_{t=1}^{T}(1-\epsilon\cdot\mathbf{P}(\hat{m}_t=1))\\&=N\cdot\prod_{t=1}^{T}(1-\epsilon\cdot\mathbf{P}(\hat{m}_t=1))\quad\text{(at the first time-step (0) the weights are 1 },w_i(0)=1)\end{aligned}$$

همچنین از بسط تیلور تابع e^{-x} داریم که :

$$e^{-x}=1-x+\frac{x^2}{2}+\text{HOT}$$
$$e^{-x}>1-x$$
$$\Rightarrow\prod_{t=1}^{T}(1-\epsilon\cdot\mathbf{P}(\hat{m}_t=1))\leq\prod_{t=1}^{T}e^{-\epsilon\cdot\mathbf{P}(\hat{m}_t=1)}=e^{-\epsilon\sum_t\mathbf{P}(\hat{m}_t=1)}$$

$$\Rightarrow\mathbf{E}\left[S_{T+1}\right]=N\cdot\prod_{t=1}^{T}(1-\epsilon\cdot\mathbf{P}(\hat{m}_t=1))\leq N\cdot e^{-\epsilon\sum_t\mathbf{P}(\hat{m}_t=1)}$$

بخش سوم

$$S_{t+1} = S_t(1 - \epsilon \cdot \mathbf{P}(\hat{m}_t = 1))$$
$$P_t(i) = \frac{w_i(t)}{s_t} \Rightarrow w_i(t) = s_t \cdot P_t(i)$$
$$\mathbf{E}\left[\hat{m}_t\right] = \sum_i p_i(t)\hat{m}_t(i)$$

خط آخر روابط بالا امیدریاضی تعداد خطا های expert، i ام در زمان t را به ما میدهد.

$$S_{t+1} = S_t - \epsilon \cdot S_t\mathbf{P}(\hat{m}_t = 1)$$
$$S_t\mathbf{P}(\hat{m}_t = 1) = \sum_i w_i(t)\mathbf{P}(\hat{m}_t = 1) = \sum_i S_tP_i(t)\mathbf{P}(\hat{m}_t = 1) = S_t\sum_i P_i(t)\mathbf{P}(\hat{m}_t = 1) = S_t\mathbf{E}\left[\hat{m}_t\right]$$
$$S_{t+1} = S_t(1 - \epsilon \cdot \mathbf{E}\left[\hat{m}_t\right]) \leq S_t \cdot e^{-\epsilon\cdot\mathbf{E}[\hat{m}_t]}$$
$$w_i(T) = \prod_{t=1}^{T}(1 - \epsilon \cdot \mathbf{P}(\hat{m}_t = 1))\,;\; w_i(T) \leq S_T = \sum_{i=1}^{N} w_i(T)$$

اگر بیانمان را کمی عوض کنیم و بروز رسانی وزن ها بعد از T زمان را به شکل زیر نشان داهیم خواهیم داشت :

$$w_T(i) = (1-\epsilon)^{M_T(i)}$$
$$(1-\varepsilon)^{M_T(i)} \leq N\exp\Big(-\varepsilon\,\mathbb{E}\Big[\sum_{t=1}^{T}\tilde{m}_t\Big]\Big) = N\exp\big(-\varepsilon\,\mathbb{E}[M_T]\big)$$
$$M_T(i)\,\log(1-\varepsilon) \;\leq\; \log N \;-\; \varepsilon\,\mathbb{E}[M_T]$$
$$-\,M_T(i)\,(\varepsilon+\varepsilon^2) \;\leq\; \log N \;-\; \varepsilon\,\mathbb{E}[M_T]$$

در رابطه آخر از این استفاده کردیم که در نزدیکی صفر داریم $-x - x^2 \leq \log(1-x)$. در آخر به فرم خواسته شده میرسیم :

$$\mathbb{E}[M_T] \leq (1+\varepsilon)M_T(i) + \frac{\log N}{\varepsilon}.$$

بخش چهارم

از رابطه ای که در بخش سوم به دست آوردیم داریم که به ازای همه i ها این نامعادله برقرار هست :

$$\mathbb{E}[M_T] \leq (1+\varepsilon)M_T(i) + \frac{\log N}{\epsilon}$$

و چون به ازای همه i ها برقرار هست پس ، بر روی مینیموم خطا ها این کران را قرار دهیم :

$$\mathbf{E}[M_T] \le \min_i\{(1+\varepsilon)M_T(i)\} + \frac{\log N}{\epsilon} \quad (*)$$

$$(1+\epsilon)M_i = M_i + \epsilon M_i \le M_i + \epsilon T \quad (\mathbf{T}\ _i \le \mathbf{M})$$

$$\xrightarrow{(*)} \mathbf{E}\left[M_T\right] \le \min_i\{M_i + \epsilon T\} + \frac{\log N}{\epsilon}$$

$$\mathbf{E}\left[M_T\right] \le \min_i\{M_i\} + \epsilon T + \frac{\log N}{\epsilon} \Rightarrow \frac{\partial}{\partial \epsilon}(\epsilon T + \frac{\log N}{\epsilon}) = T - \frac{\log N}{\epsilon^2} = 0$$

$$T = \frac{\log N}{\epsilon^2} \Rightarrow \epsilon = \sqrt{\frac{\log N}{T}} \Rightarrow \epsilon T + \frac{\log N}{\epsilon} = \sqrt{\frac{\log N}{T}}T + \frac{\log N}{\sqrt{\frac{\log N}{T}}} = 2\sqrt{T \log N}$$

$$\Rightarrow \mathbf{E}\left[M_T\right] \le \min_i M_i + 2\sqrt{T\log N}$$

بخش امتیازی

بخش اول

$$\exp(-x) = \sum_{n=0}^{\infty} \frac{(-x)^n}{n!} \Rightarrow \exp(-x) \le 1 - x + \frac{x^2}{2}$$

$$S_{t+1} = \sum_i w_{t+1}(i) \Rightarrow S_{t+1} = \sum_i w_t(i)\cdot \exp(-\epsilon l_{ti}) \le \sum_i w_t(i)\left(1 - \epsilon l_{ti} + \frac{(\epsilon l_{ti})^2}{2}\right)$$

$$w_t(i) = p_t(i)\sum_i w_t(i) = p_t(i)S_t$$

$$S_{t+1} \le \sum_i p_t(i) S_t\left(1 - \epsilon l_{ti} + \frac{(\epsilon l_{ti})^2}{2}\right) = S_t(1 - \epsilon\sum_i p_t(i)l_t(i) + \frac{\epsilon^2}{2}\sum_i p_t(i)l_t(i)^2)$$

$$\boxed{S_{t+1} \le S_t(1 - \epsilon\sum_i p_t(i)l_t(i) + \frac{\epsilon^2}{2}\sum_i p_t(i)l_t(i)^2)}$$

بخش دوم

$$\begin{aligned} S_{t+1} &\le S_t\left(\sum_i p_t(i) - \varepsilon\sum_i p_t(i)\ell_t(i) + \frac{\varepsilon^2}{2}\sum_i p_t(i)\ell_t(i)^2\right) \\ &\le S_t\left(1 - \varepsilon p_t^\top \ell_t + \frac{\varepsilon^2}{2}p_t^\top \ell_t^2\right) \\ &\le S_t \exp\left(-\varepsilon p_t^\top \ell_t + \frac{\varepsilon^2}{2}p_t^\top \ell_t^2\right). \end{aligned}$$

$$S_T \le S_1 \exp\left(-\varepsilon\sum_{t=1}^{T} p_t^\top \ell_t + \frac{\varepsilon^2}{2}\sum_{t=1}^{T} p_t^\top \ell_t^2\right)$$

$$\leq N \exp\left(-\varepsilon \sum_{t=1}^{T} p_t^\top \ell_t + \frac{\varepsilon^2}{2} \sum_{t=1}^{T} p_t^\top \ell_t^2\right)$$

$$S_T \geq \exp\left(-\varepsilon \sum_{t=1}^{T} \ell_t(i)\right)$$

$$-\varepsilon \sum_{t=1}^{T} \ell_t(i) \leq \ln(N) - \varepsilon \sum_{t=1}^{T} p_t^\top \ell_t + \frac{\varepsilon^2}{2} \sum_{t=1}^{T} p_t^\top \ell_t^2$$

$$\sum_{t=1}^{T} p_t^\top \ell_t - \sum_{t=1}^{T} \ell_t(i) \leq \frac{\varepsilon}{2} \sum_{t=1}^{T} p_t^\top \ell_t^2 + \frac{\ln(N)}{\varepsilon}$$

$$\boxed{R_T \leq \frac{\varepsilon}{2} \sum_{t=1}^{T} p_t^\top \ell_t^2 + \frac{\ln(N)}{\varepsilon}}$$

تا حدی شبیه به باند برای الگوریتم RWM می باشد با این تفاوت که یک مجموع خطا (منظور $M_i = \sum_i l_t^2 \cdot p_t$) کم دارد. یعنی tight تر هست.

بخش سوم

با توجه به اینکه $l_t(i)$ بین یک تا منفی یک قرار دارد پس

$$\frac{\epsilon}{2} \sum_t p_t^T \ell_t^2 \leq \frac{\epsilon}{2} T \leq \epsilon T$$

$$R_T \leq \epsilon T + \frac{\ln(N)}{\epsilon}$$

$$\epsilon = \sqrt{\frac{2\ln(N)}{T}} \Rightarrow R_T \leq \sqrt{2\ln(N)T} + \sqrt{\frac{\ln(N)T}{2}} \leq 2\sqrt{\ln(N)T}$$