تمرین ۲ نظریه اطلاعات، آمار و یادگیری

سوال ۱) (۱)

$$\mathbb{E}_{\theta,\tilde\theta\sim\pi}\left[G(\theta,\tilde\theta)\right]-1=\mathbb{E}_{\theta,\tilde\theta\sim\pi}\left[\mathbb{E}_{Q}\left[\frac{P_\theta P_{\tilde\theta}}{Q^2}\right]\right]-1$$

$$=\int_\Theta \pi(\theta)\,d\theta\int_\Theta \pi(\tilde\theta)\,d\tilde\theta\int_\mathcal{X} Q(x)\,\frac{P_\theta(x)P_{\tilde\theta}(x)}{Q^2(x)}\,dx-1$$

$$=\int_\mathcal{X}\left(\frac{dx}{Q(x)}\int_\Theta P_\theta(x)\,\pi(\theta)\,d\theta\times\int_\Theta P_{\tilde\theta}(x)\,\pi(\tilde\theta)\,d\tilde\theta\right)-1=\int_\mathcal{X}\frac{dx}{Q(x)}\times P_\pi(x)\times P_\pi(x)-1$$

$$=\int_\mathcal{X}\frac{P_\pi^2(x)}{Q(x)}dx-1=\boxed{\chi^2(P_\pi\|Q)}$$

سوال ۴) (۲) برهان خلف، فرض کنید دنباله‌ی A_n وجود دارد که به ازای آن $Q_n(A_n)\xrightarrow{n\to\infty}0$ ولی $P_n(A_n)\xrightarrow{n\to\infty}a$ که a صفر نیست. داریم،

$$\chi^2(P_n\|Q_n)=\int\frac{P_n^2(x)}{Q_n(x)}dx-1\geqslant\int_{A_n}\frac{P_n^2(x)}{Q_n(x)}dx-1$$

$$\int_{A_n}\frac{P_n^2(x)}{Q_n(x)}dx\times\int_{A_n}Q_n(x)dx\underset{\text{کوشی شوارتز}}{\geqslant}\left(\int_{A_n}P_n(x)\,dx\right)^2\Rightarrow\int_{A_n}\frac{P_n^2(x)}{Q_n(x)}dx\geqslant\frac{P_n^2(A_n)}{Q_n(A_n)}$$

$$\Rightarrow\chi^2(P_n\|Q_n)\geqslant\frac{P_n^2(A_n)}{Q_n(A_n)}-1\Rightarrow\lim_{n\to\infty}\chi^2(P_n\|Q_n)\geqslant\lim_{n\to\infty}\frac{P_n^2(A_n)}{Q_n(A_n)}-1=+\infty$$

که خلاف فرض $\chi^2(P_n\|Q_n)=O(1)$ است $\Rightarrow$ فرض خلف باطل.

	$\sigma_i = \sigma_j$	$\sigma_i \neq \sigma_j$
$\hat{\sigma}_i = \hat{\sigma}_j$	K_1	K_2
$\hat{\sigma}_i \neq \hat{\sigma}_j$	K_3	K_4

سوال ۲) (۱) برای حل، روی σ_i و σ_j و $\hat{\sigma}_i$ و $\hat{\sigma}_j$ حالت بندی می کنیم، که K_1 تا K_4، تعداد این حالت‌ها را نشان می دهد:

حال $G(\sigma,\hat{\sigma})$ را محاسبه می کنیم:

یعنی گراف G را به توزیع‌ها ورودی بدیم

$$G(\sigma,\hat{\sigma}) = \int \frac{P_\sigma P_{\hat{\sigma}}}{P_0}(G)\, dG = \int \prod_{i<j,\ i,j=1}^{n} \frac{P_\sigma P_{\hat{\sigma}}}{P_0}(G_{ij})\, dG_{ij} \xrightarrow[\text{یال‌ها}]{\text{استقلال}}$$

$$= \prod_{i<j,\ i,j=1}^{n} \int \frac{P_\sigma P_{\hat{\sigma}}}{\frac{P+Q}{2}}(G_{ij})\, dG_{ij} = \left(\int \frac{2P^2}{P+Q}\right)^{K_1} \left(\int \frac{2Q^2}{P+Q}\right)^{K_4} \left(\int \frac{2PQ}{P+Q}\right)^{K_2+K_3}$$

یعنی G_{ij} را به توزیع‌ها ورودی بدیم

براساس حالت‌بندی σ و $\hat{\sigma}$ که در بالا گفتیم، P_σ و $P_{\hat{\sigma}}$ را با P و Q جایگزین کردیم. حال مقدار این انتگرال‌ها را محاسبه می کنیم:

$$2\int \frac{P^2}{P+Q} = 2\int \frac{(P^2+PQ)-PQ}{P+Q} = 2\int \left(P - \frac{PQ}{P+Q}\right) = 2\int P - \int \frac{2PQ}{P+Q} = 2 - \int \frac{2PQ}{P+Q}$$

به طور مشابه $\int \frac{2Q^2}{P+Q} = 2 - \int \frac{2PQ}{P+Q}$. پس $\int \frac{2PQ}{P+Q}$ را برحسب ρ به دست می آوریم:

$$\rho = \int \frac{(P-Q)^2}{2(P+Q)} = \int \frac{P^2+Q^2+2PQ-4PQ}{2(P+Q)} = \underbrace{\int \frac{1}{2}(P+Q)}_{\frac{1}{2}(1+1)=1} - \int \frac{2PQ}{P+Q}$$

$$\Rightarrow \int \frac{2PQ}{P+Q} = \boxed{1-\rho} \quad \Rightarrow \int \frac{2P^2}{P+Q} = \int \frac{2Q^2}{P+Q} = 2-(1-\rho) = \boxed{1+\rho}$$

$$\Rightarrow G(\sigma,\hat{\sigma}) = (1-\rho)^{K_2+K_3}(1+\rho)^{K_1+K_4} \longrightarrow K \triangleq K_2+K_3 \Rightarrow K_1+K_4 = \frac{n^2}{2} - K$$

چون ماتریس قرینه است، نصف حالت‌ها را نداریم.

حال K را محاسبه می کنیم (ایده‌ی این محاسبه از کسری خوشجو):

$$K = \left|\{ i<j \mid \sigma_i \oplus \sigma_j \oplus \hat{\sigma}_i \oplus \hat{\sigma}_j = 1 \}\right|$$

جابه‌جا کردن XOR‌ها

$$= (\sigma_i \oplus \hat{\sigma}_i) \oplus (\sigma_j \oplus \hat{\sigma}_j) = 1 \Rightarrow$$

یعنی در یک نقطه مساوی باشند و در یک نقطه نه

$$\Rightarrow K = d_H(\sigma,\hat{\sigma})\,(n - d_H(\sigma,\hat{\sigma}))$$

هم چنین می دانیم $\langle \sigma,\hat{\sigma}\rangle = n - 2d_H(\sigma,\hat{\sigma})$. پس $d_H = \frac{1}{2}(n - \langle\sigma,\hat{\sigma}\rangle)$ و داریم:

$$k = \frac{1}{2}\left(n - \langle \sigma, \hat{\sigma} \rangle\right) \times \frac{1}{2}\left(n + \langle \sigma, \hat{\sigma} \rangle\right) = \frac{1}{4}\left(n^2 - \langle \sigma, \hat{\sigma} \rangle^2\right)$$

$$\Longrightarrow k_1 + k_4 = \frac{n^2}{2} - k = \frac{1}{4}\left(n^2 + \langle \sigma, \hat{\sigma} \rangle^2\right)$$

بنابراین،

$$\Theta(\sigma, \hat{\sigma}) = (1-p)^k (1+p)^{\frac{n^2}{2} - k} \overset{*}{\leq} \exp\left(-p\left(\frac{n^2 - \langle \sigma, \hat{\sigma} \rangle^2}{4}\right) + p\left(\frac{n^2 + \langle \sigma, \hat{\sigma} \rangle^2}{4}\right)\right)$$

$$= \exp\left(\frac{p}{2} \langle \sigma, \hat{\sigma} \rangle^2\right)$$

که در $*$ از نامساوی $(1-x) \leq e^{-x}$ که در سوال بعد آمده استفاده کردیم.

(۲)

$$\rho = \sum \frac{(P-Q)^2}{2(P+Q)} = \frac{\left(\frac{a-b}{n}\right)^2}{2\left(\frac{a+b}{n}\right)} + \frac{\left(1 - \frac{a}{n} + 1 - \frac{b}{n}\right)^2}{2\left(2 - \frac{a+b}{n}\right)} = \frac{1}{2n}(a-b)^2 \left[\frac{1}{a+b} + \frac{1}{2n - (a+b)}\right]$$

$$= \frac{\tau}{n} + \frac{\frac{1}{2(2n-a-b)}}{n} = \frac{\tau + o(1)}{n}$$

همانطور که می‌دانیم $\frac{1}{2(2n-a-b)} \in o(1)$ است.

(۳) می‌خواهیم نشان دهیم P_0 و P_1 از هم قابل تمیز نیستند. برای این کار نشان می‌دهیم فاصله‌ی χ^2 بین این دو توزیع وقتی $n \to \infty$ می‌رود، صفر است:

می دانیم P_1 میانگین P_σ های مختلف است (مانند P_π در سوال ۱). پس داریم:

$$\chi^2(P_1 \| P_0) = \mathbb{E}_{\sigma,\hat\sigma}\left[G(\sigma,\hat\sigma)\right]-1 \quad \text{and} \quad G(\sigma,\hat\sigma) = \mathbb{E}_{P_0}\left[\frac{P_\sigma P_{\hat\sigma}}{P_0}\right] = \int \frac{P_\sigma P_{\hat\sigma}}{P_0}$$

و در بخش ۱ نشان دادیم:

$$G(\sigma,\hat\sigma) \le \exp\left(\frac{\rho}{2}\langle\sigma,\hat\sigma\rangle^2\right)$$

پس:

$$\chi^2 \le \mathbb{E}_{\sigma,\hat\sigma}\left[\exp\left(\frac{\rho}{2}\langle\sigma,\hat\sigma\rangle^2\right)\right]-1$$

حال می‌توان $\langle\sigma,\hat\sigma\rangle$ را به صورت جمع تعدادی متغیر برنولی در نظر گرفت:

$$\langle\sigma,\hat\sigma\rangle = \sum_i \sigma_i\hat\sigma_i \qquad p(\sigma_i\hat\sigma_i = 1) = p(\sigma_i\hat\sigma_i = -1) = \frac{1}{2} \qquad \left(2\times\frac{1}{2}\times\frac{1}{2} = \frac{1}{2}\right)$$

$$Var(\sigma_i\hat\sigma_i) = \underbrace{2^2}_{\text{فاصله‌ی دو مقدار برنولی}} \times p\times(1-p) = 4\times\frac{1}{4} = 1$$

پس در $n\to\infty$ می‌توان $\langle\sigma,\hat\sigma\rangle$ را یک متغیر گاوسی در نظر گرفت؛ یا دقیق‌تر:

$$\lim_{n\to\infty}\frac{1}{\sqrt{n}}\langle\sigma,\hat\sigma\rangle \sim N(0,1) \qquad \text{(قضیه حد مرکزی)}$$

پس داریم:

$$\lim_{n\to\infty}\chi^2(P_1\|P_0) \le \lim \mathbb{E}_{Z\sim N(0,1)}\left[\exp\left(\frac{\rho}{2}\times n\times Z^2\right)\right]-1$$

$$= \lim_{n\to\infty}\mathbb{E}_Z\left[\cancel{\exp}\,\frac{1}{2}(\tau+o(1))Z^2\right]-1 \underset{o(1)\to 0}{=} \mathbb{E}_{Z\sim N}\left[\exp\left(\frac{\tau Z^2}{2}\right)\right]-1$$

$$= \int_{-\infty}^{+\infty}\frac{1}{\sqrt{2\pi}}\underbrace{e^{\frac{-z^2}{2}}e^{\frac{\tau z^2}{2}}}_{e^{\frac{-z^2}{2}(\tau-1)}}dz - 1 \overset{\sigma=\frac{1}{\sqrt{\tau-1}}}{=\!=\!=} \int_{-\infty}^{+\infty}\frac{\sigma}{\sigma\sqrt{2\pi}}e^{\frac{-z^2}{2\sigma^2}}dz - 1 = \sigma - 1$$

$$= \frac{1}{\sqrt{\tau-1}} - 1 = O(1)$$

بنابراین طبق قسمت (۲) سوال (۱)، $P_1 \triangle P_0$ و بنابراین نمی‌توان آن‌ها را از هم تمیز داد.

سوال ۳) (۱) طبق تعریف ML پاسخ رو به رو
رابطه را می دهد؛
که می توان آن را ساده‌تر کرد:

$$\hat{M}_{ML} \in \arg\max_{M} \prod_{e\in M} P(w_e) \prod_{e\notin M} Q(w_e)$$

$$\xrightarrow{\text{log صعودی}} = \arg\max_{M} \log\Big(\prod_{e\in M} P(w_e) \prod_{e\in M} Q(w_e)\Big)$$

اضافه و کم

$$= \arg\max_{M} \sum_{e\in M} \log P(w_e) + \sum_{e\notin M} \log Q(w_e) + \sum_{e\in M} \log Q(w_e) - \sum_{e\in M} \log Q(w_e)$$

$$= \arg\max_{M} \sum_{e\in M} \log \frac{P}{Q}(w_e) + \underbrace{\sum_{e} \log Q(w_e)}_{\text{مستقل از انتخاب } M} = \boxed{\arg\max_{M} \sum_{e\in M} \log \frac{P}{Q}(w_e)}$$

(۲) ابتدا باند مورد نیاز برای $\binom{n}{k}k!$ را محاسبه می کنیم،

$$\binom{n}{k}k! = \frac{n!}{(n-k)!} = \prod_{j=0}^{k-1}(n-j)$$

$$= n^k \prod_{j=0}^{k-1}\left(1-\frac{j}{n}\right) = n^k \prod_{j=0}^{k-1} e^{-j/n} = n^k e^{-\frac{k^2-k}{2n}} = \exp\left(k\log n - \frac{k^2}{2n} + \frac{k}{2n}\right)$$

الان احتمال آن را محاسبه می کنیم که $\frac{1}{2}|M^*\Delta \hat{M}_{ML}| \geqslant \beta n$ باشد (ما به این نتیجه رسیدیم که برای اثبات این حکم باید $\frac{1}{2}$ تفاضل متقارن را درنظر گرفت، یعنی تعداد یال‌هایی که از M^* در $\hat{M}_{ML}$ نیامده)

برای محاسبه‌ی این احتمال، همه‌ی $\frac{1}{2}|M^*\Delta\hat{M}_{ML}| = k$ به ازای $n \geqslant k \geqslant \beta n$ را درنظر می گیریم. برای محاسبه‌ی این احتمال نیز یک حالت دلخواه $\hat{M}$ را درنظر می گیریم و ضرب در حداکثر تعداد حالات مجاز می کنیم.

$$P\left[\frac{1}{2}|M^*\Delta\hat{M}_{ML}| \geqslant \beta n\right] = \sum_{k=\beta n}^{n} P\left[\frac{1}{2}|M^*\Delta\hat{M}_{ML}| = k\right] \leqslant \sum_{k=\beta n}^{n}\binom{n}{k}k!\; \underline{P\left[\frac{1}{2}|M^*\Delta\hat{M}| = k\right]}$$

حال برای باند زدن روی این احتمال، از نامساوی انحراف بزرگ استفاده می کنیم (که حالت خاص آن که در سوال ۲ تمرین ۱ اثبات شد). می دانیم وقتی $\frac{1}{2}|M^*\Delta\hat{M}| = k$ باشد، یعنی جمع $\log\frac{P}{Q}$ در k تا یال در $\hat{M}$ که در تطابق نبوده اند (از توزیع Q بودند) از k تا یال در M^* (از توزیع P) بیشتر بوده است که انتخاب شده اند یا به عبارتی

$$\sum_{i=1}^{k} Y_i \geqslant \sum_{i=1}^{k} X_i \iff \sum_{i=1}^{k} Y_i - X_i \geqslant 0 \xrightarrow{\text{که}} Y_i = \log\frac{P(r_i)}{Q(r_i)},\ r_i \sim Q$$

$$X_i = \log\frac{P(r_i)}{Q(r_i)},\ r_i \sim P$$

$$P\left[\sum_{i=1}^{k} Y_i - X_i \geqslant 0\right] \leqslant \exp(2k\log B)$$

و طبق تمرین قبل می دانیم:

که β ، $BC(P,Q)$ است. پس با استفاده از این موضوع ،

$$P\left[\frac{1}{2}|M^* \Delta \hat{M}_{ML}| \geq \beta n\right] \leq \sum_{k=\beta n}^{n} \binom{n}{k} k!\; P\left[\sum_{i=1}^{k} Y_i - X_i \geq 0\right]$$

$$\rightarrow \leq \sum_{k=\beta n}^{n} \exp\left(k\log n - \frac{k^2}{2n} + \frac{k}{2n} + 2k\log B\right)$$

$B \leq \frac{1+\varepsilon}{\sqrt{n}} \rightarrow \log B \leq \log(1+\varepsilon) - \frac{1}{2}\log n$

$$\leq \sum_{k=\beta n}^{n} \exp\left(\cancel{k\log n} - \frac{k^2}{2n} + \frac{k}{2n} + 2k\log(1+\varepsilon) - \cancel{k\log n}\right)$$

$\beta \geq 8\log(1+\varepsilon) \Rightarrow$
$2\log(1+\varepsilon) \leq \frac{\beta}{4}$

$$\leq \sum_{k=\beta n}^{n} \exp\left(\frac{-k^2}{2n} + \frac{k}{2n} + \frac{k\beta}{4}\right)$$

$k \leq n \Rightarrow \frac{k}{2n} \leq \frac{1}{2}$

$k \geq \beta n \Rightarrow \frac{k^2}{2n} \geq \frac{k\beta}{2}$

$\Rightarrow \frac{-k^2}{2n} \leq \frac{-k\beta}{2}$

$$\leq \sum_{k=\beta n}^{n} \exp\left(\frac{-k\beta}{2} + \frac{1}{2} + \frac{k\beta}{4}\right)$$

$\beta n \leq n \Rightarrow \beta \leq 1 \Rightarrow e^{\frac{-\beta}{4}} \leq 1$

$$= e^{\frac{1}{2}} \sum_{k=\beta n}^{n} \exp\left(\frac{-k\beta}{4}\right) \leq e^{\frac{1}{2}} \sum_{k=\beta n}^{\infty} \exp\left(\frac{-k\beta}{4}\right)$$

$\longleftarrow$ تقریب

$$\boxed{= e^{\frac{1}{2}} \times \frac{\exp\left(\frac{-\beta^2 n}{4}\right)}{1 - \exp\left(\frac{-\beta}{4}\right)}}$$

حکم ثابت شد.

(۳)

Scanned by CamScanner

سوال ۵)

۱) می دانیم بهترین خطایی در آزمون فرض می توان به دست آورد $\frac{1}{2}(1-TV(P\|Q_\theta))$ است.
بنابراین تعیین کمترین خطا معادل به دست آوردن TV است که یک f_div به صورت روبرو است.

$$TV(P,Q_\theta) = \sup_f \mathbb{E}_P[f(x)] - \mathbb{E}_Q[f(x)]$$

یا به عبارت دیگر،

$$f(x) = \frac{1}{2}|x-1| \Rightarrow D_f(P,Q) = TV(P,Q)$$

۲) نشان می دهیم در حالت بهینه (یعنی زمانی که discriminator به صورت بهینه عمل کند) تابع هدف تبدیل به انحراف ینسن شنون بین P و Q_θ می شود

$$L \triangleq \mathbb{E}_{x,z}[2\log P(z|x)] = \int_x P(x)\times\frac{1}{2}\times 2\times\log p(0|x) + Q_\theta(x)\times\frac{1}{2}\times 2\times\log p(1|x)\,dx$$

اگر قرار دهیم $D(x)=p(0|x)$ (در واقع خروجی discriminator و احتمالی است که به واقعی بودن تصویر نسبت می دهد) خواهیم داشت،

$$L = \int_x p(x)\log D(x) + Q_\theta(x)\log(1-D(x))\,dx$$

D بهینه برای هر x را می توان با مشتق گرفتن از عبارت بالا و برابر صفر گذاشتن پیدا کرد،

$$\frac{p(x)}{D(x)} - \frac{Q_\theta(x)}{1-D(x)} = 0 \Rightarrow P(x) - p(x)D(x) - Q_\theta(x)D(x) = 0 \Rightarrow D^*(x) = \frac{P(x)}{P(x)+Q_\theta(x)}$$

با قرار دادن D بهینه در L داریم،

$$L = \int_x p(x)\log\frac{P(x)}{P(x)+Q_\theta(x)} + Q_\theta(x)\log\frac{Q_\theta(x)}{P(x)+Q_\theta(x)}\,dx$$

ابتدا انحراف JS را می نویسیم تا شباهتشان را ببینیم،

$$f(x) = x\log\frac{2x}{x+1} + \log\frac{2}{x+1} \Rightarrow JS(P\|Q_\theta) = D_f(P,Q)$$

$$\Rightarrow JS(P\|Q_\theta) = \int_x p(x)\log\frac{2P(x)}{P(x)+Q(x)} + Q_\theta(x)\log\frac{2Q_\theta(x)}{P(x)+Q(x)}\,dx$$

(البته برخی JS را $\frac{1}{2}$ این مقدار می گیرند)

پس داریم،

$$L = JS(P\|Q_\theta) + \int_x p(x)\log\frac{1}{2} + Q_\theta(x)\log\frac{1}{2}\,dx$$

$$\Rightarrow \boxed{L = JS(P\|Q_\theta) - 2\log 2}$$

۳) با توجه به روابطی که برای $D(x)$ بهینه و تابع هدف به ازای آن به دست آوردیم، شبکه‌ی مولد برای آن که بتواند شبکه‌ی Discriminator را فریب بدهد لازم است توزیع دیتا را کاملا کپی کند؛ یا به عبارتی $Q_\theta = P$ کند.

آموزش یک GAN را می‌توان به صورت یک بازی minmax زیر نوشت،

$$\min_\theta \max_D \int_x p(x) \log D(x) + Q_\theta(x) \log(1-D(x))\, dx$$

$$= \min_\theta \max_D \mathbb{E}_{x\sim P}\left[\log D(x)\right] + \mathbb{E}_{x\sim Q_\theta}\left[\log(1-D(x))\right]$$

که در آن $D(x) = P(z=0 \mid x)$ است.

سوال ۶ (۱)

$$R_\phi(f) = \mathbb{E}_{XY}\left[\phi(f(X)Y)\right] = \mathbb{E}_X \mathbb{E}_{Y|X}\left[\phi(f(X)Y)\right]$$

$$= \mathbb{E}_X\left[\phi(f(x))\,P[Y=1|X=x] + \phi(-f(x))\,P[Y=-1|X=x]\right]$$

$$= \mathbb{E}_X\left[\phi(f(x))\,\eta(x) + \phi(-f(x))\,(1-\eta(x))\right]$$

$$\Rightarrow \ell_\phi(x,y) = \phi(x)y + \phi(-x)(1-y)$$

فرم تابع ℓ_ϕ :

(۲) از $D_f(P_1\|P_{-1})$ شروع می‌کنیم و نشان می‌دهیم برابر با $R^*_{prior,\phi} - R^*_\phi$ است:

$$f(t) = \ell^*_\phi(\pi)(t\pi + 1-\pi) - \inf_\alpha \pi\phi(\alpha)t + (1-\pi)\phi(-\alpha)$$

$$\Rightarrow D_f(P_1\|P_{-1}) = \mathbb{E}_{P_{-1}}\left[f\left(\frac{P_1}{P_{-1}}\right)\right]$$

$$= \mathbb{E}_{P_{-1}}\Bigg[\underbrace{\inf_\alpha\left(\pi\phi(\alpha) + \phi(-\alpha)(1-\pi)\right)}_{\ell^*_\phi(\pi)}\left(\frac{P_1(x)}{P_{-1}(x)}\pi + 1-\pi\right)\Bigg]$$

$$- \mathbb{E}_{P_{-1}}\left[\inf_\alpha\left(\pi\phi(\alpha)\frac{P_1(x)}{P_{-1}(x)}\pi + (1-\pi)\phi(-\alpha)\right)\right]$$

$$= \sum_x \underbrace{\left[\inf_\alpha \pi\phi(\alpha) + \phi(-\alpha)(1-\pi)\right]}_{R^*_{prior,\phi}}\Big(\underbrace{P_1(x)\pi + P_{-1}(x)(1-\pi)}_{\sum_y P_{x|y}P(y) = P_x}\Big) = R^*_{prior,\phi}$$

↑ به x وابسته نیست و بیرون می‌آید

$$- \sum_x\left[\inf_\alpha \pi\phi(\alpha)P_1(x) + (1-\pi)\phi(-\alpha)P_{-1}(x)\right] \rightarrow$$

به ازای هر x، یک α مینیمم پیدا می‌کنیم مانند این است که یک تابع مینیمم می‌یابیم.

$$\hookrightarrow = \inf_{\mathcal{X}\to\mathbb{R}} \sum_x \pi\phi(f(x))P_1(x) + (1-\pi)\phi(-f(x))P_{-1}(x)$$

$$= \inf_{\mathcal{X}\to\mathbb{R}} \sum_x\sum_y P(x,y)\,\phi(f(x)y) = R^*_\phi$$

$$\Rightarrow \boxed{D_f(P_1\|P_{-1}) = R^*_{prior,\phi} - R^*_\phi}$$

(ع)

$$R^*_{prior,\phi} = \inf_{\alpha} \; P(Y=1)\phi(\alpha) + P(Y=-1)\,\phi(\alpha)$$

$$= \inf_{\alpha} \; P(Y=1)\log(1+e^{+\alpha}) + P(Y=-1)\log(1+e^{-\alpha})$$

به طور کلی عبارت inf $p\log(1+e^{+\alpha}) + \bar{p}\log(1+e^{-\alpha})$ را به دست می‌آوریم (که $p+\bar{p}=1$)
که در بخش بعد هم استفاده کنیم

$$\inf_{\alpha} \; p\log(1+e^{\alpha}) + \bar{p}\log(1+e^{-\alpha})$$

$$\Rightarrow \frac{d}{d\alpha}=0 \Rightarrow p \times \frac{e^{\alpha}}{1+e^{\alpha}} + \bar{p} \times \frac{-e^{-\alpha}}{1+e^{-\alpha}} = \frac{1}{1+e^{\alpha}}\left(pe^{\alpha} - \bar{p}\right) = 0$$

$$\frac{-e^{-\alpha}}{1+e^{-\alpha}} = \frac{-1}{e^{\alpha}+1}$$

$$\Rightarrow e^{\alpha} = \frac{\bar{p}}{p}$$

$$\Rightarrow \inf = p\log\left(1+\frac{\bar{p}}{p}\right) + \bar{p}\log\left(1+\frac{p}{\bar{p}}\right) = p\log\left(\frac{1}{p}\right) + \bar{p}\log\left(\frac{1}{\bar{p}}\right)$$

$$\Rightarrow R^*_{prior,\phi} = P(Y=1)\log\frac{1}{P(Y=1)} + P(Y=-1)\log\frac{1}{P(Y=-1)} = H(Y)$$

$$R^*_{\phi} = \sum_{x} P(x) \times \left[\inf_{\alpha} \; \phi(\alpha)\,P[Y=1|x] + \phi(-\alpha)\,P[Y=-1|x]\right]$$

$$= \sum_{x} P(x)\left(P[Y=1|x]\log\frac{1}{P[Y=1|x]} + P[Y=-1|x]\log\frac{1}{P[Y=-1|x]}\right)$$

$$= \sum_{x} P(x)\, H(Y|X=x) = \mathbb{E}_X\left[H(Y|X)\right] \Rightarrow$$

$$R^*_{prior,\phi} - R^*_{\phi} = \mathbb{E}_X\left[H(Y|X)\right] - H(Y) = I(X;Y)$$

حکم ثابت شد.

سوال ۷)

(۱) نشان می دهیم تخمینگری وجود دارد که بدترین خطای از اردر $p \log$ می گیرد. پس بهترین تخمینگر اردر خطا یش کمتر مساوی خواهد بود.

تخمین‌گری که ارائه می کنیم، بیشترین مقدار X خواهد بود،

$$T(x) = \max_i X_i$$

$$\Rightarrow \mathbb{E}_\theta\left[(\theta_{max} - X_{max})^2\right] = \mathbb{E}_\theta\left[\|X-\theta\|_\infty^2\right]$$

$$X - \theta \sim N(0,1) \quad \Rightarrow \quad = \mathbb{E}_{Z \sim N(0,1)}\left[\|Z\|_\infty^2\right] \approx c \log p$$

(این موضوع که نرم بی‌نهایت یک متغیر گاوسی، امید ریاضی به‌توان ۲ اش از اردر $\log n$ دارد معروف است ولی نتوانستم آن را اثبات کنم)

پس با $\inf$ گرفتن روی T های ممکن به خطای بهتری می رسیم و می توان گفت،

$$\inf_T \sup_\theta \mathbb{E}_\theta\left[(\theta_{max} - T)^2\right] \lesssim \log p \quad \text{or} \quad \leqslant c \log p$$

(۲) برای اثبات این بخش از روش دونقطه‌ای Le-Cam استفاده می کنیم

θ_0 را بردار صفر انتخاب می کنیم و بردار θ_1 را به صورت یکنواخت از $\{2\delta e_1, \dots, 2\delta e_p\}$ انتخاب می کنیم (چون فقط درایه ی max برایمان مهم است بردار e را بردارهای یکه ی مختصات در نظر گرفتیم). حال داریم،

$$|\theta_1 - \theta_0| = \rho(\theta_1, \theta_0) = 2\delta \qquad \phi(x) = x^2$$

$$TV \leqslant \sqrt{\tfrac{1}{2} KL}$$
$$KL \leqslant \chi^2$$

$$\Rightarrow \min\max \geqslant \phi(\delta) \times \frac{1}{2}\left(1 - TV(P_{\theta_1}, P_{\theta_0})\right) \geqslant \delta^2 \times \frac{1}{2}\left(1 - \sqrt{\tfrac{1}{2}\chi^2(P_{\theta_1}, P_{\theta_0})}\right)$$

از سوال ۱ می دانیم،

$$\chi^2(P_{\theta_1}, P_{\theta_0}) = \mathbb{E}_{\theta_1, \hat{\theta}_1 \sim \pi}\left[\int_x \frac{P_{\theta_1}(x)\, P_{\hat{\theta}_1}(x)}{P_{\theta_0}(x)}\, dx\right] - 1$$

توزیع یکنواخت روی $\{2\delta e_i\}$

$$= \mathbb{E}_{\theta, \hat{\theta}}\left[\int_x \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{\|x - (\hat{\theta}_1 + \theta_1)\|^2}{2}\right) \exp\left(\langle \theta_1, \hat{\theta}_1 \rangle \times \frac{2}{2}\right) dx\right] - 1$$

احتمال اینکه درایه ی ناصفرشان یکجا باشد

$$= \mathbb{E}_{\theta, \hat{\theta}}\left[\exp(\langle \theta_1, \hat{\theta}_1 \rangle)\right] - 1 = \frac{1}{p} \times \exp(4\delta^2) + \left(1 - \frac{1}{p}\right) \times \exp(0) - 1$$

$$= \frac{1}{p}\exp(4\delta^2) - \frac{1}{p}$$

بنابراین داریم،

$$\min\max \geq \frac{\delta^2}{2}\left(1-\sqrt{\frac{1}{2}\times\frac{\exp(\epsilon\delta^2)-1}{p}}\right)$$

$$\sqrt{\frac{1}{2}\,\frac{\exp(\epsilon\delta^2)-1}{p}} = c \Rightarrow \epsilon\delta^2 = c\log p$$

$$\Rightarrow \min\max \geq \frac{\delta^2}{2}\times c = c\log p$$

تخمین‌گری نیز در بخش قبل ارائه دادیم که چنین خطای دارد. پس این اردر برای خطای minmax tight است.

سوال ۸) (۱)

$$\hat{\theta} = \arg\min_{\theta} \sum_{i=1}^{d} |y_i - \theta_i|^2 + \lambda 1[\theta_i \neq 0] \qquad d \rightarrow \theta \text{ بعد}$$

$$\Rightarrow \hat{\theta}_i = \arg\min_{\theta_i} |y_i - \theta_i|^2 + \lambda 1[\theta_i \neq 0]$$

$$\begin{cases} \theta = 0 \longrightarrow & loss_i = y_i^2 \\ \theta \neq 0 \ (\theta = y_i) \longrightarrow & loss_i = \lambda \end{cases} \Rightarrow \begin{cases} y_i^2 > \lambda \longrightarrow \theta = y_i \\ y_i^2 \leq \lambda \longrightarrow \theta = 0 \end{cases}$$

بنابراین $\boxed{\tau = \sqrt{\lambda}}$

(۲)

$$\hat{\theta} = \arg\min_{\theta} \sum_{i=1}^{d} |y_i - \theta_i|^2 + \lambda|\theta_i| \Rightarrow \hat{\theta}_i = \arg\min_{\theta_i} |y_i - \theta_i|^2 + \lambda|\theta_i|$$

در سه حالت $\theta > 0$, $\theta < 0$, $\theta = 0$ مشتق می‌گیریم:

$$\theta > 0 \Rightarrow 2(\theta_i - y_i) + \lambda = 0 \Rightarrow \boxed{\theta_i = y_i - \frac{\lambda}{2}} , \ \theta_i > 0 \Rightarrow \boxed{y_i > \frac{\lambda}{2}}$$

$$\theta < 0 \Rightarrow 2(\theta_i - y_i) - \lambda \Rightarrow \boxed{\theta_i = y_i + \frac{\lambda}{2}} \xrightarrow{\theta_i < 0} \boxed{y_i < \frac{-\lambda}{2}}$$

$$\theta = 0 \Rightarrow \frac{-\lambda}{2} \leq y_i \leq \frac{\lambda}{2}$$

$$\Rightarrow \hat{\theta}_i^{ST} = \begin{cases} y_i - \frac{\lambda}{2} & y_i \geq \frac{\lambda}{2} \\ 0 & |y_i| \leq \frac{\lambda}{2} \\ y_i + \frac{\lambda}{2} & y_i \leq \frac{-\lambda}{2} \end{cases} \Rightarrow \boxed{\tau = \frac{\lambda}{2}}$$

(۳) تا زمانی که $\tau \geq \|y\|_\infty$ است، می‌توانیم θ را صفر نگه داریم و loss صفر داشته باشیم.
هرگاه قدر مطلق یکی از درایه‌های y از τ بزرگ‌تر شود، باید درایه‌ی متناظر در θ را برابر با مقداری قرار دهیم که $y_i - \theta_i < \tau$ شود. با توجه به اینکه تنها برای ما مهم است که θ_i صفر است یا نه و مقدار آن مهم نیست، می‌توان آن را برابر با y_i گذاشت.

$$\Rightarrow \hat{\theta}_i = \begin{cases} 0 & |y_i| < \tau \\ y_i & |y_i| \geq \tau \end{cases} \Rightarrow \hat{\theta} = \theta^{HT}$$

(۴) به طور مشابه با بخش قبل، هرگاه همه‌ی درایه‌های y مقدار کوچک‌تر از τ دارند، θ_i را صفر می‌گذاریم. اما اگر $y_i > \tau$ باشد، کوچک‌ترین مقداری که برای θ_i می‌توان متصور شد تا قید برقرار شود، $y_i - \tau$ است. به طور مشابه اگر $y_i < -\tau$، لازم است θ_i را برابر با $y_i + \tau$ قرار دهیم. پس $\hat{\theta}^{AST}$ یک جواب این مسئله است.

سوال ۹

(۱) نشان می‌دهیم به ازای یک تخمین‌گر $\sup_{\theta} \mathbb{E}_{\theta}\|\hat{\theta}-\theta\|_2^2$ از اردر 1 است.
پس به ازای همه‌ی تخمین‌گرها (یا inf روی تخمین‌گرها) نیز این نامساوی برقرار خواهد بود.
تخمین‌گرمان را به این صورت طراحی می‌کنیم که $\hat{\theta}_1$ را برابر با X_1 قرار می‌دهیم و بقیه درایه‌های $\hat{\theta}$ را ($\hat{\theta}_{/1}$) را صفر می‌گذاریم. در این صورت داریم،

$$\sup_{\theta} \mathbb{E}_{\theta}\left[\|X_1-\theta_1\|_2^2 + \|\theta_{/1}\|_2^2\right]$$

$$= \mathbb{E}_{\theta}\left[|X_1-\theta_1|^2\right] + \sup_{\theta} \mathbb{E}_{\theta}\left[\|\theta_{/1}\|_2^2\right]$$

$$X_1-\theta_1 \sim N(0,1)$$

$$\Rightarrow \mathbb{E}_{\theta}\left[|X_1-\theta_1|^2\right] = \mathbb{E}\left[(X_1-\theta_1)\right]^2 + Var(X_1-\theta_1) = 1$$

$$\|\theta_{/1}\|_2 \leq 2\left(1 - \underbrace{|\theta_1| p^{-1/4}}_{\text{مثبت}}\right) \leq 2$$

$$\Rightarrow \sup_{\theta} \mathbb{E}_{\theta}\left[\|\hat{\theta}-\theta\|_2^2\right] = 1+4 = 5 \lesssim 1$$

ثابت شد.