به نام خدا

نظریه‌ی اطّلاعات، آمار و یادگیری (۱-۲۵۱۱۰)

تمرین سری سوم

ترم بهار ۰۳-۱۴۰۲

دانشکده‌ی مهندسی برق

دانشگاه صنعتی شریف

استاد: دکتر محمّدحسین یاسائی میبدی

مهلت تحویل: جمعه ۴ خرداد ۱۴۰۳ ساعت ۲۳:۵۹

($*$) مسائلی که با ستاره مشخّص شده‌اند امتیازی هستند و حل‌کردن آن‌ها نمره‌ی امتیازی خواهد داشت!

۱ TV و ارتباط آن با برخی انحراف‌ها

در این سوال به بررسی برخی از ویژگی‌های انحراف TV می‌پردازیم. می‌توانید به دلخواه به سه قسمت پاسخ دهید و مابقی نمره امتیازی خواهد داشت.

۱. فرض کنید P و Q دو توزیع احتمال بر روی $X_{۱:n} = (X_۱, \ldots, X_n) \in \mathcal{X}^n$ باشند. همچنین فرض کنید $P_i(\cdot|x_{۱:i-۱})$ توزیع احتمال شرطی متغیر X_i به شرط $X_{۱:i-۱} = x_{۱:i-۱}$ باشد ($Q_i(\cdot|x_{۱:i-۱})$ را نیز به طور مشابه در نظر بگیرید). نشان دهید:

$$\|P-Q\|_{\mathrm{TV}} \leq \sum_{i=۱}^{n} \mathbb{E}_{X_{۱:i-۱}\sim P}\left[\|P_i(\cdot|X_{۱:i-۱}) - Q_i(\cdot|X_{۱:i-۱})\|_{\mathrm{TV}}\right],$$

که در آن امید ریاضی بر روی متغیر $X_{۱:i-۱}$ برحسب توزیع P گرفته می‌شود.

۲. **نامساوی Bretagnolle-Huber**: ثابت کنید برای هر دو توزیع P و Q داریم:

$$\|P-Q\|_{\mathrm{TV}} \leq \sqrt{۱-\exp\left(-D_{\mathrm{KL}}(P\|Q)\right)} \leq ۱ - \frac{۱}{۲}D_{\mathrm{KL}}(P\|Q).$$

۳. برای هر دنباله از توزیع‌های P_n و Q_n نشان دهید هنگامی که $n \to \infty$ داریم:

$$d_{\mathrm{TV}}\left(P_n^{\otimes n}, Q_n^{\otimes n}\right) \to ۰ \quad \Leftrightarrow \quad D_{H^۲}\left(P_n, Q_n\right) = o\left(\frac{۱}{n}\right),$$

$$d_{\mathrm{TV}}\left(P_n^{\otimes n}, Q_n^{\otimes n}\right) \to ۱ \quad \Leftrightarrow \quad D_{H^۲}\left(P_n, Q_n\right) = \omega\left(\frac{۱}{n}\right),$$

که در آن $D_{H^۲}(\cdot,\cdot)$ فاصله‌ی Hellinger است.

۴. فرم وردشی زیر را برای انحراف TV ثابت کنید:

$$d_{\mathrm{TV}}\left(P_۱, P_۲\right) = \frac{۱}{۲}\inf_q \sqrt{\int_{x\in\mathcal{X}} \frac{\left(p_۱(x)-p_۲(x)\right)^۲}{q(x)}dx}.$$

راهنمایی: از نامساوی کوشی-شوارتز استفاده کنید.

۵. فرض کنید $P_{Y|X}$ یک کانال با ورودی باینری باشد. قرار دهید $P_۰ = P_{Y|X=۰}, P_۱ = P_{Y|X=۱}$. ثابت کنید

$$\frac{۱}{۲} d_{\mathrm{TV}}^۲(P_۰, P_۱) \leq I(X;Y) \leq d_{\mathrm{TV}}(P_۰, P_۱).$$

راهنمایی: برای سمت چپ از نامساوی Pinsker استفاده کنید و برای سمت راست از نامساوی بین اطلاعات متقابل و $\chi^۲$ استفاده کنید.

۲ نشت اطّلاعات

فرض کنید $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ یک گراف ساده‌ی بدون جهت متناهی باشد.
متغیّرهای تصادفی $\text{Bernoulli}(\frac{۱}{۲}) \overset{\text{i.i.d.}}{\sim} \{X_v : v \in \mathcal{V}\}$ را روی رئوس این گراف و متغیّرهای تصادفی $\{Z_e : e \in \mathcal{E}\} \overset{\text{i.i.d.}}{\sim}$ $\text{Bernoulli}(\delta)$ را بر روی یال‌های آن تعریف می‌کنیم. حال برای هر یال $e = (u,v) \in \mathcal{E}$ تعریف کنید $Y_e = X_u \oplus X_v \oplus Z_e$.
مدل نشت روی گراف $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ یک اندازه‌ی احتمال روی تمام زیرگراف‌های $\mathcal{G}$ است که در آن احتمال حضور هرکدام از یال‌های $\mathcal{G}$ به طور مستقل برابر p باشد. چنین اندازه‌ی احتمالی را با $\mathbb{P}_{(\mathcal{G},p)}$ نمایش می‌دهیم. همینطور پیشامدِ وجود مسیر بین دو زیر مجموعه‌ی $\mathcal{S}, \mathcal{S}'$ از رئوس را با $(\mathcal{S} \rightsquigarrow \mathcal{S}')$. نمایش می‌دهیم.
در این سوال قصد داریم قضیه‌ی زیر را ثابت کنیم:

قضیه ۲-۱. برای هر زیرمجموعه از رئوس مانند $\mathcal{S} \subset \mathcal{V}$ و هر رأس مانند $v \in \mathcal{V}$ داریم:

$$I(X_v; X_{\mathcal{S}}, Y_{\mathcal{E}}) \leq \mathbb{P}_{(\mathcal{G},\eta)}[v \rightsquigarrow \mathcal{S}] \log(۲),$$

که در آن $\eta = (۱ - ۲\delta)^۲$. همچنین منظور از $X_{\mathcal{S}}$ مجموعه‌ی $\{X_u : u \in \mathcal{S}\}$ است.

برای اثبات، ابتدا باید با نامساوی قوی پردازش‌داده‌ها آشنا شوید. اگر $U \to X \to Y$ یک زنجیره‌ی مارکف باشد، از نامساوی پردازش‌داده‌ها می‌دانیم: $I(U;Y) \leq I(U;X)$ حال اگر $P_{Y|X}$ ثابت باشد، می‌توانیم ضریب $\eta_{P_{Y|X}}$ را به صورت زیر تعریف کنیم:

$$\eta_{P_{Y|X}} = \sup_{P_{U,X}} \frac{I(U;Y)}{I(U;X)}$$

در این صورت برای این زنجیره‌ی مارکف همواره خواهیم داشت:

$$I(U;Y) \leq \eta_{P_{Y|X}} I(U;X).$$

می‌توان دید اگر $P_{Y|X}$ یک کانال دوتایی متقارن[1] با پارامتر δ باشد، داریم: $\eta_{P_{Y|X}} = (۱ - ۲\delta)^۲$.

۱. ثابت کنید: $I(X_v; Y_{\mathcal{E}}) = ۰$.

۲. با استقرا روی $|\mathcal{E}|$ حکم مسئله را نتیجه بگیرید.
راهنمایی: برای گام استقرا از شرطی کردن اطلاعات متقابل استفاده کنید.

۳. $(*)$ فرض کنید $\mathcal{T}$ یک درخت منتظم با ریشه‌ی ρ باشد، که در آن درجه‌ی هر رأس $(d+۱)$ است. همین‌طور فرض کنید هر یال این درخت یک کانال $\text{BSC}(\delta)$ باشد. ابتدا یک بیت با توزیع $X_\rho \sim \text{Bernoulli}(\frac{۱}{۲})$ روی ریشه‌ی درخت تولید می‌شود. سپس این بیت از طریق کانال‌هایی که روی یال‌ها قرار دارند به سمت پایین انتشار می‌یابد. اگر $\mathcal{S}_k$ مجموعه‌ی رئوس در عمق k از این درخت باشد، ثابت کنید اگر $d(۱ - ۲\delta)^۲ < ۱$ داریم:

$$d_{\mathrm{TV}}\left(X_\rho, X_{\mathcal{S}_k}\right) \xrightarrow[k\to\infty]{} ۰.$$

۳ تبحّر در اثبات نامساوی‌ها!

۱. فرض کنید $f : [۰, \infty] \mapsto \mathbb{R}_{\geq ۰} \cup \infty$ تابعی محدب باشد که $f(۱) = f'(۱) = ۰$. همینطور فرض کنید μ, ν دو اندازه‌ی احتمال روی مجموعه‌ی $\mathcal{X}$ باشند. ثابت کنید برای $M > ۱$ داریم

$$\nu\left(\frac{d\nu}{d\mu} > M\right) \leq \frac{D_f(\nu||\mu)}{f'(M)}$$

[1] Binary Symmetric Channel (BSC)

راهنمایی: از تکنیک تغییر اندازه و همچنین تغییر متغیر در انتگرال استفاده کنید.

۲. (*) فرض کنید P_{XY} توزیع مشترک (X,Y) باشد و $\mathcal{E}$ واقعه‌ای مستقل از X باشد به طوری که $P[\mathcal{E}] = ۱ - \delta$. برای توزیع دلخواه Q_Y که برای آن به صورت $P_X - a.s.$ داریم: $P_{Y|X} \ll Q_Y$، ثابت کنید:

$$D_{\mathrm{KL}}(P_Y \| Q_Y) \le \log\left(۱ + D_{\chi^۲}\left(P_{Y|\mathcal{E}} \| Q_Y\right)\right) + \delta\left(\log(\frac{۱}{\delta}) + \mathbb{E}_X[D_{\mathrm{KL}}\left(P_{Y|X} \| Q_Y\right)]\right) + \sqrt{\delta \operatorname{Var}\left[\log \frac{dP_{Y|X}}{dQ_Y}\right]}.$$

راهنمایی: می‌توانید نامساوی $D_{\mathrm{KL}}(P\|Q) \le \log(۱ + D_{\chi^۲}(P\|Q))$ را دانسته فرض کنید. از تحدّب D_{KL} و نامساوی کوشی-شوارتز استفاده کنید.

۴ اطّلاعات متقابل و خطای تخمین

فرض کنید رابطه‌ی یک کانال با نویز گوسی به صورت $Y = \sqrt{A}X + Z$ باشد که در آن X ورودی کانال، Y خروجی کانال و $Z \sim \mathcal{N}(۰, ۱)$ است. فرض کنید می‌خواهیم با توجّه به خروجی این کانال ورودی آن را با تابعی مانند $f(Y)$ تخمین بزنیم. خطای این تخمین را به صورت $\mathbb{E}\left[(X - f(Y))^۲\right]$ در نظر می‌گیریم. همینطور خطای بهینه را برای این کانال به صورت $\mathcal{M}_{\mathsf{E}}(A) = \min_f \left\{\mathbb{E}\left[(X - f(Y))^۲\right]\right\}$ تعریف می‌کنیم. در این سوال قصد داریم رابطه‌ی زیر را بین $\mathcal{M}_{\mathsf{E}}(A)$ و $I(A) \triangleq I(X; \sqrt{A}X + Z)$ اثبات کنیم:

$$\frac{d}{dA} I(A) = \mathcal{M}_{\mathsf{E}}(A).$$

۱. ابتدا ثابت کنید تابع تخمین بهینه همان $\mathbb{E}[X|Y]$ است.

۲. ثابت کنید اگر ورودی کانال $Y = \sqrt{\delta}X + Z$ توزیع گوسی داشته باشد، برای $\delta \to ۰$ داریم:

$$I(X;Y) = \frac{\delta}{۲}\,\mathbb{E}\left[(X - \mathbb{E}[X])^۲\right] + o(\delta)$$

راهنمایی: از رابطه‌ی $I(X;Y) = \mathbb{E}_X\left[D_{\mathrm{KL}}(P_{Y|X} \| P_W)\right] - D_{\mathrm{KL}}(P_Y || P_W)$ برای توزیع Z مناسب استفاده کنید.

۳. برای اثبات قضیه از ایده‌ی کانال با نویز افزایشی استفاده می‌کنیم. برای این کار از ترکیب دو کانال گوسی استفاده می‌کنیم. به این صورت که ابتدا مقداری نویز به ورودی اضافه می‌کنیم تا نسبت سیگنال به نویز برابر $A + \delta$ شود، و سپس نویز بیشتری اضافه می‌کنیم تا نسبت سیگنال به نویز به A کاهش یابد (شکل ۱ را ببینید.) ثابت کنید که برای اثبات قضیه کافیست ثابت کنیم:

$$I(X;Y_۱) - I(X;Y_۲) = \frac{\delta}{۲}\mathcal{M}_{\mathsf{E}}(A) + o(\delta).$$

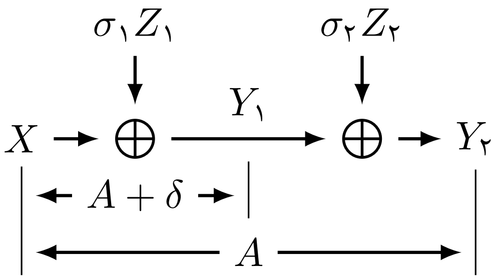
همین‌طور ثابت کنید: $I(X;Y_۱) - I(X;Y_۲) = I(X;Y_۱|Y_۲)$.

۴. رابطه‌ی زیر را اثبات کنید:

$$(A+\delta)Y_۱ = AY_۲ + \delta X + \sqrt{\delta}Z$$

که در آن Z یک نرمال استاندارد و مستقل از X است.

۵. با توجّه به قسمت‌های قبل حکم را نتیجه بگیرید.



شکل ۱: کانال با نویز افزایشی

۵ برخی خواصّ انحراف $\chi^۲$

فرض کنید یک خانواده‌ی پارامتری از توزیع‌ها به صورت $\{P_\theta : \theta \in \Theta\}$ داریم و π یک توزیع روی فضای Θ است. توزیع مخلوط P_π را به صورت زیر تعریف می‌کنیم:

$$P_\pi = \int P_\theta \pi(d\theta)$$

برای توزیع دلخواه Q تعریف می‌کنیم:

$$\mathcal{W}(\theta, \tilde{\theta}) = \mathbb{E}_Q[\frac{p_\theta p_{\tilde{\theta}}}{q^۲}]$$

۱. ثابت کنید:

$$D_{\chi^۲}(P_\pi \| Q) = \mathbb{E}_{\theta, \tilde{\theta} \overset{\text{i.i.d.}}{\sim} \pi}[\mathcal{W}(\theta, \tilde{\theta})] - ۱$$

۲. ثابت کنید اگر $D_{\chi^۲}(P_n \| Q_n) = O(۱)$ داریم:

$$\mathbb{P} \triangleleft \mathbb{Q},$$

که در آن $\mathbb{P} = \{P_n\}_{n=۱}^{\infty}$ و $\mathbb{Q} = \{Q_n\}_{n=۱}^{\infty}$.
راهنمایی: به سوال ۶ از تمرین اول مراجعه کنید.

۶ مسئله ی تشخیص در SBM

Planted Partition Model یک مدل برای تولید گراف تصادفی است. فرض کنید $\boldsymbol{\sigma} \in \{-۱, ۱\}^n$ باشد. در این صورت گراف تصادفی به صورت زیر تولید می‌شود:

$$A_{ij} \sim \begin{cases} \mathcal{P} & \sigma_i = \sigma_j \\ \mathcal{Q} & \sigma_i \neq \sigma_j \end{cases}$$

که در آن $\mathbf{A} = [A_{ij}]_{n \times n}$ ماتریس مجاورت وزن‌دار گراف است. این توزیع را با $\mathsf{G}(\boldsymbol{\sigma}, \mathcal{P}, \mathcal{Q})$ نمایش می‌دهیم. در حالتی که $\mathcal{P} \sim \mathsf{Bernoulli}(p)$ و $\mathcal{Q} \sim \mathsf{Bernoulli}(q)$ باشد، به این مدل Stochastic Block Model گفته می شود و آنرا با $\mathsf{SBM}(\boldsymbol{\sigma}, p, q)$ نمایش می‌دهیم. حال مسئله‌ی آزمون فرض دوتایی زیر را درنظر بگیرید:

$$\begin{aligned} \mathsf{H}_۰ &: \mathcal{G} \overset{\text{i.i.d.}}{\sim} R_۰ = \mathsf{G}(n, \frac{\mathcal{P}+\mathcal{Q}}{۲}) \\ \mathsf{H}_۱ &: \mathcal{G} \overset{\text{i.i.d.}}{\sim} R_۱ = \mathsf{G}(\mathcal{P}, \mathcal{Q}), \end{aligned}$$

که در آن منظور از توزیع $\mathsf{G}(\mathcal{P}, \mathcal{Q})$ این است که ابتدا بردار $\boldsymbol{\sigma}$ با توزیع Rademacher $\sigma_i \overset{\text{i.i.d.}}{\sim}$ (با احتمال برابر ±۱) تولید می‌شود و سپس $\mathcal{G}$ از توزیع $\mathsf{G}(\boldsymbol{\sigma}, \mathcal{P}, \mathcal{Q})$ نمونه‌برداری می‌شود. منظور از $\mathsf{G}(n, \frac{\mathcal{P}+\mathcal{Q}}{۲})$ نیز این است که وزن همه‌ی یال‌ها از توزیع $\frac{\mathcal{P}+\mathcal{Q}}{۲}$ می‌آید. حال می‌خواهیم در چند گام قضیه‌ی زیر را ثابت کنیم:

قضیه ۶-۱. در حالت $\mathsf{SBM}(\boldsymbol{\sigma}, p, q)$ اگر $p = \frac{a}{n}, q = \frac{b}{n}$ باشد و داشته باشیم: $\frac{(a-b)^۲}{۲(a+b)} < ۱$، در این‌صورت تشخیص بین دو فرض بالا غیرممکن می‌شود. یعنی خطای تشخیص نمی‌تواند به صفر همگرا شود، وقتی $n \to \infty$.

۱. اگر $P_{\boldsymbol{\sigma}} = \mathsf{G}(\boldsymbol{\sigma}, \mathcal{P}, \mathcal{Q})$ باشد، ثابت کنید:

$$\mathcal{W}(\boldsymbol{\sigma}, \hat{\boldsymbol{\sigma}}) = \mathbb{E}_{R_{۰}}\left[\frac{p_{\boldsymbol{\sigma}} p_{\hat{\boldsymbol{\sigma}}}}{r_{۰}^{۲}}\right] \leq \exp(\frac{\rho}{۲}\langle \boldsymbol{\sigma}, \hat{\boldsymbol{\sigma}}\rangle^{۲}),$$

که در آن:

$$\rho = \int_x \frac{(p(x) - q(x))^{۲}}{۲(p(x)+q(x))} dx.$$

۲. ثابت کنید در حالت $\mathsf{SBM}(\boldsymbol{\sigma}, p, q)$ که $p = \frac{a}{n}, q = \frac{b}{n}$، با تعریف $\tau = \frac{(a-b)^{۲}}{۲(a+b)}$ داریم:

$$\rho = \frac{\tau + o(۱)}{n}.$$

۳. با استفاده از قضیه‌ی حدّ مرکزی حکم را ثابت کنید (فرض کنید در این‌جا همگرایی در توزیع همگرایی تابع مولّد گشتاور را نتیجه می‌دهد، نیازی به اثبات این مورد نیست).

۷ تعیین ناحیه‌ی مشترک به وسیله‌ی توزیع‌های دوتایی

برای دو f–انحراف به شکل $D_f(P\|Q)$ و $D_g(P\|Q)$، ناحیه‌ی مشترک[۲] به صورت زیر تعریف می‌شود:

$$\mathcal{R} \triangleq \left\{ \Big(D_f(P\|Q), D_g(P\|Q) \Big) : P, Q \text{ are probability measures on some measurable space} \right\}.$$

همچنین ناحیه‌ی مشترک بر روی توزیع‌های k–تایی به صورت زیر تعریف می‌شود:

$$\mathcal{R}_k \triangleq \left\{ \Big(D_f(P\|Q), D_g(P\|Q) \Big) : P, Q \text{ are probability measures on } [k] \right\}.$$

در این پرسش قصد داریم نشان دهیم

$$\mathcal{R} = \mathsf{co}\left(\mathcal{R}_{۲}\right),$$

که در آن منظور از $\mathsf{co}(\mathcal{R}_{۲})$ پوش محدّب[۳] مجموعه‌ی $\mathcal{R}_{۲}$ است. به بیان دیگر برای به دست آوردن ناحیه‌ی مشترک دو f–انحراف به جای بررسی تمامی توزیع‌ها، کافی است توزیع‌های دوتایی را در نظر گرفت.

۱. نشان دهید مجموعه‌ی $\mathcal{R}$ محدّب است.

۲. نشان دهید مجموعه‌های $\mathcal{R}$ و $\mathcal{R}_k$ را می‌توان به صورت زیر نمایش داد

$$\mathcal{R} = \left\{ \begin{pmatrix} \mathbb{E}[f(X)] + \tilde{f}(۰)(۱ - \mathbb{E}[X]) \\ \mathbb{E}[g(X)] + \tilde{g}(۰)(۱ - \mathbb{E}[X]) \end{pmatrix}^{\top} : X \geq ۰,\ \mathbb{E}[X] \leq ۱ \right\}$$

$$\mathcal{R}_k = \left\{ \begin{pmatrix} \mathbb{E}[f(X)] + \tilde{f}(۰)(۱ - \mathbb{E}[X]) \\ \mathbb{E}[g(X)] + \tilde{g}(۰)(۱ - \mathbb{E}[X]) \end{pmatrix}^{\top} : \begin{array}{c} X \geq ۰,\ \mathbb{E}[X] \leq ۱,\ X \text{ takes at most } k-۱ \text{ values,} \\ \text{or,} \\ X \geq ۰,\ \mathbb{E}[X] = ۱,\ X \text{ takes at most } k \text{ values} \end{array} \right\},$$

که در آن‌ها:

$$\tilde{f}(۰) \triangleq \lim_{x \to ۰} x f(\frac{۱}{x}),$$
$$\tilde{g}(۰) \triangleq \lim_{x \to ۰} x g(\frac{۱}{x}).$$

۳. $(*)$ نشان دهید:

$$\mathcal{R} = \mathcal{R}_{۴}.$$

راهنمایی: از قضیه‌ی Fenchel-Eggleston-Caratheodory که در ادامه بیان می‌شود استفاده کنید.

[2] Joint range
[3] Convex hull

قضیه ۷-۱. فرض کنید $\mathcal{S} \subseteq \mathbb{R}^d$ و $\mathbf{x} \in \mathsf{co}(\mathcal{S})$. آن‌گاه می‌توان مجموعه‌ای مانند $\mathcal{S}'$ شامل $d+1$ نقطه به صورت $\mathcal{S}' = \{\mathbf{x}_1, \ldots, \mathbf{x}_{d+1}\} \subseteq \mathcal{S}$ یافت، به طوری که $\mathbf{x} \in \mathsf{co}(\mathcal{S}')$. همچنین اگر $\mathcal{S}$ هم‌بند باشد آن‌گاه d نقطه کافی است.

۴. $(*)$ برای $k \geq 2$ نشان دهید:

$$\mathcal{R}_{k+1} = \mathsf{co}(\mathcal{R}_k).$$

۵. حکم سوال را اثبات کنید.