به نام خدا

نظریه‌ی اطّلاعات، آمار و یادگیری (۱-۲۵۱۱۰)

تمرین سری سوم

ترم بهار ۰۴-۱۴۰۳

دانشکده‌ی مهندسی برق

دانشگاه صنعتی شریف

استاد: دکتر محمّدحسین یاسائی میبدی

مهلت تحویل: جمعه ۲ خرداد ۱۴۰۴ 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 $X_{1:n} = (X_1, \ldots, X_n) \in \mathcal{X}^n$ باشند. همچنین فرض کنید $P_i(\cdot|x_{1:i-1})$ توزیع احتمال شرطی متغیر X_i به شرط $X_{1:i-1} = x_{1:i-1}$ باشد ($Q_i(\cdot|x_{1:i-1})$ را نیز به طور مشابه در نظر بگیرید). نشان دهید:

$$\|P - Q\|_{\mathrm{TV}} \leq \sum_{i=1}^{n} \mathbb{E}_{X_{1:i-1}\sim P}\left[\|P_i(\cdot|X_{1:i-1}) - Q_i(\cdot|X_{1:i-1})\|_{\mathrm{TV}}\right],$$

که در آن امید ریاضی بر روی متغیر $X_{1:i-1}$ برحسب توزیع P گرفته می‌شود.

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

$$\|P - Q\|_{\mathrm{TV}} \leq \sqrt{1 - \exp\left(-D_{\mathrm{KL}}(P\|Q)\right)} \leq 1 - \frac{1}{2}\exp\left(-D_{\mathrm{KL}}(P\|Q)\right).$$

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

$$d_{\mathrm{TV}}\left(P_n^{\otimes n}, Q_n^{\otimes n}\right) \to 0 \quad \Leftrightarrow \quad D_{H^2}\left(P_n, Q_n\right) = o\left(\frac{1}{n}\right),$$

$$d_{\mathrm{TV}}\left(P_n^{\otimes n}, Q_n^{\otimes n}\right) \to 1 \quad \Leftrightarrow \quad D_{H^2}\left(P_n, Q_n\right) = \omega\left(\frac{1}{n}\right),$$

که در آن $D_{H^2}(\cdot,\cdot)$ فاصله‌ی Hellinger است.

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

$$d_{\mathrm{TV}}\left(P_1, P_2\right) = \frac{1}{2}\inf_q \sqrt{\int_{x\in\mathcal{X}} \frac{\left(p_1(x) - p_2(x)\right)^2}{q(x)} dx}.$$

راهنمایی: از نامساوی کوشی-شوارتز استفاده کنید.

۵. فرض کنید $P_{Y|X}$ یک کانال با ورودی باینری $X \sim Ber(۱/۲)$ باشد. قرار دهید $P_۰ = P_{Y|X=۰}, P_۱ = P_{Y|X=۱}$. ثابت کنید
$$\frac{۱}{۲} d^۲_{\mathrm{TV}}(P_۰, P_۱) \le I(X;Y) \le d_{\mathrm{TV}}(P_۰, P_۱).$$
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 $\{X_v \ : \ v \in \mathcal{V}\} \overset{\text{i.i.d.}}{\sim}$ Bernoulli($\frac{۱}{۲}$) را روی رئوس این گراف و متغیّرهای تصادفی $\{Z_e \ : \ e \in \mathcal{E}\} \overset{\text{i.i.d.}}{\sim}$ Bernoulli(δ) 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:
$$I(X_v; X_{\mathcal{S}}, Y_{\mathcal{E}}) \le \ \mathbb{P}_{(\mathcal{G},\eta)}[v \rightsquigarrow \mathcal{S}] \log(۲),$$
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: $I(U;Y) \le I(U;X)$ 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:
$$I(U;Y) \le \eta_{P_{Y|X}} I(U;X).$$
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 BSC(δ) باشد. ابتدا یک بیت با توزیع $X_\rho \sim$ 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:
$$d_{\mathrm{TV}}\left(\mathbb{P}_{X_{S_k}|X_\rho=+۱}, \mathbb{P}_{X_{S_k}|X_\rho=۰}\right) \xrightarrow[k\to\infty]{} ۰.$$

۳ تبحّر در اثبات نامساوی‌ها!

۱. فرض کنید $f : [۰, \infty] \mapsto \mathbb{R}_{\ge ۰} \cup \infty$ 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
$$\nu\left(\frac{d\nu}{d\mu} > M\right) \le \frac{M}{f(M)} D_f(\nu \| \mu)$$

[1] Binary Symmetric Channel (BSC)

راهنمایی: از تکنیک تغییر اندازه و همچنین خواص تابع محدب 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:

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

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

۴ نامساوی قوی پردازش داده‌ها

هدف ما در این سوال، آشنایی بیشتر با نامساوی قوی پردازش داده‌ها[2] است. فرض کنید تابع $f : (۰, \infty) \to \mathbb{R}$ ، تابعی محدّب باشد که در $x = ۱$ محدّب اکید است و همچنین داریم: $f(۱) = ۰$. برای متغیّرهای تصادفی $X \in \mathcal{X}$ و $Y \in \mathcal{Y}$ که در آن $|\mathcal{X}|, |\mathcal{Y}| < \infty$ و کانال $P_{Y|X}$ تعریف می‌کنیم:

$$\eta_f(P_{Y|X}, Q) \triangleq \sup_{P: ۰ < D_f(P\|Q) < \infty} \frac{D_f(P_{Y|X} \circ P \| P_{Y|X} \circ Q)}{D_f(P\|Q)},$$
$$\eta_f(P_{Y|X}) \triangleq \sup_Q \eta_f(P_{Y|X}, Q).$$

که در آن منظور از $P_{Y|X} \circ P$ توزیع القا شده بر روی متغیّر تصادفی Y از روی X با توزیع P و تحت کانال $P_{Y|X}$ است.

۱. با توجه به تعریف بالا، برای انحراف TV ثابت کنید:

$$\eta_{\mathrm{TV}}(P_{Y|X}) = \sup_{x,x'} \left\{d_{\mathrm{TV}}(P_{Y|X=x}, P_{Y|X=x'})\right\}.$$

۲. $(*)$ برای انحراف KL نشان دهید:

$$\eta_{\mathrm{KL}}(P_{Y|X}, P_X) = \sup_{U: U \to X \to Y} \frac{I(U;Y)}{I(U;X)},$$

که در آن، سوپریموم بر روی تمام زنجیره‌مارکف‌های به شکل $U \to X \to Y$ گرفته می‌شود که در آن توزیع توأم X, Y ثابت است.

۳. $(*)$ نشان دهید برای هر تابع f با ویژگی‌هایی که بیان شد، همواره داریم:

$$\eta_f(P_{Y|X}) \leq \eta_{\mathrm{TV}}(P_{Y|X}).$$

۵ 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:

$$\frac{d}{dA} I(A) = \frac{۱}{۲}\mathcal{M}_{\mathsf{E}}(A).$$

[2] Strong Data Processing Inequality

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

۲. ثابت کنید برای کانال گوسی $Y = \sqrt{\delta}X + Z$ با توزیع ورودی دلخواه، برای $\delta \to 0$ داریم:

$$I(X;Y) = \frac{\delta}{2}\,\mathbb{E}\left[(X - \mathbb{E}[X])^2\right] + o(\delta)$$

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:

$$I(X;Y_1) - I(X;Y_2) = \frac{\delta}{2}\mathcal{M}_{\mathsf{E}}(A) + o(\delta).$$

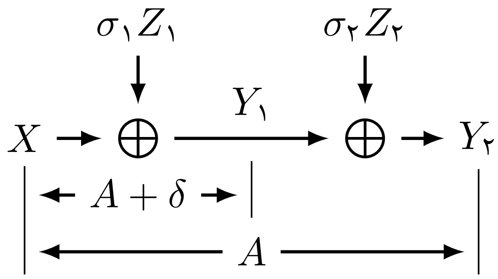
همین‌طور ثابت کنید: $I(X;Y_1) - I(X;Y_2) = I(X;Y_1|Y_2)$.

۴. رابطه‌ی زیر را اثبات کنید:

$$(A+\delta)Y_1 = AY_2 + \delta X + \sqrt{\delta}Z$$

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

۶ مسئله ی تشخیص در SBM

Planted Partition Model یک مدل برای تولید گراف تصادفی است. فرض کنید $\boldsymbol{\sigma} \in \{-1, 1\}^n$ 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:

$$\begin{aligned} \mathsf{H}_0 &: \ \mathcal{G} \overset{\text{i.i.d.}}{\sim} R_0 = \mathsf{G}(n, \frac{\mathcal{P}+\mathcal{Q}}{2}) \\ \mathsf{H}_1 &: \ \mathcal{G} \overset{\text{i.i.d.}}{\sim} R_1 = \mathsf{G}(\mathcal{P}, \mathcal{Q}), \end{aligned}$$

که در آن منظور از توزیع $\mathsf{G}(\mathcal{P}, \mathcal{Q})$ این است که ابتدا بردار $\boldsymbol{\sigma}$ با توزیع $\sigma_i \overset{\text{i.i.d.}}{\sim} \mathsf{Rademacher}$ (با احتمال برابر ± 1) تولید می‌شود و سپس $\mathcal{G}$ از توزیع $\mathsf{G}(\boldsymbol{\sigma}, \mathcal{P}, \mathcal{Q})$ نمونه‌برداری می‌شود. منظور از $\mathsf{G}(n, \frac{\mathcal{P}+\mathcal{Q}}{2})$ نیز این است که وزن همه‌ی یال‌ها از توزیع $\frac{\mathcal{P}+\mathcal{Q}}{2}$ 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:

$$\mathcal{W}(\boldsymbol{\sigma}, \hat{\boldsymbol{\sigma}}) = \mathbb{E}_{R_۰}\left[\frac{p_{\boldsymbol{\sigma}} p_{\hat{\boldsymbol{\sigma}}}}{r_۰^۲}\right] \leq \exp\left(\frac{\rho}{۲}\langle \boldsymbol{\sigma}, \hat{\boldsymbol{\sigma}}\rangle^۲\right),$$

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 تزویج بهینه

همانطور که در درس دیدیم، انحراف TV دارای تعریف معادل زیر است:

$$d_{\mathrm{TV}}(P\|Q) = \sup_E \{P(E) - Q(E)\},$$

که در آن E عضو مجموعه‌ی پیشامدهای فضای احتمال است.

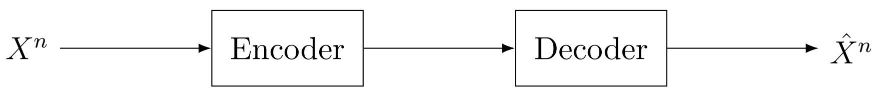
۱. با استفاده از این تعریف، قضیه‌ی Strassen را ثابت کنید:

$$d_{\mathrm{TV}}(P\|Q) = \inf_{P_{X\hat{X}} \in \Pi(P,Q)} P_{X\hat{X}}(X \neq \hat{X}),$$

که در آن

$$\Pi(P, Q) = \{P_{X\hat{X}} : \text{probability measure on } \mathcal{X}^۲, P_X = P, P_{\hat{X}} = Q\}.$$

۲. فرض کنید X^n یک منبع i.i.d. باشد و سناریوی کدینگ منبع با اعوجاج زیر را درنظر بگیرید:



هدف ما در این سوال، پیدا کردن کران بالایی برای متوسط فاصله‌ی همینگ ورودی کدگذار و خروجی کدگشا برحسب انحراف KL است. ثابت کنید توزیع $P_{X^n\hat{X}^n}$ مناسبی وجود دارد به گونه‌ای که

$$\frac{۱}{n}\mathbb{E}[d_n(X^n, \hat{X}^n)] = \frac{۱}{n}\sum_{i=۱}^{n} P_{X^n\hat{X}^n}(X_i \neq \hat{X}_i) \leq \sqrt{\frac{۱}{n} D_{\mathrm{KL}}(P_{\hat{X}^n}\|P_{X^n})}.$$

راهنمایی: از نامساوی Pinsker و قضیه‌ی Strassen استفاده کنید.

۸ انحراف Marton

۱. انحراف Marton به صورت زیر تعریف می‌شود:

$$D_{\mathrm{M}}(P\|Q) = \int \left(۱ - \frac{dP}{dQ}\right)_+^{۲} dQ.$$

ثابت کنید:

$$D_{\mathrm{M}}(P||Q) = \inf_{P_{XY}\in\Pi(P,Q)} \mathbb{E}\left[P_{XY}^{۲}(X \neq Y|Y)\right].$$

۲. انحراف Marton متقارن به صورت زیر تعریف می‌شود:

$$D_{\mathrm{SM}}(P\|Q) = D_{\mathrm{M}}(P\|Q) + D_{\mathrm{M}}(Q\|P).$$

ثابت کنید:

$$D_{\mathrm{SM}}(P\|Q) = \inf_{P_{XY}\in\Pi(P,Q)} \left\{ \mathbb{E}\left[P_{XY}^{۲}(X \neq Y|Y)\right] + \mathbb{E}\left[P_{XY}^{۲}(X \neq Y|X)\right]\right\}.$$