«بسمه تعالی»

پایان‌ترم درس احتمالات ابعاد بالا | ۶ تیر ۱۴۰۲

۱. برای مجموعه $T \subseteq \mathbb{R}^n$ و بردارهای تصادفی مستقل گوسی استاندارد $G_j, j = ۱, \cdots, n$، نشان دهید

$$\mathbb{E}\left[\sup_{v \in T} \sum_{j=۱}^{n} G_j |v_j|\right] \leq \mathbb{E}\left[\sup_{v \in T} \sum_{j=۱}^{n} G_j v_j\right].$$

تفسیری هندسی از نامساوی فوق ارائه دهید.

۲. (آ) بعد vc خانواده همه کره‌های سه بعدی را بیابید.

(ب) مجموعه نقاط

$$\{(x_۱, x_۲, \cdots, x_n) : \ x_i \in \{-۱, +۱\}\}$$

رئوس یک مکعب n بعدی را تشکیل میدهند. طول ضلع مربع برابر ۲ می باشد. حجم این مکعب n بعدی را محاسبه کنید. حال فرض کنید که به مرکز هرکدام از رئوس یک کره n بعدی به شعاع یک بزنیم. این کره ها به همدیگر مماس خواهند شد. بنابراین ۲^n کره خواهیم داشت. حال کره ای جدید به مرکز مبدا و با شعاع R در نظر بگیرید که به تمامی کره قبلی مماس باشد. شعاع این کره چیست؟ آیا این کره کاملا داخل مکعب قرار می‌گیرد؟! نسبت حجم این کره به حجم مکعب برای ابعاد بالا چیست؟ برای سادگی فرض کنید n زوج است. در این حالت حجم کره به شعاع r، برابر با $\frac{\pi^{n/۲}}{(n/۲)!} r^n$ است. ضمنا برای تخمین فاکتوریل می‌توانید از تقریب استرلینگ یا نامساوی ساده زیر استفاده کنید

$$\log n! < (n + ۱) \log(n + ۱) - n.$$

۳. در این مسئله به دنبال فشرده‌کردن داده‌ها هستیم به گونه‌ای که عملیات آماری یا یادگیری روی داده‌های فشرده شده تقریبا مشابه حالت داده‌های خام باشد.

(آ) ابتدا مسئله زیر را در نظر بگیرید. فرض کنید Z متغیرتصادفی با دامنه محدود باشد یعنی $|Z| \leq B$. همچنین فرض کنید $\gamma > B$ و U یک متغیر تصادفی یکنواخت روی $[-\gamma, \gamma]$ باشد. قرار دهید

$$\hat{Z} = \mathsf{sign}(U + Z)$$

در حقیقت $\hat{Z}$ فشرده‌شده Z توسط یک بیت است. نشان دهید

$$\mathbb{E}\left[\gamma \hat{Z}\right] = \mathbb{E}[Z]$$

(ب) مجموعه داده زیر در اختیار ما گذاشته شده است:

$$\mathcal{D} = \{(\mathbf{X}_۱, Y_۱), \cdots, (\mathbf{X}_n, Y_n)\}$$

که در آن $\mathbf{X}_k \in \mathbb{R}^d$ و $Y_k \in \mathbb{R}$ هستند. فرض کنید که رابطه بین بردار ویژگی $\mathbf{X}$ و متغیر Y (که مجموعه داده بالا هم از روی این توزیع تولید شده‌اند) توسط رابطه خطی نویزی شده زیر توصیف می‌شود:

$$Y = \Theta^\top \mathbf{X} + \epsilon$$

فرض کنید که متغیر Y و نویز ϵ هر دو زیرگوسی با پارامتر محدود $O(۱)$ باشند. به جای ذخیره کل داده علاقه مند به فشرده‌کردن آن هستیم. فعلا برای سادگی فقط Y_k را فشرده می‌کنیم. بدین منظور از فرآیند فشرده‌سازی قسمت قبل استفاده می‌کنیم. بدین منظور برای هر $k \in \{۱, ۲, \cdots, n\}$، $f(n)$ متغیر یکنواخت $(U_{k,۱}, \cdots, U_{k,f(n)})$ از $[-\gamma, \gamma]$ انتخاب و از روی آن‌ها

$$\hat{Y}_{k,j} = \mathsf{sign}(Y_k + U_{k,j}), \qquad j \in \{۱, \cdots, f(n)\}$$

می‌سازیم. دقت کنید که این فرآیند معادل با فشرده‌کردن Y_k با $f(n)$ بیت است. بدین ترتیب به مجموعه داده فشرده شده زیر می‌رسیم:

$$\mathcal{D}_c = \{(\mathbf{X}_k, \hat{Y}_{k,j}), k \in \{۱, \cdots, n\}, j \in \{۱, \cdots, f(n)\} \ \}$$

با توجه به قسمت قبل، سعی کنید که ابتدا تقریبی از داده اصلی $\mathcal{D}$ را از روی داده فشرده شده $\mathcal{D}_c$ بدست آورید. داده تقریبی شما باید به فرم زیر باشد

$$\mathcal{D}_{appr} = \{(\mathbf{X}_۱, \tilde{Y}_۱), \cdots, (\mathbf{X}_n, \tilde{Y}_n)\}$$

که در آن $\tilde{Y}_k$ باید از روی داده‌های فشرده شده محاسبه شود. سپس مقادیر γ و $f(n)$ را به گونه‌ای انتخاب کنید به طوریکه رابطه زیر برقرار باشد:

$$\tilde{Y}_k = \Theta^\top \mathbf{X}_k + \xi_k$$

به طوری که نویز ξ_k هم مشابه مدل اصلی زیرگوسی از مرتبه $O(۱)$ باشد. در نهایت چند بیت برای فشرده کردن بدون از دست دادن خواص آماری مسئله اصلی لازم است.

۴. سبد سهام را در کارهای سرمایه‌گذاری و اقتصاد و مدیریت حتما تاکنون شنیده‌اید. در این مساله سعی می‌کنیم مساله انتخاب سبد سهام را به صورت ریاضی مدل و در مورد آن تحقیق کنیم. فرض کنید یک سبد سهام با یک بردار $u \in \mathbb{R}^d$ که $\sum_{i=۱}^d u_i = ۱$ و $u_j \geq ۰$ برای هر $i \in [۱ : d]$ مدل می‌شود که در آن u_j میزان سرمایه گذاری فرد در دارایی jام را نشان می‌دهد. بردار (تصادفی) بازدهی سرمایه گذاری فرد را با $X \in \mathbb{R}^d$ نشان می‌دهیم و فرض می‌کنیم که X یک بردار زیرگوسی با پارامتر واریانس ۱ و ماتریس کواریانس نامشخص Σ است (اطلاعاتی در مورد آن نداریم!).

پاداش و ریسک را به این شکل مدل می‌کنیم: پاداش: $\mu(u) = \mathbb{E}[u^\top X]$ و ریسک: $R(u) = Var(X^\top u)$. مساله زیر را

درنظر بگیرید:

$$u^* = \arg\min_{u:\mu(u)\geq\lambda} R(u)$$

ابتدا بگویید در این مساله چه استراتژی برای سوددهی درنظر گرفته شده است؟ در ادامه فرض کنید این مساله جواب دارد، در عمل توزیع بردار X نامشخص است. فرض کنید توانسته‌ایم نمونه‌های مستقل و هم توزیع $X_۱, \dots, X_n$ از روی X را بدست بیاوریم و تقریبهای زیر را برای $\mu(u), R(u)$ زده‌ایم:

$$\hat{\mu}(u) = \bar{X}^\top u = \frac{۱}{n}\sum_{i=۱}^{n} X_i^\top u,$$
$$\hat{R}(u) = u^\top \widehat{\Sigma} u, \quad \widehat{\Sigma} = \frac{۱}{n-۱}\sum_{i=۱}^{n}(X_i - \hat{X})(X_i - \hat{X})^\top$$

ما از سبد سهام تقریبی زیر استفاده می‌کنیم:

$$\hat{u} = \arg\min_{u:\hat{\mu}(u)\geq\lambda} \hat{R}(u)$$

همچنین فرض کنید تعداد نمونه‌ها در $\log d \ll n \ll d$ صدق می‌کند.

(آ) نشان دهید برای هر u،

$$|\hat{\mu}(u) - \mu(u)| \lesssim \frac{۱}{\sqrt{n}}$$

و همچنین

$$\left|\hat{R}(u) - R(u)\right| \lesssim \frac{۱}{\sqrt{n}}$$

(ب) نشان دهید

$$\hat{R}(\hat{u}) - R(\hat{u}) \lesssim \sqrt{\frac{\log d}{n}}$$

(ج) تخمینگر زیر را درنظر بگیرید:

$$\tilde{u} = \arg\min_{u:\hat{\mu}(u)\geq\lambda-\epsilon} \hat{R}(u)$$

کوچکترین ϵ مثبت را پیدا کنید (ضریب ثابت را در نظر نگیرید!) که با احتمال ۰٫۹۹

$$R(\tilde{u}) \leq R(u^*).$$

۵. خانواده زیر از ماتریسهای مرتبه یک را در نظر بگیرید:

$$\mathbb{M}^{(n,d)}(۱) := \{A \in \mathbb{R}^{n\times d} | \mathsf{rank}(A) = ۱, \|A\|_F = ۱\}$$

کران پایین زیر را برای عدد پکینگ این مجموعه اثبات نمایید:

$$\log \mathcal{P}\left(\mathbb{M}^{(n,d)}(۱), \|.\|_F, \delta\right) \geq c(n+d) \log \frac{۱}{\delta}$$