Theoretical Stats and Machine Learning (Homework 04)

Student Name: Trang Pham

UT EID: vp22673

Question 1: Entropy method beyond uniform gradient bounds

a. Young's inequality for entropy.

First, we define the titled random density $Q := \frac{e^W}{\mathbb{E}e^W}$. We can see that $Q \geq 0$ and $\mathbb{E}[Q] = \frac{\mathbb{E}e^W}{\mathbb{E}e^W} = 1$.

Now consider the relative entropy between Y snd Q. We have:

$$\begin{aligned}\mathrm{KL}(Y||Q) &= \mathbb{E}\left[Y \log \frac{Y}{Q}\right] \\ &= \mathbb{E}\left[Y \log Y - Y \log e^W + Y \log(\mathbb{E}e^W)\right] \\ &= \mathbb{E}\left[Y \log Y - WY + Y \log(\mathbb{E}e^W)\right] \\ &= \mathbb{E}[Y \log Y] - \mathbb{E}[WY] + \log(\mathbb{E}e^W)\mathbb{E}[Y] \\ &= \mathrm{Ent}[Y] - \mathbb{E}[WY] + \log(\mathbb{E}e^W)\end{aligned}$$

because $\mathrm{Ent}[Y] = \mathbb{E}[Y \log Y] - \mathbb{E}[Y] \log \mathbb{E}[Y]$ and $\mathbb{E}[Y] = 1$ then $\mathrm{Ent}[Y] = \mathbb{E}[Y \log Y]$.

Because the relative entropy between P and Q is positive, we have $\mathrm{Ent}[Y] - \mathbb{E}[WY] + \log(\mathbb{E}e^W) \geq 0$, which implies that $\mathbb{E}[WY] \leq \log(\mathbb{E}e^W) + \mathrm{Ent}[Y]$, or equivalently, $\mathbb{E}[WY] \leq \kappa_W(1) + \mathrm{Ent}(Y)$.

The equality holds when $\mathrm{KL}(Y||Q) = 0$, which means $Y = \frac{e^W}{\mathbb{E}e^W}$.

b. A nonlinear Bernstein inequality from MLSI.

b1. By the MLSI, for any smooth function u, we have:

$$\mathrm{Ent}(e^{u(X)}) \leq C\,\mathbb{E}\Big[\|\nabla u(X)\|_2^2 e^{u(X)}\Big]$$

Now we this with $u(X) = \theta Z = \theta\big(f(X) - \mathbb{E}f(X)\big)$.

Because $\mathbb{E}f(X)$ is constant, we have:

$$\nabla u(X) = \theta \nabla f(X)$$

This implies:

$$\|\nabla u(X)\|_2^2 = \theta^2 \|\nabla f(X)\|_2^2 = \theta^2 G$$

Therefore, we can conclude that: $\mathrm{Ent}(e^{\theta Z}) \le C\,\mathbb{E}\big[\theta^2 G e^{\theta Z}\big] = C\theta^2 \mathbb{E}\big[G e^{\theta Z}\big]$.

b2. Fix $\eta > 0$. Apply part (a) with $Y = \frac{e^{\theta Z}}{\mathbb{E}e^{\theta Z}}$ and $W = \eta G$.

Then $\mathbb{E}Y = 1$, so part (a) gives

$$\mathbb{E}[WY] \le \kappa_W(1) + \mathrm{Ent}(Y).$$

Now $\mathbb{E}[WY]$. Because $W = \eta G$, we have: $\kappa_W(1) = \log \mathbb{E}e^{\eta G} = \kappa_G(\eta)$.

By the homogeneity of entropy, we have $\mathrm{Ent}(Y) = \frac{\mathrm{Ent}(e^{\theta Z})}{\mathbb{E}e^{\theta Z}}$ because $Y = \frac{e^{\theta Z}}{\mathbb{E}e^{\theta Z}}$.

Therefore:

$$\eta\,\frac{\mathbb{E}[Ge^{\theta Z}]}{\mathbb{E}e^{\theta Z}} \le \kappa_G(\eta) + \frac{\mathrm{Ent}(e^{\theta Z})}{\mathbb{E}e^{\theta Z}}$$

which implies $\frac{\mathbb{E}[Ge^{\theta Z}]}{\mathbb{E}e^{\theta Z}} \le \frac{\kappa_G(\eta)}{\eta} + \frac{\mathrm{Ent}(e^{\theta Z})}{\eta\,\mathbb{E}e^{\theta Z}}$.

b3. Divide the inequality in part (b1) by $\mathbb{E}e^{\theta Z}$, we obtain:

$$\frac{\mathrm{Ent}(e^{\theta Z})}{\mathbb{E}e^{\theta Z}} \le C\theta^2 \frac{\mathbb{E}[Ge^{\theta Z}]}{\mathbb{E}e^{\theta Z}}$$

Apply part (b2), we have:

$$\begin{aligned}
&\frac{\mathrm{Ent}(e^{\theta Z})}{\mathbb{E}e^{\theta Z}} \le C\theta^2\left(\frac{\kappa_G(\eta)}{\eta} + \frac{\mathrm{Ent}(e^{\theta Z})}{\eta\,\mathbb{E}e^{\theta Z}}\right)\\
&\iff \left(1 - \frac{C\theta^2}{\eta}\right)\frac{\mathrm{Ent}(e^{\theta Z})}{\mathbb{E}e^{\theta Z}} \le \frac{C\theta^2\kappa_G(\eta)}{\eta}\\
&\implies \frac{\mathrm{Ent}(e^{\theta Z})}{\mathbb{E}e^{\theta Z}} \le \frac{C\theta^2\kappa_G(\eta)}{\eta - C\theta^2} \qquad \text{(because } n > C\theta^2\text{)}\\
&\implies \frac{\mathrm{Ent}(e^{\theta Z})}{\theta^2\,\mathbb{E}e^{\theta Z}} \le \frac{C\,\kappa_G(\eta)}{\eta - C\theta^2} \qquad \text{(because } \theta^2 > 0\text{)}
\end{aligned}$$

b4. We have the Herbst identity:

$$\frac{d}{d\theta}\left(\frac{\psi_f(\theta)}{\theta}\right) = \frac{\mathrm{Ent}(e^{\theta Z})}{\theta^2\,\mathbb{E}e^{\theta Z}}$$

Therefore

$$\frac{\psi_f(\theta)}{\theta} - \lim_{s\to 0}\frac{\psi_f(s)}{s} = \int_0^\theta \frac{\mathrm{Ent}(e^{sZ})}{s^2\,\mathbb{E}e^{sZ}}\,ds,$$

We have $\lim_{s\to 0}\frac{\psi_f(s)}{s} = \lim_{s\to 0}\frac{\psi_f'(s)}{1} = \psi_f'(0)$. Moreover $\psi_f'(s) = \frac{d}{ds}\log\mathbb{E}e^{sZ} = \frac{\mathbb{E}[Ze^{sZ}]}{\mathbb{E}e^{sZ}}$. Therefore $\psi_f'(0) = \frac{\mathbb{E}[Z]}{1} = \mathbb{E}Z$, which implies $\lim_{s\to 0}\frac{\psi_f(s)}{s} = 0$. Therefore, the boundary term at 0 vanishes. Apply part (b3), we obtain:

$$\frac{\psi_f(\theta)}{\theta} \le \int_0^\theta \frac{C\,\kappa_G(\eta)}{\eta - Cs^2}\,ds$$

For $0 \le s \le \theta$, we have $\eta - Cs^2 \ge \eta - C\theta^2$ so $\frac{1}{\eta - Cs^2} \le \frac{1}{\eta - C\theta^2}$.

Therefore:

$$\frac{\psi_f(\theta)}{\theta} \le \int_0^\theta \frac{C\,\kappa_G(\eta)}{\eta - C\theta^2}\,ds = \frac{C\theta\,\kappa_G(\eta)}{\eta - C\theta^2}$$

which implies $\psi_f(\theta) \le \frac{C\theta^2\,\kappa_G(\eta)}{\eta - C\theta^2}$.

b5. By Chernoff's bound, for any $\theta > 0$, we have:

$$\mathbb{P}(Z \ge t) = \mathbb{P}(e^{\theta Z} \ge e^{\theta t}) \le e^{-\theta t}\,\mathbb{E}e^{\theta Z} = \exp\bigl(-\theta t + \psi_f(\theta)\bigr)$$

Apply part (b4), whenever $\eta > C\theta^2$, we have:

$$\mathbb{P}(Z \ge t) \le \exp\left(-\theta t + \frac{C\theta^2\kappa_G(\eta)}{\eta - C\theta^2}\right)$$

Because this holds for every $\eta > 0$ and every $0 \le \theta < \sqrt{\eta/C}$, we optimize over both parameters to obtain:

$$\mathbb{P}\{f(X) - \mathbb{E}f(X) \ge t\} \le \inf_{\eta>0}\ \inf_{0\le\theta<\sqrt{\eta/C}} \exp\left(-\theta t + \frac{C\theta^2\kappa_G(\eta)}{\eta - C\theta^2}\right).$$

This is the desired one-sided nonlinear Bernstein inequality.

c. [Bonus] Application: self-bounded functions.

c1. For all $\eta \ge 0$, we have: $\kappa_G(\eta) = \log\mathbb{E}e^{\eta G} \le \log\mathbb{E}e^{\eta aZ + \eta b}$.

Therefore: $\kappa_G(\eta) \le \log\mathbb{E}e^{\eta aZ+\eta b} = \log(e^{\eta b}\mathbb{E}e^{\eta aZ}) = \log(e^{\eta b}) + \log\mathbb{E}e^{\eta aZ} = b\eta + \psi_f(a\eta)$.

c2. From part (b4) and (c1), we obtain:

$$\psi_f(\theta) \leq \frac{C\theta^2 \kappa_G(\eta)}{\eta - C\theta^2} \leq \frac{C\theta^2(b\eta + \psi_f(a\eta))}{\eta - C\theta^2}$$

Choose $\eta = \frac{\theta}{a}$, we have:

$$\psi_f(\theta) \leq \frac{C\theta^2\left(b \cdot \frac{\theta}{a} + \psi_f(\theta)\right)}{\frac{\theta}{a} - C\theta^2} = \frac{C\theta\left(b\frac{\theta}{a} + \psi_f(\theta)\right)}{\frac{1}{a} - C\theta} = \frac{C\theta\left(b\frac{\theta}{a} + \psi_f(\theta)\right) \cdot a}{1 - aC\theta} = \frac{Cb\theta^2 + Ca\theta\psi_f(\theta)}{1 - aC\theta}$$

$$\Rightarrow \psi_f(\theta) \leq \frac{Cb\theta^2 + Ca\theta\psi_f(\theta)}{1 - aC\theta}$$

$$\iff \psi_f(\theta) - \frac{Ca\theta\psi_f(\theta)}{1 - aC\theta} \leq \frac{Cb\theta^2}{1 - aC\theta}$$

$$\iff \psi_f(\theta)\left(\frac{1 - 2aC\theta}{1 - aC\theta}\right) \leq \frac{Cb\theta^2}{1 - aC\theta}$$

Because $0 \leq \theta < \frac{1}{2cA}$, $a > 0$, $\theta \geq 0$ and $C > 0$, we have $1 - 2aC\theta > 0$ and $1 - aC\theta > 0$. Therefore, $\psi_f(\theta)\left(\frac{1-2aC\theta}{1-aC\theta}\right) \leq \frac{Cb\theta^2}{1-aC\theta}$ implies $\psi_f(\theta) \leq \frac{Cb\theta^2}{1-2aC\theta}$.

c3. By Chernoff's bound, for any $\theta > 0$, we have:

$$\mathbb{P}(Z \geq t) \leq \exp\bigl(-\theta t + \psi_f(\theta)\bigr)$$

From part (c2), for $0 \leq \theta < \frac{1}{2Ca}$, we have: $\psi_f(\theta) \leq \frac{Cb\theta^2}{1-2Ca\theta}$. Therefore $\mathbb{P}(Z \geq t) \leq \exp\left(-\theta t + \frac{Cb\theta^2}{1-2Ca\theta}\right)$. Choose $\theta = \min\left\{\frac{t}{4Cb}, \frac{1}{4Ca}\right\}$ then $\theta \leq \frac{1}{4Ca}$, so $1 - 2Ca\theta \geq \frac{1}{2}$, and thus $\frac{Cb\theta^2}{1-2Ca\theta} \leq 2Cb\theta^2$.

Therefore:

$$-\theta t + \frac{Cb\theta^2}{1 - 2Ca\theta} \leq -\theta t + 2Cb\theta^2$$

Also, because $\theta \leq \frac{t}{4Cb}$, we have $2Cb\theta^2 \leq \frac{\theta t}{2}$. Therefore:

$$-\theta t + 2Cb\theta^2 \leq -\frac{\theta t}{2}$$

From $\mathbb{P}(Z \geq t) \leq \exp\left(-\frac{\theta t}{2}\right)$, we substituting θ to obtain:

$$\frac{\theta t}{2} = \frac{1}{2}\, t \min\left\{\frac{t}{4Cb}, \frac{1}{4Ca}\right\} = \min\left\{\frac{t^2}{8Cb}, \frac{t}{8Ca}\right\}.$$

Therefore:

$$\mathbb{P}\{f(X)-\mathbb{E}f(X)\geq t\}\leq \exp\left(-\min\left\{\frac{t^2}{8Cb},\frac{t}{8Ca}\right\}\right)=\exp\left(-c\min\left\{\frac{t^2}{b},\frac{t}{a}\right\}\right),$$

with $c=\frac{1}{8C}$.

If $a=0$, then $G\leq b$ almost surely. In this case, $\psi_f(\theta)\leq Cb\theta^2$. Therefore, $\mathbb{P}\{f(X)-\mathbb{E}f(X)\geq t\}\leq \exp\left(-c\,\frac{t^2}{b}\right)$ which means that f(X) is sub-Gaussian.

d. [Bonus] Application: positive semidefinite quadratic forms.

d1. We have:

$$\begin{aligned}\mathbb{E}f(X)&=\mathbb{E}[X^\top AX]-\operatorname{tr}(A)\\&=\mathbb{E}[\operatorname{tr}(X^\top AX)]-\operatorname{tr}(A)\qquad\text{(because } X^TAX \text{ is a scalar)}\\&=\mathbb{E}[\operatorname{tr}(AXX^\top)]-\operatorname{tr}(A)\qquad\text{(cyclic property)}\\&=\operatorname{tr}\mathbb{E}[AXX^\top]-\operatorname{tr}(A)\\&=\operatorname{tr}(A\mathbb{E}[XX^\top])-\operatorname{tr}(A)\\&=\operatorname{tr}(AI)-\operatorname{tr}(A)\qquad\text{(because X is isotropic)}\\&=0\end{aligned}$$

Next, because A is symmetric, we have $\nabla f(X)=\nabla(X^\top AX)=2AX$ so $\|\nabla f(X)\|_2^2=\|2AX\|_2^2=4X^\top A^2X$.

Because $A\succeq 0$, we have $A^2\preceq\|A\|A$. Indeed, because $A\succeq 0$, there exists an orthogonal matrix U such that $A=U^\top\Lambda$ with $\Lambda=\operatorname{diag}(\lambda_1,\dots,\lambda_n),\ \lambda_i\geq 0$. Then $A^2=U^\top\Lambda^2U$. Because $\|A\|=\max_i\lambda_i$ and $\lambda_i\geq 0$, we have

$$\lambda_i^2\leq\|A\|\lambda_i\quad\text{for all } i,$$

which implies $\Lambda^2\preceq\|A\|\Lambda$.

Therefore:

$$A^2=U^\top\Lambda^2U\preceq U^\top(\|A\|\Lambda)U=\|A\|A$$

Then for all $X \in \mathbb{R}^n$, we obtain:

$$X^\top A^2 X \leq \|A\|\, X^\top A X$$

Apply the above result, we get:

$$\|\nabla f(X)\|_2^2 = 4X^\top A^2 X \leq 4\|A\|\, X^\top A X = 4\|A\|\big(f(X) + \operatorname{tr}(A)\big).$$

We can see that $f(X)$ is self-bounded with $a = 4\|A\|, b = 4\|A\|\ \operatorname{tr}(A)$.

d2. Apply part (c3) to the function $f(X) = X^\top A X - \operatorname{tr}(A)$ with $a = 4\|A\|, b = 4\|A\|\ \operatorname{tr}(A)$, we obtain

$$\mathbb{P}\{X^\top A X - \operatorname{tr}(A) \geq t\} \leq \exp\left(-c(C) \min\left\{\frac{t^2}{4\|A\| \operatorname{tr}(A)}, \frac{t}{4\|A\|}\right\}\right).$$

The factors 4 can be absorbed into the constant to get:

$$\mathbb{P}\{X^\top A X - \operatorname{tr}(A) \geq t\} \leq \exp\left(-c \min\left\{\frac{t^2}{\|A\| \operatorname{tr}(A)}, \frac{t}{\|A\|}\right\}\right),$$

for some constant $c > 0$ depending only on the MLSI constant C.

d3. For Gaussian X, the Hanson-Wright type bound from Homework 3 gives the sharper estimate for the standard Gaussian vector $X^\top A X - \operatorname{tr}(A)$:

$$\mathbb{P}\{X^\top A X - \operatorname{tr}(A) \geq t\} \leq \exp\left(-c \min\left\{\frac{t^2}{\|A\|_F^2}, \frac{t}{\|A\|}\right\}\right).$$

Thus the linear regime $t/\|A\|$ is the same, but the quadratic regime here uses $\|A\|\ \operatorname{tr}(A)$ instead of the sharper Gaussian quantity $\|A\|_F^2$.

If $\lambda_1, \dots, \lambda_n \geq 0$ are the eigenvalues of A, then:

$$\|A\|_F^2 = \sum_i \lambda_i^2 \leq \lambda_{\max} \sum_i \lambda_i = \|A\|\ \operatorname{tr}(A).$$

Therefore, the bound from part (d2) is weaker in the sub-Gaussian regime. The reason is that the MLSI/self-bounded argument is more general and only uses the rough estimate $A^2 \preceq \|A\|A$ while in the Gaussian case we can exploit the exact diagonal representation of the quadratic form, leading to the sharper Frobenius-scale term $\|A\|_F^2$.

Question 2. Practice with covering and packing

a. Monotonicity properties

a1. *Covering number.* By definition of covering number, $\mathcal{N}(K,d,\varepsilon)=\min\big\{|\mathcal{M}|:\mathcal{M}\subset K \text{ is an } \varepsilon\text{-net of } K\big\}$. Let $\mathcal{M}\subset K$ be an ε_1-net of K. Then for every $x\in K$, there exists some $m\in\mathcal{M}$ such that $d(x,m)\le\varepsilon_1$. Because $\varepsilon_1\le\varepsilon_2$, we also have $d(x,m)\le\varepsilon_2$. Therefore, for every $x\in K$, there exists some $m\in\mathcal{M}$ such that $d(x,m)\le\varepsilon_2$. This means that $\mathcal{M}$ is also an ε_2-net of K.

Hence the collection of ε_1-nets is contained in the collection of ε_2-nets. Taking minima over these two collections, we obtain: $\mathcal{N}(K,d,\varepsilon_2)\le\mathcal{N}(K,d,\varepsilon_1)$. Therefore, $\varepsilon\mapsto\mathcal{N}(K,d,\varepsilon)$ is weakly decreasing.

Packing number. By definition of packing number, $\mathcal{P}(K,d,\varepsilon):=\max\big\{\,|\mathcal{M}|:\mathcal{M}\subset K \text{ is } \varepsilon\text{-separated}\big\}$. Let $\mathcal{M}\subset K$ is ε_2-separated. Then for every $x,y\in\mathcal{M}$, we have $d(x,y)>\varepsilon_2$. Because $\varepsilon_2\ge\varepsilon_1$, we also have $d(x,y)>\varepsilon_1$. This means that any $\mathcal{M}$ which is ε_2-separated is also ε_1-separated. Therefore, the collection of $\mathcal{M}$ which is ε_2-separated is contained in the collection of $\mathcal{M}$ which is ε_1-separated. Take the maxima over these two collections, we obtain $\mathcal{P}(K,d,\varepsilon_2)\le\mathcal{P}(K,d,\varepsilon_1)$, which implies that $varepsilon\mapsto\mathcal{P}(K,d,\varepsilon)$ is weakly decreasing.

a2. Example. Work in $T=\mathbb{R}$ with the Euclidean metric, and let $K=[0,1]$ and $L=\{0,1\}\subset K$. Take $\varepsilon=0.6$. Then $\mathcal{N}(K,d,\varepsilon)=1$ because $B(0.5,0.6)=[-0.1,1.1]\supset[0,1]=K$, so $\{0.5\}$ is an ε-net of K. On the other hand, $\mathcal{N}(L,d,\varepsilon)=2$ because $d(0,1)=1>0.6=\varepsilon$, so no single ball of radius ε can cover both points, while $\{0,1\}$ clearly forms an ε-net of L. Therefore $L\subset K$ but $\mathcal{N}(L,d,\varepsilon)=2>1=\mathcal{N}(K,d,\varepsilon)$, which shows that the covering number is not monotone in the underlying set.

a3. Let $M\subset K$ be an $(\varepsilon/2)$-net of K with $|M|=\mathcal{N}(K,d,\varepsilon/2)$. For each $m\in M$ such that $L\cap B(m,\varepsilon/2)\neq\emptyset$, choose a representative point $\ell_m\in L\cap B(m,\varepsilon/2)$.

Let $M_L:=\{\ell_m : m\in M,\ L\cap B(m,\varepsilon/2)\neq\emptyset\}$. Then $|M_L|\le|M|$ because each m contributes at most one representative.

We claim that M_L is an ε-net of L. Take any $x\in L$. Because M is an $(\varepsilon/2)$-net of K, there exists $m\in M$ such that $d(x,m)\le\varepsilon/2$. Therefore, $x\in L\cap B(m,\varepsilon/2)$, so ℓ_m is defined and satisfies

$d(\ell_m, m) \leq \varepsilon/2$. By the triangle inequality, we have:

$$d(x, \ell_m) \leq d(x, m) + d(m, \ell_m) \leq \varepsilon/2 + \varepsilon/2 = \varepsilon.$$

Thus, for every $x \in L$, there exists $\ell_m \in M_L$ such that $d(x, \ell_m) \leq \varepsilon$, so M_L is an ε-net of L. Therefore, $\mathcal{N}(L, d, \varepsilon) \leq |M_L| \leq |M| = \mathcal{N}(K, d, \varepsilon/2)$.

b. Packing versus covering

b1. Because $\mathcal{M}$ is maximal ϵ-separated, it is enough to show that if $\mathcal{M}$ were not an ϵ-net of K, then we could add one more point of K to $\mathcal{M}$ while preserving ϵ-separation. Suppose for contradiction that $\mathcal{M}$ is not an ϵ-net. Then there exists some $x \in K$ such that for every $y \in \mathcal{M}$, $d(x, y) > \epsilon$. However, this means exactly that x is more than ϵ away from every point already in $\mathcal{M}$. Therefore, if we consider $\mathcal{M} \cup \{x\}$, then any two distinct old points in $\mathcal{M}$ are still more than ϵ apart because $\mathcal{M}$ was already ϵ-separated, and the new point x is also more than ϵ away from every point of $\mathcal{M}$. Therefore, $\mathcal{M} \cup \{x\}$ is still ϵ-separated, contradicting the maximality of $\mathcal{M}$. Therefore such an x cannot exist, so for every $x \in K$ there exists some $y \in \mathcal{M}$ with $d(x, y) \leq \epsilon$. This shows that $\mathcal{M}$ is an ϵ-net of K.

b2. To prove $\mathcal{P}(K, d, 2\epsilon) \leq \mathcal{N}(K, d, \epsilon) \leq \mathcal{P}(K, d, \epsilon)$, we establish the two inequalities separately.

Upper bound. Let $\mathcal{M} \subset K$ be a maximal ϵ-separated set. By part (b1), $\mathcal{M}$ is an ϵ-net of K. Therefore, by definition of the covering number, we have: $\mathcal{N}(K, d, \epsilon) \leq |\mathcal{M}|$. Because $\mathcal{M}$ is also ϵ-separated, its cardinality cannot exceed the packing number, which implies $|\mathcal{M}| \leq \mathcal{P}(K, d, \epsilon)$. Combining the two inequalities, we obtain $\mathcal{N}(K, d, \epsilon) \leq \mathcal{P}(K, d, \epsilon)$.

Lower bound. Let $\mathcal{C} \subset K$ be any ϵ-net of K, and let $\mathcal{S} \subset K$ be any 2ϵ-separated set. Because $\mathcal{C}$ is an ϵ-net, for every $x \in \mathcal{S}$ there exists some $c(x) \in \mathcal{C}$ such that $d(x, c(x)) \leq \epsilon$. We claim that the map $x \mapsto c(x)$ is injective. Indeed, if $x, y \in \mathcal{S}$ satisfy $c(x) = c(y) = c$, then:

$$d(x, y) \leq d(x, c) + d(c, y) \leq \epsilon + \epsilon = 2\epsilon$$

which contradicts the fact that $\mathcal{S}$ is 2ϵ-separated. Therefore, distinct points of $\mathcal{S}$ must be mapped to distinct points of $\mathcal{C}$, which implies $|\mathcal{S}| \leq |\mathcal{C}|$. Because this holds for every 2ϵ-separated set $\mathcal{S}$ and every ϵ-net $\mathcal{C}$, taking the maximum over all such $\mathcal{S}$ and the minimum over all such $\mathcal{C}$, we have:

$$\mathcal{P}(K, d, 2\epsilon) \leq \mathcal{N}(K, d, \epsilon).$$

Therefore: $\mathcal{P}(K, d, 2\epsilon) \leq \mathcal{N}(K, d, \epsilon) \leq \mathcal{P}(K, d, \epsilon)$.

From the above result, we see that the covering number is sandwiched between two packing numbers evaluated at scales ε and 2ε. Therefore, if we know $\mathcal{P}(K, \varepsilon)$, we can determine $\mathcal{N}(K, \varepsilon)$ up to a constant rescaling of ε. Particularly, packing and covering numbers have the same growth rate (up to universal constants) and therefore they are equivalent at the level of metric entropy. **c. [Bonus] Volumetric bounds in Euclidean space**

c1. Let $K \subset \mathbb{R}^n$ be measurable. We prove the volumetric bounds.

Left inequality. Let $\{x_1, \ldots, x_N\}$ be an ϵ-net of K with $N = \mathcal{N}(K, \epsilon)$. Then $K \subseteq \bigcup_{i=1}^{N} \left(x_i + \epsilon B_2^n\right)$. Therefore $\mathrm{vol}(K) \leq \sum_{i=1}^{N} \mathrm{vol}(\epsilon B_2^n) = N \ \mathrm{vol}(\epsilon B_2^n)$. Divide both sides by $\mathrm{vol}(\epsilon B_2^n)$, we obtain:

$$\frac{\mathrm{vol}(K)}{\mathrm{vol}(\epsilon B_2^n)} \leq \mathcal{N}(K, \epsilon).$$

Right inequality. Let $\{x_1, \ldots, x_M\} \subset K$ be an ϵ-separated set with $M = \mathcal{P}(K, \epsilon)$. Then the sets $x_i + (\epsilon/2)B_2^n$ are pairwise disjoint. Moreover, because $x_i \in K$, we have:

$$x_i + (\epsilon/2)B_2^n \subseteq K + (\epsilon/2)B_2^n$$

Therefore: $\bigcup_{i=1}^{M} \left(x_i + (\epsilon/2)B_2^n\right) \subseteq K + (\epsilon/2)B_2^n$.

Take volumes and using disjointness, we obtain:

$$M \ \mathrm{vol}((\epsilon/2)B_2^n) \leq \mathrm{vol}\left(K + (\epsilon/2)B_2^n\right)$$

This implies: $\mathcal{P}(K, \epsilon) \leq \frac{\mathrm{vol}\left(K + (\epsilon/2)B_2^n\right)}{\mathrm{vol}((\epsilon/2)B_2^n)}$.

c2. Applying (c1) with $K = B_2^n$, we obtain:

$$\frac{\mathrm{vol}(B_2^n)}{\mathrm{vol}(\epsilon B_2^n)} \leq \mathcal{N}(B_2^n, \epsilon) \leq \frac{\mathrm{vol}\left(B_2^n + (\epsilon/2)B_2^n\right)}{\mathrm{vol}((\epsilon/2)B_2^n)}$$

Using $\mathrm{vol}(rB_2^n) = r^n \,\mathrm{vol}(B_2^n)$, we have $\mathrm{vol}(\epsilon B_2^n) = \epsilon^n \,\mathrm{vol}(B_2^n)$.

Therefore: $\left(\frac{1}{\epsilon}\right)^n \le \mathcal{N}(B_2^n, \epsilon)$.

We also claim that $B_2^n + (\epsilon/2)B_2^n = (1+\epsilon/2)B_2^n$. Indeed, if $z \in B_2^n + (\epsilon/2)B_2^n$, then $z = x + y$ for some x, y with $\|x\|_2 \le 1$ and $\|y\|_2 \le \epsilon/2$, so by the triangle inequality:

$$\|z\|_2 \le \|x\|_2 + \|y\|_2 \le 1 + \epsilon/2,$$

which implies $z \in (1+\epsilon/2)B_2^n$. Conversely, if $z \in (1+\epsilon/2)B_2^n$, then $\|z\|_2 \le 1 + \epsilon/2$. If $z = 0$ the claim is trivial; otherwise, define:

$$x := \frac{1}{1+\epsilon/2} z, \qquad y := z - x.$$

Then $\|x\|_2 \le 1$ and $\|y\|_2 = \left\| z - \frac{1}{1+\epsilon/2} z \right\|_2 = \frac{\epsilon/2}{1+\epsilon/2} \|z\|_2 \le \epsilon/2$, so $x \in B_2^n$ and $y \in (\epsilon/2)B_2^n$, hence $z = x + y \in B_2^n + (\epsilon/2)B_2^n$. Therefore the two sets are equal.

Apply the above result, we have: $\mathcal{N}(B_2^n, \epsilon) \le \frac{(1+\epsilon/2)^n}{(\epsilon/2)^n} = \left(1 + \frac{2}{\epsilon}\right)^n$.

For $0 < \epsilon \le 1$, we have $1 + \frac{2}{\varepsilon} = \frac{\varepsilon + 2}{\varepsilon} \le \frac{3}{\varepsilon}$. Therefore:

$$\left(\frac{1}{\epsilon}\right)^n \le \mathcal{N}(B_2^n, \epsilon) \le \left(\frac{3}{\epsilon}\right)^n.$$

c3. Because $\mathcal{N}(K, \epsilon) \le \mathcal{P}(K, \epsilon)$, if we can upper bound $\mathcal{P}(S^{n-1}, \epsilon)$ then the bound for $\mathcal{N}(K, \epsilon)$ will follows. Let $\{x_1, \ldots, x_M\} \subset S^{n-1}$ be an ϵ-separated set. Then the balls $x_i + (\epsilon/2)B_2^n$ are disjoint. For any $z = x_i + u$ with $\|u\|_2 \le \epsilon/2$, we have $\|z\|_2 \le \|x_i\|_2 + \|u\|_2 \le 1 + \epsilon/2$ (by triangular inequality). Therefore: $x_i + (\epsilon/2)B_2^n \subseteq (1+\epsilon/2)B_2^n$ for all i, which yields $\bigcup_{i=1}^M (x_i + (\epsilon/2)B_2^n) \subseteq (1+\epsilon/2)B_2^n$. This implies $\mathrm{vol}(\bigcup_{i=1}^M (x_i + (\epsilon/2)B_2^n)) \le M \;\mathrm{vol}((\epsilon/2)B_2^n) \le \mathrm{vol}((1+\epsilon/2)B_2^n)$ (by union bound). Therefore: $\mathcal{P}(S^{n-1}, \epsilon) \le \left(1 + \frac{2}{\epsilon}\right)^n$, which implies $\mathcal{N}(S^{n-1}, \epsilon) \le \left(1 + \frac{2}{\epsilon}\right)^n$.

c4. From (c2), for $0 < \epsilon \le 1/2$, we have:

$$\left(\frac{1}{\epsilon}\right)^n \le \mathcal{N}(B_2^n, \epsilon) \le \left(\frac{2.5}{\epsilon}\right)^n$$
$$\iff n \log(1/\epsilon) \le H(B_2^n, \epsilon) \le n \log(1/\epsilon) + n \log 2.5.$$

Because $\log(1/\epsilon) \geq \log 2 > 0$, the term $n \log 2.5$ is bounded by a constant multiple of $n\log(1/\epsilon)$ (we have $\frac{\log 2.5}{\log(1/\epsilon)} \leq \frac{\log 2.5}{\log 2} = C$ then $n\log(2.5) \leq Cn\log(\frac{1}{\varepsilon})$). Therefore $H(B_2^n, \epsilon) \asymp n\log(1/\epsilon) \asymp n\log(e/\epsilon)$ up to absolute constants.

Question 3. Covariance estimation with sub-Gaussian random vectors

a. Coordinates and dependence

a1. We have: $\langle X, u\rangle = \sum_{i=1}^d u_i X_i$. In order to prove that X is a sub-Gaussian random vector, we need to show that $\sum_{i=1}^d u_i X_i$ is sub-Gaussian random variable for all u.

Let $Z = \sum_{i=1}^d u_i X_i$, we have:

$$\log \mathbb{E} e^{\lambda Z} = \log \mathbb{E} e^{\lambda \sum_{i=1}^d u_i X_i} = \sum_{i=1}^d \log \mathbb{E} e^{\lambda u_i X_i} \qquad \text{(because } X_i \text{ are independent)}$$

Because each X_i is sub-Gaussian, mean-zero, there exists $\sigma_i^2 > 0$ such that for all $t \in \mathbb{R}$, $\log \mathbb{E} e^{tX_i} \leq \frac{\sigma_i^2 t^2}{2}$. Taking $t = \lambda u_i$, we get: $\log \mathbb{E} e^{\lambda u_i X_i} \leq \frac{\sigma_i^2 \lambda^2 u_i^2}{2}$. Therefore:

$$\log \mathbb{E} e^{\lambda Z} = \sum_{i=1}^d \log \mathbb{E} e^{\lambda u_i X_i} \leq \sum_{i=1}^d \frac{\sigma_i^2 \lambda^2 u_i^2}{2} = \frac{\lambda^2}{2} \sum_{i=1}^d \sigma_i^2 u_i^2$$

This indicates that $\langle X, u\rangle$ is $\sum_{i=1}^d \sigma_i^2 u_i^2$ sub-Gaussian and this result holds for every u. Therefore, we can conclude that X is a sub-Gaussian random vector.

Lower bound. By definition, we have $||X||_{\psi_2} = \sup_{u \in S^{d-1}} ||\langle X, u\rangle||_{\psi_2}$, which implies that $||X||_{\psi_2} \geq ||\langle X, u\rangle||_{\psi_2}$ for all $u \in S^{d-1}$. Choose $u = e_i$, where $\{e_1, \ldots, e_d\}$ is the standard basis of $\mathbb{R}^d$ and each e_i has a 1 in the i-th coordinate and 0 elsewhere. We have $\langle X, e_i\rangle = X_i$. Therefore: $\|X\|_{\psi_2} \geq \max_{1\leq i\leq d} \|\langle X, e_i\rangle\|_{\psi_2} = \max_{1 \leq i \leq d} \|X_i\|_{\psi_2}$.

Upper bound. Fix any $u \in S^{d-1}$ and define $Y_i := u_i X_i$ for $1 \leq i \leq d$. Because the X_i are independent, mean-zero, and sub-Gaussian, the same is true for the Y_i. Moreover: $\langle X, u\rangle = \sum_{i=1}^d u_i X_i = \sum_{i=1}^d Y_i$. By the sub-Gaussian Hoeffding inequality, we have:

$$\|\langle X, u\rangle\|_{\psi_2}^2 = \left\| \sum_{i=1}^d Y_i \right\|_{\psi_2}^2 \leq C \sum_{i=1}^d \|Y_i\|_{\psi_2}^2$$

Using the homogeneity of the ψ_2-norm, we get: $\|Y_i\|_{\psi_2} = \|u_i X_i\|_{\psi_2} = |u_i|\,\|X_i\|_{\psi_2}$. Therefore:

$$\|\langle X, u\rangle\|_{\psi_2}^2 \le C \sum_{i=1}^{d} u_i^2 \|X_i\|_{\psi_2}^2 \le C \left(\max_{1\le i\le d} \|X_i\|_{\psi_2}^2 \right) \sum_{i=1}^{d} u_i^2$$

Because $u \in S^{d-1}$, we have $\sum_{i=1}^d u_i^2 = 1$, thus: $\|\langle X, u\rangle\|_{\psi_2}^2 \le C \max_{1\le i\le d} \|X_i\|_{\psi_2}^2$. Take square roots, we obtain: $\|\langle X, u\rangle\|_{\psi_2} \le C \max_{1\le i\le d} \|X_i\|_{\psi_2}$.

Finally, take the supremum over $u \in S^{d-1}$, we can conclude:

$$\|X\|_{\psi_2} = \sup_{u\in S^{d-1}} \|\langle X, u\rangle\|_{\psi_2} \le C \max_{1\le i\le d} \|X_i\|_{\psi_2}$$

a2. Consider the random vector $X = (Z, \dots, Z) \in \mathbb{R}^d$, where Z is a mean-zero sub-Gaussian random variable. Therefore, $X_i = Z$ for all i, which results in $\max_{1\le i\le d} \|X_i\|_{\psi_2} = \|Z\|_{\psi_2}$.

Moreover, for any $u \in S^{d-1}$, $\langle X, u\rangle = \sum_{i=1}^d u_i Z = \left(\sum_{i=1}^d u_i\right) Z$. Therefore:

$$\|\langle X, u\rangle\|_{\psi_2} = \left| \sum_{i=1}^{d} u_i \right| \|Z\|_{\psi_2}$$

By Cauchy-Schwarz, we have: $\left|\sum_{i=1}^d u_i\right| = |\langle u, \mathbf{1}\rangle| \le \|u\|_2 \|\mathbf{1}\|_2 = \sqrt{d}$. The equality holds if and only if u is proportional to $\mathbf{1} = (1, \dots, 1)$. Because $u \in S^{d-1}$, we obtain:

$$u = \frac{1}{\sqrt{d}}(1, \dots, 1).$$

Therefore: $\|X\|_{\psi_2} = \sup_{u\in S^{d-1}} \|\langle X, u\rangle\|_{\psi_2} = \sqrt{d}\,\|Z\|_{\psi_2}$.

Consequently: $\|X\|_{\psi_2} = \sqrt{d} \max_{1\le i\le d} \|X_i\|_{\psi_2}$, which shows that the independence assumption in part (a1) is essential.

a3. [Bonus] Canonical example.

a31. Let $X \sim \mathrm{Unif}([-1,1]^d)$. We see that each coordinate $X_i \in [-1, 1]$, therefore $X_i^2 \le 1$. By the definition of the ψ_2- norm:

$$\|X_i\|_{\psi_2} = \inf\left\{t > 0 :\ \mathbb{E}\exp(X_i^2/t^2) \le 2\right\}$$

Because $X_i^2 \le 1$, we have: $\mathbb{E}\exp(X_i^2/t^2) \le \exp(1/t^2)$. Choose $t \ge 1/\sqrt{\log 2}$ ensures that $\exp(1/t^2) \le 2$, therefore $\|X_i\|_{\psi_2} \le \frac{1}{\sqrt{\log 2}} \le C_1$ for some absolute constant C_1.

By part (a1), we have:

$$\|\langle X, u\rangle\|_{\psi_2} = \left\| \sum_{i=1}^{d} u_i X_i \right\|_{\psi_2} \le C \left(\sum_{i=1}^{d} u_i^2 \|X_i\|_{\psi_2}^2 \right)^{1/2} \le CC_1 \sum_{i=1}^{d} u_i^2 = CC_1$$

Take supremum over $u \in S^{d-1}$, we obtain: $\|X\|_{\psi_2} \le C_2$ for some absolute constant C_2 which does not depend on d or i.

Boolean cube. For the Boolean cube $\{-1, 1\}^d$, each coordinate satisfies $|X_i| \le 1$, hence $X_i^2 \le 1$. The conclusion then follows by the same argument as above.

a32. Because $\langle X, u\rangle = u^T X \sim \mathcal{N}(0, u^\top \Sigma u)$ and $\|Z\|_{\psi_2} \asymp \sigma$ for $Z \sim \mathcal{N}(0, \sigma^2)$, we obtain:

$$\|\langle X, u\rangle\|_{\psi_2} \le C\sqrt{u^\top \Sigma u}.$$

Using the variational characterization of the operator norm, we have:

$$u^\top \Sigma u \le \|\Sigma\| \qquad \text{for all } u \in S^{d-1}$$

Therefore: $\|\langle X, u\rangle\|_{\psi_2} \le C\sqrt{\|\Sigma\|}$. Take the supremum over $u \in S^{d-1}$, we conclude: $\|X\|_{\psi_2} \le C\sqrt{\|\Sigma\|}$.

b. Covariance estimation for sub-Gaussian data

b1.

$Z_i(u)$ are independent. Because $X_1, \ldots, X_N$ are independent, it follows that $Z_i(u)$ are independent.

$Z_i(u)$ are mean-zero. Because X is isotropic, we have $\mathbb{E}\langle X_i, u\rangle^2 = u^\top \mathbb{E}[X_i X_i^\top] u = u^\top I_d u = \|u\|_2^2 = 1$. Therefore $\mathbb{E} Z_i(u) = E\langle X_i, u\rangle^2 - 1 = 0$.

$Z_i(u)$ are independent sub-Gaussian random variables. By assumption, we have $\langle X_i, u\rangle$ is sub-Gaussian and satisfies $\|\langle X_i, u\rangle\|_{\psi_2} \le \|X_i\|_{\psi_2} \le K$. By the given result, if W is sub-Gaussian, then $W^2 - \mathbb{E}W^2$ is sub-exponential and $\|W^2 - \mathbb{E}W^2\|_{\psi_1} \le C\|W\|_{\psi_2}^2$.

Apply this with $W = \langle X_i, u \rangle$, we obtain:

$$\|Z_i(u)\|_{\psi_1} = \|\langle X_i, u\rangle^2 - \mathbb{E}\langle X_i, u\rangle^2\|_{\psi_1} \leq CK^2.$$

Finally, by Bernstein's inequality for independent sub-exponential random variables, for every $t \geq 0$, we have:

$$\mathbb{P}\left\{\left|\frac{1}{N}\sum_{i=1}^{N} Z_i(u)\right| \geq t\right\} \leq 2\exp\left(-cN\min\left(\frac{t^2}{K^4}, \frac{t}{K^2}\right)\right).$$

We have $\frac{1}{N}\sum_{i=1}^{N} Z_i(u) = \frac{1}{N}\sum_{i=1}^{N}\langle X_i, u\rangle^2 - 1 = u^T\left(\frac{1}{N}X_iX_i^T\right)u - u^T I_d u = u^\top(\widehat{\Sigma} - I_d)u$.
Therefore, we conclude that:

$$\mathbb{P}\left\{\left|u^\top(\widehat{\Sigma} - I_d)u\right| \geq t\right\} \leq 2\exp\left(-cN\min\left(\frac{t^2}{K^4}, \frac{t}{K^2}\right)\right).$$

b2. Continue in the isotropic case. Let $\mathcal{N}$ be a 1/4-net of S^{d-1} with $|\mathcal{N}| \leq 12^d$. By the net reduction, for any symmetric matrix A, we have: $\|A\| \leq 2\max_{u\in\mathcal{N}}|u^\top A u|$.
Apply this with $A = \widehat{\Sigma} - I_d$, we obtain:

$$\|\widehat{\Sigma} - I_d\| \leq 2\max_{u\in\mathcal{N}}\left|u^\top(\widehat{\Sigma} - I_d)u\right|.$$

Therefore, for any $t > 0$, we have:

$$\begin{aligned}\mathbb{P}\Big\{\|\widehat{\Sigma} - I_d\| \geq t\Big\} &\leq \mathbb{P}\Big\{\max_{u\in\mathcal{N}}\left|u^\top(\widehat{\Sigma} - I_d)u\right| \geq t/2\Big\}\\ &\leq \sum_{u\in\mathcal{N}}\mathbb{P}\Big\{\left|u^\top(\widehat{\Sigma} - I_d)u\right| \geq t/2\Big\} \qquad \text{(by union bound)}\end{aligned}$$

From part (b1), for each fixed $u \in S^{d-1}$, we have:

$$\mathbb{P}\Big\{\left|u^\top(\widehat{\Sigma} - I_d)u\right| \geq s\Big\} \leq 2\exp\Big(-cN\min\left(\frac{s^2}{K^4}, \frac{s}{K^2}\right)\Big)$$

Applying this with $s = t/2$, we get:

$$\mathbb{P}\Big\{\|\widehat{\Sigma} - I_d\| \geq t\Big\} \leq 2|\mathcal{N}| \exp\Big(-cN \min\left(\frac{t^2}{K^4}, \frac{t}{K^2}\right)\Big)$$

Because $|\mathcal{N}| \leq 12^d$, we have:

$$\mathbb{P}\Big\{\|\widehat{\Sigma} - I_d\| \geq t\Big\} \leq 2 \cdot 12^d \exp\Big(-cN \min\left(\frac{t^2}{K^4}, \frac{t}{K^2}\right)\Big).$$

Now choose $t = CK^2\sqrt{\frac{d}{N}}$ then $\frac{t^2}{K^4} = C^2\frac{d}{N}$, $\frac{t}{K^2} = C\sqrt{\frac{d}{N}}$. Therefore:

$$N \min\left(\frac{t^2}{K^4}, \frac{t}{K^2}\right) = \min\left(C^2 d,\ C\sqrt{dN}\right).$$

Because $N \geq d$, we have $\sqrt{dN} \geq d$, and therefore $N \min\left(\frac{t^2}{K^4}, \frac{t}{K^2}\right) \geq cd$ for some absolute constant $c > 0$. As a result, $\mathbb{P}\Big\{\|\widehat{\Sigma} - I_d\| \geq t\Big\} \leq 2 \cdot 12^d\, e^{-cd} = 2\exp\big(d \log 12 - cd\big)$.

If we choose C sufficiently large so that $c > \log 12 + 1$, we conclude that:

$$\mathbb{P}\left\{\|\widehat{\Sigma} - I_d\| \geq CK^2\sqrt{\frac{d}{N}}\right\} \leq 2e^{-d}.$$

b3. [Bonus] Let $Y := \Sigma^{-1/2}X$.

Because $\mathbb{E}[XX^\top] = \Sigma$, we have: $\mathbb{E}[YY^\top] = \Sigma^{-1/2}\,\mathbb{E}[XX^\top]\,\Sigma^{-1/2} = \Sigma^{-1/2}\Sigma\Sigma^{-1/2} = I_d$.

Moreover, $\mathbb{E}Y = \mathbb{E}(\Sigma^{-1/2}X) = \Sigma^{-1/2}\,\mathbb{E}X = 0$. Therefore, Y is isotropic.

Next, we fix $v \in S^{d-1}$ and define:

$$u := \frac{\Sigma^{-1/2}v}{\|\Sigma^{-1/2}v\|_2}$$

Then $u \in S^{d-1}$, and $\langle Y, v\rangle = v^\top\Sigma^{-1/2}X = \langle X, \Sigma^{-1/2}v\rangle = \|\Sigma^{-1/2}v\|_2\langle X, u\rangle$.

Therefore: $\|\langle Y, v\rangle\|_{\psi_2} = \|\Sigma^{-1/2}v\|_2\,\|\langle X, u\rangle\|_{\psi_2}$.

By the assumption on X, we have:

$$\|\langle X, u\rangle\|_{\psi_2} \leq K\big(\mathbb{E}\langle X, u\rangle^2\big)^{1/2}.$$

We also have: $\mathbb{E}\langle X, u\rangle^2 = u^\top \Sigma u = \frac{v^\top \Sigma^{-1/2}\Sigma\Sigma^{-1/2} v}{\|\Sigma^{-1/2}v\|_2^2} = \frac{\|v\|_2^2}{\|\Sigma^{-1/2}v\|_2^2} = \frac{1}{\|\Sigma^{-1/2}v\|_2^2}$.

Therefore:

$$\|\langle Y, v\rangle\|_{\psi_2} \leq \|\Sigma^{-1/2}v\|_2 \cdot K \cdot \frac{1}{\|\Sigma^{-1/2}v\|_2} = K.$$

Because this result holds for all $v \in S^{d-1}$, we conclude: $\|Y\|_{\psi_2} \leq K$.

Now let $Y_1, \dots, Y_N$ be the i.i.d. copies of Y defined by $Y_i := \Sigma^{-1/2}X_i$ and let $\widehat{\Sigma}_Y := \frac{1}{N}\sum_{i=1}^N Y_i Y_i^\top$.

Then

$$\widehat{\Sigma}_Y = \frac{1}{N}\sum_{i=1}^N \Sigma^{-1/2}X_i X_i^\top \Sigma^{-1/2} = \Sigma^{-1/2}\widehat{\Sigma}\Sigma^{-1/2}$$

Therefore: $\widehat{\Sigma}_Y - I_d = \Sigma^{-1/2}(\widehat{\Sigma} - \Sigma)\Sigma^{-1/2}$.

Apply part (b2) to the isotropic vector Y, we obtain:

$$\mathbb{P}\left\{\|\widehat{\Sigma}_Y - I_d\| \geq CK^2\sqrt{\frac{d}{N}}\right\} \leq 2e^{-d}, \qquad \text{provided } N \geq d.$$

On the other hand, we have: $\widehat{\Sigma} - \Sigma = \Sigma^{1/2}(\widehat{\Sigma}_Y - I_d)\Sigma^{1/2}$.

Therefore, by submultiplicativity of the operator norm (homework 1), we have:

$$\|\widehat{\Sigma} - \Sigma\| \leq \|\Sigma^{1/2}\|^2 \, \|\widehat{\Sigma}_Y - I_d\| = \|\Sigma\| \, \|\widehat{\Sigma}_Y - I_d\|.$$

This implies that: if $\|\widehat{\Sigma}_Y - I_d\| < CK^2\sqrt{\frac{d}{N}}$ then $\|\widehat{\Sigma} - \Sigma\| < CK^2\|\Sigma\|\sqrt{\frac{d}{N}}$.

That is:

$$\left\{\|\widehat{\Sigma}_Y - I_d\| < CK^2\sqrt{\frac{d}{N}}\right\} \subseteq \left\{\|\widehat{\Sigma} - \Sigma\| < CK^2\|\Sigma\|\sqrt{\frac{d}{N}}\right\}$$

Equivalently, we have:

$$\left\{\|\widehat{\Sigma} - \Sigma\| \geq CK^2\|\Sigma\|\sqrt{\frac{d}{N}}\right\}^c \supseteq \left\{\|\widehat{\Sigma}_Y - I_d\| \geq CK^2\sqrt{\frac{d}{N}}\right\}^c$$

Take probabilities and use the complement relation, we obtain:

$$\mathbb{P}\left\{\|\widehat{\Sigma} - \Sigma\| \geq CK^2\|\Sigma\|\sqrt{\frac{d}{N}}\right\} \leq \mathbb{P}\left\{\|\widehat{\Sigma}_Y - I_d\| \geq CK^2\sqrt{\frac{d}{N}}\right\} \leq 2e^{-d}$$

We can conclude:

$$\mathbb{P}\left\{\|\widehat{\Sigma}-\Sigma\| \geq CK^2\|\Sigma\|\sqrt{\frac{d}{N}}\right\} \leq 2e^{-d}, \qquad \text{provided } N \geq d$$

Final Remark. If $X \sim \mathcal{N}(0,\Sigma)$, then for every $u \in S^{d-1}$, we have $\langle X,u\rangle \sim \mathcal{N}(0,u^\top \Sigma u)$. Therefore:

$$\|\langle X,u\rangle\|_{\psi_2} \leq C\sqrt{u^\top \Sigma u} = C\big(\mathbb{E}\langle X,u\rangle^2\big)^{1/2}$$

This implies that the assumption holds with $K \geq C$, and therefore:

$$\mathbb{P}\left\{\|\widehat{\Sigma}-\Sigma\| \geq C\|\Sigma\|\sqrt{\frac{d}{N}}\right\} \leq 2e^{-d}, \qquad \text{provided } N \geq d.$$

Question 4. Learning a spike model

a. [Bonus] Projection matrices and perturbation of top eigenvectors

a1. First, we choose $s \in \{-1,1\}$ such that $\langle u, sv\rangle \geq 0$. Because $u, sv \in S^{d-1}$, there exists $\theta \in [0,\pi/2]$ such that $\langle u, sv\rangle = \cos\theta$.

Now let $w := \frac{sv-\langle u,sv\rangle u}{\|sv-\langle u,sv\rangle u\|_2}$.

Therefore, $w \perp u$, $\|w\|_2 = 1$, and $sv = \cos\theta\, u + \sin\theta\, w$.

Upper bound. Because $vv^\top = (sv)(sv)^\top$, we write:

$$P_u - P_v = uu^\top - vv^\top = uu^\top - (sv)(sv)^\top = (uu^\top - suv^\top) + (suv^\top - sv(sv)^\top) = u(u-sv)^\top + (u-sv)(sv)^\top$$

Therefore, by the triangle inequality and the fact that $\|xy^\top\| = \|x\|_2\|y\|_2$ (because $xy^\top$ is a rank one matrix), we have:

$$\|P_u - P_v\| \leq \|u(u-sv)^\top\| + \|(u-sv)(sv)^\top\| = \|u\|_2\|u-sv\|_2 + \|u-sv\|_2\|sv\|_2 = 2\|u-sv\|_2$$

Lower bound. By definition of the operator norm, we have:

$$\|P_u - P_v\| \geq \|(P_u - P_v)u\|_2.$$

Moreover, $(P_u - P_v)u = (uu^\top - (sv)(sv)^\top).u = u - (sv)(sv)^\top u = u - \langle sv, u\rangle\, sv = u - \cos\theta\, sv$. Because $sv = \cos\theta\, u + \sin\theta\, w$, we get:

$$u - \cos\theta\, sv = u - \cos\theta(\cos\theta\, u + \sin\theta\, w) = \sin^2\theta\, u - \sin\theta\cos\theta\, w$$

Therefore:

$$\|(P_u - P_v)u\|_2 = \|\sin^2\theta\, u - \sin\theta\cos\theta\, w\|_2 = \sqrt{(\sin^2\theta\, u - \sin\theta\cos\theta\, w)^\top(\sin^2\theta\, u - \sin\theta\cos\theta\, w)}$$

$$= \sqrt{\sin^4\theta\, u^\top u - 2\sin^3\theta\cos\theta\, u^\top w + \sin^2\theta\cos^2\theta\, w^\top w} = \sqrt{\sin^4\theta + \sin^2\theta\cos^2\theta} = \sin\theta.$$

This implies $\|P_u - P_v\| \geq \sin\theta$.

On the other hand, $\|u - sv\|_2^2 = \|u\|_2^2 + \|sv\|_2^2 - 2\langle u, sv\rangle = 2 - 2\cos\theta = 4\sin^2(\theta/2)$, then we have:

$$\|u - sv\|_2 = 2\sin(\theta/2).$$

Because $\theta \in [0, \pi/2]$, we have $\sin\theta \geq \sin(\theta/2)$, and therefore:

$$\|P_u - P_v\| \geq \sin\theta \geq \sin(\theta/2) \geq \frac{1}{2}\|u - sv\|_2.$$

Combine the above inequalities, we conclude: $\frac{1}{2}\|u - sv\|_2 \leq \|P_u - P_v\| \leq 2\|u - sv\|_2$.

a2. Let $A, B \in \mathbb{R}^{d\times d}$ be symmetric and $\delta := \lambda_1(A) - \lambda_2(A) > 0$. Additionally, let $v_1(A)$ and $v_1(B)$ be unit top eigenvectors. We define $P_A := v_1(A)v_1(A)^\top$ and $P_B := v_1(B)v_1(B)^\top$.

Case 1: $\|A - B\| \leq \delta/2$. By the Davis–Kahan theorem (for $k = 1$), we have:

$$\|P_A - P_B\| \leq \frac{2\|A - B\|}{\delta}$$

Apply part (a1) with $u = v_1(A)$, $v = v_1(B)$, there exists $s \in \{-1, 1\}$ such that:

$$\frac{1}{2}\|v_1(A) - sv_1(B)\|_2 \leq \|P_A - P_B\|$$

Therefore: $\|v_1(A) - sv_1(B)\|_2 \leq 2\|P_A - P_B\| \leq 4\frac{\|A-B\|}{\delta}$.

Case 2: $\|A - B\| > \delta/2$. Because $v_1(A)$ and $v_1(B)$ are unit vectors, we can apply the triangular inequality to obtain the upper bound for $\|v_1(A) - sv_1(B)\|_2$ as follows:

$$\|v_1(A) - sv_1(B)\|_2 \leq \|v_1(A)\|_2 + \|v_1(B)\|_2 = 2$$

On the other hand, $\|A - B\| > \delta/2$ implies $\frac{\|A-B\|}{\delta} > \frac{1}{2}$, so $2 \leq 4\frac{\|A-B\|}{\delta}$

Therefore:

$$\|v_1(A) - sv_1(B)\|_2 \leq 2 \leq 4\frac{\|A - B\|}{\delta}$$

Combine two cases, we conclude that there exists a constant $C > 0$ (e.g., $C = 4$) such that:

$$\|v_1(A) - sv_1(B)\|_2 \leq C\frac{\|A - B\|}{\delta}.$$

b. Learning a rank-one spike model

b1. For any $x \in \mathbb{R}^d$ with $\|x\|_2 = 1$, we have:

$$x^\top \Sigma x = x^\top I x + \beta x^\top u u^\top x = 1 + \beta \langle x, u\rangle^2$$

Therefore: $\lambda_1(\Sigma) = \max_{\|x\|_2=1} x^\top \Sigma x = 1 + \beta \max_{\|x\|_2=1} \langle x, u\rangle^2$.

By Cauchy-Swart's inequality, we have: $\langle x, u\rangle^2 \leq ||x||_2^2 ||u||_2^2 = 1$. The equality holds when $x = cu$, which implies $||x||_2 = |c| \cdot ||u||_2$. Because we have $||x||_2 = ||u||_2 = 1$, then $c \in \{1, -1\}$, or equivalently $x = u$ or $x = -u$.

Therefore $\max_{||x||_2=1} \langle x, u\rangle^2 = 1$, which implies that $\lambda_1(\Sigma) = 1 + \beta$. Moreover $v_1(\Sigma) = u$ is one of the unit eigenvector corresponding to $\lambda_1(\Sigma)$.

Compute $\lambda_2(\Sigma)$. By the min–max characterization, we have $\lambda_2(\Sigma) = \min_{\substack{E\subset\mathbb{R}^d \\ \dim(E)=d-1}} \max_{\substack{x\in E \\ \|x\|_2=1}} x^\top \Sigma x$.

$$\lambda_2(\Sigma) = \min_{\substack{E\subset\mathbb{R}^d \\ \dim(E)=d-1}} \max_{\substack{x\in E \\ \|x\|_2=1}} x^\top \Sigma x.$$

Choose the subspace $E = u^\perp$, which satisfies $\dim(E) = d - 1$. For any $x \in E$, we have $\langle x, u\rangle = 0$.

Therefore $x^\top \Sigma x = 1 + \beta\langle x, u\rangle^2 = 1$, which implies that $\max_{\substack{x\in E \\ \|x\|_2=1}} x^\top \Sigma x = 1$ and then $\lambda_2(\Sigma) \le 1$.
On the other hand, because $\Sigma = I_d + \beta uu^\top \succeq I_d$, we have $x^\top\Sigma x \ge 1 \quad$ for all $\|x\|_2 = 1$, so all eigenvalues of Σ are at least 1. Therefore $\lambda_2(\Sigma) \ge 1$.

Combine the two bounds, we conclude: $\lambda_2(\Sigma) = 1$.

b2. Let $v = v_1(\widehat{\Sigma})$. We have the eigengap of Σ $\delta = \lambda_1(\Sigma) - \lambda_2(\Sigma) = \beta$. By the covariance estimation bound, with probability at least $1 - 2e^{-d}$,

$$\|\widehat{\Sigma} - \Sigma\| \le C\Big(\sqrt{\frac{d}{n}} + \frac{d}{n}\Big)$$

By part (a2), there exists $s \in \{-1, 1\}$ such that:

$$\|u - sv\|_2 \le C\frac{\|\widehat{\Sigma} - \Sigma\|}{\beta} \le \frac{C}{\beta}\Big(\sqrt{\frac{d}{n}} + \frac{d}{n}\Big).$$

Moreover, we have: $n \ge C_1\frac{d}{\beta^2}$. Therefore:

$$\|u - sv\|_2 \le \frac{C}{\beta}\Big(\sqrt{\frac{d}{n}} + \frac{d}{n}\Big) \le \frac{C}{\beta}\left(\sqrt{\frac{d}{C_1\frac{d}{\beta^2}}} + \frac{d}{C_1\frac{d}{\beta^2}}\right) = C\left(\frac{1}{\sqrt{C_1}} + \frac{\beta}{C_1}\right)$$

Additionally, we have:

$$1+\beta = \max_{\|x\|=1} x^\top\Sigma x = u^\top\Sigma u = u^\top\mathbb{E}[X_iX_i^\top]u = \mathbb{E}[(X_i^\top u)^2] = \mathbb{E}\langle X_i, u\rangle^2 = \|\langle X_i, u\rangle\|_{L^2}^2 \lesssim \|\langle X_i, u\rangle\|_{\psi_2}^2 \le 100. \tag{1}$$

Indeed, we prove that $|Z\|_{L_2} \le \|Z\|_{\psi_2}$. Let $K := \|Z\|_{\psi_2}$. By definition of the sub-Gaussian norm, we have $\mathbb{E}\exp(Z^2/K^2) \le 2$. Use the inequality $e^x \ge 1 + x$ for all $x \ge 0$, we obtain $\exp(Z^2/K^2) \ge 1 + Z^2/K^2$. Take expectation, we have $\mathbb{E}\exp(Z^2/K^2) \ge 1 + \mathbb{E}[Z^2]/K^2$. Combine the two bounds, we get $2 \ge 1 + \mathbb{E}[Z^2]/K^2$, then $\mathbb{E}[Z^2] \le K^2$. Therefore $\|Z\|_{L_2} = (\mathbb{E}[Z^2])^{1/2} \le K = \|Z\|_{\psi_2}$.

We have 1 implies that $\beta \le 99$. Therefore: $\|u - sv\|_2 \le C\left(\frac{1}{\sqrt{C_1}} + \frac{99}{C_1}\right)$. If we choose sufficiently large C_1 such that $C\left(\frac{1}{\sqrt{C_1}} + \frac{99}{C_1}\right) \le 0.1$, we will obtain $\min_{s\in\{-1,1\}} \|u - sv\|_2 \le \|u - sv\|_2 \le 0.1$.
We then have:

$$\min_{s\in\{-1,1\}} \|u - sv\|_2 \le \|u - sv\|_2 \le \frac{C}{\beta}\Big(\sqrt{\frac{d}{n}} + \frac{d}{n}\Big) \le 0.1$$

Therefore:

$$\mathbb{P}\left(\min_{s\in\{-1,1\}}\|u-sv\|_2 \le 0.1\right) \ge \mathbb{P}\left(\|u-sv\|_2 \le 0.1\right) \ge \mathbb{P}\left(\|\widehat{\Sigma}-\Sigma\| \le C\left(\sqrt{\frac{d}{n}}+\frac{d}{n}\right)\right) \ge 1-2e^{-d}.$$

Question 5. Learning a Gaussian mixture model

a. Covariance and spike structure

a1.

Compute $\mathbb{E}X$. We have:

$$\mathbb{E}[X] = \mathbb{E}[G+\theta tu] = \mathbb{E}[G] + \mathbb{E}[\theta tu] = 0 + tu\mathbb{E}[\theta] = tu\left[\frac{1}{2}(1)+\frac{1}{2}(-1)\right] = 0$$

Compute covariance matrix of X. We have:

$$\begin{aligned}
\Sigma := \mathbb{E}[XX^\top] &= \mathbb{E}[(G+\theta tu)(G+\theta tu)^\top] \\
&= \mathbb{E}[(G+\theta tu)(G^\top+\theta tu^\top)] \\
&= \mathbb{E}[GG^\top + \theta tGu^\top + \theta tuG^\top + \theta^2t^2uu^\top] \\
&= \mathbb{E}[GG^\top] + t\mathbb{E}[\theta]\mathbb{E}[G]u^\top + tu\mathbb{E}[\theta]\mathbb{E}[G^\top] + t^2uu^\top\mathbb{E}[\theta^2] \qquad \text{(because } \theta \text{ and } G \text{ are independent)} \\
&= \operatorname{Cov}(G) + t^2uu^\top \quad (\text{since } \operatorname{Cov}(G)=\mathbb{E}[GG^\top]; \mathbb{E}[\theta]=0; \mathbb{E}[G]=0; \mathbb{E}[G^\top]=(\mathbb{E}[G])^\top=0; \mathbb{E}[\theta^2]=1) \\
&= I_d + t^2uu^\top
\end{aligned}$$

a2. Because the covariance matrix of X, $\Sigma = I_d + t^2uu^\top$ with $u \in S^{d-1}, t^2 > 0$, we apply the result of (4b2) with $\beta = t^2$ to obtain: $\lambda_1(\Sigma) = 1+t^2,\ \lambda_2(\Sigma)=1,\ v_1(\Sigma)=u$.

b. Learning the signal direction from data

b1. We have $\langle X,v\rangle = \langle G,v\rangle + \theta t\langle u,v\rangle$. We can see that $\langle G,v\rangle = v^\top G \sim \mathcal{N}(0, v^\top \operatorname{Var}(G)v) = \mathcal{N}(0,1)$ and $|\theta t\langle u,v\rangle|$ is bounded by $t|\langle u,v\rangle|$. Therefore $\langle X,v\rangle$ is sub-Gaussian. We have:

$$\|\langle X,v\rangle\|_{\psi_2} \le C(1+t|\langle u,v\rangle|).$$

On the other hand:

$$\begin{aligned}\|\langle X, v\rangle\|_{L^2}^2 &= \mathbb{E}\big[\langle X, v\rangle^2\big] \\ &= \mathbb{E}\big[(\langle G, v\rangle + \theta t\langle u, v\rangle)^2\big] \\ &= \mathbb{E}\big[\langle G, v\rangle^2 + 2\theta t\langle G, v\rangle\langle u, v\rangle + \theta^2 t^2\langle u, v\rangle^2\big] \\ &= \mathbb{E}\big[\langle G, v\rangle^2\big] + 2t\langle u, v\rangle\,\mathbb{E}\big[\theta\langle G, v\rangle\big] + t^2\langle u, v\rangle^2\mathbb{E}[\theta^2]\end{aligned}$$

Because $\langle G, v\rangle \sim N(0,1)$, then $\mathbb{E}\big[\langle G, v\rangle^2\big] = 1$. Also, because θ is independent of G, $\mathbb{E}[\theta] = 0$, and $\mathbb{E}[\langle G, v\rangle] = 0$, $\mathbb{E}\big[\theta\langle G, v\rangle\big] = \mathbb{E}[\theta]\,\mathbb{E}[\langle G, v\rangle] = 0$. Finally, $\theta \in \{-1, 1\}$, so $\theta^2 = 1$ almost surely and then $\mathbb{E}[\theta^2] = 1$. Therefore $\|\langle X, v\rangle\|_{L^2}^2 = \mathbb{E}\big[\langle X, v\rangle^2\big] = 1 + t^2\langle u, v\rangle^2$, which implies $\|\langle X, v\rangle\|_{L^2} = \sqrt{1 + t^2\langle u, v\rangle^2}$.

Moreover, we have $1 + t|\langle u, v\rangle| \leq C_1\sqrt{1 + t^2\langle u, v\rangle^2}$ where C_1 is an absolute constant. Therefore $\|\langle X, v\rangle\|_{\psi_2} \leq CC_1\|\langle X, v\rangle\|_{L^2}$.

b2. By covariance estimation, we have:

$$\|\widehat{\Sigma} - \Sigma\| \leq C\Big(\sqrt{\frac{d}{n}} + \frac{d}{n}\Big)\|\Sigma\| = C\Big(\sqrt{\frac{d}{n}} + \frac{d}{n}\Big)\ \lambda_1(\Sigma) = C\Big(\sqrt{\frac{d}{n}} + \frac{d}{n}\Big)\ (1 + t^2)$$

Because $t \geq 0.1$, we have $0.001 \leq t^2$, which implies $1 \leq 100t^2$. Therefore:

$$\|\widehat{\Sigma} - \Sigma\| \leq C\Big(\sqrt{\frac{d}{n}} + \frac{d}{n}\Big)\|\Sigma\| = C\Big(\sqrt{\frac{d}{n}} + \frac{d}{n}\Big)\ \lambda_1(\Sigma) = C\Big(\sqrt{\frac{d}{n}} + \frac{d}{n}\Big)\ 101t^2 = C\Big(\sqrt{\frac{d}{n}} + \frac{d}{n}\Big)\ t^2$$

Apply Davis-Kahan, we obtain:

$$\min_s \|u - sv\|_2 \leq \|u - sv\|_2 \leq C\frac{\|\widehat{\Sigma} - \Sigma\|}{t^2} \leq C\Big(\sqrt{\frac{d}{n}} + \frac{d}{n}\Big)$$

Moreover, we have $n \geq C_1 d$. Therefore:

$$\min_s \|u - sv\|_2 \leq C\Big(\sqrt{\frac{d}{n}} + \frac{d}{n}\Big) \leq C\Big(\sqrt{\frac{d}{C_1 d}} + \frac{d}{C_1 d}\Big) = C\Big(\sqrt{\frac{1}{C_1}} + \frac{1}{C_1}\Big)$$

Therefore: $\mathbb{P}\Big(\min_s \|u - sv\|_2 \leq C\Big(\sqrt{\frac{1}{C_1}} + \frac{1}{C_1}\Big)\Big) \geq \mathbb{P}\Big(\min_s \|u - sv\|_2 \leq C\Big(\sqrt{\frac{d}{n}} + \frac{d}{n}\Big)\Big) \geq 1 - 2e^{-d}$

With sufficiently large C_1 such that $C\Big(\sqrt{\frac{1}{C_1}} + \frac{1}{C_1}\Big)) \leq 0.1$, we can conclude:

$$\mathbb{P}\Big(\min_s \|u - sv\|_2 \leq 0.1) \geq \mathbb{P}\Big(\min_s \|u - sv\|_2 \leq C\Big(\sqrt{\frac{1}{C_1}} + \frac{1}{C_1}\Big)\Big) \geq 1 - 2e^{-d}$$

From the above result, we can see that with $n = \mathcal{O}(d)$ unlabeled samples, one can recover the direction of separation in the Gaussian mixture model up to sign.