Homework IV
Theoretical Statistics/Machine Learning

Sagar Ghosh
Department of Statistics and Data Sciences
The University of Texas at Austin
Email: sg63684@my.utexas.edu

March 27, 2026

1 Entropy method beyond uniform gradient bounds.

Throughout this porblem, we assume $X \in \mathbb{R}^n$ is a random vector satisfying the MLSI

$$\mathrm{Ent}(e^{u(X)}) \leq C\mathbb{E}\left[\|\nabla u(X)\|_2^2 e^{u(X)}\right], \quad \text{for all smooth } u : \mathbb{R}^n \to \mathbb{R},$$

where $C > 0$ is a constant. For a non-negative random variable Y, the concentration entropy $\mathrm{Ent}(Y) = \mathbb{E}[Y \log Y] - (\mathbb{E}Y)\log(\mathbb{E}Y)$. For all real valued random variable W, we define its log-MGF as $\kappa_W(\eta) = \log \mathbb{E}e^{\eta W}$, whenever this is finite. For centered quantities, we write $\psi_Z(\theta) = \log \mathbb{E}e^{\theta Z}$, where $\mathbb{E}Z = 0$. We use the following fact without proof for centered Z,

$$\frac{\mathrm{Ent}(e^{\theta Z})}{\mathbb{E}e^{\theta Z}} = \theta\psi_Z'(\theta) - \psi_Z(\theta) = \theta^2\left(\frac{\psi_Z(\theta)}{\theta}\right)'.$$

In particular, an upper bound on $\frac{\mathrm{Ent}(e^{\theta Z})}{\theta^2 \mathbb{E}e^{\theta Z}}$ can be integrated using Herbst's argument to bound $\psi_Z(\theta)$.

(a) **Young's Inequality for entropy.** We assume $Y \geq 0$ satisfying $\mathbb{E}Y = 1$. Let W be any real valued random variable. Since $Y \geq 0$ and $\mathbb{E}Y = 1$, we can take Y as a density function. Let's construct another density function from W, let $g = \frac{e^W}{\mathbb{E}e^W}$, so $g \geq 0$ and $\mathbb{E}g = 1$, i.e., g is also a density. Now, let's compute the KL divergence between Y and g.

$$0 \leq D_{KL}(Y||g) = \mathbb{E}[Y \log(Y/g)] = \mathbb{E}[Y \log Y - Y \log[e^W/\mathbb{E}e^W]] = \mathbb{E}[Y \log Y] - \mathbb{E}[YW] + \mathbb{E}[Y \log \mathbb{E}e^W]$$
$$= \mathbb{E}[Y \log Y] + \log \mathbb{E}e^W \mathbb{E}[Y] - \mathbb{E}[WY].$$

Therefore, $\mathbb{E}[YW] \leq \log \mathbb{E}e^W + \mathrm{Ent}(Y)$, since $\mathrm{Ent}(Y) = \mathbb{E}[Y \log Y]$ as $\mathbb{E}[Y] = 1$.

(b) **A non-linear Bernstein inequality from MLSI.** Let $f : \mathbb{R}^n \to \mathbb{R}$ be smooth, and define the centered random variable $Z = f(X) - \mathbb{E}f(X)$, $\psi_f(\theta) = \log \mathbb{E}e^{\theta Z}$, $G = \|\nabla f(X)\|_2^2$.

(b_1) Let's consider the function $p(Z) = \theta Z$ so, $\nabla p(Z) = \theta \nabla f(X)$. Also, using the multivariate MLSI we have $\mathrm{Ent}(e^{p(Z)}) \leq C\mathbb{E}[\|\nabla p(Z)\|_2^2 e^{p(Z)}]$, putting $p(Z) = \theta Z$ and $p'(Z) = \theta \nabla f(X)$, we have $\mathrm{Ent}(e^{\theta Z}) \leq C\theta^2 \mathbb{E}[Ge^{\theta Z}]$.

(b_2) Fix $\eta > 0$. Let $Y = \frac{e^{\theta Z}}{\mathbb{E}e^{\theta Z}}$, $W = \eta G$. Then by part (a), we have $\mathbb{E}[WY] = \frac{\mathbb{E}[We^{\theta Z}]}{\mathbb{E}e^{\theta Z}} \leq \log \mathbb{E}e^{\eta G} + \mathrm{Ent}[Y] = \kappa_G(\eta) + \frac{\mathrm{Ent}(e^{\theta Z})}{\mathbb{E}e^{\theta Z}}$. The last equality follows since Ent functional is positive homogeneous and hence $\mathrm{Ent}[e^{\theta Z}/\mathbb{E}e^{\theta Z}] = \mathrm{Ent}[e^{\theta Z}]/\mathbb{E}e^{\theta Z}$. Dividing both sides by η, we get the required inequality, i.e., $\frac{\mathbb{E}[Ge^{\theta Z}]}{\mathbb{E}e^{\theta Z}} \leq \frac{\kappa_G(\eta)}{\eta} + \frac{\mathrm{Ent}(e^{\theta Z})}{\eta \mathbb{E}e^{\theta Z}}$.

(b_3) Plugging in the result $\text{Ent}(e^{\theta Z}) \le C\theta^2 \mathbb{E}[Ge^{\theta Z}]$ from (b_1) in (b_2), we get

$$\begin{aligned}
&\frac{\mathbb{E}[Ge^{\theta Z}]}{\mathbb{E}e^{\theta Z}} \le \frac{\kappa_G(\eta)}{\eta} + \frac{\text{Ent}(e^{\theta Z})}{\eta \mathbb{E}e^{\theta Z}} \le \frac{\kappa_G(\eta)}{\eta} + \frac{C\theta^2 \mathbb{E}[Ge^{\theta Z}]}{\eta \mathbb{E}[e^{\theta Z}]} \\
\implies &\frac{\mathbb{E}[Ge^{\theta Z}]}{\mathbb{E}e^{\theta Z}} \left(1 - \frac{C\theta^2}{\eta}\right) \le \frac{\kappa_G(\eta)}{\eta} \\
\implies &\frac{\text{Ent}(e^{\theta Z})}{\theta^2 \mathbb{E}e^{\theta Z}} \le \frac{1}{\theta^2} C\theta^2 \frac{\mathbb{E}[Ge^{\theta Z}]}{\mathbb{E}e^{\theta Z}} \le C \left(1 - \frac{C\theta^2}{\eta}\right)^{-1} \frac{\kappa_G(\eta)}{\eta} = \frac{C\kappa_G(\eta)}{\eta - C\theta^2}.
\end{aligned}$$

(b_4) We have Herbst's identity : $\frac{\text{Ent}(e^{\theta Z})}{\theta^2 \mathbb{E}e^{\theta Z}} \left(\frac{\psi_Z(\theta)}{\theta}\right)'$, we have from ($b_3$) that $\left(\frac{\psi_Z(s)}{s}\right)' \le \frac{C\kappa_G(\eta)}{\eta - Cs^2}$. Integrating both sides w.r.t. s from 0 to θ, and bounding $\frac{1}{\eta - Cs^2} \le \frac{1}{\eta - C\theta^2}$, we get $\left(\frac{\psi_Z(\theta)}{\theta}\right) \le \frac{C\theta\kappa_G(\eta)}{\eta - C\theta^2}$ or $\psi_f(\theta) \le \frac{C\theta^2\kappa_G(\eta)}{\eta - C\theta^2}$.

(b_5) Note that the RHS of (b_4) does not depend on η for $\eta > C\theta^2$. hence, we can write that $\psi_f(\theta) \le \inf_{\eta : \eta \ge C\theta^2} \frac{C\theta\kappa_G(\eta)}{\eta - C\theta^2}$. Further, taking inf on both sides w.r.t. θ, we get $\inf_\theta \psi_f(\theta) \le \inf_\theta \inf_{\eta : \eta \ge C\theta^2} \frac{C\theta\kappa_G(\eta)}{\eta - C\theta^2} = \inf_{\eta > 0} \inf_{0 \le \theta \le \sqrt{\eta/C}} \frac{C\theta\kappa_G(\eta)}{\eta - C\theta^2}$ Now, using Chernoff method, we already have

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

(c) **Application: self-bounded functions.** We assume that the centered random variable Z is self-bounded in the sense that $G = \|\nabla f(X)\|_2^2 \le aZ + b$ almost surely for some constants $a, b \ge 0$.

(c_1) Suppose $\eta \ge 0$. Then,

$$\begin{aligned}
\kappa_G(\eta) &= \log \mathbb{E}e^{\eta G} \\
&\le \log \mathbb{E}e^{\eta(aZ+b)} \quad \text{[Since } e^{\eta G} \text{ is a non-negative random variable and log is monotonic]} \\
&= \eta b + \log \mathbb{E}e^{a\eta G} = \eta b + \psi_f(a\eta) \quad \text{[By the definition of } \psi\text{]}.
\end{aligned}$$

(c_2) Assume $a > 0$, from part (b_4) we have for $\eta > C\theta^2$, we have $\psi_f(\theta) \le \frac{C\theta^2\kappa_G(\eta)}{\eta - C\theta^2} \le \frac{C\theta^2[\eta b + \psi_f(a\eta)]}{\eta - C\theta^2}$, where the last inequality follows from part (c_1). Putting $\eta = \theta/a$ and changing side, we get

$$\begin{aligned}
\psi_f(\theta)[1 - 2aC\theta] &\le CB\theta^2 \\
\implies \psi_f(\theta) &\le \frac{Cb\theta^2}{1 - 2aC\theta},
\end{aligned}$$

holds true for all $\theta \in [0, 1/(2Ca))$

(c_3) Similar to part (b_5) and using the estimate from part (c_3), we get

$$\begin{aligned}\mathbb{P}[f(X) - \mathbb{E}f(X) \geq t] &\leq \inf_{\theta \in [0, 1/(2Ca))} \exp\left(-\theta t + \frac{Cb\theta^2}{1 - 2aC\theta}\right) \\ &\leq \exp\left(-c \min\left(\frac{t^2}{b}, \frac{t}{a}\right)\right),\end{aligned}$$

where the last bound we got using $\theta = \min\{t/(4Cb), 1/(4Ca)\}$. This implies, $-\theta t + \frac{Cb\theta^2}{1-2aC\theta} \leq -c\min(t^2/(2b), (t/2Ca))$, absorbing the constants, we get the required bounds. As $a \to 0$, the $\min\left\{\frac{t^2}{b}, \frac{t}{a}\right\} = \frac{t^2}{a}$, so we would recover the usual Sub-Gaussian bound, this is also confirmed by the fact that $\psi_f(\theta) \leq Cb\theta^2$, i.e., the $\log -$ CGF is bounded by a quadratic function, which explains this Sub-Gaussian behavior.

(d) **Application: Positive Semi-Definite Quadratic Forms.** Let $A \in \mathbb{R}^{n \times n}$ be positive semi-definite and let $X \in \mathbb{R}^n$ be isotropic: $\mathbb{E}[XX^\top] = I_n$. We define, $f(X) = X^\top AX - \operatorname{tr}(A)$.

(d_1) We have $\mathbb{E}[f(X)] = \mathbb{E}[X^\top AX] - \mathbb{E}[\operatorname{tr}(A)] = \mathbb{E}[\operatorname{tr}(X^\top AX)] - \operatorname{tr}(A) = \mathbb{E}[\operatorname{tr}(AXX^\top)] - \operatorname{tr}(A) = \operatorname{tr}[A\mathbb{E}[XX^\top]] - \operatorname{tr}(A) = \operatorname{tr}[A] - \operatorname{tr}[A] = 0$. We also have $\nabla f(X) = 2AX$ [Since A is symmetric]. Taking the square on both sides, we get $\|\nabla f(X)\|_2^2 = 4\|AX\|_2^2 = 4(X^\top A^\top AX) = 4(X^\top A^2 X) \leq 4\|A\|(X^\top AX) = 4\|A\|(f(X) + \operatorname{tr}(A))$, where the third equality follows from the fact that $A^2 \leq \|A\|A$, since A is positive semi-definite and $X^\top VX$ is a quadratic form. Comparing the last expression with part (c_1), we get that $f(X)$ is self-bounded with $a = 4\|A\|$ and $b = 4\|A\| \operatorname{tr}(A)$.

(d_2) Using parts (C_2) and (c_3), we get the Bernstein type inequality for $f(X)$, i.e.,

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

where $c > 0$ depends only on C, where C came first from the result in part (b_1), where we used MLSI, hence $c > 0$ depends only on the MLSI constant C.

(d_3) For X being a standard Gaussian vector, $\|\nabla f(X)\|_2^2 = 4\|AX\|_2^2 = 4\operatorname{tr}(A^\top A)XX^\top \leq 4\|A\|_F(f(X) + \|A\|_F)$, i.e., for b we will get a sharper bound as $4\|A\|_F^2$ instead of $4\|A\| \operatorname{tr}(A)$.

2 Practice with covering and packing.

Throughout we assume (T, d) is a metric space and $K \subseteq T, \epsilon > 0$. We recall the definitions:

- A set $M \subset K$ is an $\epsilon-$net of K if for every $x \in K$ there exists $y \in M$ such that $d(x, y) < \epsilon$. Equivalently the balls of radius ϵ centered at points of M covers K.
- The covering number of K is $\mathcal{N}(\epsilon, K, d) = \min\{|M| : M \subset K \text{ is an } \epsilon \text{ net of K}\}$.
- A set $M \subset K$ is ϵ separated if $d(x, y) > \epsilon$ for all distinct $x, y \in M$. The packing number of K is $\mathcal{P}(\epsilon, K, d) = \max\{|M| : M \subset K \text{ is } \epsilon \text{ separated}\}$.
- The metric entropy is the logarithm of the covering number: $\mathcal{H}(\epsilon, K, d) = \log \mathcal{N}(\epsilon, K, d)$.

(a) **Monotonicity properties.** The first goal is to get comfortable with the definitions and with a subtle point: covering numbers are monotone in the scale ϵ, but not monotone in the set K.

(a_1) Take $\epsilon_1 \leq \epsilon_2$. Then for any two points $x, y \in K$, $d(x, y) < \epsilon_1 \implies d(x, y) < \epsilon_2$, therefore any ϵ_1 net is automatically an ϵ_2 net, this means to cover up the space K we need fewer balls as ϵ gets bigger. Hence, $\epsilon \to \mathcal{N}(\epsilon, K, d)$ is decreasing. For the packing number, $d(x, y) > \epsilon_2 \implies d(x, y) > \epsilon_1$. Therefore, if two points in K are ϵ_2 distance apart, then they are automatically ϵ_1 distance apart. That is any ϵ_2 packing set is already an ϵ_1 packing set in K. This means, we will have fewer no of points in K to pack the space which are at least ϵ distance apart as ϵ gets bigger. As a result, $\epsilon \to \mathcal{P}(\epsilon, K, d)$ is decreasing.

(a_2) Consider $K = [0, 1]$ and $L = \{0, 1\}$, so $L \subsetneq K$. We note that, to cover K with a ball of radius $\epsilon = 1/2$, it is sufficient to place a ball at $1/2$, so that $B(1/2, 1/2)$ covers K, i.e., $\mathcal{N}(1/2, K, d) = 1$. But to cover L, we need to place two balls of the same radius centered at 0 and centered at 1, i.e., $L \subseteq B(0, 1/2) \cup B(1, 1/2)$, but no balls of radius $1/2$ can cover up L when placed at either 0 or 1, i.e., $\mathcal{N}(1/2, L, d) = 2$, i.e., $\mathcal{N}(1/2, L, d) > \mathcal{N}(1/2, K, d)$.

(a_3) Take an $\epsilon/2$ cover of K, say $\{x_1, ..., x_n\}$, i.e., for all $x \in K$, $\exists\ x_i \in \{x_1, ..., x_n\}$ such that $d(x, x_i) < \epsilon/2$. Now, $L \subset K$, so, $L \subset K \subseteq \cup_{i=1}^{n} B(x_i, \epsilon/2)$. Whenever the balls $B(x_i, \epsilon/2)$ intersects L, choose a representative point y_i for all $i \in \{1, 2, ..., n\}$. So, we would have a set with maximum cardinality n from L, such that for any point $l \in L$, $\exists y_j \in \{y_1, ..., y_n\}$ with $d(y_j, l) < d(y_j, x_j) + d(x_j, l) < \epsilon/2 + \epsilon/2$. So, $\{y_1, ..., y_n\}$ is an ϵ cover of L, this implies $\mathcal{N}(\epsilon, L, d) \leq \mathcal{N}(\epsilon, K, d)$.

(b) **Packing vs. Covering.** This part develops basic equivalence between covering and packing.

(b_1) Suppose M is not an ϵ net of K, i.e., $\exists x \in K$ such that there is no $y \in M$ such that $d(x,y) < \epsilon$, i.e., for all $y \in M$, $d(x,y) > \epsilon$. So, the ball $B(x,\epsilon)$ does not contain any point from M, so, $M \cup \{x\}$ is a new ϵ separated subset of K, which contradicts the maximality of M. Hence M has to be an $\epsilon-$ net.

(b_2) From part (b_1) we can see that any ϵ packing set is automatically an ϵ net of K, i.e., the minimal cardinality required to cover up K by balls of radius ϵ can not exceed the cardinality of any arbitrary ϵ packing set, i.e., $\mathcal{N}(\epsilon, K, d) \leq \mathcal{P}(\epsilon, K, d)$. For the left side inequality, we take an 2ϵ packing set of K, let $\{x_1, ..., x_n\}$. Then for any $x \in K$, x can intersect at most one of the balls $\{B(x_i, \epsilon)\}_{i=1}^n$, otherwise if x intersects two distinct balls $B(x_i,\epsilon)$ and $B(x_j,\epsilon)$, then $d(x_i,x_j) \leq d(x_i,x)+d(x_j,x) < 2\epsilon$, contradicting the assumption that $\{x_1, ..., x_n\}$ is a 2ϵ packing set of K. Therefore, any ball $B(x_i,\epsilon)$ can intersect at most one point from any ϵ net of K, in other words, $\mathcal{P}(2\epsilon, K, d) \leq \mathcal{N}(2\epsilon, K, d)$.

(c) **Volumetric bounds in Euclidean space.** For this part, we take $T = \mathbb{R}^n$ with $d(x,y) = \|x-y\|_2$, the Euclidean metric. We write $B_2^n = \{x \in \mathbb{R}^n \mid \|x\|_2 = 1\}$. For two sets $A, B \subseteq \mathbb{R}^n$, one defines the Minkowsi Sum as $A + B = \{a + b, a \in A, b \in B\}$.

(c_1) Let $\{x_1, ..., x_r\}$ be an ϵ cover of K. Then $K \subseteq \cup_{i=1}^r B(x_i, \epsilon)$. Note that, using Minkowski sum notation, we can also write $B(x_i,\epsilon) = x_i + \epsilon^n B_2^n$. Using the volumetric bound, we get $\text{Vol}(K) \leq \text{Vol}(\sup_{i=1}^r \{x_i + \epsilon^n B_2^n\}) = r\epsilon^n \text{Vol}(B_2^n)$. Changing side, we get $\frac{\text{Vol}(K)}{\text{Vol}(\epsilon B_2^n)} \leq r = \mathcal{N}(\epsilon, K, d)$. For the other part, we will use the result from (b_2), i.e., $\mathcal{P}(\epsilon, K, d) \leq \mathcal{N}(\epsilon/2, K, d)$. We also note that, the set $K + (\epsilon/2)^n B_2^n$ contains the sets $\cup_{i=1}^m B(z_i, \epsilon/2)$, where $m = |\mathcal{N}(\epsilon/2, K, d)|$ and $\{z_1, ..., z_m\}$ be the $\epsilon/2$ covering of K. Using a similar volumetric argument, we see that $m(\epsilon/2)^n \text{Vol}(B_2^n) \leq \text{Vol}(K + (\epsilon/2)^n B_2^n)$. Hence, $\mathcal{P}(\epsilon, K, d) \leq \mathcal{N}(\epsilon/2, K, d) = m \leq \frac{\text{Vol}(K+(\epsilon/2)^n B_2^n)}{\text{Vol}((\epsilon/2)B_2^n)}$.

(c_2) Choosing $K = B_2^n$ in part (c_1), we have $\frac{\text{Vol}(K)}{\text{Vol}(\epsilon B_2^n)} = \frac{1}{\epsilon^n}$. Hence, using the left hand side from part (c_1), we get $\frac{1}{\epsilon^n} \leq \mathcal{N}(\epsilon, K, d)$. From the second part, we have

$$\mathcal{P}(\epsilon, K, d) \leq \frac{\text{Vol}(K + (\epsilon/2)^n B_2^n)}{\text{Vol}((\epsilon/2)B_2^n)} = \frac{(1+\epsilon/2)^n}{(\epsilon/2)^n} = \left(1 + \frac{2}{\epsilon}\right)^n.$$

Combining both, we get

$$\left(\frac{1}{\epsilon}\right)^n \leq \mathcal{N}(\epsilon, K, d) \leq \left(1 + \frac{2}{\epsilon}\right)^n.$$

As $\epsilon < 1$ in general,

$$\left(\frac{1}{\epsilon}\right)^n \leq \mathcal{N}(\epsilon, K, d) \leq \left(\frac{3}{\epsilon}\right)^n.$$

(c_3) Take $\{y_1, ..., y_m\}$ as an ϵ packing set on the unit sphere S^{n-1}. Then the union of the balls $B(y_i, \epsilon/2)$ is contained in $B_2^n + (\epsilon/2)B_2^n$. Using a similar volumetric argument, we see that $\text{Vol}(S^{n-1}) \leq \text{Vol}(\cup_{i=1}^m B(y_i, \epsilon/2)) = m(\epsilon/2)^n \text{Vol}(B_2^n) \leq (1 + (\epsilon/2))^n \text{Vol}(B_2^n)$. This imples $\mathcal{N}(\epsilon, S^{n-1}, d) \leq \mathcal{P}(\epsilon, S^{n-1}, d) = m \leq \left(1 + \frac{2}{\epsilon}\right)^n$. Hence,

$$\left(\frac{1}{\epsilon}\right)^n \leq \mathcal{N}(\epsilon, S^{n-1}, d) \leq \left(1 + \frac{2}{\epsilon}\right)^n.$$

(c_4) Taking log in part (c_2), we get $n \log(1/\epsilon) \leq \mathcal{H}(\epsilon, B_2^n, d) = \log \mathcal{N}(\epsilon, B_2^n, d) \leq n \log(3/\epsilon)$. That is, up to a constant, we can write $\mathcal{H}(\epsilon, B_2^n, d) = \log \mathcal{N}(\epsilon, B_2^n, d) \asymp n \log(e/\epsilon)$.

3 Covariance estimation with sub-Gaussian random vectors.

Throughout, we assume X denotes a random vector in $\mathbb{R}^d$. We say that X is a sub-Gaussian random vector if every one-dimensional marginal $\langle X, u\rangle$ is a sub-Gaussian random variable. Its sub-Gaussian norm is defined by $\|X\|_{\psi_2} = \sup_{u\in\mathbb{S}^{d-1}} \|\langle X, u\rangle\|_{\psi_2}$. We also call X to be isotropic if $\mathbb{E}X = 0$ and $\mathbb{E}[XX^\top] = I_d$. We also assume $c, C > 0$ denote absolute constants whose values may differ from lines to lines.

(a) **Coordinates and dependence.** This part introduces the basic geometry of the sub-Gaussian norm for random vectors.

(a_1) Let $X_1, ...X_d$ be independent, mean-zero sub-Gaussian random variables and let $X = (X_1, ..., X_d) \in \mathbb{R}^d$. Take any $u \in \mathbb{S}^{d-1}$. Then, $\|\langle X, u\rangle\|_{\psi_2} = \|\sum_{i=1}^d u_i X_i\|_{\psi_2} \leq \max_{1\leq i\leq d} \|X_i\|_{\psi_2} \|\sum_{i=1}^d u_i\|_{\psi_2}$. But for a random vector $u \in \mathbb{S}^{d-1}$, $\|\sum_{i=1}^d u_i\|_{\psi_2} < C$ for some absolute constant $C > 0$ [Since $\|u\|_2 = 1$ and hence $\mathbb{E}[\exp(\|u\|^2/K^2)] = \mathbb{E}[\exp(1/K^2) < \infty]$ for some finite K]. Taking sup over all $u \in \mathbb{S}^{d-1}$ on both sides, we get $\|X\|_{\psi_2} = \sup_{u\in\mathbb{S}^{d-1}} \|\langle X, u\rangle\|_{\psi_2} \leq C \max_{1\leq i\leq d} \|X_i\|_{\psi_2}$. This proves that X is also sub-Gaussian, since all the X_i s are sub-gaussian. We note that, the independence assumption here is crucial since the $\log -$ CGF splits only under the independence assumption, where we can apply component wise quadratic $\log -$ CGF bounds.

To prove the left hand side of the inequality, we assume wlog that $\max_{1\leq i\leq d} \|X_i\|_{\psi_2} = \|X_1\|_{\psi_2}$. Consider, $v = \{1, 0, ..., 0\} \in \mathbb{R}^d$, the standard basis vector in the direction of the first coordinate. Then, $\|X\|_{\psi_2} = \sup_{u\in\mathbb{S}^{d-1}} \|\langle X, u\rangle\|_{psi_2} \geq \|\langle X, v\rangle\|_{psi_2} = \|X_1\|_{\psi_2} = \max_{1\leq i\leq d} \|X_i\|_{\psi_2}$.

(a_2) Let's consider this random vector $X = (X_1, X_1, ..., X_1) \in \mathbb{R}^d$. Take $v = \{1/\sqrt{d}, ..., 1/\sqrt{d}\}$. Then, $\|X\|_{\psi_2} = \sup_{u\in\mathbb{S}^{d-1}} \|\langle X, u\rangle\|_{\psi_2} \geq \|\langle X, v\rangle\|_{\psi_2} = \sqrt{d}\|X_1\|_{\psi_2} = \sqrt{d}\max_{1\leq i\leq d} \|X_i\|_{\psi_2}$. So, the RHS from part (a_1) does not hold.

(a_3) *Cannonical Examples.*

(a_{3_1}) Suppose $X \sim \text{Unif}[-1, 1]^d$. Then, $\mathbb{E}e^{\lambda X} = \int_{[-1,1]^d} e^{\lambda x} \frac{1}{2^d} dx = \left(\frac{\sinh(\lambda)}{\lambda}\right)^d$. Now, by elementary computation, we can show that, $\frac{\sinh\lambda}{\lambda} \leq \exp(\lambda^2/6)$ [By expanding the series on both sides, it is easy to see this inequality holds true]. Therefore, $\mathbb{E}e^{\lambda X} \leq \exp(d\lambda^2/6)$. This implies the $\log -$ CGF of X has a quadratic bound, i.e., $\log \mathbb{E}e^{\lambda X} \leq \frac{d\lambda^2}{6}$. Therefore, the uniform distribution on $[-1, 1]^d$ is sub-Gaussian and hence $\|X\|_{\psi_2} \leq C$. For the uniform Boolean Cube $\{-1, 1\}^d$, we have $\mathbb{E}e^{\lambda X} = \prod_{i=1}^d \frac{1}{2}[\exp(\lambda) + \exp(-\lambda)] =$

$(\cosh\lambda)^d$. Taking the log on both sides and using the elementary inequality $\cosh\lambda \le \lambda^2/2$, we get that for $Y \sim \text{unif}\{-1,+1\}^d$, $\log \mathbb{E}e^{\lambda Y} \le \frac{d\lambda^2}{2}$. Hence, Y is also sub-Gaussian, in other words, $\|Y\|_{\psi_2} \le C$.

(a_{3_2}) Using the fact that $\|X\|_{\psi_2} = \sup_{u\in\mathbb{S}^{d-1}} \|\langle X,u\rangle\|_{\psi_2}$. But $\langle X,u\rangle \sim \mathcal{N}(0, u^\top\Sigma u)$. Now, for a scalar normal variable $Y \sim \mathcal{N}(\mu,\sigma^2)$, we already have $\|Y\|_{\psi_2} \le c\sigma$. Hence, $\|\langle X,u\rangle\|_{\psi_2} \le c\sqrt{\|u^\top\Sigma u\|} \le c\sqrt{\|u^\top\|\|\Sigma\|\|u\|} = c\sqrt{\|u^\top\|\|u\|}\sqrt{\|\Sigma\|}$. Taking the norms over $\mathbb{S}^{d-1}$, we have $\|u^\top\| = \|u\| = 1$. Hence, $\|X\|_{\psi_2} \le c\sqrt{\|\Sigma\|}$.

(b) **Covariance Estimation for Sub-Gaussian data.** Let $X_1,...,X_N$ be i.i.d. copies of a random vector $X \in \mathbb{R}^d$, we define the sample covariance matrix $\hat{\Sigma} = \frac{1}{N}\sum_{i=1}^N X_iX_i^\top$. In this part, we will use the following two facts from earlier lectures/homeworks:

- If W is sub-Gaussian, then $W^2 - \mathbb{E}W^2$ is sub-Exponential and $\|W^2 - \mathbb{E}W^2\|_{\psi_1} \le C\|W\|_{\psi_2}^2$.
- If $\mathcal{N}$ is a 1/4 net of $\mathbb{S}^{d-1}$, then $|\mathcal{N}| \le 12^d$, and for every symmetric matrix A, $\|A\| \le 2\max_{u\in\mathcal{N}}|u^\top Au|$.

(b_1) We assume that X is isotropic and $\|X\|_{\psi_2} \le K$. Fix $u \in \mathbb{S}^{d-1}$ and define $Z_i(u) = \langle X,u\rangle^2 - 1$. Since X_is are i.i.d.s, then obviously Z_i s are so, since each Z_i is only a function of X_is We also, have that $\mathbb{E}[\langle X_i,u\rangle^2] = \mathbb{E}[(u^\top X_iX_i^\top u)] = \mathbb{E}[\text{tr}(u^\top X_iX_i^\top u)] = \mathbb{E}[\text{tr}(X_iX_i^\top uu^\top)] = \text{tr}[\mathbb{E}[X_iX_i^\top]uu^\top] = \text{tr}[I_duu^\top] = \text{tr}[uu^\top] = 1$, since $uu^\top$ is a rank 1 projection operator. Hence, $\mathbb{E}[Z_i(u)] = 0$. Now, X_is are sub-Gaussian, so are $\langle X_i,u\rangle$ by the law of linear transformation of normal variables. But the variance proxy for $\langle X_i,u\rangle$ does not change since u is a unit vector [it still remains as K]. Then, $\langle X_i,u\rangle^2 - \mathbb{E}[\langle X_i,u\rangle^2] = Z_i(u)$ is a sub-exponential random variable [using the stated fact before]. And, $\|Z_i(u)\|_{\psi_1} = \|\langle X_i,u\rangle^2 - 1\|_{\psi_1} \le C\|\langle X_i,u\rangle\|_{\psi_2}^2 = CK^2$. To prove the second part, we note that $\sum_{i=1}^N Z_i(u) = \sum_{i=1}^N[\langle X_i,u\rangle^1 - 1] = \sum_{i=1}^N[u^\top X_iX_i^\top u - 1] = u^\top\sum_{i=1}^N[X_iX_i^\top - I_d]u = Nu^\top(\hat{\Sigma} - I_d)u$. Now we know that $Z_i(u)$s are mean zero i.i.d. sun-exponential random variable with $\|Z_i(u)\|_{\psi_1} \le CK^2$, hence, $\sum_{i=1}^N \|Z_i(u)\|_{\psi_1}^2 \le NK^4$. Using Bernstein's inequality, we get

$$\begin{aligned}
&\mathbb{P}\left[\left|\sum_{i=1}^N Z_i(u)\right| \ge t\right] \le 2\exp\left(-c\min\left\{\frac{t^2}{NK^4}, \frac{t}{K^2}\right\}\right)\\
\implies &\mathbb{P}\left[N\left|u^\top(\hat{\Sigma} - I_d)u\right| \ge t\right] \le 2\exp\left(-c\min\left\{\frac{t^2}{NK^4}, \frac{t}{K^2}\right\}\right)\\
\implies &\mathbb{P}\left[\left|u^\top(\hat{\Sigma} - I_d)u\right| \ge \frac{t}{N}\right] \le 2\exp\left(-c\min\left\{\frac{t^2}{NK^4}, \frac{t}{K^2}\right\}\right)\\
\implies &\mathbb{P}\left[\left|u^\top(\hat{\Sigma} - I_d)u\right| \ge t\right] \le 2\exp\left(-cN\min\left\{\frac{t^2}{K^4}, \frac{t}{K^2}\right\}\right) \quad \text{[replacing } t \text{ by } Nt \text{ in the last inequality]}
\end{aligned}$$

(b_2) Let $\mathcal{N}$ be a 1/4 net of $\mathbb{S}^{d-1}$. Then, $\|\hat{\Sigma} - I_d\| = \sup_{u\in\mathbb{S}^{d-1}} \left|u^\top(\hat{\Sigma} - I_d)u\right| \leq |\mathcal{N}| \max_{u\in\mathcal{N}} \left|u^\top(\hat{\Sigma} - I_d)u\right| = 12^d \max_{u\in\mathcal{N}} \left|u^\top(\hat{\Sigma} - I_d)u\right|$. Now, applying part ($b_1$), we get

$$\mathbb{P}[\|\hat{\Sigma} - I_d \geq t\|] \leq \mathbb{P}\left(\max_{u\in\mathcal{N}} \left|u^\top(\hat{\Sigma} - I_d)u\right| \geq \frac{t}{12^d}\right) \leq 2\exp\left[-cN\min\left\{\frac{t^2}{12^{2d}K^4}, \frac{t}{12^dK^4}\right\}\right]$$

By taking $t = c'K^2\sqrt{d/N}$, we get

$$\mathbb{P}\left[\|\hat{\Sigma} - I_d\| \geq c'K^2\sqrt{\frac{d}{N}}\right] \leq 2\exp\left[-cc'N\frac{K^4t^2}{K^412^{2d}}\right] \approx 2\exp(-d),$$

where for large constant c', we can absorb the constant 12^d in it.

(b_3) We define a random variable $T = \Sigma^{-1/2}X$. Then T is isotropic, and by part (b_2), we have

$$\mathbb{P}\left[\|\hat{\Sigma}_T - I_d\| \geq c'K^2\sqrt{\frac{d}{N}}\right] \lesssim 2\exp(-d),$$

where $\hat{\Sigma}_T = \frac{1}{N}\sum_{i=1}^N T_iT_i^\top$, where T_is are i.i.d. copies of T. Now, if we replace T by X, then $\hat{\Sigma} = \Sigma\hat{\Sigma}_T$. Plugging this in the last inequality, we get

$$\mathbb{P}\left[\|\hat{\Sigma} - \Sigma\| \geq c'\|\Sigma\|K^2\sqrt{\frac{d}{N}}\right] \lesssim 2\exp(-d).$$

Now for $X \sim \mathcal{N}(0, \Sigma)$, the variance proxy becomes $\|\Sigma\|$, therefore, we can replace $k^2\|\Sigma\|$ by $\|\Sigma\|$ in the last inequality followed by the notation $\|\langle X, u\rangle\|_{\psi_2} \leq K\left(\mathbb{E}\langle X, u\rangle^2\right)^{1/2}$. Hence, we get

$$\mathbb{P}\left[\|\hat{\Sigma} - \Sigma\| \geq c'\|\Sigma\|\sqrt{\frac{d}{N}}\right] \lesssim 2\exp(-d)$$

for $N \geq d$.

4 Learning a spike model.

This exercise studies the simplest structured covariance model, known as the spike model: $\Sigma = I_d + \beta uu^\top$, where $u \in \mathbb{S}^{d-1}$ is an unknown signal direction and $\beta > 0$ is the signal to noise ratio(SNR). The leading eigenvector of Σ is exactly u, so if we can choose that sample covariance $\hat{\Sigma}$ close to Σ in operator norm, then matrix perturbation theory should imply that the top eignevector of $\hat{\Sigma}$ is close to u [up to sign]. Throughout $c, C > 0$ denote universal constants whose values may differ from lines to lines.

(a) **Projection Matrices and Perturbation of top Eigenvectors.** For a unit vector $u \in \mathbb{R}^d$, we write $P_u = uu^\top$, the orthogonal projection onto the line spanned by u.

(a_1) Let $u, v \in \mathbb{S}^{d-1}$. Then the operator $P_u - P_v = uu^\top - vv^\top$ is a rank 2 [at most rank 2, but if it has rank 1, then essentially u and v are constant multiples of each other, and there's nothing left to prove] operator. The trace of $P_u - P_v$ is 0 and since it has rank 2, it has two distinct eigenvalues which are of opposite sign but of same magnitude. We can explicitly compute the eigenvalues from the Frobenius norm of the operator $P_u - P_v$. Note that, if two eigenvalues are λ and $-\lambda$, then, $2\lambda^2 = \text{tr}[(uu^\top - vv^\top)^\top(uu^\top - vv^\top)] = 2 - 2r$, where $r = \langle v, u\rangle$ [where we note that tr of a rank l operator is l]. Therefore, $\lambda = \sqrt{1-r} = \|P_u - P_v\|$. Now, $\|u - sv\|_2^2 = (u-sv)^\top(u-sv) = 2 - 2|r|$ [since $\text{sign}(r)r = |r|$]. Therefore, to prove the LHS we need to show that $\frac{1}{2}\sqrt{2-2|r|} \le \sqrt{1-r}$, which is reduced to showing that $4r - 2|r| \le 2$, which is obviously true, since $|r| \ge r \implies 4r - 2|r| \le 2r \le 2$, as $r \le |r| = |\langle v, u\rangle| \le \|u\|_2\|v\|_2 = 1$ [By Cauchy-Schwarz inequality]. Therefore, we have proved the LHS of the inequality: $\frac{1}{2}\|u - sv\|_2 \le \|P_u - P_v\|$.

To prove the RHS, we simply write $P_u - P_v = uu^\top - vv^\top = u(u-sv)^\top + (u-sv)(sv)^\top$. Using triangle inequality and the fact that $\|u\|_2 = \|sv\|_2 = 1$, we get that $\|P_u - P_v\| \le \|u\|_2\|u - sv\|_2 + \|u - sv\|_2\|sv\|_2 = 2\|u - sv\|_2$.

(a_2) $A, B \in \mathbb{R}^{d\times d}$ are symmetric matrices with $\delta = \lambda_1(A) - \lambda_2(A) > 0$. Let P_A be the orthogonal projection on the column space of $\{v_1(A)\}$ and P_B be the orthogonal projection on the column space of $\{v_1(B)\}$. Therefore, $P_A = v_1(A)v_1(A)^\top$ and $P_B = v_1(B)v_1(B)^\top$. For a $\delta > 0$ gap at the top 1 eigenspace of A, using the Davis-Kahan result, we get $\|P_A - P_B\| \le \frac{2\|A-B\|}{\delta}$. Using part ($a_1$) we conclude that $\|v_1(A) - sv_1(B)\|_2 \le 2\|P_A - P_B\| \le 4\frac{\|A-B\|}{\delta}$.

(b) **Learning a rank-one spike model.** Let $u \in \mathbb{S}^{d-1}$ and $\beta > 0$. We consider the covariance matrix $\Sigma - I_d + \beta uu^\top$. Let $X_1, ..., X_n$ be i.i.d. mean 0 random vectors in $\mathbb{R}^d$ with covariance matrix Σ, we also define the sample

covariance $\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n} X_i X_i^\top$. Here we will us the following covariance estimation part from the previous exercise: if $\|X_i\|_{\psi_2} \leq K$, then

$$\mathbb{P}\left\{\|\hat{\Sigma} - \Sigma\| \geq CK^2\left(\sqrt{\frac{d}{n}} + \frac{d}{n}\right)\right\} \leq 2e^{-d}.$$

(b_1) Using triangle inequality of operator norm, we note that $\|\Sigma\| = \|I_d + \beta uu^\top\| \leq \|I_d\| + |\beta|\|uu^\top\| = 1 + \beta$ [Since the operator norm of $\|uu^\top\| = 1$, and is acheived at the vector u itself, and $\beta > 0$]. Therefore, the maximum eigenvalue of Σ can be $1+\beta$. Operating Σ on u, we get $\Sigma u = (I_d + \beta uu^\top)u = u + \beta u = (1+\beta)u$, since $u^\top u = \|u\|_2^2 = 1$. Therefore, the operator norm of Σ is achieved at $v_1(\Sigma) = u$, and the largest eignevalue of Σ is $1 + \beta$, i.e., $\lambda_1(\Sigma) = 1 + \beta$. To get the second largest eigenvalue of Σ, we can project Σ onto the space orthogonal to the column space spanned by u and then take the largest eigenvalue of that projection. More explicitly, $\lambda_2(\Sigma) = \lambda_1(\Sigma - (1+\beta)uu^\top) = \lambda_1(I_d - uu^\top) = 1$, where the maximum eigenvalue of $I_d - uu^\top$ is 1 and the value is attained at any vector $w \perp u$. Therefore, $\lambda_2(\Sigma) = 1$.

(b_2) In addition we assume that X_is are sub-Gaussian and satisfy $\|X_i\|_{\psi_2} \leq 10$. Let $v = v_1(\hat{\Sigma})$ be a unit top eigenvector of the sample covariance matrix. From part (a_2), we can write $\|u - sv\|_2 = \|v_1(\Sigma) - sv_1(\hat{\Sigma})\|_2 \leq \frac{C}{\beta}\|\Sigma - \hat{\Sigma}\|$ [As the eigengap at the top eigenvector is $1 + \beta - 1 = \beta$]. We also note that, $1 + \beta = u^\top \Sigma u = \mathbb{E}\langle X_i, u\rangle^2 = \|\langle X_i, u\rangle\|_{L_2}^2 \lesssim \|\langle X_i, u\rangle\|_{\psi_2}^2 \leq 100$. Hence, $\beta \lesssim 100$. Using the concentration bound on the operator norm of the difference between Σ and $\hat{\Sigma}$, we can write for $n \geq C'd/\beta^2$ [here we assume that C' is large], with probability at least $1 - 2e^{-d}$

$$\|u - sv\|_2 \leq \frac{C}{\beta}\|\Sigma - \hat{\Sigma}\| \leq C^2 * 100 * \frac{1}{\beta}\left(\sqrt{\frac{d}{n}} + \frac{d}{n}\right) \leq C^2 * 100 * \frac{1}{\beta}\left(\frac{\beta}{\sqrt{C'}} + \frac{\beta}{C'}\right) \leq 0.1,$$

by absorbing the constant within C', we can achieve this accuracy. Since this is true for any arbitrary $s \in \{\pm 1\}$, this result holds for, in particular, $\min_{s\in\{\pm1\}} \|u - sv\|_2$.

5 Learning a Gaussian mixture model.

One of the simplest models of structured high-dimensional data is a Gaussian mixture model. In the two-cluster version, one observes points drawn from one of two Gaussian distributions with different means. A basic example is $X = G + \theta tu$, where $u \in \mathbb{S}^{d-1}$ is a fixed unit vector, $t > 0$ controls the separation between two clusters. $G \sim \mathcal{N}(0, I_d)$, and $\theta \in \{\pm 1\}$ is a Rademacher random variable independent of G. Equivalently X is drawn from $\mathcal{N}(tu, I_d)$ or $\mathcal{N}(-tu, I_d)$ with probability 1/2 each. Thus the clusters are centered at $\pm tu$, and the direction u is the signal we wish to learn from the data. Since the model is symmetric under $u \to -u$, the best we can hope for is recovery of u upto sign.

Let $X_1, ..., X_n$ be i.i.d. copies of X, and define the sample covariance matrix $\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n} X_i X_i^\top$. In this problem, we will use the results derived in problem 3 and problem 4:

- Covariance estimation for sub-Gaussian data: If $Y \in \mathbb{R}^d$ is mean zero with covariance matrix Σ_Y and if $\|\langle Y, v\rangle\|_{\psi_2} \le K\|\langle Y, v\rangle\|_{L^2}$ for all $v \in \mathbb{S}^{d-1}$, then

$$\mathbb{P}\left\{\|\hat{\Sigma} - \Sigma_Y\| \ge CK^2\left(\sqrt{\frac{d}{n}} + \frac{d}{n}\right)\|\Sigma_Y\|\right\} \le 2e^{-d}. \tag{1}$$

- Davis-Kahan for top eigenvectors: If A, B are symmetric and $\lambda_1(A) - \lambda_2(A) = \delta > 0$, then

$$\min_{s\in\{\pm 1\}} \|v_1(A) - sv_1(B)\|_2 \le C\frac{\|A - B\|}{\delta}. \tag{2}$$

Throughout we will assume that $\|u\|_2 = 1$ and $t > 0.1$.

(a) **Covariance and spike structure.** This part shows that the gaussian mixture model has a rank-one spiked covariance structure.

(a_1) Conditioning on θ, we can see that $X \mid \theta \sim \mathcal{N}(\theta tu, I_d)$. Therefore, using the chain rule of expectation, we get $\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X \mid \theta]] = \mathbb{E}[\theta tu] = tu\mathbb{E}[\theta] = 0$, since $\mathbb{E}[\theta] = 1 \cdot \frac{1}{2} - 1 \cdot \frac{1}{2} = 0$. We also compute the covariance matrix of X using the law of total variance.

$$\begin{aligned}
\mathrm{Var}[X] = \mathbb{E}[XX^\top] = &= \mathbb{E}[\mathrm{Var}[X \mid \theta]] + \mathrm{Var}[\mathbb{E}[X \mid \theta]] \\
&= \mathbb{E}[I_d] + \mathrm{Var}[\theta tu] \\
&= I_d + t^2 u\mathrm{Var}[\theta]u^\top \\
&= I_d + t^2 uu^\top \quad [\text{Since } \mathrm{Var}[\theta] = \mathbb{E}[\theta^2] = 1 \cdot \frac{1}{2} + 1 \cdot \frac{1}{2} = 1].
\end{aligned}$$

(a_2) This is exactly similar to Part (b_1), Problem 4 [(b_1)]. So, we will not do it again here, just for the completion, stating the results here: $\lambda_1(\Sigma) = 1 + t^2, \lambda_2(\Sigma) = 1$ and $v_1(\Sigma) = u$.

(b) **learning the signal direction from the data.** We now show that the top eigenvector of the sample covariance recovers the cluster-separation direction.

(b_1) Let $v \in \mathbb{S}^{d-1}$. We note that $\langle X, v\rangle = \langle G, v\rangle + \theta t\langle u, v\rangle$. Using the triangle inequality involving the ψ_2 norm, we get $\|\langle X, v\rangle\|_{\psi_2} \leq \|\langle G, v\rangle\|_{\psi_2} + \|\theta t\langle u, v\rangle\|_{\psi_2}$. Since v is a unit vector and $G \sim \mathcal{N}(0, I_d)$, by the law of transformation of normal variable, we have $\langle G, v\rangle \sim \mathcal{N}(0, 1)$, so $\|\langle G, v\rangle\|_{\psi_2} \lesssim 1$. And $\|\theta t\langle u, v\rangle\|_{\psi_2} \lesssim \mathrm{Var}[\theta] t|\langle u, v\rangle| = t|\langle u, v\rangle|$. If we compute the L^2 norm of $\langle X, v\rangle$, using the triangle inequality, we see that $\langle X, v\rangle_{L_2} = \left[\mathbb{E}[\langle X, v\rangle]^2\right]^{1/2} = \left[\mathbb{E}[\langle G, v\rangle]^2 + t^2\langle u, v\rangle^2\mathbb{E}[\theta^2]\right]^{1/2} = \sqrt{1 + t^2\langle u, v\rangle^2}$. Choosing $K = \sqrt{2}$, we note that

$$K^2[1 + t^2\langle u, v\rangle^2] - [1 + t\langle u, v\rangle]^2 = 2[1 + t^2\langle u, v\rangle^2] - [1 + t\langle u, v\rangle]^2 = [1 - t\langle u, v\rangle]^2 \geq 0.$$

Therefore, for $K = \sqrt{2}$, we have that $\|\langle X, v\rangle\|_{\psi_2} \leq K\|\langle X, v\rangle\|_{L_2}$.

(b_2) We note that the eigen-gap $\delta = \lambda_1(\Sigma) - \lambda_2(\Sigma) = 1 + t^2 - 1 = t^2$. Therefore, using the result of Davis-Kahan 2, we get with probability at least $1 - 2e^{-d}$,

$$\begin{aligned}
\min_{s\in\{\pm 1\}} \|u - sv\|_2 = \min_{s\in\{\pm 1\}} \|v_1(\Sigma) - sv_1(\hat{\Sigma})\|_2 \leq C\frac{\|\Sigma - \hat{\Sigma}\|}{\delta} &= C\frac{\|\Sigma - \hat{\Sigma}\|}{t^2} \\
&\leq \frac{C}{t^2}\left[cK^2\left(\sqrt{\frac{d}{n}} + \frac{d}{n}\right)\|\Sigma\|\right] \\
&\leq \frac{C}{t^2}\left[2c\left(\frac{1}{\sqrt{c'}} + \frac{1}{c'}\right)(1 + t^2)\right] \\
&\leq 0.1,
\end{aligned}$$

where c' is a large enough constant such that $n \geq c'd$, and the first inequality follows by the Inequality 1.