Math 7252: High-Dimensional Statistics
Homework 1

due: Sep 29, 11:59 pm

Please submit electronically directly to Canvas in a PDF file.
Each problem is worth the number of points in parentheses.
The full score is 44 points; you get "A" for 22+ points, "B" for 14+ points.

1 MGF method vs. moment bounds (3)

It is natural to compare the best bound on the tails obtained via MGF and by bounding the moments. As it turns out, the moment bounds are sharper, even if we only use the integer moments.

(a) Show that if $X > 0$ a.s., then for any $u > 0$,

$$\inf_{\lambda>0} M_X(\lambda)e^{-\lambda u} \geqslant \inf_{k\in\mathbb{Z}_+} \mathbb{E}\left[X^k\right]u^{-k}.$$

(b) Show that if X is symmetric (i.e. X and $-X$ have the same distribution), then for any $u > 0$,

$$\inf_{\lambda>0} M_X(\lambda)e^{-\lambda u} \geqslant \frac{1}{2} \inf_{k\in\mathbb{Z}_+} \mathbb{E}\left[X^{2k}\right]u^{-2k}.$$

2 Convexity of the cumulant-generating function (2)

For any distribution X, the logarithm of the MGF

$$K_X(t) = \log \mathbb{E}[e^{tX}]$$

is called the cumulant-generating function, or the *log-partition function* of the distribution.

(a) Show that K_X is convex. Use Young's inequality: for $a, b \in \mathbb{R}^d$ and $p, q \in [1, \infty]$ with $\frac{1}{p} + \frac{1}{q} = 1$,

$$|a^\top b| \leqslant \|a\|_p \|b\|_q.$$

You can assume that X has a discrete distribution.

3 Gaussian tails (3)

3.1 Mills ratio

Let $\phi(\cdot)$ be the p.d.f. of $\mathcal{N}(0,1)$, i.e. $\phi(t) = \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}$. For any $u \geqslant 0$, let $\Phi(u) := \int_{t \geqslant u} \phi(t) dt$.

(*a*) Prove the following bounds (holding for all $u \geqslant 0$):

$$\left(\frac{1}{u} - \frac{1}{u^3}\right) \phi(u) \leqslant \Phi(u) \leqslant \frac{1}{u} \phi(u).$$

Hint 1: Try to prove the upper bound first.

Hint 2: Integrate by parts – first to prove the upper bound, then again for the lower bound.

(*b*) Capitalizing on the trick you have just figured out to get the lower bound from the upper bound, prove a new upper bound:

$$\Phi(u) \leqslant \left(\frac{1}{u} - \frac{1}{u^3} + \frac{3}{u^5}\right) \phi(u).$$

Note that this bound is sharper than the previous one for large enough u.

*(*c*) If we continue this approach, we obtain a power series in $1/u$ for the Mills ratio $\Phi(u)/\phi(u)$; see Theorem 2.1 from Lecture 2. Get yourself convinced in it (no need to prove).

3.2 Power series for c.d.f.

Show that

$$\frac{1}{2} - \Phi(u) = \frac{1}{\sqrt{2\pi}} \sum_{k=0}^{\infty} \frac{(-1)^k u^{2k+1}}{2^k k! (2k+1)}.$$

Hint: change variable to remove u from the integration limits; differentiate in u under the integral.

4 Paley-Zygmund and friends (3)

(i) Prove the Paley-Zygmund inequality (it can be interpreted as a counterpart of Markov: a nonnegative random variable cannot be much *smaller* than its expectation):

If X is a non-negative random variable with $\mathbb{E}[X^2] < \infty$, then for any $t \in [0,1]$ one has

$$\mathbb{P}(X \geqslant (1-t)\mathbb{E}X) \geqslant t^2 \frac{(\mathbb{E}X)^2}{\mathbb{E}[X^2]}. \tag{1}$$

(ii) Under the same assumptions, strengthen (1), to Cantelli's inequality:

$$\mathbb{P}(X \geqslant (1-t)\mathbb{E}X) \geqslant t^2 \frac{(\mathbb{E}X)^2}{t^2(\mathbb{E}X)^2 + \mathrm{Var}[X]}.$$

This new inequality is sharp – give an example where it is attained.

(iii) Now: instead of $\mathbb{E}[X^2] < \infty$, assume that $\mathbb{E}[|X|^p] < \infty$ for some $p > 1$, and generalize (1) to

$$\mathbb{P}(X \geqslant (1-t)\mathbb{E}X) \geqslant \left(t^p \frac{(\mathbb{E}X)^p}{\mathbb{E}[|X|^p]}\right)^{\frac{1}{p-1}}.$$

Note that when $p > 2$, this gives an improvement over (1) for small t, which is important in applications where X is itself the sample average of i.i.d. $Y_1, ..., Y_n$.

Hint*: use Hölder's inequality: given $p, q \geqslant 1$ such that $1/p + 1/q = 1$, and random variables U, V on the same sample space, one has $\mathbb{E}[|UV|] \leqslant (\mathbb{E}|U|^p)^{\frac{1}{p}} \cdot (\mathbb{E}|V|^q)^{\frac{1}{q}}$.*

5 Stein's paradox (5)

Consider the problem of estimating the mean μ in the multivariate Gaussian location family

$$\mathbb{P}_\mu = \mathcal{N}(\mu, \boldsymbol{I}_d), \quad \mu \in \mathbb{R}^d, \tag{2}$$

where $\boldsymbol{I}_d$ is the $d \times d$ identity matrix, from a single observation $X \sim \mathbb{P}_\mu$. Note that here, X itself is the maximum likelihood estimator (MLE) for μ. Defining for any estimator $\hat{\mu} = \hat{\mu}(X)$ of μ the variance

$$\mathrm{Var}_\mu[\hat{\mu}] := \mathbb{E}_\mu[\|\hat{\mu} - \mathbb{E}[\hat{\mu}]\|^2]$$

and the quadratic risk

$$\mathrm{Risk}_\mu[\hat{\mu}] := \mathbb{E}_\mu[\|\hat{\mu} - \mu\|^2],$$

where $\|x\| := (\sum_i x_i^2)^{1/2}$ is the Euclidean norm of $x = (x_1, ..., x_d) \in \mathbb{R}^d$, we see that for any $\mu \in \mathbb{R}^d$,

$$\mathrm{Risk}_\mu[X] = \mathrm{Var}_\mu[X] = d.$$

Intuitively, one can suspect that no better estimator of X can be found: really, what can be done with only a single observation of the mean? Yet, this turns out to be false: one may improve over the MLE uniformly on the family (2) when $d > 2$. This celebrated result was established by James and Stein in 1976, and our goal is to reproduce it. But first, let us establish the terminology.

Definition 1. An estimator $\hat{\mu}$ is *dominated* by some other estimator $\hat{\mu}'$ if $\mathrm{Risk}_\mu[\hat{\mu}'] \leqslant \mathrm{Risk}_\mu[\hat{\mu}]$ for any μ, and there exists a parameter value $\bar{\mu}$ such that $\mathrm{Risk}_{\bar{\mu}}[\hat{\mu}'] < \mathrm{Risk}_{\bar{\mu}}[\hat{\mu}]$.

Definition 2. An estimator $\hat{\mu}$ is called *admissible* if it is not dominated by any other estimator. Otherwise, it is called *inadmissible*.

As statisticians, ideally we would like to compare two estimators over the whole family at once, without specifying a value of μ. Two admissible estimators cannot be compared this way, but at the very least we can rule out any *inadmissible* estimator, as for it there exists a uniformly better one.

You will show that the MLE is inadmissible when $d \geqslant 3$, by constructing a dominating estimator.

(*a*) Consider *shrinkage estimators* $\hat{\mu} = sX$ with $s \in \mathbb{R}$, and compute their risks for any s. Show that one can restrict attention to $s \in [0, 1]$ (hence "shrinkage") by finding a dominating estimator for $\hat{\mu}$ with $s < 0$ or $s > 1$.

(*b*) Show that, for given μ, the best value of s—i.e., the one minimizing the risk—is given by

$$s^* = \frac{\|\mu\|^2}{d + \|\mu\|^2} = 1 - \frac{d}{d + \|\mu\|^2}.$$

(*c*) Unfortunately, $\hat{\mu}^* = s^* X$ is not a proper estimator. (*Why?*) Instead of it, one may consider

$$\left(1 - \frac{d}{\|X\|^2}\right) X,$$

which is an actual estimator. Can you explain the heuristic motivation behind this estimator?

$^*(d)$ Assuming that $d \geqslant 2$, derive the *James-Stein estimator*

$$\hat{\mu}^{JS} = \left(1 - \frac{d-2}{\|X\|^2}\right) X \tag{3}$$

by minimizing over $\delta \in \mathbb{R}$ the risk of the estimator

$$\hat{\mu}^{\delta} = \left(1 - \frac{\delta}{\|X\|^2}\right) X$$

for a fixed μ. In order to show that $R(\delta) = \text{Risk}_{\mu}[\hat{\mu}^{\delta}]$ is minimized at $d-2$, use Stein's lemma:

Lemma 1. *Let $X \sim \mathcal{N}(\mu, I)$ and $g(x)$ be a function on $\mathbb{R}^d$ differentiable almost everywhere, and such that $\mathbb{E}_{\mu}\left[|\frac{\partial}{\partial x_i} g(X)|\right] < \infty$ and $\mathbb{E}_{\mu}[|(X_i - \mu_i)g(X)|] < \infty$ for any $i \in [d] := \{1, 2, ..., d\}$. Then*

$$\mathbb{E}_{\mu}[(X_i - \mu_i)g(X)] = \mathbb{E}_{\mu}\left[\frac{\partial}{\partial x_i} g(X)\right], \quad i \in [d].$$

When applying Stein's lemma to the right function $g(X)$, please do check the absolute integrability conditions in its premise, and explain why the argument does not work for $d = 1$. Finally, verify that $R(\delta)$ is strictly convex when $d \geqslant 3$ (thus $\hat{\mu}^{JS}$ indeed dominates the MLE).

6 Improved union bound for the maximum of Gaussians (3)

Solve Exercise 3.1 from Lecture 5 in lecture notes [LN]. You will find the definitions and context there.

7 Lower bound for the maximum of Gaussians (3)

Show that if $Z_1, \dots, Z_n \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$, one has $\mathbb{E}[\max_{j \in [n]} Z_j] \geqslant c\sqrt{\log(n+1)}$ for some constant $c > 0$.

8 Left tail of χ^2 is lighter than its right tail (2)

Show that for any selection of signs $(s_1, \ldots, s_n) \in \{-1, 1\}^n$, the "generalized χ_n^2" random variable $\sum_{j \in [n]} s_j Z_j^2$ admits the same (right) tail bound as the bound we derived for the actual χ_n^2: w.p. $1-\delta$,

$$\sum_{j \in [n]} Z_j^2 \lesssim \sqrt{n \log(\delta^{-1})} + \log(\delta^{-1}).$$

You might argue by examining the left tail of $Z^2 \sim \chi_1^2$ and showing that its "lighter" than the right tail, by comparing the MGF for Z^2 and $-Z^2$.

9 Moment bound under the Bernstein assumption (2)

Let $Z_1, \dots, Z_n$ be independent with $\mathbb{E}[Z_j] = 0$, $\mathbb{E}[Z_j^2] = \sigma_j^2$, and $|Z_j| \leqslant B$ a.s. Using the Bernstein inequality we proved (or will soon prove) in the class, show that for $p \geqslant 1$, $S_n := \sum_{j \in [n]} Z_j$ satisfies

$$\|S_n\|_{L_p} \lesssim \left(p \sum_{j \in [n]} \sigma_j^2 \right)^{1/2} + pB.$$

10 Moment bound for the maximum of subgaussian variables (2)

Let $X_1, \dots, X_n$ satisfy $\mathbb{E}[X_j] = 0$, $\|X_j\|_{\psi_2} \leqslant \sigma_j$. Show that for $p \geqslant 1$, $M_n := \max_{j \in [n]} X_j$ satisfies

$$\|M_n\|_{L_p} \lesssim \max_{j \in [n]} \sigma_j \sqrt{p \log(n+1)}.$$

11 Optimality of the Hoeffding lemma (2+1)

Solve Exercise 1.3 from Lecture 2 in [LN]. To get all 3 points, use the method of extreme points.

12 Orlicz norms I (2)

Solve Exercises 2.1–2.2 from Lecture 3 in [LN]. You will find the definitions and context therein.

13 Orlicz norms II (2)

Solve Exercises 3.1–3.2 from Lecture 3 in [LN]. You will find the definitions and context therein.

14 Hamburger moment problem (2)

Solve Exercise 4.1 from Lecture 2 in [LN]. You will find the definitions and context therein.

15 Polynomial with independent Gaussian coefficients (3)

I encountered this hidden gem as Eq. (1.43) in A. Nemirovski's lectures `https://www2.isye.gatech.edu/~nemirovs/Lect_SaintFlour.pdf`. Let $Z_k \sim \mathcal{N}(0,1)$ i.i.d., and consider the stochastic process

$$X_t = \sum_{k \in [n]} Z_k t^k, \quad t \in [0,1].$$

Show that with probability at least $1 - \delta$,

$$\sup_{t \in [0,1]} |X_t| \lesssim \sqrt{n \log(n/\delta)}.$$

Explain why this bound must be tight up to a log-factor. *Hint*: use the extreme points method.

16 Planar Venn diagrams (4)

A (congruent) *Venn diagram* in $\mathbb{R}^d$ for n sets is the following object: you choose a "template" $A \subset \mathbb{R}^d$ and n locations $a_1, ..., a_n \in \mathbb{R}^d$ such that the shifted sets $A_1, A_2, ..., A_n$, where $A_j := \{a + a_j, a \in A\}$, intersect in all possible combinations: for any subset of indices $I \subseteq \{1, 2, ..., n\}$, the set $A_I := \cap_{i \in I} A_i$ must be nonempty. Prove the following result:

One cannot draw a planar ($d = 2$) Venn diagram for $n \geqslant 4$ sets by shifting a circle.

Use **Euler's formula**: any planar graph with V vertices, E edges, and F faces (subsets in which $\mathbb{R}^2$ is partitioned by the graph) satisfies

$$V - E + F = 2.$$

For example, in the case of a triangle $V = E = 3$ and $F = 2$.
Hint: estimate V_n, E_n, F_n in a Venn diagram for n sets in terms of $V_{n-1}, E_{n-1}, F_{n-1}$ respectively.