36-709, Spring 2019
Homework 2

Due Thursday, Feb 21 by 5:00pm in JaeHyeok's mailbox

1. Orlicz norms. We have defined sub-gaussian and sub-exponential variables in terms of bounds on the moment generating functions. There exists an equivalent and more general way of expressing these properties using *Orlicz Norms* of random variables, which is more abstract but, at the same time, leads to simpler calculation. You will explore these concepts in this exercise. First, do the following problems in the book:

   (a) 2.18 and

   (b) 2.19.

   In this context, a random variables is said to be sub-gaussian if there exists a $K > 0$ such that

   $$\mathbb{E}\left[e^{X^2/K^2}\right] \leq 2 \tag{1}$$

   and sub-exponential if there exists a constant $K' > 0$ such that

   $$\mathbb{E}\left[e^{|X|/K'}\right] \leq 2. \tag{2}$$

   If X is sub-gaussian, its *sub-gaussian norm* is the smallest K satisfying (1), which correspond to $\|X\|_{\psi_2}$. Similarly, if X is sub-exponential, its *sub-exponential norm* is $\|X\|_{\psi_1}$, the smallest K' satisfying (2).

   (c) Prove that X is sub-gaussian if and only if X^2 is sub-exponential and

   $$\|X^2\|_{\psi_1} = \|X\|_{\psi_2}^2$$

   (d) If X and Y are sub-gaussians, then XY is sub-exponential with

   $$\|XY\|_{\psi_1} \leq \|X\|_{\psi_2}\|Y\|_{\psi_2}.$$

   Hint: use the inequality $xy \leq \frac{1}{2}x^2 + \frac{1}{2}y^2$, valid for all $x, y \in \mathbb{R}$.

   The last two properties would have made problem 8 in Homework 1 easier...

   Remarks. (Please read) it is possible to show that the above definitions are equivalent to the ones given in class: see the Appendix of Chapter 2 of the textbook. In particular, if X is sub-exponential then

   $$\mathbb{E}\left[e^{\lambda X}\right] \leq \exp \lambda^2\|X\|_{\psi_1}^2, \qquad \forall|\lambda| \leq \frac{1}{\|X\|_{\psi_1}}.$$

   From this, it is possible to derive the following, equivalent, versions of Hoeffding and Bernstein inequalities which you will also find in the literature.

   - **Hoeffding inequality**. Let $X_1, \ldots, X_n$ be independent, mean-zero sub-gaussian variables. Then, there exists a universal constant $c > 0$ such that, for any $t \geq 0$,

   $$\mathbb{P}\left(\left|\sum_{i=1}^n X_i\right| \geq t\right) \leq 2\exp\left(-\frac{ct^2}{\sum_{i=1}^n \|X_i\|_{\psi_2}^2}\right)$$

- **Bernestein inequality**. Let $X_1, \dots, X_n$ be independent, mean-zero sub-exponential variables. Then, there exists a universal constant $c > 0$ such that, for any $t \geq 0$,

$$\mathbb{P}\left(\left| \sum_{i=1}^n X_i \right| \geq t \right) \leq 2 \exp \left(-c \min \left\{ -\frac{t^2}{\sum_{i=1}^n \|X_i\|_{\psi_1}^2}, \frac{t}{\sum_{i=1}^n \|X_i\|_{\psi_1}} \right\} \right)$$

In other words, mapping to the notation used in class, $\sigma = \|X\|_{\psi_2}$ and $\nu = \alpha = \|X\|_{\psi_1}$.

2. Let $(X_1, \dots, X_n)$ be independent random variables with mean zero and let $(a_1, \dots, a_n) \in \mathbb{R}^n$. Compute bounds for

$$\mathbb{P}\left(|\sum_{i=1}^n a_i X_i| \geq t \right)$$

under the assumption that the X_i's are in the class $SG(\sigma^2)$ and also under the assumption that they are in $SE(\nu^2, \alpha)$. Compare the bounds. When does one dominate the other?

3. (Reading exercise. **Not to be graded for correctness, but only for effort**)
Suppose that $X_1, \dots, X_n$ are zero-mean, independent random variables belonging to the class $SG(\sigma^2)$ and $A = (A_{i,j})$ a $n \times n$ matrix. Let

$$\|A\|_{\text{op}} = \max_{x \in \mathbb{R}^n, x \neq 0} \frac{\|Ax\|}{\|x\|}$$

and

$$\|A\|_{\text{HS}} = \sqrt{\sum_{i=1}^n \sum_{j=1}^n A_{i,j}^2}$$

be the operator and the Hiolbert-Schmidt (or Frobenius) norm of A. Notice that $\|A\|_{\text{op}}$ is also the largest absolute eigenvalue of A. The goal of this exercise is to derive an exponential inequality for the probability

$$\mathbb{P}\left(\left| X^\top A X - \mathbb{E}\left[X^\top A X \right] \right| \geq t \right), \forall t \geq 0.$$

Do so by reproducing the proof of Theorem 1.1 from the following reference, using the definition of sub-Gaussian and sub-Exponential variables given in class.

- Rudelson, M., and Vershynin, R. (2013). Hanson-Wright inequality and sub-gaussian concentration. Electron. Commun. Probab., 18(82), 1- 9.

Notice that the definitions of sub-gaussian and sub-exponential variables in this paper is different than the ones given in class and correspond to the ones in problem 3. Make sure to keep track of the constants that depend on σ^2.

4. (a) Let $X = (X_1, \dots, X_d)$ be a d-dimensional vector comprosed of independent, zero-mean, unit-variance random variables in the class $SG(\sigma^2)$. Find a bound for

$$\mathbb{P}\left(\left| \|X\| - \sqrt{d} \right| \geq t\sqrt{d} \right), \quad t > 0.$$

We sketched the proof in class. Please give the details.

(b) Let E be a d-dimensional linear subspace of $\mathbb{R}^n$ and $(X_1, \dots, X_n)$ be a vector of independent zero-mean, unit-variance sub-Gaussian random variables with sub-Gaussian parameter σ^2. Compute a bound for

$$\mathbb{P}\left(|d(X,E) - \sqrt{n-d}| \geq u\right), \quad u \geq 0$$

where $d(X,E) = \inf_{y \in E} \|X - y\|$ is the distance between X and E.
Hint: Write $d(X,E) = \|P_{E^\perp} X\|$, where $P_{E^\perp}$ is the orthogonal projection of X onto the orthogonal complement of E and work wth $d^2(X,E)$. Use the result from problem 3.

5. Look at the paper *Hsu. D., Kakade, S. M. and ZXhang, T. (2012). A tail inequality for quadratic forms of subgaussian random vectors, Electron. Commun. Probab. 17, no. 52, 1–6.*

(a) Prove Proposition 1.2. Assume the u_i's to be independent.

(b) On page 5, take a look at the subsection titled "Example: fixed-design regression with subgaussian noise". The settings considered there are thos of a linear regression with fixed design. In details, let $Y_1, \dots, Y_n$ be independent random variables and $x_1, \dots, x_n$ be fixed points in $\mathbb{R}^d$. In the regression framework, it is typically assumed that

$$Y_i = f(x_i) + \epsilon_i, \quad i = 1, \dots, n, \tag{3}$$

where $f\colon \mathbb{R}^d \to \mathbb{R}$ is an regression function and the ϵ_i's are i.i.d. sub-gaussian centered variables. (The linear regression model further assumes the parametric form $f(x) = x^\top \beta$ for some $\beta \in \mathbb{R}^d$.) Here, we do not need (3) to hold. Setting $\Sigma_n = \frac{1}{n}\sum_{i=1}^n x_i x_i^\top$ (assumed invertible), let

$$\beta = \Sigma_n^{-1}\left(\frac{1}{n}\sum_{i=1}^n x_i \mathbb{E}[Y_i]\right) \quad \text{and} \quad \hat{\beta} = \Sigma_n^{-1}\left(\frac{1}{n}\sum_{i=1}^n x_i Y_i\right)$$

be the linear regression parameters and the least squares estimator of such parameters, respectively. Recall that β is the unique minimizer of the function $\eta \mapsto \sum_{i=1}^n \frac{1}{n}\mathbb{E}\left[(Y_i - x_i^\top \eta)^2\right]$, i.e. β minimizes the expected mean squared error, or L_2 risk. We are interested in evaluating the *excess risk* of $\hat{\beta}$, which is the increase in the expected mean squared error when using $\hat{\beta}$ instead of the optimal parameter β. To that end, suppose that we could observe n new independent responses $Y' = (Y_1', \dots, Y_n')$, independent of the original Y_i's. Then, for any vector $\gamma \in \mathbb{R}^d$ we may define the **excess risk of** γ as

$$R(\gamma) = \mathbb{E}_{Y'}\left[\frac{1}{n}\sum_{i=1}^n (Y_i' - x_i^\top \gamma)\right] - \mathbb{E}_{Y'}\left[\frac{1}{n}\sum_{i=1}^n (Y_i' - x_i^\top \beta)\right],$$

where $\mathbb{E}_{Y'}$ denotes expectation with respect to Y'. The excess risk $R(\gamma)$ measures the increase in the L_2 loss we incur if instead of using the optimal vector β to predict the Y_i''s at the design points we use γ. Now, it is natural to consider the excess risk $R(\hat{\beta})$ based on the least squares estimator $\widehat{\beta}$ computed using the original sample $Y_1, \dots, Y_n$. This is of course a random variable, since it is a function of $\hat{\beta}$ (and therefore of the Y_i's). It can be shown that

$$R(\hat{\beta}) = \left\|\Sigma_n^{1/2}\left(\hat{\beta} - \beta\right)\right\|^2 = \left\|\frac{1}{n}\sum_{i=1}^n \left(\Sigma_n^{-1/2} x_i\right)(Y_i - \mathbb{E}[Y_i])\right\|^2.$$

Prove the above identities (you may want to rewrite the expression for the excess risk using matrix algebra).

(c) Fill in the details for the application of Theorem 2.1 to derive a probabilistic bound on the excess risk of the least squares estimator.

6. **Robust statistics and the median-of-mean estimator**. Suppose we observe n i.i.d. random variables with distribution P and would like to construct a $1-\alpha$ confidence set for the expected value of P, where $\alpha \in (0,1)$.

(a) If the common distribution P is in the class $SG(\sigma^2)$ provide such a confidence interval.

(b) Now let's drop the assumption that P is a $SG(\sigma^2)$ distribution and in particular allow for very thick tails.

How should we proceed?

Here is a simple method. Assume that $\mathrm{Var}[X] = \sigma^2 < \infty$. For a fixed $\alpha \in [e^{1-n/2}, 1)$, set $b = \lceil \ln(1/\alpha) \rceil$ and note that $b \leq n/2$. Next, partition $[n] = \{1, \ldots, n\}$ into b blocks $B_1, \ldots, B_b$ each of size $|B_i| \geq \lfloor n/b \rfloor \geq 2$ and compute the sample mean in each block:

$$\overline{X}_i = \frac{1}{|B_i|} \sum_{j \in B_i} X_j, \quad i = 1, \ldots, b.$$

Finally define **the median-of-means** estimator as

$$\hat{\mu} = \hat{\mu}(\alpha) = \text{median}\left\{\overline{X}_1, \ldots, \overline{X}_b\right\},$$

where, for any b-tuple of numbers $(x_1, \ldots, x_b)$,

$$\text{median}\{x_1, \ldots, x_b\} = x_{j^*},$$

with

$$|\{k \in [b] \colon x_k \leq x_{j^*}\}| \geq b/2 \quad \text{and} \quad |\{k \in [b] \colon x_k \geq x_{j^*}\}| \geq b/2,$$

(if more than one such x_{j^*} satisfies the above inequalities, pick one of them at random).
Show that the median-of-means estimator yields, up to constants, the same type of sub-Gaussian confidence interval obtained in the first part, but without requiring the assumption of sub-Gaussianity. That is, show that

$$\mathbb{P}\left(|\hat{\mu} - \mu| \geq C\sqrt{\frac{\sigma^2 \log(1/\alpha)}{n}} \right) \leq \alpha,$$

for some constant C, where $\sigma^2 = \mathrm{Var}[X]$. You may want to consult these paper:

- M. Lerasle and R. I. Oliveira (2011). Robust empirical mean estimators. https://arxiv.org/pdf/1112.3914v1.pdf
- Luc Devroye, Matthieu Lerasle, Gabor Lugosi and Roberto I. Oliveira (2016). Sub-Gaussian mean estimators. https://arxiv.org/pdf/1509.05845v1.pdf

(c) The median-of-means estimator has an obvious drawback. What is it? *Hint: think of the situation when you want to use this estimator to compute confidence intervals at different levels α and α'...*

7. **Concentration for the bins and balls problem.**
In the balls and bins problem, m balls are thrown independently and at random into n bins (meaning: each balls is equally likely to be placed in any of the n bins, independently of the placements of the other balls). Let Z denotes the number of empty bins. We are interested in bounding

$$\mathbb{P}\left(|Z - \mathbb{E}[Z]| \geq t\right), \quad \forall t \geq 0. \tag{4}$$

(a) Show that $\mathbb{E}[Z] = n(1 - 1/n)^m$.

(b) Show that

$$\mathbb{P}\left(|Z - \mathbb{E}[Z]| \geq t\right) \leq 2 \exp\left\{\frac{-2t^2}{m}\right\}, \quad \forall t \geq 0.$$

8. **Median and sample quantiles.**

(a) Suppose that $(X_1, \ldots, X_n)$ is an i.i.d. sample from a distribution P (if you like, you may assume P to be absolutely continuous). Let $X_{(1)} \leq X_{(2)} < \ldots < X_{(n)}$ be the order statistics and set $\alpha \in (0,1)$. Determine a $1-\alpha$ confidence interval for the median of P of the form

$$\left(X_{(k_1)}, X_{(k_2)}\right)$$

for some choice of $k_1 < k_2$. Determine k_1 and k_2 by relating this problem to a $\text{Bin}(n, 1/2)$ distribution and use concentration.

(b) Consider the same setting as the previous exercise and let F be the c.d.f. of P and $p \in (0,1)$. The pth quantile and p-th sample quantile are, respectively,

$$\xi_p = \inf\{x \colon F(x) \geq p\}$$

and

$$\hat{\xi}_p = \inf\{x \colon F_n(x) \geq p\},$$

res[ectively, where F_n is the sample c.d.f. (i.e. $F_n(x) = \frac{1}{n}\sum_{i=1}^n 1(X_i \leq x)$). Show that, for any $\epsilon > 0$,

$$\mathbb{P}\left(|\hat{\xi}_p - \xi_p| > \epsilon\right) \leq 2 \exp\left\{-2n\delta_\epsilon^2\right\},$$

where $\delta_\epsilon = \min\left\{F(x_p + \epsilon) - p, p - F(\xi_p - \epsilon)\right\}$.
Write, for instance, $\mathbb{P}\left(\hat{\xi}_p > \xi_p + \epsilon\right) = \mathbb{P}\left(p > F_n(\xi_p + \epsilon)\right)$. *Then, notice that* $F_n(x)$ *is a sum of i.i.d. Bernoulli's and use Hoeffding yet again...*