36-709, Spring 2019
Homework 1

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

1. **Some simple but surprising facts about high-dimensional uniform distributions.**

   (a) Let X be a random vector in $\mathbb{R}^d$ that is uniformly distributed on the Euclidean unit ball $\{x \in \mathbb{R}^d \colon \|x\| \leq 1\}$. Show that

   $$\mathbb{E}[\|X\|] = \frac{d}{d+1}.$$

   (b) Let X be a random vector in $\mathbb{R}^d$ that is uniformly distributed on the unit cube $[-1,1]^d = \{x \in \mathbb{R}^d \colon \|x\|_\infty \leq 1\}$. Show that, for any $\epsilon \in (0,1)$,

   $$\lim_{d \to \infty} \mathbb{P}\left(\sqrt{\frac{d}{3}}(1-\epsilon) \leq \|X\| \leq \sqrt{\frac{d}{3}}(1+\epsilon) \right) = 1.$$

   Hint: the simplest strategy is to us the WLLN.

2. **Reading exercise. Not to be graded for correctness, but only for effort.** In this problem you are essentially required to reproduce a proof that can be found in the references given below. My intention is for you to read up and understand the proof rather than trying to solve this problem on your own, which would be challenging (though you are welcome to this challenge). Let $X = (X_1, \ldots, X_d) \in \mathbb{R}^d$ be a zero-mean random vector with covariance matrix Σ such that $\frac{X_i}{\sqrt{\Sigma_{i,i}}}$ is sub-Gaussian with parameter σ^2, for all $i = 1, \ldots, d$. Assume we observe n i.i.d. copies of X and compute the empirical covariance matrix $\widehat{\Sigma} = \frac{1}{n}\sum_{i=1}^n X_i^\top X_i$[1]. Show that, for all $i, j \in \{1, \ldots, d\}$,

   $$\mathbb{P}\left(\left| \widehat{\Sigma}_{i,j} - \Sigma_{i,j} \right| > \epsilon \right) \leq C_1 e^{-\epsilon^2 n C_2},$$

   for some constants C_1 and C_2. Conclude that

   $$\max_{i,j} \left| \widehat{\Sigma}_{i,j} - \Sigma_{i,j} \right| = O\left(\sqrt{\frac{\log d + \log n}{n}} \right),$$

   with probability at least $1 - \frac{1}{n}$. Thus, estimation of the covariance matrix in the L_∞ norm is possible even when d is much larger then n. Of course, you may wonder whether this is a good enough guarantee. In few weeks we will look at consistency rates for covariance estimation under more sensible norms and we will see that the requirements on d are much more stringent.

   You will definitely need to use the results in Problem 8 and you may want to take a look at these references:

   - Lemma 12 in Yuan. M. (2010). High Dimensional Inverse Covariance Matrix Estimation via Linear Programming, JMLR, 11, 2261-2286.
   - Lemma 1 in Ravikumar, P., Wainwright, M.J., Raskutti, G. and Yu, B. (2011). EJS, 5, 935-980.

[1] If mean of each X_i were not zero, then we would use $\widehat{\Sigma} = \frac{1}{n}\sum_{i=1}^n (X_i - \overline{X}_n)^\top (X_i \overline{X}_n)$ instead. We would obtain the same rate but the arguments would be slightly more complicated.

- Lemma A.3 in Bickel, P.J. and Levina, E. (2008). Regularized estimation of large covariance matrices, teh Annals of Statistics, 36(1), 199-227.

3. **Moments versus Cernoff bounds.** Show that moment bounds for tail probabilities are always better than Cramér–Chernoff bounds. More precisely, let Y be a nonnegative random variable and let $t > 0$. The best moment bound for the tail probability $\mathbb{P}Y \geq t$ is $\min_q \mathbb{E}[Y^q]t^{-q}$ where the minimum is taken over all positive integers. The best Cramér–Chernoff bound is $\inf_{\lambda>0} \mathbb{E}e^{\lambda(Y-t)}$. Prove that

$$\min_q \mathbb{E}[Y^q]t^{-q} \leq \inf_{\lambda>0} \mathbb{E}e^{\lambda(Y-t)}.$$

(See Philips, T.K. and Nelson, R. (1995). The moment bound is tighter than Chernoff's bound for positive tail probabilities. *The American Statistician*, 49, 175—178.)

4. **From tail bounds to moment bounds and high probability bounds.** In this exercise, you may think of the random variable X below as an average of n independent random variables, though this is not necessarily the case.

   (a) Suppose that the random variable X has mean zero and satisfies the inequality

   $$\mathbb{P}(|X| \geq t) \leq c_1 e^{-c_2 n t^a}, \quad \forall t > 0$$

   where $a \in \{1, 2\}$, n is a positive integer and c_1 and c_2 are positive numbers.

   i. Show that, when $a = 2$, $\mathbb{V}[X] \leq \frac{c_1}{nc_2}$.

   ii. Show that

   $$\mathbb{E}[|X|] \leq c_3 n^{-1/a}$$

   and express c_3 as a function of c_1 and c_2.

   (b) (From Hoeffding/Bernstein exponential inequality to high probability bounds). Suppose that, for all $t > 0$, some constants $a > 0$, $b > 0$, $c > 0$ and a $d \geq 0$ and a positive integer n, the random variable X is such that

   $$\mathbb{P}(|X| \geq t) \leq a \exp\left\{-\frac{nbt^2}{c + dt}\right\}.$$

   Then show that, for any $\delta \in (0, 1)$,

   $$|X| \leq \sqrt{\frac{c}{nb} \ln \frac{a}{\delta}} + \frac{d}{nb} \ln \frac{a}{\delta},$$

   with probability at least $1 - \delta$.
   If the above exponential inequality holds for $X - \theta$ instead of just X, where θ is a number (e.g., the mean or median of X), then the last bound would immediately give a $1 - \delta$ confidence interval for θ.

5. **Vectors in high-dimensions.** The results in this exercise provide the conceptual basis for the 2005 paper *Geometric representation of high dimension, low sample size data,* by Peter Hall, J. S. Marron and Amnon Neeman, published in the *Journal of the Royal Statistical Society,* Series B (Statistical Methodology), Vol. 67, No. 3, pp. 427-444. This paper is regarded by some as a seminal contribution to high-dimensional statistics. Check it out (especially sections 1 through 3).

(a) Let $X \sim N_d(0, I_d)$, where I_d is the d-dimensional identity matrix. Then, $\|X\|^2 = \sum_{i=1}^d X_i^2 \sim \chi_d^2$. Show that, for any $\epsilon \in (0,1)$

$$\mathbb{P}\left(\left|\|X\|^2 - d\right| \geq d\epsilon\right) \leq 2e^{-d\epsilon^2/8}.$$

You can use the following fact: the moment generating function of a χ_d^2 is $(1-2\lambda)^{-d/2}$ for all $\lambda < 1/2$. Alternatively, use the version of Bernstein inequality for sum of sub-exponential variables given in class. This results says that, in high dimensions, X is concentrated around a sphere of radius $\sqrt{d}$. Informaly, $\|X\| \sim \sqrt{d}$ with high probability.

See, e.g., Lemma 2 in *A Probabilistic Analysis of EM for Mixtures of Separated, Spherical Gaussians, by S. Dasgupta and L. Schulman, JMLR, 8, 203–26, 2007.*

Of course, the assumption of Gaussianity of X can be relaxed and a similar bound holds for any random d-dimensional vector of independent zero mean sub-Gaussian variables (using Bernstein inequality). Soon, we will derive a more general concentration bound for the norm $\|X\|$ of a sub-Gaussian vector X using maximal inequalities.

(b) Now assume that X and Y are both $N_d(0, I_d)$ and are independent. Below you are allowed to proceed in a heuristics manner, even though the arguments can be made rigorous. Show that

$$\mathbb{E}[(X^\top Y)^2] = d.$$

In fact, it can be shown that $|X^\top Y|$ concentrates around $\sqrt{d}$ as well, i.e. $|X^\top Y| \sim \sqrt{d}$ with high probability (again, this is an informal statemnent).

Using this fact and part (a), argue informally that

$$\frac{|X^\top Y|}{\|X\|\|Y\|} \sim \frac{1}{\sqrt{d}},$$

with high probability. Thus conclude that in high-dimensions, independent isotropic (i.e. having the identity matrix as the covariance matrix) Gaussian vectors arenearly orthogonal[2] with high probability; the higher the dimension, the more orthognal random isotropic vectors are.

You may use the fact that if $X \sim N_d(, I_n)$, then $\frac{X}{\|X\|}$ and $\|X\|$ are independent.

Again, the assumption of Gaussianity can be replaced by that of sub-Gaussianity.

6. **Random Projection and the Celebrated Johnson-Lindenstruass Lemma.**

See *D. Achlioptas, Database friendly random projections: Johnson-Lindenstrauss with binary coins, Journal of Computer and System Sciences 66 (2003) 671–687.*

Suppose we have a (deterministic) vector x in $\mathbb{R}^D$ and, for $\epsilon \in (0, 1/2)$ we would like to find a random mapping $f \colon \mathbb{R}^D \to \mathbb{R}^d$, where d is smaller than D, such that

$$(1-\epsilon)\|f(x)\|^2 \leq \|x\|^2 \leq (1+\epsilon)\|f(x)\|^2$$

with high probability. One way is to construct a $d \times D$ matrix A with iid entries from the $N(0,1)$ distribution and then take

$$f(x) = \frac{1}{\sqrt{d}} Ax, \quad x \in \mathbb{R}^D.$$

You can think of f as being a random projection from a high-dimensional space $\mathbb{R}^D$ into the smaller space $\mathbb{R}^d$.

Show that

[2]Recall that $x, y \in \mathbb{R}^d$ are orthognsal vectors whenever $x^\top y = 0$.

(a) $\|x\|^2 = \mathbb{E}\left[\|f(x)\|^2\right]$.

(b) For each $\epsilon \in (0, 1/2)$

$$\mathbb{P}\Big(\big|\|f(x)\|^2 - \|x\|^2\big| > \epsilon \|x\|^2 \Big) < 2 \exp\left\{ -d/4(\epsilon^2 - \epsilon^3) \right\}.$$

(c) Using the above result, show that, if we are given n deterministic vectors $(x_1, \ldots, x_n)$ in $\mathbb{R}^D$ and we compute their projections $f(x_1), \ldots, f(x_n)$ in $\mathbb{R}^d$, we are guaranteed that the all the pairwise squared distances between the projected points are distorted by at most a factor of $\epsilon \in (0, 1/2)$ with probability at least $1 - \delta$ if $d \geq \frac{4(\log(1/\delta) + 2\log(n))}{\epsilon^2 - \epsilon^3}$. That is,

$$\|x_i - x_j\|^2 (1 - \epsilon) \leq \|f(x_i) - f(x_j)\|^2 \leq \|x_i - x_j\|^2 (1 + \epsilon), \quad \forall i \neq j,$$

with probability at least $1 - \delta$.

For parts (a) and (b) proceed as follows: show that the squared norm of $\frac{\sqrt{d} f(x)}{\|x\|}$ is equal in distribution to the sum of d squared standard normals, and therefore has a χ^2_d distribution.
In your subsequent derivation, you may use the following facts:

(a) The mfg of a χ^2_1 at any $\lambda < 1/2$ is $(1 - 2\lambda)^{-1/2}$.

(b) For any $\epsilon \in (0, 1/2)$, setting $\lambda = \frac{\epsilon}{2(1+\epsilon)} < 1/2$, we get

$$\frac{e^{-2(1+\epsilon)\lambda}}{1 - 2\lambda} = (1 + \epsilon) e^{-\epsilon} < e^{-1/2(\epsilon^2 - \epsilon^3)}$$

and setting $\lambda = \frac{\epsilon}{2(1-\epsilon)} < 1/2$ we get

$$\frac{e^{2(1-\epsilon)\lambda}}{1 + 2\lambda} = (1 - \epsilon) e^{\epsilon} < e^{-1/2(\epsilon^2 - \epsilon^3)}$$

What is striking about this result is that the dimension D of the original space does not appear anywhere in these bounds!

This is an instance of what is also known as the Johnson-Lindenstruass Lemma, which loosely speaking, states that a random projection of n points from a high-dimensional space into a d dimensional space preserves the pairwise squared distances up to a multiplicative factor of ϵ with high probability if d is of order $\frac{\log n}{\epsilon^2}$, independently of the dimension of the original space.

Notice that instead of using independent $N(0, 1)$ variables to populate A, we could have used any sub-Gaussian distribution.

7. **The Bretagnole-Huber-Carol Inequality.** Let $(X_1, \ldots, X_k)$ have a multinomial distribution with parameters $(p_1, \ldots, p_k) \in (0, 1)^k$ and n. Show that

$$\mathbb{P}\left(\sum_{i=1}^{k} |X_i - np_i| \geq 2\sqrt{n} t \right) \leq 2^k \exp^{-2t^2}, \quad \forall t > 0.$$

8. **Squares and products of sub-gaussians are sub-exponentials.** Below, X and Y are centered random variables.

(a) Show that if $X \in SG(\sigma^2)$ than $X^2 \in SE(\nu^2, \alpha)$, for some α and ν, both depending on σ.

(b) Show that if $X \in SG(\sigma^2)$ and $Y \in SG(\tau^2)$, then $XY \in SE(\nu^2, \alpha)$, some α and ν, both depending on σ, both depending on σ and τ (see exercise 2.13 (d) in the book.)

Hint: For this problem you may find it helpful to use the following:

(a) **The C_r inequality:** *If X and Y are random variables such that $\mathbb{E}|X|^r < \infty$ and $\mathbb{E}|Y|^r < \infty$, where $r \geq 1$, then*

$$\mathbb{E}|X + Y|^r \leq 2^{r-1} \left(\mathbb{E}|X|^r + \mathbb{E}|Y|^r \right)$$

(b) The bound

$$\mathbb{E}|X|^r \leq (2\sigma^2)^{r/2} r\Gamma(r/2) \quad r \geq 1,$$

proved in class.

9. **Sub-Gaussian concentration around the mean implies sub-Gaussian concentration around the median, and vice versa.**

(a) Exercise 2.14 from Chapter 1. In part (c) you do not need to prove the claims about c_3 and c_2 and, similarly, in part (d) you do need to prove the claims about c_1 and c_2.

(b) Assume that $P(|X - \mu| \geq t) \leq Ce^{-ct^2}$ for all $t \geq 0$ and some positive C and c, where $\mu = \mathbb{E}[X]$. Find a bound for $|\mu - m_X|$, where m_X is a median for X. Assuming similarly that $P(|X - m_X| \geq t) \leq Ce^{-ct^2}$, find a bound for $|\mu - m_X|$.