36-755, Fall 2017
Homework 1n Solution

Due Sep 20.

Points: 100 pts total for the assignment.

1. On the MLE in parametric models.

   (a) Recall that the Kullback-Leibler (KL) divergence between two probability measures P and Q on some measurable space $(\mathcal{X}, \mathcal{B})$ with densities p and q with respect to a common dominating measure μ is

   $$K(P,Q) = \begin{cases} \int_{\mathcal{X}} p(x) \log\left(\frac{p(x)}{q(x)}\right) d\mu(x) & \text{if } P << Q \\ \infty & \text{otherwise.} \end{cases}$$

   Use Jensen inequality to show that $K(P,Q) \geq 0$ with equality if and only if $P = Q$.

   (b) Assume that $\mathcal{P} = \{P_\theta, \theta \in \Theta\}$ is a parametric model over the sample space $(\mathcal{X}, \mathcal{B})$, such that $P_\theta << \mu$ for all $\theta \in \Theta$, for some σ-finite dominating measure μ. Assume also that all the P_θ's have the same support and $\theta \neq \theta'$ implies that $P_\theta \neq P_{\theta'}$. Let $\mathbb{X}_n = (X_1, \ldots, X_n) \overset{id}{\sim} P_{\theta_0}$ for some $\theta_0 \in \Theta$ and write

   $$L_n(\theta; \mathbb{X}_n) = \prod_i^n p_\theta(X_i),$$

   for the likelihood function at $\theta \in \Theta$, where p_θ is the density of P_θ with respect to μ
   Use the law of large numbers to show that, for any $\theta \neq \theta_0$ in Θ,

   $$\lim_{n\to\infty} \mathbb{P}\left(L_n(\mathbb{X}_n; \theta_0) > L_n(\mathbb{X}_n; \theta)\right) = 1$$

   The previous result offers an asymptotic justification of why in this case the MLE is a sensible choice. *Hint: express the inequality in term of log-likelihood ratio and show that the ratio converges in probability to $K(P_{\theta_0}, P_\theta)$. You can use the law of large numbers.*

Points: 16 pts = 10 + 6.

Solution.
(a)
First, when P is not absolutely continuous with respect to Q, then $P \neq Q$ and $K(P,Q) = \infty > 0$ holds.
Second, when $P \ll Q$, then applying Jensen inequality on concave function $\varphi(y) = \log y$ gives

$$\begin{aligned} K(P,Q) &= \int p(x) \log\left(\frac{p(x)}{q(x)}\right) d\mu(x) = -\int p(x) \log\left(\frac{q(x)}{p(x)}\right) d\mu(x) \\ &= -\mathbb{E}_P\left[\log\left(\frac{q(X)}{p(X)}\right)\right] \\ &\geq -\log\left(\mathbb{E}_P\left[\frac{q(X)}{p(X)}\right]\right) \\ &= -\log\left(\int p(x)\frac{q(x)}{p(x)} d\mu(x)\right) = -\log 1 = 0. \end{aligned}$$

And since $\varphi(y) = \log y$ is strictly concave, equality holds if and only if $\frac{q(X)}{p(X)}$ is a point mass under P, i.e. if and only if there exists $a \in \mathbb{R}$ such that $P\left(\frac{q(X)}{p(X)} = a\right) = 1$. Then $\mathbb{E}_P\left[\frac{q(X)}{p(X)}\right] = \mathbb{E}_P[a] = 1$, so $a = 1$. Hence $\frac{q(x)}{p(x)} = 1$ a.s. with respect to P, i.e. $p(x) = q(x)$ a.e. with respect to μ on $\{x : p(x) > 0\}$. Then

$$\begin{aligned} 0 &= \int_{p(x)=0} p(x)d\mu(x) = 1 - \int_{p(x)>0} p(x)d\mu(x) \\ &= 1 - \int_{p(x)>0} q(x)d\mu(x) = \int_{p(x)>0} q(x)d\mu(x), \end{aligned}$$

so $q(x) = 0$ a.e. with respect to μ on $\{x : p(x) = 0\}$. Hence $p(x) = q(x)$ a.e. with respect to μ, and hence $P = Q$ holds.
From the first and the second, $K(P, Q) \geq 0$ holds with equality if and only if $P = Q$.
(b)
Note that condition $L_n(\mathbb{X}_n; \theta_0) > L_n(\mathbb{X}_n; \theta)$ can be equivalently written as

$$\begin{aligned} L_n(\mathbb{X}_n; \theta_0) > L_n(\mathbb{X}_n; \theta) &\iff \log L_n(\mathbb{X}_n; \theta_0) > \log L_n(\mathbb{X}_n; \theta) \\ &\iff \sum_{i=1}^n \log p_{\theta_0}(X_i) > \sum_{i=1}^n \log p_\theta(X_i) \\ &\iff \frac{1}{n}\sum_{i=1}^n \log \frac{p_{\theta_0}(X_i)}{p_\theta(X_i)} > 0 \\ &\iff \frac{1}{n}\sum_{i=1}^n Y_i > 0, \end{aligned}$$

where $Y_i = \log \frac{p_{\theta_0}(X_i)}{p_\theta(X_i)}$ for $i = 1, \cdots, n$. Then since all P_θ has same support, $P_{\theta_0} \ll P_\theta$ holds, so expectation of Y_i under P_{θ_0} can be computed as

$$\mathbb{E}_{\theta_0}[Y_i] = \mathbb{E}_{\theta_0}\left[\log \frac{p_{\theta_0}(X_i)}{p_\theta(X_i)}\right] = K(P_{\theta_0}, P_\theta) \in (0, \infty).$$

Hence by weak law of large number,

$$\frac{1}{n}\sum_{i=1}^n Y_i \xrightarrow{p} K(P_{\theta_0}, P_\theta) > 0,$$

and hence $\frac{1}{n}\sum_{i=1}^n Y_i \rightsquigarrow K(P_{\theta_0}, P_\theta)$ as well. since 0 is a continuous point of the cdf of constant random variable $K(P_{\theta_0}, P_\theta)$, so

$$\lim_{n\to\infty} P\left(\frac{1}{n}\sum_{i=1}^n Y_i > 0\right) = \lim_{n\to\infty} P\left(K(P_{\theta_0}, P_\theta) > 0\right) = 1.$$

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 random vector with covariance matrix Σ such that $\frac{X_i}{\sqrt{\Sigma_{i,i}}}$ is sub-Gaussian with parameter ν^2, for all $i = 1, \ldots, d$. Assume we observe n i.i.d. copies of X and compute the empirical covariance matrix $\widehat{\Sigma}$. 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

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

Thus, estimation of the covariance matrix in the L_∞ norm is possible even when d is much larger then n. Of course, you may ask yourself 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 may want to look 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.
- 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.

Points: 10 pts.

Solution.

Let $X^{(k)} = (X_{i1}, \cdots, X_{ik})$ be k-th sample. Note that the distribution of $\hat{S}_{ij} := \frac{1}{n}\sum_{k=1}^{n}(X_{ik} - \bar{X}_i)(X_{jk} - \bar{X_j})$ does not depend on $\mathbb{E}[X]$, so we can assume $\mathbb{E}[X] = 0$ by letting $Y := X - \mathbb{E}[X]$ if necessary.

First, we fix $i, j \in \{1, \ldots, d\}$ and note that the difference $\hat{S}_{ij} - \Sigma_{ij}$ can be factorized as

$$\hat{S}_{ij} - \Sigma_{ij} = \frac{1}{n}\sum_{k=1}^{n} X_{ik}X_{jk} - \mathbb{E}\left[X_i X_j\right] \tag{1}$$

$$+ \left(\frac{1}{n}\sum_{k=1}^{n} X_{ik}\right)\left(\frac{1}{n}\sum_{k=1}^{n} X_{jk}\right). \tag{2}$$

Consider (1) first. Since $\frac{X_i}{\sqrt{\Sigma_{i,i}}}, \frac{X_j}{\sqrt{\Sigma_{j,j}}} \in SG(\nu^2)$, $X_i \in SG(\nu^2\Sigma_{ii})$ and $X_j \in SG(\nu^2\Sigma_{jj})$ holds. Then from Claim below,

$$X_i X_j \in SE\left(121\nu^4\Sigma_{ii}\Sigma_{jj}, 11\nu^2\sqrt{\Sigma_{ii}\Sigma_{jj}}\right).$$

And then from Problem 7 Details,

$$\sum_{k=1}^{n} X_{ik}X_{jk} \in SE\left(121n\nu^4\Sigma_{ii}\Sigma_{jj}, 11\nu^2\sqrt{\Sigma_{ii}\Sigma_{jj}}\right),$$

and hence

$$\frac{1}{n}\sum_{k=1}^{n} X_{ik}X_{jk} \in SE\left(\frac{121\nu^4\Sigma_{ii}\Sigma_{jj}}{n}, \frac{11\nu^2\sqrt{\Sigma_{ii}\Sigma_{jj}}}{n}\right).$$

Then $\frac{\frac{121\nu^4\Sigma_{ii}\Sigma_{jj}}{n}}{\frac{11\nu^2\sqrt{\Sigma_{ii}\Sigma_{jj}}}{n}} = 11\nu^2\sqrt{\Sigma_{ii}\Sigma_{jj}}$, so by applying sub-exponential tail bound,

$$\begin{aligned}\mathbb{P}\left(\left|\frac{1}{n}\sum_{k=1}^{n} X_{ik}X_{jk} - \mathbb{E}\left[X_iX_j\right]\right| \geq t\right) &\leq \begin{cases}\exp\left(-\frac{nt^2}{242\nu^4\Sigma_{ii}\Sigma_{jj}}\right) & \text{if } 0 \leq t \leq 11\nu^2\sqrt{\Sigma_{ii}\Sigma_{jj}},\\ \exp\left(-\frac{nt}{22\nu^2\sqrt{\Sigma_{ii}\Sigma_{jj}}}\right) & \text{if } t \geq 11\nu^2\sqrt{\Sigma_{ii}\Sigma_{jj}}.\end{cases}\\ &= \exp\left(-\frac{nt}{22\nu^2\sqrt{\Sigma_{ii}\Sigma_{jj}}}\min\left\{\frac{t}{11\nu^2\sqrt{\Sigma_{ii}\Sigma_{jj}}}, 1\right\}\right) \qquad (3)\end{aligned}$$

Now, consider (2). From Problem 7 Details, $\sum_{k=1}^{n} X_{ik} \in SG(n\nu^2\Sigma_{ii})$ and $\sum_{k=1}^{n} X_{jk} \in SG(n\nu^2\Sigma_{jj})$, and hence

$$\frac{1}{n}\sum_{k=1}^{n} X_{ik} \in SG\left(\frac{\nu^2\Sigma_{ii}}{n}\right) \qquad \text{and} \qquad \frac{1}{n}\sum_{k=1}^{n} X_{ik} \in SG\left(\frac{\nu^2\Sigma_{jj}}{n}\right).$$

Then from Claim below,

$$\left(\frac{1}{n}\sum_{k=1}^{n} X_{ik}\right)\left(\frac{1}{n}\sum_{k=1}^{n} X_{jk}\right) \in SE\left(\frac{121\nu^4\Sigma_{ii}\Sigma_{jj}}{n}, \frac{11\nu^2\sqrt{\Sigma_{ii}\Sigma_{jj}}}{n}\right).$$

Hence by applying sub-exponential tail bound,

$$\begin{aligned}\mathbb{P}\left(\left|\left(\frac{1}{n}\sum_{k=1}^{n} X_{ik}\right)\left(\frac{1}{n}\sum_{k=1}^{n} X_{jk}\right)\right| \geq t\right) &\leq \begin{cases}\exp\left(-\frac{nt^2}{242\nu^4\Sigma_{ii}\Sigma_{jj}}\right) & \text{if } 0 \leq t \leq 11\nu^2\sqrt{\Sigma_{ii}\Sigma_{jj}},\\ \exp\left(-\frac{nt}{22\nu^2\sqrt{\Sigma_{ii}\Sigma_{jj}}}\right) & \text{if } t \geq 11\nu^2\sqrt{\Sigma_{ii}\Sigma_{jj}}.\end{cases}\\ &\leq \exp\left(-\frac{nt}{22\nu^2\sqrt{\Sigma_{ii}\Sigma_{jj}}}\min\left\{\frac{t}{11\nu^2\sqrt{\Sigma_{ii}\Sigma_{jj}}}, 1\right\}\right) \qquad (4)\end{aligned}$$

Then we combine (3) and (4) with union bound to get upper bound of probability for fixed i, j as

$$\begin{aligned}&\mathbb{P}\left(\left|\hat{S}_{ij} - \Sigma_{ij}\right| \geq t\right)\\ &\leq \mathbb{P}\left(\left|\frac{1}{n}\sum_{k=1}^{n} X_{ik}X_{jk} - \mathbb{E}\left[X_iX_j\right]\right| \geq \frac{t}{2}\right) + \mathbb{P}\left(\left|\left(\frac{1}{n}\sum_{k=1}^{n} X_{ik}\right)\left(\frac{1}{n}\sum_{k=1}^{n} X_{jk}\right)\right| \geq \frac{t}{2}\right)\\ &\leq 2\exp\left(-\frac{nt}{22\nu^2\sqrt{\Sigma_{ii}\Sigma_{jj}}}\min\left\{\frac{t}{11\nu^2\sqrt{\Sigma_{ii}\Sigma_{jj}}}, 1\right\}\right)\end{aligned}$$

And hence again by applying union bound,

$$\begin{aligned}\mathbb{P}\left(\max_{i,j}\left|\hat{S}_{ij}-\Sigma_{ij}\right|\geq t\right) &\leq \sum_{i,j}\mathbb{P}\left(\left|\hat{S}_{ij}-\Sigma_{ij}\right|\geq t\right)\\ &\leq \sum_{i,j}2\exp\left(-\frac{nt}{22\nu^2\sqrt{\Sigma_{ii}\Sigma_{jj}}}\min\left\{\frac{t}{11\nu^2\sqrt{\Sigma_{ii}\Sigma_{jj}}},1\right\}\right)\\ &\leq d(d+1)\exp\left(-\frac{nt}{22\nu^2\max_i\{\Sigma_{ii}\}}\min\left\{\frac{t}{11\nu^2\max_i\{\Sigma_{ii}\}},1\right\}\right).\end{aligned}$$

Now for any $\epsilon>0$, $d(d+1)\exp\left(-\frac{nt}{22\nu^2\max_i\{\Sigma_{ii}\}}\min\left\{\frac{t}{11\nu^2\max_i\{\Sigma_{ii}\}},1\right\}\right)\leq\epsilon$ if $t\leq 11\nu^2\max_i\{\Sigma_{ii}\}$ and $t\geq\sqrt{\frac{242\nu^4\max_i\{\Sigma_{ii}^2\}\log\left(\frac{d(d+1)}{\epsilon}\right)}{n}}$ holds or $t\geq 11\nu^2\max_i\{\Sigma_{ii}\}$ and $t\geq\frac{22\nu^2\max_i\{\Sigma_{ii}\}\log\left(\frac{d(d+1)}{\epsilon}\right)}{n}$ holds. Both conditions are satisfied when $t\geq 44\nu^2\max_i\{\Sigma_{ii}\}\left(2+\log\left(\frac{1}{\epsilon}\right)\right)\max\left\{\sqrt{\frac{\log(d+1)}{n}},\frac{\log(d+1)}{n}\right\}$. Hence

$$\mathbb{P}\left(\max_{i,j}\left|\hat{S}_{ij}-\Sigma_{ij}\right|\geq 44\nu^2\max_i\{\Sigma_{ii}\}\left(2+\log\left(\frac{1}{\epsilon}\right)\right)\max\left\{\sqrt{\frac{\log(d+1)}{n}},\frac{\log(d+1)}{n}\right\}\right)\leq\epsilon.$$

And hence

$$\max_{i,j}\left|\hat{S}_{ij}-\Sigma_{ij}\right|=O_P\left(\max\left\{\sqrt{\frac{\log(d+1)}{n}},\frac{\log(d+1)}{n}\right\}\right).$$

Claim. If $X\in SG(\sigma_1^2)$ and $Y\in SG(\sigma_2^2)$ with $\mathbb{E}[X]=\mathbb{E}[Y]=0$, then

$$XY\in SE\left((11\sigma_1\sigma_2)^2,11\sigma_1\sigma_2\right).$$

I used the bound that $\mathbb{E}|X|^r\leq(2\sigma_1^2)^{\frac{r}{2}}r\Gamma\left(\frac{r}{2}\right)\quad r\geq 1$.

$$\begin{aligned}\mathbb{E}\left[\exp\left(\lambda(XY-\mathbb{E}\left[XY\right])\right)\right] &\leq \sum_{r=0}^{\infty}\frac{1}{r!}|\lambda|^r\mathbb{E}\left[(XY-\mathbb{E}\left[XY\right])^r\right]\\ &\leq 1+\sum_{r=2}^{\infty}\frac{1}{r!}|\lambda|^r2^{r-1}\left(\mathbb{E}\left[|XY|^r\right]+\left(\mathbb{E}\left[|XY|\right]\right)^r\right)\\ &\leq 1+\sum_{r=2}^{\infty}\frac{1}{r!}|\lambda|^r2^r\sqrt{\mathbb{E}\left[|X|^{2r}\right]\mathbb{E}\left[|Y|^{2r}\right]}\\ &\leq 1+\sum_{r=2}^{\infty}\frac{1}{r!}|\lambda|^r2^r(2\sigma_1\sigma_2)^r(2r)\Gamma(r)\\ &= 1+\sum_{r=2}^{\infty}2(4\sigma_1\sigma_2|\lambda|)^r\\ &= 1+\frac{32\sigma_1^2\sigma_2^2\lambda^2}{1-4\sigma_1\sigma_2|\lambda|}\end{aligned}$$

Now, for $|\lambda| \leq \frac{1}{11\sigma_1\sigma_2}$, above can be bounded as

$$\begin{aligned}\mathbb{E}\left[\exp\left(\lambda(X^2 - \mathbb{E}\left[X^2\right])\right)\right] &\leq 1 + \frac{32\sigma_1^2\sigma_2^2\lambda^2}{1 - 4\sigma_1\sigma_2|\lambda|} \\ &\leq 1 + \frac{11}{7} \times 32\sigma_1^2\sigma_2^2\lambda^2 \\ &\leq \exp\left(\frac{1}{2}\lambda^2 \left(11\sigma_1\sigma_2\right)^2\right).\end{aligned}$$

Hence $X^2 \in SE(\nu^2, \alpha)$ where $\nu = \alpha = 11\sigma_1\sigma_2$.

3. (Sampling with replacement). Let $\mathcal{X}$ a finite set with N elements. Let $X_1, \ldots, X_n$ be a random sample without replacement from $\mathcal{X}$ and $Y_1, \ldots, Y_n$ be a random sample with replacement from $\mathcal{X}$. Show that, for any convex function $f\colon \mathbb{R}^n \to \mathbb{R}$,

$$\mathbb{E}\left[f\left(\sum_{i=1}^n X_i\right)\right] \leq \mathbb{E}\left[f\left(\sum_{i=1}^n Y_i\right)\right].$$

Use this result to show that all the inequalities derived for the sums of independent random variables $\{Y_1, \ldots, Y_n\}$ using Chernoff's bounding techniques remain true also for the sums of the X_i's. (see *Hoeffding, W. (1963). Probility Inequalities for sums of Bounded Random Variables, by W. Hoeffding, JASA, 58, 13–30., 1963*).
Actually, this is a special case of a more general property known as negative dependence. The punchline is that negatively dependent variables obeys the same Chernoff's bound as independent variables. Perhaps the most prominent example of negatively dependent variables is that of Multinomial variables. For more information see `http://www.brics.dk/RS/96/25/BRICS-RS-96-25.pdf`

Points: 10 pts.

Solution.
Let $[n] := \{1, \ldots, n\}$, $F_{N,n} := \{\sigma : [n] \to [N]\}$ be the set of functions from $[n]$ to $[N]$, and let $I_{N,n} := \{\sigma \in F_{N,n} |\, \sigma \text{ is injective}\}$ be the set of injective functions from $[n]$ to $[N]$. Let $P(N,n) := |I_{N,n}|$ be the permutation number, and let $\mathcal{X} = \{c_1, \ldots, c_N\}$. Then

$$\mathbb{E}\left[f\left(\sum_{i=1}^n X_i\right)\right] = \frac{1}{P(N,n)} \sum_{\sigma \in I_{N,n}} f\left(\sum_{i=1}^n c_{\sigma(i)}\right),$$

while

$$\mathbb{E}\left[f\left(\sum_{i=1}^n Y_i\right)\right] = \frac{1}{N^n} \sum_{\sigma \in F_{N,n}} f\left(\sum_{i=1}^n c_{\sigma(i)}\right).$$

Then there exists $\tilde{f} : \mathcal{X}^n \to \mathbb{R}$ symmetric function satisfying

$$\frac{1}{P(N,n)} \sum_{\sigma \in I_{N,n}} \tilde{f}\left(c_{\sigma(1)}, \ldots, c_{\sigma(n)}\right) = \frac{1}{N^n} \sum_{\sigma \in F_{N,n}} f\left(\sum_{i=1}^n c_{\sigma(i)}\right), \tag{5}$$

where

$$\tilde{f}\left(x_1, \ldots, x_n\right) = \sum_{r_1, \ldots, r_n} p(r_1, \ldots, r_n) f\left(\sum_{i=1}^n r_i x_i\right)$$

with the sum taken over all nonnegative integers $r_1, \ldots, r_n$ with $\sum r_i = n$, and the coefficients p being independent of f. Then applying constant function $f(x) = 1$ gives

$$\sum_{r_1,\ldots,r_n} p(r_1, \ldots, r_n) = 1.$$

And applying $f(x) = x$ implies $\tilde{f}$ is a symmetric linear function , then (5) implies that $\tilde{f}(x_1, \ldots, x_n) = x_1 + \cdots + x_n$. Hence this gives

$$\sum_{r_1,\ldots,r_n} p(r_1, \ldots, r_n) \left(\sum_{i=1}^{n} r_i x_i \right) = \sum_{i=1}^{n} x_i.$$

Hence Jensen's inequality on a convex function f gives

$$\begin{aligned} \tilde{f}\left(x_1, \ldots, x_n\right) &= \sum_{r_1,\ldots,r_n} p(r_1, \ldots, r_n) f\left(\sum_{i=1}^{n} r_i x_i \right) \\ &\geq f\left(\sum_{r_1,\ldots,r_n} p(r_1, \ldots, r_n) \left(\sum_{i=1}^{n} r_i x_i \right) \right) = f\left(\sum_{i=1}^{n} x_i \right). \end{aligned}$$

And hence $\mathbb{E}\left[\tilde{f}\left(X_1, \ldots, X_n\right)\right] \geq \mathbb{E}\left[f\left(\sum_{i=1}^{n} X_i\right)\right]$ holds. Then (5) implies $\mathbb{E}\left[\tilde{f}\left(X_1, \ldots, X_n\right)\right] = \mathbb{E}\left[f\left(\sum_{i=1}^{n} Y_i\right)\right]$, and hence

$$\mathbb{E}\left[f\left(\sum_{i=1}^{n} X_i \right) \right] \leq \mathbb{E}\left[f\left(\sum_{i=1}^{n} Y_i \right) \right].$$

Let $\mu = \mathbb{E}[X_i]$. For Chernoff's bounding techniques, note that $x \mapsto \exp\left(\lambda(x - n\mu)\right)$ is a convex function. Hence the tail probability for the sum of the X_i's can be bounded as

$$\begin{aligned} \mathbb{P}\left(\sum_{i=1}^{n} X_i - n\mu \geq t \right) &= \mathbb{P}\left(\exp\left(\lambda \left(\sum_{i=1}^{n} X_i - n\mu \right) \right) \geq e^{\lambda t} \right) \\ &\leq e^{-\lambda t} \mathbb{E}\left[\exp\left(\lambda \left(\sum_{i=1}^{n} X_i - n\mu \right) \right) \right] \text{ (markov)} \\ &\leq e^{-\lambda t} \mathbb{E}\left[\exp\left(\lambda \left(\sum_{i=1}^{n} Y_i - n\mu \right) \right) \right] \text{ (from above)} \\ &= e^{-\lambda t} \mathbb{E}\left[\exp\left(n\lambda(X_1 - \mu) \right) \right]. \text{ (independence)} \end{aligned}$$

Hence Chernoff's bounding techniques works for the sum of the samples without replacements as well.

4. From tail bounds to moment bounds and high probability bounds.

   (a) Suppose that, the random variable X satisfies the inequality

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

   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

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

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 $t > 0$, and some positive constants a, b, c and a non-negative constant d,

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

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 $1-\delta$.

Points: 11 pts = 6 + 5.

Solution.

(a)

For any constant $p \geq 1$, consider $\mathbb{E}\left[|X|^p\right]$. Applying the formula $\mathbb{E}\left[|X|^p\right] = \int_0^\infty \mathbb{P}\left(|X|^p \geq t\right) dt$ gives

$$\begin{aligned}
\mathbb{E}\left[|X|^p\right] &= \int_0^\infty \mathbb{P}\left(|X|^p \geq t\right) dt \\
&= \int_0^\infty \mathbb{P}\left(|X| \geq t^{\frac{a}{p}}\right) dt \\
&\leq \int_0^\infty c_1 \exp\left(-c_2 n t^{\frac{a}{p}}\right) dt.
\end{aligned}$$

For applying change of variables, if we let $c_2 n t^{\frac{a}{p}} = x$, then $dt = \frac{p}{a}\frac{x^{\frac{p}{a}-1}}{(c_2 n)^{\frac{p}{a}}}dt$, so

$$\begin{aligned}
\mathbb{E}\left[|X|^p\right] &\leq \int_0^\infty \frac{pc_1}{a(c_2 n)^{\frac{p}{a}}} x^{\frac{p}{a}-1} e^{-x} dx \\
&= \frac{pc_1}{a(c_2 n)^{\frac{p}{a}}}\Gamma\left(\frac{p}{a}\right).
\end{aligned}$$

(i)

Applying $p = a = 2$ to above equation gives

$$\mathbb{E}\left[X^2\right] \leq \frac{c_1}{nc_2},$$

and hence

$$\mathbb{V}[X] \leq \mathbb{E}\left[X^2\right] \leq \frac{c_1}{nc_2}.$$

(ii)

Note that $(\mathbb{E}[|X|])^p \leq \mathbb{E}[|X|^p]$ holds by applying Jensen's inequality on $f(x) = x^p$, hence

$$\mathbb{E}[|X|] \leq (\mathbb{E}[|X|^p])^{\frac{1}{p}} \leq c_2^{-\frac{1}{a}} \left(c_1 \frac{p}{a} \Gamma\left(\frac{p}{a}\right)\right)^{\frac{1}{p}} n^{-\frac{1}{a}}.$$

Since this holds for any $p \geq 1$, so

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

with

$$c_3 = c_2^{-\frac{1}{a}} \min_{p \geq 1} \left\{ \left(c_1 \frac{p}{a} \Gamma\left(\frac{p}{a}\right)\right)^{\frac{1}{p}} \right\}.$$

(b)
Note that $t \geq 0$, so

$$\delta = a \exp\left\{-\frac{nbt^2}{c+dt}\right\} \iff \frac{nbt^2}{c+dt} = \ln\left(\frac{a}{\delta}\right)$$
$$\iff t = \frac{d \ln\left(\frac{a}{\delta}\right) + \sqrt{d^2 \ln^2\left(\frac{a}{\delta}\right) + 4nbc \ln\left(\frac{a}{b}\right)}}{2nb}.$$

Hence

$$\mathbb{P}\left(|X| \geq \frac{d \ln\left(\frac{a}{\delta}\right) + \sqrt{d^2 \ln^2\left(\frac{a}{\delta}\right) + 4nbc \ln\left(\frac{a}{b}\right)}}{2nb}\right) \leq \delta$$

holds. Now,

$$\frac{d \ln\left(\frac{a}{\delta}\right) + \sqrt{d^2 \ln^2\left(\frac{a}{\delta}\right) + 4nbc \ln\left(\frac{a}{b}\right)}}{2nb} < \frac{d \ln\left(\frac{a}{\delta}\right) + \sqrt{d^2 \ln^2\left(\frac{a}{\delta}\right)} + \sqrt{4nbc \ln\left(\frac{a}{b}\right)}}{2nb}$$
$$= \sqrt{\frac{c}{nb} \ln \frac{a}{\delta}} + \frac{d}{nb} \ln \frac{a}{\delta},$$

And hence

$$\mathbb{P}\left(|X| > \sqrt{\frac{c}{nb} \ln \frac{a}{\delta}} + \frac{d}{nb} \ln \frac{a}{\delta}\right) \leq \mathbb{P}\left(|X| \geq \frac{d \ln\left(\frac{a}{\delta}\right) + \sqrt{d^2 \ln^2\left(\frac{a}{\delta}\right) + 4nbc \ln\left(\frac{a}{b}\right)}}{2nb}\right) \leq \delta$$

holds. i.e. with probability at least $1 - \delta$, $|X| \leq \sqrt{\frac{c}{nb} \ln \frac{a}{\delta}} + \frac{d}{nb} \ln \frac{a}{\delta}$ holds.

5. Let X be distributed like a $N_d(0, I_d)$, where I_d is the d-dimensional identity matrix. Then, $\|X\|^2 = \sum_{i=1}^d X_i^2 \sim \chi_d^2$.

   (a) 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 $\lambda < 1/2$. Alternstively, 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 $\sqrt{d}$.
   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.*
   You should convince yourself that the same result holds for any vector X whose entries are i.i.d. sub-Gaussians.

(b) Now assume that X and Y are both $\in N_d(0, I_d)$ and are independent. Argue **very informally (it is OK to use heuristics)** that

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

with high probability. Thus conclude that in high-dimensions, independent isotropic Gaussian vectors are orthogonal with high probability, the more so the higher the dimension.
You may use the fact that if $X \sim N_d(0, I_n)$, then X and $\|X\|$ are independent.
Again, the assumption of Gaussianity can be replaced by that of sub-Gaussianity.

Points: 16 pts = 8 + 8.

Solution.
(a)
Since $||X||^2 \sim \chi^2_d$, its mgf is $m_{||X||^2}(\lambda) = (1-2\lambda)^{-d/2}$ for $\lambda \in \left(-\frac{1}{2}, \frac{1}{2}\right)$, and hence $\mathbb{E}\left[||X||^2\right] = \frac{d}{d\lambda} m_{||X||^2}(0) = d$. Therefore,

$$\mathbb{E}\left[\exp\left(\lambda(||X||^2 - d)\right)\right] = (1-2\lambda)^{-d/2} e^{-\lambda d}.$$

Now, note that let $f(x) = \log(1-x)+x+x^2$ for $x \in \left[-\frac{1}{2}, \frac{1}{2}\right]$, then $f'(x) = -\frac{1}{1-x}+1+2x = \frac{x(1-2x)}{1-x}$, so f is decreasing on $x \in \left[-\frac{1}{2}, 0\right]$ and is increasing on $x \in \left[0, \frac{1}{2}\right]$. Hence $f(x) \geq f(0) = 0$, i.e. $\log(1-x) \geq -x - x^2$ holds for $x \in \left[-\frac{1}{2}, \frac{1}{2}\right]$. Hence for $\lambda \in \left[-\frac{1}{4}, \frac{1}{4}\right]$,

$$\begin{aligned}
\mathbb{E}\left[\exp\left(\lambda(||X||^2 - d)\right)\right] &= \exp\left(-\frac{d}{2}\log(1-2\lambda) - \lambda d\right) \\
&\leq \exp\left(-\frac{d}{2}\left(-2\lambda - 4\lambda^2\right) - \lambda d\right) \\
&= e^{4d\lambda^2/2}.
\end{aligned}$$

Hence $||X||^2$ is sub-exponential with $(\nu, \alpha) = (2\sqrt{d}, 4)$. Then since $\epsilon < 1$, $d\epsilon < d = \frac{(2\sqrt{d})^2}{4}$, hence from sub-exponential tail bound,

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

By applying the same sub-exponential tail bound to $-||X||^2$, we have

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

hence we get the desired inequality as

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

(b)
Note that $\frac{Y}{\|Y\|}$ is supported on $\mathbb{S}^{d-1} = \{x \in \mathbb{R}^d : \|x\| = 1\}$. Now, for any $a, b \in \mathbb{S}^{d-1}$, there exists $R \in O(d) = \{A \in \mathbb{R}^{d\times d} : A^\top A = AA^\top = I_d\}$ such that $b = R^\top a$, i.e. R is a rotation on $\mathbb{S}^{d-1}$. Then note that the norm of X is invariant under the rotation R, i.e.

$$\|RX\|^2 = X^\top R^\top RX = X^\top X = \|X\|^2,$$

and the distribution of X is invariant under the rotation R, i.e.

$$RX \sim N\left(0,\, RI_dR^\top\right) \overset{d}{=} N(0,\, I_d).$$

Hence the distributions of $\left(\frac{X}{\|X\|}\right)^\top a$ and $\left(\frac{X}{\|X\|}\right)^\top b$ are the same, i.e.

$$\left(\frac{X}{\|X\|}\right)^\top b = \left(\frac{X}{\|X\|}\right)^\top R^\top a = \left(\frac{RX}{\|RX\|}\right)^\top a \overset{d}{=} \left(\frac{X}{\|X\|}\right)^\top a.$$

Hence this together with X and Y being independent implies $\left(\frac{X}{\|X\|}\right)^\top \left(\frac{Y}{\|Y\|}\right) \mid \frac{Y}{\|Y\|} = a$ has same distribution for all $a \in \mathbb{S}^{d-1}$. Hence let $e = (1, 0, \ldots, 0) \in \mathbb{S}^{d-1}$, then

$$\begin{aligned}\left(\frac{X}{\|X\|}\right)^\top \left(\frac{Y}{\|Y\|}\right) &\overset{d}{=} \left(\frac{X}{\|X\|}\right)^\top e \\ &= \frac{X_1}{\sqrt{\sum_{i=1}^d X_i^2}}.\end{aligned}$$

Now, note that $\frac{1}{d}\sum_{i=1}^d X_i^2 \overset{\text{a.s.}}{\to} 1$ by law of large numbers, hence by Slutsky's theorem,

$$\sqrt{d}\frac{|X_1|}{\sqrt{\sum_{i=1}^d X_i^2}} = \frac{|X_1|}{\sqrt{\frac{1}{d}\sum_{i=1}^d X_i^2}} \rightsquigarrow |X_1|.$$

Hence for any $f : \mathbb{N} \to \mathbb{R}_+$ with $f(d) \in \Omega\left(\frac{1}{\sqrt{d}}\right)$ (i.e. decreasing slower than $\frac{1}{\sqrt{d}}$),

$$\mathbb{P}\left(\frac{|X_1|}{\sqrt{\sum_{i=1}^d X_i^2}} > f(d)\right) = \mathbb{P}\left(\sqrt{d}\frac{|X_1|}{\sqrt{\sum_{i=1}^d X_i^2}} > \sqrt{d}f(d)\right) \to 0$$

as $d \to \infty$, and for any $f : \mathbb{N} \to \mathbb{R}_+$ with $f(d) \in o\left(\frac{1}{\sqrt{d}}\right)$ (i.e. decreasing faster than $\frac{1}{\sqrt{d}}$),

$$\mathbb{P}\left(\frac{|X_1|}{\sqrt{\sum_{i=1}^d X_i^2}} < f(d)\right) = \mathbb{P}\left(\sqrt{d}\frac{|X_1|}{\sqrt{\sum_{i=1}^d X_i^2}} < \sqrt{d}f(d)\right) \to 0$$

as $d \to \infty$. Hence

$$\frac{|X^\top Y|}{\|X\|\|Y\|} \overset{d}{=} \frac{|X_1|}{\sqrt{\sum_{i=1}^d X_i^2}} \sim \frac{1}{\sqrt{d}}.$$

6. Suppose that $X_1, \ldots, X_n$ are such that $X_i \in SG(\sigma_i^2)$, not necessarily independent. Show that $\sum_{i=1}^n X_i \in SG(\tau^2)$ and find τ. What if $X_i \in SE(\tau_i^2, \alpha_i)$ for all i?

Points: 12 pts.

Solution.

Let $p_1, \cdots, p_n \in [1, \infty]$ be satisfying $\sum_{i=1}^{n} \frac{1}{p_i} = 1$. By using Hlder's inequality,

$$\begin{aligned}
&\mathbb{E}\left[\exp\left(\lambda\left(\sum_{i=1}^{n} X_i - \mathbb{E}\left[\sum_{i=1}^{n} X_i\right]\right)\right)\right] \\
&= \mathbb{E}\left[\exp\left(\lambda\left(X_1 - \mathbb{E}[X_1]\right)\right)\exp\left(\lambda\sum_{i=2}^{n}\left(X_i - \mathbb{E}[X_i]\right)\right)\right] \\
&\leq \mathbb{E}\left[\exp\left(p_1\lambda\left(X_1 - \mathbb{E}[X_1]\right)\right)\right]^{\frac{1}{p_1}} \mathbb{E}\left[\exp\left((1-p_1)\lambda\sum_{i=2}^{n}\left(X_i - \mathbb{E}[X_i]\right)\right)\right]^{\frac{1}{1-p_1}}.
\end{aligned}$$

By applying Hlder's inequality repeatedly, we get

$$\mathbb{E}\left[\exp\left(\lambda\left(\sum_{i=1}^{n} X_i - \mathbb{E}\left[\sum_{i=1}^{n} X_i\right]\right)\right)\right] = \prod_{i=1}^{n}\mathbb{E}\left[\exp\left(p_i\lambda\left(X_i - \mathbb{E}[X_i]\right)\right)\right]^{\frac{1}{p_i}}.$$

Note that $SG(\sigma_i^2) = SE(\sigma_i^2, 0)$, so we can generally solve for $SE(\tau_i^2, \alpha)$. If $X_i \in SE(\tau_i^2, \alpha_i)$, then $\mathbb{E}\left[\exp\left(p_i\lambda\left(X_i - \mathbb{E}[X_i]\right)\right)\right] \leq \exp\left(\frac{p_i^2\lambda^2\tau_i^2}{2}\right)$ for $|\lambda| \leq \frac{1}{p_i\alpha_i}$ holds for each $i = 1, \cdots, n$, hence

$$\begin{aligned}
\mathbb{E}\left[\exp\left(\lambda\left(\sum_{i=1}^{n} X_i - \mathbb{E}\left[\sum_{i=1}^{n} X_i\right]\right)\right)\right] &\leq \prod_{i=1}^{n}\left(\exp\left(\frac{p_i^2\lambda^2\tau_i^2}{2}\right)\right)^{\frac{1}{p_i}} \\
&= \exp\left(\frac{\lambda^2}{2}\sum_{i=1}^{n} p_i\tau_i^2\right),
\end{aligned}$$

for $|\lambda| \leq \min_{1\leq i\leq n}\left\{\frac{1}{p_i\alpha_i}\right\}$. Hence

$$\sum_{i=1}^{n} X_i \in SE\left(\sum_{i=1}^{n} p_i\tau_i^2, \max_{1\leq i\leq n}\{p_i\alpha_i\}\right)$$

for any $p_1, \cdots, p_n \in [1, \infty]$ satisfying $\sum_{i=1}^{n} \frac{1}{p_i} = 1$.

If we are in particular interested in minimizing $\sum_{i=1}^{n} p_i\tau_i^2$, we need to minimize $\sum_{i=1}^{n} p_i\tau_i^2$ with conditions $\sum_{i=1}^{n} \frac{1}{p_i} = 1$. Minimum occurs when $p_i = \frac{\sum_{j=1}^{n}\tau_j}{\tau_i}$, hence plugging in those values gives

$$\sum_{i=1}^{n} X_i \in SE\left(\left(\sum_{i=1}^{n}\tau_i\right)^2, \max_{1\leq i\leq n}\left\{\frac{\alpha_i}{\tau_i}\right\}\sum_{j=1}^{n}\tau_j\right).$$

Hence when $X_i \in SG(\sigma^2)$, then plugging $\tau_i^2 = \sigma_i^2$ and $\alpha_i = 0$ gives

$$\sum_{i=1}^{n} X_i \in SG\left(\left(\sum_{i=1}^{n}\sigma_i\right)^2\right).$$

Details.

Note that this constant $\left(\sum_{i=1}^n \sigma_i\right)^2$ is strict: when $Z \sim N(0,1)$ and $X_i = \sigma_i Z$, then $X_i \in SG(\sigma_i^2)$ and

$$\mathbb{E}\left[\exp\left(\lambda \sum_{i=1}^n X_i\right)\right] = \mathbb{E}\left[\exp\left(\lambda\left(\sum_{i=1}^n \sigma_i\right) Z\right)\right] = \exp\left(\frac{1}{2}\lambda^2 \left(\sum_{i=1}^n \sigma_i\right)^2\right).$$

Also, note that if X_i were independent, then for $X_i \in SE(\tau_i^2, \alpha_i)$

$$\begin{aligned}\mathbb{E}\left[\exp\left(\lambda\left(\sum_{i=1}^n X_i - \mathbb{E}\left[\sum_{i=1}^n X_i\right]\right)\right)\right] &= \prod_{i=1}^n \mathbb{E}\left[\exp\left(\lambda\left(X_i - \mathbb{E}[X_i]\right)\right)\right] \\ &\leq \prod_{i=1}^n \exp\left(\frac{\lambda^2 \tau_i^2}{2}\right) \\ &= \exp\left(\frac{\lambda^2}{2}\sum_{i=1}^n p_i \tau_i^2\right),\end{aligned}$$

for $|\lambda| \leq \min\left\{\frac{1}{\alpha_i}\right\}$, hence

$$\sum_{i=1}^n X_i \in SE\left(\sum_{i=1}^n \tau_i^2, \max_{1\leq i\leq n}\{\alpha_i\}\right),$$

and if $X_i \in SG(\sigma_i^2)$, then

$$\sum_{i=1}^n X_i \in SG\left(\sum_{i=1}^n \sigma_i^2\right).$$

Then $\sum_{i=1}^n \sigma_i^2 \leq \left(\sum_{i=1}^n \sigma_i\right)^2$, so having an additional independence assumption results in a better constant.

7. (Random Projection and the Johnson-Lindenstruass Lemma).

   See *D. Achlioptas, Database friendly random projections: Johnson-Lindenstrauss with binary coins, Journal of Computer and System Sciences 66 (2003) 671687.*

   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

   (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.

Points: 13 pts = 5 + 5 + 3.

Solution.

(a)

Note that ith element of $f(x)$ is $(f(x))_i = \frac{1}{\sqrt{d}} \sum_{j=1}^{D} A_{ij} x_j$, and since A_{ij} are i.i.d. $N(0, 1)$, the distribution of $(f(x))_i$ becomes

$$(f(x))_i \sim N\left(0, \frac{1}{d} \sum_{j=1}^{D} x_j^2\right) = N\left(0, \frac{||x||^2}{d}\right).$$

And since $\{(f(x))_i\}_{i=1}^d$ are independent, hence

$$\frac{\sqrt{d}f(x)}{||x||} \sim N\left(0, I_{d\times d}\right),$$

where $I_{d\times d}$ is d by d identity matrix. And hence

$$\frac{d||f(x)||^2}{||x||^2} = \left(\frac{\sqrt{d}f(x)}{||x||}\right)^\top \frac{\sqrt{d}f(x)}{||x||} \sim \chi_d^2.$$

Therefore, $\mathbb{E}\left[\frac{d||f(x)||^2}{||x||^2}\right] = \frac{d}{d\lambda} m_{\chi_d^2}(0) = d$, i.e.

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

(b)
Note that $\mathbb{E}\left[\exp\left(\lambda\frac{d||f(x)||^2}{||x||^2}\right)\right] = m_{\chi_d^2}(\lambda) = (1-2\lambda)^{-\frac{d}{2}}$. Hence for any $\lambda \in \left(0, \frac{1}{2}\right)$,

$$\begin{aligned}
\mathbb{P}\left(||f(x)||^2 > (1+\epsilon)||x||^2\right) &= \mathbb{P}\left(\frac{d||f(x)||^2}{||x||^2} > (1+\epsilon)d\right) \\
&\leq \exp\left(-\lambda(1+\epsilon)d\right)\mathbb{E}\left[\exp\left(\lambda\frac{d||f(x)||^2}{||x||^2}\right)\right] \\
&= (1-2\lambda)^{-\frac{d}{2}}\exp\left(-\lambda(1+\epsilon)d\right).
\end{aligned}$$

Now plugging in $\lambda = \frac{\epsilon}{2(1+\epsilon)} < \frac{1}{2}$ gives

$$\begin{aligned}
\mathbb{P}\left(||f(x)||^2 > (1+\epsilon)||x||^2\right) &\leq (1-2\lambda)^{-\frac{d}{2}}\exp\left(-\lambda(1+\epsilon)d\right). \\
&= (1+\epsilon)^{\frac{d}{2}}\exp\left(-\frac{\epsilon d}{2}\right) \\
&< \exp\left(-\frac{d}{4}(\epsilon^2-\epsilon^3)\right).
\end{aligned}$$

Also for any $\lambda \in \left(0, \frac{1}{2}\right)$,

$$\begin{aligned}
\mathbb{P}\left(||f(x)||^2 < (1-\epsilon)||x||^2\right) &= \mathbb{P}\left(-\frac{d||f(x)||^2}{||x||^2} > -(1-\epsilon)d\right) \\
&\leq \exp\left(\lambda(1-\epsilon)d\right)\mathbb{E}\left[\exp\left(-\lambda\frac{d||f(x)||^2}{||x||^2}\right)\right] \\
&= (1+2\lambda)^{-\frac{d}{2}}\exp\left(\lambda(1-\epsilon)d\right).
\end{aligned}$$

Now plugging in $\lambda = \frac{\epsilon}{2(1-\epsilon)} < \frac{1}{2}$ gives

$$\begin{aligned}
\mathbb{P}\left(||f(x)||^2 < (1-\epsilon)||x||^2\right) &\leq (1+2\lambda)^{-\frac{d}{2}}\exp\left(\lambda(1-\epsilon)d\right). \\
&= (1-\epsilon)^{\frac{d}{2}}\exp\left(\frac{\epsilon d}{2}\right) \\
&< \exp\left(-\frac{d}{4}(\epsilon^2-\epsilon^3)\right).
\end{aligned}$$

Hence combining these gives

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

(c)

Now note that since f is a linear function, $f(x_i) - f(x_j) = f(x_i - x_j)$. Since there are $\frac{n(n-1)}{2}$ pairs of $\{x_i, x_j\}$, hence

$$\begin{aligned}
&\mathbb{P}\Big(\big| \|f(x_i) - f(x_j)\|^2 - \|x_i - x_j\|^2 \big| > \epsilon \|x_i - x_j\|^2 \, \exists i, j \Big) \\
&\leq \sum_{x_i, x_j} \mathbb{P}\Big(\big| \|f(x_i) - f(x_j)\|^2 - \|x_i - x_j\|^2 \big| > \epsilon \|x_i - x_j\|^2 \Big) \\
&< n(n-1) \exp\left\{ -\frac{d(\epsilon^2 - \epsilon^3)}{4} \right\} \\
&< \exp\left(2 \log n - \frac{d(\epsilon^2 - \epsilon^3)}{4} \right).
\end{aligned}$$

Hence when $d \geq \frac{4(\log(1/\delta) + 2\log(n))}{\epsilon^2 - \epsilon^3}$, then

$$\exp\left(2 \log n - \frac{d(\epsilon^2 - \epsilon^3)}{4} \right) \leq \exp\left(2 \log n - \left(2 \log n + \log\left(\frac{1}{\delta} \right) \right) \right) = \delta$$

holds, so

$$\mathbb{P}\Big(\big| \|f(x_i) - f(x_j)\|^2 - \|x_i - x_j\|^2 \big| \leq \epsilon \|x_i - x_j\|^2 \, \forall i, j \Big) \geq 1 - \delta,$$

i.e.

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

holds with probability at least $1 - \delta$.

8. Show that if $X \in SG(\sigma^2)$ than $X^2 \in SE(\nu^2, \alpha)$ where

$$\nu = \alpha = 16\sigma^2.$$

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

(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 following bound, proved in class

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

Feel free to prove the claim by other methods and/or by obtaining sharper (smaller) bounds on ν^2 and/or $1/\alpha$.

Points: 12 pts.

Solution.

$$\begin{aligned}
\mathbb{E}\left[\exp\left(\lambda(X^2-\mathbb{E}\left[X^2\right])\right)\right] &\leq \sum_{r=0}^{\infty}\frac{1}{r!}|\lambda|^r\mathbb{E}\left[\left(X^2-\mathbb{E}\left[X^2\right]\right)^r\right] \\
&\leq 1+\sum_{r=2}^{\infty}\frac{1}{r!}|\lambda|^r 2^{r-1}\left(\mathbb{E}\left[X^{2r}\right]+\left(\mathbb{E}\left[X^2\right]\right)^r\right) \\
&\leq 1+\sum_{r=2}^{\infty}\frac{1}{r!}|\lambda|^r 2^r(2\sigma^2)^r(2r)\Gamma(r) \\
&= 1+\sum_{r=2}^{\infty}2(4\sigma^2|\lambda|)^r \\
&= 1+\frac{32\sigma^4\lambda^2}{1-4\sigma^2|\lambda|}
\end{aligned}$$

Now, for $|\lambda| \leq \frac{1}{11\sigma^2}$, above can be bounded as

$$\begin{aligned}
\mathbb{E}\left[\exp\left(\lambda(X^2-\mathbb{E}\left[X^2\right])\right)\right] &\leq 1+\frac{32\sigma^4\lambda^2}{1-4\sigma^2|\lambda|} \\
&\leq 1+\frac{11}{7}\times 32\sigma^4\lambda^2 \\
&\leq \exp\left(\frac{1}{2}\lambda^2\left(11\sigma^2\right)^2\right).
\end{aligned}$$

Hence $X^2 \in SE(\nu^2,\alpha)$ where $\nu=\alpha=11\sigma^2$.