Homework 13
High-Dimensional Probability for Data Science, Fall 2023

Hints are in the back of this homework set.

As in the previous homework sets, $C, C_1, C_2, \dots$ and $c, c_1, c_2, \dots$ denote positive absolute constants of your choice.

Next, we will prove a few basic facts about the operator norm. They will be helpful in the further problems.

1. A few identities for the operator norm

(a) Let A be any matrix. Explain why the following identities hold:

$$\|A\|^2 = \lambda_1\left(AA^\mathsf{T}\right) = \lambda_1\left(A^\mathsf{T}A\right) = \left\|AA^\mathsf{T}\right\| = \left\|A^\mathsf{T}A\right\|.$$

(b) Deduce from (a) that the operator norm is invariant under transposition, i.e.

$$\|A^\mathsf{T}\| = \|A\|.$$

(c) Let S be a $d_1 \times d_1$ symmetric matrix, and T be a $d_2 \times d_2$ symmetric matrix. Prove that the $(d_1+d_2) \times (d_1+d_2)$ block diagonal matrix

$$Y = \begin{bmatrix} S & 0 \\ 0 & T \end{bmatrix} \quad \text{satisfies} \quad \|Y\| = \|S\| \vee \|T\|.$$

Here $\vee$ is a shorthand for maximum, i.e. $a \vee b = \max(a,b)$.

(d) Let A be a $d_1 \times d_2$ matrix, and B be a $d_2 \times d_1$ matrix. Prove that the $(d_1+d_2) \times (d_1+d_2)$ block matrix

$$Z = \begin{bmatrix} 0 & A \\ B & 0 \end{bmatrix} \quad \text{satisfies} \quad \|Z\| = \|A\| \vee \|B\|.$$

So far, we proved matrix concentration inequalities—Hoeffding and Bernstein–for *symmetric* matrices only. The symmetry was crucial in the proof: it was required in Lieb's trace inequality. However, *Girko's hermitization trick* allows us to automatically extend all those results to nonsymmetric matrices. In fact, arbitrary rectangular $d_1 \times d_2$ matrices B will be allowed.

Girko's hermitization trick consists of replacing B by the $(d_1+d_2) \times (d_1+d_2)$ block matrix

$$H(B) := \begin{bmatrix} 0 & B \\ B^\mathsf{T} & 0 \end{bmatrix}.$$

Then $H(B)$ is obviously symmetric. Moreover, by Problem 1 parts (d) and (b), hermitization preserves the operator norm: $\|H(B)\| = \|B\|$.

Using hermetization trick, one can extend many results from symmetric to general matrices. Let's do this for the expected form of matrix Hoeffding inequality (Lecture 22, p.3).

2. Matrix Hoeffding inequality for rectangular matrices

Let ε_i be independent Rademacher random variables, i.e. each ε_i takes values ± 1 independently with probability $1/2$. Let B_i be (fixed) $d_1 \times d_2$ matrices. Prove that with probability at least 0.9 we have the following bound:

$$\left\| \sum_{i=1}^{n} \varepsilon_i B_i \right\| \le C\sigma \sqrt{\log(d_1 + d_2)} \quad \text{where } \sigma^2 = \left\| \sum_{i=1}^{n} B_i B_i^{\mathsf{T}} \right\| \vee \left\| \sum_{i=1}^{n} B_i^{\mathsf{T}} B_i \right\|.$$

Some real-world problems require us to operate with huge matrices (for example, matrices of interactions between hundreds of thousands of individuals). Some matrices are so huge that computing their eigenvectors and eigenvalues is practically impossible. In such cases, one can try to reduce the size of a matrix by *subsampling*: include only a small random sample of rows or columns.

In the following problem, we consider a very tall $N \times d$ matrix A; think of $N \gg n$. We show how to approximately compute the singular values of A from a smaller matrix B obtained by subsampling $n = O(d \log d)$ rows of A.

3. Subsampling a matrix

Let A be a $d \times N$ matrix whose columns A_i have the same length, i.e. $\|A_i\|_2 = \|A_j\|_2$ for all $i = 1, \ldots, N$. Consider the $d \times n$ matrix R obtained by including n randomly chosen columns of A, where each time a column is chosen independently, with replacement, and with the same probability $1/N$. Let $B = \sqrt{N/n}\, R$. Show that, as long as $n \ge C d \log d$ with a sufficiently large absolute constant C, the singular values satisfy

$$\max_{i=1,\ldots,n} \left| \sigma_i(A)^2 - \sigma_i(B)^2 \right| \le 0.1\, \sigma_1(A)^2$$

with probability at least 0.9.

The general covariance estimation result (Lecture 23) guarantees that the covariance of a d-dimensional distribution can be accurately estimated from a sample of size $n = O(d \log d)$. You will now show that the logarithmic oversampling factor $\log d$ is in general necessary.

4. Logarithmic oversampling is sometimes unavoidable

(a) Check that for any matrix A, all of its entries are bounded by the operator norm, i.e.

$$\max_{i,j} |A_{ij}| \le \|A\|.$$

(b) Let $e_1, \dots, e_d$ be the canonical basis of $\mathbb{R}^d$. Consider a random vector X that is uniformly distributed in the set $\{\sqrt{d}\, e_1, \dots, \sqrt{d}\, e_d\}$, i.e. X takes each of the n values $\sqrt{d}\, e_k$ with probability $1/d$. Check that

$$\Sigma = \mathbb{E}\, X X^\mathsf{T} = I_d.$$

(c) Argue[1] that, unless $n > cd \log d$, the sample covariance matrix Σ_n has at least one zero diagonal entry with probability at least 0.9.

(d) Conclude that if $n < cd \log d$, we have $\|\Sigma_n - \Sigma\| \geq \|\Sigma\|$ with probability at least 0.9, so the covariance estimation fails.

TURN OVER FOR HINTS

[1] In solving part (c), you can assume without proof the following classical result in probability about the *coupon collector's problem*. Suppose each day we receive a coupon in mail. There are d types of coupons, and each day's coupon is one of these types independently with equal probability. Then, with probability at least 0.9, we need at least $cd \log d$ days to obtain a full collection of all d types of coupons.

Hints for Problem 1

(a) For the first two identities, use the definition of the operator norm (Lecture 18). For the next two identities, use the result of Homework 10, Problem 3(a).

(c) Note that the spectrum of Y is the union of the spectra of S and T. Then recall the result of the result of Homework 10, Problem 3(a) again.

(d) Compute ZZ^{T}; it should be a block-diagonal matrix. Use parts (c) and (a) to compute its norm and relate it to the norms of A and B.

Hints for Problem 2

Apply matrix Hoeffding inequality $\|\sum_{i=1}^{n} \varepsilon_i A_i\| \leq 10\sigma\sqrt{\log d}$ (Lecture 22) for the Hermiticized matrices, i.e. for $A_i := H(B_i)$. As we noted, hermitization does not change the norm, so $\|\sum_{i=1}^{n} \varepsilon_i A_i\| = \|\sum_{i=1}^{n} \varepsilon_i B_i\|$. To compute $\sigma^2 = \left\|\sum_{i=1}^{n} \varepsilon_i A_i^2\right\|$, compute A_i^2, sum up, and use Problem 1(c).

Hints for Problem 3

Consider a random vector $X \in \mathbb{R}^d$ that is uniformly distributed on the set of rows of A, i.e. X takes values A_k^{T} with probability $1/N$ for each $k = 1, \ldots, n$. Check that $\Sigma = \frac{1}{N} AA^{\mathsf{T}}$ and $\Sigma_n = \frac{1}{n} RR^{\mathsf{T}}$. Apply the general covariance estimation result (Lecture 23) to get $\|AA^{\mathsf{T}} - BB^{\mathsf{T}}\| \leq 0.1 \|AA^{\mathsf{T}}\|$. Finally, use Weyl's inequality (Lecture 20) and recall how the singular values of A and B are related to the eigenvalues of AA^{T} and BB^{T}.

Hints for Problem 4

(d) The matrix Σ_n has a zero diagonal entry by (c), but the corresponding entry of Σ equals 1 by (b). The result now follows from (a).