Assignment 2

Random Matrix Theory in Data Science and Statistics

(EN.553.796, Fall 2026)

Assigned: September 23, 2026   Due: October 8, 2026, 9:00AM

Solve all problems. Each problem is worth an equal amount towards your grade.

Submit solutions in LaTeX. Write in complete sentences. Include and justify all steps of your arguments, but avoid writing excessive discussion that is not contributing to your solution. You are welcome to include images if you think that will help explain your solutions.

Problem 1 (High-dimensional geometry)**.** This problem studies another interesting manifestation of the behaviors of volumes and angles in high dimension.

1. Let $\boldsymbol{v}_1, \dots, \boldsymbol{v}_N$ be independent uniform samples from $\{\pm 1/\sqrt{d}\}^d \subset \mathbb{S}^{d-1}$, where $d \geq 1$ and $N \geq 2$. For $0 < \epsilon < 1$, show that

$$\mathbb{P}\left[\max_{i \neq j} |\langle \boldsymbol{v}_i, \boldsymbol{v}_j \rangle| > \epsilon\right] \leq N^2 \exp(-d\epsilon^2/2).$$

   Deduce that, for any $\epsilon > 0$, there is some constant $c = c(\epsilon) > 0$ such that, for all d sufficiently large, there exist $N \geq \exp(cd)$ vectors $\boldsymbol{v}_1, \dots, \boldsymbol{v}_N \in \mathbb{S}^{d-1}$ with $\max_{i \neq j} |\langle \boldsymbol{v}_i, \boldsymbol{v}_j \rangle| \leq \epsilon$.

2. Suppose that $\boldsymbol{v}_1, \dots, \boldsymbol{v}_N \in \mathbb{S}^{d-1}$ have $\max_{i \neq j} |\langle \boldsymbol{v}_i, \boldsymbol{v}_j \rangle| \leq \epsilon$ for some $\epsilon \in (0, 1)$. Show that there is a constant $C = C(\epsilon) > 0$ such that $N \leq \exp(Cd)$.

   (Hint: Make an argument involving volumes and packings.)

3. Given $\boldsymbol{v}_1, \dots, \boldsymbol{v}_N \in \mathbb{S}^{d-1}$, define the *Gram matrix* $\boldsymbol{M} \in \mathbb{R}^{N \times N}_{\mathsf{sym}}$ to have entries $M_{ij} = \langle \boldsymbol{v}_i, \boldsymbol{v}_j \rangle$. Note that $M_{ii} = \|\boldsymbol{v}_i\|^2 = 1$, while in the setting of Part 1 $\boldsymbol{M}$ can be exponentially large in d and have every off-diagonal entry at most any given $\epsilon > 0$ in magnitude. Show that, nonetheless,

$$\lambda_1(\boldsymbol{M}) \geq \frac{N}{d},$$

   so $\boldsymbol{M}$ is not close to the identity matrix in terms of its eigenvalues once $N \gg d$.

Problem 2 (Haar measures)**.** We have repeatedly seen various notions of orthogonal invariance and the Haar measures on different sets (the unique ones invariant under orthogonal transformations) in class. This problem will ask you to work out some of their properties.

1. Let $\boldsymbol{U} \sim \mathsf{Haar}(\mathsf{Stief}(m, d))$ for some $1 \le d \le m$, and let $\boldsymbol{x} \in \mathbb{S}^{m-1} \subset \mathbb{R}^m$ be deterministic. Let $\boldsymbol{P} := \boldsymbol{U}\boldsymbol{U}^\top \in \mathbb{R}^{m \times m}_{\mathsf{sym}}$ and note that this is a projection matrix. It has the law of the random projection matrix that $\frac{1}{m}\boldsymbol{G}^\top \boldsymbol{G}$ is close to for $\boldsymbol{G} \sim \mathcal{N}(0,1)^{\otimes d \times m}$ when $m \gg d$, as we discussed in class. Let $\boldsymbol{g} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_m)$. Show that
$$\mathsf{Law}(\|\boldsymbol{P}\boldsymbol{x}\|^2) = \mathsf{Law}\left(\frac{g_1^2 + \cdots + g_d^2}{g_1^2 + \cdots + g_d^2 + g_{d+1}^2 + \cdots + g_m^2}\right).$$
State and prove a concentration inequality expressing that, if d and m are both large, then $\|\boldsymbol{P}\boldsymbol{x}\|^2$ is close to $\frac{d}{m}$ with high probability. Try to make your probability bound have exponentially decaying tails.

   (**Hint:** Recall that $\boldsymbol{U}$ has the same law as $\boldsymbol{Q} \sim \mathsf{Haar}(\mathcal{O}(m))$ truncated to the first d columns. Express this fact in linear algebraic terms, writing $\mathsf{Law}(\boldsymbol{U}) = \mathsf{Law}(\boldsymbol{Q}\boldsymbol{U}^{(0)})$ for a suitable deterministic $\boldsymbol{U}^{(0)}$. Plug that into the expression $\|\boldsymbol{P}\boldsymbol{x}\|^2 = \|\boldsymbol{U}\boldsymbol{U}^\top \boldsymbol{x}\|^2$.)

2. Let $\boldsymbol{Q} \sim \mathsf{Haar}(\mathcal{O}(d))$. Calculate the entrywise moments
$$\mathbb{E}Q_{ij}, \quad \mathbb{E}Q_{ij}Q_{k\ell}, \quad \mathbb{E}Q_{ij}Q_{k\ell}Q_{mn}, \quad \mathbb{E}Q_{ij}Q_{k\ell}Q_{mn}Q_{rs},$$
in terms of d and the indices $i, j, k, \ell, m, n, r, s \in [d]$.

   (**Hint:** Use that the Haar measure is invariant, so $\mathsf{Law}(\boldsymbol{A}\boldsymbol{Q}\boldsymbol{B}) = \mathsf{Law}(\boldsymbol{Q})$ for all $\boldsymbol{A}, \boldsymbol{B} \in \mathcal{O}(d)$. Make some simple choices of $\boldsymbol{A}$ and $\boldsymbol{B}$, and also use the relation $\boldsymbol{Q}^\top\boldsymbol{Q} = \boldsymbol{I}_d$, expanded into polynomials of the entries of $\boldsymbol{Q}$. You should find that, for most choices of indices, these moments are zero.)

Problem 3 (Perturbation of eigenvalues)**.** In this problem, you will prove a perturbation bound for eigenvalues and use it to control averages of functions of eigenvalues. Recall that $\|\boldsymbol{A}\|_F = (\sum_{i,j=1}^d A_{ij}^2)^{1/2} = (\mathsf{Tr}(\boldsymbol{A}^2))^{1/2}$ is the *Frobenius norm* of a symmetric matrix $\boldsymbol{A} \in \mathbb{R}^{d \times d}_{\mathsf{sym}}$.

1. Let $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{d \times d}_{\mathsf{sym}}$. Show the perturbation inequality
$$\min_{\pi \text{ permutation of } [d]} \sum_{i=1}^d (\lambda_i(\boldsymbol{A}) - \lambda_{\pi(i)}(\boldsymbol{B}))^2 \le \|\boldsymbol{A} - \boldsymbol{B}\|_F^2.$$
You may use the following two facts:

   (a) *Birkhoff–von Neumann Theorem*: The set of doubly stochastic $d \times d$ matrices (i.e., $\boldsymbol{P} \in \mathbb{R}^{d \times d}$ such that $P_{ij} \ge 0$ for all $i, j \in [d]$, $\sum_j P_{ij} = 1$ for all $i \in [d]$, and $\sum_i P_{ij} = 1$ for all $j \in [d]$) is the convex hull of the set of the $d \times d$ permutation matrices (those $\boldsymbol{P}$ with exactly one 1 in each row and each column and all other entries 0, of which there are $d!$).

(b) A linear function over a convex polytope is maximized at one of its vertices.

(**Hint:** Write the spectral decomposition of $\boldsymbol{A}$ and $\boldsymbol{B}$. Consider the matrix of $\langle \boldsymbol{u}_i, \boldsymbol{v}_j \rangle^2$ for $\boldsymbol{u}_i$ eigenvectors of $\boldsymbol{A}$ and $\boldsymbol{v}_j$ eigenvectors of $\boldsymbol{B}$.)

2. Show that the minimum above is achieved by π the identity permutation, and therefore in fact we have

$$\sum_{i=1}^{d} (\lambda_i(\boldsymbol{A}) - \lambda_i(\boldsymbol{B}))^2 \leq \|\boldsymbol{A} - \boldsymbol{B}\|_F^2.$$

Recall that, in our notation, the eigenvalues are always sorted as $\lambda_1 \geq \cdots \geq \lambda_d$.

(**Hint:** This does not have to do with linear algebra; it is a general fact about minimizing $\sum_{i=1}^{d} (a_i - b_{\pi(i)})^2$ for sorted sequences $a_1 \geq \cdots \geq a_d$ and $b_1 \geq \cdots \geq b_d$.)

3. Let f be a smooth and compactly supported function. Show that there is a constant $K = K(f)$ depending only on f such that, for any $d \geq 1$ and $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{d \times d}_{\mathsf{sym}}$,

$$\left| \frac{1}{d} \sum_{i=1}^{d} f(\lambda_i(\boldsymbol{A})) - \frac{1}{d} \sum_{i=1}^{d} f(\lambda_i(\boldsymbol{B})) \right| \leq \frac{K}{\sqrt{d}} \|\boldsymbol{A} - \boldsymbol{B}\|_F.$$