Assignment 1

Random Matrix Theory in Data Science and Statistics

(EN.553.796, Fall 2024)

Assigned: September 12, 2024   Due: 12pm EST, September 30, 2024

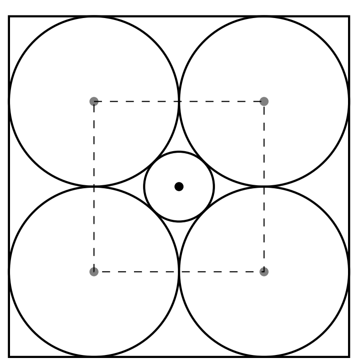
Solve any four out of the five problems. If you solve more, we will grade the first four solutions you include. Each problem is worth an equal amount towards your grade.

Submit solutions in LaTeX. Write in complete sentences. Include and justify all steps of your arguments, but try to avoid writing excessive explanation that is not contributing to our understanding your solution. You are welcome to include images if you think that will help explain your solutions (and you must include the plots we ask for if you choose to solve the numerical problem).

Problem 1 (High-dimensional geometry)**.** This problem will walk you through a few counter-intuitive properties of high-dimensional Euclidean space.

1. Consider the box B in $\mathbb{R}^d$ of side length 2, centered at the origin, with vertices at the points $(\pm 1, \dots, \pm 1)$. For each $\boldsymbol{s} \in \{\pm 1\}^d$, let $S_{\boldsymbol{s}}$ be the sphere of radius $\frac{1}{2}$ centered at the point $\frac{1}{2}\boldsymbol{s}$. These are 2^d spheres packed on a cubic lattice into the box B. Consider the sphere S' centered at the origin that is tangent to every $S_{\boldsymbol{s}}$. Find d_0 such that, if $d < d_0$, then S' is contained in B, but if $d \geq d_0$, then S' is *not* contained in B.

   For your reference, the case $d = 2$ looks as follows. The innermost circle is S', and the four around it are $S_{(\pm 1, \pm 1)}$. The larger outermost square is B; the smaller square in a dotted line is just for reference to show how the centers of the latter circles are arranged. Your task is to show that, in high dimension, the innermost circle is not contained in the outermost square (!).



2. Show that there is a constant $c > 0$ such that, for all $\epsilon \in (0, 1)$, for d sufficiently large, there are at least $N = \exp(c\epsilon^2 d)$ unit vectors $\boldsymbol{v}_1, \dots, \boldsymbol{v}_N$ in $\mathbb{R}^d$ such that $|\langle \boldsymbol{v}_i, \boldsymbol{v}_j \rangle| \le \epsilon$ for all $1 \le i < j \le N$ (i.e., such that the $\boldsymbol{v}_i$ are almost orthogonal). You may look up and use Hoeffding's inequality.

   (Hint: Consider random vectors. Choose a convenient distribution to work with. Note, though, that the question is not making a probabilistic statement.)

3. Consider the shape $\Delta^d \subset \mathbb{R}^d$ that is the convex hull of the points $\boldsymbol{0}, \boldsymbol{e}_1, \boldsymbol{e}_1 + \boldsymbol{e}_2, \dots, \boldsymbol{e}_1 + \boldsymbol{e}_2 + \cdots + \boldsymbol{e}_d$ (the origin plus the "partial sums" of the standard basis vectors). This is a *simplex* or high-dimensional tetrahedron, though not an equilateral one: edges have lengths varying among $1 = \sqrt{1}, \sqrt{2}, \dots, \sqrt{d}$. Compute the volume of Δ^d. What is the side length of a cube in $\mathbb{R}^d$ with the same volume? (Give an asymptotic approximation as $d \to \infty$.)

   (Hint: For the volume computation, consider gluing several copies of Δ^d together to tile a more familiar object.)

Problem 2 (Johnson-Lindenstrauss and some linear algebra)**.** In this problem, you will prove a lower bound on the dimension required for embedding a particular point cloud that almost matches the Johnson-Lindenstrauss lemma. Along the way, you will see some linear algebra that you might not have been introduced to before.

1. Let $\lambda_1, \dots, \lambda_n \ge 0$. Show that

$$\|\lambda\|_0 := \#\{i : \lambda_i \neq 0\} \ge \frac{(\sum_{i=1}^n \lambda_i)^2}{\sum_{i=1}^n \lambda_i^2} = \frac{\|\lambda\|_1^2}{\|\lambda\|_2^2}. \tag{1}$$

   Reinterpret this as a relationship between the rank, trace, and Frobenius norm of a positive semidefinite matrix.

2. Suppose $\boldsymbol{X} \in \mathbb{R}^{n \times n}_{\mathsf{sym}}$ has $\boldsymbol{X} \succeq \boldsymbol{0}$, $X_{ii} = 1$ for all $i \in [n]$, and $|X_{ij}| \le 1/\sqrt{n}$ for all $i \neq j$. Show that $\mathsf{rank}(\boldsymbol{X}) \ge n/2$.

3. For $k \ge 1$ and $\boldsymbol{X} \in \mathbb{R}^{n \times n}_{\mathsf{sym}}$ with $\boldsymbol{X} \succeq \boldsymbol{0}$, write $\boldsymbol{X}^{\odot k}$ for the matrix that has entries $(\boldsymbol{X}^{\odot k})_{ij} = X_{ij}^k$, i.e., for the entrywise kth power of $\boldsymbol{X}$ (note that $\boldsymbol{X}^{\odot k} \neq \boldsymbol{X}^k$). Show that $\boldsymbol{X}^{\odot k} \succeq 0$, and that $\mathsf{rank}(\boldsymbol{X}^{\odot k}) \le \binom{\mathsf{rank}(\boldsymbol{X})+k}{k}$.

   (Hint: View $\boldsymbol{X}$ as a Gram matrix, $X_{ij} = \langle \boldsymbol{v}_i, \boldsymbol{v}_j \rangle$, and write $\boldsymbol{X}^{\odot k}$ in the same way. If you do this at all, the first part will follow (be sure to explain why). If you do it carefully, the second part will follow as well.)

4. Show that there are constants $c, \epsilon_0 > 0$ such that the following holds. For all $0 < \epsilon < \epsilon_0$ (i.e., ϵ sufficiently small), there exists $n_0 = n_0(\epsilon)$ such that, if $n \ge n_0(\epsilon)$ (i.e., n sufficiently large depending on ϵ), then there exists no pairwise ϵ-faithful embedding (that is, one preserving pairwise distances up to a factor of $1 \pm \epsilon$, as in the Johnson-Lindenstrauss lemma) of $\boldsymbol{0}, \boldsymbol{e}_1, \dots, \boldsymbol{e}_n \in \mathbb{R}^n$ into fewer than $\frac{c}{\log(1/\epsilon)} \cdot \frac{\log n}{\epsilon^2}$ dimensions.

Thus, the dimension of the embedding provided by the Johnson-Lindenstrauss lemma for these points is tight up to a factor of $\log(1/\epsilon)$.

(**Hint:** Form the correlation matrix of the embeddings of the $\boldsymbol{e}_i$. Raise it to a large enough entrywise power that Part 2 applies. Compare with Part 3.)

Problem 3 (Properties of Gaussian measure)**.** The Gaussian measure is the most important one in probability theory, if not all of mathematics. Here you will derive some of its algebraic properties that we will use in class and future assignments.

1. Let $\boldsymbol{g} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma})$ for some $\boldsymbol{\Sigma} \in \mathbb{R}^{d\times d}_{\mathsf{sym}}$ with $\boldsymbol{\Sigma} \succeq \boldsymbol{0}$ (i.e., $\boldsymbol{\Sigma}$ is positive semidefinite). Prove that, for any smooth function $f : \mathbb{R}^d \to \mathbb{R}$ with $|f(\boldsymbol{x})| \leq C\|\boldsymbol{x}\|^K$ for some $C, K > 0$ and all $\boldsymbol{x} \in \mathbb{R}^d$, we have

$$\mathbb{E}[g_i f(\boldsymbol{g})] = \sum_{j=1}^{d} \Sigma_{ij}\mathbb{E}[\partial_j f(\boldsymbol{g})] = (\boldsymbol{\Sigma}\,\mathbb{E}[\nabla f(\boldsymbol{g})])_i \tag{2}$$

   where $\partial_i f$ is the partial derivative with respect to the ith argument and ∇ is the gradient (the second equality is just by the definition of gradient).

   (**Hint:** Integrate by parts. You might also find it useful to first treat the case $\boldsymbol{\Sigma} = \boldsymbol{I}_d$, and then to observe that $\boldsymbol{g}$ and $\boldsymbol{\Sigma}^{1/2}\boldsymbol{h}$ for $\boldsymbol{h} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_d)$ have the same law.)

2. Let $g \sim \mathcal{N}(0, 1)$ (a Gaussian scalar, not a vector). Prove that, for $k \geq 1$, $\mathbb{E}g^{2k-1} = 0$ and $\mathbb{E}g^{2k} = \prod_{i=1}^{k}(2i-1)$.

3. Let $\boldsymbol{g} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma})$ as in Part 1, and let $1 \leq i_1 < \cdots < i_k \leq d$. Let $\mathcal{M}$ be the set of all *matchings* of the set $I = \{i_1, \dots, i_k\}$: a matching is a set of disjoint pairs $\{i_a, i_b\}$ whose union is I. For example, the three matchings of $\{1, 2, 3, 4\}$ are $\{\{1, 2\}, \{3, 4\}\}$, $\{\{1, 3\}, \{2, 4\}\}$, and $\{\{1, 4\}, \{2, 3\}\}$. Prove that

$$\mathbb{E}\left[\prod_{a=1}^{k} g_{i_a}\right] = \sum_{M\in\mathcal{M}} \prod_{\{a,b\}\in M} \Sigma_{ab}. \tag{3}$$

   (For example, one case of the claim is that $\mathbb{E}g_1g_2g_3g_4 = \Sigma_{12}\Sigma_{34} + \Sigma_{13}\Sigma_{24} + \Sigma_{14}\Sigma_{23}$.) Generalize this to allow for repetitions among the $i_1, \dots, i_k$. Try to be precise. Explain why Part 2 is a special case of this latter generalization.

   (**Hint:** Induction.)

4. Suppose that μ is a probability measure on $\mathbb{R}$ with a smooth density $\rho(x)$ that has $\rho(x) > 0$ for all $x \in \mathbb{R}$. Suppose that, for any smooth and compactly supported $f : \mathbb{R} \to \mathbb{R}$, $\mathbb{E}_{g\sim\mu}[gf(g)] = \mathbb{E}_{g\sim\mu}[f'(g)]$. Show that $\mu = \mathcal{N}(0, 1)$ (i.e., that the converse of the $d = 1$ case of Part 1 holds).

   (**Hint:** Write the expectations as integrals involving ρ. Integrate by parts.)

Problem 4 (Random determinants)**.** Let $\boldsymbol{G} \in \mathbb{R}^{d\times d}$ have i.i.d. entries distributed as $\mathcal{N}(0,1)$ (with no symmetry constraint). You will study the random variable $|\det(\boldsymbol{G})|$, one interpretation of which is the volume of the random parallelopiped generated by d independent standard Gaussian vectors.

1. Show that $\mathbb{E}[|\det(\boldsymbol{G})|] \leq \sqrt{d!}$. (Do something much simpler than using Part 2 below.)

2. Let $\boldsymbol{g}^{(k)} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_k)$ for $k = 1, \dots, d$, drawn independently (that is, $\boldsymbol{g}^{(k)}$ is a standard Gaussian vector in $\mathbb{R}^k$). Show that $|\det(\boldsymbol{G})|$ has the same law as $\prod_{k=1}^d \|\boldsymbol{g}^{(k)}\|$.

   (**Hint:** Consider the QR decomposition of $\boldsymbol{G}$.)

3. Show that, for a constant $c > 0$, $\mathbb{E}[|\det(\boldsymbol{G})|] \geq \frac{c}{d}\sqrt{d!}$, almost matching the upper bound of Part 1.

   (**Hint:** Prove that $\sqrt{x} \geq \frac{1}{2}(1+x-(x-1)^2)$ for all $x \geq 0$. Apply this with $x := \|\boldsymbol{g}^{(k)}\|^2/k$.)

4. Show that, for all square $\boldsymbol{G}$ (not random), $|\det(\boldsymbol{G})| = \prod_{i=1}^d \sigma_i(\boldsymbol{G})$, where σ_i are the singular values. By computing $\mathbb{E}|\det(\boldsymbol{G})|^2$ (for random Gaussian $\boldsymbol{G}$ again now), make a heuristic but intuitively justified prediction for the limiting value of

$$\lim_{d\to\infty} \mathbb{E}\left[\frac{1}{d}\sum_{i=1}^d \log\left(\lambda_i\left(\frac{1}{d}\boldsymbol{G}^\top \boldsymbol{G}\right)\right)\right] = ???. \tag{4}$$

   Optionally, if you don't mind skipping ahead a little, confirm that your prediction is compatible with the Marchenko-Pastur limit theorem.

   (**Hint:** Make a heuristic leap of the form $\mathbb{E}[\log(\cdots)] \approx \log(\mathbb{E}[\cdots])$. Don't be afraid. If you like, speculate about when you expect this to be accurate.)

Problem 5 (Numerical study of eigenvalue spacing)**.** This problem will concern the *Gaussian orthogonal ensemble (GOE)* that we will study in class soon. This distribution, denoted $\mathsf{GOE}(n)$, is the law of $\boldsymbol{W} \in \mathbb{R}^{n\times n}$ where $W_{ii} \sim \mathcal{N}(0,2)$ and $W_{ij} = W_{ji} \sim \mathcal{N}(0,1)$, all independently drawn for $1 \leq i \leq j \leq n$. Note that this $\boldsymbol{W}$ is almost surely a symmetric matrix, and thus has real eigenvalues $\lambda_1 \geq \cdots \geq \lambda_n$, whose distribution we will study.

Below, we write $\chi^2(d, \sigma^2)$ for the law of $\|\boldsymbol{g}\|^2 = g_1^2 + \cdots + g_d^2$ for $\boldsymbol{g} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_d)$ (the usual χ^2 distribution, but allowing for rescaling). Similarly, we write $\chi(d, \sigma^2)$ for the law of $\|\boldsymbol{g}\|$.

1. One small theoretical task: show that the *eigenvalue spacing* $\lambda_1 - \lambda_2 \geq 0$ when $\boldsymbol{W} \sim \mathsf{GOE}(2)$ (a 2×2 matrix) has the law $\chi(2, \sigma^2)$ for some σ^2 (calculate and give this value). Look up and write down the density of this distribution—you will need it later.

2. On the computer, sample $\boldsymbol{W} \sim \mathsf{GOE}(n)$ for a sequence of growing n. Go at least up to $n = 1000$. Plot histograms of the eigenvalues $\lambda_1 \geq \cdots \geq \lambda_n$ for a few growing n and make sure you observe convergence to a semicircle shape. Also plot λ_1 and λ_n versus n, taking the mean over several random trials for each value of n and including error bars. Make a prediction about the typical scaling of λ_1 and λ_n (each of the form $\mathbb{E}\lambda_i \sim a_i n^{b_i}$ for $a, b \in \mathbb{R}$), each supported by a convincing plot.

3. Now fix a large n, at least $n = 1000$ (the larger the better), and plot a histogram of the distribution of the *bulk spacing* $\lambda_{n/2} - \lambda_{n/2+1}$ over many independent draws of $\boldsymbol{W}$ (at least 2000). Come up with a procedure to try to find a good σ^2 to approximate this distribution by $\chi(2, \sigma^2)$ (i.e., to approximate the distribution of spacings for large n by a rescaling of the closed form distribution of spacings you found for $n = 2$ in Part 1). You can define a reasonable "loss function" of σ^2 and use any optimization library your language has to minimize it, for instance. Draw a plot to illustrate the quality of the fit.

4. Consider another distribution of $(\lambda_1, \dots, \lambda_n)$ where λ_i are chosen uniformly at random independently in the interval that you conjectured $[\lambda_n, \lambda_1]$ to scale like in Part 1 (that is, n independent draws from a distribution of the form $\mathsf{Unif}([a_n n^{b_n}, a_1 n^{b_1}])$). Repeat the spacing experiment: fix the same n as in Part 3, sample n independent numbers $\lambda_1, \dots, \lambda_n$ uniformly at random in the predicted interval, sort them to form $\widetilde{\lambda}_1 \leq \cdots \leq \widetilde{\lambda}_n$, and plot a histogram of the spacing $\widetilde{\lambda}_{n/2} - \widetilde{\lambda}_{n/2+1}$ over many independent trials of this procedure. Comment on the differences between the distribution of actual eigenvalue spacings and the distribution of spacings under this alternative model. What does this say about the structure of the eigenvalues? (Focus on the behavior of these distributions near zero.)

5. Download a list of the imaginary parts of the first 100 000 non-trivial zeros of the Riemann zeta function (the famous Riemann Hypothesis conjectures that the real parts of all such zeros are equal to $\frac{1}{2}$) from this website:

   https://www-users.cse.umn.edu/~odlyzko/zeta_tables/zeros1

   Calculate the differences between consecutive values (the distances between consecutive zeros along the imaginary axis.). Plot a histogram of the spacings. You should observe similar qualitative phenomena to before. Repeat the procedure you chose before to find σ^2 to fit the density of $\chi(2, \sigma^2)$ to this distribution. Find another d such that $\chi(d, \sigma^2)$ for some σ^2 achieves an exceptionally good fit. Illustrate the best choice of d (and σ^2) by plotting this density over the histogram of spacings of zeros.