Assignment 1

Random Matrix Theory in Data Science and Statistics

(EN.553.796, Fall 2025)

Assigned: September 8, 2025 Due: 11:59pm EST, September 22, 2025

Solve all three 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 (Subgaussianity)**.** We call a random variable *centered* if its expectation is zero. A centered random scalar $X \in \mathbb{R}$ is called σ^2-*subgaussian* if $\mathbb{E}\exp(\lambda X) \le \exp(\sigma^2\lambda^2/2)$ for all $\lambda \in \mathbb{R}$. A centered random vector $\boldsymbol{x} \in \mathbb{R}^d$ is called the same if, for all $\boldsymbol{u} \in \mathbb{R}^d$ with $\|\boldsymbol{u}\| = 1$, $\langle \boldsymbol{u}, \boldsymbol{x}\rangle$ is a σ^2-subgaussian random scalar.[1]

1. Show that if X is a centered random scalar that is σ^2-subgaussian, then aX is $a^2\sigma^2$-subgaussian, and that if $X_1, \dots, X_n$ are independent and each X_i is σ_i^2-subgaussian, then $\sum_{i=1}^n X_i$ is $(\sum_{i=1}^n \sigma_i^2)$-subgaussian. Conclude that a random vector with independent entries that are each centered and σ^2-subgaussian is itself σ^2-subgaussian.

2. Show that the following three conditions on a centered random scalar X are equivalent:

   (a) There exists $\sigma^2 > 0$ such that X is σ^2-subgaussian.

   (b) There exists $c > 0$ such that, for all $t > 0$, $\mathbb{P}[|X| \ge t] \le 2\exp(-ct^2)$.

   (c) There exists $\lambda > 0$ such that $\mathbb{E}\exp(\lambda X^2) < \infty$.

3. For X a centered subgaussian scalar, define

$$\|X\|_{\mathsf{sg}} := \inf\{t > 0 : \mathbb{E}\exp(X^2/t^2) \le 2\}.$$

   Note that this is well-defined by Part 2. Show that $\|\cdot\|_{\mathsf{sg}}$ is a norm on the space of centered subgaussian random variables (over a given probability space).

   (**Hint:** Prove and use that $(X+Y)^2 \le (1+a)X^2 + (1+\frac{1}{a})Y^2$ for any $a > 0$ (you might be more familiar with the case $a = 1$), together with Hölder's inequality.)

[1] It is also sometimes useful to say that an X not necessarily centered is σ^2-subgaussian if $X - \mathbb{E}X$ satisfies our condition, but we will not do that in this assignment.

4. Suppose that $\boldsymbol{x} \in \mathbb{R}^d$ is a centered σ^2-subgaussian random vector. Show that there are constants $c, C > 0$ such that

$$\mathbb{P}[\|\boldsymbol{x}\| > C\sigma\sqrt{d}] \le \exp(-cd).$$

That is, a high-dimensional $O(1)$-subgaussian vector has norm $O(\sqrt{d})$ with high probability.

(**Hint:** Use ϵ-nets and adapt our discretization of the operator norm from lecture to apply instead to the Euclidean norm $\|\boldsymbol{x}\|$ of a vector.)

5. Let $\boldsymbol{X} \in \mathbb{R}^{d\times d}_{\mathsf{sym}}$ be a centered random symmetric matrix. Write $\mathsf{vec}(\boldsymbol{X}) \in \mathbb{R}^{d(d+1)/2}$ for the (symmetric) *vectorization* of $\boldsymbol{X}$, the vector formed by concatenating the entries on and above the diagonal of $\boldsymbol{X}$ (choose any order of entries you like). Show that there are constants $c, C > 0$ such that, if $\mathsf{vec}(\boldsymbol{X})$ is a σ^2-subgaussian random vector, then

$$\mathbb{P}[\|\boldsymbol{X}\| > C\sigma\sqrt{d}] \le \exp(-cd).$$

In particular, this implies that a high-dimensional random symmetric matrix with independent $O(1)$-subgaussian entries has operator norm $O(\sqrt{d})$ with high probability.

(**Hint:** Use ϵ-nets again.)

Problem 2 (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 $f(\boldsymbol{x}) \le C\|\boldsymbol{x}\|^K$ for some $C, K > 0$, we have

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

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 $k \ge 1$, $\mathbb{E}g^{2k-1} = 0$ and $\mathbb{E}g^{2k} = \prod_{i=1}^{k}(2i-1)$.

3. In the setting of Part 2, $\mathbb{E}g^{2k} \le (2k)^k$. Show that, in fact, there exists a constant $C > 0$ such that, if X is σ^2-subgaussian (see Problem 1), then $\mathbb{E}|X|^k \le (C\sigma^2 k)^{k/2}$ for all $k \ge 1$.

(**Hint:** Use the identity that, if $Y \ge 0$, then $\mathbb{E}Y = \int_0^\infty \mathbb{P}[Y > t]\,dt$. If you have not seen this before, sketch a proof of it as well.)

4. Let $g \sim \mathcal{N}(0, \Sigma)$ as in Part 1, and let $1 \le i_1 < \cdots < i_k \le d$. Let $\mathcal{M}$ be the set of all *matchings* of the set $I = \{i_1, \ldots, i_k\}$: 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

$$\mathbb{E}\left[\prod_{a=1}^{k} g_{i_a}\right] = \sum_{M \in \mathcal{M}} \prod_{\{a,b\} \in M} \Sigma_{ab}. \tag{2}$$

(For example, one case of the claim is that $\mathbb{E} g_1 g_2 g_3 g_4 = \Sigma_{12}\Sigma_{34} + \Sigma_{13}\Sigma_{24} + \Sigma_{14}\Sigma_{23}$.) Generalize this to allow for repetitions among the $i_1, \ldots, i_k$. Try to be precise. Explain why Part 2 is a special case of this latter generalization.

(**HINT:** Induction.)

Problem 3 (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 $\lambda_1, \ldots, \lambda_n \ge 0$. Show that

$$\|\lambda\|_0 := \#\{i : \lambda_i \neq 0\} \ge \frac{(\sum_{i=1}^n \lambda_i)^2}{\sum_{i=1}^n \lambda_i^2} = \frac{\|\lambda\|_1^2}{\|\lambda\|_2^2}. \tag{3}$$

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

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

3. Let $v \in \mathbb{R}^r$. Consider the *monomials* in the entries of v, polynomials of the form $v_1^{a_1} \cdots v_r^{a_r}$ for $a_i \ge 0$. The *degree* of such a monomial is $a_1 + \cdots + a_r$. Show that the number of different monomials of degree exactly k is $\binom{r+k-1}{k}$, and show the bound

$$\binom{r+k-1}{k} \le \left(e \cdot \frac{r+k-1}{k}\right)^k.$$

(**HINT:** For the bound, first use the series expansion of $\exp(k)$ to show that $k! \ge (k/e)^k$.)

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

(**HINT:** View X as a Gram matrix, $X_{ij} = \langle v_i, v_j \rangle$ for $v_i \in \mathbb{R}^{\mathsf{rank}(X)}$, and write $X^{\odot k}$ in the same way. If you do this at all, the first part will follow, but be sure to explain why. If you do it carefully, the second part will follow as well. To do that, observe and find a way to use the formal resemblance with the first result of Part 3.)

5. Let $\boldsymbol{y}_0 := \mathbf{0}$ and $\boldsymbol{y}_1 := \boldsymbol{e}_1, \ldots, \boldsymbol{y}_n := \boldsymbol{e}_n$, a point cloud of $n+1$ vectors in $\mathbb{R}^n$ ($\boldsymbol{e}_i$ is the ith standard basis vector, with 1 in the ith coordinate and 0 in all others). Show that there are constants $c, \epsilon_0 > 0$ such that, for all $\epsilon < \epsilon_0$ and $d \leq \frac{c \log n}{\epsilon^2 \log(1/\epsilon)}$, there do not exist $\boldsymbol{z}_0, \boldsymbol{z}_1, \ldots, \boldsymbol{z}_n \in \mathbb{R}^d$ such that

$$(1-\epsilon)\|\boldsymbol{y}_i - \boldsymbol{y}_j\|^2 \leq \|\boldsymbol{z}_i - \boldsymbol{z}_j\|^2 \leq (1+\epsilon)\|\boldsymbol{y}_i - \boldsymbol{y}_j\|^2 \text{ for all } i, j \in \{0, 1, \ldots, n\}.$$

Conclude that the dimension of the embedding satisfying the above property that is described by the Johnson-Lindenstrauss lemma for these points is tight up to a factor of $O(\log(1/\epsilon))$.

(**HINT:** Form the matrix $\boldsymbol{X} \in \mathbb{R}^{n \times n}_{\mathsf{sym}}$ with

$$X_{ij} := \frac{\langle \boldsymbol{z}_i - \boldsymbol{z}_0, \boldsymbol{z}_j - \boldsymbol{z}_0 \rangle}{\|\boldsymbol{z}_i - \boldsymbol{z}_0\| \cdot \|\boldsymbol{z}_j - \boldsymbol{z}_0\|}$$

for each $i, j \in \{1, \ldots, n\}$ (omitting the zero index). Bound the off-diagonal entries of $\boldsymbol{X}$. Raise $\boldsymbol{X}$ to a large enough entrywise power to derive a contradiction between Parts 2 and 4 of the problem. The bound from Part 3 will be useful.)