# CS 229, Spring 2023
# Problem Set 0: Linear Algebra, Multivariable Calculus, and Probability

---

**Ungraded: optionally due Wednesday, April 12 at 11:59 pm on Gradescope.**

**Notes:**
(1) These questions require thought, but do not require long answers. Please be as concise as possible.
(2) If you have a question about this homework, we encourage you to post your question on our Ed at `https://edstem.org/us/courses/37893/discussion/`.
(3) If you missed the first lecture or are unfamiliar with the collaboration or honor code policy, please read the policy before you start.
(4) This specific homework is ***not graded***, but we encourage you to solve each of the problems to brush up on your linear algebra and probability. Some of them may even be useful for subsequent problem sets. It also serves as your introduction to using Gradescope for submissions. We strongly suggest you use LaTeX to write your problem set solutions (not only is it helpful for this class, but it is a good skill to learn). However, if you are scanning your document by cellphone, please use a scanning app such as CamScanner. There will not be any late days allowed for this particular assignment.

**Honor code:** We strongly encourage students to form study groups. Students may discuss and work on homework problems in groups. However, each student must write down the solution independently, and without referring to written notes from the joint session. Each student must understand the solution well enough in order to reconstruct it by him/herself. It is an honor code violation to copy, refer to, or look at written or code solutions from a previous year, including but not limited to: official solutions from a previous year, solutions posted online, and solutions you or someone else may have written up in a previous year. Furthermore, it is an honor code violation to post your assignment solutions online, such as on a public git repo. We run plagiarism-detection software on your code against past solutions as well as student submissions from previous years. Please take the time to familiarize yourself with the Stanford Honor Code [1] and the Stanford Honor Code[2] as it pertains to CS courses.

[1] `https://communitystandards.stanford.edu/policies-and-guidance/honor-code`
[2] `web.stanford.edu/class/archive/cs/cs106b/cs106b.1164/handouts/honor-code.pdf`

1. **[0 points] Gradients and Hessians**

   Recall that a matrix $A \in \mathbb{R}^{n \times n}$ is *symmetric* if $A^T = A$, that is, $A_{ij} = A_{ji}$ for all $i, j$. Also recall the gradient $\nabla f(x)$ of a function $f : \mathbb{R}^n \to \mathbb{R}$, which is the $n$-vector of partial derivatives

   $$\nabla f(x) = \begin{bmatrix} \frac{\partial}{\partial x_1} f(x) \\ \vdots \\ \frac{\partial}{\partial x_n} f(x) \end{bmatrix} \quad \text{where} \quad x = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}.$$

   The hessian $\nabla^2 f(x)$ of a function $f : \mathbb{R}^n \to \mathbb{R}$ is the $n \times n$ symmetric matrix of twice partial derivatives,

   $$\nabla^2 f(x) = \begin{bmatrix} \frac{\partial^2}{\partial x_1^2} f(x) & \frac{\partial^2}{\partial x_1 \partial x_2} f(x) & \cdots & \frac{\partial^2}{\partial x_1 \partial x_n} f(x) \\ \frac{\partial^2}{\partial x_2 \partial x_1} f(x) & \frac{\partial^2}{\partial x_2^2} f(x) & \cdots & \frac{\partial^2}{\partial x_2 \partial x_n} f(x) \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2}{\partial x_n \partial x_1} f(x) & \frac{\partial^2}{\partial x_n \partial x_2} f(x) & \cdots & \frac{\partial^2}{\partial x_n^2} f(x) \end{bmatrix}.$$

   (a) Let $f(x) = \frac{1}{2} x^T A x + b^T x$, where $A$ is a symmetric matrix and $b \in \mathbb{R}^n$ is a vector. What is $\nabla f(x)$?

   **Answer:**

   $$\nabla f(x) = Ax + b$$

   (b) Let $f(x) = g(h(x))$, where $g : \mathbb{R} \to \mathbb{R}$ is differentiable and $h : \mathbb{R}^n \to \mathbb{R}$ is differentiable. What is $\nabla f(x)$?

   **Answer:**

   $$\nabla f(x) = g'(h(x)) \times \nabla h(x)$$

   (c) Let $f(x) = \frac{1}{2} x^T A x + b^T x$, where $A$ is symmetric and $b \in \mathbb{R}^n$ is a vector. What is $\nabla^2 f(x)$?

   **Answer:**

   $$\nabla^2 f(x) = A$$

   (d) Let $f(x) = g(a^T x)$, where $g : \mathbb{R} \to \mathbb{R}$ is continuously differentiable and $a \in \mathbb{R}^n$ is a vector. What are $\nabla f(x)$ and $\nabla^2 f(x)$? (*Hint:* your expression for $\nabla^2 f(x)$ may have as few as 11 symbols, including $'$ and parentheses.)

   **Answer:**

   $$\nabla f(x) = g'(a^T x) \times a$$

   $$\nabla^2 f(x) = g''(a^T x) \times a a^T$$

2. **[0 points] Positive definite matrices**

A matrix $A \in \mathbb{R}^{n \times n}$ is *positive semi-definite* (PSD), denoted $A \succeq 0$, if $A = A^T$ and $x^T Ax \geq 0$ for all $x \in \mathbb{R}^n$. A matrix $A$ is *positive definite*, denoted $A \succ 0$, if $A = A^T$ and $x^T Ax > 0$ for all $x \neq 0$, that is, all non-zero vectors $x$. The simplest example of a positive definite matrix is the identity $I$ (the diagonal matrix with 1s on the diagonal and 0s elsewhere), which satisfies $x^T Ix = \|x\|_2^2 = \sum_{i=1}^n x_i^2$.

(a) Let $z \in \mathbb{R}^n$ be an $n$-vector. Show that $A = zz^T$ is positive semidefinite.

**Answer:** Firstly observe $A = A^T$ as $A^T = (zz^T)^T = zz^T = A$. Secondly observe $x^T Ax \geq 0$ for all $x \in \mathbb{R}^n$ as $x^T Ax = x^T zz^T x = (z^T x)^T (z^T x) = \hat{x}^T \hat{x} = \sum_i^n \hat{x}_i^2$ and $\sum_i^n \hat{x}_i^2 \geq 0$ for all $\hat{x} \in \mathbb{R}^n$. Thus, $A$ is positive semidefinite.

(b) Let $z \in \mathbb{R}^n$ be a *non-zero* $n$-vector. Let $A = zz^T$. What is the null-space of $A$? What is the rank of $A$?

**Answer:** Since $A = zz^T$, all of the columns/rows of $A$ will be linear multiples of one another. Thus, the rank of $A$ is always 1. The null-space of $A$ will be the set of all vectors orthogonal to $z$, and will have dimension $n - 1$ by the Rank-Nullity Theorem.

(c) Let $A \in \mathbb{R}^{n \times n}$ be positive semidefinite and $B \in \mathbb{R}^{m \times n}$ be arbitrary, where $m, n \in \mathbb{N}$. Is $BAB^T$ PSD? If so, prove it. If not, give a counterexample with explicit $A, B$.

**Answer:** $BAB^T$ is PSD. To see why, observe that $x^T BAB^T x = (B^T x)^T A(B^T x) = \hat{x}^T A\hat{x} \geq 0$ for any $x \in \mathbb{R}^m$ and associated $\hat{x} \in mathbbR^n$ (with the last inequality following from the initial assumption that $A$ is PSD). Therefore, $BAB^T$ is PSD.

3. **[0 points] Eigenvectors, eigenvalues, and the spectral theorem**

The eigenvalues of an $n \times n$ matrix $A \in \mathbb{R}^{n \times n}$ are the roots of the characteristic polynomial $p_A(\lambda) = \det(\lambda I - A)$, which may (in general) be complex. They are also defined as the values $\lambda \in \mathbb{C}$ for which there exists a vector $x \in \mathbb{C}^n$ such that $Ax = \lambda x$. We call such a pair $(x, \lambda)$ an *eigenvector, eigenvalue* pair. In this question, we use the notation $\text{diag}(\lambda_1, \ldots, \lambda_n)$ to denote the diagonal matrix with diagonal entries $\lambda_1, \ldots, \lambda_n$, that is,

$$\text{diag}(\lambda_1, \ldots, \lambda_n) = \begin{bmatrix} \lambda_1 & 0 & 0 & \cdots & 0 \\ 0 & \lambda_2 & 0 & \cdots & 0 \\ 0 & 0 & \lambda_3 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \lambda_n \end{bmatrix}.$$

(a) Suppose that the matrix $A \in \mathbb{R}^{n \times n}$ is diagonalizable, that is, $A = T\Lambda T^{-1}$ for an invertible matrix $T \in \mathbb{R}^{n \times n}$, where $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_n)$ is diagonal. Use the notation $t^{(i)}$ for the columns of $T$, so that $T = [t^{(1)} \ \cdots \ t^{(n)}]$, where $t^{(i)} \in \mathbb{R}^n$. Show that $At^{(i)} = \lambda_i t^{(i)}$, so that the eigenvalues/eigenvector pairs of $A$ are $(t^{(i)}, \lambda_i)$.

**Answer:**

$$\begin{aligned} A &= T\Lambda T^{-1} \\ AT &= T\Lambda T^{-1} T \\ AT &= T\Lambda \\ [At^{(1)} \cdots At^{(n)}] &= [t^{(1)T}\lambda_1 \cdots t^{(n)T}\lambda_n] \\ [At^{(1)} \cdots At^{(n)}] &= [\lambda_1 t^{(1)} \cdots \lambda_n t^{(n)}] \\ At^{(i)} &= \lambda_i t^{(i)} \text{ for } i \in [1, n] \end{aligned}$$

A matrix $U \in \mathbb{R}^{n \times n}$ is orthogonal if $U^T U = I$. The spectral theorem, perhaps one of the most important theorems in linear algebra, states that if $A \in \mathbb{R}^{n \times n}$ is symmetric, that is, $A = A^T$, then $A$ is *diagonalizable by a real orthogonal matrix*. That is, there are a diagonal matrix $\Lambda \in \mathbb{R}^{n \times n}$ and orthogonal matrix $U \in \mathbb{R}^{n \times n}$ such that $U^T AU = \Lambda$, or, equivalently,

$$A = U\Lambda U^T.$$

Let $\lambda_i = \lambda_i(A)$ denote the $i$th eigenvalue of $A$.

(b) Let $A$ be symmetric. Show that if $U = [u^{(1)} \ \cdots \ u^{(n)}]$ is orthogonal, where $u^{(i)} \in \mathbb{R}^n$ and $A = U\Lambda U^T$, then $u^{(i)}$ is an eigenvector of $A$ and $Au^{(i)} = \lambda_i u^{(i)}$, where $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_n)$.

**Answer:** Since $U$ is orthogonal, it follows that $U^{-1} = U^T$ ; hence, $A = U\Delta U^T$ is equivalent to $A = U\Delta U^{-1}$. This form is equivalent to the case $A = T\Delta T^{-1}$ from part (a), and the rest of the proof follows in analogous fashion.

(c) Show that if $A$ is PSD, then $\lambda_i(A) \geq 0$ for each $i$.

**Answer:** If $A$ is PSD, $x^T Ax \geq 0$ for any vectors $x$. Substituting $A = U\Lambda U^T$, it follows that $x^T U\Lambda U^T x \geq 0$ for any vectors $x$. Re-formulating, we observe $x^T U\Lambda U^T x = \hat{x}^T \Lambda \hat{x} = \sum_{i=1}^n \lambda_i \hat{x}_i^2 \geq 0$ for any vector $\hat{x}$. Since $\hat{x}_i^2 \geq 0$ for all $i$ in the previous inequality, for said inequality to be true, it must be the case that $\lambda_i(A) \geq 0$ for each $i$.

4. **[0 points] Probability and multivariate Gaussians**

   Suppose $X = (X_1, ..X_n)$ is sampled from a multivariate Gaussian distribution with mean $\mu$ in $\mathbb{R}^n$ and covariance $\Sigma$ in $S_+^n$ (i.e. $\Sigma$ is positive semidefinite). This is commonly also written as $X \sim \mathcal{N}(\mu, \Sigma)$.

   (a) Describe the random variable $Y = X_1 + X_2 + \ldots + X_n$. What is the mean and variance? Is this a well known distribution, and if so, which?

   **Answer:** $Y$ has a mean of $||\mu||_1$ and a variance of $||\Sigma||_1$. $Y$ is indeed represented by a well-known distribution - i.e., the Gaussian distribution.

   (b) Now, further suppose that $\Sigma$ is invertible. Find $\mathbb{E}[X^T\Sigma^{-1}X]$. (Hint: use the property of trace that $x^TAx = \text{tr}(x^TAx)$).

   **Answer:**

$$\begin{aligned}
\mathbb{E}[X^T\Sigma^{-1}X] &= \mathbb{E}[\mathsf{tr}(X^T\Sigma^{-1}X)] \\
&= \mathbb{E}[\mathsf{tr}(\Sigma^{-1}XX^T)] \\
&= \mathsf{tr}(\Sigma^{-1}\mathbb{E}[XX^T]) \\
&= \mathsf{tr}(\Sigma^{-1}(\Sigma + \mu\mu^T)) \\
&= \mathsf{tr}(I + \Sigma^{-1}\mu\mu^T)) \\
&= n + \mu^T\Sigma^{-1}\mu
\end{aligned}$$