*Fall 2022* *Instructor: Ashwin Pananjady*

# 7750: Mathematical Foundations of Machine Learning

Linear algebra and probability for data analysis

## Homework 0

*Released: Aug 21* *Due: Aug 26, 11:59pm ET*

**Note:** This homework will not be graded, but you are expected to submit solutions in the format detailed in the syllabus. You will receive 2/50 points alloted to your HW score as bonus just for a submission that follows all the instructions.

**Objective.** To solve some problems in linear algebra, calculus, and probability, and to set up and use Jupyter notebooks. The problems are deliberately dry with keywords, so that you can quickly find resources to help fill any gaps in your knowledge. Do not be worried if you cannot solve some problems; the primary point is to give you the opportunity to go back and refresh. We will post detailed solutions to these problems after the deadline, and also discuss them briefly in class.

**Resources.** The following resources may be helpful to get up to speed if you feel yourself struggling with portions of this:

1. "Essence of linear algebra" playlist on Youtube by 3Blue1Brown.
2. A handbook of mathematics for ML, by Garrett Thomas `https://gwthomas.github.io/docs/math4ml.pdf`. Not all of this is really required, but it has a nice exposition of basics in calculus, linear algebra, and probability, and can be used as a handbook over the course of the semester.
3. Matrix cookbook `https://www.math.uwaterloo.ca/~hwolkowi/matrixcookbook.pdf`. Again, most things here will not be used, but this is a useful handbook.
4. Trefethen and Bau `http://people.maths.ox.ac.uk/~trefethen/text.html`. These contain some linear algebra notes that are relatively advanced. We will use it later on to discuss a nice geometric exposition of the singular value decomposition.
5. For Python, there are a variety of blogs that will be thrown up just by a Google search "Python for machine learning beginner". You should feel free to use whatever you find convenient to learn basic syntax. There will be demos later that will talk you through code snippets and the use of particular packages. Use Problem 3 for a recommended installation of Jupyter notebooks (if you don't have a working version already).

**Notation:** Capital boldface letters will be matrices, and small boldface letter will be vectors. Capital letters (not boldface) will be random variables (and sometimes random vectors), and small letters will typically be scalars. Dimensions of matrices and vectors will be specified when needed, but for the most part, you should be able to intuit these yourself (and doing this is a useful exercise). The set of real numbers is represented by the set $\mathbb{R}$, $d$-dimensional vectors by $\mathbb{R}^d$ and $n \times d$ matrices by $\mathbb{R}^{n \times d}$. The symbol $:=$ denotes a definition; the left-hand-size is defined to be the right-hand-side.

**Problem 1 (Calculus and linear algebra).**

(a) For a vector $\mathbf{w} \in \mathbb{R}^n$ and another vector $\mathbf{a} \in \mathbb{R}^n$, consider the function $f(\mathbf{w}) := f(w_1, \ldots, w_n) = \langle \mathbf{a}, \mathbf{w} \rangle := \sum_{i=1}^n a_i w_i$. What is the partial derivative of $f$ with respect to $w_i$? The *gradient* of a function is the collection of its partial derivatives when viewed as a vector. What is the gradient $\nabla f$? Recall that the gradient is (usually) a function of the $\mathbf{w}$ variable, but what do you observe in this case?

**Solution:** Let us solve this from first principles (but you don't need to have done this if you remember your partial derivatives). From the definition of a partial derivative, we obtain

$$\begin{aligned}\frac{\partial f}{\partial w_i}(w_1, \ldots, w_n) &= \lim_{\varepsilon \to 0} \frac{f(w_1, \ldots, w_i + \varepsilon, \ldots, w_n) - f(w_1, \ldots, w_i, \ldots, w_n)}{\varepsilon} \\ &= \lim_{\varepsilon \to 0} \frac{a_i(w_i + \varepsilon) - a_i w_i}{\varepsilon} \\ &= a_i\end{aligned}$$

Thus, the gradient is given by $\nabla f(w_1, \ldots, w_n) = \left[\frac{\partial f}{\partial w_1}, \frac{\partial f}{\partial w_2}, \ldots, \frac{\partial f}{\partial w_n}\right]^T = [a_1, a_2, \ldots, a_n]^T = \mathbf{a}$.

The gradient $\nabla f$ is a constant vector (independent of $\mathbf{w}$) since the function $f$ is linear in $\mathbf{w}$.

(b) Compute the gradient of the function $f(w_1, w_2, w_3) = w_1 w_2 + w_2 w_3 + w_1 w_3$.

**Solution:** $\nabla f(w_1, w_2, w_3) = [w_2 + w_3, w_1 + w_3, w_1 + w_2]^T$.

We can derive this again from first principles. Take the first partial derivative as an example:

$$\begin{aligned}\frac{\partial f}{\partial w_1}(w_1, w_2, w_3) &= \lim_{\varepsilon \to 0} \frac{f(w_1 + \varepsilon, w_2, w_3) - f(w_1, w_2, w_3)}{\varepsilon} \\ &= \lim_{\varepsilon \to 0} \frac{(w_1 + \varepsilon) w_2 + w_2 w_3 + (w_1 + \varepsilon) w_3 - (w_1 w_2 + w_2 w_3 + w_1 w_3)}{\varepsilon} \\ &= \lim_{\varepsilon \to 0} \frac{\varepsilon (w_2 + w_3)}{\varepsilon} \\ &= w_2 + w_3.\end{aligned}$$

Unlike the previous case, the gradient now depends on the argument $\mathbf{w}$.

(c) Suppose we have the linear equation $\mathbf{y} = \mathbf{A}\mathbf{v}$, where $\mathbf{y}$ and $\mathbf{v}$ are vectors and $\mathbf{A}$ is a matrix. Argue that $\mathbf{y}$ can be written as a linear combination of the columns of $\mathbf{A}$.

**Solution:** Suppose $\mathbf{A} \in \mathbb{R}^{n \times m}$, $\mathbf{v} \in \mathbb{R}^m$, $\mathbf{y} \in \mathbb{R}^n$ and $\mathbf{y} = \mathbf{A}\mathbf{v}$.

Let $A_i$ denote the $i$-th column of $\mathbf{A}$, and let $a_{ij}$ denote the $(i, j)$ entry of $\mathbf{A}$. Let $v_i$ denote the $i$-th entry of $\mathbf{v}$.

We claim that

$$\mathbf{y} = \sum_{j=1}^m v_j A_j.$$

To prove this, it suffices to verify that $\mathbf{A}\mathbf{v} = \sum_{j=1}^m v_j A_j$. The $i$-th entry of the vector $\mathbf{A}\mathbf{v}$ equals $\sum_{j=1}^m a_{ij} v_j$. The $i$-th entry of the vector $\sum_{j=1}^m v_j A_j$ equals $\sum_{j=1}^m a_{ij} v_j$. This establishes the claim.

(d) Given any vector $\mathbf{v}$ with real entries, show that $\mathbf{v}^\top \mathbf{v}$ is always non-negative.

**Solution:** Let $v_i$ denote the $i$-th entry of $\mathbf{v}$. Then $\mathbf{v}^\top \mathbf{v} = \sum_{i=1}^n v_i^2 \geq 0$.

(e) A square and symmetric matrix $\mathbf{X} \in \mathbb{R}^{d \times d}$ is one that satisfies $\mathbf{X}^\top = \mathbf{X}$. In addition, $\mathbf{X}$ is said to be positive semidefinite (PSD) (written $\mathbf{X} \succeq 0$) if, for all vectors $\mathbf{v} \in \mathbb{R}^d$, we have $\mathbf{v}^T \mathbf{X} \mathbf{v} = \sum_{i=1}^d \sum_{j=1}^d \mathbf{X}_{i,j} v_i v_j \geq 0$.

Suppose a matrix $\mathbf{A}$ can be written as $\mathbf{A} = \mathbf{B}^\top \mathbf{B}$ for another matrix $\mathbf{B}$ (not necessarily square). Argue that $\mathbf{A}$ is square, symmetric, and PSD.

Hint: You may want to try this problem for when $\mathbf{B}$ is just a $1 \times d$ vector. For the general case, think about whether you can express $\mathbf{v}^\top \mathbf{A} \mathbf{v}$ in terms of $\mathbf{B}\mathbf{v}$.

**Solution:** Let $\mathbf{B} \in \mathbb{R}^{n \times m}$. For any $\mathbf{v} \in \mathbb{R}^m$, we have $\mathbf{v}^\top \mathbf{A} \mathbf{v} = \mathbf{v}^\top \mathbf{B}^\top \mathbf{B} \mathbf{v} = (\mathbf{B}\mathbf{v})^\top \mathbf{B}\mathbf{v} \geq 0$, where the last inequality follows from part (d) by viewing $\mathbf{B}\mathbf{v}$ as a vector.

(f) Recall that a matrix $\mathbf{A}$ has (scalar) eigenvalue $\lambda$ associated with an eigenvector $\mathbf{v}$ if $\mathbf{A}\mathbf{v} = \lambda \mathbf{v}$. To avoid identifiability issues, we also assume that the $\ell_2$ norm of $\mathbf{v}$ is equal to 1, i.e., $\|\mathbf{v}\|_2 = \sqrt{\sum_i v_i^2} = 1$. Compute eigenvectors and eigenvalues for the following matrix by hand:

$$\mathbf{\Sigma} = \begin{bmatrix} 3 & 1 \\ 1 & 3 \end{bmatrix}$$

**Solution:** Expanding $\mathbf{\Sigma}\mathbf{v} = \lambda \mathbf{v}$, we obtain the simultaneous equations

$$\begin{aligned} 3v_1 + v_2 &= \lambda v_1 \\ v_1 + 3v_2 &= \lambda v_2. \end{aligned}$$

Subtracting the second from the first then yields $(2 - \lambda)(v_1 - v_2) = 0$. Consequently, either $\lambda = 2$ or $v_1 = v_2$. These two cases will yield the two pairs we seek:

First eigenvalue/eigenvector pair: We have $\lambda = 2$, so that returning to the above equations, we have $v_1 = -v_2$. Normalizing this eigenvector to the unit sphere yields the corresponding eigenvector $\mathbf{v} = (1/\sqrt{2}, -1/\sqrt{2})$. (this is only well-defined up to a sign).

Second eigenvalue/eigenvector pair: In this case, we have $v_1 = v_2$, and normalizing to the unit sphere yields the eigenvector $\mathbf{v} = (1/\sqrt{2}, 1/\sqrt{2})$ (well-defined up to a sign). Solving the above system with $v_1 = v_2$ then yields $\lambda = 4$ as the corresponding eigenvalue.

(g) Let $\mathbf{A}$ be an invertible matrix. Show that if $\mathbf{v}$ is an eigenvector of $\mathbf{A}$ with eigenvalue $\lambda$, then it is also an eigenvector of $\mathbf{A}^{-1}$ with eigenvalue $\lambda^{-1}$.

**Solution:** Since $\mathbf{v}$ is an eigenvector of $\mathbf{A}$ with eigenvalue $\lambda$, we have $\mathbf{A}\mathbf{v} = \lambda \mathbf{v}$. Since $\mathbf{A}$ is invertible, $\lambda \neq 0$. We may then multiply both sides of this equation by $\mathbf{A}^{-1}$ to obtain

$$\mathbf{A}^{-1}\mathbf{A}\mathbf{v} = \mathbf{A}^{-1}\lambda \mathbf{v} \implies \mathbf{v} = \lambda \mathbf{A}^{-1}\mathbf{v}.$$

Since $\lambda$ is a scalar non-zero value, rearranging yields $\mathbf{A}^{-1}\mathbf{v} = \lambda^{-1} \cdot \mathbf{v}$, as desired.

**Problem 2 (Probability).**

(a) The conditional probability of an event $\mathcal{E}$ given another event $\mathcal{F}$ is given by $\Pr(\mathcal{E}|\mathcal{F}) = \Pr(\mathcal{E} \cap \mathcal{F})/\Pr(\mathcal{F})$. Use this fact to derive Bayes' rule:

$$\Pr(\mathcal{F}|\mathcal{E}) = \Pr(\mathcal{E}|\mathcal{F}) \cdot \Pr(\mathcal{F})/\Pr(\mathcal{E}),$$

which makes sense provided we are not dividing by zero.

Hint: Use the fact that intersections commute, i.e., the event $\mathcal{E} \cap \mathcal{F}$ is identical to the event $\mathcal{F} \cap \mathcal{E}$.

**Solution:** Apply the hint to see that

$$\Pr(\mathcal{F}|\mathcal{E}) \cdot \Pr(\mathcal{E}) = \Pr(\mathcal{F} \cap \mathcal{E}) = \Pr(\mathcal{E} \cap \mathcal{F}) = \Pr(\mathcal{E}|\mathcal{F}) \cdot \Pr(\mathcal{F}).$$

Dividing through $\Pr(\mathcal{E})$ yields the claim.

(b) Suppose we have two random variables $X$ and $Y$ taking real values (i.e., in the set $\mathbb{R}$) and having finite expectations $\mathbb{E}[X]$ and $\mathbb{E}[Y]$, respectively. Can you write $\mathbb{E}[X+Y]$ in terms of expectations of the individual random variables? Can you write $\mathbb{E}[XY]$ in terms of expectations of the individual random variables?

**Solution:** It follows from the linearity of expectation that $\mathbb{E}[X+Y] = \mathbb{E}[X] + \mathbb{E}[Y]$.

If the random variables $X$ and $Y$ are independent (or more generally, uncorrelated), then $\mathbb{E}[XY] = \mathbb{E}[X] \cdot \mathbb{E}[Y]$.

In general, we cannot express $\mathbb{E}[XY]$ in terms of $\mathbb{E}[X]$ and $\mathbb{E}[Y]$. For an explicit example, take $X$ to be a *Rademacher* random variable, taking the value $-1$ with probability $1/2$ and $1$ with probability $1/2$. Now consider two choices for $Y$:

(i) $Y$ is a Rademacher independent of $X$.
(ii) $Y = X$.

In both cases, we have $\mathbb{E}[X] = \mathbb{E}[Y] = 0$. However, in case (i), $\mathbb{E}[XY] = 0$ but in case (ii), $\mathbb{E}[XY] = \mathbb{E}[X^2] = 1$.

(c) The variance of a random variable is given by $\mathbb{E}[(X - \mathbb{E}[X])^2]$. Argue that this is always non-negative. Suppose we have two independent random variables $X$ and $Y$ with variances $\sigma_1^2$ and $\sigma_2^2$, respectively. What is the variance of $X+Y$? What is the variance of $X-Y$?

**Solution:** Note that $(X - \mathbb{E}[X])^2$ is always positive. The expectation is just the average of these values over some distribution, and so $\mathbb{E}\left[(X - \mathbb{E}[X])^2\right] \geq 0$.

We use $\mathrm{Var}(X)$ to denote the variance of the random variable $X$.

To compute the variance of $X+Y$, note that

$$\begin{aligned}
\mathbb{E}\left[(X+Y-\mathbb{E}[X+Y])^2\right] &= \mathbb{E}\left[(X-\mathbb{E}[X]+Y-\mathbb{E}[Y])^2\right] \\
&= \mathbb{E}\left[(X-\mathbb{E}[X])^2\right] + 2\mathbb{E}\left[(X-\mathbb{E}[X]) \cdot (Y-\mathbb{E}[Y])\right] + \mathbb{E}\left[(Y-\mathbb{E}[Y])^2\right] \\
&= \mathbb{E}\left[(X-\mathbb{E}[X])^2\right] + \mathbb{E}\left[(Y-\mathbb{E}[Y])^2\right]
\end{aligned}$$

The last equality follows from the independence of $X, Y$ from which we have

$$\mathbb{E}\left[(X - \mathbb{E}\left[X\right]) \cdot (Y - \mathbb{E}\left[Y\right])\right] = \mathbb{E}\left[(X - \mathbb{E}\left[X\right])\right] \cdot \mathbb{E}\left[(Y - \mathbb{E}\left[Y\right])\right] = 0.$$

To compute the variance of $X - Y$, note that

$$\begin{aligned}
\mathbb{E}\left[(X - Y - \mathbb{E}\left[X - Y\right])^2\right] &= \mathbb{E}\left[(X - \mathbb{E}\left[X\right] + \mathbb{E}\left[Y\right] - Y)^2\right] \\
&= \mathbb{E}\left[(X - \mathbb{E}\left[X\right])^2\right] + 2\mathbb{E}\left[(X - \mathbb{E}\left[X\right]) \cdot (\mathbb{E}\left[Y\right] - Y)\right] + \mathbb{E}\left[(Y - \mathbb{E}\left[Y\right])^2\right] \\
&= \mathbb{E}\left[(X - \mathbb{E}\left[X\right])^2\right] + \mathbb{E}\left[(Y - \mathbb{E}\left[Y\right])^2\right],
\end{aligned}$$

where the middle term is equal to zero for the same reason as before.

(d) Suppose $X$ is a random variable with expectation 0 and variance 1. What are the expectation and variance of $Y = aX + b$ (here, $a$ and $b$ are some fixed real numbers, and your answer will be in terms of $(a, b)$).

**Solution:** We use the *linearity* of expectation (hint in part (e) below) to write

$$\mathbb{E}\left[Y\right] = \mathbb{E}\left[aX + b\right] = a \cdot \mathbb{E}\left[X\right] + b = b,$$

where in the last step we used that $\mathbb{E}[X] = 0$. For the variance, write

$$\mathrm{Var}\left(Y\right) = \mathbb{E}\left[(Y - \mathbb{E}\left[Y\right])^2\right] = \mathbb{E}\left[(aX + b - a \cdot \mathbb{E}\left[X\right] - b)^2\right] = a^2 \cdot \mathbb{E}\left[(X - \mathbb{E}\left[X\right])^2\right] = a^2 \cdot \mathrm{Var}\left(X\right) = a^2,$$

where the last step uses the fact that $X$ has unit variance.

(e) Now consider an $n$-dimensional vector of random variables (i.e. each entry of the vector is itself a random variable) $Z \in \mathbb{R}^n$. Suppose $Z$ has zero expectation, i.e., $\mathbb{E}[Z_i] = \mu_i = 0$ for all $1 \leq i \leq n$ and that its *covariance matrix* has entries $\mathbf{\Sigma}_{i,j} = \mathbb{E}[(Z_i - \mu_i)(Z_j - \mu_j)]$. What is the expectation and covariance matrix of $W = \mathbf{A}Z + \mathbf{b}$, where $\mathbf{A}$ is an $n \times n$ matrix and $\mathbf{b}$ is an $n$-dimensional vector?

Hint: Use the fact that the expectation is *linear*: $\mathbb{E}[aX + bY] = a\mathbb{E}[X] + b\mathbb{E}[Y]$.

**Solution:** Let $a_{ij}$ denote the $(i, j)$ entry of $\mathbf{A}$ and $b_i$ denote the $i$-th entry of $\mathbf{b}$. Then $W_i = \sum_{k=1}^{n} a_{ik} Z_k + b_i$ for $1 \leq i \leq n$. By the linearity of expectation, we have

$$\mathbb{E}\left[W_i\right] = \sum_{k=1}^{n} a_{ik} \mathbb{E}\left[Z_k\right] + b_i = \sum_{k=1}^{n} a_{ik} \mu_k + b_i.$$

Writing this in vector form, we obtain

$$\mathbb{E}\left[W\right] = \mathbf{A}\mathbb{E}\left[Z\right] + \mathbf{b}.$$

By definition, the $(i, j)$th entry of the covariance matrix equals $\mathbb{E}\left[(W_i - \mathbb{E}\left[W_i\right]) \cdot (W_j - \mathbb{E}\left[W_j\right])\right]$. Note that

$$W_i - \mathbb{E}\left[W_i\right] = \sum_{k=1}^{n} a_{ik} \left(Z_k - \mu_k\right) \text{ and } W_j - \mathbb{E}\left[W_j\right] = \sum_{l=1}^{n} a_{jl} \left(Z_l - \mu_l\right).$$

We can use this expression to write

$$(W_i - \mathbb{E}[W_i]) \cdot (W_j - \mathbb{E}[W_j]) = \left(\sum_{k=1}^{n} a_{ik}(Z_k - \mu_k)\right) \times \left(\sum_{l=1}^{n} a_{jl}(Z_l - \mu_l)\right)$$
$$= \sum_{k=1}^{n} a_{ik}a_{jk}(Z_k - \mu_k)^2 + \sum_{k \neq l} a_{ik}a_{jl}(Z_k - \mu_k)(Z_l - \mu_l)$$

Take expectation on both sides to obtain

$$\mathbb{E}[(W_i - \mathbb{E}[W_i]) \cdot (W_j - \mathbb{E}[W_j])] = \mathbb{E}\left[\sum_{k=1}^{n} a_{ik}a_{jk}(Z_k - \mu_k)^2\right] + \mathbb{E}\left[\sum_{k \neq l} a_{ik}a_{jl}(Z_k - \mu_k)(Z_l - \mu_l)\right]$$
$$= \sum_{k=1}^{n} a_{ik}a_{jk}\mathbf{\Sigma}_{kk} + \sum_{k \neq l} a_{ik}a_{jl}\mathbf{\Sigma}_{lk}$$

Writing the above in matrix form, the covariance matrix is $\mathbf{A\Sigma A}^\top$.

**Note:** The calculations above could have been done directly in terms of vectors instead of writing scalar forms. We have $\mathbb{E}[\mathbf{A}Z + \mathbf{b}] = \mathbf{A}\mathbb{E}[Z] + \mathbf{b} = \mathbf{A}\mu + \mathbf{b} = \mathbf{b}$.

The covariance matrix of $W$ is given by

$$\mathbb{E}[(W - \mathbb{E}[W])(W - \mathbb{E}[W])^\top] = \mathbb{E}[(\mathbf{A}Z)(\mathbf{A}Z)^\top] = \mathbb{E}[\mathbf{A}ZZ^\top\mathbf{A}^\top] = \mathbf{A}\mathbb{E}[ZZ^\top]\mathbf{A}^\top = \mathbf{A\Sigma A}^\top.$$

**Problem 3 (Python installation and basic simulation).** It is strongly recommended that you use Jupyter notebooks for your coding assignments. This exercise guides you through an installation through Anaconda (please do this only if you do not have a working version of Jupyter already), and asks some basic questions about generating data and plotting. The latter questions will only make use of the numpy package.

(a) Please install Python using Anaconda if you haven't already, documentation can be found here: `https://docs.anaconda.com/anaconda/`. Review the following tutorial on using Python through Jupyter notebooks: `https://cs4540-f18.github.io/notes/python-basics` If you really need to refresh, here is a more elementary tutorial on Python programming: `https://www.learnpython.org/`

(b) Sample 100 points from a 2-dimensional multivariate Gaussian distribution $\mathcal{N}(0, \mathbf{I})$, where $\mathbf{I}$ is the $2 \times 2$ identity matrix. Store these points in a $100 \times 2$ dimensional matrix.

**Solution:** See accompanying Python notebook.

(c) Produce a 2-dimensional scatter plot of all the points you just sampled.

**Solution:** See accompanying Python notebook.

(d) Suppose for the moment that you didn't know the mean or covariance matrix of the Gaussian distribution from which the points were sampled. What is a reasonable *estimate* of the 2-dimensional mean vector from the data that you sampled? Compute this estimate and report it. Analogously, what is a reasonable *estimate* of the $2 \times 2$ covariance matrix from the data that you sampled? Compute this estimate and report it.

**Solution:** See accompanying Python notebook.

(e) Compute the $\ell_2$ distance between your mean estimate and the true mean $(0, 0)$ and report it.
**Solution:** See accompanying Python notebook.

(f) Intuitively, what do expect will happen to this error if you drew 1000 samples instead?
**Solution:** See accompanying Python notebook.