Homework #0

CSE 546: Machine Learning
Madeline Brown
Due: **Wednesday** October 02, 2024 11:59pm
38 points

Collaborators and external resources: perplexity.ai for obtaining definitions and figuring out how to add vectors in Python for A9.

Probability and Statistics

A1. *[2 points]* (From Murphy Exercise 2.4.) After your yearly checkup, the doctor has bad news and good news. The bad news is that you tested positive for a serious disease, and that the test is 99% accurate (i.e., the probability of testing positive given that you have the disease is 0.99, as is the probability of testing negative given that you don't have the disease). The good news is that this is a rare disease, striking only one in 10,000 people. What are the chances that you actually have the disease?

What to Submit:

- Final Answer
- Corresponding Calculations

To find the chances that you have the disease given that you've tested positive (+), we will use Bayes' theorem:

$$\mathbb{P}(\text{have disease}|+) = \frac{\mathbb{P}(+|\text{have disease}) \cdot \mathbb{P}(\text{have disease})}{\mathbb{P}(+)}.$$

We know from the problem description that $\mathbb{P}(+|\text{have disease}) = 0.99$, $\mathbb{P}(-|\text{don't have disease}) = 0.99$, and $\mathbb{P}(\text{have disease}) = 0.0001$. Then

$$\begin{aligned}\mathbb{P}(+) &= \mathbb{P}(+|\text{have disease}) \cdot \mathbb{P}(\text{have disease}) + \mathbb{P}(+|\text{don't have disease}) \cdot \mathbb{P}(\text{don't have disease}) \\ &= (0.99 \cdot 0.0001) + (0.01 \cdot 0.9999) \\ &= 0.010098.\end{aligned}$$

Thus, we get

$$\mathbb{P}(\text{have disease}|+) = \frac{0.99 \cdot 0.0001}{0.010098} \approx 0.0098$$

So the probability that you actually have the disease given that you've tested positive is about 0.98%.

A2. For any two random variables X, Y the *covariance* is defined as $\text{Cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]$. You may assume X and Y take on a discrete values if you find that is easier to work with.

a. *[1 point]* If $\mathbb{E}[Y \mid X = x] = x$ show that $\text{Cov}(X, Y) = \mathbb{E}\Big[(X - \mathbb{E}[X])^2\Big]$.

Proof. From the Ed discussion, we use an equivalent definition of covariance: $\text{Cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y]$. We further use the Law of Total Expectation $\mathbb{E}[XY] = \mathbb{E}[\mathbb{E}[XY|X]]$ and the assumption $\mathbb{E}[Y \mid X = x] = x$ to prove the claim:

$$\begin{aligned}
\text{Cov}(X, Y) &= \mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y] \\
&= \mathbb{E}[\mathbb{E}[XY|X]] - \mathbb{E}[X]\,\mathbb{E}[Y] \text{(Law of Total Expectation)} \\
&= \mathbb{E}[X\,\mathbb{E}[Y|X]] - \mathbb{E}[X]\,\mathbb{E}[Y] \text{ (X is constant given X)} \\
&= \mathbb{E}[X \cdot X] - \mathbb{E}[X]\,\mathbb{E}[Y] \text{ since } \mathbb{E}[Y \mid X = x] = x \\
&= \mathbb{E}\big[X^2\big] - \mathbb{E}[X]^2 \\
&= \mathbb{E}\big[X^2\big] - 2\,\mathbb{E}[X]^2 + \mathbb{E}[X]^2 \\
&= \mathbb{E}\Big[X^2 - 2X\,\mathbb{E}[X] + \mathbb{E}[X]^2\Big] \\
&= \mathbb{E}\big[(X - \mathbb{E}[X])^2\big]
\end{aligned}$$

□

b. *[1 point]* If X, Y are independent show that $\text{Cov}(X, Y) = 0$.

Proof. If X and Y are independent, then the expectation of their product is equal to the product of their expectations: $\mathbb{E}[XY] = \mathbb{E}[X]\,\mathbb{E}[Y]$. Then using the definition of covariance, we have

$$\text{Cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y] = \mathbb{E}[X]\,\mathbb{E}[Y] - \mathbb{E}[X]\,\mathbb{E}[Y] = 0.$$

□

What to Submit:

- **Parts a-b:** Proofs

A3. Let X and Y be independent random variables with PDFs given by f and g, respectively. Let h be the PDF of the random variable $Z = X + Y$.

a. *[1 point]* Show that $h(z) = \int_{-\infty}^{\infty} f(x)g(z-x)\,\mathrm{d}x$. (If you are more comfortable with discrete probabilities, you can instead derive an analogous expression for the discrete case, and then you should give a one sentence explanation as to why your expression is analogous to the continuous case.).

Proof. Recall that the CDF is the the integral of the PDF. We first write the CDF $H(z)$ for Z, using the fact that $y = z - x$, as

$$H(z) = \mathbb{P}(Z \leq z) = \mathbb{P}(X + Y \leq z) = \int_{-\infty}^{\infty} \int_{-\infty}^{z-x} f(x)g(y)dydx.$$

Then, we can take the derivative of this with respect to z to obtain the PDF $h(z)$ for Z:

$$\begin{aligned} h(z) = \frac{d}{dz}H(z) &= \int_{-\infty}^{\infty} \frac{d}{dz} \int_{-\infty}^{z-x} f(x)g(y)dydx \\ &= \int_{-\infty}^{\infty} f(x)g(z-x)dx, \end{aligned}$$

where the last equality comes from the fundamental theorem of calculus. We note that this is nonnegative because f and g are both PDFs and are thus nonnegative. Last, we just need to make sure $h(z)$ integrates to 1, then we will know that it fits the definition of PDF!

Using a change of variables with $u = z - x$ and thus $du = dz$, as well as the fact that both f and g are PDFs and integrate to 1 themselves, we have:

$$\begin{aligned} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x)g(z-x)dxdz &= \int_{-\infty}^{\infty} f(x) \left(\int_{-\infty}^{\infty} g(z-x)dz \right) dx \\ &= \int_{-\infty}^{\infty} f(x) \left(\int_{-\infty}^{\infty} g(u)du \right) dx \\ &= \int_{-\infty}^{\infty} f(x)dx = 1. \end{aligned}$$

This confirms the claim. □

b. *[1 point]* If X and Y are both independent and uniformly distributed on $[0, 1]$ (i.e. $f(x) = g(x) = 1$ for $x \in [0, 1]$ and 0 otherwise) what is h, the PDF of $Z = X + Y$?

We start with the above result that $h(z) = \int_{-\infty}^{\infty} f(x)g(z-x)\,\mathrm{d}x$. This is nonzero for sure when $x \in [0, 1]$, but we could also have $z \in [0, 2]$, and still get a nonzero result. Note that since we evaluate $z - x$ inside of g, this means that we can split into one case where $x \in [0, z]$ for $z \in [0, 1]$ and another where $x \in [z - 1, 1]$ for $z \in (1, 2]$. Evaluating $h(z)$ over these two domains, using the fact that $f(x) = g(x) = 1$ for $x \in [0, 1]$, gives

$$\int_0^z f(x)g(z-x)dx = \int_0^z dx = z$$

and

$$\int_{z-1}^1 f(x)g(z-x)dx = \int_{z-1}^1 dx = 2 - z.$$

Thus, we have that the PDF of Z is

$$h(z) = \begin{cases} z & z \in [0, 1] \\ 2 - z & z \in (1, 2]. \end{cases} \tag{1}$$

What to Submit:

- **Part a:** Proof
- **Part b:** Formula for PDF Z and corresponding calculations

A4. Let $X_1, X_2, ..., X_n \sim \mathcal{N}(\mu, \sigma^2)$ be i.i.d random variables. Compute the following:

a. *[1 point]* $a \in \mathbb{R}, b \in \mathbb{R}$ such that $aX_1 + b \sim \mathcal{N}(0, 1)$.

By rules of addition of normal random variables from lecture 1, we know that $aX_1 + b \sim \mathcal{N}(a\mu + b, a^2\sigma^2)$. Since we want $a\mu + b = 0$ and $a^2\sigma^2 = 1$, then we have $a\mu = -b$ and $a\sigma = \pm 1$. Then $a = \pm\frac{1}{\sigma}$ and $b = \mp\frac{\mu}{\sigma}$ are solutions in $\mathbb{R}$ for $aX_1 + b \sim \mathcal{N}(0, 1)$.

b. *[1 point]* $\mathbb{E}[X_1 + 2X_2], \text{Var}[X_1 + 2X_2]$.

By linearity of expectation, knowing that the expectation for $X_1, X_2, ..., X_n$ are all μ, we have that $\mathbb{E}[X_1 + 2X_2] = \mu + 2\mu = 3\mu$.

When adding variances, we must square the coefficient term, so knowing that the variances for $X_1, X_2, ..., X_n$ are all σ^2, we have $\text{Var}[X_1 + 2X_2] = \sigma^2 + 2^2\sigma^2 = 5\sigma^2$.

c. *[2 points]* Setting $\widehat{\mu}_n = \frac{1}{n}\sum_{i=1}^n X_i$, the mean and variance of $\sqrt{n}(\widehat{\mu}_n - \mu)$.

From lecture, we know that $\mathbb{E}[\widehat{\mu}_n] = \mu$. Then by linearity of expectation and $\mathbb{E}[\mu] = \mu$, we have

$$\mathbb{E}\big[\sqrt{n}(\widehat{\mu}_n - \mu)\big] = \sqrt{n}\,\mathbb{E}[\widehat{\mu}_n - \mu] = \sqrt{n}(\mathbb{E}[\widehat{\mu}_n] - \mu) = \sqrt{n}(\mu - \mu) = 0.$$

Additionally, we know the variance is $\text{Var}[\widehat{\mu}_n] = \text{Var}\big[\frac{1}{n}\sum_{i=1}^n X_i\big] = \frac{1}{n^2}\sum_{i=1}^n \text{Var}[X_i] = \frac{1}{n^2}n\sigma^2 = \frac{\sigma^2}{n}$. Using the fact that μ is a constant, we have

$$\text{Var}\big[\sqrt{n}(\widehat{\mu}_n - \mu)\big] = n\,\text{Var}[\widehat{\mu}_n - \mu] = n\,\text{Var}[\widehat{\mu}_n] = n\frac{\sigma^2}{n} = \sigma^2.$$

What to Submit:

- **Part a:** a, b, and the corresponding calculations
- **Part b:** $\mathbb{E}[X_1 + 2X_2]$, $\text{Var}[X_1 + 2X_2]$
- **Part c:** $\mathbb{E}[\sqrt{n}(\widehat{\mu}_n - \mu)]$, $\text{Var}[\sqrt{n}(\widehat{\mu}_n - \mu)]$
- **Parts a-c** Corresponding calculations

Linear Algebra and Vector Calculus

A5. Let $A = \begin{bmatrix} 1 & 2 & 1 \\ 1 & 0 & 3 \\ 1 & 1 & 2 \end{bmatrix}$ and $B = \begin{bmatrix} 1 & 2 & 3 \\ 1 & 0 & 1 \\ 1 & 1 & 2 \end{bmatrix}$. For each matrix A and B:

a. *[2 points]* What is its rank?

The rank of both matrices is 2. We can see this by noting that for A, if we denote the rows as R_1, R_2, and R_3, that $R_1 = 2R_3 - R_2$. Noting the fact that rows 2 and 3 are independent (given that the second entry for R_2 is 0 while both of the other rows have nonzero entries, so no linear combination of R_1 or R_3 can equal R_2) we have that there are just two linearly independent rows, so the rank of A is 2.

By similar logic for B, noting that we also have $R_1 = 2R_3 - R_2$, the rank of B is also 2.

b. *[2 points]* What is a (minimal size) basis for its column span?

Since both A and B are of rank 2, their minimal size bases will have two linearly independent vectors. We just pick any two linearly independent columns, which happen to be the same for A and B. Let

$$\mathcal{B}_A = \left\{ \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}, \begin{bmatrix} 2 \\ 0 \\ 1 \end{bmatrix} \right\}$$

and

$$\mathcal{B}_B = \left\{ \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}, \begin{bmatrix} 2 \\ 0 \\ 1 \end{bmatrix} \right\}$$

be the bases for A and B, respectively. Note that they are equal since the first two columns of A and B are equal.

What to Submit:

- **Parts a-b:** Solution and corresponding calculations

A6. Let $A = \begin{bmatrix} 0 & 2 & 4 \\ 2 & 4 & 2 \\ 3 & 3 & 1 \end{bmatrix}$, $b = \begin{bmatrix} -2 & -2 & -4 \end{bmatrix}^\top$, and $c = \begin{bmatrix} 1 & 1 & 1 \end{bmatrix}^\top$.

a. *[1 point]* What is Ac?

We can compute this directly:

$$Ac = \begin{bmatrix} 0 & 2 & 4 \\ 2 & 4 & 2 \\ 3 & 3 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 0+2+4 \\ 2+4+2 \\ 3+3+1 \end{bmatrix} = \begin{bmatrix} 6 \\ 8 \\ 7 \end{bmatrix}.$$

b. *[2 points]* What is the solution to the linear system $Ax = b$?

To find the solution to

$$Ax = b$$

$$\begin{bmatrix} 0 & 2 & 4 \\ 2 & 4 & 2 \\ 3 & 3 & 1 \end{bmatrix} x = \begin{bmatrix} -2 \\ -2 \\ -4 \end{bmatrix}$$

we will take the inverse of A, which will give us x because $x = A^{-1}Ax = A^{-1}b$. By the Ed discussion, we use a calculator to find this inverse, which gives

$$A^{-1} = \begin{bmatrix} 1/8 & -5/8 & 3/4 \\ -1/4 & 3/4 & -1/2 \\ 3/8 & -3/8 & 1/4 \end{bmatrix}.$$

Then

$$x = A^{-1}b = \begin{bmatrix} 1/8 & -5/8 & 3/4 \\ -1/4 & 3/4 & -1/2 \\ 3/8 & -3/8 & 1/4 \end{bmatrix} \begin{bmatrix} -2 \\ -2 \\ -4 \end{bmatrix} = \begin{bmatrix} -1/4+5/4-3 \\ 1/2-3/2+2 \\ -3/4+3/4-2 \end{bmatrix} = \begin{bmatrix} -2 \\ 1 \\ -1 \end{bmatrix}.$$

This vector x also gives us $Ax = b$, so the solution is correct!

What to Submit:

- **Parts a-b:** Solution and corresponding calculations

A7. For possibly non-symmetric $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{n \times n}$ and $c \in \mathbb{R}$, let $f(x, y) = x^\top \mathbf{A} x + y^\top \mathbf{B} x + c$. Define

$$\nabla_z f(x, y) = \begin{bmatrix} \frac{\partial f}{\partial z_1}(x, y) & \frac{\partial f}{\partial z_2}(x, y) & \dots & \frac{\partial f}{\partial z_n}(x, y) \end{bmatrix}^\top \in \mathbb{R}^n .$$

Note: If you are unfamiliar with gradients, you may find the resources available on the course website useful. Section 4 of Zico Kolter and Chuong Do's Linear Algebra Review and Reference may be particularly helpful.

a. *[2 points]* Explicitly write out the function $f(x, y)$ in terms of the components $A_{i,j}$ and $B_{i,j}$ using appropriate summations over the indices.

Breaking down the individual parts, we have

$$x^\top \mathbf{A} x = x^\top \begin{bmatrix} \sum_{j=1}^n A_{1,j} x_j \\ \sum_{j=1}^n A_{2,j} x_j \\ \vdots \end{bmatrix} = \sum_{i=1}^n x_i \left(\sum_{j=1}^n A_{i,j} x_j \right) = \sum_{i=1}^n \sum_{j=1}^n x_i A_{i,j} x_j,$$

and then we similarly have

$$y^\top \mathbf{B} x = \sum_{i=1}^n \sum_{j=1}^n y_i B_{i,j} x_j.$$

Thus,

$$f(x, y) = x^\top \mathbf{A} x + y^\top \mathbf{B} x + c = \sum_{i=1}^n \sum_{j=1}^n x_i A_{i,j} x_j + \sum_{i=1}^n \sum_{j=1}^n y_i B_{i,j} x_j + c.$$

b. *[2 points]* What is $\nabla_x f(x, y)$ in terms of the summations over indices *and* vector notation?

By the identities in section 7.8.7.2, and from lecture 3, we have

$$\frac{\partial (x^\top \mathbf{A} x)}{\partial x} = (\mathbf{A} + \mathbf{A}^\top) x$$

and

$$\frac{\partial (y^\top \mathbf{B} x)}{\partial x} = (\mathbf{B}^\top) y.$$

Then in vector notation, the gradient with respect to x is

$$\nabla_x f(x, y) = (\mathbf{A} + \mathbf{A}^\top) x + \mathbf{B}^\top y.$$

In terms of summation notation, this is equivalent to

$$\nabla_x f(x, y) = \begin{bmatrix} \sum_{j=1}^n (A_{1,j} + A_{j,1}) x_j + \sum_{i=1}^n B_{i,1} y_i \\ \sum_{j=1}^n (A_{2,j} + A_{j,2}) x_j + \sum_{i=1}^n B_{i,2} y_i \\ \vdots \\ \sum_{j=1}^n (A_{n,j} + A_{j,n}) x_j + \sum_{i=1}^n B_{i,n} y_i \end{bmatrix}.$$

c. *[2 points]* What is $\nabla_y f(x, y)$ in terms of the summations over indices *and* vector notation?

Notice that y is only present in one term of $f(x, y)$. We can see by the sum notation in part (a) that this term is

$$\sum_{i=1}^{n}\sum_{j=1}^{n} y_i B_{i,j} x_j = \sum_{i=1}^{n} y_i \sum_{j=1}^{n} B_{i,j} x_j$$

so that the coefficient of each y_i is $\sum_{j=1}^{n} B_{i,j}x_j$, thus this is what is left when we differentiate $f(x, y)$ with respect to each y_i. Notice also that $\sum_{j=1}^{n} B_{i,j}x_j = (\mathbf{B}x)_i$, where $(\mathbf{B}x)_i$ is row i of vector $\mathbf{B}x$. Then the summation and vector form of $\nabla_y f(x, y)$ are

$$\nabla_y f(x, y) = \begin{bmatrix} \sum_{j=1}^{n} B_{1,j}x_j \\ \sum_{j=1}^{n} B_{2,j}x_j \\ \vdots \\ \sum_{j=1}^{n} B_{n,j}x_j \end{bmatrix} = \mathbf{B}x.$$

What to Submit:

- **Part a:** Explicit formula for $f(x, y)$
- **Parts b-c:** Summation form and corresponding calculations. Summation form includes writing out what each component of the resultant vector is, where each component is expressed as a summation. Intermediate components may be indicated by ellipses, like in the equation given in the problem description.
- **Parts b-c:** Vector form and corresponding calculations. Vector form includes writing the final answer only in terms of products, sums (or differences), and/or transposes of the input matrices and vectors.

A8. Show the following:

a. *[2 points]* Let $g\colon \mathbb{R} \to \mathbb{R}$ and $v, w \in \mathbb{R}^n$ such that $g(v_i) = w_i$ for $i \in [n]$. Find an expression for g such that $\text{diag}(v)^{-1} = \text{diag}(w)$.

Claim: $g(x) = \frac{1}{x}$.

Proof. Recall that $\text{diag}(v)$ is an $n \times n$ matrix with the elements of v on its diagonal:

$$\text{diag}(v) = \begin{bmatrix} v_1 & 0 & \cdots & 0 \\ 0 & v_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & v_n \end{bmatrix}.$$

Note that $\text{diag}(v)^{-1}$ exists because the columns of $\text{diag}(v)$ are independent (assuming all entries of v are nonzero, per the Ed discussion). The inverse $\text{diag}(v)^{-1}$ is the diagonal matrix with the reciprocals of the entries of v on the diagonal, as we can see that this would give $\text{diag}(v)\text{diag}(v)^{-1} = \text{diag}(v)^{-1}\text{diag}(v) = \mathbf{I}$.

$$\text{diag}(v)^{-1} = \begin{bmatrix} 1/v_1 & 0 & \cdots & 0 \\ 0 & 1/v_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1/v_n \end{bmatrix}.$$

If $g(x) = 1/x$ is so that $g(v_i) = w_i$ for $i \in [n]$, then

$$\text{diag}(v)^{-1} = \begin{bmatrix} 1/v_1 & 0 & \cdots & 0 \\ 0 & 1/v_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1/v_n \end{bmatrix} = \begin{bmatrix} g(v_1) & 0 & \cdots & 0 \\ 0 & g(v_2) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & g(v_n) \end{bmatrix} = \begin{bmatrix} w_1 & 0 & \cdots & 0 \\ 0 & w_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & w_n \end{bmatrix} = \text{diag}(w),$$

as desired. □

b. *[2 points]* Let $\mathbf{A} \in \mathbb{R}^{n \times n}$ be orthonormal and $x \in \mathbb{R}^n$. An orthonormal matrix is a square matrix whose columns and rows are orthonormal vectors, such that $\mathbf{A}\mathbf{A}^\top = \mathbf{A}^\top\mathbf{A} = \mathbf{I}$ where $\mathbf{I}$ is the identity matrix. Show that $||\mathbf{A}x||_2^2 = ||x||_2^2$.

Proof. Recall that $\|x\|_2^2 = x_1^2 + x_2^2 + ... + x_n^2 = x^\top x$. Then

$$\|\mathbf{A}\|_2^2 = (\mathbf{A}x)^\top \mathbf{A}x = x^\top \mathbf{A}^\top \mathbf{A}x = x^\top \mathbf{I}x = x^\top x = \|x\|_2^2,$$

where we used the definition of orthogonal matrix that $\mathbf{A}^\top\mathbf{A} = \mathbf{I}$ and the rule of transpose of a product $(\mathbf{A}\mathbf{B})^\top = \mathbf{B}^\top\mathbf{A}^\top$. □

c. *[2 points]* Let $\mathbf{B} \in \mathbb{R}^{n \times n}$ be invertible and symmetric. A symmetric matrix is a square matrix satisfying $\mathbf{B} = \mathbf{B}^\top$. Show that $\mathbf{B}^{-1}$ is also symmetric.

Proof. Per the Ed discussion, we use the fact that the inverse of the transpose of a matrix is the same as the transpose of the inverse: $(\mathbf{B}^{-1})^\top = (\mathbf{B}^\top)^{-1}$. Since $\mathbf{B}$ is symmetric, $\mathbf{B} = \mathbf{B}^\top$, we have

$$(\mathbf{B}^{-1})^\top = (\mathbf{B}^\top)^{-1} = \mathbf{B}^{-1}.$$

Since $(\mathbf{B}^{-1})^\top = \mathbf{B}^{-1}$, this shows that $\mathbf{B}^{-1}$ is symmetric. □

d. *[2 points]* Let $\mathbf{C} \in \mathbb{R}^{n \times n}$ be positive semi-definite (PSD). A positive semi-definite matrix is a symmetric matrix satisfying $x^\top \mathbf{C} x \geq 0$ for any vector $x \in \mathbb{R}^n$. Show that its eigenvalues are non-negative.

Proof. Let v be an eigenvector of $\mathbf{C}$ with eigenvalue λ. Then by definition, $\mathbf{C}v = \lambda v$. By multiplying on the left by $v^\top$, we get

$$v^\top \mathbf{C} v = v^\top \lambda v = \lambda \|v\|_2^2.$$

Since $v^\top \mathbf{C} v \geq 0$, then $\lambda \|v\|_2^2 \geq 0$. Then since $\|v\|_2^2 \geq 0$ for any vector v, then we must also have $\lambda \geq 0$. Since λ was arbitrary, this shows that any eigenvalue of $\mathbf{C}$ is nonnegative. □

What to Submit:

- **Part a:** Explicit formula for g
- **Parts a-d:** Proof

Programming

These problems are available in a .zip file, with some starter code. All coding questions in this class will have starter code. **Before attempting these problems, you will need to set up a Conda environment that you will use for every assignment in the course. Unzip the HW0-A.zip file and read the instructions in the README file to get started.**

A9. For $\nabla_x f(x, y)$ as solved for in Problem 7:

a. *[1 point]* Using native Python, implement `vanilla_solution` using your `vanilla_matmul` and `vanilla_transpose` functions.

b. *[1 point]* Now implement `numpy_version` using NumPy functions.

c. *[1 point]* Report the difference in wall-clock time for parts a-b, and discuss reasons for the observed difference.

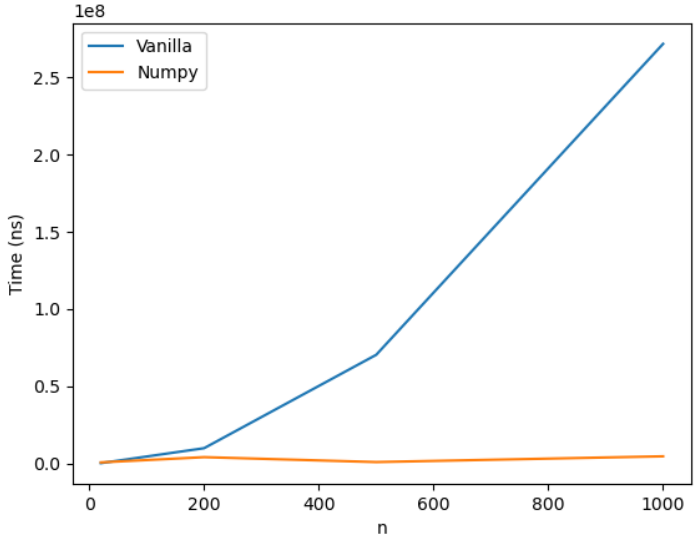

Figure 1: Plot showing the wall-clock time differences for parts a-b

```
Time for vanilla implementation: 0.126ms
Time for numpy implementation: 0.618ms
Time for vanilla implementation: 9.82ms
Time for numpy implementation: 4.006ms
Time for vanilla implementation: 70.318ms
Time for numpy implementation: 0.823ms
Time for vanilla implementation: 271.76ms
Time for numpy implementation: 4.55ms
```

Figure 2: Screenshot of the time differences reported for vanilla solution and numpy solution

For $n = 20, 200, 500, 1000$, respectively, the difference in wallclock milliseconds between the vanilla solution and numpy solution are $-0.492, 5.814, 69.495$, and 267.21. The numpy solution is almost always faster, especially for very large n. The vanilla solution is slower because it has to loop through all the elements in the matrices/vectors in order to multiply or add them together, while numpy is able to apply operations to vectors themselves. This is why as n gets large, the vanilla solution takes more time, since it has a lot more individual floats it has to address, while the numpy solution just deals with the same number of vectors and matrices themselves.

What to Submit:

- **Part c:** Plot that shows the difference in wall-clock time for parts a-b
- **Part c:** Explanation for the difference (1-2 sentences)
- **Code** on Gradescope through coding submission

A10. Two random variables X and Y have equal distributions if their CDFs, F_X and F_Y, respectively, are equal, i.e. for all x, $|F_X(x) - F_Y(x)| = 0$. The central limit theorem says that the sum of k independent, zero-mean, variance $1/k$ random variables converges to a (standard) Normal distribution as k tends to infinity. We will study this phenomenon empirically (you will use the Python packages NumPy and Matplotlib). Each of the following subproblems includes a description of how the plots were generated; these have been coded for you. The code is available in the .zip file. In this problem, you will add to our implementation to explore **matplotlib** library, and how the solution depends on n and k.

a. *[2 points]* For $i = 1, \ldots, n$ let $Z_i \sim \mathcal{N}(0, 1)$. Let $x \mapsto F(x)$ denote the true CDF from which each Z_i is drawn (i.e., Gaussian). Define $\widehat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\{Z_i \leq x\}$ for $x \in \mathbb{R}$ and we will choose n large enough such that, for all $x \in \mathbb{R}$,

$$\sqrt{\mathbb{E}\left[\left(\widehat{F}_n(x) - F(x)\right)^2\right]} \leq 0.0025 \ .$$

Plot $x \mapsto \widehat{F}_n(x)$ for x ranging from -3 to 3.

From the code, to satisfy std$= 0.0025$, we have that the minimum n is $n = 40,000$.

b. *[2 points]* Define $Y^{(k)} = \frac{1}{\sqrt{k}} \sum_{i=1}^{k} B_i$ where each B_i is equal to -1 and 1 with equal probability and the B_i's are independent. We know that each $\frac{1}{\sqrt{k}} B_i$ is zero-mean and has variance $1/k$. For each $k \in \{1, 8, 64, 512\}$ we will generate n (same as in part a) independent copies $Y^{(k)}$ and plot their empirical CDF on the same plot as part a.

As k gets larger, the empirical CDF gets closer and closer to the actual Gaussian CDF in shape. This is consistent with the idea from the central limit theorem that the sum of k independent, zero-mean, variance $1/k$ random variables will converge to a standard normal distribution as k tends to infinity.

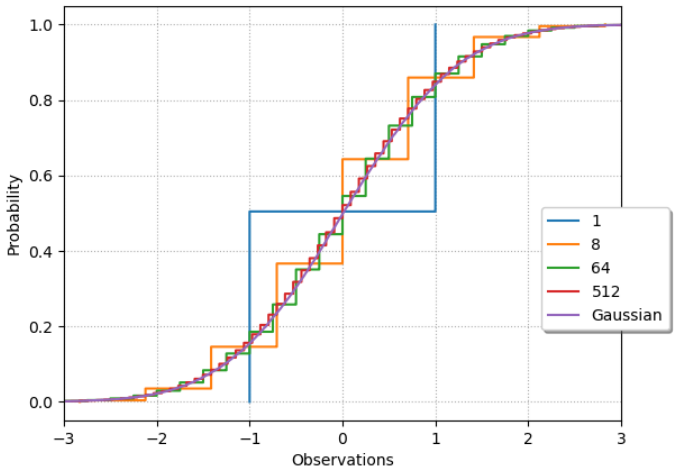

Figure 3: Plot of $x \mapsto \widehat{F}_n(x)$ for $x \in [-3, 3]$, generated by clt_with_cdfs.py

What to Submit:

- **Part a:** Value for n (You can simply print the value determined by the code provided for you). **Part b:** In 1-2 sentences: How does the empirical CDF change with k?
- **Parts a and b:** Plot of $x \mapsto \widehat{F}_n(x)$ for $x \in [-3, 3]$
- **Code** on Gradescope through coding submission