

Due Wednesday, February 8th at 11:59 pm

- Homework 2 is an entirely written assignment; no coding involved.
- We prefer that you typeset your answers using \LaTeX or other word processing software. If you haven't yet learned \LaTeX , one of the crown jewels of computer science, now is a good time! Neatly handwritten and scanned solutions will also be accepted.
- In all of the questions, **show your work**, not just the final answer.
- **Start early. This is a long assignment. Most of the material is prerequisite material not covered in lecture; you are responsible for finding resources to understand it.**

Deliverables:

1. Submit a PDF of your homework to the Gradescope assignment entitled "HW2 Write-Up". You may typeset your homework in \LaTeX or Word (submit PDF format, **not** .doc/.docx format) or submit neatly handwritten and scanned solutions. **Please start each question on a new page.** If there are graphs, include those graphs in the correct sections. **Do not** put them in an appendix. We need each solution to be self-contained on pages of its own.
 - In your write-up, please state whom you had discussions with (not counting course staff) about the homework contents.
 - In your write-up, please copy the following statement and sign your signature next to it. (Mac Preview, PDF Expert, and FoxIt PDF Reader, among others, have tools to let you sign a PDF file.) We want to make it *extra* clear so that no one inadvertently cheats.
"I certify that all solutions are entirely in my own words and that I have not looked at another student's solutions. I have given credit to all external sources I consulted."

1 Identities and Inequalities with Expectation

For this exercise, the following identity might be useful: for a probability event A , $\mathbb{P}(A) = \mathbb{E}[\mathbf{1}\{A\}]$, where $\mathbf{1}\{\cdot\}$ is the indicator function.

1. Let X be a random variable with density $f(x) = \lambda e^{-\lambda x} \mathbf{1}\{x > 0\}$. Show that $\mathbb{E}[X^k] = k!/\lambda^k$ for integer $k \geq 0$. *Hint:* One way to attempt this problem is by using induction on k .
2. For any non-negative random variable X and constant $t > 0$, show that $\mathbb{P}(X \geq t) \leq \mathbb{E}[X]/t$. This is known as Markov's inequality.
Hint: show that for $a, b > 0$, $\mathbf{1}\{a \geq b\} \leq a/b$.

3. For any non-negative random variable X , prove the identity

$$\mathbb{E}[X] = \int_0^{\infty} \mathbb{P}(X \geq t) dt.$$

You may assume that X admits a density function $f(x)$; but we note that there is a proof of the identity that works even if X does not have a density function.

4. For any non-negative random variable X with finite variance (i.e., $\mathbb{E}[X^2] < \infty$), prove that

$$\mathbb{P}(X > 0) \geq \frac{(\mathbb{E}[X])^2}{\mathbb{E}[X^2]}.$$

Hint: Use the Cauchy–Schwarz inequality $\langle u, v \rangle^2 \leq \langle u, u \rangle \langle v, v \rangle$. You have most likely seen it applied when the inner product is the real dot product; however, it holds for arbitrary inner products. Without proof, use the fact that the expectation $\mathbb{E}[UV]$ is a valid inner product of random variables U and V .

(Note that by assumption we know $\mathbb{P}(X \geq 0) = 1$, so this inequality is indeed quite powerful.)

5. For a random variable X with finite variance and $\mathbb{E}[X] = 0$, prove that

$$\mathbb{P}(X \geq t) \leq \frac{\mathbb{E}[X^2]}{\mathbb{E}[X^2] + t^2} \text{ for any } t \geq 0$$

Hint: Try using logic similar to Question 1.4 on $t - X$.

2 Probability Potpourri

1. Recall that the covariance of two random variables X and Y is defined to be $\text{Cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]$. For a multivariate random variable Z (i.e., each component of the vector Z is a random variable), we define the square covariance matrix Σ with entries $\Sigma_{ij} = \text{Cov}(Z_i, Z_j)$. Concisely, $\Sigma = \mathbb{E}[(Z - \mu)(Z - \mu)^\top]$, where μ is the mean value of the (column) vector Z . Show that the covariance matrix is always positive semidefinite (PSD). You can use the definition of PSD matrices in Q4.2.
2. The probability that an archer hits her target when it is windy is 0.4; when it is not windy, her probability of hitting the target is 0.7. On any shot, the probability of a gust of wind is 0.3. Find the probability that
 - (i) on a given shot there is a gust of wind and she hits her target.
 - (ii) she hits the target with her first shot.
 - (iii) she hits the target exactly once in two shots.
 - (iv) On an occasion when she missed, there was no gust of wind.
3. An archery target is made of 3 concentric circles of radii $1/\sqrt{3}$, 1 and $\sqrt{3}$ feet. Arrows striking within the inner circle are awarded 4 points, arrows within the middle ring are awarded 3 points, and arrows within the outer ring are awarded 2 points. Shots outside the target are awarded 0 points.

Consider a random variable X , the distance of the strike from the center in feet, and let the probability density function of X be

$$f(x) = \begin{cases} \frac{2}{\pi(1+x^2)} & x > 0 \\ 0 & \text{otherwise} \end{cases}$$

What is the expected value of the score of a single strike?

4. A random variable Z is said to be drawn from the Poisson distribution with parameter $\lambda > 0$ if it takes values in nonnegative integers with probability $\mathbb{P}(Z = k) = \frac{\lambda^k e^{-\lambda}}{k!}$. Let X and Y be two independent Poisson random variables with parameters $\lambda > 0$ and $\mu > 0$ respectively. Derive an expression for $\mathbb{P}(X | X + Y = n)$. What well-known probability distribution is this? What are its parameters?

3 Properties of the Normal Distribution (Gaussians)

1. Prove that $\mathbb{E}[e^{\lambda X}] = e^{\sigma^2 \lambda^2 / 2}$, where $\lambda \in \mathbb{R}$ is a constant, and $X \sim \mathcal{N}(0, \sigma^2)$. As a function of λ , $\mathbb{E}[e^{\lambda X}]$ is also known as the *moment-generating function*.
2. *Concentration inequalities* are inequalities that place upper bounds on the likelihood that a random variable X is far away from its mean μ , written $\mathbb{P}(|X - \mu| \geq t)$, with a falling exponential function ae^{-bt^2} having constants $a, b > 0$. Such inequalities imply that X is very likely to be close to its mean μ . To make a tight bound, we want a to be as small and b to be as large as possible.

For $t > 0$ and $X \sim \mathcal{N}(0, \sigma^2)$, prove that $\mathbb{P}(X \geq t) \leq \exp(-t^2/2\sigma^2)$, then show that $\mathbb{P}(|X| \geq t) \leq 2 \exp(-t^2/2\sigma^2)$.

Hint: Consider using Markov's inequality and the result from Question 3.1.

3. Let $X_1, \dots, X_n \sim \mathcal{N}(0, \sigma^2)$ be i.i.d. (independent and identically distributed). Find a concentration inequality, similar to Question 3.2, for the average of n Gaussians: $\mathbb{P}(\frac{1}{n} \sum_{i=1}^n X_i \geq t)$? What happens as $n \rightarrow \infty$?

Hint: Without proof, use the fact that linear combinations of i.i.d. Gaussian-distributed variables are also Gaussian-distributed. Be warned that summing two Gaussian variables does **not** mean that you can sum their probability density functions (no no no!).

4. Let $X \in \mathbb{R}^n \sim \mathcal{N}(0, \sigma^2 I_n)$ be an n -dimensional Gaussian random variable, where I_n denotes the $n \times n$ identity matrix. You may interpret X as a (column) vector whose entries are i.i.d. real values drawn from the scalar Gaussian $\mathcal{N}(0, \sigma^2)$. Given a constant (i.e., not random) matrix $A \in \mathbb{R}^{n \times n}$ and a constant vector $b \in \mathbb{R}^n$, derive the mean (which is a vector) and covariance matrix of $Y = AX + b$. Without proof, use the fact that any linear transformation of a Gaussian random variable is also a Gaussian random variable.

5. Let vectors $u, v \in \mathbb{R}^n$ be constant (i.e., not random) and orthogonal (i.e., $\langle u, v \rangle = u \cdot v = 0$). Let $X = (X_1, \dots, X_n)$ be a vector of n i.i.d. standard Gaussians, $X_i \sim \mathcal{N}(0, 1), \forall i \in [n]$. Let $u_x = \langle u, X \rangle$ and $v_x = \langle v, X \rangle$. Are u_x and v_x independent? Explain. If X_1, \dots, X_n are independently but not identically distributed, say $X_i \sim \mathcal{N}(0, i)$, does the answer change?

Hint: Two Gaussian random variables are independent if and only if they are uncorrelated.

6. Prove that $\mathbb{E}[\max_{1 \leq i \leq n} |X_i|] \leq C \sqrt{\log(2n)}\sigma$ for some constant $C \in \mathbb{R}$, where $X_1, \dots, X_n \sim \mathcal{N}(0, \sigma^2)$ are i.i.d. (Interestingly, a similar lower bound holds: $\mathbb{E}[\max_{1 \leq i \leq n} |X_i|] \geq C' \sqrt{\log(2n)}\sigma$ for some C' ; but you don't need to prove the lower bound).

Hint: Use Jensen's inequality: $f(\mathbb{E}[Y]) \leq \mathbb{E}[f(Y)]$ for any convex function f .

4 Linear Algebra Review

1. First we review some basic concepts of rank and elementary matrix operations. Let $A \in \mathbb{R}^{m \times n}$ and $B \in \mathbb{R}^{n \times p}$. Let I_n denote the $n \times n$ identity matrix.

(a) Perform elementary row and column operations to transform $\begin{bmatrix} I_n & 0 \\ 0 & AB \end{bmatrix}$ to $\begin{bmatrix} B & I_n \\ 0 & A \end{bmatrix}$.

(b) Use part (a) to prove that $\text{rank } A + \text{rank } B - n \leq \text{rank}(AB) \leq \min\{\text{rank } A, \text{rank } B\}$.

(c) Using only the ideas of the row space and column space of A (and their relationship to matrix-vector multiplication), explain why $\text{rank}(A^T A) = \text{rank } A$.

2. Let $A \in \mathbb{R}^{n \times n}$ be a symmetric matrix. Prove equivalence between these three different definitions of positive semidefiniteness (PSD). Note that when we talk about PSD matrices, they are defined to be symmetric matrices. There are nonsymmetric matrices that exhibit PSD properties, like the first definition below, but not all three.

(a) For all $x \in \mathbb{R}^n$, $x^T A x \geq 0$.

(b) All the eigenvalues of A are nonnegative.

(c) There exists a matrix $U \in \mathbb{R}^{n \times n}$ such that $A = U U^T$.

Positive semidefiniteness will be denoted as $A \geq 0$.

3. Recall that the (Frobenius) inner product between two matrices of the same dimensions $A, B \in \mathbb{R}^{m \times n}$ is defined to be $\langle A, B \rangle = \text{trace}(A^T B)$ (sometimes written $\langle A, B \rangle_F$ to be clear). The Frobenius norm of a matrix is defined to be $\|A\|_F = \sqrt{\sum_{i=1}^m \sum_{j=1}^n |A_{ij}|^2}$. Prove some of the following matrix identities. The Cauchy–Schwarz inequality and the identities above may be helpful to you.

(a) $x^T A y = \langle A, x y^T \rangle$ for all $x \in \mathbb{R}^m, y \in \mathbb{R}^n, A \in \mathbb{R}^{m \times n}$.

(b) $\langle A, B \rangle \leq \|A\|_F \|B\|_F$ for all $A, B \in \mathbb{R}^{m \times n}$.

(c) If A and B are symmetric PSD matrices, then $\text{trace}(AB) \geq 0$, where $\text{trace } M$ denotes the trace of M .

(d) If $A, B \in \mathbb{R}^{n \times n}$ are real symmetric matrices with $\lambda_{\max}(A) \geq 0$ and B being PSD, then $\langle A, B \rangle \leq \sqrt{n} \lambda_{\max}(A) \|B\|_F$.

Hint: Construct a PSD matrix using $\lambda_{\max}(A)$

4. If $M - N \geq 0$ and both M and N are positive definite, is $N^{-1} - M^{-1}$ PSD? Show your work.
5. Let $A \in \mathbb{R}^{m \times n}$ be an arbitrary matrix. Recall that the maximum singular value of A is defined as $\sigma_{\max}^2(A) = \lambda_{\max}(A^T A) = \lambda_{\max}(A A^T)$. Please prove that:

$$\sigma_{\max}(A) = \max_{u \in \mathbb{R}^m, v \in \mathbb{R}^n, \|u\|=1, \|v\|=1} (u^T A v).$$

5 Matrix/Vector Calculus and Norms

1. Consider a 2×2 matrix A , written in full $\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}$, and two arbitrary 2-dimensional vectors x, y . Calculate the derivative of

$$\sin(A_{11}^2 + e^{A_{11}+A_{22}}) + x^\top Ay,$$

with respect to the matrix A . Hint: The dimension of the derivative should match that of A and use the chain rule.

2. Aside from norms on vectors, we can also impose norms on matrices. Besides the Frobenius norm, the most common kind of norm on matrices is called the induced norm. Induced norms are defined to be

$$\|A\|_p = \sup_{x \neq 0} \frac{\|Ax\|_p}{\|x\|_p}$$

where the notation $\|\cdot\|_p$ on the right-hand side denotes the vector ℓ_p -norm. Please give the closed-form (or the most simple) expressions for the following induced norms of $A \in \mathbb{R}^{m \times n}$.

- (a) $\|A\|_2$. (Hint: Similar to Question 4.5)
 - (b) $\|A\|_\infty$.
3. (a) Let $\alpha = \sum_{i=1}^n y_i \ln \beta_i$ for $y, \beta \in \mathbb{R}^n$. What are the partial derivatives $\frac{\partial \alpha}{\partial \beta_i}$?
 (b) Let $\gamma = A\rho + b$ for $b \in \mathbb{R}^n, \rho \in \mathbb{R}^m, A \in \mathbb{R}^{n \times m}$. What are the the partial derivatives $\frac{\partial \gamma_i}{\partial \rho_j}$?
 (c) Given $x \in \mathbb{R}^n, y \in \mathbb{R}^m, z \in \mathbb{R}^k$ and $y = f(x), f: \mathbb{R}^n \mapsto \mathbb{R}^m, z = g(y), g: \mathbb{R}^m \mapsto \mathbb{R}^k$. Please write the Jacobian $\frac{dz}{dx}$ as the product of two other matrices. What are these matrices?
 (d) Given $x \in \mathbb{R}^n, y, z \in \mathbb{R}^m$, and $y = f(x), z = g(x)$. Write the gradient $\nabla_x y^\top z$ in terms of y and z and some other terms.
 4. Consider a differentiable function $f: \mathbb{R}^n \mapsto \mathbb{R}$. Suppose this function admits a unique global optimum $x^* \in \mathbb{R}^n$. Suppose that for some spherical region $\mathcal{X} = \{x \mid \|x - x^*\|^2 \leq D\}$ around x^* for some constant D , the Hessian matrix H of the function $f(x)$ is PSD and its maximum eigenvalue is 1. Prove that

$$f(x) - f(x^*) \leq \frac{D}{2}$$

for every $x \in \mathcal{X}$. Hint: Look up Taylor's Theorem with Remainder. Use Mean Value Theorem on the second order term instead of the first order term, which is what is usually done.

5. Let $X \in \mathbb{R}^{n \times d}$ be a *design matrix*, consisting of n sample points (one per row of X), each of which has d features. Let $y \in \mathbb{R}^n$ be a vector of labels. We wish to find the *best linear approximation*, i.e., we want to find the w that minimizes the cost function $L(w) = \|y - Xw\|_2^2$. Assuming that X has full column rank, compute $w^* = \operatorname{argmin}_w L(w)$ in terms of X and y .

6. (**Optional bonus question** worth 1 point. This question contains knowledge that goes beyond the scope of this course, and is intended as an exercise to really make you comfortable with matrix calculus). Consider a differentiable function $f : \mathbb{R}^{m \times n} \mapsto \mathbb{R}$ that is Lipschitz continuous. You are given arbitrary matrices $X, X' \in \mathbb{R}^{m \times n}$ where $X' = X - \eta \nabla_X f(X)$ for some constant $\eta > 0$. Let $X(t) = X - t\eta \nabla_X f(X)$ for all $t \in [0, 1]$. By the Lipschitz property and some assumptions on η , we know that

$$\|\nabla_X f(X(t)) - \nabla_X f(X)\|_F \leq t \|\nabla_X f(X)\|_F.$$

Prove that

$$f(X') \leq f(X) - \frac{\eta}{2} \|\nabla_X f(X)\|_F^2.$$

Hint: Invoke the fundamental theorem of calculus.

6 Gradient Descent

Consider the optimization problem $\min_{x \in \mathbb{R}^n} \frac{1}{2} x^\top A x - b^\top x$, where $A \in \mathbb{R}^{n \times n}$ is a PSD matrix with $0 < \lambda_{\min}(A) \leq \lambda_{\max}(A) < 1$.

1. Find the optimizer x^* (in closed form).
2. Solving a linear system directly using Gaussian elimination takes $O(n^3)$ time, which may be wasteful if the matrix A is sparse. For this reason, we will use gradient descent to compute an approximation to the optimal point x^* . Write down the update rule for gradient descent with a step size of 1 (i.e., taking a step whose length is the length of the gradient).
3. Show that the iterates $x^{(k)}$ satisfy the recursion $x^{(k)} - x^* = (I - A)(x^{(k-1)} - x^*)$.
4. Using Question 4.5, prove $\|Ax\|_2 \leq \lambda_{\max}(A)\|x\|_2$.
Hint: Use the fact that, if λ is an eigenvalue of A , then λ^2 is an eigenvalue of A^2 .
5. Using the previous two parts, show that for some $0 < \rho < 1$,

$$\|x^{(k)} - x^*\|_2 \leq \rho \|x^{(k-1)} - x^*\|_2.$$

6. Let $x^{(0)} \in \mathbb{R}^n$ be the starting value for our gradient descent iterations. If we want a solution $x^{(k)}$ that is $\epsilon > 0$ close to x^* , i.e. $\|x^{(k)} - x^*\|_2 \leq \epsilon$, then how many iterations of gradient descent should we perform? In other words, how large should k be? Give your answer in terms of ρ , $\|x^{(0)} - x^*\|_2$, and ϵ .