CS 189/289A Introduction to Machine Learning
Spring 2025 Jonathan Shewchuk

HW2: I ♥ Math

Due Wednesday, February 12 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

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 main question Q2, Q3, etc. on a new page.** (You don't have to start Q2.2, Q2.3, etc. on a new page.)

- 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 Honor Code

1. **List all collaborators. If you worked alone, then you must explicitly state so.**

2. **Declare and sign the following statement**:

   "I certify that all solutions in this document are entirely my own and that I have not looked at anyone else's solution. I have given credit to all external sources I consulted."

   Signature : ____

   While discussions are encouraged, *everything* in your solution must be your (and only your) creation. Furthermore, all external material (i.e., *anything* outside lectures and assigned readings, including figures and pictures) should be cited properly. We wish to remind you that the consequences of academic misconduct are *particularly severe*!

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 Q3.2.

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. Consider a discrete random variable X that takes on a value from a finite sample space $\mathcal{Z}$. For a value $x \in \mathcal{Z}$, let $p(x)$ be the probability that X takes the value x. The *entropy* of X is

$$H(X) = -\sum_{x \in \mathcal{Z}} p(x) \ln p(x).$$

   The base of the logarithm depends on context; here we use the natural logarithm $\ln p = \log_e p$. (Information theory often uses $\log_2$ so that the units are bits; thermodynamics uses the natural logarithm.) To handle a probability of zero, we adopt the convention that $0 \cdot \ln 0 = 0$. (If you're inclined, try calculating $\lim_{p \to 0^+} p \ln p$ to see why this makes sense.) Intuitively, entropy measures how unpredictable a random variable is. We will use it in decision trees.

   (i) First, let's consider a random variable from a Bernoulli distribution, which has only two possible states. Let $X \sim \text{Bernoulli}(p)$, $p \in (0, 1)$. Show that $H(X)$ is concave in p. (That is, $-H(X)$ is a convex function of p.)

(ii) This concavity propery generalizes to discrete distributions with more than two states: the entropy of any discrete random variable is concave in its probability mass function (PMF; you don't need to prove this). Consider a sample space $\mathcal{Z}$ with n states, a discrete distribution over those states, and a random variable X drawn from that distribution. Show that among all possible PMFs over n states, the entropy $H(X)$ is maximized by the **uniform** distribution. What is $H(X)$ for that distribution?

Hint: There are n probabilities, which all add to one. Your goal is to show that the probabilities are all equal at the maximum. If you know what Lagrange multipliers are, you are welcome to use them, but they aren't necessary.

3 Linear Algebra Review

1. First we review some basic concepts of rank. Recall that elementary matrix operations do not change a matrix's rank. 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[1] to transform $\begin{bmatrix} I_n & 0 \\ 0 & AB \end{bmatrix}$ to $\begin{bmatrix} B & I_n \\ 0 & A \end{bmatrix}$.

   (b) Let's find lower and upper bounds on $\operatorname{rank}(AB)$. Use part (a) to prove that $\operatorname{rank} A + \operatorname{rank} B - n \le \operatorname{rank}(AB)$. Then use what you know about the relationship between the column space (range) and/or row space of AB and the column/row spaces for A and B to argue that $\operatorname{rank}(AB) \le \min\{\operatorname{rank} A, \operatorname{rank} B\}$.

   (c) If a matrix A has rank r, then some $r \times r$ submatrix M of A has a nonzero determinant. Use this fact to show the standard facts that the dimension of A's column space is at least r, and the dimension of A's row space is at least r. (Note: You will not receive credit for other ways of showing those same things.)

   (d) It is a fact that $\operatorname{rank}(A^\top A) = \operatorname{rank} A$; here's one way to see that. We've already seen in part (b) that $\operatorname{rank}(A^\top A) \le \operatorname{rank} A$. Suppose that $\operatorname{rank}(A^\top A)$ were strictly less than $\operatorname{rank} A$. What would that tell us about the relationship between the column space of A and the null space of $A^\top$? What standard fact about the fundamental subspaces of A says that relationship is impossible?

   (e) Given a set of vectors $S \subseteq \mathbb{R}^n$, let $AS = \{Av : v \in s\}$ denote the subset of $\mathbb{R}^m$ found by applying A to every vector in S. In terms of the ideas of the column space (range) and row space of A: What is $A\mathbb{R}^n$, and why? (Hint: what are the definitions of column space and row space?) What is $A^\top A\mathbb{R}^n$, and why? (Your answer to the latter should be purely in terms of the fundamental subspaces of A itself, not in terms of the fundamental subspaces of $A^\top A$.)

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 $x \in \mathbb{R}^n$, $x^\top Ax \ge 0$.

   (b) All the eigenvalues of A are nonnegative.

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

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

3. Consider the quadratic form $Q_A(x) = x^\top Ax$. **Geometrically** describe the following sets of solutions. For example, consider

$$A = I \quad \text{and} \quad x \in \mathbb{R}^2.$$

[1] If you're not familiar with these, https://stattrek.com/matrix-algebra/elementary-operations is a decent introduction.

A geometric description of the solutions to $Q_A(x) = 1$ would be, “a circle centered at the origin with radius 1.” Make sure your description is complete enough to fully describe the set of solutions. If the set of solutions is an ellipse, describe the directions and lengths of the major and minor axes.

(a) $A = 2I, x \in \mathbb{R}^3, Q_A(x) = 32.$

(b) $A = aa^\top$ where $a \in \mathbb{R}^3$ is a nonzero vector, $x \in \mathbb{R}^3$, $Q_A(x) = 25.$

(c) $A = \begin{bmatrix} 5 & 1 \\ 1 & 5 \end{bmatrix}, x \in \mathbb{R}^2, Q_A(x) = 24.$

(d) $A = \begin{bmatrix} 5 & 1 \\ 1 & 5 \end{bmatrix}, x \in \mathbb{R}^2, Q_A(x) = 0.$

4. The Frobenius inner product between two matrices of the same dimensions $A, B \in \mathbb{R}^{m\times n}$ is

$$\langle A, B \rangle = \text{trace}\,(A^\top B) = \sum_{i=1}^{m} \sum_{j=1}^{n} A_{ij} B_{ij},$$

where trace M denotes the *trace* of M, which you should look up if you don’t already know it. (The norm is sometimes written $\langle A, B \rangle_F$ to be clear.) The Frobenius norm of a matrix is

$$||A||_F = \sqrt{\langle A, A \rangle} = \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} |A_{ij}|^2}.$$

Prove the following. The Cauchy-Schwarz inequality, the cyclic property of the trace, and the definitions in part 2 above may be helpful to you.

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

(b) If A and B are symmetric PSD matrices, then trace$\,(AB) \geq 0$.

(c) 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 Matrix/Vector Calculus

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

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

   with respect to the matrix A.
   Hint: The gradient has the same dimensions as A. Use the chain rule.

2. (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.

3. 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 Linear Neural Networks

Let's apply the multivariate chain rule to a "simple" type of neural network called a *linear neural network*. They're not very powerful, as they can learn only linear regression functions or decision functions, but they're a good stepping stone for understanding more complicated neural networks. We are given an $n \times d$ *design matrix* X. Each row of X is a training point, so X represents n training points with d features each. We are also given an $n \times k$ matrix Y. Each row of Y is a set of k labels for the corresponding training point in X. Our goal is to learn a $k \times d$ matrix W of weights[2] such that

$$Y \approx XW^\top.$$

If n is larger than d, typically there is no W that achieves equality, so we seek an approximate answer. We do that by finding the matrix W that minimizes the *cost function*

$$\mathrm{RSS}(W) = \|XW^\top - Y\|_F^2. \tag{1}$$

This is a classic *least-squares linear regression* problem; most of you have seen those before. But we are solving k linear regression problems simultaneously, which is why Y and W are matrices instead of vectors.

Linear neural networks. Instead of optimizing W over the space of $k \times d$ matrices directly, we write the W we seek as a product of multiple matrices. This parameterization is called a *linear neural network*.

$$W = \mu(W_L, W_{L-1}, \ldots, W_2, W_1) = W_L W_{L-1} \cdots W_2 W_1.$$

Here, μ is called the *matrix multiplication map* (hence the Greek letter mu) and each W_j is a real-valued $d_j \times d_{j-1}$ matrix. Recall that W is a $k \times d$ matrix, so $d_L = k$ and $d_0 = d$. L is the number of *layers* of "connections" in the neural network. You can also think of the network as having $L + 1$ layers of units: $d_0 = d$ units in the *input layer*, d_1 units in the first *hidden layer*, d_{L-1} units in the last hidden layer, and $d_L = k$ units in the *output layer*.

We collect all the neural network's weights in a *weight vector* $\theta = (W_L, W_{L-1}, \ldots, W_1) \in \mathbb{R}^{d_\theta}$, where $d_\theta = d_L d_{L-1} + d_{L-1} d_{L-2} + \ldots + d_1 d_0$ is the total number of real-valued weights in the network. Thus we can write $\mu(\theta)$ to mean $\mu(W_L, W_{L-1}, \ldots, W_1)$. But you should imagine θ as a column vector: we take all the components of all the matrices $W_L, W_{L-1}, \ldots, W_1$ and just write them all in one very long column vector. Given a fixed weight vector θ, the linear neural network takes an *input vector* $x \in \mathbb{R}^{d_0}$ and returns an *output vector* $y = W_L W_{L-1} \cdots W_2 W_1 x = \mu(\theta)x \in \mathbb{R}^{d_L}$.

Now our goal is to find a weight vector θ that minimizes the composition $\mathrm{RSS} \circ \mu$—that is, it minimizes the cost function

$$J(\theta) = \mathrm{RSS}(\mu(\theta)).$$

We are substituting a linear neural network for W and optimizing the weights in θ instead of directly optimizing the components of W. This makes the optimization problem harder to solve, and you

[2] The reason for the transpose on $W^\top$ is because we think in terms of applying W to an individual training point. Indeed, if $X_i \in \mathbb{R}^d$ and $Y_i \in \mathbb{R}^k$ respectively denote the i-th rows of X and Y transposed to be column vectors, then we can write $Y_i \approx WX_i$. For historical reasons, most papers in the literature use design matrices whose rows are sample points, rather than columns.

would never solve least-squares linear regression problems this way in practice; but again, it is a good exercise to work toward understanding the behavior of "real" neural networks in which μ is *not* a linear function.

We would like to use a gradient descent algorithm to find θ, so we will derive $\nabla_\theta J$ as follows.

1. The gradient $G = \nabla_W \mathrm{RSS}(W)$ is a $k \times d$ matrix whose entries are $G_{ij} = \partial \mathrm{RSS}(W)/\partial W_{ij}$, where $\mathrm{RSS}(W)$ is defined by Equation (1). Write two explicit formulas for $\nabla_W \mathrm{RSS}(W)$. First, derive a formula for each G_{ij} using summations, simplified as much as possible. Use that result to find a simple formula for $\nabla_W \mathrm{RSS}(W)$ in matrix notation with no summations.

2. Directional derivatives are closely related to gradients. The notation $\mathrm{RSS}'_{\Delta W}(W)$ denotes the directional derivative of $\mathrm{RSS}(W)$ in the direction ΔW, and the notation $\mu'_{\Delta\theta}(\theta)$ denotes the directional derivative of $\mu(\theta)$ in the direction $\Delta\theta$.[3] Informally speaking, the directional derivative $\mathrm{RSS}'_{\Delta W}(W)$ tells us how much $\mathrm{RSS}(W)$ changes if we increase W by an infinitesimal displacement $\Delta W \in \mathbb{R}^{k \times d}$. (However, any ΔW we can actually specify is not actually infinitesimal; $\mathrm{RSS}'_{\Delta W}(W)$ is a local linearization of the relationship between W and $\mathrm{RSS}(W)$ at W. To a physicist, $\mathrm{RSS}'_{\Delta W}(W)$ tells us the initial velocity of change of $\mathrm{RSS}(W)$ if we start changing W with velocity ΔW.)

   Show how to write $\mathrm{RSS}'_{\Delta W}(W)$ as a Frobenius inner product of two matrices, one related to part (a).

3. In principle, we could take the gradient $\nabla_\theta\, \mu(\theta)$, but we would need a 3D array to express it! As I don't know a nice way to write it, we'll jump directly to writing the directional derivative $\mu'_{\Delta\theta}(\theta)$. Here, $\Delta\theta \in \mathbb{R}^{d_\theta}$ is a weight vector whose matrices we will write $\Delta\theta = (\Delta W_L, \Delta W_{L-1}, \ldots, \Delta W_1)$. Show that

$$\mu'_{\Delta\theta}(\theta) = \sum_{j=1}^{L} W_{>j} \Delta W_j W_{<j}$$

   where $W_{>j} = W_L W_{L-1} \cdots W_{j+1}$, $W_{<j} = W_{j-1} W_{j-2} \cdots W_1$, and we use the convention that $W_{>L}$ is the $d_L \times d_L$ identity matrix and $W_{<1}$ is the $d_0 \times d_0$ identity matrix.

   Hint: although μ is not a linear function of θ, μ is linear in any *single* W_j; and any directional derivative of the form $\mu'_{\Delta\theta}(\theta)$ is linear in $\Delta\theta$ (for a fixed θ).

4. Recall the chain rule for scalar functions, $\frac{\mathrm{d}}{\mathrm{d}x} f(g(x))|_{x=x_0} = \frac{\mathrm{d}}{\mathrm{d}y} f(y)|_{y=g(x_0)} \cdot \frac{\mathrm{d}}{\mathrm{d}x} g(x)|_{x=x_0}$. There is a multivariate version of the chain rule, which we hope you remember from some class you've taken, and the multivariate chain rule can be used to chain directional derivatives. Write out the chain rule that expresses the directional derivative $J'_{\Delta\theta}(\theta)|_{\theta=\theta_0}$ by composing your directional derivatives for RSS and μ, evaluated at a weight vector θ_0. (Just write the pure form of the chain rule without substituting the values of those directional derivatives; we'll substitute the values in the next part.)

[3] "ΔW" and "$\Delta\theta$" are just variable names that remind us to think of these as small displacements of W or θ; the Greek letter delta is not an operator nor a separate variable.

5. Now substitute the values you derived in parts (b) and (c) into your expression for $J'_{\Delta\theta}(\theta)$ and use it to show that

$$\begin{aligned}\nabla_\theta J(\theta) \quad = \quad & (2\,(\mu(\theta)\,X^\top - Y^\top)XW_{<L}^\top, \ldots, \\ & 2W_{>j}^\top\,(\mu(\theta)\,X^\top - Y^\top)XW_{<j}^\top, \ldots, \\ & 2W_{>1}^\top\,(\mu(\theta)\,X^\top - Y^\top)X).\end{aligned}$$

This gradient is a vector in $\mathbb{R}^{d_\theta}$ written in the same format as $(W_L, \ldots, W_j, \ldots, W_1)$. Note that the values $W_{>j}$ and $W_{<j}$ here depend on θ.

Hint: you might find the cyclic property of the trace handy.

6 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, $M_X(\lambda) = \mathbb{E}[e^{\lambda X}]$ is also known as the *moment-generating function*.

2. 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 $X = (X_1, \ldots, X_n)$ 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 $X_1, \ldots, X_n$ are independently but not identically distributed, say $X_i \sim \mathcal{N}(0, i)$, does the answer change? *Hint*: two jointly normal random variables are independent if and only if they are uncorrelated.

7 The Multivariate Normal Distribution

The multivariate normal distribution with mean $\mu \in \mathbb{R}^d$ and positive definite covariance matrix $\Sigma \in \mathbb{R}^{d\times d}$, denoted $\mathcal{N}(\mu, \Sigma)$, has the probability density function

$$f(x; \mu, \Sigma) = \frac{1}{\sqrt{(2\pi)^d |\Sigma|}} \exp\left(-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)\right).$$

Here $|\Sigma|$ denotes the determinant of Σ. You may use the following facts without proof.

- The volume under the normal PDF is 1.

$$\int_{\mathbb{R}^d} f(x)\,\mathrm{d}x = \int_{\mathbb{R}^d} \frac{1}{\sqrt{(2\pi)^d|\Sigma|}} \exp\left\{-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)\right\} \mathrm{d}x = 1.$$

- The change-of-variables formula for integrals: let f be a smooth function from $\mathbb{R}^d \to \mathbb{R}$, let $A \in \mathbb{R}^{d\times d}$ be an invertible matrix, and let $b \in \mathbb{R}^d$ be a vector. Then, performing the change of variables $x \mapsto z = Ax + b$,

$$\int_{\mathbb{R}^d} f(x)\,\mathrm{d}x = \int_{\mathbb{R}^d} f(A^{-1}z - A^{-1}b)\,|A^{-1}|\,\mathrm{d}z.$$

1. Let $X \sim \mathcal{N}(\mu, \Sigma)$. Use a suitable change of variables to show that $\mathbb{E}[X] = \mu$.

2. Use a suitable change of variables to show that $\mathrm{Var}(X) = \Sigma$, where the variance of a vector-valued random variable X is

$$\mathrm{Var}(X) = \mathrm{Cov}(X, X) = \mathbb{E}[(X-\mu)\,(X-\mu)^\top] = \mathbb{E}[XX^\top] - \mu\,\mu^\top.$$

   Hints: Every symmetric, positive semidefinite matrix Σ has a symmetric, positive definite square root $\Sigma^{1/2}$ such that $\Sigma = \Sigma^{1/2}\Sigma^{1/2}$. Note that Σ and $\Sigma^{1/2}$ are invertible. After the change of variables, you will have to find another variance $\mathrm{Var}(Z)$; if you've chosen the right change of variables, you can solve that by solving the integral for each diagonal component of $\mathrm{Var}(Z)$ and a second integral for each off-diagonal component. The diagonal components will require integration by parts.

8 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 3.4, 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.

A Vector Calculus Appendix

Let us first understand the definition of the derivative. Let $f : \mathbb{R}^d \to \mathbb{R}$ denote a scalar function. Then the derivative $\frac{\partial f}{\partial \mathbf{x}}$ is an operator that finds us the best local first-order linear approximation of f at $\mathbf{x}$. By local, we mean adding a **small** perturbation $\Delta \in \mathbb{R}^d$ to $\mathbf{x}$. That is,

$$f(\mathbf{x} + \Delta) = f(\mathbf{x}) + \frac{\partial f}{\partial \mathbf{x}} \Delta + o(||\Delta||) \tag{2}$$

where $o(||\Delta||)$ stands for any term $r(\Delta)$ such that $r(\Delta)/||\Delta|| \to 0$ as $||\Delta|| \to 0$. An example of such a term is a quadratic term like $||\Delta||^2$. Let us quickly verify that $r(\Delta) = ||\Delta||^2$ is indeed an $o(||\Delta||)$ term. As $||\Delta|| \to 0$, we have

$$\frac{r(\Delta)}{||\Delta||} = \frac{||\Delta||^2}{||\Delta||} = ||\Delta|| \to 0,$$

thereby verifying our claim. As a rule of thumb, any term that has a higher-order dependence on $||\Delta||$ than linear is $o(||\Delta||)$ and is ignored to compute the derivative.[4]

We call $\frac{\partial f}{\partial \mathbf{x}}$ the *derivative of f at* $\mathbf{x}$. Sometimes we use $\frac{df}{d\mathbf{x}}$ but it we use ∂ to indicate that f may depend on some other variable too. (But to define $\frac{\partial f}{\partial \mathbf{x}}$, we study changes in f with respect to changes in only $\mathbf{x}$.)

Since Δ is a column vector the vector $\frac{\partial f}{\partial \mathbf{x}}$ should be a row vector so that $\frac{\partial f}{\partial \mathbf{x}}\Delta$ is a scalar. The gradient of f at $\mathbf{x}$ is defined to be the transpose of this derivative. That is $\nabla f(\mathbf{x}) = \left(\frac{\partial f}{\partial \mathbf{x}}\right)^\top$. One way to compute the derivative is to expand out $f(\mathbf{x} + \Delta)$ and guess from the expression. We call this method *computation via first principle*. But we also have nice element-wise definitions.

We now write down some formulas that would be helpful to compute different derivatives in various settings where a solution via first principle might be hard to compute. We will also distinguish between the derivative, gradient, Jacobian, and Hessian in our notation.

1. Let $f : \mathbb{R}^d \to \mathbb{R}$ denote a scalar function. Let $\mathbf{x} \in \mathbb{R}^d$ denote a vector and $\mathbf{A} \in \mathbb{R}^{d \times d}$ denote a matrix. We have

$$\frac{\partial f}{\partial \mathbf{x}} \in \mathbb{R}^{1 \times d} \quad \text{such that} \quad \frac{\partial f}{\partial \mathbf{x}} = \left[\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, \dots, \frac{\partial f}{\partial x_d}\right] \tag{3}$$

$$\nabla_{\mathbf{x}} f = \left(\frac{\partial f}{\partial \mathbf{x}}\right)^\top = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \\ \vdots \\ \frac{\partial f}{\partial x_d} \end{bmatrix}. \tag{4}$$

[3] Good resources for matrix calculus are:

- The Matrix Cookbook: `https://www.math.uwaterloo.ca/~hwolkowi/matrixcookbook.pdf`
- Wikipedia: `https://en.wikipedia.org/wiki/Matrix_calculus`
- Khan Academy: `https://www.khanacademy.org/math/multivariable-calculus/multivariable-derivatives`
- YouTube: `https://www.youtube.com/playlist?list=PLSQl0a2vh4HC5feHa6Rc5c0wbRTx56nF7`.

[4] Note that $r(\Delta) = \sqrt{||\Delta||}$ is not an $o(||\Delta||)$ term. Since for this case, $r(\Delta)/||\Delta|| = 1/\sqrt{||\Delta||} \to \infty$ as $||\Delta|| \to 0$.

2. Let $y : \mathbb{R}^{m\times n} \to \mathbb{R}$ be a scalar function defined on the space of $m \times n$ matrices. Then its derivative is an $n \times m$ matrix and is given by

$$\frac{\partial y}{\partial \mathbf{B}} \in \mathbb{R}^{n\times m} \quad \text{such that} \quad \left[\frac{\partial y}{\partial \mathbf{B}}\right]_{ij} = \frac{\partial y}{\partial B_{ji}}. \tag{5}$$

   An argument via first principle follows.

$$y(\mathbf{B} + \Delta) = y(\mathbf{B}) + \operatorname{trace}\left(\frac{\partial y}{\partial \mathbf{B}}\,\Delta\right) + o(\|\Delta\|). \tag{6}$$

3. For $\mathbf{z} : \mathbb{R}^d \to \mathbb{R}^k$ a vector-valued function; its derivative $\frac{\partial \mathbf{z}}{\partial \mathbf{x}}$ is an operator such that it can help find the change in function value at $\mathbf{x}$, up to first order, when we add a little perturbation Δ to $\mathbf{x}$:

$$\mathbf{z}(\mathbf{x} + \Delta) = \mathbf{z}(\mathbf{x}) + \frac{\partial \mathbf{z}}{\partial \mathbf{x}}\,\Delta + o(\|\Delta\|). \tag{7}$$

   A formula for the same can be derived as

$$D(\mathbf{z}) = \frac{\partial \mathbf{z}}{\partial \mathbf{x}} \in \mathbb{R}^{k\times d} = \begin{bmatrix} \frac{\partial z_1}{\partial \mathbf{x}} \\ \frac{\partial z_2}{\partial \mathbf{x}} \\ \vdots \\ \frac{\partial z_k}{\partial \mathbf{x}} \end{bmatrix} = \begin{bmatrix} \frac{\partial z_1}{\partial x_1} & \frac{\partial z_1}{\partial x_2} & \cdots & \frac{\partial z_1}{\partial x_d} \\ \frac{\partial z_2}{\partial x_1} & \frac{\partial z_2}{\partial x_2} & \cdots & \frac{\partial z_2}{\partial x_d} \\ \vdots & & & \\ \frac{\partial z_k}{\partial x_1} & \frac{\partial z_k}{\partial x_2} & \cdots & \frac{\partial z_k}{\partial x_d} \end{bmatrix}, \tag{8}$$

$$\text{that is} \qquad [D(\mathbf{z})]_{ij} = \left[\frac{\partial \mathbf{z}}{\partial \mathbf{x}}\right]_{ij} = \frac{\partial z_i}{\partial x_j}. \tag{9}$$

4. The Hessian of f is defined as

$$H(f) = \nabla^2 f(\mathbf{x}) = D(\nabla f)^\top = \begin{bmatrix} \frac{\partial z_1}{\partial x_1} & \frac{\partial z_2}{\partial x_1} & \cdots & \frac{\partial z_d}{\partial x_1} \\ \frac{\partial z_1}{\partial x_2} & \frac{\partial z_2}{\partial x_2} & \cdots & \frac{\partial z_d}{\partial x_2} \\ \vdots & & & \\ \frac{\partial z_1}{\partial x_d} & \frac{\partial z_2}{\partial x_d} & \cdots & \frac{\partial z_d}{\partial x_d} \end{bmatrix} = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_d} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2^2} & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_d} \\ \vdots & & & \\ \frac{\partial^2 f}{\partial x_d \partial x_1} & \frac{\partial^2 f}{\partial x_d \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_d^2}. \end{bmatrix} \tag{10}$$

   The ”first principle” form is

$$\nabla f(\mathbf{x} + \Delta) = \nabla f(\mathbf{x}) + \nabla^2 f(\mathbf{x})\Delta + o(\|\Delta\|)$$

   or, by taking a second-order expansion of f,

$$f(\mathbf{x} + \Delta) = f(\mathbf{x}) + \nabla f(\mathbf{x})\Delta + \frac{1}{2}\Delta^\top \nabla^2 f(\mathbf{x})\Delta + o(\|\Delta\|^2).$$

   For sufficiently smooth functions (when the mixed derivatives are equal), the Hessian is a symmetric matrix by Clairaut’s theorem. In such cases (which covers most cases for our purposes), the convention with transposes does not matter.