CSCI 145 Problem 3: Three Gradients You'll Use All Semester

July 20, 2026

Submission Instructions

Please upload your work to Gradescope by **11:59pm on DATE**.

- Turn in a single PDF of *handwritten* solutions. Typed (including LaTeX) solutions will not be accepted.
- Problems are grouped into three-week **cycles**. Once per cycle, you will meet with the grader and present the solution to *one* problem from that cycle, selected at random when you arrive — so bring your work and be ready to explain any of them.

Problem 3: Three Gradients You'll Use All Semester

Linear regression, logistic regression, a neural network, and a reinforcement-learning policy are all trained by the same loop: write down a loss, take its gradient, and step downhill. Between them they use three gradient formulas, and this problem derives all three from the single definition we wrote in lecture.

For $f : \mathbb{R}^d \to \mathbb{R}$, the gradient $\nabla_{\mathbf{x}} f$ is the unique vector satisfying, for every direction $\mathbf{v} \in \mathbb{R}^d$,

$$f(\mathbf{x} + \epsilon \mathbf{v}) = f(\mathbf{x}) + \epsilon \, \langle \nabla_{\mathbf{x}} f, \mathbf{v} \rangle + o(\epsilon),$$

where $o(\epsilon)$ collects terms that shrink faster than ϵ (like ϵ^2). Every part below uses the same recipe: expand $f(\mathbf{x}+\epsilon\mathbf{v})$, and whatever vector ends up paired with $\mathbf{v}$ inside the inner product of the ϵ term *is* the gradient.

Part A: The Linear Function

Let $f(\mathbf{x}) = \mathbf{a}^\top \mathbf{x}$ for a fixed vector $\mathbf{a} \in \mathbb{R}^d$. Expand $f(\mathbf{x} + \epsilon\mathbf{v})$ and conclude that

$$\nabla_{\mathbf{x}} \, \mathbf{a}^\top \mathbf{x} = \mathbf{a}.$$

(You should find that the $o(\epsilon)$ term is exactly zero, because a linear function is its own first-order approximation.)

Part B: The Quadratic Form

Let $f(\mathbf{x}) = \mathbf{x}^\top \mathbf{A} \mathbf{x}$ for a fixed square matrix $\mathbf{A} \in \mathbb{R}^{d \times d}$. Expand $f(\mathbf{x} + \epsilon\mathbf{v})$, collect the terms linear in ϵ, and show that

$$\nabla_{\mathbf{x}} \, \mathbf{x}^\top \mathbf{A} \mathbf{x} = (\mathbf{A} + \mathbf{A}^\top)\mathbf{x}.$$

(Hint: $\mathbf{v}^\top \mathbf{A} \mathbf{x}$ is a scalar, and a scalar equals its own transpose.) Check the formula when $d = 1$ and $\mathbf{A} = a$, then write down the special case we will use constantly, where $\mathbf{A}$ is symmetric, and say in one line why it is the matrix version of $\frac{d}{dx} a x^2 = 2ax$.

Part C: Least Squares

Let $f(\mathbf{x}) = \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2$ for a fixed matrix $\mathbf{A} \in \mathbb{R}^{n \times d}$ and vector $\mathbf{b} \in \mathbb{R}^n$. First, expand the squared norm to show

$$f(\mathbf{x}) = \mathbf{x}^\top (\mathbf{A}^\top \mathbf{A}) \mathbf{x} - 2\mathbf{b}^\top \mathbf{A} \mathbf{x} + \|\mathbf{b}\|_2^2.$$

Along the way, use the outer-product form of matrix multiplication from lecture to write

$$\mathbf{A}^\top \mathbf{A} = \sum_{i=1}^{n} \mathbf{a}_i \mathbf{a}_i^\top,$$

where $\mathbf{a}_i \in \mathbb{R}^d$ is the ith *row* of $\mathbf{A}$ written as a column, so that $\mathbf{a}_i^\top$ is that row; check that the two sides have the same (j, k) entry. Then apply Parts A and B, noting that $\mathbf{A}^\top \mathbf{A}$ is symmetric, to conclude

$$\nabla_{\mathbf{x}} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2 = 2\mathbf{A}^\top (\mathbf{A}\mathbf{x} - \mathbf{b}).$$

Finally, show that for *every* direction $\mathbf{v} \in \mathbb{R}^d$,

$$\mathbf{v}^\top (\mathbf{A}^\top \mathbf{A})\mathbf{v} = \|\mathbf{A}\mathbf{v}\|_2^2 \geq 0,$$

and conclude in one sentence what that sign says about the *shape* of f along every direction, and why an algorithm that repeatedly steps against the gradient should be expected to reach the bottom.

Where this goes next. Setting $2\mathbf{A}^\top(\mathbf{A}\mathbf{x}-\mathbf{b}) = \mathbf{0}$ gives the *normal equations* of the Linear Models unit in one line, and the "expand, collect the ϵ term" recipe is how we will differentiate through entire neural networks. The matrix $\mathbf{A}^\top\mathbf{A}$ also opens the course's spectral thread: its eigenvalues are the squared singular values of $\mathbf{A}$ (next lecture), and their ratio is the condition number that decides how fast gradient descent converges (Problem 8).