Homework 9: Gradients and Gradient Descent

EECS 245, Fall 2025 at the University of Michigan

due Friday, November 7th, 2025 at 11:59PM Ann Arbor Time

Write your solutions to the following problems by either typing them up or handwriting them on another piece of paper. Homeworks are due to Gradescope by 11:59PM on the due date. See the **syllabus** for details on the slip day policy.

Homework will be evaluated not only on the correctness of your answers, but on your ability to present your ideas clearly and logically. You should always explain and justify your conclusions, using sound reasoning. Your goal should be to convince the reader of your assertions. If a question does not require explanation, it will be explicitly stated.

Before proceeding, make sure you're familiar with the **collaboration policy**.

Total Points: 10 + 12 + 9 + 13 + 12 + 12 + 12 = 80, **but max score is 60!**

Note: We've read your feedback on the Homework 8 Survey (thank you!), and are making two changes to this homework:

- To encourage you to review the solutions to previous homeworks, the first problem on the homework will ask you to reflect on some of your answers to Homework 8.
- Since recent homeworks have been taking longer than expected for students to complete, and given the impending Midterm 2, **for this homework only**, we've lowered the "denominator" from 80 to 60. **This effectively allows you to skip 1-2 problems on the homework and still earn a full score.** That said, all of the problems are relevant to Midterm 2, so if you don't complete them, make sure to review their solutions.

Problem 1: Homework 8 Solutions Review (10 pts)

Once they're available (Monday morning), review the solutions to Homework 8. Pick **two problem parts** (for example, Problem 2c and Problem 4a) from Homework 8 in which your solutions have the most room for improvement, i.e. where they have unsound reasoning, could be significantly more efficient or clearer, etc. Include a screenshot of your solution to each problem part, and in a few sentences, explain what was deficient and how it could be fixed.

Alternatively, if you think one of your solutions is significantly better than the posted one, copy it here and explain why you think it is better. If you didn't do Homework 8, choose two problem parts from it that look challenging to you, and in a few sentences, explain the key ideas behind their solutions in your own words.

This idea was borrowed from EECS 376. Thank you to those who suggested we borrow it in the Homework 8 Survey!

Problem 2: Gradient Descent Fundamentals (12 pts)

Let $f(\vec{x}) = (x_1 - 5)^2 + (x_1^2 - x_2)^2 + 1$.

a) (4 pts) Find $\nabla f(\vec{x})$, the gradient of $f(\vec{x})$.

b) (4 pts) Find the equation of the tangent plane to $f(\vec{x})$ at $\vec{x} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$. *Hint: See the solutions to Lab 9.*

c) (4 pts) To minimize $f(\vec{x})$, we'll use gradient descent. Perform one iteration of gradient descent by hand, using the initial guess $\vec{x}^{(0)} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$ and learning rate $\alpha = \frac{1}{2}$. What is $\vec{x}^{(1)}$?

Problem 3: Regularization (9 pts)

Suppose we'd like to perform multiple linear regression using the $n \times (d+1)$ design matrix X, observation vector $\vec{y} \in \mathbb{R}^n$, and parameter vector $\vec{w} \in \mathbb{R}^{d+1}$.

Instead of minimizing mean squared error to find $\vec{w}^*$, suppose we'd like to minimize the following **regularized objective function**:

$$R_{\text{ridge}}(\vec{w}) = \|\vec{y} - X\vec{w}\|^2 + \lambda\|\vec{w}\|^2$$

where $\lambda \geq 0$ is a constant. The $+\lambda\|\vec{w}\|^2$ term is called the **regularization term**.

The vector $\vec{w}^*_{\text{ridge}}$ that minimizes $R_{\text{ridge}}(\vec{w})$ will be, in general, different than the vector $\vec{w}^*$ that minimizes mean squared error without the added $+\lambda\|\vec{w}\|^2$ term, and will thus have a higher mean squared error (meaning, worse predictions on the training data). But, it turns out that the vector $\vec{w}^*_{\text{ridge}}$ **may** make better predictions on unseen test data, if we choose λ carefully, by forcing the model to be more simple and less overfit to the training data.

We will explore this idea in more depth in Homework 10, once we have enough background to fully explore. In this problem, you'll get started on your journey to understand regularization.

a) (6 pts) Find $\nabla R_{\text{ridge}}(\vec{w})$, the gradient of $R_{\text{ridge}}(\vec{w})$.
Hint: Most of the steps involved were done in Chapter 4.1, but you'll need to redo the work yourself and extend it slightly.

b) (3 pts) Find $\vec{w}^*_{\text{ridge}}$, the vector that minimizes $R_{\text{ridge}}(\vec{w})$.
Hint: Your answer should be such that if $\lambda = 0$, then $\vec{w}^*_{ridge}$ is the same as the vector $\vec{w}^*$ that minimizes mean squared error without the added $+\lambda\|\vec{w}\|^2$ term.

One of the side benefits of adding this regularization term is that a unique solution for $\vec{w}^*_{\text{ridge}}$ exists, **even if X is not full rank**!

Problem 4: Product and Chain Rules (13 pts)

Our goal in this problem is to study the behavior of the function

$$f(\vec{x}) = \frac{\vec{x}^T A \vec{x}}{\vec{x}^T \vec{x}}$$

where $x \in \mathbb{R}^n$ and A is a symmetric $n \times n$ matrix (meaning $A = A^T$). This function, called the **Rayleigh quotient**, will play an important role in Chapter 5 of the course, when we eventually study the **dimensionality reduction** problem first introduced in Chapter 1.1.

But first, we have to get a handle on a few gradient rules.

a) (4 pts) As described in the Norm and Chain Rule in Chapter 4.1, the chain rule for gradients says that if

- $g : \mathbb{R}^d \to \mathbb{R}$ is a **vector**-to-scalar function, and
- $h : \mathbb{R} \to \mathbb{R}$ is a **scalar**-to-scalar function,

then the gradient of the **vector**-to-scalar function $f(\vec{x}) = h(g(\vec{x}))$ is given by

$$\nabla f(\vec{x}) = \left(\frac{\mathrm{d}h}{\mathrm{d}x}(g(\vec{x})) \right) \nabla g(\vec{x})$$

or, perhaps more intuitively,

$$\nabla f(\vec{x}) = h'(g(\vec{x})) \nabla g(\vec{x})$$

Note that we need to pay close attention to the types of functions we're working with. $h(g(\vec{x}))$ is well-defined, but $g(h(\vec{x}))$ is not, since h doesn't take in vectors (it takes in scalars).

Find the gradients of each of the following functions.

(i) $f_1(\vec{x}) = \log(\vec{x}^T A \vec{x})$, where $\vec{x} \in \mathbb{R}^n$ and A is a symmetric $n \times n$ matrix

(ii) $f_2(\vec{x}) = e^{-\sin(\vec{a}^T \vec{x})}$, where $\vec{x}, \vec{a} \in \mathbb{R}^n$

Hint: You can use any of the three important gradient rules from Chapter 4.1 without proof.

b) (4 pts) The product rule for gradients is a natural extension of the product rule for derivatives. If $f(\vec{x}) = g(\vec{x})h(\vec{x})$, then

$$\nabla f(\vec{x}) = \nabla(g(\vec{x})h(\vec{x})) = g(\vec{x})\nabla h(\vec{x}) + h(\vec{x})\nabla g(\vec{x})$$

Find the gradients of each of the following functions.

(i) $f_3(\vec{x}) = (\vec{a} \cdot \vec{x})(\vec{b} \cdot \vec{x})$, where $\vec{x}, \vec{a}, \vec{b} \in \mathbb{R}^n$

(ii) $f_4(\vec{x}) = \vec{a}^T \vec{x} \vec{x}^T A \vec{x}$, where $\vec{x}, \vec{a} \in \mathbb{R}^n$ and A is a symmetric $n \times n$ matrix

c) (5 pts) Putting together the chain rule and product rule, find the gradient of

$$f(\vec{x}) = \frac{\vec{x}^T A \vec{x}}{\vec{x}^T \vec{x}}$$

where $x \in \mathbb{R}^n$ and A is a symmetric $n \times n$ matrix.

Problem 5: Implementing Gradient Descent (12 pts)

This problem involves writing code and submitting it to the Gradescope autograder. The goal of this problem is to give you a taste of how to implement gradient descent to find optimal model parameters.

There are two ways to access the supplemental Jupyter Notebook:

- **Option 1**: Click **here** to open `hw09.ipynb` on DataHub. Before doing so, read the instructions on the **Tech Support** page on how to use the DataHub.
- **Option 2**: Set up a Jupyter Notebook environment locally, use `git` to clone our course repository, and open `homeworks/hw09/hw09.ipynb`. For instructions on how to do this, see the **Tech Support** page of the course website.

This problem is entirely autograded; to receive credit for Problem 5 of this homework, you'll need to submit your completed notebook to the autograder on Gradescope. Your submission time for Homework 9 is the **latter** of your PDF and code submission times.

Problem 6: Convexity (12 pts)

In Chapter 4.3 (and Tuesday's video lecture), we'll introduce the formal definition of **convexity**. Intuitively, a function $f : \mathbb{R}^d \to \mathbb{R}$ is convex if its graph is a bowl-shaped surface. Formally, f is convex if and only if, for all $\vec{x}$ and $\vec{y}$ in its domain, and for any $t \in [0, 1]$,

$$f(t\vec{x} + (1-t)\vec{y}) \leq tf(\vec{x}) + (1-t)f(\vec{y})$$

This is a formal way of saying that when you connect any two points on the graph of f with a line segment, the line segment lies on or above the graph of f, never below.

The second derivative test for convexity is more convenient, but it doesn't apply to non-differentiable functions, e.g. $f(x) = |x|$ is convex, but it isn't differentiable.

For each statement below, prove that the statement is true using the formal definition above, or give a counterexample.

a) (4 pts) The sum of two convex functions must also be convex.

b) (4 pts) The difference of two convex functions must also be convex.

c) (4 pts) Suppose $f(x)$ and $g(x)$ are both scalar-to-scalar convex functions and that, for some scalar a, $f(a) = g(a)$. Then, $h(x)$ is also convex, where

$$h(x) = \begin{cases} f(x) & x \leq a \\ g(x) & x > a \end{cases}$$

Hint: The statement is false, so focus your energy on finding a counterexample.

Problem 7: Jensen's Inequality (12 pts)

As we've seen several times, the variance of a dataset $x_1, x_2, ..., x_n$ is defined

$$\sigma_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

where $\bar{x} = \text{Mean}(x_1, x_2, ..., x_n)$. By expanding the summation (as we've done in several homework problems), we find that

$$\sigma_x^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2$$

Another way of expressing this equation is

$$\sigma_x^2 = \text{Mean}(x_1^2, ..., x_n^2) - (\text{Mean}(x_1, ..., x_n))^2$$

Since $\sigma_x^2 \geq 0$, this implies that

$$\text{Mean}(x_1^2, ..., x_n^2) - (\text{Mean}(x_1, ..., x_n))^2 \geq 0 \implies \text{Mean}(x_1^2, ..., x_n^2) \geq (\text{Mean}(x_1, ..., x_n))^2$$

The inequality on the last line can be expressed more generally as

$$\boxed{\text{Mean}(g(x_1), g(x_2), ..., g(x_n)) \geq g(\text{Mean}(x_1, ..., x_n))}$$

The inequality above is known as Jensen's inequality, and is true for all **convex functions** $g(x)$. Let's see how we can use Jensen's inequality to prove something useful!

a) (2 pts) Using the second derivative test for convexity, prove that $g(x) = -\log(x)$ is convex across its entire domain, $(0, \infty)$. (It would be difficult to prove this using the formal definition of convexity.)

b) (4 pts) Using Jensen's inequality with $g(x) = -\log(x)$, prove that, for any dataset of positive numbers $x_1, x_2, ..., x_n$,

$$\frac{x_1 + x_2 + ... + x_n}{n} \geq (x_1 \cdot x_2 \cdot ... \cdot x_n)^{1/n}$$

The quantity on the left is the familiar arithmetic mean (AM), while the quantity on the right is known as the geometric mean (GM) of $x_1, x_2, ..., x_n$. The entire inequality above is known as the "AM-GM inequality".

c) (6 pts) Using Jensen's inequality with **some** convex function $g(x)$, prove that the arithmetic mean is greater than or equal to the harmonic mean for any dataset of positive numbers $x_1, x_2, ..., x_n$, i.e.

$$\frac{x_1 + x_2 + ... + x_n}{n} \geq \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + ... + \frac{1}{x_n}}$$

Note that you must prove your choice of function $g(x)$ is convex (using the second derivative test for convexity)!

Hint: You can use a function that is only convex on an interval, as long as the only inputs you pass into that function are some from that interval.