Mathematics for Machine Learning
Problem Set 3
COMS 3770: Summer A 2025

Instructor: Samuel Deng
samdeng@cs.columbia.edu
Due: 06/20/2025 11:59 PM ET.

Problem 1

A closer look at multivariable derivatives (25 points total).

In class, we introduced a couple of notions of "derivative" for multivariable functions. This problem will clarify these notions by working with a concrete example. In the most general case, a *multivariable function* $\mathbf{f} : \mathbb{R}^d \to \mathbb{R}^n$ maps from d variables to n variables. We will consider the following multivariable function $\mathbf{f} : \mathbb{R}^2 \to \mathbb{R}^2$, where $d = 2$ and $n = 2$:

$$\mathbf{f}(x, y) = \begin{bmatrix} xy \\ x^2 - y^2 \end{bmatrix}, \text{ with } f_1(x, y) = xy \quad \text{and} \quad f_2(x, y) = x^2 - y^2. \tag{1}$$

Recall from lecture that the reason why multivariable derivatives are more complicated is that there are infinitely many directions in which we can move in multiple dimensions, while, in one dimension, there were only two directions: the "negative" direction and the "positive" direction. Because of this, our first notion of derivative for multivariable functions is the *directional derivative*, the rate of change of $\mathbf{f} : \mathbb{R}^d \to \mathbb{R}^n$ in a specific direction, $\mathbf{v} \in \mathbb{R}^d$. Recall from lecture that the directional derivative of $\mathbf{f}$ at $\mathbf{x}_0$ in the direction $\mathbf{v} \in \mathbb{R}^d$ is

$$\lim_{\delta \to 0} \frac{\mathbf{f}(\mathbf{x}_0 + \delta \mathbf{v}) - \mathbf{f}(\mathbf{x}_0)}{\delta}.$$

In calculus with a single variable, $f : \mathbb{R} \to \mathbb{R}$ with $d = 1$ and $n = 1$, a "direction" $v \in \mathbb{R}$ is simply a scalar. Because of this, the directional derivative should simply generalize the single variable definition of the derivative. Problem 1(a) shows you that it indeed does.

Problem 1(a) [3 points] Consider the single-variable function $f(x) = x^2$. Find the "directional derivative" of f at $x_0 = 1$ in the direction $v = 1$ using the definition. Find the "directional derivative" of f at $x_0 = 1$ in the direction $v = -1$ using the definition. You may use any properties of limits from single-variable calculus without proof.

The directional derivative becomes more interesting when we compute it for multivariable functions.

Problem 1(b) [3 points] Compute the directional derivative of $\mathbf{f}$ in Equation (1) above at $\mathbf{x}_0 = (1, 1)$ in the direction $\mathbf{v} = (1, 1)$. Compute the directional derivative of in Equation (1) above at $\mathbf{x}_0 = (1, 1)$ in the direction $\mathbf{u} = (1/\sqrt{2}, 1/\sqrt{2})$, which is the normalized unit vector $\mathbf{u} = \mathbf{v}/\|\mathbf{v}\|$. You may use the simple fact that, to take the limit

of a vector valued function, you may take the limit of each component individually: $\lim_{\delta \to b} \mathbf{f}(\delta) = (\lim_{\delta \to b} f_1(\delta), \ldots, \lim_{\delta \to b} f_d(\delta))$.

The computation in Problem 1(b) suggests that, if we only care about the direction in which $\mathbf{f}$ is changing, we need only specify directional derivatives with unit vectors. Some of our favorite unit vectors are the unit basis vectors, $\mathbf{e}_1 = (1, 0, \ldots, 0), \ldots, \mathbf{e}_d = (0, 0, \ldots, 1)$. This leads to our second notion of multivariable derivative, the *partial derivative*. Recall from lecture that, for a function $\mathbf{f} : \mathbb{R}^d \to \mathbb{R}^n$, the ith partial derivative of $\mathbf{f}$ at $\mathbf{x} \in \mathbb{R}^d$ is the directional derivative at $\mathbf{x}$ in the direction $\mathbf{e}_i \in \mathbb{R}^d$:

$$\frac{\partial \mathbf{f}}{\partial x_i} := \lim_{\delta \to 0} \frac{\mathbf{f}(\mathbf{x} + \delta \mathbf{e}_i) - \mathbf{f}(\mathbf{x})}{\delta} = \lim_{\delta \to 0} \frac{\mathbf{f}(x_1, \ldots, x_i + \delta, \ldots, x_d) - \mathbf{f}(x_1, \ldots, x_i, \ldots, x_d)}{\delta}.$$

The partial derivative of a vector-valued function $\mathbf{f} : \mathbb{R}^d \to \mathbb{R}^n$ is a vector $\frac{\partial \mathbf{f}}{\partial x_i} \in \mathbb{R}^n$; the partial derivative of a scalar-valued function $f : \mathbb{R}^d \to \mathbb{R}$ is a scalar $\frac{\partial f}{\partial x_i} \in \mathbb{R}$. We'll use the definition once, for practice, in Problem 1(c), but, for all intents and purposes, there's a much easier way to calculate a formula for the partial derivative. Mechanically, we take the ith partial derivative by taking the derivative with respect to x_i and holding all the other variables constant.

Problem 1(c) [3 points] Compute a formula for the first partial derivative, $\frac{\partial \mathbf{f}}{\partial x} \in \mathbb{R}^2$, of the function in Equation (1) at the point $(x, y) = (1, 1)$ using the formal definition of partial derivative (write your setup and steps down in your assignment). Compute the second partial derivative, $\frac{\partial \mathbf{f}}{\partial y} \in \mathbb{R}^2$ at the point $(x, y) = (1, 1)$, of the function in Equation (1) using the usual shortcut: taking the derivative with respect to y and keeping x constant.

Associated with every vector-valued function $\mathbf{f} : \mathbb{R}^d \to \mathbb{R}^n$ is the *Jacobian matrix*, the $n \times d$ matrix composed of all the function's partial derivatives as its columns. We will denote this matrix $[\nabla \mathbf{f}(\mathbf{x}_0)] \in \mathbb{R}^{n \times d}$, where the brackets are there to remind you that this is indeed a matrix. This should now ring some alarm bells to recall that central theorem of linear algebra you proved on PS1: all matrices can be associated with a linear transformation, and vice versa. Indeed, the Jacobian is no exception — it is associated with a very particular linear transformation for $\mathbf{f}$: its (total) derivative.

Problem 1(d) [3 points] State the formula for the 2×2 Jacobian matrix of the function $\mathbf{f}$ in Equation (1) (by "formula for" we mean just using x and y for the variables). State the Jacobian matrix at the point $(x, y) = (1, 1)$. From this problem forward, you may take partial derivatives the usual way: taking the derivative with respect to your variable of choice while keeping the others constant.

The all-encompassing notion of the derivative in multiple variables is the *total derivative*, which, in many cases, is just referred to as the derivative for reasons we shall see soon. Recall

from lecture that for a function $\mathbf{f} : \mathbb{R}^d \to \mathbb{R}^n$ and a point $\mathbf{x}_0 \in \mathbb{R}^d$, if there exists some linear transformation $D\mathbf{f}_{\mathbf{x}_0} : \mathbb{R}^d \to \mathbb{R}^n$ such that

$$\lim_{\mathbf{d}\to\mathbf{0}} \frac{1}{\|\mathbf{d}\|} \left((\mathbf{f}(\mathbf{x}_0 + \mathbf{d}) - \mathbf{f}(\mathbf{x}_0)) - D\mathbf{f}_{\mathbf{x}_0}(\mathbf{d}) \right) = \mathbf{0}, \tag{2}$$

then $\mathbf{f}$ is said to be *differentiable* at $\mathbf{x}_0$ and has the *(total) derivative* $D\mathbf{f}_{\mathbf{x}_0}$. In this definition, $\mathbf{d} \in \mathbb{R}^d$ is the change in $\mathbf{x}_0$; we used $\mathbf{d}$ instead of δ for clarity. The first part, $\mathbf{f}(\mathbf{x}_0+\mathbf{d})-\mathbf{f}(\mathbf{x}_0)$, is the change in $\mathbf{f}$ going from $\mathbf{x}_0$ to $\mathbf{x}_0+\mathbf{d}$. The second part, $D\mathbf{f}_{\mathbf{x}_0}(\mathbf{d})$, is the linear transformation applied to the change in $\mathbf{x}_0$ to $\mathbf{x}_0 + \mathbf{d}$. From the perspective of an "origin point" $\mathbf{x}_0$ and *any* "destination point" $\mathbf{x}$, the definition above says that $\mathbf{f}$ is differentiable if we can find some linear transformation $D\mathbf{f}_{\mathbf{x}_0}$ such that, as $\mathbf{d} = \mathbf{x} - \mathbf{x}_0$ gets smaller and smaller,

$$D\mathbf{f}_{\mathbf{x}_0}(\mathbf{x} - \mathbf{x}_0) \approx \mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{x}_0).$$

What's a good candidate for what this linear transformation $D\mathbf{f}_{\mathbf{x}_0}$ might be? Well, we happen to have a matrix, the Jacobian, and we know that all matrices are associated with a linear transformation — could this be the one? Indeed, you will now prove the following theorem:

Theorem 1 (The Jacobian and the derivative). If $\mathbf{f} : \mathbb{R}^d \to \mathbb{R}^n$ is differentiable at $\mathbf{x}_0$, then all the partial derivatives of $\mathbf{f}$ at $\mathbf{x}_0$ exist, and the *unique* linear transformation $D\mathbf{f}_{\mathbf{x}_0}$ is given by the Jacobian $[\nabla\mathbf{f}(\mathbf{x}_0)]$ of $\mathbf{f}$.

In Theorem 1 above, we assume differentiability and need to prove that $D\mathbf{f}_{\mathbf{x}_0}$ is indeed given by the Jacobian. That is, we can assume that there does exist some (currently mysterious) $D\mathbf{f}_{\mathbf{x}_0}$ that satisfies Equation (2). Recall from our "central theorem of linear algebra" you proved in PS1 that $D\mathbf{f}_{\mathbf{x}_0} : \mathbb{R}^d \to \mathbb{R}^n$ is associated with a *unique* matrix, which we'll denote $[D\mathbf{f}_{\mathbf{x}_0}] \in \mathbb{R}^{n\times d}$, using brackets to differentiate the matrix from the linear transformation itself. The transformation and the matrix agree on all inputs:

$$D\mathbf{f}_{\mathbf{x}_0}(\mathbf{u}) = [D\mathbf{f}_{\mathbf{x}_0}]\mathbf{u} \quad \text{for all } \mathbf{u} \in \mathbb{R}^d.$$

Therefore, our strategy will be to prove that $[D\mathbf{f}_{\mathbf{x}_0}] = [\nabla\mathbf{f}(\mathbf{x}_0)]$, the Jacobian, and we will do this by showing that they agree on all inputs:

$$[D\mathbf{f}_{\mathbf{x}_0}]\mathbf{u} = [\nabla\mathbf{f}(\mathbf{x}_0)]\mathbf{u} \quad \text{for all } \mathbf{u} \in \mathbb{R}^d.$$

Problem 1(e) [3 points] Prove that, for any two matrices $\mathbf{A} \in \mathbb{R}^{n\times d}$ and $\mathbf{B} \in \mathbb{R}^{n\times d}$, if $\mathbf{A}\mathbf{e}_i = \mathbf{B}\mathbf{e}_i$ for all standard basis vectors $\mathbf{e}_1, \ldots, \mathbf{e}_d$, then $\mathbf{A}\mathbf{u} = \mathbf{B}\mathbf{u}$ for any $\mathbf{u} \in \mathbb{R}^d$.

Problem 1(e) shows us that we only need to demonstrate that

$$[D\mathbf{f}_{\mathbf{x}_0}]\mathbf{e}_i = [\nabla\mathbf{f}(\mathbf{x}_0)]\mathbf{e}_i$$

for every standard basis vector $\mathbf{e}_i$ to prove Theorem 1. This will be our new goal.

Problem 1(f) [2 points] Let $\mathbf{d} = \alpha \mathbf{e}_i$ for a scalar $\alpha \in \mathbb{R}$ in Equation (2); this limit exists because we are assuming differentiability for Theorem 1. Prove that, in this case, Equation (2) simplifies to:

$$\lim_{\alpha \to 0} \frac{\mathbf{f}(\mathbf{x}_0 + \alpha \mathbf{e}_i) - \mathbf{f}(\mathbf{x}_0) - D\mathbf{f}_{\mathbf{x}_0}(\alpha \mathbf{e}_i)}{\alpha} = \mathbf{0}.$$

Note that $\alpha \mathbf{e}_i \to \mathbf{0}$ is equivalent to $\alpha \to 0$ once we fix $\mathbf{e}_i$.

Problem 1(f) reduces our problem from thinking of the limit from any arbitrary direction $\mathbf{d} \in \mathbb{R}^d$ to just the standard basis directions. It also shows us the way forward to prove Theorem 1.

Problem 1(g) [5 points] Use Problem 1(f) to conclude that

$$[D\mathbf{f}_{\mathbf{x}_0}]\mathbf{e}_i = [\nabla \mathbf{f}(\mathbf{x}_0)]\mathbf{e}_i,$$

proving Theorem 1. State why this also proves that all partial derivatives of $\mathbf{f}$ at $\mathbf{x}_0$ exist. *Hint:* Linearity might help.

By proving Theorem (1), you've unveiled why a solid understanding of linear algebra is so important for multivariable differential calculus: in a nutshell, multivariable differential calculus is the study of replacing nonlinear functions with linear transformations that approximate them well — these linear transformations are the derivatives.

One subtle point here is that Theorem 1 states that, when a function *is* differentiable (i.e. when the limit in Equation (2) exists for some $D\mathbf{f}_{\mathbf{x}_0}$), the Jacobian is definitely the approximating linear transformation:

$$\text{differentiable} \implies \text{Jacobian is the derivative.}$$

But this seems a bit strange — saying that a function is differentiable means that we are saying that the limit in Equation (2) exists in the first place, but in what cases might that limit exist?

$$? \implies \text{differentiable.}$$

We won't prove this fact, but, as we saw in class, one sufficient condition for a function to be differentiable is that the function is $\mathcal{C}^1$, or *continuously differentiable*. This means that all of its partial derivatives exist and are continuous.

$$\mathbf{f} \in \mathcal{C}^1 \implies \text{differentiable.}$$

One could check if a function is $\mathcal{C}^1$ by taking all its partial derivatives (perhaps in the usual mechanical way) and checking if they are continuous. [1]That is outside of the scope of this

[1]If you're curious, this footnote contains some examples of functions that behave quite pathologically.

course, because all of the functions we'll study and care about in this course are $\mathcal{C}^1$, and, typically, so are many of the non-pathological functions we typically care about in nature and engineering. Functions composed of polynomials, sin, cos, logarithms, and exponentials are all $\mathcal{C}^1$ in their domains. You can rest assured that, at least for us, any function we encounter can be reliably approximated locally by its derivative, the Jacobian.

Problem 1(h) [3 points] Prove that $\mathbf{f}$ in Equation (1) is in $\mathcal{C}^1$. You may use, without proof, the fact that any linear function is continuous everywhere (but to use this claim, you must prove that the function you apply it on is a linear function).

Example 1: $f(x,y) = \frac{x^2 y}{x^2+y^2}$ for $(x,y) \neq 0$ and $f(x,y) = 0$ for $(x,y) = 0$. This is a function that is continuous, has both partial derivatives existing, but is not differentiable. The problem is that its partial derivatives are discontinuous.
Example 2: $f(x,y) = \frac{3x^2 y - y^3}{x^2+y^2}$ when $(x,y) \neq (0,0)$ and $f(x,y) = 0$ for $(x,y) = (0,0)$. This function has all directional derivatives but it isn't differentiable at the origin.

Problem 2

Calculus with quadratic forms (25 points total).

In this problem, we will focus on multivariable calculus with quadratic functions and quadratic forms. Recall from lecture and PS2 that a quadratic form is the "quadratic part" of a quadratic function in multiple variables. A *quadratic form* is a polynomial function $f : \mathbb{R}^d \to \mathbb{R}$ with terms all of degree two. Some examples of quadratic forms $f : \mathbb{R}^2 \to \mathbb{R}$ include $f(x_1, x_2) = x_1^2 + x_2^2$ or $f(x_1, x_2) = 4x_1^2 + x_1x_2 - x_2^2$. Recall from PS2 that any quadratic form $f : \mathbb{R}^d \to \mathbb{R}$ can be written in terms of a symmetric matrix $\mathbf{A} \in \mathbb{R}^{d\times d}$ as:

$$f(\mathbf{x}) = \mathbf{x}^\top \mathbf{A}\mathbf{x}.$$

In lecture, we found that, for fixed $\mathbf{X} \in \mathbb{R}^{n\times d}$ and $\mathbf{y} \in \mathbb{R}^n$, the familiar least squares objective $f(\mathbf{w}) = \|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2$ can be expanded as:

$$f(\mathbf{w}) = \mathbf{w}^\top \mathbf{X}^\top \mathbf{X}\mathbf{w} - 2\mathbf{w}^\top(\mathbf{X}^\top \mathbf{y}) + \mathbf{y}^\top \mathbf{y}. \tag{3}$$

Notice that the quadratic term is indeed a quadratic form: $\mathbf{X}^\top\mathbf{X}$ is a $d \times d$ symmetric matrix, no matter what $\mathbf{X}$ is. In all, $f(\mathbf{w})$ is simply a quadratic function in multiple variables $\mathbf{w} = (w_1, \ldots, w_d)$. Quadratic functions arise everyhwere in optimization and machine learning, particularly because Taylor approximations allow us to study many functions with their first and second-order approximations. Therefore, doing calculus fluently with them is important.

First, however, a technical note about the gradient is in order. Recall that, for a scalar-valued function $f : \mathbb{R}^d \to \mathbb{R}$, the *gradient* at $\mathbf{x}$ is the vector $\nabla f(\mathbf{x}) = (x_1, \ldots, x_d) \in \mathbb{R}^d$ constructed by taking all the partial derivatives of f and stacking them into a vector:

$$\nabla f(\mathbf{x}) = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \vdots \\ \frac{\partial f}{\partial x_d} \end{bmatrix}.$$

Scalar-valued functions $f : \mathbb{R}^d \to \mathbb{R}$ are, of course, vector-valued functions with a single output, so all the definitions of derivative above in Problem 1 apply to them with $n = 1$. One subtle point to note, though, is that through our definitions above, technically the appropriate Jacobian (and, hence, derivative) of a scalar-valued function should be $[\nabla f(\mathbf{x})] \in \mathbb{R}^{1\times d}$, a matrix with a single row. This is just the gradient (which is a vector), but transposed: $[\nabla f(\mathbf{x})] = \nabla f(\mathbf{x})^\top$. Therefore, when we talk about the derivative of a scalar-valued function at $\mathbf{x}$, we are technically talking about the linear transformation $Df_\mathbf{x} : \mathbb{R}^d \to \mathbb{R}$ obtained by taking (the standard Euclidean) inner product with $\nabla f(\mathbf{x})$:

$$Df_\mathbf{x}(\mathbf{u}) = \nabla f(\mathbf{x})^\top \mathbf{u} \quad \text{for all } \mathbf{u} \in \mathbb{R}^d.$$

Technically, then, the derivative of a scalar-valued multivariable function is the the *linear functional* (defined in PS2) determined by the gradient vector $\nabla f(\mathbf{x})$. For all intents and purposes,

though, it is usually convention to just call the gradient the "derivative," because it is second-nature to take inner products with it.

Multivariable calculus generalizes all the familiar derivative rules we saw in single-variable calculus, like the sum rule, product rule, quotient rule, and chain rule in the ways you might expect. We won't prove all of these properties, but we'll prove two that are important for taking the gradient of a quadratic form.

The most basic property of derivatives from single-variable calculus is that constant functions have a derivative of 0 everywhere. That is, for $f(x) = C$ for some $C \in \mathbb{R}$,

$$f'(x) = 0.$$

The multivariable equivalent of this is that $\nabla f(\mathbf{x}) = \mathbf{0}$ if $f(\mathbf{x}) = C$ for some $C \in \mathbb{R}$. To prove that this is indeed true, we must appeal to the multivariable definition of the derivative (restated in Problem 1 in Equation (2)). In order to prove that these properties are true, we simply need to plug in a "guess" at what the derivative might be into Equation (2) and verify that the limit is indeed 0. These guesses are, of course, informed by the equivalent properties from single-variable calculus.

Problem 2(a) [3 points] Let $f : \mathbb{R}^d \to \mathbb{R}$ be a constant function $f(\mathbf{x}) = C$ for some $C \in \mathbb{R}$. A reasonable "guess" at the gradient of such a function is $\mathbf{0} \in \mathbb{R}^d$, the zero vector. Prove, using the definition of the total derivative in Equation (2) that, for any $\mathbf{x} \in \mathbb{R}^d$, the constant function $f(\mathbf{x}) = C$ is differentiable with gradient $\nabla f(\mathbf{x}) = \mathbf{0}$.

Second, we know from single-variable calculus that if $f : \mathbb{R} \to \mathbb{R}$ and $g : \mathbb{R} \to \mathbb{R}$ are differentiable at x with derivatives $f'(x)$ and $g'(x)$, then so is $f + g$, with derivative:

$$(f + g)'(x) = f'(x) + g'(x).$$

The derivative in multiple variables is no different.

Problem 2(b) [3 points] Let $f : \mathbb{R}^d \to \mathbb{R}$ and $g : \mathbb{R}^d \to \mathbb{R}$ be two scalar-valued multivariable functions. Suppose f and g are both differentiable at $\mathbf{a}$ with gradients $\nabla f(\mathbf{a})$ and $\nabla g(\mathbf{a})$. Prove that their sum, the function $(f + g) : \mathbb{R}^d \to \mathbb{R}$, defined by

$$(f + g)(\mathbf{x}) = f(\mathbf{x}) + g(\mathbf{x}) \quad \text{for all } \mathbf{x} \in \mathbb{R}^d$$

is differentiable at $\mathbf{a}$ with gradient $\nabla(f + g)(\mathbf{x}) = \nabla f(\mathbf{x}) + \nabla g(\mathbf{x})$.

Hint: Because f and g are both differentiable at $\mathbf{a}$, Equation (2) applies to both of them at $\mathbf{a}$. Use these to prove the claim.

Now that we have the familiar sum rule of multivariable calculus from Problem 2(b), we can rest assured that taking the derivative of Equation (3) can be done term-by-term. That is,

$$\nabla_{\mathbf{w}} f(\mathbf{w}) = \nabla_{\mathbf{w}} \mathbf{w}^\top \mathbf{X}^\top \mathbf{X}\mathbf{w} - \nabla_{\mathbf{w}} 2\mathbf{w}^\top (\mathbf{X}^\top \mathbf{y}) + \nabla_{\mathbf{w}} \mathbf{y}^\top \mathbf{y}, \tag{4}$$

where $\nabla_{\mathbf{w}}$ is notation that reminds us that the variable we're taking a derivative respect to is $\mathbf{w}$. The last term in Equation (4), as a function of $\mathbf{w}$, can be expressed as $f(\mathbf{w}) = \mathbf{y}^\top \mathbf{y}$, where $\mathbf{y}^\top \mathbf{y} \in \mathbb{R}$ is just some scalar. Therefore, Problem 2(a) tells us that $\nabla_{\mathbf{w}} \mathbf{y}^\top \mathbf{y} = 0$. We are left with:

$$\nabla_{\mathbf{w}} f(\mathbf{w}) = \nabla_{\mathbf{w}} \mathbf{w}^\top \mathbf{X}^\top \mathbf{X}\mathbf{w} - \nabla_{\mathbf{w}} 2\mathbf{w}^\top (\mathbf{X}^\top \mathbf{y}).$$

Now, let us prove a couple of useful vector calculus gradient identities that will allow us to take care of these last two terms. The first identity concerns the gradient of linear functionals, which are just linear functions of the form $f : \mathbb{R}^d \to \mathbb{R}$.

Problem 2(c) [3 points] Let $\mathbf{a} \in \mathbb{R}^d$ be any fixed vector. Consider the function $f : \mathbb{R}^d \to \mathbb{R}$ defined by

$$f(\mathbf{x}) = \mathbf{a}^\top \mathbf{x}.$$

Prove that $\nabla f(\mathbf{x}) = \mathbf{a}$ at any $\mathbf{x} \in \mathbb{R}^d$. State why, if we expressed $f(\mathbf{x}) = \mathbf{x}^\top \mathbf{a}$, we would still have $\nabla f(\mathbf{x}) = \mathbf{a}$.

Hint: To prove this, use the definition of the standard Euclidean inner product (dot product) and take partial derivatives.

Problem 2(c) should be seen as analogous to the single-variable differentiation rule that linear functions through the origin have derivative equal to their slope:

$$\frac{d}{dx} ax = a.$$

Problem 2(d) [3 points] Let $\mathbf{A} \in \mathbb{R}^{d \times d}$ be any fixed square matrix (*not* necessarily symmetric). Consider the function $f : \mathbb{R}^d \to \mathbb{R}$ defined by

$$f(\mathbf{x}) = \mathbf{x}^\top \mathbf{A}\mathbf{x}.$$

Prove that $\nabla f(\mathbf{x}) = (\mathbf{A} + \mathbf{A}^\top)\mathbf{x}$ at any $\mathbf{x} \in \mathbb{R}^d$. State why $\nabla f(\mathbf{x}) = 2\mathbf{A}\mathbf{x}$ if $\mathbf{A}$ is symmetric. You may use any definition of matrix-vector product to prove this claim.

For symmetric matrices, Problem 2(d) gives us the derivative of quadratic forms. It should be seen as analogous to the single-variable differentiation rule that quadratic functions with no linear or constant terms obey:

$$\frac{d}{dx} ax^2 = 2ax.$$

Together, Problems 2(a)-(d) give you all the ingredients to derive the gradient of the least squares objective. The least squares objective is an instance of a general quadratic function in multiple variables. In the last half of this problem, we will derive a general formula for minimizing such functions.

Let $\mathbf{A} \in \mathbb{R}^{d \times d}$ be a symmetric matrix, let $\mathbf{b} \in \mathbb{R}^d$ be a vector, and let $c \in \mathbb{R}$ be a scalar. Then, any multiavariable quadratic function $f : \mathbb{R}^d \to \mathbb{R}$ can be written as:

$$f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^\top \mathbf{A}\mathbf{x} - \mathbf{b}^\top \mathbf{x} + c. \tag{5}$$

The 1/2 in front of the quadratic form term is sometimes written for mathematical convenience with the gradients. You will prove the following general claim:

Theorem 2 (Optimizing positive definite quadratic functions). Consider any multivariable quadratic function $f : \mathbb{R}^d \to \mathbb{R}$, written as

$$f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^\top \mathbf{A}\mathbf{x} - \mathbf{b}^\top \mathbf{x} + c.$$

If $\mathbf{A} \in \mathbb{R}^{d \times d}$ is positive definite, then the global minimizer of $f(\mathbf{x})$ exists and is $\mathbf{x}^* = \mathbf{A}^{-1}\mathbf{b}$, with the minimum value

$$f(\mathbf{x}^*) = -\frac{1}{2}\mathbf{b}^\top \mathbf{A}^{-1}\mathbf{b} + c \leq f(\mathbf{x}) \quad \text{for all } \mathbf{x} \in \mathbb{R}^d.$$

In the following problems, take the following theorem to be true (we will prove this later in the course).

Theorem 3 (First-order necessary condition for optima). Let $f : \mathbb{R}^d \to \mathbb{R}$ be a differentiable function at all $\mathbf{x} \in \mathbb{R}^d$. Let $\mathbf{x}^* \in \mathbb{R}^d$ be a *local minimum:* there exists some δ such that $f(\mathbf{x}^*) \leq f(\mathbf{x})$ for all $\mathbf{x} \in B_\delta(\mathbf{x}^*)$, where $B_\delta(\mathbf{x}^*)$ is the δ-neighborhood (defined in class) around $\mathbf{x}^*$. Then, $\nabla f(\mathbf{x}^*) = \mathbf{0}$.

Problem 2(e) [4 points] Prove that the gradient at any $\mathbf{x} \in \mathbb{R}^d$ of any quadratic function, written as Equation (5), is:

$$\nabla f(\mathbf{x}) = \mathbf{A}\mathbf{x} - \mathbf{b}.$$

Suppose that $\mathbf{A}$ is positive definite. State why $\mathbf{A}$ must be invertible, and prove that the only candidate global minimum is $\mathbf{x}^* = \mathbf{A}^{-1}\mathbf{b}$. State why this is the unique candidate choice for a global minimum. You may use Theorem 3.

Problem 2(e) gives us a candidate for a minimum, but notice that Theorem 3 only tells us that it is locally a minimum. That is, we are only guaranteed that $\mathbf{x}^*$ is a minimum for points in some neighborhood around $\mathbf{x}^*$. To show that it is a *global* minimum, we usually need some other condition on the function. One such condition that is useful for our purposes is coerciveness. A function $f : \mathbb{R}^d \to \mathbb{R}$ is said to be *coercive* if

$$f(\mathbf{x}) \to \infty \text{ as } \|\mathbf{x}\| \to \infty.$$

As a small example, $f(x_1, x_2) = x_1^2 + x_2^2$ is coercive, because $f(x_1, x_2) = \left(\sqrt{x_1^2 + x_2^2}\right)^2 = \|\mathbf{x}\|^2$, and clearly $f \to \infty$ as $\|\mathbf{x}\|^2 \to \infty$. Coercive functions are functions that "coerce" their output values to increase as their inputs increase in magnitude. Intuitively, one class of coercive functions are the "bowl-shaped" ones, as they increase to infinity in all directions as we increase the size of the inputs. We still don't have a formal definition for "bowl-shaped" yet, but an important sub-class of bowl-shaped functions are the positive definite quadratic forms. Problem 2(f) shows that positive definite quadratic forms are coercive.

Problem 2(f) [3 points] Prove that for a positive definite matrix $\mathbf{A} \in \mathbb{R}^{d \times d}$, any function of the form

$$f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^\top \mathbf{A}\mathbf{x} - \mathbf{b}^\top \mathbf{x} + c$$

is coercive. This will require a couple of ingredients you have already proven or seen, namely: (i) The spectral theorem (ii) Cauchy-Schwarz and (iii) the property that any orthogonal matrix $\mathbf{V} \in \mathbb{R}^{d \times d}$ preserves lengths under projections: $\|\mathbf{V}^\top \mathbf{x}\| = \|\mathbf{x}\|$. If you haven't seen this last property before, it has an easy proof you can try for yourself (try squaring both sides).

It turns out that a very useful property of coercive functions is that coercive functions *always* have at least one global minimizer.

Theorem 4 (Coercive functions and global minima). If a function $f : \mathbb{R}^d \to \mathbb{R}$ is coercive then it has at least one global minimizer, i.e. there exists $\mathbf{x}^* \in \mathbb{R}^d$ such that

$$f(\mathbf{x}^*) \leq f(\mathbf{x}) \quad \text{for all } \mathbf{x} \in \mathbb{R}^d.$$

We won't prove this, but this seems intuitively true — if a function only grows as we let its inputs grow in magnitude, it better have an input where its function value is smallest somewhere.

Problem 2(g) [3 points] Conclude using the last couple of problems and Theorem 4 that Theorem 2 is true.

Theorem 2 is very handy in machine learning; keep this tucked away in your back pocket. For instance, because we proved this general theorem, we can go back around and try to apply it immediately to the least squares objective function.

Problem 2(h) [3 points] Verify that the least squares objective $f(\mathbf{w})$ in Equation (3) can be written as in Equation (5). What is $\mathbf{A}$, $\mathbf{b}$, and c? State what must be true about the matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$ in the least squares objective for Theorem 2 to apply and state the global minimizer $\mathbf{w}^* \in \mathbb{R}^d$ and global minimum value $f(\mathbf{w}^*) \in \mathbb{R}$ when it does apply.

Problem 3

The gradient, Taylor's Theorem, and gradient descent [25 points total].

In this problem, we will try to get a bit more intuition for *gradient descent*, the main workhorse algorithm for much of modern machine learning. In the grand scheme of things, gradient descent is also the other "narrative" that our course is based around (next to least squares regression), and it unifies many of the concepts we'll be learning moving forward. Recall that the algorithm for gradient descent can simply be written as follows:

Algorithm 1 Gradient Descent

Input: Function $f : \mathbb{R}^d \to \mathbb{R}$. Initial point $\mathbf{x}_0 \in \mathbb{R}^d$. Positive step size $\eta > 0$. Positive stopping condition $\epsilon > 0$.

1: **for** $t = 1, 2, 3, \ldots$ **do**
2: Compute $\mathbf{x}_t \leftarrow \mathbf{x}_{t-1} - \eta \nabla f(\mathbf{x}_{t-1})$.
3: **if** $\|\nabla f(\mathbf{x}_t)\| < \epsilon$ **then**
4: **return** $f(\mathbf{x}_t)$.
5: **end if**
6: **end for**

In particular, we saw in lecture that the idea of a first-order Taylor approximation was crucial to our understanding of gradient descent. At any time $t \geq 0$ in gradient descent, we are at the point $\mathbf{x}_t \in \mathbb{R}^d$. The idea is that, to get to the next point, we want to use local first-order information to move in a direction $\mathbf{d} \in \mathbb{R}^d$ that gets us to the next point $\mathbf{x}_{t+1} = \mathbf{x}_t + \mathbf{d}$. Intuitively, the *first-order Taylor approximation* told us that, as long as η is small enough, then $\mathbf{x}_t + \mathbf{d}$ is close to $\mathbf{x}_t$, and

$$f(\mathbf{x}_{t+1}) = f(\mathbf{x}_t + \mathbf{d}) \approx f(\mathbf{x}_t) + \nabla f(\mathbf{x}_t)^\top \mathbf{d}.$$

What should $\mathbf{d} \in \mathbb{R}^d$ be then? Our idea was to use a greedy step — we should move in the *local* direction that makes the function smallest the fastest. That is, we want to find the direction of *steepest descent*, which is often used as a one-line summary of Algorithm 1:

Move in the direction of steepest descent.

It turns out that this direction is the negative of the function's *gradient* at $\mathbf{x}_t$, which is given by the vector $\nabla f(\mathbf{x}_t) \in \mathbb{R}^d$. This is the first mystery in gradient descent that isn't immediately obvious from the definition of the gradient: *why is the negative gradient the direction of steepest descent?* To descend, we should go in the negative direction, so the equivalent question is: *why is the gradient the direction of steepest ascent?*

One useful property we saw in class was that, for differentiable functions, the *directional derivative* can be obtained by just taking a dot product with the gradient (more generally, they can

be found by matrix-vector product with the Jacobian). Recall from lecture and Problem 1 that the directional derivative of f at $\mathbf{x}_0$ in the direction $\mathbf{v} \in \mathbb{R}^d$ is

$$\lim_{\delta \to 0} \frac{f(\mathbf{x}_0 + \delta \mathbf{v}) - f(\mathbf{x}_0)}{\delta}.$$

We will prove the following claim:

Theorem 5 (Directional derivatives from gradient). Let $\mathbf{f} : \mathbb{R}^d \to \mathbb{R}^n$ be differentiable at $\mathbf{x}_0 \in \mathbb{R}^d$. Then, all directional derivatives of $\mathbf{f}$ at $\mathbf{x}_0$ exist and the directional derivative at $\mathbf{x}_0$ in the direction $\mathbf{v}$ is given by the formula

$$\lim_{\delta \to 0} \frac{f(\mathbf{x}_0 + \delta \mathbf{v}) - f(\mathbf{x}_0)}{\delta} = \nabla f(\mathbf{x}_0)^\top \mathbf{v}.$$

In the definition of the derivative in Equation (2), define the remainder as the numerator term:

$$R(\mathbf{d}) = (f(\mathbf{x}_0 + \mathbf{d}) - f(\mathbf{x}_0)) - \nabla f(\mathbf{x}_0)^\top \mathbf{d}. \tag{6}$$

Problem 3(a) [3 points] State why, if $f : \mathbb{R}^d \to \mathbb{R}$ is differentiable at $\mathbf{x}_0$, its remainder $R(\mathbf{d})$ in Equation (6) satisfies

$$\lim_{\mathbf{d} \to \mathbf{0}} \frac{R(\mathbf{d})}{\|\mathbf{d}\|} = 0.$$

Take $\mathbf{d} = \delta \mathbf{v}$ to be the direction given in the directional derivative definition, where $\delta \in \mathbb{R}$ and $\mathbf{v} \in \mathbb{R}^d$. Prove that

$$\|\mathbf{v}\| \frac{R(\delta \mathbf{v})}{\delta \|\mathbf{v}\|} = \frac{f(\mathbf{x}_0 + \delta \mathbf{v}) - f(\mathbf{x}_0)}{\delta} - \nabla f(\mathbf{x}_0)^\top \mathbf{v}.$$

Linearity might be helpful. Use the property that

$$\lim_{\delta \to 0} \frac{R(\delta \mathbf{v})}{\delta \|\mathbf{v}\|} = 0 \quad \text{if} \quad \lim_{\mathbf{d} \to \mathbf{0}} \frac{R(\mathbf{d})}{\|\mathbf{d}\|} = 0$$

to conclude that Theorem 5 is true.

Problem 3(a) shows us that we can get *any* of a function's directional derivatives by taking a standard Euclidean inner product (dot product) with the gradient. Now, we want to formalize and prove the often cited property that

The gradient is the direction of steepest ascent.

Now, recall from PS1 that the dot product is intimately connected with the angle between two vectors and the idea of projection. Formally, for any two vectors $\mathbf{u}, \mathbf{v} \in \mathbb{R}^d$, the dot product can be written as

$$\mathbf{u}^\top \mathbf{v} = \|\mathbf{u}\| \|\mathbf{v}\| \cos\theta, \tag{7}$$

where θ is the angle formed between $\mathbf{u}$ and $\mathbf{v}$, drawn as vectors. Because we only care about the *direction* of steepest ascent, we will restrict ourselves to the unit vectors $\mathcal{S} := \{\mathbf{v} \in \mathbb{R}^d : \|\mathbf{v}\| = 1\}$. Recall from class that we can express optimization problems with an *objective function* $f : \mathbb{R}^d \to \mathbb{R}$ and a *constraint set*, $\mathcal{C} \subseteq \mathbb{R}^d$, as follows:

$$\begin{aligned} \underset{\mathbf{x} \in \mathbb{R}^d}{\text{maximize}} \quad & f(\mathbf{x}) \\ \text{subject to} \quad & \mathbf{x} \in \mathcal{C}. \end{aligned}$$

Problem 3(b) [4 points] Let $f : \mathbb{R}^d \to \mathbb{R}$ be a function with the gradient $\nabla f(\mathbf{x}_0) \in \mathbb{R}^d$ at a point $\mathbf{x}_0$. Consider the following optimization problem:

$$\begin{aligned} \underset{\mathbf{v} \in \mathbb{R}^d}{\text{maximize}} \quad & \nabla f(\mathbf{x}_0)^\top \mathbf{v} \\ \text{subject to} \quad & \|\mathbf{v}\| = 1 \end{aligned}$$

Prove that this is equivalent to the optimization problem:

$$\begin{aligned} \underset{\mathbf{v} \in \mathbb{R}^d}{\text{maximize}} \quad & \|\nabla f(\mathbf{x}_0)\| \cos \theta_{\mathbf{v}} \\ \text{subject to} \quad & \|\mathbf{v}\| = 1, \end{aligned}$$

where $\theta_{\mathbf{v}}$ is the angle between $\mathbf{v}$ and $\nabla f(\mathbf{x}_0)$. We have written $\theta_{\mathbf{v}}$ because the angle clearly still depends on $\mathbf{v}$. At what setting of $\theta_{\mathbf{v}}$ do we achieve a maximum? Why does that setting of $\theta_{\mathbf{v}}$ mean that the maximizer, $\mathbf{v}^*$, satisfies

$$\mathbf{v}^* = \frac{\nabla f(\mathbf{x}_0)}{\|\nabla f(\mathbf{x}_0)\|}?$$

You may answer this with geometric intuition.

Recall from Theorem 5 that the directional derivative, which is just the rate of change of f in the direction of $\mathbf{v}$, is obtained exactly by that optimization problem in Problem 3(b). Therefore, by finding $\mathbf{v}$ that maximizes the above problem, you found the direction in which the function changes the *fastest*. In short, you found the direction of steepest ascent, which was just the normalized gradient.

Problem 3(c) [3 points] Using Problem 3(b), prove that the actual value of the directional derivative at $\mathbf{x}_0$ is $\|\nabla f(\mathbf{x}_0)\|$.

Problem 3(c) tells us that, moving in the direction of steepest ascent, the rate of change of the function $f : \mathbb{R}^d \to \mathbb{R}$ is the norm of the gradient itself! Therefore, the gradient not only tells us the direction of steepest ascent, but also how fast we ascend.

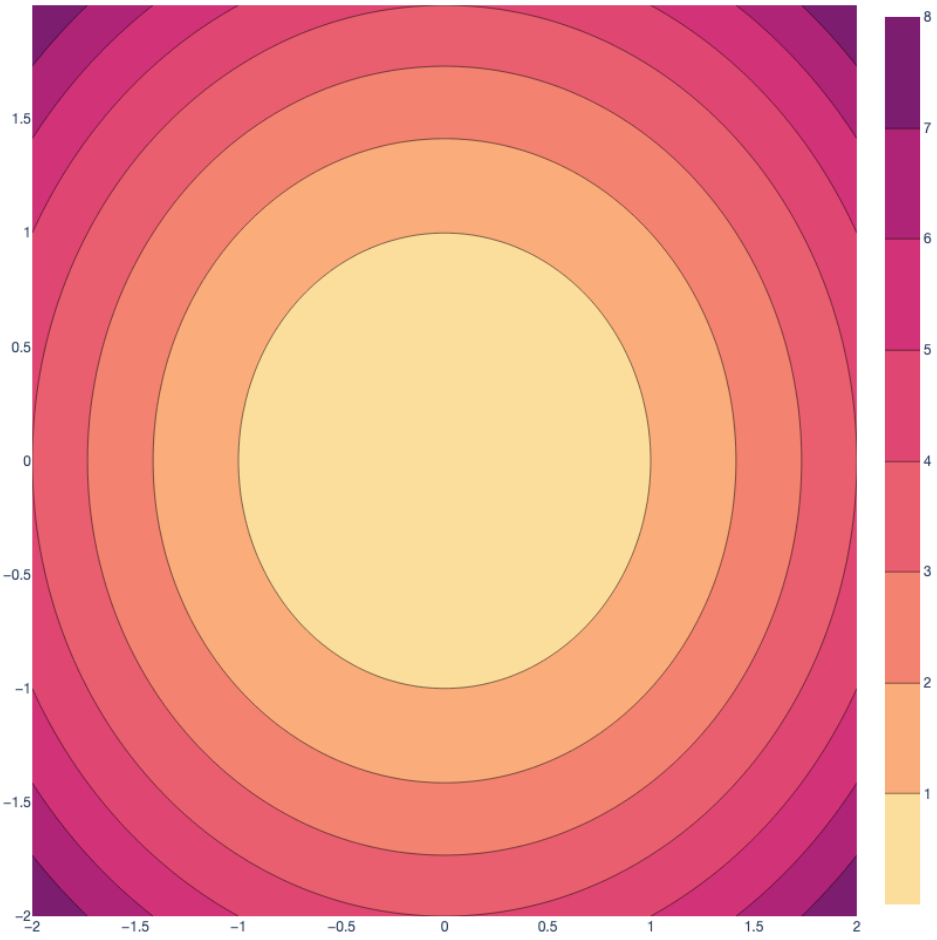

Figure 1: Contour plot of $f(x, y) = x^2 + y^2$ in Problem 3(d).

Problem 3(d) [3 points] Consider the function $f : \mathbb{R}^2 \to \mathbb{R}$ given by

$$f(x, y) = x^2 + y^2.$$

our classic "bowl-shaped" positive definite quadratic function. Compute the gradient $\nabla f(\mathbf{x}_0)$ at the point $\mathbf{x}_0 = (1, 1)$. State the unit direction of steepest ascent at $\mathbf{x}_0$ and what the rate of change is by taking that direction. You don't need to turn this in, but draw on some scratch paper the contour plot in Figure 1 and the direction of steepest ascent at $(1, 1)$.

It may help your intuition to draw some other directions at $\mathbf{x}_0$ in which our function may change.

Problem 3(e) [3 points] Consider the function $f : \mathbb{R}^2 \to \mathbb{R}$ given by

$$f(x, y) = x^2 + y^2.$$

our classic "bowl-shaped" positive definite quadratic function. Compute the directional derivatives at $\mathbf{x}_0 = (1, 1)$ in the directions

$$\mathbf{v}_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \quad \mathbf{v}_2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix} \quad \mathbf{v}_3 = \begin{bmatrix} -1/\sqrt{2} \\ 1/\sqrt{2} \end{bmatrix}$$

using Theorem 5. Also draw these on your contour plot, continuing from Problem 3(d). You don't have to turn the drawings in.

To relate this back to PS1 and the intuition from dot product you gained from there, another way of viewing a dot product is a projection of a vector onto another. Maximizing the dot product between two vectors is the same as minimizing the length of the *projection* of one vector onto the other. From that perspective, finding the maximizing $\mathbf{v} \in \mathbb{R}^d$ is the same as finding the vector that, when projected onto the span of $\nabla f(\mathbf{x}_0)$, maximizes the length of the projection: $\|\Pi_{\nabla f(\mathbf{x}_0)}(\mathbf{v})\|$. Of course, choosing $\mathbf{v}$ to be exactly *on* the span of $\nabla f(\mathbf{x}_0)$ (i.e., parallel) maximizes that projection.

Finally, let's consider performing gradient descent on this simple function. Typically, tuning the learning rate $\eta \in \mathbb{R}$ for gradient descent (see Algorithm 1) is a dark art, but, in certain cases, we have some principled ways of choosing η. One particular case is when we have a *β-smooth function*.

Problem 3(f) [3 points] Consider the function $f : \mathbb{R}^2 \to \mathbb{R}$ given by

$$f(x, y) = x^2 + y^2.$$

Compute and state its Hessian. Prove that this function is indeed β-smooth, and state the appropriate value of β.

Let's consider some approximations of this function and how accurate they are. To do this, we should employ *Taylor's Theorem*, which gives us an exact expression for our approximation error (remainder).

Problem 3(g) [3 points] Write the first-order Taylor approximation of

$$f(x, y) = x^2 + y^2$$

at the point $\mathbf{x}_0 = (1, 1)$. Prove that using (Lagrange's form of) Taylor's Theorem to capture the first-order approximation at any point $f(\mathbf{x})$ plus the error term simply gives us back the function itself. Why is this the case?

Finally, we will specify the gradient descent algorithm for our test function. In this case, we know the β-smoothness of the function, so this allows us to write the gradient descent algorithm exactly, with a principled choice of η.

Problem 3(h) [3 points] Using the choice $\eta = 1/\beta$ suggested by the theorem we proved in class about gradient descent making function values smaller, write out the gradient descent update step (Step 2 in Algorithm 1) for this function, using its specific gradient $\nabla f(\mathbf{x}_t)$. Then, write out the first iteration $\mathbf{x}_1$ of gradient descent when starting with $\mathbf{x}_0 = (1, 1)$. Can you apply Theorem 2 in Problem 2 to this function? If so, state what $\mathbf{A}$, $\mathbf{b}$ and c are, as well as the global minimizer $\mathbf{x}^*$ and the minimum value $f(\mathbf{x}^*)$.

Programming Part

Implementing gradient descent (25 points total). In this problem, you will implement gradient descent from scratch and experiment a bit with it on some simple functions.

In order to start this programming part, download the file `ps3.ipynb` from Course Content on the course webpage. Your submission for this part will be the same `ps3.ipynb` file modified with your code; see HW Submission on the course webpage for additional instructions.