Name: Wang Andie

CS 189/289A Introduction to Machine Learning
Fall 2024 Jennifer Listgarten, Saeed Saremi **DIS0**

1 Vector Calculus

Below, $\mathbf{x} \in \mathbb{R}^d$ means that $\mathbf{x}$ is a $d \times 1$ column vector with real-valued entries. Likewise, $\mathbf{A} \in \mathbb{R}^{d \times d}$ means that $\mathbf{A}$ is a $d \times d$ matrix with real-valued entries. In this course, we will by convention consider vectors to be column vectors.

Consider $\mathbf{x}, \mathbf{w} \in \mathbb{R}^d$ and $\mathbf{A} \in \mathbb{R}^{d \times d}$. In the following questions, $\nabla_{\mathbf{x}}$ denotes the gradient with respect to $\mathbf{x}$, which, by convention, is a column vector. See the appendix for more details on definitions for vector calculus.

Calculate the following derivatives.

(a) $\nabla_{\mathbf{x}}(\mathbf{w}^\top \mathbf{x})$

$\nabla_x(w^T x) = \left(\frac{\partial (w^T x)}{\partial x}\right)^T = w$

(b) $\nabla_{\mathbf{x}}(\mathbf{w}^\top \mathbf{A}\mathbf{x})$

$\nabla_x(w^T A x) = A^T w$

(c) $\nabla_{\mathbf{A}}(\mathbf{w}^\top \mathbf{A}\mathbf{x})$

$w^T A x = \sum_{i=1}^{n} \sum_{j=1}^{m} w_i A_{ij} \cdot x_j$

$\nabla_A(w^T A x) = w x^T$

(d) $\nabla_{\mathbf{x}}(\mathbf{x}^\top \mathbf{A}\mathbf{x})$

$\nabla_x(x^TAx) = (A+A^T)x$

(e) $\nabla^2_{\mathbf{x}}(\mathbf{x}^\top \mathbf{A}\mathbf{x})$

$\nabla_x(x^TAx) = (A+A^T)x$

$\nabla^2_x(x^TAx) = \left(\frac{\partial(A+A^T)x}{\partial x}\right)^T$

$= A+A^T$

(f) Now let's apply our identities derived above to a practical problem. Given a design matrix $X \in \mathbb{R}^{nxd}$ and a label vector $Y \in \mathbb{R}^n$, the ordinary least squares regression problem is

$$w^* = \underset{w}{\operatorname{argmin}} \frac{1}{2}||Xw - Y||_2^2$$

Using parts (a)–(e), derive a necessary condition for w^*. *Note: We do not necessarily assume X is full rank! Hint: A necessary condition for a minimum point of a function is that it is a critical point, i.e. where the gradient is 0.*

$g(w) = \nabla_w\left(\frac{1}{2}||Xw-Y||_2^2\right) = \left(\frac{\partial\left[\frac{1}{2}||Xw-Y||_2^2\right]}{\partial w}\right)^T$

$= \left(\frac{\partial\left(\frac{1}{2}||Xw-Y||_2^2\right)}{\partial(Xw-Y)} \cdot \frac{\partial(Xw-Y)}{\partial w}\right)^T$

$= \left((Xw-Y)^T \cdot X\right)^T$

$= X^T(Xw-Y)$

$g(w^*) = X^T(Xw^*-Y) = 0 \qquad w^* = (X^TX)^{-1}X^TY$ if X^TX is invertible

2 Back to Basics: Linear Algebra

Let $X \in \mathbb{R}^{n \times m}$. When we write $\subseteq$, it means "is a subspace of." We study a few important subspaces in the theory of linear maps:

- The **columnspace**, also called the range or span, of X is $Range(X) := \{Xv : v \in \mathbb{R}^m\} \subseteq \mathbb{R}^n$. Consists of all vectors in the span (the set of all linear combinations) of the columns of X.
- The **rowspace** is $Row(X) := \{X^\top v : v \in \mathbb{R}^n\} \subseteq \mathbb{R}^m$. Consists of all vectors in the span of the rows of X.
- The **nullspace**, also called the kernel, of X is $\mathcal{N}(x) := \{v \in \mathbb{R}^m : Xv = 0\} \subseteq \mathbb{R}^m$.
- The **orthogonal complement** of a subspace U in some vector space V is a subspace, denoted $U^\perp$, such that $u \in U, v \in U^\perp \implies u \cdot v = 0$ and U and $U^\perp$ together span V. (These facts imply that $\dim U + \dim U^\perp = \dim V$. It also implies that $U^{\perp\perp} = U$.) For example, in the three-dimensional Euclidean space $V = \mathbb{R}^3$, if U is a plane through the origin, then $U^\perp$ is a line through the origin perpendicular to U.

For this problem we do not assume that X has full rank.

(a) Show that the following facts are true.

(i) $Row(X) = Range(X^\top)$

$Range(X^T) := \{X^T v : v \in R^n\} \subseteq R^m$

$Row(X) = Range(X^T)$

(ii) $\mathcal{N}(X)^\perp = Row(X)$.

$\forall v_1 \in Row(X), v_2 \in N(X), v_1 \cdot v_2 = v_1^T v_2 = (Xv_0)^T v_2 = v_0^T (X v_2) = 0 \Rightarrow v_1 \in N(X)^\perp \implies Row(X) \subseteq N(X)^\perp$

According to the defination of Null Space, $Row(X)^\perp = N(X)$

$\forall u \in R^m, u = u_x + u_y, u_x \in Row(X), u_y \in Row(X)^\perp$

$\forall v \in N(X)^\perp, v = v_x + v_y, v_x \in Row(X), v_y \in Row(X)^\perp$

$v \in N(X)^\perp, v_y \in N(X) \Rightarrow v \cdot v_y = 0 \Rightarrow (v_x + v_y) \cdot v_y = 0 \Rightarrow \|v_y\|_2^2 = 0 \Rightarrow v_y = 0$

$v_x \in Row(X), v_y \in Row(X)^\perp \Rightarrow v_x \cdot v_y = 0$

$v = v_x + v_y = v_x + 0 = v_x \in Row(X) \Rightarrow N(X)^\perp \subseteq Row(X)$ Therefore, $N(X)^\perp = Row(X)$

(iii) $\mathcal{N}(X^\top X) = \mathcal{N}(X)$. *Hint: if $v \in \mathcal{N}(X^\top X)$, then $v^\top X^\top X v = 0$.*

$\forall v_1 \in N(X), Xv_1 = 0, X^T X v_1 = 0, v_1 \in N(X^T X)$

$\forall v_2 \in N(X^T X), X^T X v_2 = 0, v_2^T X^T X v_2 = 0, \|Xv_2\|_2^2 = 0, Xv_2 = 0$

$v_2 \in N(X)$

$N(X^T X) = N(X)$

(b) We now prove an important result of linear algebra, the rank-nullity theorem. Let $Rank(X) = \dim Range(X) = \dim Row(X)$ and $Nullity(X) = \dim \mathcal{N}(X)$. (The fact that $\dim Range(X) = \dim Row(X)$—that is, the dimension spanned by the rows equals the dimension spanned by the columns—is itself a pretty important result, which you should always remember when you hear the word "rank.") The rank-nullity theorem says that for any $X \in \mathbb{R}^{n \times m}$,

$$Rank(X) + Nullity(X) = m.$$

Use the above results to prove this theorem. *Hint: Use the orthogonal complement of the nullspace to connect the rank to the nullity.*

$\dim N(X) + \dim N(X)^{\perp} = m$

$Rank(X) + Nullity(X) = m$

Gilbert Strang has proposed that a collection of four facts be called "fundamental theorem of linear algebra." Two of these facts are the rank-nullity theorem, part (b), and the fact that the rowspace is the orthogonal complement of the nullspace, part (a)(ii). The other two facts are related to the singular value decomposition, which we'll learn late in the semester.

3 Eigenvalues

(a) Let $\mathbf{A}$ be an invertible matrix. Show that if $\mathbf{v}$ is an eigenvector of $\mathbf{A}$ with eigenvalue λ, then it is also an eigenvector of $\mathbf{A}^{-1}$ with eigenvalue λ^{-1}.

$Av = \lambda v \Rightarrow A^{-1}Av = A^{-1}\lambda v \Rightarrow v = \lambda A^{-1}v \Rightarrow \lambda^{-1}v = A^{-1}v$

(b) A symmetric matrix $\mathbf{A}$ is said to be positive semidefinite (PSD) ($\mathbf{A} \succeq 0$) if $\forall \mathbf{v} \neq 0, \mathbf{v}^\top \mathbf{A}\mathbf{v} \geq 0$. Show that $\mathbf{A}$ is PSD if and only if all of its eigenvalues are nonnegative.

Hint: Use the eigendecomposition of the matrix $\mathbf{A}$.

$A = Q\Lambda Q^T$

$v^T A v = v^T Q \Lambda Q^T v$

$z = Q^T v$

$v^T A v = z^T \Lambda z \geq 0$

$\therefore z^T \Lambda z = \sum \lambda_i z_i^2 \geq 0$

All $\lambda_i \geq 0$

(c) Let $\mathbf{A}$ be a PSD matrix. Show that its eigenvalues are equal to its singular values.

$A = Q\Lambda Q^{-1}$

$A^2 = Q\Lambda^2 Q^{-1}$

$\therefore \lambda_A \geq 0$

$\therefore \lambda_A = \sqrt{\lambda_{A^2}}$

4 Probability Review

There are n archers all shooting at the same target (bulls-eye) of radius 1. Let the score for a particular archer be defined to be the distance away from the center (the lower the score, the better, and 0 is the optimal score). Each archer's score is independent of the others, and is distributed uniformly between 0 and 1. What is the expected value of the worst (highest) score?

(a) Define a random variable Z equal to the worst (highest) score, in terms of random variables that indicate each archer's score.

$Z = \max\{X_i \mid i \in [0, n], i \in \mathbb{Z}\}$

(b) Derive the Cumulative Distribution Function (CDF) of Z. *Hint: Recall the CDF of a random variable Z is given by* $F(z) = P(Z \leq z)$

$F(z) = P(Z \leq z) = z^n$

$\frac{dF(z)}{dz} = nz^{n-1}$

(c) Let X be a non-negative random variable. The Tail-Sum formula states that

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

Using both the Tail-Sum formula and the CDF of Z you derived, calculate the expected value of Z. *Hint: Write* $\mathbb{P}(X \geq t)$ *in terms of the CDF of X.*

$E[X] = \int_0^1 P(X \geq t)\, dt = \int_0^1 (1 - t^n)\, dt$

$= \left(t - \frac{1}{n+1} t^{n+1}\right)\Big|_0^1 = \frac{n}{n+1}$

(d) Consider what happens to $\mathbb{E}[Z]$ as $n \to \infty$. Does this match your intuition?

Yes, $n \to \infty$, the max will more be more close to 1.

5 Vector Calculus Appendix [1]

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 can help find the change in function value at $\mathbf{x}$, up to first order, when we add a little perturbation $\Delta \in \mathbb{R}^d$ to $\mathbf{x}$. That is,

$$f(\mathbf{x} + \Delta) = f(\mathbf{x}) + \frac{\partial f}{\partial \mathbf{x}} \Delta + o(\|\Delta\|) \tag{1}$$

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.[2]

We call $\frac{\partial f}{\partial \mathbf{x}}$ the *derivative of f at $\mathbf{x}$*. Sometimes we use $\frac{df}{d\mathbf{x}}$ but 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 $\mathbf{x}$ only.)

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. So one way to compute the derivative is to expand out $f(\mathbf{x} + \Delta)$ and guess the expression for the derivative. We call this method *computation via first principle*.

The gradient of f at $\mathbf{x}$ is defined to be the transpose of this derivative. That is $\nabla f(\mathbf{x}) = \left(\frac{\partial f}{\partial \mathbf{x}}\right)^\top$.

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 the vector input to f. We have

$$\frac{\partial f}{\partial \mathbf{x}} \in \mathbb{R}^{1 \times d} \quad \text{such that} \quad \frac{\partial f}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, \ldots, \frac{\partial f}{\partial x_d} \end{bmatrix} \tag{2}$$

$$\nabla_{\mathbf{x}} f = \left(\frac{\partial f}{\partial \mathbf{x}}\right)^\top = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \\ \vdots \\ \frac{\partial f}{\partial x_d} \end{bmatrix}. \tag{3}$$

[1] 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`.

[2] 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. Let $\mathbf{B}$ denote the matrix input to f. Then, the derivative of f with respect to $\mathbf{B}$ is an $n \times m$ matrix given by

$$\frac{\partial y}{\partial \mathbf{B}} \in \mathbb{R}^{n\times m} \quad \text{such that} \quad \left[\frac{\partial y}{\partial \mathbf{B}}\right]_{ij} = \frac{\partial y}{\partial B_{ji}}, \tag{4}$$

As in the vector case above, the gradient and derivative with respect to a matrix are also transposes of each other. So, the gradient of f with respect to $\mathbf{B}$ is an $m \times n$ matrix given by

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

An argument via first principle follows as:

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

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 that 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

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

That is,

$$[J(\mathbf{z})]_{ij} = \left[\frac{\partial \mathbf{z}}{\partial \mathbf{x}}\right]_{ij} = \frac{\partial z_i}{\partial x_j}. \tag{9}$$

4. The Hessian of $f : \mathbb{R}^d \to \mathbb{R}$ is the generalization of a second derivative and is defined as

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

A first principle definition is

$$\nabla f(\mathbf{x} + \Delta) \approx \nabla f(\mathbf{x}) + \nabla^2 f(\mathbf{x})\Delta \tag{11}$$

or equivalently

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

For sufficiently smooth functions (when the mixed derivatives are equal), the Hessian is a symmetric matrix. Most of the functions we cover in this class will have symmetric Hessians.