CS 189:289A Introduction to Machine Learning
Spring 2023 Jonathan Shewchuk

DIS2

1 Back to Basics: Linear Algebra

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

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

Solution: This follows immediately from the definitions: $Row(X) = \{X^\top v : v \in \mathbb{R}^n\} = Range(X^\top)$. Intuitively, the rows of X are the columns of $X^\top$, and vice versa.

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

Solution: v is in the nullspace of X if and only if $Xv = 0$, which is true if and only if $\langle X_i, v\rangle = 0$ for every row X_i of X. That is, v is perpendicular to each row of X. It follows that v is also perpendicular to any vector in the span of the rows of X, i.e. any vector in the rowspace $Row(X)$, which means v is in the orthogonal component of $Row(X)$, $Row(X)^\perp$. We can write $\mathcal{N}(X) = Row(X)^\perp$ to express the fact that v is in the nullspace of X if and only if v is in the orthogonal complement of the span of the rows of X. This is equivalent to the statement $\mathcal{N}(X)^\perp = Row(X)$.

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

Solution: If v is in the nullspace of X (that is $Xv = 0$), then $(X^\top X)v = X^\top(Xv) = X^\top 0 = 0$, meaning v is also in the nullspace of $X^\top X$. Proving the reverse implication is a bit harder.

If v is in the nullspace of $X^\top X$, i.e. $X^\top Xv = 0$, then we have $v^\top X^\top Xv = v^\top 0 = 0$. Observe that $v^\top X^\top Xv = \|Xv\|_2^2$. It follows that $\|Xv\|_2^2 = 0$, which implies that $Xv = 0$, meaning v is in the nullspace of 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.*

Solution: Since $\mathcal{N}(X)$ is a subspace of $\mathbb{R}^m$, it has a complementary subspace $\mathcal{N}(X)^\perp$ with the property that

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

From (ii) we know $\dim \mathcal{N}(X)^\perp = \dim Row(X) = Rank(X)$, yielding

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

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.

2 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}.

Solution: By definition, this means $\mathbf{Av} = \lambda\mathbf{v}$. Then

$$\mathbf{v} = \mathbf{A}^{-1}\mathbf{Av} = \mathbf{A}^{-1}(\lambda\mathbf{v}) = \lambda\mathbf{A}^{-1}\mathbf{v}$$

We know $\lambda \neq 0$ since $\mathbf{A}$ is invertible, so division by λ is valid, giving $\lambda^{-1}\mathbf{v} = \mathbf{A}^{-1}\mathbf{v}$.

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

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

Solution: *Start with the reverse direction. We wish to prove: if the eigenvalues are nonnegative,* $\mathbf{A}$ *is PSD.*

The spectral theorem of $\mathbf{A}$ allows us to decompose a symmetric matrix $\mathbf{A}$ into $\mathbf{U}\Lambda\mathbf{U}^\top$, where Λ is diagonal with eigenvalues λ_i on the diagonal and $\mathbf{U}$ is orthonormal. Define $\mathbf{z} = \mathbf{U}^\top\mathbf{v}$; since $\mathbf{U}$ is orthonormal, there exists a one-to-one mapping between all $\mathbf{z}, \mathbf{v}$.

$$\mathbf{v}^\top\mathbf{Av} = \mathbf{v}^\top(\mathbf{U}\Lambda\mathbf{U}^\top)\mathbf{v} = \mathbf{z}^\top\Lambda\mathbf{z} = \Sigma_{i=1}^n \lambda_i z_i^2$$

We assume $\lambda_i \geq 0$, so $\forall \mathbf{v}, \mathbf{v}^\top\mathbf{Av} = \sum_{i=1}^n \lambda_i z_i^2 \geq 0$, which is the definition of PSD.

Next, take the forward direction. We wish to prove: if $\mathbf{A}$ *is PSD, the eigenvalues are nonnegative.*

Since $\mathbf{A}$ is PSD, we know $\forall \mathbf{x}, \mathbf{x}^\top\mathbf{Ax} \geq 0$. So for all i, take the ith eigenvector $\mathbf{u}_i$ for $\mathbf{A}$. Then,

$$\mathbf{u}_i^\top\mathbf{Au}_i = \mathbf{u}_i^\top(\lambda_i\mathbf{u}_i) = \lambda_i\mathbf{u}_i^\top\mathbf{u}_i = \lambda_i\|\mathbf{u}_i\|_2^2 \geq 0$$

Since $\lambda_i\|\mathbf{u}_i\|_2^2 \geq 0$ and $\|\mathbf{u}_i\|_2^2 \geq 0$, we must have that $\lambda_i \geq 0$

3 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.

Solution: $Z = \max\{X_1, \ldots, X_n\}$.

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

Solution:

$$F(z) = P(Z \leq z) = P(X_1 \leq z)\, P(X_2 \leq z) \cdots P(X_n \leq z) = \prod_{i=1}^{n} P(X_i \leq z)$$

Since each X_i is uniformly distributed between 0 and 1, $P(X_i \leq z) = z$. Thus,

$$F(z) = \begin{cases} 0 & \text{if } z < 0, \\ z^n & \text{if } 0 \leq z \leq 1, \\ 1 & \text{if } z > 1. \end{cases}$$

(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.*

Solution:

$$\begin{aligned}
\mathbb{E}[Z] &= \int_0^\infty \mathbb{P}(Z \geq t)\, dt \\
&= \int_0^\infty (1 - \mathbb{P}(Z < t))\, dt \\
&= \int_0^\infty (1 - F(t))\, dt \\
&= \int_0^1 (1 - t^n)\, dt + \int_1^\infty (1 - 1)\, dt \\
&= \frac{n}{n+1}
\end{aligned}$$

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

Solution: $\mathbb{E}[Z]$ increases as n increases, and as $n \to \infty$, $\mathbb{E}[Z] \to 1$. This makes intuitive sense because increasing the number of archers increases the likelihood that more extreme values are encountered, which causes the max to tend towards the positive extreme (in this case, $Z = 1$).

4 Vector Calculus [1]

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.

Solution: 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 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 $\nabla f(\mathbf{x}) = \left(\frac{\partial f}{\partial \mathbf{x}}\right)^\top$. So 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*.

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

[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 $||\Delta|| \to 0$.

matrix. We have

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

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

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

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

An argument via first principle follows.

$$y(\mathbf{B} + \Delta) = y(\mathbf{B}) + \text{trace}\left(\frac{\partial y}{\partial \mathbf{B}}\, \Delta\right) + o(\|\Delta\|). \tag{5}$$

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 $\mathbf{x}$:

$$\mathbf{z}(\mathbf{x} + \Delta) = \mathbf{z}(\mathbf{x}) + \frac{\partial \mathbf{z}}{\partial \mathbf{x}}\, \Delta + o(\|\Delta\|). \tag{6}$$

A formula for the same can be derived as

$$J(\mathbf{z}) = \frac{\partial \mathbf{z}}{\partial \mathbf{x}} \in \mathbb{R}^{k\times d} = \begin{bmatrix} \frac{\partial z_1}{\partial \mathbf{x}} \\ \frac{\partial z_2}{\partial \mathbf{x}} \\ \vdots \\ \frac{\partial z_k}{\partial \mathbf{x}} \end{bmatrix} = \begin{bmatrix} \frac{\partial z_1}{\partial x_1} & \frac{\partial z_1}{\partial x_2} & \cdots & \frac{\partial z_1}{\partial x_d} \\ \frac{\partial z_2}{\partial x_1} & \frac{\partial z_2}{\partial x_2} & \cdots & \frac{\partial z_2}{\partial x_d} \\ \vdots & & & \\ \frac{\partial z_k}{\partial x_1} & \frac{\partial z_k}{\partial x_2} & \cdots & \frac{\partial z_k}{\partial x_d} \end{bmatrix}, \tag{7}$$

$$\text{that is} \quad [J(\mathbf{z})]_{ij} = \left[\frac{\partial \mathbf{z}}{\partial \mathbf{x}}\right]_{ij} = \frac{\partial z_i}{\partial x_j}. \tag{8}$$

4. However, the Hessian of f is defined as

$$H(f) = \nabla^2 f(\mathbf{x}) = J(\nabla f)^\top = \begin{bmatrix} \frac{\partial z_1}{\partial x_1} & \frac{\partial z_2}{\partial x_1} & \cdots & \frac{\partial z_d}{\partial x_1} \\ \frac{\partial z_1}{\partial x_2} & \frac{\partial z_2}{\partial x_2} & \cdots & \frac{\partial z_d}{\partial x_2} \\ \vdots & & & \\ \frac{\partial z_1}{\partial x_d} & \frac{\partial z_2}{\partial x_d} & \cdots & \frac{\partial z_d}{\partial x_d} \end{bmatrix} = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_d} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2^2} & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_d} \\ \vdots & & & \\ \frac{\partial^2 f}{\partial x_d \partial x_1} & \frac{\partial^2 f}{\partial x_d \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_d^2}. \end{bmatrix} \tag{9}$$

A first principle definition is

$$\nabla f(\mathbf{x} + \Delta) \approx \nabla f(\mathbf{x}) + \nabla^2 f(\mathbf{x})\Delta \tag{10}$$

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 and in such cases (which cover a lot of cases in daily use) the convention does not matter.

5. The following linear algebra formulas are also helpful:

$$(\mathbf{A}\mathbf{x})_i = \sum_{j=1}^{d} A_{ij}x_j, \quad \text{and,} \tag{11}$$

$$(\mathbf{A}^\top\mathbf{x})_i = \sum_{j=1}^{d} \mathbf{A}^\top_{ij}x_j = \sum_{j=1}^{d} A_{ji}x_j. \tag{12}$$

Calculate the following derivatives.

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

Solution: We discuss two ways to solve the problem.

Using computation via first principle: We use $f(\mathbf{x}) = \mathbf{w}^\top\mathbf{x}$. Then we have

$$f(\mathbf{x}+\Delta) = \mathbf{w}^\top(\mathbf{x}+\Delta) = \mathbf{w}^\top\mathbf{x} + \mathbf{w}^\top\Delta = f(\mathbf{x}) + \mathbf{w}^\top\Delta.$$

Comparing with equation (1), we conclude that

$$\frac{\partial \mathbf{w}^\top\mathbf{x}}{\partial \mathbf{x}} = \mathbf{w}^\top \quad \textit{and thus} \quad \nabla_{\mathbf{x}}(\mathbf{w}^\top\mathbf{x}) = \left(\frac{\partial \mathbf{w}^\top\mathbf{x}}{\partial \mathbf{x}}\right)^\top = \mathbf{w}.$$

Using the formula (2)**:** The idea is to use $f = \mathbf{w}^\top\mathbf{x}$ and apply equation (2). Note that $\mathbf{w}^\top\mathbf{x} = \sum_j w_j x_j$. Hence, we have

$$\frac{\partial f}{\partial x_i} = \frac{\partial \sum_j w_j x_j}{\partial x_i} = w_i.$$

Thus, we find that

$$\frac{\partial \mathbf{w}^\top\mathbf{x}}{\partial \mathbf{x}} = \frac{\partial \sum_j w_j x_j}{\partial \mathbf{x}} = \left[\frac{\partial \sum_j w_j x_j}{\partial x_1}, \frac{\partial \sum_j w_j x_j}{\partial x_2}, \ldots, \frac{\partial \sum_j w_j x_j}{\partial x_d}\right] = \left[w_1, w_2, \ldots, w_d\right] = \mathbf{w}^\top.$$

And $\nabla_{\mathbf{x}}(\mathbf{w}^\top\mathbf{x}) = \frac{\partial \mathbf{w}^\top\mathbf{x}}{\partial \mathbf{x}}^\top = \mathbf{w}$.

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

Solution: We discuss three ways to solve the problem.

Using part (a): Note that we can solve this question simply by using part (a). We substitute $\mathbf{u} = \mathbf{A}^\top \mathbf{w}$ to obtain that $f(\mathbf{x}) = \mathbf{u}^\top \mathbf{x}$. Now from part (a), we conclude that

$$\begin{aligned}\nabla_{\mathbf{x}}(\mathbf{w}^\top \mathbf{A}\mathbf{x}) &= \nabla_{\mathbf{x}}(\mathbf{u}^\top \mathbf{x}) \\ &= \mathbf{u} \\ &= \mathbf{A}^\top \mathbf{w}.\end{aligned}$$

Using computation via first principle: Taking $f(\mathbf{x}) = \mathbf{w}^\top \mathbf{A}\mathbf{x}$ and expanding, we have

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

Comparing with equation (1), we conclude that

$$\frac{\partial \mathbf{w}^\top \mathbf{A}\mathbf{x}}{\partial \mathbf{x}} = \mathbf{w}^\top \mathbf{A} \quad \textit{and} \quad \nabla_{\mathbf{x}}(\mathbf{w}^\top \mathbf{A}\mathbf{x}) = \left(\frac{\partial \mathbf{w}^\top \mathbf{A}\mathbf{x}}{\partial \mathbf{x}}\right)^\top = \mathbf{A}^\top \mathbf{w}.$$

Using the formula (2)**:** The idea is to use $f(\mathbf{x}) = \mathbf{w}^\top \mathbf{A}\mathbf{x}$, and apply equation (2). Using the fact that $\mathbf{w}^\top \mathbf{A}\mathbf{x} = \sum_{i=1}^d \sum_{j=1}^d w_i A_{ij} x_j$, we find that

$$\frac{\partial f}{\partial x_j} = \frac{\partial \sum_{i=1}^d \sum_{j=1}^d w_i A_{ij} x_j}{\partial x_j} = \frac{\partial \sum_{j=1}^d x_j (\sum_{i=1}^d A_{ij} w_i)}{\partial x_j} = \sum_{i=1}^d A_{ij} w_i = \sum_{i=1}^d A_{ji}^\top w_i = (\mathbf{A}^\top \mathbf{w})_j,$$

where in the last step we have used equation (12). Consequently, we have

$$\frac{\partial (\mathbf{w}^\top \mathbf{A}\mathbf{x})}{\partial \mathbf{x}} = \left[(\mathbf{A}^\top \mathbf{w})_1, (\mathbf{A}^\top \mathbf{w})_2, \ldots, (\mathbf{A}^\top \mathbf{w})_d\right] = (\mathbf{A}^\top \mathbf{w})^\top = \mathbf{w}^\top \mathbf{A},$$

and

$$\nabla_{\mathbf{x}}(\mathbf{w}^\top \mathbf{A}\mathbf{x}) = \left(\frac{\partial (\mathbf{w}^\top \mathbf{A}\mathbf{x})}{\partial \mathbf{x}}\right)^\top = \mathbf{A}^\top \mathbf{w}.$$

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

Solution:

We discuss two approaches to solve this problem.

Using computation via first principle (5)**:** Treating $y = \mathbf{w}^\top \mathbf{A}\mathbf{x}$ as a function of A and expanding with respect to change in A, we have

$$y(\mathbf{A} + \Delta) = \mathbf{w}^\top (\mathbf{A} + \Delta)\mathbf{x} = \mathbf{w}^\top \mathbf{A}\mathbf{x} + \mathbf{w}^\top \Delta x.$$

Note that, for two matrices $M \in \mathbb{R}^{m \times n}$ and $N \in \mathbb{R}^{n \times m}$, we have

$$\text{trace}(\mathbf{MN}) = \text{trace}(\mathbf{NM}).$$

Since $\mathbf{w}^\top \Delta\mathbf{x}$ is a scalar, we can write $\mathbf{w}^\top \Delta\mathbf{x} = \text{trace}(\mathbf{w}^\top \Delta\mathbf{x})$. And using the trace trick, we obtain

$$\mathbf{w}^\top \Delta\mathbf{x} = \text{trace}(\mathbf{w}^\top \Delta\mathbf{x}) = \text{trace}(\mathbf{x}\mathbf{w}^\top \Delta).$$

Thus, we have

$$y(\mathbf{A}+\Delta) = \mathbf{w}^\top(\mathbf{A}+\Delta)\mathbf{x} = \mathbf{w}^\top\mathbf{A}\mathbf{x} + \mathbf{w}^\top\Delta\mathbf{x} = y(\mathbf{A}) + \text{trace}(\mathbf{x}\mathbf{w}^\top\Delta),$$

which on comparison with equation (5) yields that

$$\frac{\partial(\mathbf{w}^\top\mathbf{A}\mathbf{x})}{\partial\mathbf{A}} = \mathbf{x}\mathbf{w}^\top \quad \textit{and} \quad \nabla_{\mathbf{A}}(\mathbf{w}^\top\mathbf{A}\mathbf{x}) = \left[\frac{\partial(\mathbf{w}^\top\mathbf{A}\mathbf{x})}{\partial\mathbf{A}}\right]^\top = \mathbf{w}\mathbf{x}^\top.$$

Using the formula (4)**:** We use $y = \mathbf{w}^\top\mathbf{A}\mathbf{x}$ and apply the formula (4). We have $\mathbf{w}^\top\mathbf{A}\mathbf{x} = \sum_{i=1}^d \sum_{j=1}^d w_i A_{ij} x_j$ and hence

$$\left[\frac{\partial(\mathbf{w}^\top\mathbf{A}\mathbf{x})}{\partial\mathbf{A}}\right]_{ij} = \frac{\partial(\mathbf{w}^\top\mathbf{A}\mathbf{x})}{\partial A_{ji}} = w_j x_i = (\mathbf{x}\mathbf{w}^\top)_{ij}.$$

Consequently, we have

$$\frac{\partial(\mathbf{w}^\top\mathbf{A}\mathbf{x})}{\partial\mathbf{A}} = [(\mathbf{x}\mathbf{w}^\top)_{ij}] = \mathbf{x}\mathbf{w}^\top,$$

and thereby $\nabla_{\mathbf{A}}(\mathbf{w}^\top\mathbf{A}\mathbf{x}) = \mathbf{w}\mathbf{x}^\top$.

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

Solution:

We provide a few ways to solve this problem. Taking $f(\mathbf{x}) = \mathbf{x}^\top\mathbf{A}\mathbf{x}$ and expanding, we have

$$\begin{aligned} f(\mathbf{x}+\Delta) &= (\mathbf{x}+\Delta)^\top\mathbf{A}(\mathbf{x}+\Delta) \\ &= \mathbf{x}^\top\mathbf{A}\mathbf{x} + \Delta^\top\mathbf{A}\mathbf{x} + \mathbf{x}^\top\mathbf{A}\Delta + \Delta^\top\mathbf{A}\Delta \\ &= f(\mathbf{x}) + (\mathbf{x}^\top\mathbf{A}^\top + \mathbf{x}^\top\mathbf{A})\Delta + O(\|\Delta\|^2) \end{aligned}$$

which yields

$$\begin{aligned} \frac{\partial(\mathbf{x}^\top\mathbf{A}\mathbf{x})}{\partial\mathbf{x}} &= \mathbf{x}^\top(\mathbf{A}^\top + \mathbf{A}) \qquad \textit{and}, \\ \nabla_{\mathbf{x}}(\mathbf{x}^\top\mathbf{A}\mathbf{x}) &= \left[\frac{\partial(\mathbf{x}^\top\mathbf{A}\mathbf{x})}{\partial\mathbf{x}}\right]^\top = (\mathbf{A}+\mathbf{A}^\top)\mathbf{x}. \end{aligned}$$

Using the chain rule, and parts (b) and (c): We have

$$\frac{\partial(\mathbf{x}^\top\mathbf{A}\mathbf{x})}{\partial\mathbf{x}} = \frac{\partial\,\mathbf{w}^\top\mathbf{A}\mathbf{x}}{\partial\mathbf{x}}(\mathbf{x})\bigg|_{\mathbf{w}=\mathbf{x}} + \frac{\partial\,\mathbf{w}^\top\mathbf{A}\mathbf{x}}{\partial\mathbf{w}}(\mathbf{w})\bigg|_{\mathbf{w}=\mathbf{x}} = \mathbf{w}^\top\mathbf{A}|_{\mathbf{w}=\mathbf{x}} + \mathbf{x}^\top\mathbf{A}^\top|_{\mathbf{w}=\mathbf{x}} = \mathbf{x}^\top(\mathbf{A}+\mathbf{A}^\top)$$

and thereby $\nabla_{\mathbf{x}}(\mathbf{x}^\top\mathbf{A}\mathbf{x}) = \left[\frac{\partial(\mathbf{x}^\top\mathbf{A}\mathbf{x})}{\partial\mathbf{x}}\right]^\top = (\mathbf{A}+\mathbf{A}^\top)\mathbf{x}$.

Using the product rule: We have

$$\frac{\partial(\mathbf{x}^\top \mathbf{A}\mathbf{x})}{\partial \mathbf{x}} = \mathbf{x}^\top \frac{\partial(\mathbf{A}\mathbf{x})}{\partial \mathbf{x}} + (\mathbf{A}\mathbf{x})^\top \frac{\partial(\mathbf{x})}{\partial \mathbf{x}} = \mathbf{x}^\top \mathbf{A} + \mathbf{x}^\top \mathbf{A}^\top = \mathbf{x}^\top (\mathbf{A} + \mathbf{A}^\top)$$

and thereby $\nabla_{\mathbf{x}}(\mathbf{x}^\top \mathbf{A}\mathbf{x}) = \left[\frac{\partial(\mathbf{x}^\top \mathbf{A}\mathbf{x})}{\partial \mathbf{x}}\right]^\top = (\mathbf{A} + \mathbf{A}^\top)\mathbf{x}$.

Using the formula (2)**:** We have $\mathbf{x}^\top \mathbf{A}\mathbf{x} = \sum_{i=1}^d \sum_{j=1}^d x_i A_{ij} x_j$. For any given index ℓ, we have

$$\mathbf{x}^\top \mathbf{A}\mathbf{x} = A_{\ell\ell} x_\ell^2 + x_\ell \sum_{j \neq \ell} (A_{j\ell} + A_{\ell j}) x_j + \sum_{i \neq \ell} \sum_{j \neq \ell} x_i A_{ij} x_j.$$

Thus we have

$$\frac{\partial \mathbf{x}^\top \mathbf{A}\mathbf{x}}{\partial x_\ell} = 2A_{\ell\ell} x_\ell + \sum_{j \neq \ell} (A_{j\ell} + A_{\ell j}) x_j = \sum_{j=1}^d (A_{j\ell} + A_{\ell j}) x_j = ((\mathbf{A}^\top + \mathbf{A})\mathbf{x})_\ell.$$

And consequently

$$\begin{aligned}
\frac{\partial \mathbf{x}^\top \mathbf{A}\mathbf{x}}{\partial \mathbf{x}} &= \left[\frac{\partial \mathbf{x}^\top \mathbf{A}\mathbf{x}}{\partial x_1}, \frac{\partial \mathbf{x}^\top \mathbf{A}\mathbf{x}}{\partial x_2}, \ldots, \frac{\partial \mathbf{x}^\top \mathbf{A}\mathbf{x}}{\partial x_d}\right] \\
&= \left[((\mathbf{A}^\top + \mathbf{A})\mathbf{x})_1, ((\mathbf{A}^\top + \mathbf{A})\mathbf{x})_2, \ldots, ((\mathbf{A}^\top + \mathbf{A})\mathbf{x})_d\right] \\
&= ((\mathbf{A}^\top + \mathbf{A})\mathbf{x})^\top \\
&= \mathbf{x}^\top (\mathbf{A} + \mathbf{A}^\top),
\end{aligned}$$

and hence $\nabla_{\mathbf{x}}(\mathbf{x}^\top \mathbf{A}\mathbf{x}) = \left[\frac{\partial(\mathbf{x}^\top \mathbf{A}\mathbf{x})}{\partial \mathbf{x}}\right]^\top = (\mathbf{A} + \mathbf{A}^\top)\mathbf{x}$.

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

Solution:

We discuss two ways to solve this problem.

Using computation via first principle: We expand $z(\mathbf{x}) = \nabla f(\mathbf{x}) = (\mathbf{A} + \mathbf{A}^\top)\mathbf{x}$ and find that

$$z(\mathbf{x} + \Delta) = (\mathbf{A} + \mathbf{A}^\top)\mathbf{x} + (\mathbf{A} + \mathbf{A}^\top)\Delta = \nabla f(\mathbf{x}) + (\mathbf{A} + \mathbf{A}^\top)\Delta.$$

Relating with equation (10), we obtain that $\nabla^2 f(\mathbf{x}) = \mathbf{A} + \mathbf{A}^\top$.

Using the formula (9)**:** A straight forward computation yields that

$$\frac{\partial^2 f}{\partial x_i \partial x_j} = A_{ij} + A_{ji}$$

and hence

$$\nabla^2 f(\mathbf{x}) = \left[\frac{\partial^2 f}{\partial x_i \partial x_j}\right] = [(A_{ij} + A_{ji})] = \mathbf{A} + \mathbf{A}^\top.$$

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

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

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.*

Solution: Let $L(w) = \frac{1}{2}\|Xw - Y\|_2^2$. From calculus, we know a necessary condition of any potential solution w^* is that it must be a critical point of L, that is $\nabla_w L(w^*) = 0$. (It turns out, this function is also convex, so this is also a sufficient condition). Thus, we have

$$\begin{aligned}
\nabla_w L(w) &= \nabla_w \frac{1}{2}\|Xw - Y\|_2^2 \\
&= \nabla_w \frac{1}{2}(Xw - Y)^\top (Xw - Y) \\
&= \frac{1}{2}\nabla_w (w^\top X^\top Xw - 2Y^\top Xw + Y^\top Y) \\
&= \frac{1}{2}\nabla_w (w^\top X^\top Xw) - \nabla_w (Y^\top Xw) + \frac{1}{2}\nabla_w (Y^\top Y)
\end{aligned}$$

Because $Y^\top Y$ is constant w.r.t w, that term disappears from the gradient. Additionally, from (e), and the fact that $X^\top X$ is symmetric, we know

$$\nabla_w (w^\top X^\top Xw) = (X^\top X + (X^\top X)^\top)w = 2X^\top Xw$$

Finally, we apply (b) to the second term to get

$$\nabla_w L(w) = X^\top Xw - X^\top Y$$

Setting the gradient equal to zero we arrive at the necessary (and sufficient) condition

$$X^\top (Xw^* - Y) = 0$$

Note: If X is full rank then $X^\top X$ is invertible, and we can solve for w^* exactly: $w^* = (X^\top X)^{-1} X^\top Y$. Otherwise, there may be infinite possible w that satisfy the above condition

Note for the mathematically adventurous: The above condition says the residual error vector, $Xw^* - y$, is in the null space of $X^\top$. However, by the fundamental theorem of linear algebra, we know that $\mathcal{N}(X^\top) \perp \mathit{Range}(X)$. Thus, the above condition is equivalent to saying that the error vector of the optimal projection onto $\mathit{Range}(X)$ is orthogonal to $\mathit{Range}(X)$, hence the term "orthogonal projection."