1 Back to Basics: Linear Algebra

Let \( X \in \mathbb{R}^{m \times n} \). We do not assume that \( X \) has full rank.

(a) Give the definition of the rowspace, column space, and nullspace of \( X \).

Solution: The rowspace is the span (or the set of all linear combinations) of the rows of \( X \), the column space is the span of the columns of \( X \), also known as \( \text{Range}(X) \), and the nullspace is the set of vectors \( v \) such that \( Xv = 0 \), also known as \( \text{N}(X) \).

(b) Check the following facts:

(a) The rowspace of \( X \) is the column space of \( X^\top \), and vice versa.

Solution: The rows of \( X \) are the columns of \( X^\top \), and vice versa.

(b) The nullspace of \( X \) and the row space of \( X \) are orthogonal complements.

Solution: \( v \) is in the nullspace of \( X \) if and only if \( Xv = 0 \), which is true if and only if for every row \( X_i \) of \( X \), \( \langle X_i, v \rangle = 0 \). This is precisely the condition that \( v \) is perpendicular to each row of \( X \). This means 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 \( X \), i.e. the orthogonal complement of the rowspace of \( X \).

(c) The nullspace of \( X^\top X \) is the same as the nullspace of \( X \). Hint: if \( v \) is in the nullspace of \( X^\top X \), then \( v^\top X^\top Xv = 0 \).

Solution: If \( v \) is in the nullspace of \( X \), then \( X^\top Xv = X^\top 0 = 0 \). On the other hand, if \( v \) is in the nullspace of \( X^\top X \), then \( v^\top X^\top Xv = v^\top 0 = 0 \). Then, \( v^\top X^\top Xv = \|Xv\|_2^2 = 0 \), which implies that \( Xv = 0 \).

(d) The column space and rowspace of \( X^\top X \) are the same, and are equal to the rowspace of \( X \). Hint: Use the relationship between nullspace and rowspace.

Solution: \( X^\top X \) is symmetric, and by part (a),

\[
\text{rowspace}(X^\top X) = \text{column space}((X^\top X)^\top) = \text{column space}(X^\top X)
\]

By part (b), (c), then (b) again,

\[
\text{rowspace}(X^\top X) = \text{nullspace}(X^\top X)^\perp = \text{nullspace}(X)^\perp = \text{rowspace}(X),
\]

where \((\cdot)^\perp\) denotes orthogonal complement.
2 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 $Z$ that corresponds with the worst (highest) score.

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

(b) Derive the Cumulative Distribution Function (CDF) of $Z$.

**Solution:**

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

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

(c) Derive the Probability Density Function (PDF) of $Z$.

**Solution:**

$$f(z) = \frac{d}{dz} F(z) = \begin{cases} 
    nz^{n-1} & \text{if } 0 \leq z \leq 1, \\
    0 & \text{otherwise}.
\end{cases}$$

(d) Calculate the expected value of $Z$.

**Solution:**

$$E[Z] = \int_{-\infty}^{\infty} z f(z) \, dz = \int_{0}^{1} zn^{n-1} \, dz = n \int_{0}^{1} z^n \, dz = \frac{n}{n+1}.$$
Below, \( \mathbf{x} \in \mathbb{R}^d \) means that \( \mathbf{x} \) is a \( d \times 1 \) column vector with real-valued entries. Likewise, \( A \in \mathbb{R}^{d \times d} \) means that \( A \) 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 \( A \in \mathbb{R}^{d \times d} \). In the following questions, \( \frac{\partial}{\partial \mathbf{x}} \) denotes the derivative with respect to \( \mathbf{x} \), while \( \nabla \mathbf{x} \) denotes the gradient with respect to \( \mathbf{x} \).

**Solution:** Let us first understand the definition of the derivative. Let \( f : \mathbb{R}^d \rightarrow \mathbb{R} \) 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 \( r(\Delta)/\|\Delta\| \rightarrow 0 \) as \( \|\Delta\| \rightarrow 0 \). 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 \( \|\Delta\| \rightarrow 0 \), we have

\[
r(\Delta)/\|\Delta\| = \frac{\|\Delta\|^2}{\|\Delta\|} = \|\Delta\| \rightarrow 0,
\]

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.

We call \( \frac{\partial f}{\partial \mathbf{x}} \) the **derivative of \( f \) at \( \mathbf{x} \)**. Sometimes we use \( \frac{df}{dx} \) but we also use \( \partial \) to indicate that \( f \) may depend on some other variable too. (But to define \( \frac{df}{dx} \), we study changes in \( f \) with respect to changes in only \( \mathbf{x} \).)

Since \( \Delta \) 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 \( \nabla f(\mathbf{x}) = \left( \frac{\partial f}{\partial \mathbf{x}} \right)^T \).

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 \( f : \mathbb{R}^d \rightarrow \mathbb{R} \) denote a scalar function. Let \( \mathbf{x} \in \mathbb{R}^d \) denote a vector and \( A \in \mathbb{R}^{d \times d} \) denote a matrix. We have

\[
\frac{\partial f}{\partial \mathbf{x}} \in \mathbb{R}^{1 \times d}
\]

such that

\[
\frac{\partial f}{\partial \mathbf{x}} = \left[ \frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, \ldots, \frac{\partial f}{\partial x_d} \right] \tag{2}
\]

---

1. Good resources for matrix calculus are:
   - The Matrix Cookbook: [https://www.math.uwaterloo.ca/~hwolkowi/matrixcookbook.pdf](https://www.math.uwaterloo.ca/~hwolkowi/matrixcookbook.pdf)
   - Khan Academy: [https://www.khanacademy.org/math/multivariable-calculus/multivariable-derivatives](https://www.khanacademy.org/math/multivariable-calculus/multivariable-derivatives)
   - YouTube: [https://www.youtube.com/playlist?list=PLSQl0a2vh4HC5feHa6RC5c0wbRTx56nF7](https://www.youtube.com/playlist?list=PLSQl0a2vh4HC5feHa6RC5c0wbRTx56nF7)

2. Note that \( r(\Delta) = \sqrt{\|\Delta\|} \) is not an \( o(\|\Delta\|) \) term. Since for this case, \( r(\Delta)/\|\Delta\| = 1/\sqrt{\|\Delta\|} \rightarrow \infty \) as \( \|\Delta\| \rightarrow 0 \).
\[ \nabla_x f = \left( \frac{\partial f}{\partial x} \right)^\top = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \\ \vdots \\ \frac{\partial f}{\partial x_d} \end{bmatrix}. \] (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

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

3. For \( z : \mathbb{R}^d \to \mathbb{R}^k \) a vector-valued function; its derivative \( \frac{\partial z}{\partial x} \) is an operator such that it can help find the change in function value at \( x \), up to first order, when we add a little perturbation \( \Delta \) to \( x \):

\[ z(x + \Delta) = z(x) + \frac{\partial z}{\partial x} \Delta + o(\|\Delta\|). \] (5)

A formula for the same can be derived as

\[ J(z) = \frac{\partial z}{\partial x} \in \mathbb{R}^{k \times d} = \begin{bmatrix} \frac{\partial z_1}{\partial x_1} & \frac{\partial z_2}{\partial x_1} & \cdots & \frac{\partial z_k}{\partial x_1} \\ \frac{\partial z_1}{\partial x_2} & \frac{\partial z_2}{\partial x_2} & \cdots & \frac{\partial z_k}{\partial x_2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial z_1}{\partial x_d} & \frac{\partial z_2}{\partial x_d} & \cdots & \frac{\partial z_k}{\partial x_d} \end{bmatrix}, \] (6)

that is \[ [J(z)]_{ij} = \frac{\partial z_i}{\partial x_j}. \] (7)

4. However, the Hessian of \( f \) is defined as

\[ H(f) = \nabla^2 f(x) = J(\nabla f)^\top = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1 \partial x_1} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_d} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_d} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_d \partial x_1} & \frac{\partial^2 f}{\partial x_d \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_d \partial x_d} \end{bmatrix}. \] (8)

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:

\[ (Ax)_i = \sum_{j=1}^{d} A_{ij}x_j, \quad \text{and,} \]

\[ (A^\top x)_i = \sum_{j=1}^{d} A_{ji}^\top x_j = \sum_{j=1}^{d} A_{ji}x_j. \] (9) (10)
Derive the following derivatives.

(a) $\frac{\partial w^\top x}{\partial x}$ and $\nabla_x (w^\top x)$

**Solution:**
The idea is to use $f = w^\top x$ and apply equation (2). Note that $w^\top x = \sum_j w_j x_j$. 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

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

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

(b) $\frac{\partial (w^\top Ax)}{\partial x}$ and $\nabla_x (w^\top Ax)$

**Solution:** We discuss two ways to solve the problem.

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

$$\frac{\partial (u^\top x)}{\partial x} = u^\top = w^\top A \quad \text{and} \quad \nabla_x (w^\top Ax) = \left( \frac{\partial (w^\top Ax)}{\partial x} \right)^\top = A^\top w.$$

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

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

where in the last step we have used equation (10). Consequently, we have

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

and

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

(c) $\frac{\partial (w^\top Ax)}{\partial w}$ and $\nabla_w (w^\top Ax)$

**Solution:** We discuss two ways to solve the problem.
Using part (a) and (b): Note that we can solve this question simply by using part (a) and (b). We have $(w^\top Ax) = (x^\top A^\top w)$, since for a scalar $\alpha$, we have $\alpha = \alpha^\top$. And in part (b), reversing the roles of $x$ and $w$, we obtain that

$$ \frac{\partial w^\top Ax}{\partial w} = \frac{\partial x^\top A^\top w}{\partial w} = x^\top A^\top $$ and $\nabla_w (w^\top Ax) = \left( \frac{\partial w^\top Ax}{\partial w} \right)^\top = Ax$.

Using the formula (2): Using a similar idea as in the previous part, we have

$$ \frac{\partial f}{\partial w_i} = \frac{\partial}{\partial w_i} \left( \sum_{i=1}^d \sum_{j=1}^d w_i A_{ij} x_j \right) = \frac{\partial}{\partial w_i} \left( \sum_{i=1}^d (\sum_{j=1}^d A_{ij} x_j) \right) = \sum_{j=1}^d A_{ij} x_j = (Ax)_i, $$

where in the last step we have used equation (9). Consequently, we have

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

and

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

(d) $\frac{\partial (w^\top Ax)}{\partial A}$ and $\nabla_A (w^\top Ax)$

Solution:

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

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

Consequently, we have

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

and thereby $\nabla_A (w^\top Ax) = wx^\top$.

(e) $\frac{\partial (x^\top Ax)}{\partial x}$ and $\nabla_x (x^\top Ax)$

Solution:

We provide two ways to solve this problem.
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 \( \ell \), we have

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

Thus we have

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

And consequently

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

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

Using the product rule: Let

\[
g(\mathbf{x}) = \mathbf{x},
\]

\[
h(\mathbf{x}) = \mathbf{A} \mathbf{x}.
\]

We have that

\[
\frac{dg(\mathbf{x})}{d\mathbf{x}} = \mathbf{I},
\]

\[
\frac{dh(\mathbf{x})}{d\mathbf{x}} = \mathbf{A}.
\]

The product rule says that

\[
\frac{d(\mathbf{x}^\top \mathbf{A} \mathbf{x})}{d\mathbf{x}} = \frac{d(g(\mathbf{x})^\top h(\mathbf{x}))}{d\mathbf{x}} = g(\mathbf{x})^\top \frac{dh(\mathbf{x})}{d\mathbf{x}} + h(\mathbf{x})^\top \frac{dg(\mathbf{x})}{d\mathbf{x}} = \mathbf{x}^\top \mathbf{A} + \mathbf{x}^\top \mathbf{A}^\top \mathbf{I} = \mathbf{x}^\top (\mathbf{A} + \mathbf{A}^\top),
\]

and hence \( \nabla_\mathbf{x} (\mathbf{x}^\top \mathbf{A} \mathbf{x}) = (\mathbf{A} + \mathbf{A}^\top) \mathbf{x} \)

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

Solution:
Using the formula (8): A straight forward computation yields that

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

and hence

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