CS 189 Spring 2018 — Introduction to Machine Learning — 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, $\frac{\partial}{\partial \mathbf{x}}$ denotes the derivative with respect to $\mathbf{x}$, while $\nabla_{\mathbf{x}}$ denote the gradient with respect to $\mathbf{x}$. Compute the following:

Solution: A good resource for matrix calculus is the matrix cookbook: `https://www.math.uwaterloo.ca/~hwolkowi/matrixcookbook.pdf` and the wikipedia page: `https://en.wikipedia.org/wiki/Matrix_calculus`.

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 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 $\mathbf{x}$. That is

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

where the term $o(\|\Delta\|)$ 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 thumb rule, any term that has a higher order dependence on $\|\Delta\|$ than linear is $o(\|\Delta\|)$ and is ignored to compute the derivative. [1]

We call $\frac{\partial f}{\partial \mathbf{x}}$ as the derivative of f at $\mathbf{x}$. Ideally, we should use $\frac{df}{d\mathbf{x}}$ but it is okay to 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 $\mathbf{x}$.)

Note that for $\frac{\partial f}{\partial \mathbf{x}}\Delta$ to be a scalar, the vector $\frac{\partial f}{\partial \mathbf{x}}$ should be a row vector, since Δ is a column vector. The gradient of f at $\mathbf{x}$ is *defined* as the transpose of this derivative. That is $\nabla f(\mathbf{x}) = \frac{\partial f}{\partial \mathbf{x}}^T$. 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 asolution via first principle might be hard to compute. We will also distinguish between the derivative, gradient, Jacobian and Hessian in our notation.

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

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 matrix. We have

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

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

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 is follows:

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

3. For $\mathbf{z} : \mathbb{R}^d \to \mathbb{R}^k$ is a vector-valued function, then 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

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

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

4. However, the Hessian of f is defined as

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

A first principle definition is given as:

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

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:

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

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

(a) $\frac{\partial \mathbf{w}^T\mathbf{x}}{\partial \mathbf{x}}$ and $\nabla_{\mathbf{x}}(\mathbf{w}^T\mathbf{x})$

Solution: We discuss two ways to solve the problem.

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

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

Comparing with equation (1), we conclude that

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

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

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

Thus, we find that

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

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

(b) $\frac{\partial (\mathbf{w}^T\mathbf{A}\mathbf{x})}{\partial \mathbf{x}}$ and $\nabla_{\mathbf{x}}(\mathbf{w}^T\mathbf{A}\mathbf{x})$

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 $\mathbf{u} = \mathbf{A}^T\mathbf{w}$ to obtain that $f(\mathbf{x}) = \mathbf{u}^T\mathbf{x}$. Now from part (a), we conclude that

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

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

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

Comparing with equation (1), we conclude that

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

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

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

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

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

and

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

(c) $\frac{\partial(\mathbf{w}^T\mathbf{A}\mathbf{x})}{\partial\mathbf{w}}$ and $\nabla_{\mathbf{w}}(\mathbf{w}^T\mathbf{A}\mathbf{x})$

Solution: We discuss three 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 $(\mathbf{w}^T\mathbf{A}\mathbf{x}) = (\mathbf{x}^T\mathbf{A}^T\mathbf{w})$, since for a scalar α, we have $\alpha = \alpha^T$. And in part (b), reversing the roles of $\mathbf{x}$ and $\mathbf{w}$, we obtain that

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

Using the first principle: We now consider f as a function of $\mathbf{w}$. Taking $f(\mathbf{w}) = \mathbf{w}^T\mathbf{A}\mathbf{x}$ and expanding, we have

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

Comparing with equation (1), we conclude that

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

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

$$\frac{\partial f}{\partial w_i} = \frac{\partial \sum_{i=1}^d \sum_{j=1}^d w_i A_{ij} x_j}{\partial w_i} = \frac{\partial \sum_{i=1}^d w_i (\sum_{j=1}^d A_{ij} x_j)}{\partial w_i} = \sum_{j=1}^d A_{ij} x_j = (\mathbf{A}\mathbf{x})_i,$$

where in the last step we have used equation (11). Consequently, we have

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

and

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

(d) $\frac{\partial(\mathbf{w}^T\mathbf{A}\mathbf{x})}{\partial \mathbf{A}}$ and $\nabla_{\mathbf{A}}(\mathbf{w}^T\mathbf{A}\mathbf{x})$

Solution:

We discuss two approaches to solve this problem.

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

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

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

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

Thus, we have

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

which on comparison with equation (5) yields that

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

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

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

Consequently, we have

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

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

(e) $\frac{\partial(\mathbf{x}^T\mathbf{A}\mathbf{x})}{\partial \mathbf{x}}$ and $\nabla_{\mathbf{x}}(\mathbf{x}^T\mathbf{A}\mathbf{x})$

Solution:

We provide three ways to solve this problem.

Using the first principle: Taking $f(\mathbf{x}) = \mathbf{x}^T\mathbf{A}\mathbf{x}$ and expanding, we have

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

which yields

$$\frac{\partial(\mathbf{x}^T\mathbf{A}\mathbf{x})}{\partial \mathbf{x}} = \mathbf{x}^T(\mathbf{A}^T + \mathbf{A}) \qquad and,$$

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

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

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

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

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

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

Thus we have

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

And consequently

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

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

(f) $\nabla^2_\mathbf{x}(\mathbf{x}^T\mathbf{A}\mathbf{x})$

Solution:

We discuss two ways to solve this problem.

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

$$z(\mathbf{x}+\Delta) = (\mathbf{A}+\mathbf{A}^T)\mathbf{x} + (\mathbf{A}+\mathbf{A}^T)\Delta.$$

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

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

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

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 $\mathbf{A}\mathbf{v} = \lambda\mathbf{v}$. Then

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

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

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

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

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

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

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

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

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

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

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

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