CS 189/289A Introduction to Machine Learning
Fall 2023 Jennifer Listgarten, Jitendra Malik DIS0

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, columnspace, and nullspace of X.

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

(b) Check (write an informal proof for) the following facts:

(a) The rowspace of X is the columnspace 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 rowspace 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 $v^\top X^\top X v = 0$.*

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

(d) The columnspace 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),

$$\text{rowspace}(X^\top X) = \text{columnspace}((X^\top X)^\top) = \text{columnspace}(X^\top X)$$

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 $()^\perp$ 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:

$$E[Z] = \int_{-\infty}^{\infty} z\, f(z)\, \mathrm{d}z = \int_0^1 z n z^{n-1}\, \mathrm{d}z = n \int_0^1 z^n\, \mathrm{d}z = \frac{n}{n+1}.$$

3 Vector Calculus

[1]

[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`.

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 $\nabla_{\mathbf{x}}$ denotes the gradient with respect to $\mathbf{x}$. Recall that $\nabla_{\mathbf{x}} f = \left(\frac{\partial f}{\partial \mathbf{x}}\right)^\top$.

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 we also 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 $\nabla f(\mathbf{x}) = \left(\frac{\partial f}{\partial \mathbf{x}}\right)^\top$.

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 matrix. We have

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

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

[2] 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. Then its derivative is an $n \times m$ matrix and is given by

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

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 $\mathbf{x}$:

$$\mathbf{z}(\mathbf{x} + \Delta) = \mathbf{z}(\mathbf{x}) + \frac{\partial \mathbf{z}}{\partial \mathbf{x}} \Delta + o(\|\Delta\|). \tag{5}$$

A formula for the same can be derived as

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

$$\text{that is} \quad [J(\mathbf{z})]_{ij} = \left[\frac{\partial \mathbf{z}}{\partial \mathbf{x}}\right]_{ij} = \frac{\partial z_i}{\partial x_j}. \tag{7}$$

4. However, the Hessian of f is defined as

$$H(f) = \nabla^2 f(\mathbf{x}) = J(\nabla f)^\top = \begin{bmatrix} \frac{\partial z_1}{\partial x_1} & \frac{\partial z_2}{\partial x_1} & \cdots & \frac{\partial z_d}{\partial x_1} \\ \frac{\partial z_1}{\partial x_2} & \frac{\partial z_2}{\partial x_2} & \cdots & \frac{\partial z_d}{\partial x_2} \\ \vdots \\ \frac{\partial z_1}{\partial x_d} & \frac{\partial z_2}{\partial x_d} & \cdots & \frac{\partial z_d}{\partial x_d} \end{bmatrix} = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_d} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2^2} & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_d} \\ \vdots \\ \frac{\partial^2 f}{\partial x_d \partial x_1} & \frac{\partial^2 f}{\partial x_d \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_d^2}. \end{bmatrix} \tag{8}$$

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:

$$(\mathbf{A}\mathbf{x})_i = \sum_{j=1}^{d} A_{ij} x_j, \quad \text{and,} \tag{9}$$

$$(\mathbf{A}^\top \mathbf{x})_i = \sum_{j=1}^{d} \mathbf{A}_{ij}^\top x_j = \sum_{j=1}^{d} A_{ji} x_j. \tag{10}$$

Derive the following derivatives.

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

Solution:

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

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

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

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 (10). 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) $\frac{\partial (\mathbf{w}^\top \mathbf{A}\mathbf{x})}{\partial \mathbf{w}}$ and $\nabla_{\mathbf{w}}(\mathbf{w}^\top \mathbf{A}\mathbf{x})$

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

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

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 (9). Consequently, we have

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

and

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

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

Solution:

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

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

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 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

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

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

Using the product rule: Let

$$\begin{aligned} g(\mathbf{x}) &= \mathbf{x}, \\ h(\mathbf{x}) &= \mathbf{A}\mathbf{x}. \end{aligned}$$

We have that

$$\begin{aligned} \frac{dg(\mathbf{x})}{d\mathbf{x}} &= \mathbf{I}, \\ \frac{dh(\mathbf{x})}{d\mathbf{x}} &= \mathbf{A}. \end{aligned}$$

The product rule says that

$$\begin{aligned}\frac{d(\mathbf{x}^\top \mathbf{A}\mathbf{x})}{d\mathbf{x}} = \frac{d(g(\mathbf{x})^\top h(\mathbf{x}))}{d\mathbf{x}} &= g(\mathbf{x})^\top \frac{dh(\mathbf{x})}{d\mathbf{x}} + h(\mathbf{x})^\top \frac{dg(\mathbf{x})}{d\mathbf{x}} \\ &= \mathbf{x}^\top \mathbf{A} + \mathbf{x}^\top \mathbf{A}^\top \mathbf{I} \\ &= \mathbf{x}^\top(\mathbf{A} + \mathbf{A}^\top),\end{aligned}$$

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

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