Maths Knowledge Overview - COMP24112

Tingting Mu TINGTINGMU@MANCHESTER.AC.UK
Department of Computer Science
University of Manchester
Manchester M13 9PL, UK

Editor: NA

1. Linear Algebra Basics

1.1 Basic Concepts and Notations

A ***matrix*** is a rectangular array of numbers arranged in rows and columns. By $\mathbf{X} \in R^{m \times n}$, we denote a matrix $\mathbf{X}$ with m rows and n columns of real-valued numbers. The notation $\mathbf{X} = [x_{ij}]$ (or $\mathbf{X} = [x_{i,j}]$) indicates that the element of $\mathbf{X}$ at its i-th row and j-th column is denoted by x_{ij} (or $x_{i,j}$):

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & x_{13} & \cdots & x_{1n} \\ x_{21} & x_{22} & x_{23} & \cdots & x_{2n} \\ x_{31} & x_{32} & x_{33} & \cdots & x_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{m1} & x_{m2} & x_{m3} & \cdots & x_{mn} \end{bmatrix}. \tag{1}$$

For instance,

$$\mathbf{A} = \begin{bmatrix} 1.2 & 4 & -0.4 \\ 3 & 0 & 1 \end{bmatrix} \tag{2}$$

is a 2×3 matrix containing two rows and three columns. Given a matrix $\mathbf{X}$, the notation $\mathbf{X}_{:,i}$ is usually used to denote its i-th column. Its i-th row can be denoted by $\mathbf{X}_{i,:}$. Its element at the i-th row and j-th column, which is referred to as the ij-th element, can be denoted by $\mathbf{X}_{ij}$.

A ***row vector*** is a matrix with one row. By $\boldsymbol{x} = [x_1, x_2, \ldots, x_n]$, we denote a row vector of dimension n. For instance, the 2nd row of the matrix $\mathbf{A}$ in Eq. (2) is

$$\mathbf{A}_{2,:} = [3,\ 0,\ 1].$$

A ***column vector*** is a matrix with one column. By $\boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$, we denote a column vector of dimension n. For instance, the 3rd column of the matrix $\mathbf{A}$ in Eq. (2) is

$$\mathbf{A}_{:,3} = \begin{bmatrix} -0.4 \\ 1 \end{bmatrix}.$$

The i-th element of a ***vector*** $\boldsymbol{x}$, which can be either a row or column vector, is denoted by x_i.

A matrix with the same number of rows and columns is called a ***square matrix***. A square matrix with ones on the diagonal and zeros everywhere else is called the ***identity*** matrix, typically denoted by $\mathbf{I}$:

$$\mathbf{I} = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix}. \tag{3}$$

An identity matrix of size n is denoted by $\mathbf{I}_n \in R^{n\times n}$. A matrix with all the non-diagonal elements equal to 0 is called a ***diagonal matrix***, typically denoted by $\mathbf{D} = \text{diag}([d_1, d_2, \dots, d_n])$:

$$\mathbf{D} = \begin{bmatrix} d_1 & 0 & 0 & \cdots & 0 \\ 0 & d_2 & 0 & \cdots & 0 \\ 0 & 0 & d_3 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & d_n \end{bmatrix}. \tag{4}$$

Clearly, $\mathbf{I} = \text{diag}([1, 1, \dots, 1])$. A diagonal matrix formed from the n-dimensional vector $\boldsymbol{x}$ is diag($\boldsymbol{x}$), written as

$$\text{diag}(\boldsymbol{x}) = \begin{bmatrix} x_1 & 0 & 0 & \cdots & 0 \\ 0 & x_2 & 0 & \cdots & 0 \\ 0 & 0 & x_3 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & x_n \end{bmatrix}. \tag{5}$$

1.2 Matrix Operations

A summary of some frequently used matrix operations is provided below.

- The ***transpose*** of a matrix $\mathbf{X}$, denoted by $\mathbf{X}^T$, is formed by "flipping" the rows and columns: $\left(\mathbf{X}^T\right)_{ij} = \mathbf{X}_{ji}$. For instance,

$$\begin{bmatrix} 1 & 0 & 0 & -7 \\ -2 & 4 & 1 & 0 \end{bmatrix}^T = \begin{bmatrix} 1 & -2 \\ 0 & 4 \\ 0 & 1 \\ -7 & 0 \end{bmatrix}. \tag{6}$$

  It has the property of $\left(\mathbf{X}^T\right)^T = \mathbf{X}$.

- The ***sum*** operation is applied to two matrices of the same size. Given two $m \times n$ matrices $\mathbf{X}$ and $\mathbf{Y}$, their sum is calculated entrywise such that $(\mathbf{X} + \mathbf{Y})_{ij} = \mathbf{X}_{ij} + \mathbf{Y}_{ij}$. For instance,

$$\begin{bmatrix} 1 & 0 & 0 & -7 \\ -2 & 4 & 1 & 0 \end{bmatrix} + \begin{bmatrix} 0 & 0 & 0 & 1 \\ 1 & 2 & 1 & 1 \end{bmatrix} = \begin{bmatrix} 1+0 & 0+0 & 0+0 & -7+1 \\ -2+1 & 4+2 & 1+1 & 0+1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & -6 \\ -1 & 6 & 2 & 1 \end{bmatrix}. \tag{7}$$

  It has the property of $(\mathbf{X} + \mathbf{Y})^T = \mathbf{X}^T + \mathbf{Y}^T$.

- The product of a number (also called a scalar) and a matrix is referred to as ***scalar multiplication***. Given a scalar c and a matrix $\mathbf{X}$, their scalar multiplication is computed by multiplying every entry of $\mathbf{X}$ by c such that $(c\mathbf{X})_{ij} = c(\mathbf{X})_{ij}$. For instance,

$$2\begin{bmatrix} 1 & 0 & 0 & -7 \\ -2 & 4 & 1 & 0 \end{bmatrix} = \begin{bmatrix} 2\times 1 & 2\times 0 & 2\times 0 & 2\times(-7) \\ 2\times(-2) & 2\times 4 & 2\times 1 & 2\times 0 \end{bmatrix} = \begin{bmatrix} 2 & 0 & 0 & -14 \\ -4 & 8 & 2 & 0 \end{bmatrix}. \quad (8)$$

It has the property of $(c\mathbf{X})^T = c\mathbf{X}^T$.

- The ***multiplication*** operation is defined over two matrices where the number of columns of the left matrix has to be the same as the number of rows of the right matrix. Given an $m \times n$ matrix $\mathbf{X}$ and an $n \times p$ matrix $\mathbf{Y}$, their multiplication is denoted by $\mathbf{XY}$, where

$$(\mathbf{XY})_{ij} = \sum_{k=1}^{n} \mathbf{X}_{ik}\mathbf{Y}_{kj}. \quad (9)$$

An illustration example of calculating the multiplication of a 4×2 matrix $\mathbf{A} = [a_{i,j}]$ and a 2×3 matrix $\mathbf{B} = [b_{i,j}]$ is shown in Figure 1.

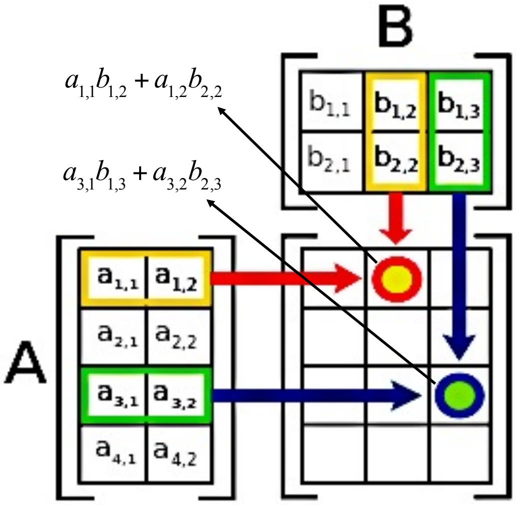

Figure 1: An illustration of calculating matrix multiplication. The figure is adapted from the Wikipedia page on matrix multiplication.

Given matrices $\mathbf{A} \in R^{m\times n}$, $\mathbf{B} \in R^{n\times p}$, $\mathbf{C} \in R^{n\times p}$ and $\mathbf{D} \in R^{p\times q}$, some properties of the matrix multiplication are shown in the following:

$$\begin{aligned} \mathbf{A}(\mathbf{B}+\mathbf{C}) &= \mathbf{AB}+\mathbf{AC}, & (10)\\ (\mathbf{B}+\mathbf{C})\mathbf{D} &= \mathbf{BD}+\mathbf{CD}, & (11)\\ (\mathbf{AB})\mathbf{D} &= \mathbf{A}(\mathbf{BD}), & (12)\\ (\mathbf{AB})^T &= \mathbf{B}^T\mathbf{A}^T. & (13) \end{aligned}$$

- The ***trace*** operation is defined for a square matrix $\mathbf{X} \in R^{n\times n}$, denoted by $\text{tr}(\mathbf{X})$. It is the sum of all the diagonal elements in the matrix, given by

$$\text{tr}(\mathbf{X}) = \sum_{i=1}^{n} \mathbf{X}_{ii}. \tag{14}$$

Given two square matrices $\mathbf{X}$ and $\mathbf{Y}$ of size n, and two matrices $\mathbf{A} \in R^{m\times n}$ and $\mathbf{B} \in R^{n\times m}$ some properties of the trace are shown in the following:

$$\begin{aligned} \text{tr}(\mathbf{X}) &= \text{tr}\left(\mathbf{X}^T\right), & (15)\\ \text{tr}(\mathbf{X}+\mathbf{Y}) &= \text{tr}(\mathbf{X}) + \text{tr}(\mathbf{Y}), & (16)\\ \text{tr}(\mathbf{AB}) &= \text{tr}(\mathbf{BA}). & (17) \end{aligned}$$

- The ***inverse*** of a square matrix $\mathbf{X}$ of size n is denoted by $\mathbf{X}^{-1}$, which is the unique matrix such that

$$\mathbf{X}\mathbf{X}^{-1} = \mathbf{X}^{-1}\mathbf{X} = \mathbf{I}. \tag{18}$$

Non-square matrices do not have inverses by definition. For some square matrices, their inverse may not exist. We say that $\mathbf{X}$ is ***invertible*** (or ***non-singular***) if $\mathbf{X}^{-1}$ exists, and ***non-invertible*** (or ***singular***) otherwise. Given two invertible square matrices $\mathbf{X}$ and $\mathbf{Y}$ of the same size, there are some useful properties of the inverse:

$$\begin{aligned} \left(\mathbf{X}^{-1}\right)^{-1} &= \mathbf{X}, & (19)\\ \left(\mathbf{X}^{-1}\right)^{T} &= \left(\mathbf{X}^{T}\right)^{-1}, & (20)\\ (\mathbf{XY})^{-1} &= \mathbf{Y}^{-1}\mathbf{X}^{-1}. & (21) \end{aligned}$$

- Given two n-dimensional column vectors $\boldsymbol{x}$ and $\boldsymbol{y}$, the quantity $\boldsymbol{x}^T\boldsymbol{y}$ is called the ***inner product*** (or ***dot product***) of the two vectors, which is a real number computed by

$$\boldsymbol{x}^T\boldsymbol{y} = [x_1, x_2, \ldots, x_n]\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \sum_{i=1}^{n} x_i y_i. \tag{22}$$

- A ***norm*** of a vector $\boldsymbol{x}$ is informally a measure of the "length" of the vector, and is usually denoted by $\|\boldsymbol{x}\|$. Assuming $\boldsymbol{x}$ is an n-dimensional column vector, the commonly used ***Euclidean norm*** (or called l_2***-norm***) is given by

$$\|\boldsymbol{x}\|_2 = \sqrt{\sum_{i=1}^{n} x_i^2} = \sqrt{\boldsymbol{x}^T\boldsymbol{x}}. \tag{23}$$

Another example of the norm is the l_1***-norm***, given by

$$\|\boldsymbol{x}\|_1 = \sum_{i=1}^{n} |x_i|. \tag{24}$$

- A norm can also be defined for a matrix. For example, the ***Frobenius norm*** of an $m \times n$ matrix $\mathbf{X}$ is given by

$$\|\mathbf{X}\|_F = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n} \mathbf{X}_{ij}^2} = \sqrt{\text{tr}\left(\mathbf{X}^T\mathbf{X}\right)} = \sqrt{\text{tr}\left(\mathbf{X}\mathbf{X}^T\right)}. \tag{25}$$

1.3 Symmetric Matrices

Given a square matrix $\mathbf{X} \in R^{n\times n}$, it is ***symmetric*** if $\mathbf{X} = \mathbf{X}^T$. For instance, the following 4×4 matrix is symmetric:

$$\begin{bmatrix} 1 & 0 & 0 & -7 \\ 0 & 4 & 3 & 0 \\ 0 & 3 & 2 & 1 \\ -7 & 0 & 1 & -1.6 \end{bmatrix}. \tag{26}$$

Given an arbitrary square matrix $\mathbf{X} \in R^{n\times n}$, the matrix $\mathbf{X} + \mathbf{X}^T$ is symmetric.

2. Calculus Basics

2.1 Derivative and Differentiation Rules

Given a function of a real variable $f(x) : R \to R$, its ***derivative*** $f'(x)$ (or $\frac{df}{dx}$ in Leibniz's notation) measures the rate at which the function value changes with respect to the change of the input variable x, where

$$f'(x) = \frac{df}{dx} = \lim_{\Delta x \to 0} \frac{f(x + \Delta x) - f(x)}{\Delta x}. \tag{27}$$

This gives the trivial case that the derivative of a constant function is zero. The tangent line to the graph of a function $f(x)$ at a chosen input value is the straight line that "just touches" the function curve at that point. The slope of the tangent line is equal to the derivative of the function at the chosen value (see Figure 2 for example).

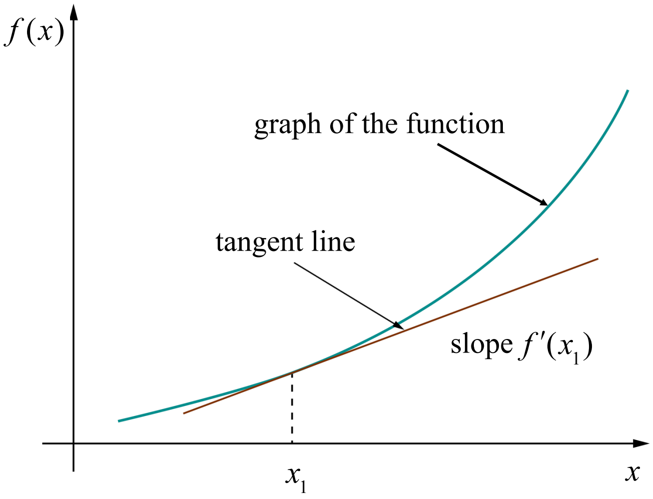

Figure 2: Geometric illustration of the derivative of a single-variable function.

The process of finding a derivative is called differentiation. Here is a summary of rules for computing the derivative of a function in calculus, referred to as ***differentiation rules***.

- Linearity: For any functions $f(x)$ and $g(x)$ and any real numbers a and b, the derivative of the function $h(x) = af(x) + bg(x)$ with respect to x is

$$h'(x) = af'(x) + bg'(x). \tag{28}$$

  Its special cases include the constant factor rule $(af)' = af'$, the sum rule $(f + g)' = f' + g'$, and the subtraction rule $(f - g)' = f' - g'$.

- Product rule: For any functions $f(x)$ and $g(x)$, the derivative of the function $h(x) = f(x)g(x)$ with respect to x is

$$h'(x) = f'(x)g(x) + f(x)g'(x). \tag{29}$$

- Chain rule: For any functions $f(x)$ and $g(x)$, the derivative of the function $h(x) = f(g(x))$ with respect to x is

$$h'(x) = f'(g(x))g'(x). \tag{30}$$

- Inverse function rule: If the function $f(x)$ has an inverse function $g(x)$, which means that $g(f(x)) = x$ and $f(g(y)) = y$, the derivative of $g(x)$ with respect to x is

$$g'(x) = \frac{1}{f'(g(x))}. \tag{31}$$

- Quotient rule: For any function $g(x) \neq 0$ and for any function $f(x)$, the derivative of the function $h(x) = \frac{f(x)}{g(x)}$ with respect to x is

$$h'(x) = \frac{f'(x)g(x) - f(x)g'(x)}{(g(x))^2}. \tag{32}$$

  Its special case is the reciprocal rule, where the derivative of the function $g(x) = \frac{1}{f(x)}$ with respect to x is $g'(x) = -\frac{f'(x)}{(f(x))^2}$.

Utilising the differentiation rules, most derivative computations can eventually be based on the computation of ***derivatives of some common functions***. Table 1 provides an incomplete list showing some frequently used single-variable functions and their derivatives.

Functions	Derivatives	Functions	Derivatives
x^r	rx^{r-1}	e^x	e^x
$\ln(x)$	$\frac{1}{x}$	a^x	$a^x \ln(a)$
$\sin(x)$	$\cos(x)$	$\cos(x)$	$-\sin(x)$

Table 1: Some frequently used single-variable functions and their derivatives.

2.2 Partial Derivative and Gradient

Given a function of multiple real variables $f(x_1, x_2, \ldots, x_n)$, its ***partial derivative*** f'_{x_i} (or denoted by $\frac{\partial f}{\partial x_i}$), where $i = 1, 2, \ldots, n$, is its derivative with respect to one of those variables, with the others held constant. For instance, given a function $f(x, y, z) = x^2 + 3xy + z + 1$, we have $\frac{\partial f}{\partial x} = 2x + 3y$, $\frac{\partial f}{\partial y} = 3x$ and $\frac{\partial f}{\partial z} = 1$.

The ***gradient*** is a multi-variable generalisation of the derivative, which is defined on a function of multiple variables $f(x_1, x_2, \ldots, x_n)$. The multi-variable function can be viewed as a function $f(\boldsymbol{x}) : R^n \to R$ taking the vector $\boldsymbol{x} = [x_1, x_2, \ldots, x_n]$ as the input. Its gradient is denoted by $\triangledown_{\boldsymbol{x}} f$ and is defined from the partial derivatives:

$$\triangledown_{\boldsymbol{x}} f = \left[\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, \ldots, \frac{\partial f}{\partial x_n}\right]. \tag{33}$$

It can be seen that a derivative is a scalar-valued function, while a gradient is a vector-valued function. For instance, the gradient of the function $f(x, y, z) = x^2 + 3xy + z + 1$ is $[2x + 3y, 3x, 1]$.

If a function $f(\mathbf{X}) : R^{m \times n} \to R$ takes an $m \times n$ matrix $\mathbf{X} = [x_{ij}]$ as the input. The gradient of f with respect to the matrix $\mathbf{X}$ is defined as the matrix of partial derivatives, given as

$$\triangledown_{\mathbf{X}} f = \begin{bmatrix} \frac{\partial f}{\partial x_{11}} & \frac{\partial f}{\partial x_{12}} & \cdots & \frac{\partial f}{\partial x_{1n}} \\ \frac{\partial f}{\partial x_{21}} & \frac{\partial f}{\partial x_{22}} & \cdots & \frac{\partial f}{\partial x_{2n}} \\ \frac{\partial f}{\partial x_{31}} & \frac{\partial f}{\partial x_{32}} & \cdots & \frac{\partial f}{\partial x_{3n}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f}{\partial x_{m1}} & \frac{\partial f}{\partial x_{m2}} & \cdots & \frac{\partial f}{\partial x_{mn}} \end{bmatrix}. \tag{34}$$

3. Linear and Quadratic Functions

3.1 Linear Function

Let $\boldsymbol{w} \in R^n$ denote a known n-dimensional vector. For an input column vector $\boldsymbol{x} \in R^n$, the following function

$$f(\boldsymbol{x}) = \sum_{i=1}^{n} w_i x_i = \boldsymbol{w}^T \boldsymbol{x} \tag{35}$$

is a ***linear function*** of $\boldsymbol{x}$. The partial derivative of this function is

$$\frac{\partial f(\boldsymbol{x})}{\partial x_i} = \frac{\partial}{\partial x_i}\left(\sum_{i=1}^{n} w_i x_i\right) = w_i, \text{ for } i = 1, 2, \ldots n. \tag{36}$$

The gradient of $f(\boldsymbol{x})$ with respect to the input column vector $\boldsymbol{x}$ is

$$\triangledown_{\boldsymbol{x}} f(\boldsymbol{x}) = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \\ \vdots \\ \frac{\partial f}{\partial x_n} \end{bmatrix} = \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_n \end{bmatrix} = \boldsymbol{w}. \tag{37}$$

Note that the function $f(\boldsymbol{x})$ can also be written as $f(\boldsymbol{x}) = \boldsymbol{x}^T \boldsymbol{w}$, and its gradient with respect to $\boldsymbol{x}$ is $\boldsymbol{w}$.

3.2 Quadratic Function

Let $\mathbf{A} = [a_{ij}]$ denote an $n \times n$ square matrix. For an input column vector $\boldsymbol{x} \in R^n$, the following function

$$f(\boldsymbol{x}) = \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} x_i x_j = \boldsymbol{x}^T \mathbf{A} \boldsymbol{x} \tag{38}$$

is a ***quadratic function*** of $\boldsymbol{x}$. To compute the partial derivative of this function with respect to an element x_k in the input vector $(k = 1, 2, \dots n)$, we consider separately the terms that contain x_k and x_k^2, also the terms that do not contain x_k. This gives

$$\begin{aligned}
\frac{\partial f(\boldsymbol{x})}{\partial x_k} &= \frac{\partial}{\partial x_k} \left(\sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} x_i x_j \right) \\
&= \frac{\partial}{\partial x_k} \left(a_{kk} x_k^2 + \sum_{i \neq k} a_{ik} x_i x_k + \sum_{j \neq k} a_{kj} x_k x_j + \sum_{i \neq k} \sum_{j \neq k} a_{ij} x_i x_j \right) \\
&= 2 a_{kk} x_k + \sum_{i \neq k} a_{ik} x_i + \sum_{j \neq k} a_{kj} x_j \\
&= \sum_{i=1}^{n} a_{ik} x_i + \sum_{j=1}^{n} a_{kj} x_j && (39) \\
&= \mathbf{A}_{:,k}^T \boldsymbol{x} + \mathbf{A}_{k,:} \boldsymbol{x}. && (40)
\end{aligned}$$

The gradient of $f(\boldsymbol{x})$ with respect to $\boldsymbol{x}$ is

$$\nabla_{\boldsymbol{x}} f(\boldsymbol{x}) = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \\ \vdots \\ \frac{\partial f}{\partial x_n} \end{bmatrix} = \begin{bmatrix} \mathbf{A}_{:,1}^T \boldsymbol{x} + \mathbf{A}_{1,:} \boldsymbol{x} \\ \mathbf{A}_{:,2}^T \boldsymbol{x} + \mathbf{A}_{2,:} \boldsymbol{x} \\ \vdots \\ \mathbf{A}_{:,n}^T \boldsymbol{x} + \mathbf{A}_{n,:} \boldsymbol{x} \end{bmatrix} = \begin{bmatrix} \mathbf{A}_{:,1}^T \boldsymbol{x} \\ \mathbf{A}_{:,2}^T \boldsymbol{x} \\ \vdots \\ \mathbf{A}_{:,n}^T \boldsymbol{x} \end{bmatrix} + \begin{bmatrix} \mathbf{A}_{1,:} \boldsymbol{x} \\ \mathbf{A}_{2,:} \boldsymbol{x} \\ \vdots \\ \mathbf{A}_{n,:} \boldsymbol{x} \end{bmatrix} = \mathbf{A}^T \boldsymbol{x} + \mathbf{A} \boldsymbol{x}. \tag{41}$$

A special case of the quadratic function is $f(\boldsymbol{x}) = \boldsymbol{x}^T \boldsymbol{x}$, where $\mathbf{A}$ is an identity matrix. Its gradient with respect to $\boldsymbol{x}$ is therefore $\nabla_{\boldsymbol{x}} f(\boldsymbol{x}) = \mathbf{I}^T \boldsymbol{x} + \mathbf{I} \boldsymbol{x} = 2\boldsymbol{x}$.