2 Convexity, Derivative

2.1 Lines, hyperplanes and linear varieties

- The line segment between two points $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$ is the set,

$$\{\boldsymbol{z} \in \mathbb{R}^n : \boldsymbol{z} = \alpha \boldsymbol{x} + (1-\alpha)\boldsymbol{y}, \alpha \in [0,1]\}.$$

- A hyperplane of the space $\mathbb{R}^n$, is the set of all points $\boldsymbol{x} = [x_1, x_2, \ldots, x_n]^\top$ that satisfy the linear equation

$$u_1 x_1 + u_2 x_2 + \cdots + u_n x_n = v,$$

where at least one of the u_i is nonzero. The hyperplane is defined by

$$\left\{\boldsymbol{x} \in \mathbb{R}^n : \boldsymbol{u}^\top \boldsymbol{x} = v\right\},$$

where

$$\boldsymbol{u} = [u_1, u_2, \ldots, u_n]^\top .$$

- Two half-spaces, postive half-space and negative half-space are

$$H_+ = \left\{\boldsymbol{x} \in \mathbb{R}^n : \boldsymbol{u}^\top \boldsymbol{x} \geq v\right\},$$

$$H_- = \left\{\boldsymbol{x} \in \mathbb{R}^n : \boldsymbol{u}^\top \boldsymbol{x} \leq v\right\}.$$

- A linear variety is a set of form

$$\{\boldsymbol{x} \in \mathbb{R}^n : \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}\},$$

for some matrix $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ and vector $\boldsymbol{b} \in \mathbb{R}^n$.

2.2 Convex sets

- A point $\boldsymbol{w} = \alpha \boldsymbol{u} + (1-\alpha)\boldsymbol{v}$ (where $\alpha \in [0,1]$) is called a convex combination of the points $\boldsymbol{u}$ and $\boldsymbol{u}$.

- A set $\Theta \subset \mathbb{R}^n$ is convex if for all $\boldsymbol{u}, \boldsymbol{v} \in \Theta$, the *line segment* between $\boldsymbol{u}$ and $\boldsymbol{v}$ is in Θ.

  That is, Θ is *convex* if and only if $\alpha \boldsymbol{u} + (1-\alpha)\boldsymbol{v} \in \Theta$ for all $\boldsymbol{u}, \boldsymbol{v} \in \Theta$ and $\alpha \in (0,1)$. Examples of convex sets include the following:

  - The empty set
  - A set consisting of a single point
  - A line or a line segment
  - A subspace
  - A hyperplane
  - A linear variety
  - A half-space
  - $\mathbb{R}^n$

♠ $\mathcal{THEOREM}$4.3 Convex subsets of $\mathbb{R}^n$ have the following properties:

a. If Θ is a *convex set* and β is a real number, then the set

$$\beta\Theta = \{\boldsymbol{x} : \boldsymbol{x} = \beta\boldsymbol{v}, \boldsymbol{v} \in \Theta\}$$

is also convex.

b. If Θ_1 and Θ_2 are *convex sets*, then the set

$$\Theta_1 + \Theta_2 = \{\boldsymbol{x} : \boldsymbol{x} = \boldsymbol{v}_1 + \boldsymbol{v}_2, \boldsymbol{v}_1 \in \Theta_1, \boldsymbol{v}_2 \in \Theta_2\}$$

is also convex.

c. The intersection of any collection of *convex sets* is convex.

- An extreme point $\boldsymbol{x}$ in a *convex set* Θ, if there are no two distinct points $\boldsymbol{u}$ and $\boldsymbol{v}$ in Θ such that $\boldsymbol{x} = \alpha\boldsymbol{u} + (1-\alpha)\boldsymbol{v}$ for some $\alpha \in (0, 1)$.

2.3 Differentiation rules

- A function $f : \mathbb{R}^n \to \mathbb{R}$ follows,

$$f(\boldsymbol{x}) = f\left(\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}\right) = a_1x_1 + a_2x_2 + \cdots + a_nx_n = \begin{bmatrix} a_1 & \cdots & a_n \end{bmatrix}\begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}$$

$$= \begin{bmatrix} & \boldsymbol{a}^\top & \end{bmatrix}\begin{bmatrix} \vdots \\ \boldsymbol{x} \\ \vdots \end{bmatrix} = \begin{bmatrix} & \boldsymbol{x}^\top & \end{bmatrix}\begin{bmatrix} \vdots \\ \boldsymbol{a} \\ \vdots \end{bmatrix}.$$

- A matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$, $\mathbf{A}$ is $m \times n$ matrix,

$$\mathbf{A} = \begin{bmatrix} \vdots & \vdots & & \vdots \\ \boldsymbol{a}_{*1} & \boldsymbol{a}_{*2} & \cdots & \boldsymbol{a}_{*n} \\ \vdots & \vdots & & \vdots \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{12} & \cdots & a_{mn} \end{bmatrix} = \begin{bmatrix} \boldsymbol{a}_1^\top \\ \boldsymbol{a}_2^\top \\ \vdots \\ \boldsymbol{a}_m^\top \end{bmatrix}.$$

- A function $\boldsymbol{g} : \mathbb{R}^n \to \mathbb{R}^m$ and a matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$, $\mathbf{A}\boldsymbol{x}$ is a column vector whose element is a scalar $g_\star(\boldsymbol{x})$.

$$\mathbf{A}\boldsymbol{x} = \begin{bmatrix} \boldsymbol{a}_1^\top \\ \vdots \\ \boldsymbol{a}_m^\top \end{bmatrix}\boldsymbol{x} = \begin{bmatrix} \boldsymbol{a}_1^\top\boldsymbol{x} \\ \vdots \\ \boldsymbol{a}_m^\top\boldsymbol{x} \end{bmatrix} = \begin{bmatrix} \boxed{a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n} \\ \vdots \\ \boxed{a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n} \end{bmatrix} = \begin{bmatrix} g_1(\boldsymbol{x}) \\ \vdots \\ g_m(\boldsymbol{x}) \end{bmatrix} = \boldsymbol{g}(\boldsymbol{x}).$$

- To be noted, in this course, we write the **derivative** $Df(\boldsymbol{x})$ as a row vector, and write the **gradient** $\nabla f(\boldsymbol{x})$ as a column vector.

Types of Matrix Derivatives[1]

Types	Scalar	Vector	Matrix
Scalar	(1) $\frac{\mathrm{d}y}{\mathrm{d}x}$ $\quad \frac{\mathrm{d}f(x)}{\mathrm{d}x}$	$\frac{\mathrm{d}\boldsymbol{y}}{\mathrm{d}x} = \left[\frac{\partial y_i}{\partial x}\right]$ $\quad \frac{\mathrm{d}\boldsymbol{g}(t)}{\mathrm{d}t}$ (3)	$\frac{\mathrm{d}\mathbf{Y}}{\mathrm{d}x} = \left[\frac{\partial y_{ij}}{\partial x}\right], \frac{\mathrm{d}\boldsymbol{A}(t)}{\mathrm{d}t}$
Vector	$\frac{\mathrm{d}y}{\mathrm{d}\boldsymbol{x}} = \left[\frac{\partial y}{\partial x_j}\right]$ (2) $\quad D_{\boldsymbol{x}} f(\boldsymbol{x}) = \left[\cdot\cdot \frac{\partial f(\boldsymbol{x})}{\partial x_j} \cdot\cdot\right]$ $\quad \nabla_{\boldsymbol{x}} f(\boldsymbol{x}) = \begin{bmatrix} \cdot \\ \frac{\partial f(\boldsymbol{x})}{\partial x_j} \\ \cdot \end{bmatrix}$	$\frac{\mathrm{d}\boldsymbol{y}}{\mathrm{d}\boldsymbol{x}} = \left[\frac{\partial y_i}{\partial x_j}\right]$ $\quad D\boldsymbol{g}(\boldsymbol{x}) = \left[\frac{\partial g_i(\boldsymbol{x})}{\partial x_j}\right]$ (4)	
Matrix	$\frac{\mathrm{d}y}{\mathrm{d}\mathbf{X}} = \left[\frac{\partial y}{\partial x_{ji}}\right]$ $\quad D_{\mathbf{X}} f = \left[\frac{\partial f}{\partial x_{ji}}\right]$		

(1) Given $f : \mathbb{R} \to \mathbb{R}$, if the limit exists, the derivative of f is a function $f' : \mathbb{R} \to \mathbb{R}$ given by

$$D_x(f(x)) = \frac{\mathrm{d}f}{\mathrm{d}x} = f'(x) = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h}.$$

(2) Given $f : \mathbb{R}^n \to \mathbb{R}$, consider a scalar $f(\boldsymbol{x}) = a_1 x_1 + a_2 x_2 + \cdots + a_n x_n = \boldsymbol{a}^\top \boldsymbol{x}$.

For **derivative** rule (2),

$$D_{\boldsymbol{x}} f(\boldsymbol{x}) = D_{\boldsymbol{x}}(\boldsymbol{a}^\top \boldsymbol{x}) = \left[\boxed{\frac{\partial f}{\partial x_1}(\boldsymbol{x}) \; \frac{\partial f}{\partial x_2}(\boldsymbol{x}) \; \cdots \; \frac{\partial f}{\partial x_n}(\boldsymbol{x})} \right] = \begin{bmatrix} a_1 & \cdots & a_n \end{bmatrix} = \boldsymbol{a}^\top.$$

For **gradient** rule (2), if $f : \mathbb{R}^n \to \mathbb{R}$ is differentiable, then the *gradient* of f is a function $\nabla f : \mathbb{R}^n \to \mathbb{R}^n$ given by

$$\nabla_{\boldsymbol{x}} f(\boldsymbol{x}) = \left[\boxed{\begin{matrix} \frac{\partial f}{\partial x_1}(\boldsymbol{x}) \\ \vdots \\ \frac{\partial f}{\partial x_n}(\boldsymbol{x}) \end{matrix}} \right] = \begin{bmatrix} a_1 \\ \vdots \\ a_n \end{bmatrix} = \boldsymbol{a} = D_{\boldsymbol{x}}(f(\boldsymbol{x}))^\top.$$

(3) Given $\boldsymbol{g} : \mathbb{R} \to \mathbb{R}^m$, here $t \in \mathbb{R}$ is a scalar. $\boldsymbol{g}(t)$ is a column vector.

$$\boldsymbol{g}(t) = \begin{bmatrix} g_1(t) \\ g_2(t) \\ \vdots \\ g_m(t) \end{bmatrix}, \quad D_t \boldsymbol{g}(t) = \begin{bmatrix} \frac{\mathrm{d}}{\mathrm{d}t} g_1(t) \\ \vdots \\ \frac{\mathrm{d}}{\mathrm{d}t} g_m(t) \end{bmatrix} = \begin{bmatrix} g_1'(t) \\ \vdots \\ g_m'(t) \end{bmatrix},$$

(4) Consider $\boldsymbol{g} : \mathbb{R}^n \to \mathbb{R}^m$, here $\boldsymbol{x} \in \mathbb{R}^n$ is a vector. Since $g_i(\boldsymbol{x})$ is a scalar, $\boldsymbol{g} = [g_1, \ldots, g_m]^\top$, $\boldsymbol{g}(\boldsymbol{x})$ is a column vector.

$$\boldsymbol{g}(\boldsymbol{x}) = \begin{bmatrix} g_1(\boldsymbol{x}) \\ g_2(\boldsymbol{x}) \\ \vdots \\ g_m(\boldsymbol{x}) \end{bmatrix}, D_{\boldsymbol{x}} \boldsymbol{g}(\boldsymbol{x}) = \begin{bmatrix} D_{\boldsymbol{x}} g_1(x_1, x_2, \cdots, x_n) \\ D_{\boldsymbol{x}} g_2(x_1, x_2, \cdots, x_n) \\ \vdots \\ D_{\boldsymbol{x}} g_m(x_1, x_2, \cdots, x_n) \end{bmatrix} = \begin{bmatrix} \boxed{\frac{\partial}{\partial x_1} g_1 \; \frac{\partial}{\partial x_2} g_1 \; \cdots \; \frac{\partial}{\partial x_n} g_1} \\ \boxed{\frac{\partial}{\partial x_1} g_2 \; \frac{\partial}{\partial x_2} g_2 \; \cdots \; \frac{\partial}{\partial x_n} g_2} \\ \vdots \quad \vdots \quad \ddots \quad \vdots \\ \boxed{\frac{\partial}{\partial x_1} g_m \; \frac{\partial}{\partial x_2} g_m \; \cdots \; \frac{\partial}{\partial x_n} g_m} \end{bmatrix} = \mathbf{J}.$$

The matrix $\mathbf{J}$ is called the Jacobian matrix, or derivative matrix, of function $\boldsymbol{g}$.

[1] Ref: Thomas P. Minka, "Old and New Matrix Algebra Useful for Statistics", 2000

- If all elements in $\boldsymbol{g}(\boldsymbol{x})$ are linear combination of $\boldsymbol{x}$,

$$\boldsymbol{g}(\boldsymbol{x}) = \begin{bmatrix} g_1(\boldsymbol{x}) \\ \vdots \\ g_m(\boldsymbol{x}) \end{bmatrix} = \begin{bmatrix} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n \\ \vdots \\ a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n \end{bmatrix} = \begin{bmatrix} \boldsymbol{a}_1^\top \boldsymbol{x} \\ \vdots \\ \boldsymbol{a}_m^\top \boldsymbol{x} \end{bmatrix} = \begin{bmatrix} \boldsymbol{a}_1^\top \\ \vdots \\ \boldsymbol{a}_m^\top \end{bmatrix} \boldsymbol{x} = \mathbf{A}\boldsymbol{x}.$$

Then, the derivative of $\mathbf{A}\boldsymbol{x}$ is equivalent to $D_{\boldsymbol{x}}\boldsymbol{g}(\boldsymbol{x})$,

$$D_{\boldsymbol{x}}(\boldsymbol{g}(\boldsymbol{x})) = \underbrace{\frac{\mathrm{d}}{\mathrm{d}\boldsymbol{x}}(\mathbf{A}\boldsymbol{x})}_{\substack{\text{Notation not used} \\ \text{in this course}}} = D_{\boldsymbol{x}}(\mathbf{A}\boldsymbol{x}) = \begin{bmatrix} D_{\boldsymbol{x}}(\boldsymbol{a}_1^\top \boldsymbol{x}) \\ D_{\boldsymbol{x}}(\boldsymbol{a}_2^\top \boldsymbol{x}) \\ \vdots \\ D_{\boldsymbol{x}}(\boldsymbol{a}_m^\top \boldsymbol{x}) \end{bmatrix} = \begin{bmatrix} \boldsymbol{a}_1^\top \\ \boldsymbol{a}_2^\top \\ \vdots \\ \boldsymbol{a}_m^\top \end{bmatrix} = \mathbf{A}.$$

- In summary, the derivative rules are listed as,

$$D(\boldsymbol{a}^\top \boldsymbol{x}) = \boldsymbol{a}^\top, \qquad \boxed{(2) f : \mathbb{R}^n \to \mathbb{R}, f(\boldsymbol{x}) = \boldsymbol{a}^\top \boldsymbol{x}}$$

$$D(\boldsymbol{g}(t)) = \begin{bmatrix} \vdots \\ g_*'(t) \\ \vdots \end{bmatrix}, \qquad \boxed{(3) \boldsymbol{g} : \mathbb{R} \to \mathbb{R}^n, \boldsymbol{g}(t) = \begin{bmatrix} \vdots \\ g_*(t) \\ \vdots \end{bmatrix}}$$

$$D(\mathbf{A}\boldsymbol{x}) = \mathbf{A}, \qquad \boxed{(4) \boldsymbol{g} : \mathbb{R}^n \to \mathbb{R}^m, \boldsymbol{g}(\boldsymbol{x}) = \mathbf{A}\boldsymbol{x}}$$

$$D(\mathbf{A}(\alpha\boldsymbol{x})) = \alpha\mathbf{A},$$

$$\frac{\mathrm{d}}{\mathrm{d}\alpha}(\mathbf{A}(\alpha\boldsymbol{x})) = \mathbf{A}\boldsymbol{x},$$

$$\nabla \boldsymbol{a}^\top \boldsymbol{x} = \boldsymbol{a}, \qquad \boxed{(2) f : \mathbb{R}^n \to \mathbb{R}, f(\boldsymbol{x}) = \boldsymbol{a}^\top \boldsymbol{x}}$$

$$\nabla \mathbf{A}\boldsymbol{x} = \mathbf{A}^\top, \qquad \boxed{(4) \boldsymbol{g} : \mathbb{R}^n \to \mathbb{R}^m, \boldsymbol{g}(\boldsymbol{x}) = \mathbf{A}\boldsymbol{x}}$$

$$\nabla \mathbf{A}(\alpha\boldsymbol{x}) = \alpha\mathbf{A}^\top.$$

- Note that for $\underline{f : \mathbb{R}^n \to \mathbb{R}}$, we have

$$\nabla f(\boldsymbol{x}) = D f(\boldsymbol{x})^\top.$$

2.4 Differentiation rules on composite function

- To differentiate the composite function, $h(t) = f(\boldsymbol{g}(t))$ is differentiable on (a, b), and

$$f(\boldsymbol{g}(t)) = f\left(\begin{bmatrix} g_1(t) \\ g_2(t) \\ \vdots \\ g_m(t) \end{bmatrix}\right) = a_1 g_1(t) + a_2 g_2(t) + \cdots + a_m g_m(t).$$

- The differentiated composite function with **derivative** rule is

$$h'(t) = D_{\boldsymbol{g}} f(\boldsymbol{g}(t)) D_t \boldsymbol{g}(t) = \nabla f(\boldsymbol{g}(t))^\top \begin{bmatrix} g_1'(t) \\ \vdots \\ g_m'(t) \end{bmatrix} = \begin{bmatrix} \frac{\mathrm{d}}{\mathrm{d}g_1} f(\boldsymbol{g}(t)) & \cdots & \frac{\mathrm{d}}{\mathrm{d}g_m} f(\boldsymbol{g}(t)) \end{bmatrix} \begin{bmatrix} g_1'(t) \\ \vdots \\ g_m'(t) \end{bmatrix}.$$

- Consider Hessian matrix, which is second order derivative of scalar. Noted that, $D\left(f(\boldsymbol{x})\right)$ is spreading the derivative of the polynomials on the horizontal direction. Thus, we would like to ensure each entry is located on a vertical direction, then the entry could be applied to conduct derivative, as

$$D^2\left(f(\boldsymbol{x})\right) = D\left(Df\left(\boldsymbol{x}\right)^\top\right) = D(\nabla f(\boldsymbol{x})) = \begin{bmatrix} D\left(\frac{\partial f}{\partial x_1}\right) \\ D\left(\frac{\partial f}{\partial x_2}\right) \\ \vdots \\ D\left(\frac{\partial f}{\partial x_n}\right) \end{bmatrix} = \begin{bmatrix} \boxed{\frac{\partial^2 f}{\partial x_1 \partial x_1} \quad \frac{\partial^2 f}{\partial x_2 \partial x_1} \quad \cdots \quad \frac{\partial^2 f}{\partial x_n \partial x_1}} \\ \boxed{\frac{\partial^2 f}{\partial x_1 \partial x_2} \quad \frac{\partial^2 f}{\partial x_2 \partial x_2} \quad \cdots \quad \frac{\partial^2 f}{\partial x_n \partial x_2}} \\ \vdots \qquad \vdots \qquad \ddots \qquad \vdots \\ \boxed{\frac{\partial^2 f}{\partial x_1 \partial x_n} \quad \frac{\partial^2 f}{\partial x_2 \partial x_n} \quad \cdots \quad \frac{\partial^2 f}{\partial x_n \partial x_n}} \end{bmatrix}.$$

2.5 Differentiation product rules

i) Let $f : \mathbb{R} \to \mathbb{R}$ and $g : \mathbb{R} \to \mathbb{R}$ be two differentiable functions, $x \in \mathbb{R}$,

$$D\Big(f(x)g(x)\Big) = f(x)Dg(x) + g(x)Df(x),$$
$$\nabla\Big(f(x)g(x)\Big) = f(x)\nabla g(x) + g(x)\nabla f(x).$$

ii) Let $f : \mathbb{R}^n \to \mathbb{R}$ and $g : \mathbb{R}^n \to \mathbb{R}$ be two differentiable functions, $\boldsymbol{x} \in \mathbb{R}^n$,

$$D\Big(f(\boldsymbol{x})g(\boldsymbol{x})\Big) = f(\boldsymbol{x})\begin{bmatrix} \quad Dg(\boldsymbol{x}) \quad \end{bmatrix} + g(\boldsymbol{x})\begin{bmatrix} \quad Df(\boldsymbol{x}) \quad \end{bmatrix},$$
$$\nabla\Big(f(\boldsymbol{x})g(\boldsymbol{x})\Big) = f(\boldsymbol{x})\begin{bmatrix} \\ \nabla g(\boldsymbol{x}) \\ \\ \end{bmatrix} + g(\boldsymbol{x})\begin{bmatrix} \\ \nabla f(\boldsymbol{x}) \\ \\ \end{bmatrix}.$$

iii) Let $\boldsymbol{f} : \mathbb{R}^n \to \mathbb{R}^m$ and $\boldsymbol{g} : \mathbb{R}^n \to \mathbb{R}^m$ be two differentiable functions, $\boldsymbol{x} \in \mathbb{R}^n$,

$$D\Big(\boldsymbol{f}(\boldsymbol{x})^\top \boldsymbol{g}(\boldsymbol{x})\Big) = \boldsymbol{f}(\boldsymbol{x})^\top D\boldsymbol{g}(\boldsymbol{x}) + \boldsymbol{g}(\boldsymbol{x})^\top D\boldsymbol{f}(\boldsymbol{x}),$$
$$\nabla\Big(\boldsymbol{f}(\boldsymbol{x})^\top \boldsymbol{g}(\boldsymbol{x})\Big) = \boldsymbol{f}(\boldsymbol{x})^\top \nabla\boldsymbol{g}(\boldsymbol{x}) + \boldsymbol{g}(\boldsymbol{x})^\top \nabla\boldsymbol{f}(\boldsymbol{x}).$$

- Based on the above **derivative** rule, we have

  1. Consider $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ be a given matrix and $\boldsymbol{y} \in \mathbb{R}^m$ a given vector. Then,

$$D\left(\boldsymbol{y}^\top \boldsymbol{A}\boldsymbol{x}\right) = \boldsymbol{y}^\top \boldsymbol{A},$$
$$D\left(\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x}\right) = \boldsymbol{x}^\top\left(\boldsymbol{A} + \boldsymbol{A}^\top\right). \qquad \boxed{\text{if } m = n}$$

  2. Consider $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ be a given matrix and $\boldsymbol{y} \in \mathbb{R}^n$ a given vector. Then,

$$D\left(\boldsymbol{y}^\top \boldsymbol{x}\right) = \boldsymbol{y}^\top.$$

  3. Consider if $\boldsymbol{Q}$ is a symmetric matrix, then

$$D\left(\boldsymbol{x}^\top \boldsymbol{Q}\boldsymbol{x}\right) = 2\boldsymbol{x}^\top \boldsymbol{Q}.$$

  In particular,

$$D\left(\boldsymbol{x}^\top \boldsymbol{x}\right) = 2\boldsymbol{x}^\top.$$

- Based on the above **gradient** rule, we have

1. Consider $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ be a given matrix and $\boldsymbol{y} \in \mathbb{R}^m$ a given vector. Then,

$$\nabla\left(\boldsymbol{y}^\top \boldsymbol{A}\boldsymbol{x}\right) = \boldsymbol{A}^\top \boldsymbol{y},$$
$$\nabla\left(\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x}\right) = \left(\boldsymbol{A} + \boldsymbol{A}^\top\right)\boldsymbol{x}. \qquad \boxed{\text{if } m = n}$$

2. Consider $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ be a given matrix and $\boldsymbol{y} \in \mathbb{R}^n$ a given vector. Then,

$$\nabla\left(\boldsymbol{y}^\top \boldsymbol{x}\right) = \boldsymbol{y}.$$

3. Consider if $\boldsymbol{Q}$ is a symmetric matrix, then

$$\nabla\left(\boldsymbol{x}^\top \boldsymbol{Q}\boldsymbol{x}\right) = 2\boldsymbol{Q}\boldsymbol{x}.$$

In particular,

$$\nabla\left(\boldsymbol{x}^\top \boldsymbol{x}\right) = 2\boldsymbol{x}.$$

2.6 Derivation details

- Let $\boldsymbol{f} : \mathbb{R}^n \to \mathbb{R}^m$ and $\boldsymbol{g} : \mathbb{R}^n \to \mathbb{R}^m$ be two differentiable functions, $\boldsymbol{x} \in \mathbb{R}^n$,

$$D\Big(\boldsymbol{f}(\boldsymbol{x})^\top \boldsymbol{g}(\boldsymbol{x})\Big) = \boldsymbol{f}(\boldsymbol{x})^\top D\boldsymbol{g}(\boldsymbol{x}) + \boldsymbol{g}(\boldsymbol{x})^\top D\boldsymbol{f}(\boldsymbol{x}).$$

Proof. $\boldsymbol{f} : \mathbb{R}^n \to \mathbb{R}^m$ and $\boldsymbol{g} : \mathbb{R}^n \to \mathbb{R}^m$, $\boldsymbol{x} \in \mathbb{R}^n$, write the derivative as

$$\begin{aligned}
D\Big(\boldsymbol{f}(\boldsymbol{x})^\top \boldsymbol{g}(\boldsymbol{x})\Big) &= D\left(\begin{bmatrix} f_1(\boldsymbol{x}) & \cdots & f_m(\boldsymbol{x}) \end{bmatrix} \begin{bmatrix} g_1(\boldsymbol{x}) \\ \vdots \\ g_m(\boldsymbol{x}) \end{bmatrix}\right) \\
&= D\left(\begin{bmatrix} f_1(x_1, x_2, \ldots, x_n) & \cdots & f_m(x_1, x_2, \ldots, x_n) \end{bmatrix} \begin{bmatrix} g_1(x_1, x_2, \ldots, x_n) \\ \vdots \\ g_m(x_1, x_2, \ldots, x_n) \end{bmatrix}\right) \\
&= D\Big(f_1(\boldsymbol{x})g_1(\boldsymbol{x}) + \cdots + f_m(\boldsymbol{x})g_m(\boldsymbol{x})\Big) \\
&= f_1(\boldsymbol{x})Dg_1(\boldsymbol{x}) + \cdots + f_m(\boldsymbol{x})Dg_m(\boldsymbol{x}) \\
&\qquad\qquad + g_1(\boldsymbol{x})Df_1(\boldsymbol{x}) + \cdots + g_m(\boldsymbol{x})Df_m(\boldsymbol{x}) \\
&= \begin{bmatrix} f_1(\boldsymbol{x}) & \cdots & f_m(\boldsymbol{x}) \end{bmatrix} \begin{bmatrix} Dg_1(\boldsymbol{x}) \\ \vdots \\ Dg_m(\boldsymbol{x}) \end{bmatrix} + \begin{bmatrix} g_1(\boldsymbol{x}) & \cdots & g_m(\boldsymbol{x}) \end{bmatrix} \begin{bmatrix} Df_1(\boldsymbol{x}) \\ \vdots \\ Df_m(\boldsymbol{x}) \end{bmatrix} \\
&= {}^{1\backslash m}\begin{bmatrix} \boldsymbol{f}(\boldsymbol{x})^\top \end{bmatrix}^{m\backslash n} \begin{bmatrix} D\boldsymbol{g}(\boldsymbol{x}) \end{bmatrix} + {}^{1\backslash m}\begin{bmatrix} \boldsymbol{g}(\boldsymbol{x})^\top \end{bmatrix}^{m\backslash n} \begin{bmatrix} D\boldsymbol{f}(\boldsymbol{x}) \end{bmatrix}.
\end{aligned}$$

- Let $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ be a given matrix and $\boldsymbol{y} \in \mathbb{R}^{m}$ a given vector. Then,

$$\begin{aligned} D\left(\boldsymbol{y}^\top \boldsymbol{A}\boldsymbol{x}\right) &= \boldsymbol{y}^\top \boldsymbol{A}, \\ D\left(\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x}\right) &= \boldsymbol{x}^\top \left(\boldsymbol{A} + \boldsymbol{A}^\top\right). \qquad \boxed{\text{if } m = n} \end{aligned}$$

Proof,

$$\begin{aligned} D\left(\boldsymbol{y}^\top \mathbf{A}\boldsymbol{x}\right) = D\Big(\boldsymbol{y}^\top (\mathbf{A}\boldsymbol{x})\Big) &= D\Big(\boldsymbol{f}(\boldsymbol{y})^\top \big(\boldsymbol{g}(\boldsymbol{x})\big)\Big) \\ &= \boldsymbol{f}(\boldsymbol{y})^\top D\boldsymbol{g}(\boldsymbol{x}) + \boldsymbol{g}(\boldsymbol{x})^\top D\boldsymbol{f}(\boldsymbol{y}) \\ &= \boldsymbol{y}^\top D(\mathbf{A}\boldsymbol{x}) + (\mathbf{A}\boldsymbol{x})^\top [\mathbf{0}] \\ &= \boldsymbol{y}^\top \mathbf{A}. \end{aligned}$$

If $\boldsymbol{A} \in \mathbb{R}^{n\times n}$,

$$\begin{aligned} D\left(\boldsymbol{x}^\top \mathbf{A}\boldsymbol{x}\right) = D\Big(\boldsymbol{x}^\top (\mathbf{A}\boldsymbol{x})\Big) &= D\Big(\boldsymbol{f}(\boldsymbol{x})^\top \big(\boldsymbol{g}(\boldsymbol{x})\big)\Big) \\ &= \boldsymbol{f}(\boldsymbol{x})^\top D\boldsymbol{g}(\boldsymbol{x}) + \boldsymbol{g}(\boldsymbol{x})^\top D\boldsymbol{f}(\boldsymbol{x}) \\ &= \boldsymbol{x}^\top D(\mathbf{A}\boldsymbol{x}) + (\mathbf{A}\boldsymbol{x})^\top D(\mathbf{I}_n\boldsymbol{x}) \\ &= \boldsymbol{x}^\top \mathbf{A} + \boldsymbol{x}^\top \mathbf{A}^\top \mathbf{I} \\ &= \boldsymbol{x}^\top \left(\mathbf{A} + \mathbf{A}^\top\right). \end{aligned}$$

- It follows that if $\boldsymbol{Q}$ is a symmetric matrix, then

$$\begin{aligned} D\left(\boldsymbol{x}^\top \boldsymbol{Q}\boldsymbol{x}\right) &= 2\boldsymbol{x}^\top \boldsymbol{Q}, \\ D\left(\boldsymbol{x}^\top \boldsymbol{x}\right) &= 2\boldsymbol{x}^\top. \end{aligned}$$

Proof,

$$\begin{aligned} D\left(\boldsymbol{x}^\top \mathbf{Q}\boldsymbol{x}\right) &= D\Big(\boldsymbol{x}^\top (\mathbf{Q}\boldsymbol{x})\Big) \\ &= \boldsymbol{x}^\top \left(\mathbf{Q} + \mathbf{Q}^\top\right) \\ &= 2\boldsymbol{x}^\top \mathbf{Q}, \end{aligned}$$

$$\begin{aligned} D\left(\boldsymbol{x}^\top \boldsymbol{x}\right) = D\left(\boldsymbol{x}^\top \mathbf{I}\boldsymbol{x}\right) &= D\Big(\boldsymbol{x}^\top (\mathbf{I}\boldsymbol{x})\Big) \\ &= \boldsymbol{x}^\top \left(\mathbf{I} + \mathbf{I}^\top\right) \\ &= 2\boldsymbol{x}^\top. \end{aligned}$$

- Derivative of scalar by scalar,

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}\alpha}\left((\alpha\boldsymbol{x})^\top \mathbf{Q}(\alpha\boldsymbol{x})\right) &= (\alpha\boldsymbol{x})^\top \frac{\mathrm{d}}{\mathrm{d}\alpha}(\mathbf{Q}(\alpha\boldsymbol{x})) + (\mathbf{Q}(\alpha\boldsymbol{x}))^\top \frac{\mathrm{d}}{\mathrm{d}\alpha}(\mathbf{I}(\alpha\boldsymbol{x})) \\ &= (\alpha\boldsymbol{x})^\top \mathbf{Q}\boldsymbol{x} + (\alpha\boldsymbol{x})^\top \mathbf{Q}^\top \boldsymbol{x} \\ &= 2(\alpha\boldsymbol{x})^\top \mathbf{Q}\boldsymbol{x} \\ &= 2\alpha\boldsymbol{x}^\top \mathbf{Q}\boldsymbol{x} \end{aligned}$$

- Hand writing derivation: given $\underline{y} \in \mathbb{R}^m, \underline{A} \in \mathbb{R}^{m\times n}, \underline{x} \in \mathbb{R}^n$, write the derivative as

$$
\begin{aligned}
D_{\boldsymbol{x}}\left(\boldsymbol{y}^\top \mathbf{A}\boldsymbol{x}\right) &= D_{\underline{x}}\left(\underline{y}^\top \underline{A}\ \ \underline{x}\right) \\
&= D\left({}^{1\backslash m}\begin{bmatrix} \ \ \underline{y}^\top \ \ \end{bmatrix} {}^{m\backslash n}\begin{bmatrix} \\ \quad \underline{A} \quad \\ \\ \end{bmatrix} {}^{n\backslash 1}\begin{bmatrix} \\ \underline{x} \\ \\ \end{bmatrix} \right) \\
&= {}^{1\backslash m}\begin{bmatrix} \ \ \underline{y}^\top \ \ \end{bmatrix} D\left({}^{m\backslash n}\begin{bmatrix} \\ \quad \underline{A} \quad \\ \\ \end{bmatrix} {}^{n\backslash 1}\begin{bmatrix} \\ \underline{x} \\ \\ \end{bmatrix} \right) \\
&\quad + \left({}^{m\backslash n}\begin{bmatrix} \\ \quad \underline{A} \quad \\ \\ \end{bmatrix} {}^{n\backslash 1}\begin{bmatrix} \\ \underline{x} \\ \\ \end{bmatrix} \right)^\top {}^{m\backslash n}\begin{bmatrix} \\ \quad D\underline{y} \quad \\ \\ \end{bmatrix} \\
&= {}^{1\backslash m}\begin{bmatrix} \ \ \underline{y}^\top \ \ \end{bmatrix} {}^{m\backslash n}\begin{bmatrix} \\ \quad \underline{A} \quad \\ \\ \end{bmatrix} + {}^{1\backslash n}\begin{bmatrix} \ \ \underline{x}^\top \ \ \end{bmatrix} {}^{n\backslash m}\begin{bmatrix} \\ \quad \underline{A}^\top \quad \\ \\ \end{bmatrix} {}^{m\backslash n}\begin{bmatrix} \\ \quad \underline{0} \quad \\ \\ \end{bmatrix} \\
&= {}^{1\backslash m}\begin{bmatrix} \ \ \underline{y}^\top \ \ \end{bmatrix} {}^{m\backslash n}\begin{bmatrix} \\ \quad \underline{A} \quad \\ \\ \end{bmatrix} \\
&= \boldsymbol{y}^\top \mathbf{A}.
\end{aligned}
$$

[Ref]: Edwin K.P. Chong, Stanislaw H. Żak, "PART I MATHEMATICAL REVIEW" in "An introduction to optimization", 4th Edition, John Wiley and Sons, Inc. 2013.