Appendix C

Concepts from multivariate calculus

In this appendix we provide a summary of several useful facts and results (without proof) from differential calculus of several variables. We assume familiarity with vectors and matrices (Appendices A and B), since they provide useful tools for expressing properties of these functions in a concise manner.

C.1 Gradient and first-order directional derivatives

Let $\phi : \mathbb{R}^n \longrightarrow \mathbb{R}$ for some integer $n \geq 1$, that is, ϕ is a scalar-valued function of a vector. As usual, an element $\mathbf{x} \in \mathbb{R}^n$ is denoted by $\mathbf{x} = (x_1, x_2, \cdots, x_n)^{\mathrm{T}}$, where T denotes the transpose. Let $\|\mathbf{x}\|$ denote a *norm* of $\mathbf{x}$. The function ϕ is said to be *continuous* at the point $\mathbf{x}$ if and only if for every $\epsilon > 0$, there exists a $\delta > 0$, such that $|\phi(\mathbf{y}) - \phi(\mathbf{x})| < \epsilon$ for all $\mathbf{y}$ such that $\|\mathbf{y} - \mathbf{x}\| < \delta$. That is, $\phi(\mathbf{y})$ is "close" to $\phi(\mathbf{x})$ whenever $\mathbf{y}$ is "close" to $\mathbf{x}$ in *every direction*. It is possible for a function of several variables to be *continuous* in each of its component variables separately without being continuous as a function of several variables. As an example, consider (assuming $n = 2$)

$$\phi(\mathbf{x}) = \begin{cases} \frac{x_1 x_2}{x_1^2 + x_2^2} & \text{if } x_1^2 + x_2^2 > 0 \\ 0 & \text{otherwise.} \end{cases} \tag{C.1.1}$$

When $x_2 = 0$, considered as a function of x_1, $\phi(\mathbf{x})$ is identically zero for all x_1 and hence is continuous for all x_1, including $x_1 = 0$. A similar conclusion follows when $\phi(\mathbf{x})$ is considered as a function of x_2 for a fixed $x_1 = 0$. Thus, $\phi(\mathbf{x})$ is continuous at the origin when considered as functions of x_1 and x_2 separately. But as a function of two variables, $\phi(\mathbf{x})$ is *not* continuous at the origin. For, along the line $x_1 = x_2$, it can be verified that

$$\phi(\mathbf{x}) = \begin{cases} \frac{1}{2} \text{ when } x_1 = x_2 \neq 0 \\ 0 \text{ for } \quad x_1 = x_2 = 0 \end{cases}$$

Thus, there is a discontinuity in $\phi(\mathbf{x})$ in the direction $x_1 = x_2$ and hence $\phi(\mathbf{x})$ is *not* continuous in $\mathbb{R}^2$.

In the case of single variables, existence of derivatives implies continuity. A function of several variables may possess partial derivatives with respect to each of the variables and yet may not be continuous in $\mathbb{R}^n$. For the function ϕ in (C.1.1), it can be verified that the partial derivatives with respect to x_1 and x_2 exist and are equal to zero at the origin and yet it is not continuous at the origin, as seen above.

Obviously partial derivatives relate to properties only along the coordinate directions and hence do not constitute an analog of the derivative. In search of the analog of derivative, we turn to the notion of differentiability. Recall that when $n = 1$, we say that a function is *differentiable* at x if and only if there exists a unique number $\phi'(x)$, called the *derivative* of ϕ at x, such that for all t small

$$\phi(x + t) - \phi(x) = \phi'(x)t + \text{HOT}(t) \tag{C.1.2}$$

where HOT(t) denoting *higher order terms* in t is such that

$$\lim_{t \to 0} \frac{\text{HOT}(t)}{t} \to 0.$$

Analogously, we say $\phi : \mathbb{R}^n \longrightarrow \mathbb{R}$ is *differentiable* at $\mathbf{x} \in \mathbb{R}^n$, if and only if, there exists a vector $\mathbf{u}$ such that for every $\mathbf{z} \in \mathbb{R}^n$

$$\phi(\mathbf{x} + \mathbf{z}) - \phi(\mathbf{x}) = \langle \mathbf{u}, \mathbf{z} \rangle + \text{HOT}(\mathbf{z}) \tag{C.1.3}$$

where $\langle \mathbf{u}, \mathbf{z} \rangle = \mathbf{u}^{\mathrm{T}}\mathbf{z}$ is the *inner product* of $\mathbf{u}$ and $\mathbf{z}$ and HOT($\mathbf{z}$) denoting the higher order terms in the components of $\mathbf{z}$ is such that

$$\lim_{\|\mathbf{z}\| \to 0} \frac{\text{HOT}(\mathbf{z})}{\|\mathbf{z}\|} = 0. \tag{C.1.4}$$

Comparing (C.1.2) with (C.1.3), we must expect the vector $\mathbf{u}$ to play the role of derivative of ϕ. We now list several properties of interest to us.

If the vector $\mathbf{u}$ in (C.1.3) exists, then it is *unique*. For, let $\mathbf{v}$ be another vector such that

$$\phi(\mathbf{x} + \mathbf{z}) - \phi(\mathbf{x}) = \langle \mathbf{v}, \mathbf{z} \rangle + \text{HOT}(\mathbf{z}). \tag{C.1.5}$$

Subtracting (C.1.5) from (C.1.3), we obtain

$$\langle \mathbf{u} - \mathbf{v}, \mathbf{z} \rangle = \text{HOT}(\mathbf{z}). \tag{C.1.6}$$

Combining (C.1.4) and (C.1.6), it follows that $\mathbf{u} = \mathbf{v}$, and hence the uniqueness.

Gradient The unique vector $\mathbf{u}$ in (C.1.3), if it exists, is called the *gradient* of ϕ at $\mathbf{x}$ and is denoted by $\nabla\phi(\mathbf{x})$.

Condition for differentiability If $\phi : \mathbb{R}^n \longrightarrow \mathbb{R}$ has continuous partial derivatives in the neighborhood of $\mathbf{x}$, then ϕ is *differentiable* at $\mathbf{x}$, and the *gradient* is given by:

$$\nabla\phi(\mathbf{x}) = \left(\frac{\partial \phi}{\partial x_1}, \frac{\partial \phi}{\partial x_2}, \ldots, \frac{\partial \phi}{\partial x_n} \right)^{\mathrm{T}}. \tag{C.1.7}$$

Also, if ϕ is differentiable at $\mathbf{x}$, then it is continuous at $\mathbf{x}$. Let $\phi_1(\mathbf{x})$ and $\phi_2(\mathbf{x})$ be functions from $\mathbb{R}^n$ to $\mathbb{R}$, satisfying the conditions for differentiability. Then,

$$\nabla(\phi_1(\mathbf{x}) + \phi_2(\mathbf{x})) = \nabla\phi_1(\mathbf{x}) + \nabla\phi_2(\mathbf{x})$$
$$\nabla(c\phi_1(\mathbf{x})) = c\nabla\phi_1(\mathbf{x}).$$

That is, ∇ is a *linear operator* (Appendix B). Also

$$\nabla(\phi_1(\mathbf{x})\phi_2(\mathbf{x})) = \phi_1(\mathbf{x})\nabla\phi_2(\mathbf{x}) + \phi_2(\mathbf{x})\nabla\phi_1(\mathbf{x}).$$

Directional derivative Let $\mathbf{z}$ denote an arbitrary **unit vector** in $\mathbb{R}^n$. The *directional derivative* of ϕ at the point $\mathbf{x}$ in the direction $\mathbf{z}$ denoted by $\phi'(\mathbf{x};\mathbf{z})$ is defined by

$$\phi'(\mathbf{x};\mathbf{z}) = \lim_{\Delta t\to 0} \frac{\phi(\mathbf{x}+\mathbf{z}\Delta t) - \phi(\mathbf{x})}{\Delta t}, \tag{C.1.8}$$

assuming that the limit exists and is finite.

Let $g(t) = \phi(\mathbf{x} + \mathbf{z}t)$. Then, from

$$\frac{\phi(\mathbf{x}+\mathbf{z}\Delta t) - \phi(\mathbf{x})}{\Delta t} = \frac{g(\Delta t) - g(0)}{\Delta t}$$

it follows that

$$\phi'(\mathbf{x};\mathbf{z}) = g'(0).$$

More generally, let $\mathbf{y} = \mathbf{x} + \mathbf{z}t$. Then

$$\phi'(\mathbf{y};\mathbf{z}) = \lim_{\Delta t\to 0} \frac{\phi(\mathbf{y}+\mathbf{z}\Delta t) - \phi(\mathbf{y})}{\Delta t} = \lim_{\Delta t\to 0} \frac{\phi(t+\Delta t) - g(t)}{\Delta t} = g'(t)$$

That is, the directional derivative of ϕ is the ordinary derivative of an auxiliary function g.

Let $\mathbf{e}_i$, $1 \le i \le n$, denote the ith *unit* vector in $\mathbb{R}^n$. Then, $\phi'(\mathbf{x};\mathbf{e}_i)$ is the *partial derivative* of ϕ with respect to the variable x_i denoted by $\partial\phi/\partial x_i$.

Existence of partial derivatives of ϕ does *not*, however, guarantee the existence of directional derivatives. For, consider the function ϕ in (C.1.1). It can be verified that at the origin $\mathbf{x} = (0,0)^{\mathrm{T}}$, both the partial derivatives exist and are equal to zero. But the directional derivative of ϕ at the origin does *not* exist in the direction $\mathbf{z} = \frac{1}{\sqrt{2}}(1,1)^{\mathrm{T}}$, since ϕ is discontinuous in this direction as seen above.

Conditions for the existence of directional derivative If $\phi : \mathbb{R}^n \longrightarrow \mathbb{R}$ is such that it has continuous partial derivatives at $\mathbf{x}$, then ϕ has directional derivative $\phi'(\mathbf{x};\mathbf{z})$ for every unit vector $\mathbf{z}$.

Combining this with the condition for differentiability, it is seen that in this case ϕ is also differentiable and the directional derivatives $\phi'(\mathbf{x};\mathbf{z})$ and the gradient $\nabla\phi(\mathbf{x})$ are related through

$$\phi'(\mathbf{x};\mathbf{z}) = \langle \mathbf{z}, \nabla\phi(\mathbf{x})\rangle = \mathbf{z}^{\mathrm{T}}\nabla\phi(\mathbf{x}) = \sum_{i=1}^{n} z_i \frac{\partial\phi}{\partial x_i}. \tag{C.1.9}$$

Recall from Appendix A that the inner product of a vector with a unit vector denotes the projection of that vector onto the unit vector. Accordingly, the directional derivative of $\phi(\mathbf{x})$ along the direction $\mathbf{z}$ is the projection of the gradient $\nabla\phi(\mathbf{x})$ onto $\mathbf{z}$.

Again considering the function ϕ in (C.1.1), we leave it to the reader to verify that $\partial\phi/\partial x_i$ is **not** continuous at $\mathbf{x} = (0,\ 0)^{\mathrm{T}}$, and hence ϕ is **not** differentiable at the origin,

Direction of maximum rate of change A differentiable function changes most rapidly in the direction of the gradient. For, from (C.1.9) we have, since $\|\mathbf{z}\| = 1$,

$$|\phi'(\mathbf{x};\mathbf{z})| = \|\nabla\phi(\mathbf{x})\| \cos\theta$$

where θ is the angle between $\mathbf{z}$ and $\nabla\phi(\mathbf{x})$ (Appendix A). Clearly, this is maximum when $\theta = 0$. This property is the basis for the optimization procedures described in Chapters 10 through 12.

If the directional derivative $\phi'(\mathbf{x};\mathbf{z})$ exists and is continuous, then it is **linear** in $\mathbf{z}$. That is,

$$\phi'(\mathbf{x};\mathbf{z}_1 + \mathbf{z}_2) = \phi'(\mathbf{x};\mathbf{z}_1) + \phi'(\mathbf{x};\mathbf{z}_2)$$

$$\phi'(\mathbf{x};c\mathbf{z}) = c\phi'(\mathbf{x};\mathbf{z})$$

for any real constant c.

Property C.1.1 The following statements are equivalent:

(a) The function ϕ belongs to the class C_1.
(b) All the directional derivatives of ϕ exist and are continuous.
(c) All the partial derivatives of ϕ exist and are continuous.

Chain rule Let $x_i : \mathbb{R} \longrightarrow \mathbb{R}$ for $1 \leq i \leq n$ and $\phi : \mathbb{R}^{n+1} \longrightarrow \mathbb{R}$. Consider the composite function

$$\phi(x_1(t), x_2(t), \ldots, x_n(t), t) \tag{C.1.10}$$

Then,

$$\frac{\mathrm{d}\phi}{\mathrm{d}t} = \frac{\partial\phi}{\partial x_1}\frac{\mathrm{d}x_1}{\mathrm{d}t} + \frac{\partial\phi}{\partial x_2}\frac{\mathrm{d}x_2}{\mathrm{d}t} + \cdots + \frac{\partial\phi}{\partial x_n}\frac{\mathrm{d}x_n}{\mathrm{d}t} + \frac{\partial\phi}{\partial t}.$$

For example, $\mathbf{x}(t) = (x_1(t), \ldots, x_n(t))^{\mathrm{T}}$ could denote the state of a dynamic system evolving according to a differential equation.

C.2 Hessian and second-order directional derivative

Let $\mathbf{z}$ and $\mathbf{w}$ be two unit vectors in $\mathbb{R}^n$. Then the **directional derivative of $\phi(\mathbf{x})$ of second order** with respect to the directions $\mathbf{z}$ and $\mathbf{w}$, denoted by $\phi''(\mathbf{x};\mathbf{z},\mathbf{w})$ is

defined as the directional derivative in the direction $\mathbf{w}$ of the directional derivative in the direction $\mathbf{z}$ of ϕ at the point $\mathbf{x}$. Thus,

$$\begin{aligned}\phi''(\mathbf{x};\mathbf{z},\mathbf{w}) = \mathbf{w}^{\mathrm{T}}\nabla(\mathbf{z}^{\mathrm{T}}\nabla\phi(\mathbf{x})) &= \textstyle\sum_{i=1}^{n} w_i \frac{\partial}{\partial x_i}\left(\sum_{j=1}^{n} z_j \frac{\partial\phi}{\partial x_j}\right)\\ &= \textstyle\sum_{i=1}^{n}\sum_{j=1}^{n} w_i z_j \left(\frac{\partial^2\phi}{\partial x_i \partial x_j}\right)\\ &= \mathbf{w}^{\mathrm{T}}[\nabla^2\phi(\mathbf{x})]\mathbf{z} \qquad \text{(C.2.1)}\end{aligned}$$

where

$$\nabla^2\phi(\mathbf{x}) = \left[\frac{\partial^2\phi}{\partial x_i \partial x_j}\right] \quad \text{for} \quad 1 \le i, j \le n \qquad \text{(C.2.2)}$$

is an $n \times n$ symmetric matrix of second partial derivatives of $\phi(\mathbf{x})$, called the **Hessian** of $\phi(\mathbf{x})$. Notice that $\phi''(\mathbf{x};\mathbf{z},\mathbf{w})$ is linear in $\mathbf{z}$ and $\mathbf{w}$ and the right-hand side expression in (C.2.1) is often called the **bilinear form** to emphasize the linearity in $\mathbf{z}$ and $\mathbf{w}$. When $\mathbf{z} = \mathbf{w}$, then

$$\phi''(\mathbf{x};\mathbf{z},\mathbf{w}) = \mathbf{z}^{\mathrm{T}}[\nabla\phi(\mathbf{x})]\mathbf{z} \qquad \text{(C.2.3)}$$

is a **quadratic** form in $\mathbf{z}$.

Property C.2.1 The following conditions are equivalent:

(a) The function ϕ is of class C_2.
(b) The second-order directional derivatives exist and are continuous in $\mathbf{z}$ and $\mathbf{w}$.
(c) All the partial derivatives $\frac{\partial^2\phi}{\partial x_i \partial x_j}$ of second order exist and are continuous.

Continuing in this fashion, one may define the class C_r consisting of functions with continuous partial derivatives of order $r > 2$.

C.3 Vector-valued function of a vector

Let $\phi : \mathbb{R}^n \longrightarrow \mathbb{R}^m$, where $\phi(\mathbf{x}) = (\phi_1(\mathbf{x}), \phi_2(\mathbf{x}), \ldots, \phi_m(\mathbf{x}))^{\mathrm{T}}$ be a vector-valued function, where $\mathbf{x} = (x_1, x_2, \ldots, x_n)^{\mathrm{T}}$. Consider a unit vector $\mathbf{z} \in \mathbb{R}^n$. Assume that each component function $\phi_i(\mathbf{x})$, $1 \le i \le m$, satisfy the conditions for the existence of directional derivatives. Then, the directional derivative $\phi'(\mathbf{x};\mathbf{z})$ of $\phi(\mathbf{x})$ in the direction $\mathbf{z}$ is a **column vector** defined by

$$\phi'(\mathbf{x};\mathbf{z}) = (\phi'_1(\mathbf{x};\mathbf{z}), \phi'_2(\mathbf{x};\mathbf{z}), \ldots, \phi'_m(\mathbf{x};\mathbf{z}))^{\mathrm{T}}.$$

Since each $\phi_i : \mathbb{R}^n \longrightarrow \mathbb{R}$, from the definition (C.1.9), we obtain

$$\begin{aligned}\phi'(\mathbf{x};\mathbf{z}) &= \begin{pmatrix} \phi'_1(\mathbf{x};\mathbf{z}) \\ \phi'_2(\mathbf{x};\mathbf{z}) \\ \vdots \\ \phi'_m(\mathbf{x};\mathbf{z}) \end{pmatrix} = \begin{pmatrix} \mathbf{z}^{\mathrm{T}}\nabla\phi_1(\mathbf{x}) \\ \mathbf{z}^{\mathrm{T}}\nabla\phi_2(\mathbf{x}) \\ \vdots \\ \mathbf{z}^{\mathrm{T}}\nabla\phi_m(\mathbf{x}) \end{pmatrix} \\ &= \begin{pmatrix} \frac{\partial\phi_1}{\partial x_1} & \frac{\partial\phi_1}{\partial x_2} & \cdots & \frac{\partial\phi_1}{\partial x_n} \\ \frac{\partial\phi_2}{\partial x_1} & \frac{\partial\phi_2}{\partial x_2} & \cdots & \frac{\partial\phi_2}{\partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial\phi_m}{\partial x_1} & \frac{\partial\phi_m}{\partial x_2} & \cdots & \frac{\partial\phi_m}{\partial x_n} \end{pmatrix} \begin{pmatrix} \mathbf{z}_1 \\ \mathbf{z}_2 \\ \vdots \\ \mathbf{z}_n \end{pmatrix} \\ &= \mathbf{D}_\phi(\mathbf{x})\mathbf{z}\end{aligned} \tag{C.3.1}$$

where

$$\mathbf{D}_\phi(\mathbf{x}) = \left[\frac{\partial\phi_i}{\partial x_j}\right],\ 1 \le i \le m;\ 1 \le j \le n \tag{C.3.2}$$

is an $m \times n$ matrix, called the **Jacobian** of $\phi(\mathbf{x})$.

Notice that when $m = 1$, $\mathbf{D}_\phi(\mathbf{x}) = [\nabla\phi(\mathbf{x})]^{\mathrm{T}}$, the transpose of the gradient of $\phi(\mathbf{x})$. Thus, (C.3.1) can be succinctly written as

$$\phi'(\mathbf{x};\mathbf{z}) = \mathbf{D}_\phi(\mathbf{x})\mathbf{z}. \tag{C.3.3}$$

Let $\mathbf{z}$ and $\mathbf{w}$ be two unit vectors in $\mathbb{R}^n$. Then, the second-order directional derivation of ϕ in the directions $\mathbf{z}$ and $\mathbf{w}$ denoted by $\phi''(\mathbf{x};\mathbf{z},\mathbf{w})$, is defined as the directional derivative in the direction $\mathbf{w}$ of the vector-valued function $\phi'(\mathbf{x};\mathbf{z})$ defined in (C.3.1). Thus,

$$\phi''(\mathbf{x};\mathbf{z},\mathbf{w}) = \begin{pmatrix} \mathbf{w}^{\mathrm{T}}\nabla\phi'_1(\mathbf{x};\mathbf{z}) \\ \mathbf{w}^{\mathrm{T}}\nabla\phi'_2(\mathbf{x};\mathbf{z}) \\ \vdots \\ \mathbf{w}^{\mathrm{T}}\nabla\phi'_m(\mathbf{x};\mathbf{z}) \end{pmatrix}. \tag{C.3.4}$$

Recall that, for $1 \le k \le m$,

$$\phi'_k(\mathbf{x};\mathbf{z}) = \sum_{j=1}^{n} z_j \frac{\partial\phi_k}{\partial x_j},$$

and so $\nabla\phi'_k(\mathbf{x};\mathbf{z})$ is the gradient vector of $\phi'_k(\mathbf{x};\mathbf{z})$. We have

$$\begin{aligned}\nabla\phi'_k(\mathbf{x};\mathbf{z}) &= \nabla\left[\textstyle\sum_{j=1}^{n} z_j \frac{\partial\phi_k}{\partial x_j}\right] \\ &= \left(\tfrac{\partial}{\partial x_1}\textstyle\sum_{j=1}^{n} z_j \frac{\partial\phi_k}{\partial x_j}, \tfrac{\partial}{\partial x_2}\textstyle\sum_{j=1}^{n} z_j \frac{\partial\phi_k}{\partial x_j}, \ldots, \tfrac{\partial}{\partial x_n}\textstyle\sum_{j=1}^{n} z_j \frac{\partial\phi_k}{\partial x_j}\right)^{\mathrm{T}} \\ &= \nabla^2\phi_k(\mathbf{x})\mathbf{z}\end{aligned} \tag{C.3.5}$$

where $\nabla^2\phi_k(\mathbf{x})$ is the Hessian of $\phi_k(\mathbf{x})$.

Table C.3.1 *Gradients of some useful linear and quadratic functions*

Function ϕ	Gradient
$\phi(\mathbf{x}) = \mathbf{a}^{\mathrm{T}}\mathbf{x},\ \ \mathbf{a} \in \mathbb{R}^n, \mathbf{x} \in \mathbb{R}^n$	$\nabla\phi(\mathbf{x}) = \mathbf{a}$
$\phi(\mathbf{x}) = \mathbf{a}^{\mathrm{T}}\mathbf{h}(\mathbf{x}),\ \ \mathbf{a} \in \mathbb{R}^n, \mathbf{x} \in \mathbb{R}^n$ $\mathbf{h}(\mathbf{x}) = (h_1(\mathbf{x}), h_2(\mathbf{x}), \ldots, h_n(\mathbf{x}))^{\mathrm{T}}$	$\nabla\phi(\mathbf{x}) = \mathbf{D}_{\mathbf{h}}^{\mathrm{T}}(\mathbf{x})\mathbf{a}$, where $\mathbf{D}_{\mathbf{h}}(\mathbf{x}) \in \mathbb{R}^{n\times n}$ is the Jacobian of $\mathbf{h}(\mathbf{x})$
$\phi(\mathbf{x}) = \frac{1}{2}\mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x} - \mathbf{b}^{\mathrm{T}}\mathbf{x}, \mathbf{x} \in \mathbb{R}^n, \mathbf{b} \in \mathbb{R}^n$, $\mathbf{A} \in \mathbb{R}^{n\times n}$	$\nabla\phi(\mathbf{x}) = \mathbf{A}_s\mathbf{x} - \mathbf{b}$, where $\mathbf{A}_s = \frac{\mathbf{A}+\mathbf{A}^{\mathrm{T}}}{2}$, is the symmetric part of $\mathbf{A}$.
$\phi(\mathbf{x}) = \frac{1}{2}\mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x} - \mathbf{b}^{\mathrm{T}}\mathbf{x}, \mathbf{x} \in \mathbb{R}^n, \mathbf{b} \in \mathbb{R}^n$, $\mathbf{A} \in \mathbb{R}^{n\times n}$, symmetric	$\nabla\phi(\mathbf{x}) = \mathbf{A}\mathbf{x} - \mathbf{b}$
$\phi(\mathbf{x}) = \frac{1}{2}\mathbf{h}^{\mathrm{T}}(\mathbf{x})\mathbf{A}\mathbf{h}(\mathbf{x}), \mathbf{x} \in \mathbb{R}^n, \mathbf{A} \in \mathbb{R}^{n\times n}$, symmetric, $\mathbf{h}(\mathbf{x}) = (h_1(\mathbf{x}), h_2(\mathbf{x}), \ldots, h_n(\mathbf{x}))^{\mathrm{T}}$	$\nabla\phi(\mathbf{x}) = \mathbf{D}_{\mathbf{h}}^{\mathrm{T}}\mathbf{A}\mathbf{h}(\mathbf{x})$, where $\mathbf{D}_{\mathbf{h}}(\mathbf{x}) \in \mathbb{R}^{n\times n}$ is the Jacobian of $\mathbf{h}(\mathbf{x})$
$\phi(\mathbf{x}) = \frac{1}{2}\mathbf{h}^{\mathrm{T}}(\mathbf{x})\mathbf{A}\mathbf{g}(\mathbf{x}), \mathbf{x} \in \mathbb{R}^n, \mathbf{A} \in \mathbb{R}^{n\times m}$, $\mathbf{h}(\mathbf{x}) = (h_1(\mathbf{x}), h_2(\mathbf{x}), \ldots, h_n(\mathbf{x}))^{\mathrm{T}}$, $\mathbf{g}(\mathbf{x}) = (g_1(\mathbf{x}), g_2(\mathbf{x}), \ldots, g_m(\mathbf{x}))^{\mathrm{T}}$	$\nabla\phi(\mathbf{x}) = \frac{1}{2}[\mathbf{D}_{\mathbf{h}}^{\mathrm{T}}(\mathbf{x})\mathbf{A}\mathbf{g}(\mathbf{x}) + \mathbf{D}_{\mathbf{g}}^{\mathrm{T}}(\mathbf{x})\mathbf{A}^{\mathrm{T}}\mathbf{h}(\mathbf{x})]$, where $\mathbf{D}_{\mathbf{h}}(\mathbf{x}) \in \mathbb{R}^{n\times n}$ is the Jacobian of $\mathbf{h}(\mathbf{x})$, and $\mathbf{D}_{\mathbf{g}}(\mathbf{x}) \in \mathbb{R}^{m\times n}$ is the Jacobian of $\mathbf{g}(\mathbf{x})$
$\phi(\mathbf{x}) = \frac{1}{2}\mathbf{a}^{\mathrm{T}}\mathbf{b}(\mathbf{x}), \mathbf{x} \in \mathbb{R}^n, \mathbf{a} \in \mathbb{R}^n$, $\mathbf{b}(\mathbf{x}) = (b_1(\mathbf{x}), b_2(\mathbf{x}), \ldots, b_n(\mathbf{x}))^{\mathrm{T}}$, $b_i(\mathbf{x}) = (\mathbf{x})^{\mathrm{T}}\mathbf{B}_i\mathbf{x}, \mathbf{B}_i \in \mathbb{R}^{n\times n}$, symmetric, $1 \le i \le n$	$\nabla\phi(\mathbf{x}) = \sum_{i=1}^{n} a_i\mathbf{B}_i\mathbf{x}$
$\eta(\mathbf{x}) = \phi o\psi(\mathbf{x}) = \phi(\psi(\mathbf{x}))$, where $\phi : \mathbb{R}^n \longrightarrow \mathbb{R}^p$, $\psi : \mathbb{R}^m \longrightarrow \mathbb{R}^n$, and n, m, and p are positive integers. Then, $\eta : \mathbb{R}^m \longrightarrow \mathbb{R}^p$ is called the composite function.	$\mathbf{D}_\eta(\mathbf{x}) = \mathbf{D}_\psi(\mathbf{y})\mathbf{D}_\phi(\mathbf{x})$, where $\mathbf{y} = \phi(\mathbf{x})$ and $\mathbf{D}_\phi \in \mathbb{R}^{p\times n}$, $\mathbf{D}_\psi \in \mathbb{R}^{n\times m}$, and $\mathbf{D}_\eta \in \mathbb{R}^{p\times m}$ are the Jacobians of ϕ, ψ, and η, respectively

Combining (C.3.4) and (C.3.5), we obtain

$$\phi''(\mathbf{x};\mathbf{z},\mathbf{w}) = \begin{pmatrix} \mathbf{w}^{\mathrm{T}}\nabla^2\phi_1(\mathbf{x})\mathbf{z} \\ \mathbf{w}^{\mathrm{T}}\nabla^2\phi_2(\mathbf{x})\mathbf{z} \\ \vdots \\ \mathbf{w}^{\mathrm{T}}\nabla^2\phi_m(\mathbf{x})\mathbf{z} \end{pmatrix}. \tag{C.3.6}$$

Symbolically, we denote the vector of second variation of $\phi : \mathbb{R}^n \longrightarrow \mathbb{R}^m$ as

$$\phi''(\mathbf{x};\mathbf{z},\mathbf{w}) = \mathbf{D}_\phi^2(\mathbf{x};\mathbf{z},\mathbf{w}) \tag{C.3.7}$$

and when $\mathbf{w} = \mathbf{z}$, then as

$$\phi''(\mathbf{x};\mathbf{z},\mathbf{z}) = \mathbf{D}_\phi^2(\mathbf{x};\mathbf{z}). \tag{C.3.8}$$

Table C.3.1 contains a listing of gradients of many functions of interest in the main body of this book.

C.4 Taylor series

Scalar-valued function of a vector Let $\phi : \mathbb{R}^n \to \mathbb{R}$ be in class C_2. Then,

$$\phi(\mathbf{x}+\mathbf{y}) \approx \phi(\mathbf{x}) + \langle \mathbf{y}, \nabla\phi(\mathbf{x})\rangle + \frac{1}{2}\langle \mathbf{y}, \nabla^2\phi(\mathbf{x})\mathbf{y}\rangle \tag{C.4.1}$$

for any $\mathbf{y} \in \mathbb{R}^n$, such that $\|\mathbf{y}\|$ is small.

Vector-valued function of a vector Let $\phi : \mathbb{R}^n \longrightarrow \mathbb{R}^m$ be such that each component of ϕ is a class C_2 function. Then, for any vector $\mathbf{y} \in \mathbb{R}^n$

$$\phi(\mathbf{x}+\mathbf{y}) \approx \phi(\mathbf{x}) + \mathbf{D}_\phi(\mathbf{x})\mathbf{y} + \frac{1}{2}\mathbf{D}^2_\phi(\mathbf{x};\mathbf{y}). \tag{C.4.2}$$

The first term $\phi(\mathbf{x})$ on the right-hand side of (C.4.1) and (C.4.2) is independent of $\mathbf{y}$. The second term – $\langle \mathbf{y}, \nabla\phi(\mathbf{x})\rangle$ in (C.4.1) and $\mathbf{D}_\phi(\mathbf{x})\mathbf{y}$ in (C.4.2) is **linear** in $\mathbf{y}$. The third term – $\frac{1}{2}\langle \mathbf{y}, \nabla^2\phi(\mathbf{x})\mathbf{y}\rangle$ in (C.4.1) and $\frac{1}{2}\mathbf{D}^2_\phi(\mathbf{x};\mathbf{y})$ in (C.4.2) is **quadratic** in $\mathbf{y}$ where $\mathbf{D}^2_\phi(\mathbf{x};\mathbf{y})$ is defined in (C.3.8). If we exclude the quadratic terms in $\mathbf{y}$ from the right-hand side of (C.4.1) and (C.4.2), the resulting two-term expansion of $\phi(\mathbf{x}+\mathbf{y})$ is called the **first-order** Taylor series. The three-term expansion as given in (C.4.1) and (C.4.2) has come to be known as the **second-order** Taylor series.

C.5 First and second variations

In this section we develop basic properties relating to the notion of **first** and **second variations** of functions. It will be shown that there is an intimate relation between the notion of first/second variation and the first-order/second-order directional derivatives.

Scalar-valued function Let $\phi : \mathbb{R}^n \longrightarrow \mathbb{R}$ and let $\delta\mathbf{x} = (\delta x_1, \delta x_2, \ldots, \delta x_n)^{\mathrm{T}}$ denote a small increment[†] or **variation** in $\mathbf{x} \in \mathbb{R}^n$. Let $\Delta\phi(\mathbf{x}) = \phi(\mathbf{x}+\delta\mathbf{x}) - \phi(\mathbf{x})$ denote the **actual change** or variation that $\phi(\mathbf{x})$ suffers resulting from the variation $\delta\mathbf{x}$ in $\mathbf{x}$, where it is assumed that $\|\delta\mathbf{x}\|$ is small. We can approximate $\Delta\phi(\mathbf{x})$ using the first-order Taylor series as

$$\Delta\phi(\mathbf{x}) \approx \delta\phi(\mathbf{x}) = \langle \delta\mathbf{x}, \nabla\phi(\mathbf{x})\rangle \tag{C.5.1}$$

where $\delta\phi(\mathbf{x})$ is called the **first variation** in ϕ induced by $\delta\mathbf{x}$. That is, the first variation in $\phi(\mathbf{x})$ is defined as the first-order approximation to the actual change, $\Delta\phi$ that ϕ suffers resulting from the variation $\delta\mathbf{x}$ in $\mathbf{x}$. Comparing this with the

[†] A note on the notation is in order. It is customary to use Δ to denote the actual change and δ as the first variation operator. For independent variables $(\Delta x_1, \Delta x_2, \ldots, \Delta x_n)^{\mathrm{T}} = \Delta\mathbf{x} = \delta\mathbf{x} = (\delta x_1, \delta x_2, \ldots, \delta x_n)^{\mathrm{T}}$, but for dependent variables $\delta\phi$ is an approximation for the actual change $\Delta\phi$.

definition of the directional derivative in (C.1.9), it follows that

$$\delta\phi(\mathbf{x}) = \|\delta\mathbf{x}\|\,\phi'\left(\mathbf{x}, \frac{\delta\mathbf{x}}{\|\delta\mathbf{x}\|}\right). \tag{C.5.2}$$

Consequently, the **first variation operator** δ is **linear**. That is, if $\phi_1, \phi_2 : \mathbb{R}^n \longrightarrow \mathbb{R}$, then

$$\delta(a\phi_1(\mathbf{x}) + b\phi_2(\mathbf{x})) = a\delta\phi_1(\mathbf{x}) + b\delta\phi_2(\mathbf{x}) \tag{C.5.3}$$

for any real constants a and b. It can be verified that

$$\left.\begin{aligned} &\delta(a) = 0 \quad \text{where } a \text{ is a constant} \\ \textit{and} \quad &\delta(\phi_1(\mathbf{x})\phi_2(\mathbf{x})) = \phi_1(\mathbf{x})\delta\phi_2(\mathbf{x}) + [\delta\phi_1(\mathbf{x})]\phi_2(\mathbf{x}) \end{aligned}\right\} \tag{C.5.4}$$

The first variation of the first variation is called the **second variation** and is denoted by $\delta^2\phi(\mathbf{x})$. Clearly,

$$\begin{aligned} \delta^2\phi(\mathbf{x}) &= \delta[\delta\phi(\mathbf{x})] \\ &= \delta[\langle\delta\mathbf{x}, \nabla\phi(\mathbf{x})\rangle] \\ &= \langle\delta\mathbf{x}, \delta[\nabla\phi(\mathbf{x})]\rangle \\ &= \textstyle\sum_{i=1}^n \delta x_i \delta\left(\frac{\partial\phi}{\partial x_i}\right) \\ &= \textstyle\sum_{i=1}^n \delta x_i \left[\sum_{j=1}^n \delta x_j \left(\frac{\partial^2\phi}{\partial x_i \partial x_j}\right)\right] \\ &= \textstyle\sum_{i=1}^n \sum_{j=1}^n \delta x_i \delta x_j \left[\frac{\partial^2\phi}{\partial x_i \partial x_j}\right] \\ &= \langle\delta\mathbf{x}, \nabla^2\phi(\mathbf{x})\delta\mathbf{x}\rangle \end{aligned} \tag{C.5.5}$$

where, recall that, $\nabla^2\phi(\mathbf{x})$ is the Hessian of $\phi(\mathbf{x})$. Comparing the expression for the second variation with that of the second-order directional derivative in (C.2.3), it follows that

$$\delta^2\phi(\mathbf{x}) = \|\delta\mathbf{x}\|^2\phi''\left(\mathbf{x}; \frac{\delta\mathbf{x}}{\|\delta\mathbf{x}\|}, \frac{\delta\mathbf{x}}{\|\delta\mathbf{x}\|}\right). \tag{C.5.6}$$

It can be verified that δ^2 is also a linear operator, that is,

$$\delta^2(a\phi_1(\mathbf{x}) + b\phi_2(\mathbf{x})) = a\delta^2\phi_1(\mathbf{x}) + b\delta^2\phi_2(\mathbf{x}) \tag{C.5.7}$$

where a and b are real constants.

Also

$$\delta^2(\phi_1(\mathbf{x})\phi_2(\mathbf{x})) = [\delta^2\phi_1(\mathbf{x})]\phi_2(\mathbf{x}) + 2[\delta\phi_1(\mathbf{x})][\delta\phi_2(\mathbf{x})] + \phi_1(\mathbf{x})[\delta^2\phi_2(\mathbf{x})] \tag{C.5.8}$$

Vector-valued functions Let $\phi : \mathbb{R}^n \longrightarrow \mathbb{R}^m$, where $\phi(\mathbf{x}) = (\phi_1(\mathbf{x}), \phi_2(\mathbf{x}), \ldots, \phi_m(\mathbf{x}))^{\mathrm{T}}$ and $\mathbf{x} \in \mathbb{R}^n$. Then, the first variation $\delta\phi(\mathbf{x})$ is a vector given

Table C.5.1 *Variations of some useful functions*

Function ϕ	First Variation
$\phi(\mathbf{x}) = \langle \mathbf{a}, \mathbf{x} \rangle$ where $\mathbf{a}, \mathbf{x} \in \mathbb{R}^n$	$\delta\phi = \langle \mathbf{a}, \delta\mathbf{x} \rangle$
$\phi(\mathbf{x}) = \frac{1}{2}\langle \mathbf{x}, \mathbf{A}\mathbf{x} \rangle = \frac{1}{2}\mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x}$, where $\mathbf{A} \in \mathbb{R}^{n\times n}$ is a symmetric and positive definite matrix.	$\delta\phi = \langle \delta\mathbf{x}, \mathbf{A}\mathbf{x} \rangle$
$\phi(\mathbf{x}) = \frac{1}{2}(\mathbf{z} - \mathbf{H}\mathbf{x})^{\mathrm{T}}(\mathbf{z} - \mathbf{H}\mathbf{x})$ where $\mathbf{z} \in \mathbb{R}^m$, $\mathbf{x} \in \mathbb{R}^n$ and $\mathbf{H} \in \mathbb{R}^{m\times n}$.	$\delta\phi(\mathbf{x}) = \langle \mathbf{H}^{\mathrm{T}}(\mathbf{H}\mathbf{x} - \mathbf{z}),\ \delta\mathbf{x} \rangle$.

by

$$\delta\phi = \begin{pmatrix} \delta\phi_1 \\ \delta\phi_2 \\ \vdots \\ \delta\phi_m \end{pmatrix} = \begin{pmatrix} \langle \delta\mathbf{x},\ \nabla\phi_1(\mathbf{x}) \rangle \\ \langle \delta\mathbf{x},\ \nabla\phi_2(\mathbf{x}) \rangle \\ \vdots \\ \langle \delta\mathbf{x},\ \nabla\phi_m(\mathbf{x}) \rangle \end{pmatrix} = \mathbf{D}_\phi(\mathbf{x})\delta\mathbf{x} \tag{C.5.9}$$

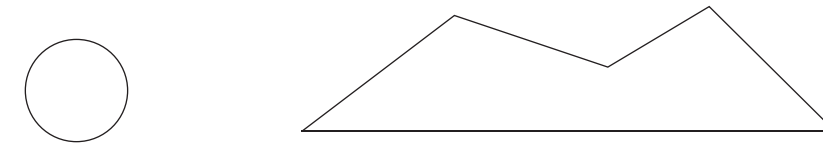
where $\mathbf{D}_\phi \in \mathbb{R}^{m\times n}$ is the Jacobian of ϕ. Using (C.5.5), the second variation of ϕ is given by

$$\begin{aligned} \delta^2\phi(\mathbf{x}) = \delta[\delta\phi(\mathbf{x})] &= \begin{pmatrix} \delta[\delta\phi_1(\mathbf{x})] \\ \delta[\delta\phi_2(\mathbf{x})] \\ \vdots \\ \delta[\delta\phi_m(\mathbf{x})] \end{pmatrix} \\ &= \begin{pmatrix} \langle \delta\mathbf{x},\ [\nabla^2\phi_1(\mathbf{x})]\delta(\mathbf{x}) \rangle \\ \langle \delta\mathbf{x},\ [\nabla^2\phi_2(\mathbf{x})]\delta(\mathbf{x}) \rangle \\ \vdots \\ \langle \delta\mathbf{x},\ [\nabla^2\phi_m(\mathbf{x})]\delta(\mathbf{x}) \rangle \end{pmatrix}. \end{aligned} \tag{C.5.10}$$

Comparing this for the second variation with the expressions for the second-order directional derivatives in (C.3.6) – (C.3.8), it readily follows that

$$\delta^2\phi(\mathbf{x}) = \|\delta\mathbf{x}\|^2 \begin{pmatrix} \langle \mathbf{z},\ \nabla^2\phi_1(\mathbf{x})\mathbf{z} \rangle \\ \langle \mathbf{z},\ \nabla^2\phi_2(\mathbf{x})\mathbf{z} \rangle \\ \vdots \\ \langle \mathbf{z},\ \nabla^2\phi_m(\mathbf{x})\mathbf{z} \rangle \end{pmatrix} = \|\delta\mathbf{x}\|^2 \mathbf{D}^2_\phi(\mathbf{x}; \mathbf{z}) \tag{C.5.11}$$

where the unit vector $\mathbf{z} = \delta\mathbf{x}/\|\delta\mathbf{x}\|$. Variations of several functions of interest are given in Table C.5.1.



(a) Convex set (b) Non-convex set

Fig. C.6.1 Examples of convex and non-convex sets.

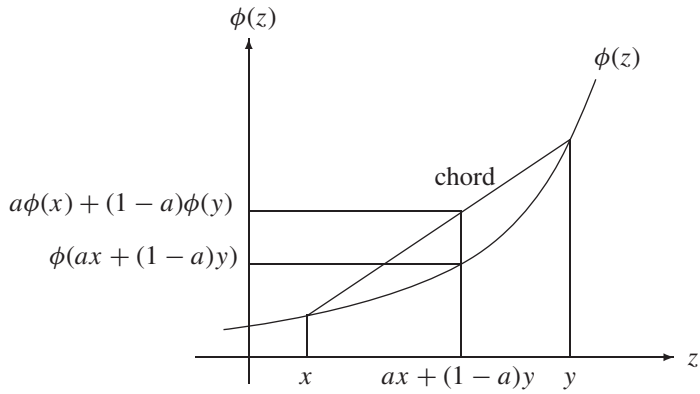


Fig. C.6.2 An Example of a convex function.

C.6 Convex functions

Let $\mathcal{S}$ be a subset of $\mathbb{R}^n$. Then $\mathcal{S}$ is said to be a **convex set** if for every pair of points $\mathbf{x}$ and $\mathbf{y}$ in $\mathcal{S}$

$$a\mathbf{x} + (1-a)\mathbf{y} \in \mathcal{S} \quad \text{for all } a \in [0,1]. \tag{C.6.1}$$

That is, the line segment connecting $\mathbf{x}$ and $\mathbf{y}$ lies entirely in $\mathcal{S}$. Refer to Figure C.6.1 for examples of convex and non-convex sets in $\mathbb{R}^2$.

A function $\phi : \mathcal{S} \longrightarrow \mathbb{R}$ is said to be a **convex function** if

$$\phi(a\mathbf{x} + (1-a)\mathbf{y}) \leq a\phi(\mathbf{x}) + (1-a)\phi(\mathbf{y}) \tag{C.6.2}$$

for all $a \in [0,1]$ and for every pair of points $\mathbf{x}, \mathbf{y} \in \mathcal{S}$. The function ϕ is said to be **strictly convex**, if (C.6.2) holds with strict inequality for all $a \in [0,1]$ and for every pair of points $\mathbf{x}$ and $\mathbf{y} \in \mathcal{S}$. An example of a convex function is given in Figure C.6.2.

It is clear from the definition that the function lies below the **chord** joining the points $(\mathbf{x}, \phi(\mathbf{x}))$ and $(\mathbf{y}, \phi(\mathbf{y}))$. A function ϕ is said to be **concave** if $-\phi$ is convex. Thus, $\phi(x) = x^2$ is convex, and $\phi(x) = -x^2$ is concave. But $\phi(x) = x^3$ is neither convex nor concave.

We now state several properties without proof.

(a) Let $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m$ be a set of m points lying in a convex set $\mathcal{S} \in \mathbb{R}^n$. Let $a_1, a_2, \ldots, a_m$ be a set of non-negative real numbers such that $\sum_{i=1}^m a_i = 1$. If $\phi : \mathcal{S} \longrightarrow \mathbb{R}$ is a convex function, then

$$\phi\left(\sum_{i=1}^m a_i \mathbf{x}_i\right) \leq \sum_{i=1}^m a_i \phi(\mathbf{x}_i). \tag{C.6.3}$$

(b) A linear function is convex.

(c) A weighted sum of convex functions with positive weights is convex.

(d) Let $\phi : \mathbb{R}^n \longrightarrow \mathbb{R}^m$ given by $\phi(\mathbf{x}) = (\phi_1(\mathbf{x}), \phi_2(\mathbf{x}), \ldots, \phi_m(\mathbf{x}))^{\mathrm{T}}$. Let $\mathbf{a} = (a_1, a_2, \ldots, a_m)^{\mathrm{T}}$, be any real vector, and define

$$\mathcal{S} = \left\{\mathbf{x} \in \mathbb{R}^n | \phi_i(\mathbf{x}) \leq a_i \text{ for } 1 \leq i \leq m\right\}. \tag{C.6.4}$$

Then, $\mathcal{S}$ is a convex set if each $\phi_i(\mathbf{x})$ is a convex function for $1 \leq i \leq m$.

(e) Intersection of convex sets is a convex set.

(f) Let ϕ and $\mathbf{a}$ be as defined in Property (4) above. Then,

$$\mathcal{S} = \left\{\mathbf{x} \in \mathbb{R}^n | \phi_i(\mathbf{x}) = a_i \text{ for } 1 \leq i \leq m\right\}. \tag{C.6.5}$$

is a convex set if each $\phi_i(\mathbf{x})$ is a linear function.

(g) **Condition for global minimum** Let ϕ be a convex function over a convex set $\mathcal{S}$. Then, ϕ has a unique minimum.

To verify this, assume the contrary. Let $\mathbf{z}$ be a global minimum and $\mathbf{y}$ be a local minimum, that is $\phi(\mathbf{z}) < \phi(\mathbf{y})$. Since both $\mathcal{S}$ and ϕ are convex, we have, for any $a \in [0, 1]$

$$\begin{aligned}\phi(a\mathbf{z} + (1-a)\mathbf{y}) &\leq a\phi(\mathbf{z}) + (1-a)\phi(\mathbf{y}) \\ &< a\phi(\mathbf{z}) + (1-a)\phi(\mathbf{z}) = \phi(\mathbf{y}).\end{aligned}$$

Now, given any $\epsilon > 0$, we can choose a such that $(a\mathbf{z} + (1-a)\mathbf{y})$ is at distance ϵ from $\mathbf{y}$. Thus, there is a point close to $\mathbf{y}$ where the function value is less than $\phi(\mathbf{y})$. This contradicts the assumption that ϕ attains a local minimum at $\mathbf{y}$. Hence, the uniqueness of the minimum.

(h) If $\phi : \mathbb{R}^n \longrightarrow \mathbb{R}$ is convex, then it is continuous.

(i) Let $\mathbf{I}$ be an interval of the real line, and let $\phi : \mathbf{I} \longrightarrow \mathbb{R}$ be a convex function. If $x < y < z$, then

$$\frac{\phi(y) - \phi(x)}{y - x} \leq \frac{\phi(z) - \phi(x)}{z - x} \leq \frac{\phi(z) - \phi(y)}{z - y}.$$

(j) Let $\mathcal{S}$ be a convex set in $\mathbb{R}^n$, and $\phi : \mathcal{S} \longrightarrow \mathbb{R}$, be a continuously differentiable function. Then, ϕ is convex, if and only if

$$\phi(\mathbf{y}) \geq \phi(\mathbf{x}) + (\mathbf{y} - \mathbf{x})^{\mathrm{T}} \nabla\phi(\mathbf{x}) \tag{C.6.6}$$

for all $\mathbf{x}, \mathbf{y} \in \mathcal{S}$.

(k) If strict inequality holds in (C.6.6), then ϕ is strictly convex.

(l) Let $\phi : \mathbb{R}^n \longrightarrow \mathbb{R}$ be a convex function. Then $\nabla\phi(\mathbf{x}) = 0$, if and only if, $\mathbf{x}$ is a global minimum.

(m) Let $\phi : \mathbb{R}^n \longrightarrow \mathbb{R}$ be twice continuously differentiable. Then
   (a) $\phi(\mathbf{x})$ is convex if the Hessian $\nabla^2\phi(\mathbf{x})$ is non-negative definite for all $\mathbf{x}$.
   (b) $\phi(\mathbf{x})$ is strictly convex if $\nabla^2\phi(\mathbf{x})$ is positive definite.

(n) Let $\mathbf{A}$ be a real symmetric matrix of order n, and let $\phi(\mathbf{x}) = \mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x}$. Then $\phi(\mathbf{x})$ is strictly convex if and only if $\mathbf{A}$ is positive definite. Since $\mathbf{I}$, the identity matrix is positive definite, clearly $\phi(\mathbf{x}) = \mathbf{x}^{\mathrm{T}}\mathbf{I}\mathbf{x} = \mathbf{x}^{\mathrm{T}}\mathbf{x} = \|\mathbf{x}\|_2^2$ is strictly convex. It can be verified that any vector norm is convex.

C.7 Function of a matrix

Matrix-valued function of a scalar

Let $x \in \mathbb{R}$. For each pair of indices ij where $1 \leq i,\ j \leq n$ define $F_{ij} : \mathbb{R} \to \mathbb{R}$ and define a matrix $\mathbf{F}(x) = [F_{ij}(x)] \in \mathbb{R}^{n\times n}$ of these functions. The derivation of $\mathbf{F}(x)$ with respect to x is an $n \times n$ matrix and is given by

$$\frac{\mathrm{d}\mathbf{F}(x)}{\mathrm{d}x} = \left[\frac{\mathrm{d}\mathbf{F}_{ij}(x)}{dx}\right]. \tag{C.7.1}$$

As an example, let $n = 2$ and

$$\mathbf{F} = \begin{bmatrix} 1 + x^2 & 3 - x^3 \\ x^3 & x^4 - x^3 \end{bmatrix}.$$

Then

$$\frac{\mathrm{d}\mathbf{F}}{\mathrm{d}x} = \begin{bmatrix} 2x & -3x^2 \\ 3x^2 & 4x^3 - 3x^2 \end{bmatrix}.$$

Scalar-valued function of a matrix

Let $\mathbf{X} = [x_{ij}] \in \mathbb{R}^{n\times n}$. Define $F : \mathbb{R}^{n\times n} \to \mathbb{R}$ be the scalar-valued function of the elements of the matrix $\mathbf{X}$. Examples of such functions include the trace and determinant of matrices. The derivative of F with respect to the matrix $\mathbf{X}$ denoted by $\partial F/\partial\mathbf{X}$ is a matrix given by

$$\frac{\partial F}{\partial \mathbf{X}} = \left[\frac{\partial F}{\partial x_{ij}}\right]. \tag{C.7.2}$$

For example, let $n = 2$ and

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \end{bmatrix}.$$

Consider

$$F(\mathbf{X}) = x_{11}^2 + x_{22}^2 + x_{11}x_{22} - x_{12}x_{21}.$$

Then

$$\frac{\partial F(\mathbf{X})}{\partial \mathbf{X}} = \begin{bmatrix} 2x_{11} & 0 \\ 0 & 2x_{22} \end{bmatrix} + \begin{bmatrix} x_{22} & -x_{21} \\ -x_{12} & x_{11} \end{bmatrix}.$$

Matrix-valued function of a matrix

Let $\mathbf{X} \in \mathbb{R}^{n\times n}$ and $\mathbf{F} : \mathbb{R}^{n\times n} \to \mathbb{R}^{n\times n}$ where $\mathbf{F}(\mathbf{X}) = [F_{ij}(\mathbf{X})]$ and for each ij, $F_{ij} : \mathbb{R}^{n\times n} \to \mathbb{R}$ is a scalar valued function of a matrix $\mathbf{X}$. In our development we are specifically interested in special function of the matrix $\mathbf{F}(\mathbf{X})$. Recall that the trace of $\mathbf{F}(\mathbf{X})$, denoted by $\text{tr}[\mathbf{F}(\mathbf{X})]$ is given by

$$\text{tr}[\mathbf{F}(\mathbf{X})] = \sum_{i=1}^{n} F_{ii}(\mathbf{X}) \tag{C.7.3}$$

Then, the derivative of this scalar valued function of a matrix is given by

$$\frac{\partial\, \text{tr}[\mathbf{F}(\mathbf{X})]}{\partial \mathbf{X}} = \frac{\partial}{\partial \mathbf{X}} \left[\sum_{i=1}^{n} F_{ii}(\mathbf{X}) \right] = \sum_{i=1}^{n} \frac{\partial F_{ii}(\mathbf{X})}{\partial \mathbf{X}} \tag{C.7.4}$$

which is clearly the sum of n matrices.

We now illustrate the computation of these derivatives for special cases of interest to us.

(a) Let $\mathbf{A} \in \mathbb{R}^{n\times n}$, $\mathbf{X} \in \mathbb{R}^{n\times n}$ and $\mathbf{F}(\mathbf{X}) = \mathbf{A}\mathbf{X}$. Then from

$$\mathbf{F}(\mathbf{X}) = \mathbf{A}\mathbf{X} = \begin{bmatrix} \mathbf{A}_{1*} \\ \mathbf{A}_{2*} \\ \vdots \\ \mathbf{A}_{n*} \end{bmatrix} \begin{bmatrix} \mathbf{X}_{*1} & \mathbf{X}_{*2} & \cdots & \mathbf{X}_{*n} \end{bmatrix}$$

we obtain that the

$$\text{tr}[\mathbf{F}(\mathbf{X})] = \sum_{k=0}^{n-1} \mathbf{A}_{k*}\, \mathbf{X}_{*k}. \tag{C.7.5}$$

Thus, the gradient of the scalar $\mathbf{A}_{k*}\, \mathbf{X}_{*k}$ w.r.t. the column vector $\mathbf{X}_{*k}$ is given by

$$\nabla[\mathbf{A}_{k*}\, \mathbf{X}_{*k}] = \mathbf{A}_{k*}^{\text{T}}$$

Now combining each of these columns in a matrix we readily obtain

$$\frac{\partial\, \text{tr}(\mathbf{X})}{\partial \mathbf{X}} = \begin{bmatrix} \mathbf{A}_{1*}^{\text{T}} & \mathbf{A}_{2*}^{\text{T}} & \cdots & \mathbf{A}_{n*}^{\text{T}} \end{bmatrix}. \tag{C.7.6}$$

Remark C.7.1 Since $\text{tr}(\mathbf{A}) = \text{tr}(\mathbf{A}^{\text{T}})$ and $\text{tr}(\mathbf{A}\,\mathbf{B}) = \text{tr}(\mathbf{B}\,\mathbf{A})$, we can obtain a variety of interesting corollaries.

$$\begin{aligned}\frac{\partial\,\text{tr}(\mathbf{AX})}{\partial\mathbf{X}} = \mathbf{A}^{\text{T}} &= \frac{\partial\,\text{tr}(\mathbf{XA})}{\partial\mathbf{X}} \\ &= \frac{\partial\,\text{tr}(\mathbf{A}^{\text{T}}\mathbf{X}^{\text{T}})}{\partial\mathbf{X}} = \frac{\partial\,\text{tr}(\mathbf{X}^{\text{T}}\mathbf{A}^{\text{T}})}{\partial\mathbf{X}}\end{aligned} \tag{C.7.7}$$

and

$$\frac{\partial\,\text{tr}(\mathbf{A}^{\text{T}}\mathbf{X})}{\partial\mathbf{X}} = \mathbf{A} = \frac{\partial\,\text{tr}(\mathbf{XA}^{\text{T}})}{\partial\mathbf{X}} = \frac{\partial\,\text{tr}(\mathbf{AX}^{\text{T}})}{\partial\mathbf{X}} = \frac{\partial\,\text{tr}(\mathbf{X}^{\text{T}}\mathbf{A})}{\partial\mathbf{X}}. \tag{C.7.8}$$

Similarly, it can be verified that

$$\frac{\partial\,\text{tr}[\mathbf{ABX}]}{\partial\mathbf{X}} = \mathbf{B}^{\text{T}}\mathbf{A}^{\text{T}} = \frac{\partial\,\text{tr}[\mathbf{XAB}]}{\partial\mathbf{X}} = \frac{\partial\,\text{tr}[\mathbf{BXA}]}{\partial\mathbf{X}}. \tag{C.7.9}$$

(b) Let $\mathbf{A} \in \mathbb{R}^{n\times n}$, $\mathbf{X} \in \mathbb{R}^{n\times n}$ and $\mathbf{F}(\mathbf{X}) = \mathbf{X}^{\text{T}}\mathbf{AX}$. Then

$$\mathbf{F}(\mathbf{X}) = \begin{bmatrix}\mathbf{X}_{*1}^{\text{T}} \\ \mathbf{X}_{*2}^{\text{T}} \\ \vdots \\ \mathbf{X}_{*n}^{\text{T}}\end{bmatrix} \mathbf{A}\left[\mathbf{X}_{*1}\ \mathbf{X}_{*2}\ \cdots\ \mathbf{X}_{*n}\right]$$

and

$$\text{tr}[\mathbf{F}(\mathbf{X})] = \sum_{k=0}^{n-1} \mathbf{X}_{*k}^{\text{T}}\,\mathbf{AX}_{*k}. \tag{C.7.10}$$

The gradient of the scalar $\mathbf{X}_{*k}^{\text{T}}\mathbf{AX}_{*k}$ w.r.t. the column vector $\mathbf{X}_{*k}$ is given by

$$\nabla[\mathbf{X}_{*k}^{\text{T}}\,\mathbf{A}\,\mathbf{X}_{*k}] = 2\mathbf{AX}_{*k}. \tag{C.7.11}$$

Again combining these columns in a matrix we obtain

$$\begin{aligned}\frac{\partial\,\text{tr}(\mathbf{X}^{\text{T}}\mathbf{AX})}{\partial\mathbf{X}} &= 2\mathbf{A}\left[\mathbf{X}_{*1}\ \mathbf{X}_{*2}\ \cdots\ \mathbf{X}_{*n}\right] \\ &= 2\mathbf{AX}.\end{aligned} \tag{C.7.12}$$

The following formulae readily follow from this:

$$\begin{aligned}\frac{\partial\,\text{tr}[\mathbf{XX}^{\text{T}}\mathbf{A}]}{\partial\mathbf{X}} = 2\mathbf{AX} &= \frac{\partial\,\text{tr}[\mathbf{AXX}^{\text{T}}]}{\partial\mathbf{X}} \\ &= \frac{\partial\,\text{tr}[\mathbf{X}^{\text{T}}\mathbf{A}^{\text{T}}\mathbf{X}]}{\partial\mathbf{X}} = \frac{\partial\,\text{tr}[\mathbf{XX}^{\text{T}}\mathbf{A}]}{\partial\mathbf{X}} = \frac{\partial\,\text{tr}[\mathbf{AXX}^{\text{T}}]}{\partial\mathbf{X}}.\end{aligned} \tag{C.7.13}$$

(c) Let $\mathbf{A} \in \mathbb{R}^{n\times n}$, $\mathbf{X} \in \mathbb{R}^{n\times n}$ and $\mathbf{F}(\mathbf{X}) = \mathbf{X}\mathbf{A}\mathbf{X}^{\mathrm{T}}$. Then

$$\mathbf{F}(\mathbf{X}) = \begin{bmatrix} \mathbf{X}_{1*} \\ \mathbf{X}_{2*} \\ \vdots \\ \mathbf{X}_{n*} \end{bmatrix} \mathbf{A} \begin{bmatrix} \mathbf{X}_{1*}^{\mathrm{T}} & \mathbf{X}_{2*}^{\mathrm{T}} & \cdots & \mathbf{X}_{n*}^{\mathrm{T}} \end{bmatrix}$$

and

$$\mathrm{tr}[\mathbf{F}(\mathbf{X})] = \sum_{k=0}^{n-1} \mathbf{X}_{k*}\mathbf{A}\mathbf{X}_{k*}^{\mathrm{T}} \tag{C.7.14}$$

from which we obtain the gradient of the kth term on the right-hand side w.r.t. $\mathbf{X}_{k*}^{\mathrm{T}}$ as

$$\nabla[\mathbf{X}_{k*}\,\mathbf{A}\mathbf{X}_{k*}^{\mathrm{T}}] = 2\,\mathbf{A}\mathbf{X}_{k*}^{\mathrm{T}}. \tag{C.7.15}$$

In the matrix of the derivative of $\mathrm{tr}[\mathbf{X}\mathbf{A}\mathbf{X}^{\mathrm{T}}]$ w.r.t. $\mathbf{X}$, recall that the derivative with respect to a row vector X_{k*} will appear as a row. Thus, the kth row of the derivative we are seeking is given by the transpose of the r.h.s. of (C.7.15), namely $2\mathbf{X}_{k*}\,\mathbf{A}^{\mathrm{T}}$. Combining all these rows, we get

$$\frac{\partial\,\mathrm{tr}[\mathbf{X}\mathbf{X}^{\mathrm{T}}\mathbf{A}]}{\partial\mathbf{X}} = 2\begin{bmatrix} \mathbf{X}_{1*} \\ \mathbf{X}_{2*} \\ \vdots \\ \mathbf{X}_{n*} \end{bmatrix} \mathbf{A}^{\mathrm{T}} = 2\,\mathbf{X}\mathbf{A}^{\mathrm{T}}. \tag{C.7.16}$$

Notes and references

There are numerous books on multivariate calculus and we mention only two – Apostol (1957) and Sikorski (1969). Rockafellar (1970) is an excellent reference for convexity.