CRITICAL POINTS

TSOGTGEREL GANTUMUR

Abstract. Let u be a continuously differentiable function. Then we know from the level surface theorem that near any point x with $\nabla u(x) \neq 0$, the level sets of u are smooth surfaces. Hence in a certain sense, "interesting things" happen only at or near the *critical points*, that is, the points x_* such that $\nabla u(x_*) = 0$. In fact, in most situations, understanding a function can be identified with locating its critical points and studying its behaviour near the critical points. Critical points also arise in optimization problems, where one is tasked to investigate maximum and minimum values of a function.

Contents

1. Univariate functions 1
2. The gradient test 3
3. The Weierstrass existence theorem 6
4. Lagrange multipliers on surfaces 10
5. Lagrange multipliers on curves 12
6. Diagonalization of symmetric matrices 14
7. Second order derivatives 17
8. The Hessian test 20

1. Univariate functions

In this section, we collect here some results on critical points of univariate functions, that have generalization to higher dimensions. Note that there are many results special to 1 dimension, such as Rolle's theorem, which we omit.

The following fundamental theorem was established by Weierstrass in 1861.

Theorem 1.1 (Extreme value)**.** *Let $f : [a, b] \to \mathbb{R}$ be a continuous function. Then there exists $c \in [a, b]$ such that $f(x) \leq f(c)$ for all $x \in [a, b]$.*

The point c with the property given by the preceding theorem is called a *maximizer* of f over $[a, b]$, and the value $f(c)$ is called a *maximum value*, or simply a *maximum*.

Example 1.2. It is important that we have an interval of the form $[a, b]$ in Theorem 1.1. For example, the function $f(x) = x$ does not have any maximizers in the interval $[0, 1)$. Moreover, a discontinuous function may not have maximizers even over an interval of the form $[a, b]$. An example would be

$$f(x) = \begin{cases} x, & \text{for } 0 \leq x < 1, \\ 0, & \text{for } x = 1, \end{cases}$$

over the interval $[0, 1]$.

The following result is called the *first derivative test*, or *Fermat's theorem*.

Date: 2019/10/19.

Theorem 1.3. *Let $f : (a, b) \to \mathbb{R}$ be a function, and let $c \in (a, b)$ be a maximizer of f, in the sense that $f(x) \leq f(c)$ for all $x \in (a, b)$. Suppose that f is differentiable at c. Then $f'(c) = 0$.*

Remark 1.4. At least in principle, the first derivative test gives a way to find the maximizers and minimizers of a differentiable function. Namely, let a continuous function $f : [a, b] \to \mathbb{R}$ be given. Then the extreme value theorem (Theorem 1.1) guarantees the existence of a maximizer $\xi \in [a, b]$. If $\xi \in (a, b)$ and if f is differentiable in (a, b), then $f'(\xi) = 0$. In other words, all maximizers located in the interior (a, b) can be found by comparing the values $f(c)$ at the *critical points*, which are by definition the solutions $c \in (a, b)$ of the equation $f'(c) = 0$.

Example 1.5. Consider $f(x) = (x - 1)^2$ in the interval $[0, 3]$. By the Weierstrass theorem, there exists a maximizer (and a minimizer) of f in $[0, 3]$. By the first derivative test, any *interior* maximizer (and minimizer) must satisfy $f'(x) = 2(x - 1) = 0$, that is, $x = 1$. Now we simply list the values of f at $x = 1$ and at the boundary points of $[0, 3]$, as

$$f(0) = 1, \qquad f(1) = 0, \qquad f(3) = 4,$$

from which it is clear that the maximum of f over $[0, 3]$ occurs at the boundary point $x = 3$, and the minimum of f over $[0, 3]$ occurs at the interior critical point $x = 1$.

If $f : (a, b) \to \mathbb{R}$ is a function differentiable in (a, b), then the derivative $g = f'$ is a function $g : (a, b) \to \mathbb{R}$. Hence it makes sense to talk about differentiability of f', which leads to the notion of higher order derivatives.

Definition 1.6. We say that $f : (a, b) \to \mathbb{R}$ is *twice differentiable* at $x \in (a, b)$, if there exists $\varepsilon > 0$ such that f is differentiable in $(x - \varepsilon, x + \varepsilon)$, and if f' is differentiable at x. We call

$$f''(x) \equiv \frac{\mathrm{d}^2 f}{\mathrm{d}x^2}(x) = g'(x), \tag{1}$$

the *second order derivative* of f at x.

Example 1.7. For $f(x) = x^3$, we have $f'(x) = 3x^2$ and $f''(x) = 6x$. Hence f is twice differentiable at each $x \in \mathbb{R}$, i.e., it is twice differentiable in $\mathbb{R}$.

Differentiable functions are well approximated locally by linear functions. Intuitively speaking, twice differentiable functions should be close to quadratic functions.

Theorem 1.8. *a) If $f : (a, b) \to \mathbb{R}$ is twice differentiable at $c \in (a, b)$, then there is a function $\psi : (a, b) \to \mathbb{R}$ that is continuous at c with $\psi(c) = 0$, such that*

$$f(x) = f(c) + f'(c)(x - c) + \frac{f''(c)}{2}(x - c)^2 + \psi(x)(x - c)^2, \qquad x \in (a, b). \tag{2}$$

b) If $f : (a, b) \to \mathbb{R}$ is twice differentiable in (a, b), and $y, c \in (a, b)$, then there exists $\xi \in (y, c) \cup (c, y)$, such that

$$f(x) = f(c) + f'(c)(x - c) + \frac{f''(\xi)}{2}(x - c)^2. \tag{3}$$

Example 1.9. The quadratic approximation of $f(x) = e^x$ based at $x = 0$ is

$$e^x \approx 1 + x + \frac{x^2}{2}. \tag{4}$$

The following result is known as the *second derivative test*.

Theorem 1.10. *Let $f : (a, b) \to \mathbb{R}$ be a function, and suppose that f is twice differentiable at $c \in (a, b)$. Then the following are true.*

a) If c is a maximizer of f over the interval (a, b), in the sense that $f(x) \leq f(c)$ for all $x \in (a, b)$, then $f'(c) = 0$ and $f''(c) \leq 0$.

b) If $f'(c) = 0$ and $f''(c) < 0$, then there exists $\varepsilon > 0$ such that $f(c) > f(x)$ for all $x \in (c-\varepsilon, c+\varepsilon)$.

If there exists $\varepsilon > 0$ such that c is a maximizer of f over $(c-\varepsilon, c+\varepsilon)$, then c is called a *local maximizer* of f. Moreover, if $f(c) > f(x)$ for all $x \in (c-\varepsilon, c+\varepsilon)$ then c is called a *strict maximizer* of f over $(c-\varepsilon, c+\varepsilon)$. Thus, in b) of the preceding theorem, we may say that c is a local strict maximizer of f.

Example 1.11. The function $f(x) = \cos x$ has local strict maximizers at $x = 2\pi n$, $n \in \mathbb{Z}$, and local strict minimizers at $x = \pi + 2\pi n$, $n \in \mathbb{Z}$.

2. The gradient test

Let $K \subset \mathbb{R}^n$, and let $u : K \to \mathbb{R}$. We say that u has a (global) *maximum* at $y \in K$ if $u(x) \leq u(y)$ for all $x \in K$. In this situation, the point y is called a *maximizer* of u, while the value $u(y)$ is called a *maximum value*, or simply a *maximum*.

Moreover, we say that u has a *local maximum* at $y \in K$ if there exists $\delta > 0$ such that $u(x) \leq u(y)$ for all $x \in Q_\delta(y) \cap K$. Following the preceding pattern, the point y is called a *local maximizer* of u, while the value $u(y)$ is called a *local maximum value*, or simply a *local maximum*. All these notions have their equivalents with regard to *minimums*.

The following result is called the *gradient test*, or *Fermat's theorem*.

Theorem 2.1. *Let $U \subset \mathbb{R}^n$ be open, and let $u : U \to \mathbb{R}$ be differentiable. If u has a local maximum at $y \in U$, then we have*

$$\nabla u(y) = 0. \tag{5}$$

Proof. We will show that $D_V u(y) = 0$ for any direction $V \in \mathbb{R}^n$, and since V is arbitrary, it would imply that $\nabla u(y) = 0$.

Let $V \in \mathbb{R}^n$, and let $\gamma : (-\varepsilon, \varepsilon) \to \mathbb{R}^n$ be a differentiable curve satisfying $\gamma(0) = y$ and $\gamma'(0) = V$. Put $g(t) = u(\gamma(t))$, and note that $g'(0) = D_V u(y)$. Anticipating a contradiction, suppose that $g'(0) > 0$. By definition, we have $g(t) = g(0) + h(t)t$ with $h(t) \to g'(0)$ as $t \to 0$. Hence by continuity, there exists $t > 0$ arbitrarily small, such that $h(t) > \frac{1}{2}g'(0) > 0$. This gives $g(t) > g(0) + \frac{1}{2}g'(0)t$, meaning that $g(0)$ cannot be a local maximum. The case $g'(0) < 0$ can be treated similarly, by considering values $g(t)$ with small $t < 0$. □

Definition 2.2. With u as in the preceding theorem, if $\nabla u(x) = 0$, then x is called a *critical point* of u, and the value $u(x) \in \mathbb{R}$ is called a *critical value* of u.

Example 2.3. Let $f(x, y) = 5x - 7y + 4xy - 7x^2 + 4y^2$ be a function defined in the unit square $0 \leq x \leq 1$, $0 \leq y \leq 1$. Let us try to find the maximum and minimum values of f and where they occur.

Suppose that there exist a maximizer and a minimizer in the (closed) square

$$\bar{Q} = \{(x, y) : 0 \leq x \leq 1,\, 0 \leq y \leq 1\}. \tag{6}$$

This supposition will be justified in the next section. If there is a maximizer (or a minimizer) in the interior

$$Q = \{(x, y) : 0 < x < 1,\, 0 < y < 1\},$$

then by Theorem 2.1, it must be a critical point. From the equations

$$\begin{aligned} \frac{\partial}{\partial x} f(x, y) &= 5 + 4y - 14x = 0, \\ \frac{\partial}{\partial y} f(x, y) &= -7 + 4x + 8y = 0, \end{aligned} \tag{7}$$

it follows that $(x^*, y^*) = (\frac{17}{32}, \frac{39}{64})$ is the only critical point of f in Q. For later reference, let us compute

$$f(x^*, y^*) = f(\tfrac{17}{32}, \tfrac{39}{64}) = -\tfrac{103}{128}. \tag{8}$$

Let us also compute the values of f at the four corners of the square $\bar{Q}$:

$$f(0,0) = 0, \qquad f(0,1) = -3, \qquad f(1,0) = -2, \qquad f(1,1) = -1. \tag{9}$$

It remains to check the four sides of $\bar{Q}$. At the bottom edge $\ell_1 = \{0 < x < 1,\, y = 0\}$, the values of f are recorded in the single variable function

$$g_1(x) = f(x, 0) = 5x - 7x^2. \tag{10}$$

It is easy to find the critical point $x_1^* = \frac{5}{14}$, which gives

$$f(x_1^*, 0) = g_1(x_1^*) = \tfrac{25}{28}. \tag{11}$$

At the top edge $\ell_2 = \{0 < x < 1,\, y = 1\}$, we have

$$g_2(x) = f(x, 1) = 5x - 7 + 4x - 7x^2 + 4 = -7x^2 + 9x - 3, \tag{12}$$

whose critical point is $x_2^* = \frac{9}{14}$, with the corresponding value

$$f(x_2^*, 1) = g_2(x_2^*) = -\tfrac{3}{28}. \tag{13}$$

As for the left edge $\ell_3 = \{x = 0,\, 0 < y < 1\}$, we have

$$g_3(y) = f(0, y) = -7y + 4y^2. \tag{14}$$

The critical point is $y_3^* = \frac{7}{8}$, and the function value is

$$f(0, y_3^*) = g_3(y_3^*) = -\tfrac{49}{16}. \tag{15}$$

Finally, at the right edge $\ell_4 = \{x = 1,\, 0 < y < 1\}$, the values of f are

$$g_4(y) = f(1, y) = 5 - 7y + 4y - 7 + 4y^2 = 4y^2 - 3y - 2. \tag{16}$$

The only critical point of g_4 is $y_4^* = \frac{3}{8}$, with the value

$$f(1, y_4^*) = g_4(y_4^*) = -\tfrac{41}{16}. \tag{17}$$

Now, by comparing the values (8), (9), (11), (13), (15), and (17), we conclude that

- If f has a maximum over $\bar{Q}$, then the maximum value of f over $\bar{Q}$ is $\frac{25}{28}$, which occurs at $(x_1^*, 0) = (\frac{5}{14}, 0)$.
- If f has a minimum over $\bar{Q}$, then the minimum value of f over $\bar{Q}$ is $-\frac{49}{16}$, which occurs at $(0, y_3^*) = (0, \frac{7}{8})$.

Example 2.4. Let us try to find the maximum and minimum values of the function

$$f(x, y) = 5x^2 - 22xy + 5y^2 + 8,$$

in the disk $x^2 + y^2 \le 25$.

Suppose that there exist a maximizer and a minimizer in the (closed) disk

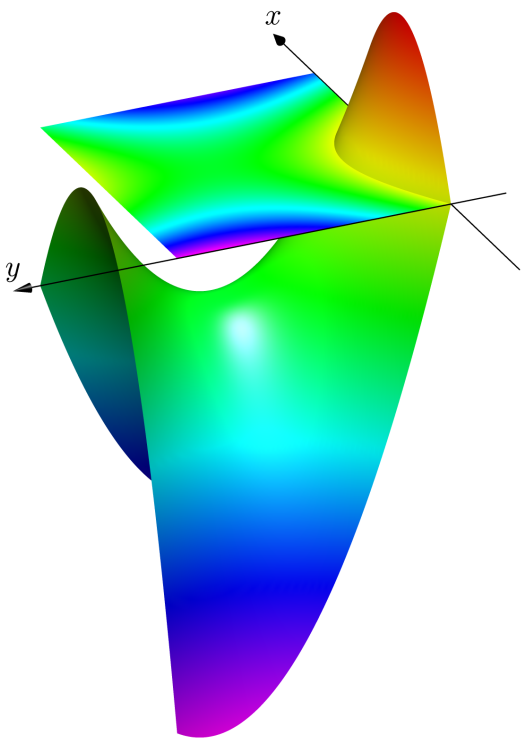
$$\bar{D} = \{(x, y) : x^2 + y^2 \le 25\}. \tag{18}$$

This supposition will be justified in the next section. If there is a maximizer (or a minimizer) in the interior

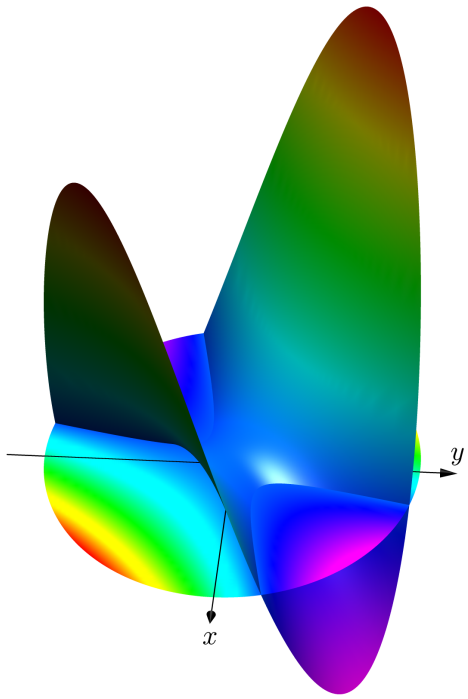
$$D = \{(x, y) : x^2 + y^2 < 25\},$$

then by Theorem 2.1, it must be a critical point. From the equations

$$\frac{\partial}{\partial x} f(x, y) = 10x - 22y = 0, \qquad \frac{\partial}{\partial y} f(x, y) = -22x + 10y = 0, \tag{19}$$



(A) The graph and a colour density plot of the function $f(x, y)$ from Example 2.3 over the unit square $\bar{Q} = \{0 \le x \le 1,\ 0 \le y \le 1\}$.



(B) The graph and a colour density plot of the function $f(x, y)$ from Example 2.4 over the disk $\bar{D} = \{x^2 + y^2 \le 25\}$.

FIGURE 1. Illustrations for Examples 2.3 and 2.4.

it follows that $(x^*, y^*) = (0, 0)$ is the only critical point of f in D. For later reference, let us compute

$$f(x^*, y^*) = f(0, 0) = 8. \tag{20}$$

Now we parameterize the boundary of $\bar{D}$ as

$$(x(t), y(t)) = (5\cos t, 5\sin t), \qquad t \in \mathbb{R}. \tag{21}$$

Note that for any t_1 and t_2 satisfying $t_2 = t_1 + 2\pi n$ with some integer n, we have $x(t_1) = x(t_2)$ and $y(t_1) = y(t_2)$, meaning that the parameter values t_1 and t_2 correspond to the same point on the boundary (circle) of $\bar{D}$. The values of f along the boundary gives rise to the function

$$\begin{aligned} g(t) = f(x(t), y(t)) &= 5 \cdot 25\cos^2 t - 22 \cdot 25 \sin t \cos t + 5 \cdot 25 \sin^2 t + 8 \\ &= 125 - 22 \cdot 25 \sin t \cos t + 8 = 133 - 275 \sin 2t, \end{aligned} \tag{22}$$

where we have used the identity $\sin 2t = 2\sin t \cos t$. From the properties of sine, we infer that the minimum of g is obtained at $2t = \frac{\pi}{2} + 2\pi n$ for integer n, and the maximum of g is obtained at $2t = \frac{3\pi}{2} + 2\pi n$ for integer n. In terms of t, the minimum is at $t = \frac{\pi}{4} + \pi n$, which means that there are two minimizers corresponding to $t_1 = \frac{\pi}{4}$ and to $t_2 = \frac{\pi}{4} + \pi = \frac{5\pi}{4}$. Similarly, the maximum is obtained at $t = \frac{3\pi}{4} + \pi n$, giving two maximizers corresponding to $t_3 = \frac{3\pi}{4}$ and to $t_4 = \frac{3\pi}{4} + \pi = \frac{7\pi}{4}$. The values of f at these points are

$$g(t_1) = g(t_2) = 133 - 275 = -142, \qquad \text{and} \qquad g(t_3) = g(t_4) = 133 + 275 = 408. \tag{23}$$

A comparison of the preceding values with (20) makes it clear that

- If f has a minimum over $\bar{D}$, then the minimum value of f over $\bar{D}$ is -142, which occurs at $(5\cos t_1, 5\sin t_1) = (\frac{5}{\sqrt{2}}, \frac{5}{\sqrt{2}})$ and at $(5\cos t_2, 5\sin t_2) = (-\frac{5}{\sqrt{2}}, -\frac{5}{\sqrt{2}})$.
- If f has a maximum over $\bar{D}$, then the maximum value of f over $\bar{D}$ is 408, which occurs at $(5\cos t_3, 5\sin t_3) = (-\frac{5}{\sqrt{2}}, \frac{5}{\sqrt{2}})$ and at $(5\cos t_4, 5\sin t_4) = (\frac{5}{\sqrt{2}}, -\frac{5}{\sqrt{2}})$.

3. The Weierstrass existence theorem

The "missing link" we needed in Example 2.3 and Example 2.4 is provided by an extension of the Weierstrass extreme value theorem (Theorem 1.1) to multi-dimensions. Before presenting the theorem, let us fix some terminology.

Definition 3.1. A set $B \subset \mathbb{R}^n$ is said to be *bounded* if $B \subset Q_R(0)$ for some $R > 0$.

Intuitively speaking, a bounded set is a set that does not contain a "point at infinity." For example, the open unit disk $\mathbb{D} = \{(x, y) : x^2 + y^2 < 1\}$ and the closed unit disk $\bar{\mathbb{D}} = \{(x, y) : x^2 + y^2 < 1\}$ are both bounded. However, the complement $\mathbb{R}^n \setminus \mathbb{D}$ is unbounded.

Definition 3.2. A *closed set* is by definition the complement of an open set, that is, a set of the form $A = \mathbb{R}^n \setminus B$, where B is open.

Intuitively, a closed set is a set that includes its boundary. Examples of closed sets are $\bar{\mathbb{D}} = \{(x, y) : x^2 + y^2 < 1\}$ and $\bar{\mathbb{Q}} = [0, 1]^2$. Since the empty set is open, $\mathbb{R}^n$ is closed. Note that $\mathbb{R}^n$ is also open.

Remark 3.3. Let $g_1, \dots, g_m$ be continuous functions of n variables. Then the set defined by

$$\begin{cases} g_1(x_1, \dots, x_n) \le 0 \\ \dots \\ g_m(x_1, \dots, x_n) \le 0 \end{cases}$$

is closed. On the other hand, the set defined by

$$\begin{cases} g_1(x_1, \dots, x_n) < 0 \\ \dots \\ g_m(x_1, \dots, x_n) < 0 \end{cases}$$

is open.

Theorem 3.4 (Weierstrass)**.** *Let $f : K \to \mathbb{R}$ be a continuous function, where $K \subset \mathbb{R}^n$ is a closed and bounded set. Then there exists $c \in K$ such that $f(x) \le f(c)$ for all $x \in K$.*

Example 3.5. The suppositions we have made in Example 2.3 and Example 2.4 are now justified, since each of these examples involves a continuous function f over either the closed unit square $\bar{Q}$, or the closed unit disk $\bar{D}$, which are obviously bounded sets.

Example 3.6. Find the coordinates of the point (x, y, z) on the plane $z = x + y + 4$ which is closest to the origin.

The square of the distance from a point (x, y, z) on the plane to the origin $(0, 0, 0)$ is

$$f(x, y) = x^2 + y^2 + (x + y + 4)^2. \tag{24}$$

To find the critical points, we set up the equations

$$\begin{aligned} \frac{\partial}{\partial x} f(x, y) &= 2x + 2(x + y + 4) = 4x + 2y + 8 = 0, \\ \frac{\partial}{\partial y} f(x, y) &= 2y + 2(x + y + 4) = 2x + 4y + 8 = 0, \end{aligned} \tag{25}$$

whose only solution is

$$x = y = -\tfrac{4}{3}. \tag{26}$$

We infer that the only critical point of $f(x, y)$ is $(x^*, y^*) = (-\frac{4}{3}, -\frac{4}{3})$, and that

$$f(x^*, y^*) = \tfrac{16}{3}. \tag{27}$$

The question is: Is this the minimum value of f? To answer this question, we first try to show that $f(x, y)$ is large when the point (x, y) is far away from the origin. If (x, y) is *outside* the open disk $D_R = \{(x, y) : x^2 + y^2 < R^2\}$ of radius $R > 0$, that is, if $x^2 + y^2 \geq R^2$, then we have

$$f(x, y) = x^2 + y^2 + (x + y + 4)^2 \geq x^2 + y^2 \geq R^2. \tag{28}$$

In particular, fixing $R = 4$, we get

$$f(x, y) \geq 16 > \tfrac{16}{3} = f(x^*, y^*) \qquad \text{for} \quad x^2 + y^2 \geq R^2, \tag{29}$$

and since (x^*, y^*) is in the disk D_R, we conclude that any possible minimizer must be contained in the disk D_R. Now we apply the Weierstrass existence theorem in the *closed* disk $\bar{D}_R = \{(x, y) : x^2 + y^2 \leq R^2\}$, to infer that there exists a minimizer of f over the closed disk $\bar{D}_R$. By (29), a minimizer cannot be on the boundary of $\bar{D}_R$, so it must be in the open disk D_R. This means that any minimizer must be a critical point of f in D_R, but we know that there is only one critical point, implying that there is only one minimizer of f over $\bar{D}_R$, and the minimizer is the point (x^*, y^*). Note that at this point all we know is that (x^*, y^*) is the minimizer of f over $\bar{D}_R$. However, invoking (29) once again, we conclude that (x^*, y^*) is indeed the minimizer of f over $\mathbb{R}^2$. Finally, since we were asked to find the coordinates of the point (x, y, z), we note that the minimizer (x^*, y^*) corresponds to the point

$$(x, y, z) = (-\tfrac{4}{3}, -\tfrac{4}{3}, \tfrac{4}{3}), \tag{30}$$

on the plane $z = x + y + 4$.

Example 3.7. Find the maximum and minimum values of

$$f(x, y) = \frac{x + y}{2 + x^2 + y^2}.$$

The critical points of f must satisfy

$$\begin{aligned} \frac{\partial}{\partial x} f(x, y) &= \frac{2 + x^2 + y^2 - (x + y) \cdot 2x}{(2 + x^2 + y^2)^2} = \frac{2 - x^2 + y^2 - 2xy}{(2 + x^2 + y^2)^2} = 0, \\ \frac{\partial}{\partial y} f(x, y) &= \frac{2 + x^2 + y^2 - (x + y) \cdot 2y}{(2 + x^2 + y^2)^2} = \frac{2 + x^2 - y^2 - 2xy}{(2 + x^2 + y^2)^2} = 0, \end{aligned} \tag{31}$$

which imply the equations

$$\begin{aligned} 2 - x^2 + y^2 - 2xy &= 0, \\ 2 + x^2 - y^2 - 2xy &= 0. \end{aligned} \tag{32}$$

By adding and subtracting one of the equations from the other, we arrive at

$$xy = 1, \qquad x^2 - y^2 = 0, \tag{33}$$

whose only solutions are

$$(x_1, y_1) = (1, 1), \qquad \text{and} \qquad (x_2, y_2) = (-1, -1). \tag{34}$$

For the values of f at the critical points, we have

$$f(x_1, y_1) = \frac{1}{2}, \qquad \text{and} \qquad f(x_2, y_2) = -\frac{1}{2}. \tag{35}$$

The question is: Are they the maximum and minimum values of f? To answer this question, we first try to show that $f(x, y)$ is close to 0 when the point (x, y) is far away from the origin.

Given (x,y), let $r=\sqrt{x^2+y^2}$ be the distance from (x,y) to the origin $(0,0)$. Then we have $|x|\leq r$ and $|y|\leq r$, and so

$$|f(x,y)|=\frac{|x+y|}{2+x^2+y^2}\leq\frac{|x|+|y|}{2+x^2+y^2}\leq\frac{2r}{2+r^2}. \tag{36}$$

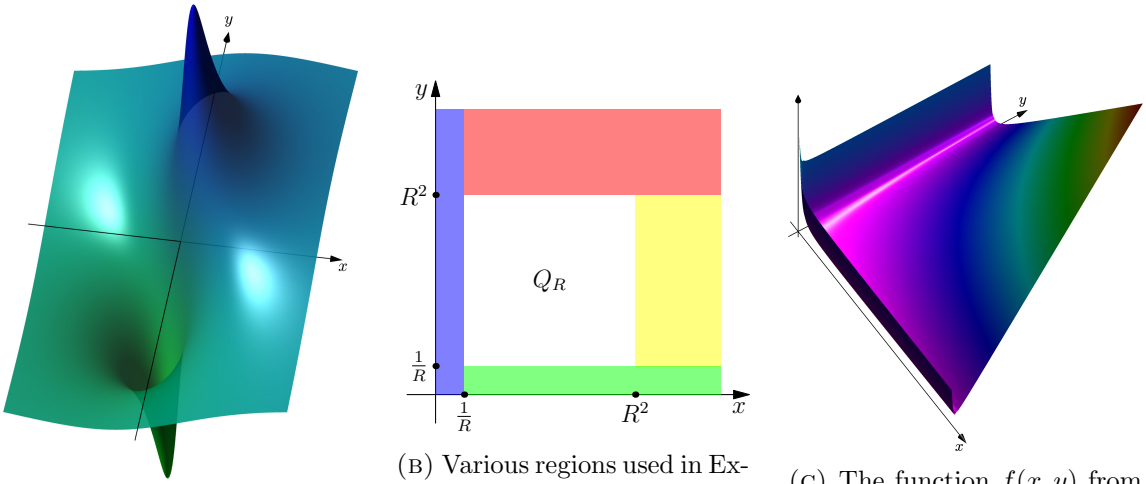
Moreover, if (x,y) is *outside* the open disk $D_R=\{(x,y):x^2+y^2<R^2\}$ of radius $R>0$, that is, if $r\geq R$, then we have

$$|f(x,y)|\leq\frac{2r}{2+r^2}\leq\frac{2r}{r^2}=\frac{2}{r}\leq\frac{2}{R}. \tag{37}$$

In particular, fixing $R=8$, we get

$$|f(x,y)|\leq\frac{1}{4}\qquad\text{for}\quad x^2+y^2\geq R^2. \tag{38}$$

Since both points (x_1,y_2) are (x_2,y_2) are in the disk D_R, we conclude that any possible maximizers and minimizers must be contained in the disk D_R. Now we apply the Weierstrass existence theorem in the *closed* disk $\bar{D}_R=\{(x,y):x^2+y^2\leq R^2\}$, to infer that there exist a maximizer and a minimizer of f over the closed disk $\bar{D}_R$. By (38), neither a maximizer nor a minimizer can be on the boundary of $\bar{D}_R$, so they must be in the open disk D_R. This means that any maximizer must be a critical point of f in D_R, and comparing the values (35), we infer that there is only one maximizer of f over $\bar{D}_R$, and the maximizer is the point (x_1,y_1). Similarly, there is only one minimizer of f over $\bar{D}_R$, and the minimizer is the point (x_2,y_2). Note that at this point all we know is that (x_1,y_1) is the maximizer of f over $\bar{D}_R$ and that (x_2,y_2) is the minimizer of f over $\bar{D}_R$. However, invoking (38) once again, we conclude that (x_1,y_1) is indeed the maximizer of f over $\mathbb{R}^2$ and that (x_2,y_2) is indeed the minimizer of f over $\mathbb{R}^2$. The final answer is that the maximum value of f in $\mathbb{R}^2$ is $\frac{1}{2}$, and the minimum value is $-\frac{1}{2}$.



(A) The graph of the function $f(x,y)$ from Example 3.7, over a square centred at the origin.

(B) Various regions used in Example 3.8.

(C) The function $f(x,y)$ from Example 3.8.

FIGURE 2. Illustrations for Examples 3.7 and 3.8.

Example 3.8. Find the most economical dimensions of a closed rectangular box of volume 3 cubic units if the cost of the material per square unit for (i) the top and bottom is 2, (ii) the front and back is 2 and (iii) the other two sides is 8.

Let us denote the width of the box by x, the height by z, and the depth by y. Then the combined area of the top and bottom faces is $2xy$, the area of the front and back faces is $2xz$, and the area of the other two sides is $2yz$. Thus the problem is to find the minimizer of

$$F(x,y,z) = 4xy + 4xz + 16yz, \qquad \text{subject to} \quad xyz = 3. \tag{39}$$

We look for the solution satisfying $x > 0$, $y > 0$, and $z > 0$, because if one of x, y, and z is 0, the volume of the box cannot be equal to 3. Then by using the volume constraint $xyz = 3$, we can express z in terms of x and y, resulting in the reformulation of the problem as minimizing

$$f(x,y) = 4xy + \frac{12}{y} + \frac{48}{x}, \tag{40}$$

over the quadrant $H = \{(x,y) : x > 0, y > 0\}$. First, let us find the critical points of f. The relevant equations are

$$\begin{aligned} \frac{\partial}{\partial x} f(x,y) &= 4y - \frac{48}{x^2} = 0, \\ \frac{\partial}{\partial y} f(x,y) &= 4x - \frac{12}{y^2} = 0, \end{aligned} \tag{41}$$

which lead us to $x^2 y = 12$ and $xy^2 = 3$. If we divide one equation by the other, we get $x = 4y$, and this in turn yields that

$$(x^*, y^*) = (\sqrt[3]{48}, \sqrt[3]{\tfrac{3}{4}}), \tag{42}$$

is the only critical point of f over H. Note that the corresponding z-value is $z^* = y^* = \sqrt[3]{48}$. (The variables y and z play indistinguishable roles in the original problem, so for quick calculations, we could have set $y = z$ from the beginning and could have transformed the whole problem into a single variable minimization problem.)

The question is now if (x^*, y^*) is indeed a minimizer of f over H. Intuitively, from (40) it is clear that $f(x,y)$ tends to ∞ if $x > 0$ and $y > 0$ are small, or if they are large. To make it precise, given $R > 0$, let

$$Q_R = \{(x,y) : \frac{1}{R} < x < R^2, \ \frac{1}{R} < y < R^2\}. \tag{43}$$

We want to show that if the point (x,y) is outside the square Q_R with $R > 0$ large, then $f(x,y)$ is large. Suppose that $x \le \frac{1}{R}$ (Figure 2(b), blue region). Then we have

$$f(x,y) = 4xy + \frac{12}{y} + \frac{48}{x} \ge \frac{48}{x} \ge 48R. \tag{44}$$

Similarly, for $y \le \frac{1}{R}$ (Figure 2(b), green region plus part of the blue region), we have

$$f(x,y) = 4xy + \frac{12}{y} + \frac{48}{x} \ge \frac{12}{y} \ge 12R. \tag{45}$$

Now suppose that $x > \frac{1}{R}$ and $y \ge R^2$ (Figure 2(b), red region). Then we have

$$f(x,y) = 4xy + \frac{12}{y} + \frac{48}{x} \ge 4xy \ge 4 \cdot \frac{1}{R} \cdot R^2 = 4R. \tag{46}$$

Finally, for $y > \frac{1}{R}$ and $x \ge R^2$ (Figure 2(b), yellow region plus part of the red region), we have

$$f(x,y) = 4xy + \frac{12}{y} + \frac{48}{x} \ge 4xy \ge 4 \cdot R^2 \cdot \frac{1}{R} = 4R, \tag{47}$$

and a combination of the last four formulas gives

$$f(x,y) \ge 4R, \qquad \text{for} \quad (x,y) \notin Q_R. \tag{48}$$

Therefore, by choosing $R > 0$ sufficiently large, we can ensure that $f(x, y) > f(x^*, y^*)$ for all (x, y) outside Q_R, meaning that any minimizer of f over H must be contained in Q_R. Let us fix such a value for R. Then as usual, the Weierstrass existence theorem guarantees the existence of a minimizer of f over the closed set $\bar{Q}_R$, where

$$\bar{Q}_R = \{(x, y) : \frac{1}{R} \leq x \leq R^2, \ \frac{1}{R} \leq y \leq R^2\}. \tag{49}$$

We have chosen $R > 0$ so large that $f(x, y) > f(x^*, y^*)$ for all (x, y) outside Q_R, which rules out the possibility that a minimizer over $\bar{Q}_R$ is on the boundary of $\bar{Q}_R$. Hence all minimizers are in Q_R, and at least one such minimizer exists. Since Q_R is open, all minimizers must be critical points, but we have only one critical point, thus we infer that (x^*, y^*) is the only minimizer of f over $\bar{Q}_R$. Finally, recalling that $f(x, y) > f(x^*, y^*)$ for all (x, y) outside Q_R, we conclude that (x^*, y^*) is the only minimizer of f over H.

4. Lagrange multipliers on surfaces

If u is defined in some set U, and we are only interested in the restriction of u to a subset $K \subset U$, then we indicate it by expressions such as local maximum *with respect to* K, and local maximum *over* K.

Theorem 4.1. *Let $U \subset \mathbb{R}^n$ be open, and let $M \subset U$ be a hypersurface. Suppose that $u : U \to \mathbb{R}$ is differentiable. If u has a local maximum over M at $y \in M$, then we have*

$$D_V u(y) = 0 \qquad \text{for all} \quad V \in T_y M. \tag{50}$$

Proof. Let $V \in T_y M$, and let $\gamma : (-\varepsilon, \varepsilon) \to M$ be a differentiable curve satisfying $\gamma(0) = y$ and $\gamma'(0) = V$. Put $g(t) = u(\gamma(t))$, and note that $g'(0) = D_V u(y)$. Anticipating a contradiction, suppose that $g'(0) > 0$. By definition, we have $g(t) = g(0) + h(t)t$ with $h(t) \to g'(0)$ as $t \to 0$. Hence by continuity, there exists $t > 0$ arbitrarily small, such that $h(t) > \frac{1}{2} g'(0) > 0$. This gives $g(t) > g(0) + \frac{1}{2} g'(0)t$, meaning that $g(0)$ cannot be a local maximum. The case $g'(0) < 0$ can be treated similarly, by considering values $g(t)$ with small $t < 0$. □

Definition 4.2. Let $U \subset \mathbb{R}^n$ be open, and let $M \subset U$ be a hypersurface. Suppose that $u : U \to \mathbb{R}$ is differentiable. If $D_V u(x) = 0$ for all $V \in T_x M$, then $x \in M$ is called a *critical point* of u over M, and the value $u(x) \in \mathbb{R}$ is called a *critical value* of u over M.

Remark 4.3. Let $\Psi : \Omega \to M$ be a local parameterization of M, and let $x = \Psi(\xi)$ for some $\xi \in \Omega$. Since $D_V u(x) = Du(x)V$ and the columns of $D\Psi(\xi)$ form a basis of $T_x M$, we infer that x is a critical point of u over M if and only if $Du(x)D\Psi(\xi) = 0$. Furthermore, noting that $Dv(\xi) = Du(x)D\Psi(\xi)$ for $v = u \circ \Psi$, we conclude that x is a critical point of u over M if and only if ξ is a critical point of v in Ω.

Definition 4.4. Given a hypersurface $M \subset \mathbb{R}^n$ and its point $p \in M$, the *conormal space* of M at p is defined as

$$N_p^* M = \{\alpha \in \mathbb{R}^{1 \times n} : \alpha V = 0 \text{ for all } V \in T_p M\}. \tag{51}$$

Remark 4.5. Let $\Psi : \Omega \to M$ be a local parameterization of M, and let $p = \Psi(q)$, $q \in \Omega$. Then $\alpha \in N_p^* M$ if and only if $\alpha D\Psi(q) = 0$, that is, $\alpha \in \operatorname{coker} D\Psi(q)$. So we have

$$N_p^* M = \operatorname{coker} D\Psi(q). \tag{52}$$

Now let $U \subset \mathbb{R}^n$ be open, and suppose that $M \cap U$ is described by the equation $\phi(x) = 0$, where $\phi : U \to \mathbb{R}$ is a continuously differentiable function with $\nabla\phi(x) \neq 0$ for all $x \in M \cap U$. We know that $T_p M = \ker \nabla\phi(p)$. For any $\beta \in \mathbb{R}^{1 \times k}$, we have $\beta \nabla\phi(p) D\Psi(q) = 0$, and hence $\beta \nabla\phi(p) \in N_p^* M$. In other words,

$$\operatorname{coran} \nabla\phi(p) \subset N_p^* M. \tag{53}$$

By assumption, $\nabla\phi(p) \neq 0$, and so the row rank must be 1, that is, $\dim(\mathrm{coran}\nabla\phi(p)) = 1$. On the other hand, the column rank of $D\Psi(q)$ is $m = n-1$, and so $\dim(\mathrm{coran}D\Psi(q)) = m$. Invoking the rank-nullity theorem, the dimension of $\mathrm{coker}D\Psi(q)$ is $n - m = 1$. Since we have the inclusion $\mathrm{coran}\nabla\phi(p) \subset \mathrm{coker}D\Psi(q)$ with dimensions agreeing, we conclude that

$$N_p^*M = \mathrm{coran}\nabla\phi(p) = \mathrm{coker}D\Psi(q). \tag{54}$$

The following result is now straightforward.

Theorem 4.6 (Lagrange multipliers)**.** *Let $U \subset \mathbb{R}^n$ be open, and let $M \subset U$ be a hypersurface. Suppose that $u : U \to \mathbb{R}$ is differentiable. Then $p \in M$ is a critical point of u over M if and only if $\nabla u(p) \in N_p^*M$. In addition, suppose that M is described by the equation $\phi(x) = 0$, where $\phi : U \to \mathbb{R}$ is a continuously differentiable function with $\nabla\phi(x) \neq 0$ for all $x \in M$. Then $\nabla u(p) \in N_p^*M$ if and only if there exists $\lambda \in \mathbb{R}$ such that*

$$\nabla u(p) = \lambda\nabla\phi(p). \tag{55}$$

Proof. By definition, p is a critical point if and only if $\nabla u(p)V = 0$ for all $V \in T_pM$. The latter condition is equivalent to $\nabla u(p) \in N_p^*M$. By (54), we have $\nabla u(p) \in N_p^*M$ if and only if $\nabla u(p)$ is a scalar multiple of $\nabla\phi(p)$. □

Remark 4.7. If ϕ is a continuous function, then the set $M = \{x : \phi(x) = 0\}$ is closed. Therefore in order to apply the Weierstrass existence theorem, one has only to ensure boundedness. If M is not bounded, an often useful and simple method is to consider the set $M \cap \bar{B}_R$ with $R > 0$ large enough, where $\bar{B}_R$ is the closed ball of radius R, centred at a suitable point. Then one needs to deal with the remaining portion $M \setminus \bar{B}_R$ of M in some other way.

Example 4.8. Let us apply the Lagrange multiplier appoach to Example 3.6: Find the coordinates of the point (x, y, z) on the plane $z = x + y + 4$ which is closest to the origin.

We consider the function

$$f(x, y, z) = x^2 + y^2 + z^2, \tag{56}$$

on the plane $P = \{(x, y, z) \in \mathbb{R}^3 : \phi(x, y, z) = 0\}$, where $\phi(x, y, z) = x + y + 4 - z$. By the Lagrange multipliers theorem, $(x, y, z) \in P$ is a critical point of f on P if and only if there exists $\lambda \in \mathbb{R}$ such that $Df(x, y, z) = \lambda D\phi(x, y, z)$, that is,

$$(2x, 2y, 2z) = \lambda(1, 1, -1). \tag{57}$$

This yields $(x, y, z) = \frac{\lambda}{2}(1, 1, -1)$, and by invoking $(x, y, z) \in P$, we find

$$\tfrac{\lambda}{2} + \tfrac{\lambda}{2} + 4 - (-\tfrac{\lambda}{2}) = 0. \tag{58}$$

Hence we have $\lambda = -\frac{8}{3}$, and so $(x, y, z) = (-\frac{4}{3}, -\frac{4}{3}, \frac{4}{3})$ is the only critical point. From here on, we can proceed as in Example 3.6.

Example 4.9. Find the maximum and minimum values of $f(x, y, z) = 2x - y + 4z$ on the sphere $x^2 + y^2 + z^2 = 1$. Here you can use any type of reasonings, including geometric ones.

Let $\phi(x, y, z) = x^2 + y^2 + z^2 - 1$. the sphere $S^2 = \{(x, y, z) \in \mathbb{R}^3 : \phi(x, y, z) = 0\}$ is closed and bounded, by the Weierstrass theorem, there exist a maximizer and a minimizer of f over S^2, and by the first derivative test, these points must be critical points of f on S^2. Then by the Lagrange multiplier theorem, $(x, y, z) \in S^2$ is a critical point if and only if there exists $\lambda \in \mathbb{R}$ such that

$$Df(x, y, z) = \lambda D\phi(x, y, z). \tag{59}$$

We compute $Df(x, y, z) = (2, -1, 4)$ and $D\phi(x, y, z) = (2x, 2y, 2z)$, and hence (59) becomes $(2, -1, 4) = \lambda(2x, 2y, 2z)$. From this, it is clear that $\lambda \neq 0$, and so $(x, y, z) = \lambda^{-1}(1, -\frac{1}{2}, 2)$. Invoking $(x, y, z) \in S^2$, we infer

$$1 = x^2 + y^2 + z^2 = \frac{1 + (\frac{1}{2})^2 + 2^2}{\lambda^2} = \frac{21}{4\lambda}, \tag{60}$$

yielding $\lambda = \pm\frac{\sqrt{21}}{2}$. Now we can find

$$(x, y, z) = \lambda^{-1}(1, -\tfrac{1}{2}, 2) = (\pm\tfrac{2}{\sqrt{21}}, \mp\tfrac{1}{\sqrt{21}}, \pm\tfrac{4}{\sqrt{21}}). \tag{61}$$

The maximum and the minimum of f on S^2 must be among these two points. We evaluate f at these two points, as

$$f(\pm\tfrac{2}{\sqrt{21}}, \mp\tfrac{1}{\sqrt{21}}, \pm\tfrac{4}{\sqrt{21}}) = \pm\sqrt{21}, \tag{62}$$

and compare the values, to conclude that $(\frac{2}{\sqrt{21}}, -\frac{1}{\sqrt{21}}, \frac{4}{\sqrt{21}})$ is the maximizer of f on S^2, and $(-\frac{2}{\sqrt{21}}, \frac{1}{\sqrt{21}}, -\frac{4}{\sqrt{21}})$ is the minimizer of f on S^2. The maximum and minimum *values* of f on S^2 are of course $\pm\sqrt{21}$.

5. Lagrange multipliers on curves

Hypersurfaces include curves in $\mathbb{R}^2$ and surfaces in $\mathbb{R}^3$. These sets are one dimension lower than the ambient space (i.e., of codimension 1), and we get away with a scalar Lagrange multiplier, as we have seen in the preceding section. However, as curves in $\mathbb{R}^3$ are of codimension 2, they need a 2-dimensional Lagrange multiplier.

Definition 5.1. Given a curve $L \subset \mathbb{R}^3$ and its point $p \in L$, the *conormal space* of L at p is defined as

$$N_p^* L = \{\alpha \in \mathbb{R}^{1\times 3} : \alpha V = 0 \text{ for all } V \in T_p L\}. \tag{63}$$

Remark 5.2. Let $\gamma : (a, b) \to L$ be a local parameterization of L, and let $p = \gamma(t)$, $t \in (a, b)$. Then $\alpha \in N_p^* M$ if and only if $\alpha\gamma'(t) = 0$, that is, $\alpha \in \operatorname{coker}\gamma(t)$. So we have

$$N_p^* M = \operatorname{coker}\gamma(t). \tag{64}$$

Now let $U \subset \mathbb{R}^3$ be open, and suppose that $M \cap U$ is described by the equation $\phi(x) = 0$, where $\phi : U \to \mathbb{R}^2$ is a continuously differentiable function with $D\phi(x)$ surjective for all $x \in M \cap U$. We know that $T_p M = \ker D\phi(p)$. For any $\beta \in \mathbb{R}^{1\times 2}$, we have $\beta D\phi(p)\gamma'(t) = 0$, and hence $\beta D\phi(p) \in N_p^* M$. In other words,

$$\operatorname{coran} D\phi(p) \subset N_p^* M. \tag{65}$$

By assumption, the column rank of $D\phi(p)$ is 2, and so the row rank must also be 2, that is, $\dim(\operatorname{coran} D\phi(p)) = 2$. On the other hand, the column rank of $\gamma'(t)$ is $m = 1$, and so $\dim(\operatorname{coran}\gamma(t)) = 1$. Invoking the rank-nullity theorem, the dimension of $\operatorname{coker}\gamma'(t)$ is $3 - 1 = 2$. Since we have the inclusion $\operatorname{coran} D\phi(p) \subset \operatorname{coker}\gamma'(t)$ with dimensions agreeing, we conclude that

$$N_p^* M = \operatorname{coran} D\phi(p) = \operatorname{coker}\gamma'(t). \tag{66}$$

The following result is now straightforward.

Theorem 5.3 (Lagrange multipliers)**.** *Let $U \subset \mathbb{R}^3$ be open, and let $L \subset U$ be a smooth curve. Suppose that $u : U \to \mathbb{R}$ is differentiable. Then $p \in L$ is a critical point of u over L if and only if $\nabla u(p) \in N_p^* M$. In addition, suppose that L is described by the equation $\phi(x) = 0$, where $\phi : U \to \mathbb{R}^2$ is a continuously differentiable function with $D\phi(x)$ surjective for all $x \in L$. Then $\nabla u(p) \in N_p^* M$ if and only if there exists $\lambda_1, \lambda_2 \in \mathbb{R}$ such that*

$$\nabla u(p) = \lambda_1 \nabla\phi_1(p) + \lambda_2 \nabla\phi_2(p). \tag{67}$$

Proof. By definition, p is a critical point if and only if $\nabla u(p)V = 0$ for all $V \in T_p M$. The latter condition is equivalent to $\nabla u(p) \in N_p^* M$. By (66), we have $\nabla u(p) \in N_p^* M$ if and only if $\nabla u(p)$ is a linear combination of the rows of $D\phi(p)$. □

Remark 5.4. Surjectivity of $D\phi(x)$ is equivalent to any (hence both) of the following.

- $\nabla\phi_1(x)$ and $\nabla\phi_2(x)$ are linearly independent.
- There is a 2×2 submatrix of $D\phi(x)$ that is nonsingular.

Example 5.5. Let us find the highest points on the curve given by

$$\begin{cases} 4x+9y+z=0, \\ z=2x^2+3y^2. \end{cases} \tag{68}$$

The curve is the 0-locus of the function

$$\phi(x,y,z)=\begin{pmatrix} 4x+9y+z \\ 2x^2+3y^2-z \end{pmatrix}, \tag{69}$$

whose derivative is

$$D\phi(x,y,z)=\begin{pmatrix} 4 & 9 & 1 \\ 4x & 6y & -1 \end{pmatrix}. \tag{70}$$

The determinants of all 2×2 submatrices of $D\phi$ are

$$\begin{aligned} \det D_{x,y}\phi(x,y,z) &= \det\begin{pmatrix} 4 & 9 \\ 4x & 6y \end{pmatrix} = 12(2y-3x), \\ \det D_{x,z}\phi(x,y,z) &= \det\begin{pmatrix} 4 & 1 \\ 4x & -1 \end{pmatrix} = -4(1+x), \\ \det D_{y,z}\phi(x,y,z) &= \det\begin{pmatrix} 9 & 1 \\ 6y & -1 \end{pmatrix} = -3(3+2y). \end{aligned} \tag{71}$$

If $x\neq -1$ or $y\neq -\frac{3}{2}$, then either the second or the third submatrix is nonsingular. If $(x,y)=(-1,-\frac{3}{2})$, then $2y-3x=-2+3\cdot\frac{3}{2}=\frac{5}{2}$, and so the first submatrix is nonsingular. This means that $D\phi(x,y,z)$ is surjective at every $(x,y,z)\in\mathbb{R}^3$. We conclude that

$$L=\{(x,y,z):\phi(x,y,z)=0\} \tag{72}$$

is a smooth curve in $\mathbb{R}^3$.

Our task is to maximize the function

$$u(x,y,z)=z, \tag{73}$$

over L. The gradient of u is

$$\nabla u(x,y,z)=\begin{pmatrix} 0 & 0 & 1 \end{pmatrix}, \tag{74}$$

meaning that the critical point of u over L must satisfy

$$\begin{pmatrix} 0 & 0 & 1 \end{pmatrix}=\lambda_1\begin{pmatrix} 4 & 9 & 1 \end{pmatrix}+\lambda_2\begin{pmatrix} 4x & 6y & -1 \end{pmatrix}, \tag{75}$$

for some real numbers λ_1 and λ_2, cf. Theorem 5.3. It is easy to see that $\lambda_2=0$ is not possible, and hence assuming $\lambda_2\neq 0$, we get

$$x=-\frac{\lambda_1}{\lambda_2},\qquad y=-\frac{3\lambda_1}{2\lambda_2}=\frac{3x}{2}. \tag{76}$$

Substituting these into (68), we have

$$\begin{aligned} z=2x^2+3y^2=2x^2+\frac{27x^2}{4}=\frac{35x^2}{4}, \\ z=-4x-9y=-4x-\frac{27x}{2}=-\frac{35x}{2}, \end{aligned} \tag{77}$$

which yield

$$x^2=-2x. \tag{78}$$

This leads to the two critical points (x_1,y_1,z_1) and (x_2,y_2,z_2) on L, where

$$x_1=0,\quad y_1=0,\quad z_1=0, \tag{79}$$

and

$$x_2 = -2, \quad y_2 = -3, \quad z_2 = 35. \tag{80}$$

Since $z_2 > z_2$, the point $(-2, -3, 35)$ is clearly the highest point on L.

6. Diagonalization of symmetric matrices

As a nice application of the theory we have developed so far, in this section, we will prove that any symmetric matrix A can be written as

$$A = Q\Lambda Q^\top, \tag{81}$$

with a diagonal matrix

$$\Lambda = \begin{pmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \lambda_n \end{pmatrix}, \tag{82}$$

and a matrix Q satisfying $Q^\top Q = QQ^\top = I$. The numbers $\lambda_1, \dots, \lambda_n$ are called the *eigenvalues* of A, and the columns $q_1, \dots, q_n \in \mathbb{R}^n$ of Q are called the *eigenvectors* of A. Note that (81) can also be written as

$$AQ = Q\Lambda, \tag{83}$$

which is equivalent to

$$Aq_k = \lambda_k q_k, \qquad k = 1, \dots, n. \tag{84}$$

Now, the condition $Q^\top Q = I$ means

$$q_k^\top q_\ell = \delta_{k\ell}, \tag{85}$$

that is, the columns of Q are pairwise orthogonal, and normalized so that their lengths are all equal to 1. We say that the set $\{q_k\}$ is *orthonormal*. An arbitrary vector $v \in \mathbb{R}^n$ can be written in this orthonormal basis as

$$v = Iv = QQ^\top v = (q_1^\top v)q_1 + \dots + (q_n^\top v)q_n, \tag{86}$$

where the quantities $q_k^\top v$ can be called the coordinates of v with respect to the basis $\{q_k\}$. Then the application of A to v becomes

$$Av = Q\Lambda Q^\top v = (\lambda_1 q_1^\top v)q_1 + \dots + (\lambda_n q_n^\top v)q_n. \tag{87}$$

Thus, the application of A to v is simply the scaling of v by the factors $\lambda_1, \dots, \lambda_n$ respectively in the directions $q_1, \dots, q_n$.

Remark 6.1. It is obvious that if $Au = \lambda u$ and $Av = \lambda v$, then we have

$$A(\alpha u + \beta v) = \lambda(\alpha u + \beta v), \tag{88}$$

for any $\alpha, \beta \in \mathbb{R}$. This means that the collection of all eigenvectors corresponding to a particular eigenvalue forms a linear space.

Remark 6.2. If we have numbers $\lambda_1, \dots, \lambda_n$, and orthonormal set of vectors $q_1, \dots, q_n$ satisfying (84), then we would immediately have the diagonalization (81). The equation (84) is equivalent to $(A - \lambda_k I)q_k = 0$, and as $q_k \neq 0$, this means that the matrix $A - \lambda_k I$ is singular. Hence the eigenvalues satisfy the so-called *characteristic equation*

$$\det(A - \lambda I) = 0, \tag{89}$$

which is an n-th degree polynomial equation for λ. After finding the eigenvalues from the characteristic equation, one may try to find the eigenvectors by solving $Aq = \lambda_k q$ for q. However, in general, one does not obtain n linearly independent eigenvectors. For example, the matrix $A = \begin{pmatrix} 2 & 1 \\ 0 & 2 \end{pmatrix}$ has the double eigenvalue $\lambda_1 = \lambda_2 = 2$, but solving $Aq = 2q$ only yields

$q = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$ and its scalar multiples. Such matrices are called *defective*. The diagonalization (81) is possible for symmetric matrices means that symmetric matrices are never defective.

The following lemma answers the question why orthogonality is a natural property of the eigenvectors of a symmetric matrix.

Lemma 6.3. *If A is symmetric, then eigenvectors corresponding to different eigenvalues are orthogonal to each other. That is, if $Au = \lambda u$ and $Av = \mu v$ with $\lambda \neq \mu$, then $u^\top v = 0$.*

Proof. First, observe that

$$v^\top Au = \lambda v^\top u, \qquad \text{and} \qquad u^\top Av = \mu u^\top v. \tag{90}$$

Since

$$v^\top u = (v^\top u)^\top = u^\top v, \tag{91}$$

and

$$v^\top Au = (v^\top Au)^\top = u^\top A^\top v = u^\top Av, \tag{92}$$

we have

$$\lambda v^\top u = \mu u^\top v, \qquad \text{or} \qquad (\lambda - \mu) u^\top v = 0, \tag{93}$$

which establishes the proof. □

Finally, we state and prove the main result of this section.

Theorem 6.4. *Let A be an $n \times n$ symmetric matrix. Then A admits an orthonormal set of n eigenvectors.*

Proof. The case $n = 1$ is trivial: If $A = [a]$ then take $\lambda_1 = a$ and $q_1 = 1$.

In the following, we will treat the cases $n = 2$ and $n = 3$ in detail. Let $n = 2$, and let

$$A = \begin{pmatrix} a & b \\ b & c \end{pmatrix}, \qquad u = \begin{pmatrix} x \\ y \end{pmatrix}. \tag{94}$$

Then consider the function

$$f(u) = f(x, y) = u^\top Au = ax^2 + 2bxy + cy^2, \tag{95}$$

over the unit circle

$$S^1 = \{(x, y) \in \mathbb{R}^2 : x^2 + y^2 = 1\}. \tag{96}$$

As S^1 is closed and bounded, by the Weierstrass existence theorem, f takes its maximum at some point $u_1 = (x_1, y_1) \in S^1$. This point u_1 must be a critical point of f over S^1. Then by the Lagrange multiplier theorem, there exists a number $\lambda_1 \in \mathbb{R}$ such that

$$\nabla f(x_1, y_1) = \lambda_1 \nabla g(x_1, y_1), \tag{97}$$

where

$$g(x, y) = x^2 + y^2 - 1. \tag{98}$$

Since

$$\nabla f(x, y) = (2ax + 2by, 2bx + 2cy) = 2Au, \tag{99}$$

and

$$\nabla g(x, y) = (2x, 2y) = 2u, \tag{100}$$

the equation (97) implies

$$Au_1 = \lambda_1 u_1, \tag{101}$$

that is, (λ_1, u_1) is in fact an eigenpair for A. To find another eigenpair, let $u_2 \in S^1$ be such that $u_2 \perp u_1$, and let $v = Au_2$. Then we have

$$u_1^\top v = u_1^\top Au_2 = (A^\top u_1)^\top Au_2 = (Au_1)^\top Au_2 = \lambda_1 u_1^\top u_2 = 0, \tag{102}$$

meaning that there is some number $\lambda_2 \in \mathbb{R}$ such that $v = \lambda_2 u_2$. Thus

$$Au_2 = \lambda_2 u_2, \tag{103}$$

showing that (λ_2, u_2) is in fact an eigenpair for A.

Now let $n = 3$, and consider the function

$$f(u) = u^\top A u, \qquad u \in \mathbb{R}^3, \tag{104}$$

over the unit sphere

$$S^2 = \{(x, y, z) \in \mathbb{R}^3 : x^2 + y^2 + z^2 = 1\}. \tag{105}$$

As S^2 is closed and bounded, by the Weierstrass existence theorem, f takes its maximum at some point $u_1 = (x_1, y_1, z_1) \in S^2$. This point must be a critical point of f over S^2. Then by the Lagrange multiplier theorem, there exists a number $\lambda_1 \in \mathbb{R}$ such that

$$\nabla f(u_1) = \lambda_1 \nabla g(u_1), \tag{106}$$

where

$$g(u) = g(x, y, z) = x^2 + y^2 + z^2 - 1. \tag{107}$$

Since

$$\nabla f(x, y) = 2Au, \qquad \text{and} \qquad \nabla g(u) = 2u, \tag{108}$$

the equation (106) implies

$$Au_1 = \lambda_1 u_1, \tag{109}$$

that is, (λ_1, u_1) is in fact an eigenpair for A. To find a second eigenpair, let

$$L = \{u \in \mathbb{R}^3 : u \perp u_1\}, \tag{110}$$

and consider f over the circle $S^2 \cap L$. This circle can be described as

$$\begin{cases} g(x, y, z) = 0, \\ g_1(x, y, z) = 0, \end{cases} \tag{111}$$

where

$$g_1(x, y, z) = xx_1 + yy_1 + zz_1. \tag{112}$$

As $S^2 \cap L$ is closed and bounded, by the Weierstrass existence theorem, f takes its maximum at some point $u_2 \in S^2 \cap L$. This point must be a critical point of f over $S^2 \cap L$. Then by the Lagrange multiplier theorem for curves in $\mathbb{R}^3$, there exist numbers λ_2 and μ such that

$$\nabla f(u_2) = \lambda_2 \nabla g(u_2) + \mu \nabla g_1(u_2). \tag{113}$$

which is the same thing as

$$2Au_2 = 2\lambda_2 u_2 + \mu u_1, \tag{114}$$

because

$$\nabla g_1(u) = u_1. \tag{115}$$

If we multiply (114) by $u_1^\top$ from the left, we get

$$2u_1^\top A u_2 = 2\lambda_2 u_1^\top u_2 + \mu u_1^\top u_1, \tag{116}$$

and since

$$u_1^\top A u_2 = (A^\top u_1)^\top u_2 = (Au_1)^\top u_2 = \lambda_1 u_1^\top u_2 = 0, \tag{117}$$

we conclude that $\mu u_1^\top u_1$ or $\mu = 0$. Thus (114) becomes

$$Au_2 = \lambda_2 u_2, \tag{118}$$

that is, (λ_2, u_2) is in fact an eigenpair for A. To find a third eigenpair, let $u_3 \in S^2$ be such that $u_3 \perp u_1$ and $u_3 \perp u_2$, and let $v = Au_3$. Then we have

$$u_1^\top v = u_1^\top A u_3 = (A^\top u_1)^\top A u_3 = (Au_1)^\top A u_3 = \lambda_1 u_1^\top u_3 = 0, \tag{119}$$

and similarly $u_2^\top v = 0$. This means that $v \perp u_1$ and $v \perp u_2$, i.e., there is some number $\lambda_3 \in \mathbb{R}$ such that $v = \lambda_3 u_3$. Thus

$$Au_3 = \lambda_3 u_3, \tag{120}$$

showing that (λ_3, u_3) is in fact an eigenpair for A.

The general case proceeds exactly as above, by looking at the function

$$f(u) = u^\top A u, \qquad u \in \mathbb{R}^n, \tag{121}$$

over the unit sphere

$$S^{n-1} = \{u \in \mathbb{R}^n : u^\top u = 1\}. \tag{122}$$

We maximize f over S^{n-1} to find the first eigenpair (λ_1, u_1). Then we maximize f over $S^{n-1} \cap L_1$, where L_1 is the plane orthogonal to u_1:

$$L_1 = \{u \in \mathbb{R}^n : g_1(u) = u^\top u_1 = 0\}. \tag{123}$$

This process gives the second eigenpair (λ_2, u_2). Next, we maximize f over $S^{n-1} \cap L_1 \cap L_2$, where L_2 is the plane orthogonal to u_2:

$$L_2 = \{u \in \mathbb{R}^n : g_2(u) = u^\top u_2 = 0\}. \tag{124}$$

We call this maximizer $u_3 \in S^{n-1} \cap L_1 \cap L_2$, and this must satisfy

$$\nabla f(u_3) = \lambda_3 \nabla g(u_3) + \mu_1 \nabla g_1(u_3) + \mu_2 \nabla g_2(u_3), \tag{125}$$

for some numbers λ_3, μ_1, and μ_2, which implies

$$2Au_3 = 2\lambda_3 u_3 + \mu_1 u_1 + \mu_2 u_2. \tag{126}$$

As in (116), we multiply this by $u_1^\top$ from the left, to get $\mu_1 = 0$. Then we multiply by $u_2^\top$ from the left, to get $\mu_2 = 0$, yielding

$$Au_3 = \lambda_3 u_3. \tag{127}$$

This process continues until we find $n-1$ eigenpairs $(\lambda_1, u_1), \ldots, (\lambda_{n-1}, u_{n-1})$, and in the last step, we consider a vector $u_n \in S^{n-1}$ such that $u_n \perp u_1, \ldots, u_n \perp u_{n-1}$, and as in (119), show that $v = Au_n$ is also orthonogal to all $u_1, \ldots, u_{n-1}$. Hence $v = \lambda_n u_n$ for some number λ_n. □

7. Second order derivatives

Before extending the second derivative test to multi-dimensions, in this section, we shall generalize the notion of second order derivatives.

Let $f : \mathbb{R}^n \to \mathbb{R}$ be differentiable everywhere in $\mathbb{R}^n$. By the convention that the elements of $\mathbb{R}^n$ are column vectors, it is natural to consider $Df(x)$ as a row vector, that is, $Df(x) \in \mathbb{R}^{1\times n}$ for $x \in \mathbb{R}^n$. Hence, Df can be considered as a function $Df : \mathbb{R}^n \to \mathbb{R}^{1\times n}$, and we can talk about its differentiability. Then the derivative $D^2 f(y) = DDf(y)$, if exists, should be a linear map $\Lambda : \mathbb{R}^n \to \mathbb{R}^{1\times n}$, satisfying

$$Df(x) = Df(y) + \Lambda(x - y) + e(x), \tag{128}$$

with $e(x) \to 0$ faster than $x - y$ as $x \to y$. Now, any such map can be written as

$$\Lambda(h) = h^\top A, \qquad h \in \mathbb{R}^n, \tag{129}$$

for some matrix $A \in \mathbb{R}^{n\times n}$. In view of this, we are going to identify $D^2 f(x)$ with the matrix A, and write

$$Df(x) = Df(y) + (x - y)^\top D^2 f(y) + e(x). \tag{130}$$

Taking the transpose of this equation, we get the equivalent form

$$Df(x)^\top = Df(y)^\top + D^2 f(y)^\top (x - y) + e(x), \tag{131}$$

which means that $D^2 f(y)^\top$ corresponds to the Jacobian matrix of the function $(Df)^\top$ at y.

Definition 7.1. Let $K \subset \mathbb{R}^n$, and let $f : K \to \mathbb{R}$ be differentiable everywhere in K. If $Df : K \to \mathbb{R}^{1\times n}$ is differentiable at $y \in K$ in the sense (130), then we say that f is *twice differentiable at* y, and call $D^2 f \in \mathbb{R}^{n\times n}$ the *second order derivative of* f *at* y.

Remark 7.2. In light of (131), f is twice differentiable at y if and only if there is a matrix $H \in \mathbb{R}^{n\times n}$ such that

$$B(x) = B(y) + H^\top (x - y) + e(x), \tag{132}$$

with $e(x) \to 0$ faster than $x - y$ as $x \to y$, where $B(x) = Df(x)^\top$ is simply the transpose of the Jacobian of f. Note that the Jacobian of f is the row vector consisting of the partial derivatives of f, and so its transpose $B(x)$ is a column vector. Hence $H^\top$ is given by the Jacobian matrix of the function $B(x)$, that is,

$$H^\top = \begin{pmatrix} \partial_1\partial_1 f(y) & \partial_2\partial_1 f(y) & \dots & \partial_n\partial_1 f(y) \\ \partial_1\partial_2 f(y) & \partial_2\partial_2 f(y) & \dots & \partial_n\partial_2 f(y) \\ \dots & \dots & \dots & \dots \\ \partial_1\partial_n f(y) & \partial_n\partial_1 f(y) & \dots & \partial_n\partial_n f(y) \end{pmatrix}, \tag{133}$$

or

$$H = \begin{pmatrix} \partial_1\partial_1 f(y) & \partial_1\partial_2 f(y) & \dots & \partial_1\partial_n f(y) \\ \partial_2\partial_1 f(y) & \partial_2\partial_2 f(y) & \dots & \partial_2\partial_n f(y) \\ \dots & \dots & \dots & \dots \\ \partial_n\partial_1 f(y) & \partial_n\partial_2 f(y) & \dots & \partial_n\partial_n f(y) \end{pmatrix}. \tag{134}$$

This matrix is called the *Hessian of* f *at* y. Note that if $D^2 f(y)$ exists, then $D^2 f(y) = H$, but without additional assumptions, the existence of H does not imply the existence of $D^2 f(y)$.

Example 7.3. (a) Let $f : \mathbb{R}^2 \to \mathbb{R}$ be defined by $f(x, y) = ax^2 + 2bxy + cy^2$, where $a, b, c \in \mathbb{R}$ are constants. We can compute the Jacobian matrix of f at (x, y) as

$$J_f(x, y) = \begin{pmatrix} 2ax + 2by & 2bx + 2cy \end{pmatrix} \in \mathbb{R}^{1\times 2}. \tag{135}$$

Since this depends on $(x, y) \in \mathbb{R}^2$ continuously, we conclude that f is differentiable everywhere in $\mathbb{R}^2$, with $Df(x, y) = J_f(x, y)$. Now, the Jacobian matrix of $(Df)^\top$ is

$$J(x, y) = \begin{pmatrix} 2a & 2b \\ 2b & 2c \end{pmatrix}, \tag{136}$$

and since it is continuous in $\mathbb{R}^2$, we conclude that f is twice differentiable in $\mathbb{R}^2$ with $D^2 f(x, y) = J^\top = J$.

(b) More generally, let $A \in \mathbb{R}^{n\times n}$, and let $f : \mathbb{R}^n \to \mathbb{R}$ be defined by $f(x) = x^\top Ax$, i.e.,

$$f(x) = \sum_{i,k=1}^{n} a_{ik} x_i x_k, \tag{137}$$

where $a_{ik} \in \mathbb{R}$ are the elements of A. Note that for $i \neq k$, the combination $x_i x_k$ appears in the sum twice, so that we can in fact write

$$f(x) = \sum_{i=1}^{n} a_{ii} x_i^2 + \sum_{i=1}^{n} \sum_{k=1}^{i-1} (a_{ik} + a_{ki}) x_i x_k. \tag{138}$$

In any case, we have

$$\partial_j f(x) = \sum_{i,k=1}^{n} a_{ik}(\delta_{kj} x_i + \delta_{ij} x_k) = \sum_{i=1}^{n} a_{ij} x_i + \sum_{k=1}^{n} a_{jk} x_k, \tag{139}$$

and hence

$$Df(x) = x^\top A + x^\top A^\top = x^\top (A + A^\top). \tag{140}$$

For the transpose, we would have $Df(x)^\top = (A + A^\top)x$, and the Jacobian of $(Df)^\top$ is simply $J = A + A^\top$. The Hessian of f is then the transpose of J, which is $H = A + A^\top$. We conclude that f is twice differentiable in $\mathbb{R}^n$ with $D^2 f(x) = A + A^\top$. Note that the Hessian is always symmetric no matter what A is, but this would have been clear from (138), which shows that replacing A by $\frac{1}{2}(A + A^\top)$ does not change f.

The following is a practical criterion on twice differentiability.

Remark 7.4. Let $Q = (a_1, b_1) \times \ldots \times (a_n, b_n)$ be an n-dimensional rectangular domain, and let $f : Q \to \mathbb{R}$ be a function. Assume that the Hessian $H(x)$ exists at every $x \in Q$, and $H(x)$ depends on x continuously. The existence of the Hessian in particular guarantees the existence of all first order partial derivatives of f in Q. Let $B(x) \in \mathbb{R}^n$ be the (column) vector consisting of all first order partial derivatives of f at x. Then $H(x)^\top$ is the Jacobian of the mapping $B : Q \to \mathbb{R}^n$ at x, and continuity of H implies that B is differentiable in Q, with $DB = H^\top$. In particular, B is continuous in Q. Now, since $B(x)^\top$ is the Jacobian of f at x, this implies that f is differentiable in Q, with $Df = B^\top$. As we have already established that B is differentiable in Q, we conclude that Df is differentiable in Q, and that $D^2 f = H$.

Differentiable functions are well approximated locally by linear functions. Intuitively speaking, twice differentiable functions should be close to quadratic functions.

Remark 7.5. Let $f : K \to \mathbb{R}$ with $K \subset \mathbb{R}^n$, and for $V \in \mathbb{R}^n$ fixed, suppose that $D_V f(x)$ exists at each $x \in K$. Then the dependence of $D_V f(x)$ on x can naturally be considered as a function $D_V f : K \to \mathbb{R}$. Hence one can talk about its directional differentiability, that is, about whether $D_W D_V f(x)$ exists for $x \in K$ and $W \in \mathbb{R}^n$. Now suppose that f is twice differentiable at y, i.e.,

$$Df(x) = Df(y) + (x - y)^\top D^2 f(y) + e(x). \tag{141}$$

If we multiply this from the right by $V \in \mathbb{R}^n$, we get

$$D_V f(x) = D_V f(y) + (x - y)^\top D^2 f(y)V + e(x)V, \tag{142}$$

where we have taken into account that $D_V f(x) = Df(x)V$. Next, we put $x = y + tW$ with $W \in \mathbb{R}^n$ and $t \in \mathbb{R}$ small, and infer

$$D_V f(y + tW) = D_V f(y) + tW^\top D^2 f(y)V + e(y + tW)V, \tag{143}$$

with $e(y + tW) \to 0$ faster than tW. This implies that

$$D_W D_V f(y) = W^\top D^2 f(y)V. \tag{144}$$

Remark 7.6. Let $Q \subset \mathbb{R}^n$ be an n-dimensional rectangular domain, and let $f : Q \to \mathbb{R}$ be twice differentiable everywhere in Q. Fix some $y \in K$ and $V \in \mathbb{R}^n$, and consider the function $g(t) = f(y + tV)$. We know that

$$\begin{aligned} g'(t) &= D_V f(y + tV) = Df(y + tV)V, \\ g''(t) &= D_V^2 f(y + tV) = V^\top D^2 f(y + tV)V, \end{aligned} \tag{145}$$

provided that $D^2 f(y + tV)$ exists. Invoking Theorem 1.8a), we get

$$f(y + tV) \approx f(y) + tDf(y)V + \frac{t^2}{2} V^\top D^2 f(y)V \qquad \text{as} \quad t \to 0. \tag{146}$$

So not surprisingly, the existence of $D^2 f(y)$ guarantees that $f(y + tV)$ can be well approximated by a quadratic polynomial in t. Supposing that $y + V \in Q$, we can also apply Theorem 1.8b), which yields

$$f(y + V) = f(y) + Df(y)V + \frac{1}{2} V^\top D^2 f(y + sV)V, \tag{147}$$

for some $s \in (0,1)$. This gives a quantitative information on the size of the error of the linear approximation $f(y+V) \approx f(y) + Df(y)V$.

Exercise 7.7. Introduce the notion of third derivative $D^3 f$, and give a quantitative information on the error of the quadratic approximation

$$f(y+V) \approx f(y) + Df(y)V + \frac{1}{2} V^\top D^2 f(y) V, \tag{148}$$

in the spirit of (147).

The following is a famous theorem which says that directional differentiations commute with each other. The latter would in particular imply that the Hessian matrix is symmetric.

Theorem 7.8. *We have $D_V D_W f = D_W D_V f$ wherever the two sides are defined.*

Exercise 7.9. Consider the function $f : \mathbb{R}^2 \to \mathbb{R}$ defined by

$$f(x,y) = \begin{cases} xy & \text{for } -|x| < y < |x|, \\ 0 & \text{otherwise.} \end{cases} \tag{149}$$

Check if the partial derivatives $\frac{\partial^2 f}{\partial x \partial y}$ and $\frac{\partial^2 f}{\partial y \partial x}$ exist at the origin, and if $\frac{\partial^2 f}{\partial x \partial y} = \frac{\partial^2 f}{\partial y \partial x}$ there.

8. The Hessian test

Let $f : U \to \mathbb{R}$ be a function twice differentiable at $c \in U$, where $U \subset \mathbb{R}^n$ is an open set, and suppose that c is a critical point of f. Then in view of Remark 7.6, the behaviour of f near c is dictated by the quadratic function

$$u(x) = f(c+x) - f(c) = x^\top A x, \tag{150}$$

where $A = \frac{1}{2} D^2 f(c)$ is symmetric. Note that since

$$Du(x) = x^\top (A + A^\top) = 2x^\top A = 2(Ax)^\top, \tag{151}$$

the only critical point of u is at $x = 0$, provided that A is nonsingular.

Example 8.1. For $n = 2$, the function under consideration is

$$u(x,y) = \begin{pmatrix} x \\ y \end{pmatrix}^\top \begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = ax^2 + (b+c)xy + dy^2. \tag{152}$$

Since u depends on b and c only through the combination $b + c$, we can and will assume that $b = c$, so that the matrix $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix} = \begin{pmatrix} a & b \\ b & d \end{pmatrix}$ is symmetric. Consider the following matrices

$$A_1 = \begin{pmatrix} 2 & 0 \\ 0 & 3 \end{pmatrix}, \qquad A_2 = \begin{pmatrix} -2 & 0 \\ 0 & -3 \end{pmatrix}, \qquad A_3 = \begin{pmatrix} 0 & 0 \\ 0 & 3 \end{pmatrix}, \qquad A_4 = \begin{pmatrix} 2 & 0 \\ 0 & -3 \end{pmatrix}, \tag{153}$$

with the corresponding functions

$$u_1(x,y) = 2x^2 + 3y^2, \qquad u_2(x,y) = -2x^2 - 3y^2,$$
$$u_3(x,y) = 3y^2, \qquad u_4(x,y) = 2x^2 - 3y^2.$$

It is obvious that:

- u_1 has its unique global minimum at $(0,0)$.
- u_2 has its unique global maximum at $(0,0)$.
- u_3 has a global minimum at *any* of the points of the y-axis.
- u_4 has its unique point critical at $(0,0)$, but this point is neither a maximum nor a minimum. In particular, $x = 0$ is the unique minimizer of $u_4(x,0)$, while $y = 0$ is the unique maximizer of $u_4(0,y)$. Such points are called *saddle points*.

Remark 8.2. As we proved in Section 6, any symmetric matrix A can be written as

$$A = Q^\top \Lambda Q, \tag{154}$$

with a diagonal matrix

$$\Lambda = \begin{pmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \lambda_n \end{pmatrix}, \tag{155}$$

and a matrix Q satisfying $Q^\top Q = I$. With this at hand, we can write (150) as

$$u(x) = x^\top A x = x^\top Q^\top \Lambda Q x = (Qx)^\top \Lambda Q x. \tag{156}$$

If we introduce the new coordinates $y = Qx$, then $x = Q^\top y$, and so

$$u(x) = u(Q^\top y) = y^\top \Lambda y = \lambda_1 y_1^2 + \lambda_2 y_2^2 + \dots + \lambda_n y_n^2. \tag{157}$$

Thus the situation basically reduces to the previous example:

- All $\lambda_k > 0$ if and only if u has a unique minimum at $x = 0$.
- All $\lambda_k < 0$ if and only if u has a unique maximum at $x = 0$.
- There exist positive as well as negative eigenvalues iff $x = 0$ is a saddle point.
- There exists a zero eigenvalue iff u has a degenerate behavior at $x = 0$.

Definition 8.3. Let $K \subset \mathbb{R}^n$, and let $u : K \to \mathbb{R}$. We say that u has a *local strict maximum* at $y \in K$ if there exists $\delta > 0$ such that $u(x) < u(y)$ for all $x \in Q_\delta(y) \cap K \setminus \{y\}$. The notion of *local strict minimum* may be defined similarly. A *saddle point* is a critical point that is not a local maximum or minimum.

Theorem 8.4. *Let $f : U \to \mathbb{R}$ be a function continuously differentiable in an open set $U \subset \mathbb{R}^n$, and suppose that f is twice differentiable at $y \in U$, where y is a critical point of f.*

a) If $D^2 f(y)$ is positive definite, in the sense that all eigenvalues of $D^2 f(y)$ are positive, then y is a local strict minimizer of f.

b) If $D^2 f(y)$ is negative definite, in the sense that all eigenvalues of $D^2 f(y)$ are negative, then y is a local strict maximizer of f.

c) If $D^2 f(y)$ is indefinite, in the sense that $D^2 f(y)$ has positive as well as negative eigenvalues, then y is a saddle point of f.

Proof. Recall from Remark 7.5, namely (143), that

$$D_V f(y + tW) = D_V f(y) + tW^\top D^2 f(y) V + e(y + tW)V, \tag{158}$$

with $e(y + tW) \to 0$ faster than tW. Putting $W = V$, let us write it as

$$D_V f(y + tV) = D_V f(y) + tV^\top D^2 f(y) V + e(y + tV)V. \tag{159}$$

Given any $\varepsilon > 0$, we have

$$|e(y + tV)V| \le \varepsilon t |V|_{\max}^2, \tag{160}$$

for all $V \in \mathbb{R}^n$ sufficiently small and for all $0 \le t \le 1$, where $|V|_{\max}$ is the absolute value of the largest component of V in absolute value:

$$|V|_{\max} = \max_k |V_k|. \tag{161}$$

Thus taking into account that $D_V f(y) = 0$, we get

$$|D_V f(y + tV) - tV^\top D^2 f(y) V| \le \varepsilon t |V|_{\max}^2, \tag{162}$$

Recall from Remark 7.6 that if $g(t) = f(y + tV)$ then

$$g'(t) = D_V f(y + tV), \tag{163}$$

so that

$$tV^\top D^2 f(y)V - \varepsilon t|V|_{\max}^2 \le g'(t) \le tV^\top D^2 f(y)V + \varepsilon t|V|_{\max}^2. \tag{164}$$

If we integrate this from $t = 0$ to $t = 1$, we get

$$\tfrac{1}{2}V^\top D^2 f(y)V - \tfrac{1}{2}\varepsilon|V|_{\max}^2 \le f(y+V) - f(y) \le \tfrac{1}{2}V^\top D^2 f(y)V + \tfrac{1}{2}\varepsilon|V|_{\max}^2, \tag{165}$$

meaning that for all sufficiently small $V \in \mathbb{R}^n$, the difference $f(y+V) - f(y)$ behaves like the quadratic form $\frac{1}{2}V^\top D^2 f(y)V$, up to the small error term $\frac{1}{2}\varepsilon|V|_{\max}^2$. □

Remark 8.5. Let us perform a more complete analysis of the case $n = 2$. To this end, let $A = \begin{pmatrix} a & b \\ b & c \end{pmatrix}$, and consider

$$u(x,y) = \begin{pmatrix} x \\ y \end{pmatrix}^\top \begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = ax^2 + 2bxy + cy^2. \tag{166}$$

We want to derive a criterion based on the determinant

$$D = ac - b^2, \tag{167}$$

that would tell us about the nature of the point $(x, y) = (0, 0)$.

First, suppose that $a = 0$, that is,

$$u(x,y) = 2bxy + cy^2 = y(2bx + cy). \tag{168}$$

If $b = 0$, or equivalently, $D = -b^2 = 0$, then

$$u(x,y) = cy^2, \tag{169}$$

meaning that $(x, y) = (0, 0)$ has a degenerate behaviour, because the point $(x, 0)$ for each x is a global minimizer. On the other hand, if $b \neq 0$, or equivalently, $D = -b^2 < 0$, then $(x, y) = (0, 0)$ is a saddle point, because $u(x, y)$ would be positive or negative according to whether y and $2bx + cy$ have the same sign.

Now suppose that $a \neq 0$, and write

$$\begin{aligned} u(x,y) &= a(x^2 + 2\frac{b}{a}xy) + cy^2 = a\big(x + \frac{b}{a}y\big)^2 - \frac{b^2}{a^2}y^2 + cy^2 \\ &= a\big(x + \frac{b}{a}y\big)^2 + \frac{ac - b^2}{a^2}y^2 = a\big(x + \frac{b}{a}y\big)^2 + \frac{D}{a^2}y^2. \end{aligned} \tag{170}$$

From this it is clear that

- A is positive definite iff $a > 0$ and $D > 0$.
- A is negative definite iff $a < 0$ and $D > 0$.
- A is indefinite iff $D < 0$. This is regardless of $a = 0$ or not.

Example 8.6. Consider

$$f(x,y) = xye^{-x^2-y^2}. \tag{171}$$

First, not that this function is smooth, meaning that we can take as many derivatives we want. We can compute the partial derivatives as

$$\partial_x f(x,y) = y(1 - 2x^2)e^{-x^2-y^2}, \qquad \partial_y f(x,y) = x(1 - 2y^2)e^{-x^2-y^2}, \tag{172}$$

which yields the 5 critical points given by

$$(x,y) = (0,0), \qquad (x,y) = (\pm\tfrac{1}{\sqrt{2}}, \pm\tfrac{1}{\sqrt{2}}). \tag{173}$$

Now we compute the Hessian as

$$H(x,y) = e^{-x^2-y^2} \begin{pmatrix} 2xy(2x^2 - 3) & 1 - 2x^2 - 2y^2 + 4x^2y^2 \\ 1 - 2x^2 - 2y^2 + 4x^2y^2 & 2xy(2y^2 - 3) \end{pmatrix}. \tag{174}$$

At $(x, y) = (0, 0)$, this is

$$H(0,0) = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \tag{175}$$

whose determinant is $D = \det H(0,0) = -1$. Hence $(x, y) = (0, 0)$ is a saddle point. At any of the other critical points $(x, y) = (\pm\frac{1}{\sqrt{2}}, \pm\frac{1}{\sqrt{2}})$, we have

$$H(x,y) = e^{-x^2-y^2} \begin{pmatrix} -4xy & 0 \\ 0 & -4xy \end{pmatrix}, \tag{176}$$

ans so $D = \det H(x, y) = 16x^2y^2 > 0$. Therefore, the nature of the point $(x, y) = (\pm\frac{1}{\sqrt{2}}, \pm\frac{1}{\sqrt{2}})$ will now be determined by the sign of the upper left entry, $a = -4xye^{-x^2-y^2}$. Thus, the points $\pm(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}})$ are local strict maximizers, and the points $\pm(\frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}})$ are local strict minimizers.

Remark 8.7. There is a general criterion for $n \times n$ matrices that extends what we have done in Remark 8.5 for 2×2 matrices. Let A be a symmetric $n \times n$ matrix, and let A_k be the $k \times k$ submatrix located in the upper left corner of A. Then we have the following.

- A is positive definite iff $\det A_k > 0$ for $k = 1, \ldots, n$.
- A is negative definite iff $(-1)^k \det A_k > 0$ for $k = 1, \ldots, n$.