The case $\boldsymbol{N = M - 1}$

In this case the image of $\boldsymbol{f}$ is a surface in $\mathbb{R}^M$. If we consider the tangent plane of this surface at $\boldsymbol{x}_0$, then clearly

$$\left\{\boldsymbol{y}_0 + t\,\frac{\partial \boldsymbol{f}}{\partial x_1}(\boldsymbol{x}_0) |\, t \in \mathbb{R}\right\}, \ldots, \left\{\boldsymbol{y}_0 + t\,\frac{\partial \boldsymbol{f}}{\partial x_N}(\boldsymbol{x}_0) |\, t \in \mathbb{R}\right\} \tag{1}$$

are N lines tangent to this surface at $\boldsymbol{x}_0$, which means the surface should be given through

$$\boldsymbol{n} \cdot (\boldsymbol{y} - \boldsymbol{y}_0) = 0 \tag{2}$$

where the vector $\boldsymbol{n} \in \mathbb{R}^M$ satisfies

$$\boldsymbol{n} \cdot \frac{\partial \boldsymbol{f}}{\partial x_i}(\boldsymbol{x}_0) = 0 \tag{3}$$

for all $i = 1, 2, \ldots, N$. This vector $\boldsymbol{n}$ is called the "normal direction" of the surface at $\boldsymbol{x}_0$.

Exercise 1. Prove that $\boldsymbol{n}$ is parallel to the following vector:

$$\det \begin{pmatrix} \boldsymbol{e}_1 & \frac{\partial f_1}{\partial x_1} & & \frac{\partial f_1}{\partial x_N} \\ \vdots & \vdots & \cdots & \vdots \\ \boldsymbol{e}_M & \frac{\partial f_M}{\partial x_1} & & \frac{\partial f_M}{\partial x_N} \end{pmatrix} \tag{4}$$

where $\boldsymbol{e}_1, \ldots, \boldsymbol{e}_M$ are the coordinate directions in $\mathbb{R}^M$. This determinant is formal, for example

$$\det \begin{pmatrix} \boldsymbol{e}_1 & 3 \\ \boldsymbol{e}_2 & 2 \end{pmatrix} = 2\,\boldsymbol{e}_1 - 3\,\boldsymbol{e}_2. \tag{5}$$

Example 1. Consider the function

$$\boldsymbol{f}(t) := \begin{pmatrix} \cos t \\ \sin t \end{pmatrix}. \tag{6}$$

We calculate

$$\det \begin{pmatrix} \boldsymbol{e}_1 & f_1'(t) \\ \boldsymbol{e}_2 & f_2'(t) \end{pmatrix} = (\cos t)\,\boldsymbol{e}_1 + (\sin t)\,\boldsymbol{e}_2. \tag{7}$$

Now if we denote $x = \cos t, y = \sin t$, then the vector $\boldsymbol{n}$ at $(x_0, y_0) = (\cos t_0, \sin t_0)$ is $x_0\,\boldsymbol{e}_1 + y_0\,\boldsymbol{e}_2 = \left(\begin{smallmatrix} x_0 \\ y_0 \end{smallmatrix}\right)$, the normal direction to the tangent.

Exercise 2. Consider the function

$$\boldsymbol{f}(\theta, \varphi) = \begin{pmatrix} \cos\theta\cos\varphi \\ \sin\theta\cos\varphi \\ \sin\varphi \end{pmatrix}. \tag{8}$$

Find the normal direction to its tangent plane at $\boldsymbol{x}_0 := \boldsymbol{f}(\theta_0, \varphi_0)$ and give geometric explanation of your result.

The case $\boldsymbol{M = 1}$

Recall that in this case our function takes the form

$$f(x_1, \ldots, x_N) \tag{9}$$

and the matrix representation of the differential $Df(\boldsymbol{x}_0)$ is a vector, which we will call the "gradient" of f at $\boldsymbol{x}_0$.

$$(\mathrm{grad}\, f)(\boldsymbol{x}_0) := \begin{pmatrix} \frac{\partial f}{\partial x_1}(\boldsymbol{x}_0) \\ \vdots \\ \frac{\partial f}{\partial x_N}(\boldsymbol{x}_0) \end{pmatrix}. \tag{10}$$

Remark 2. The case $M = 1$ can be related to this case as a special case as follows: Given $f: \mathbb{R}^N \mapsto \mathbb{R}$, we define a new function $\boldsymbol{g}: \mathbb{R}^N \mapsto \mathbb{R}^{N+1}$ as follows:

$$\boldsymbol{g}(x_1, \ldots, x_N) = \begin{pmatrix} x_1 \\ \vdots \\ x_N \\ f(x_1, \ldots, x_N) \end{pmatrix}. \tag{11}$$

However this identification is not very helpful for us in understanding the relation between the function and the gradient.

Remark 3. To avoid much confusion in future mathematics study, it is important to treat the linear function $Df(\boldsymbol{x}_0)$ and its matrix/vector representation $(\mathrm{grad}\, f)(\boldsymbol{x}_0)$ as different objects.

Remark 4. Often the notation ∇f is used for $\mathrm{grad}\, f$. However getting too used to this notation (∇f) will cause much difficulty in differential geometry.

Remark 5. We see that even when f is not differentiable, we may still be able to define the gradient vector – we only need the existence of all partial derivatives!

To understand the geometric meaning of the gradient $\mathrm{grad}\, f$ (not the differential Df!), we consider the graph of f, as a subset of $\mathbb{R}^{N+1}$:

$$x_{N+1} = f(x_1, \ldots, x_N). \tag{12}$$

Lemma 6. *Consider the surface $f(x_1, \ldots, x_N) = c$ for some $c \in \mathbb{R}$. Then the gradient vector* $\mathrm{grad}\, f$ *is perpendicular to this surface.*

Proof. Consider any curve $(x_1(t), \ldots, x_N(t))$ on the surface, that is

$$f(x_1(t), \ldots, x_N(t)) = c. \tag{13}$$

Taking $\frac{\mathrm{d}}{\mathrm{d}t}$ we have, thanks to the chain rule:

$$\frac{\partial f}{\partial x_1}(x_1(t), \ldots, x_N(t))\, x_1'(t) + \cdots + \frac{\partial f}{\partial x_N}(x_1(t), \ldots, x_N(t))\, x_N'(t) = 0 \tag{14}$$

which gives

$$(\mathrm{grad}\, f)(x_1(t), \ldots, x_N(t)) \cdot \boldsymbol{x}'(t) = 0 \tag{15}$$

and conclusion follows. $\square$

Remark 7. The justification of the claim: $f(x_1, \ldots, x_N) = c$ is a surface, needs Implicit function theorem, which we will discuss next week.

Properties and applications

Mean Value Theorem

Theorem 8. (MVT) *Let $f\colon \mathbb{R}^N \mapsto \mathbb{R}$. Let $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^N$ and assume f is differentiable at every point on the line segment $S := \{t\,\boldsymbol{x} + (1-t)\,\boldsymbol{y} |\, t \in [0,1]\}$. Then there is $\boldsymbol{\xi} \in S$ such that*

$$f(\boldsymbol{x}) - f(\boldsymbol{y}) = D f(\boldsymbol{\xi})(\boldsymbol{x} - \boldsymbol{y}). \tag{16}$$

Proof. Consider $g(t) := f(t\,\boldsymbol{x} + (1-t)\,\boldsymbol{y})$. □

Exercise 3. Fill in the details of the proof. In particular, why is g differentiable?

Corollary 9. *Let $f\colon E \mapsto \mathbb{R}$ be differentiable and E be convex. Then for any $\boldsymbol{x}, \boldsymbol{y} \in E$, there is $\boldsymbol{\xi} \in E$ such that*

$$f(\boldsymbol{x}) - f(\boldsymbol{y}) = D f(\boldsymbol{\xi})(\boldsymbol{x} - \boldsymbol{y}). \tag{17}$$

Problem 1. Critique the following claim:

> Let $\boldsymbol{f}\colon \mathbb{R}^N \mapsto \mathbb{R}^M$. Let $\boldsymbol{x}$, $\boldsymbol{y} \in \mathbb{R}^N$ and assume f is differentiable at every point on the line segment $S := \{t\,\boldsymbol{x} + (1-t)\,\boldsymbol{y} |\, t \in [0,1]\}$. Then there is $\boldsymbol{\xi} \in S$ such that
>
> $$\boldsymbol{f}(\boldsymbol{x}) - \boldsymbol{f}(\boldsymbol{y}) = D f(\boldsymbol{\xi})(\boldsymbol{x} - \boldsymbol{y}). \tag{18}$$

If it is correct, prove it; Otherwise find a counter-example. (Hint: You may want to consider the function $(\cos t, \sin t)$.

Problem 2. Critique the following claim:

> Let $\boldsymbol{f}\colon \mathbb{R}^N \mapsto \mathbb{R}^M$. Let $\boldsymbol{x}$, $\boldsymbol{y} \in \mathbb{R}^N$ and assume f is differentiable at every point on the line segment $S := \{t\,\boldsymbol{x} + (1-t)\,\boldsymbol{y} |\, t \in [0,1]\}$. Then
>
> $$\|\boldsymbol{f}(\boldsymbol{x}) - \boldsymbol{f}(\boldsymbol{y})\| \leqslant M\, \|\boldsymbol{x} - \boldsymbol{y}\| \tag{19}$$
>
> where M can be taken as
>
> $$M = \sup_{\boldsymbol{z} \in S} \left(\sum_{i,j} a_{ij}^2(\boldsymbol{z}) \right)^{1/2} \tag{20}$$
>
> with $A = (a_{ij})$ the Jacobian matrix of $\boldsymbol{f}$.

If it is correct, prove it; Otherwise find a counter-example.

Question 10. *Hadamard's theorem?*

A first look at optimization theory

In this subsection we focus on the case $f\colon E \subseteq \mathbb{R}^N \mapsto \mathbb{R}$.

Local and global optima

Definition 11. *Let $f\colon E \subseteq \mathbb{R}^N \mapsto \mathbb{R}$. $\boldsymbol{x}_0 \in E$ is said to be*

- *a global maximum of f if*

$$\forall \boldsymbol{x} \in E, \qquad f(\boldsymbol{x}_0) \geqslant f(\boldsymbol{x}). \tag{21}$$

- *a local maximum of f if there is $r > 0$ such that*

$$\forall \boldsymbol{x} \in B(\boldsymbol{x}_0, r) \cap E, \qquad f(\boldsymbol{x}_0) \geqslant f(\boldsymbol{x}). \tag{22}$$

- *a global minimum of f if*

$$\forall \boldsymbol{x} \in E, \qquad f(\boldsymbol{x}_0) \leqslant f(\boldsymbol{x}). \tag{23}$$

- *a local minimum of f if there is $r > 0$ such that*

$$\forall \boldsymbol{x} \in B(\boldsymbol{x}_0, r) \cap E, \qquad f(\boldsymbol{x}_0) \leqslant f(\boldsymbol{x}). \tag{24}$$

Theorem 12. *If $\boldsymbol{x}_0$ is a global maximum/minimum, then it is a local maximum/minimum.*

Exercise 4. Prove the above theorem.

Theorem 13. *Let $f: E \subseteq \mathbb{R}^N \mapsto \mathbb{R}$ be differentiable and $\boldsymbol{x}_0 \in E^o$ is a local maximum or minimum, then* $(\operatorname{grad} f)(\boldsymbol{x}_0) = \mathbf{0}$.

Proof. Wlog $\boldsymbol{x}_0$ is a local maximum. Assume the contrary, that is $\boldsymbol{v} := -(\operatorname{grad} f)(\boldsymbol{x}_0) \neq \mathbf{0}$. Then we have

$$\frac{\partial f}{\partial \boldsymbol{v}}(\boldsymbol{x}_0) = (D f)(\boldsymbol{x}_0)(\boldsymbol{v}) = (\operatorname{grad} f)(\boldsymbol{x}_0) \cdot (-\boldsymbol{v}) = -\|\boldsymbol{v}\|^2 < 0. \tag{25}$$

Now by definition of directional derivative:

$$\frac{\partial f}{\partial \boldsymbol{v}}(\boldsymbol{x}_0) := \lim_{h \longrightarrow 0} \frac{f(\boldsymbol{x}_0 + h\, \boldsymbol{v}) - f(\boldsymbol{x}_0)}{h} \tag{26}$$

we see that there is $\delta > 0$ such that for all $0 < h < \delta$,

$$f(\boldsymbol{x}_0 + h\, \boldsymbol{v}) < f(\boldsymbol{x}_0). \tag{27}$$

This contradicts $\boldsymbol{x}_0$ being local maximum. □

Example 14. Consider $f(x, y) = \sin x \sin y \sin(x + y)$. Find its maximum/minimum on

$$E := \{(x, y) | \, x \geqslant 0, y \geqslant 0, x + y \leqslant \pi\} \tag{28}$$

Solution. As f is continuous and E is closed and bounded and thus compact, we know that f reaches its maximum and minimum on E. There are two cases for $\boldsymbol{x}_0$ which is maximum/minimum: Either $\boldsymbol{x}_0 \in E^o$ and thus satisfy $(\operatorname{grad} f)(\boldsymbol{x}_0) = \mathbf{0}$, or $\boldsymbol{x}_0 \in \partial E$.

We first look for local maximum/minimum in the interior of the domain:

$$E^o := \{(x, y) | \, x > 0, y > 0, x + y < \pi\}. \tag{29}$$

Taking partial derivative we obtain

$$\frac{\partial f}{\partial x} = \cos x \sin y \sin(x + y) + \sin x \sin y \cos(x + y) = \sin(2\, x + y) \sin y \tag{30}$$

$$\frac{\partial f}{\partial y} = \sin(x + 2\, y) \sin x. \tag{31}$$

Setting both to 0 we have the following possible cases:

- $\sin(2x+y) = \sin(x+2y) = 0$: We must have $x+2y = y+2x = \pi \Longrightarrow x = y = \frac{\pi}{3}$.
- $\sin y = \sin(x+2y) = 0$: There is no solution in the interior.
- $\sin(2x+y) = \sin x = 0$: There is no solution in the interior.
- $\sin x = \sin y = 0$: There is no solution in the interior.

Summarizing, we see that in the interior the only candidiate is $\left(\frac{\pi}{3}, \frac{\pi}{3}\right)$ which gives $f\left(\frac{\pi}{3}, \frac{\pi}{3}\right) = -\frac{3\sqrt{3}}{8}$. On the other hand we have

$$\forall (x,y) \in \partial E, \qquad f(x,y) = 0. \tag{32}$$

Thus the maximizer of f is any point on the boundary, with maximum 0; The minimizer of f is $\left(\frac{\pi}{3}, \frac{\pi}{3}\right)$ with minimum $-\frac{3\sqrt{3}}{8}$.

Question 15. (Pareto optimal) *Consider functions $f_1, \ldots, f_M : \mathbb{R}^N \mapsto \mathbb{R}$. A point $\boldsymbol{x}_0 \in \mathbb{R}^N$ is called a "local Pareto maximizer" if there is $r > 0$ such that for all $\boldsymbol{x} \in B(\boldsymbol{x}_0, r)$, if there is $i \in \{1, 2, \ldots, M\}$ such that $f_i(\boldsymbol{x}) > f_i(\boldsymbol{x}_0)$, then there must be another $j \in \{1, 2, \ldots, M\}$ such that $f_i(\boldsymbol{x}_0) > f_i(\boldsymbol{x})$. Explore the possibility of getting a necessary condition in terms of the Jacobian matrix of $\boldsymbol{f} := \begin{pmatrix} f_1 \\ \vdots \\ f_M \end{pmatrix}$ for $\boldsymbol{x}_0 \in \mathbb{R}^N$ to be a local Pareto maximizer.*

Convex optimization

Definition 16. (Convex function) *A function $f: A \subseteq \mathbb{R}^N \mapsto \mathbb{R}$ is convex if and only if*

i. A is a convex set.

ii. $\forall \boldsymbol{x}, \boldsymbol{y} \in A$, and $\forall t \in [0,1]$, $f(t\,\boldsymbol{x} + (1-t)\,\boldsymbol{y}) \leqslant t\,f(\boldsymbol{x}) + (1-t)\,f(\boldsymbol{y})$.

Theorem 17. *If f is convex, then any of its local minimum is also global.*

Proof. Assume the contrary. Then there must be two local minimizers $\boldsymbol{x}, \boldsymbol{y}$ such that $f(\boldsymbol{x}) \neq f(\boldsymbol{y})$. Wlog assume $f(\boldsymbol{x}) < f(\boldsymbol{y})$. Now for any $r > 0$, take $t \in (0,1)$ such that

$$t > 1 - \frac{r}{\|\boldsymbol{x} - \boldsymbol{y}\|}. \tag{33}$$

Then we have

$$\|(t\,\boldsymbol{x} + (1-t)\,\boldsymbol{y}) - \boldsymbol{x}\| = (1-t)\,\|\boldsymbol{x} - \boldsymbol{y}\| < r \Longrightarrow t\,\boldsymbol{x} + (1-t)\,\boldsymbol{y} \in B(\boldsymbol{x}, r). \tag{34}$$

Now by convexity of f we have

$$f(t\,\boldsymbol{x} + (1-t)\,\boldsymbol{y}) \leqslant t\,f(\boldsymbol{x}) + (1-t)\,f(\boldsymbol{y}) < t\,f(\boldsymbol{x}) + (1-t)\,f(\boldsymbol{x}) = f(\boldsymbol{x}), \tag{35}$$

contradicting the assumption that $\boldsymbol{x}$ is a local minimizer. ☐

Theorem 18. *Let A be open and $f: A \subseteq \mathbb{R}^N \mapsto \mathbb{R}$ be differentiable. Then f is convex if and only if for all $\boldsymbol{x}, \boldsymbol{y} \in A$,*

$$(D f(\boldsymbol{x}) - D f(\boldsymbol{y}))(\boldsymbol{x} - \boldsymbol{y}) \geqslant 0. \tag{36}$$

Remark 19. Such $D f$ is said to be a "monotone".

Proof.

- If. Take any $\boldsymbol{x}, \boldsymbol{y} \in A$. Define

$$g(t) := f(t\,\boldsymbol{x} + (1-t)\,\boldsymbol{y}). \tag{37}$$

  All we need to show is for all $t \in (0,1)$,

$$g(t) \leqslant (1-t)\,g(0) + t\,g(1). \tag{38}$$

  Now fix any $t \in (0,1)$. By the one dimensional mean value theorem, we have

$$g(t) - g(0) = g'(\xi_1)\,t, \qquad g(1) - g(t) = g'(\xi_2)\,(1-t) \tag{39}$$

  where $\xi_1 \in (0,t)$, $\xi_2 \in (t,1)$. Then we have

$$(1-t)\,[g(t) - g(0)] + t\,[g(t) - g(1)] = t\,(1-t)[g'(\xi_1) - g'(\xi_2)]. \tag{40}$$

  So all we need to show is $g'(\xi_2) - g'(\xi_1) \geqslant 0$. To do this, apply the chain rule to g:

$$g'(t) = \frac{\mathrm{d} f(t\,\boldsymbol{x} + (1-t)\,\boldsymbol{y})}{\mathrm{d}t} = D f(t\,\boldsymbol{x} + (1-t)\,\boldsymbol{y})(\boldsymbol{x} - \boldsymbol{y}). \tag{41}$$

  Now, denoting $\boldsymbol{x}_i := \xi_i\,\boldsymbol{x} + (1-\xi_i)\,\boldsymbol{y}$

$$g'(\xi_2) - g'(\xi_1) = [D f(\boldsymbol{x}_2) - D f(\boldsymbol{x}_1)](\boldsymbol{x} - \boldsymbol{y}) = (\xi_2 - \xi_1)\,[D f(\boldsymbol{x}_2) - D f(\boldsymbol{x}_1)](\boldsymbol{x}_2 - \boldsymbol{x}_1) \tag{42}$$

  which $\geqslant 0$ due to $\xi_2 > \xi_1$.

- Only if. Take any $\boldsymbol{x}, \boldsymbol{y} \in A$. Define

$$g(t) := f(t\,\boldsymbol{x} + (1-t)\,\boldsymbol{y}). \tag{43}$$

  Then g is convex (see exercise below) and differentiable. Convexity of g now gives

$$g(t) \leqslant (1-t)\,g(0) + t\,g(1) \Longrightarrow \frac{g(t) - g(0)}{t} \leqslant g(1) - g(0) \tag{44}$$

  which means

$$g'(0) \leqslant g(1) - g(0). \tag{45}$$

  Similarly we have

$$\frac{g(t) - g(1)}{t-1} \geqslant g(1) - g(0) \Longrightarrow g'(1) \geqslant g(1) - g(0). \tag{46}$$

Thus

$$g'(1) \geqslant g'(0) \tag{47}$$

which translates back to:

$$D f(\boldsymbol{x})(\boldsymbol{x}-\boldsymbol{y}) \geqslant D f(\boldsymbol{y})(\boldsymbol{x}-\boldsymbol{y}) \tag{48}$$

and the proof ends. □

Exercise 5. Let $f: \mathbb{R}^N \mapsto \mathbb{R}$. Let $\boldsymbol{x}_0, \boldsymbol{v} \in \mathbb{R}^N$ and define $g: \mathbb{R} \mapsto \mathbb{R}$ by

$$g(t) = f(\boldsymbol{x}_0 + t\,\boldsymbol{v}). \tag{49}$$

a) Prove that if f is convex then g is convex.

b) If for every $\boldsymbol{x}_0, \boldsymbol{v} \in \mathbb{R}^N$ the g as defined above is convex, can we conclude f is convex? Justify.