# Lecture Notes - Multivariate Calculus

MA Math Camp 2022
Columbia University

Andrea Ciccarone*

This Version : August 24, 2022

## Contents

**1 Derivatives in one dimension** **2**
1.1 Definition . . . 2
1.2 First order expansions and affine approximations . . . 3
1.3 L'Hospital Rule . . . 4
1.4 Mean Value Theorem . . . 5

**2 Derivatives in higher dimensions** **6**
2.1 Total derivatives . . . 6
2.2 Partial Derivatives . . . 8
2.3 Directional Derivatives* . . . 10
2.4 Chain Rule . . . 11

**3 Higher Order Derivatives and Taylor Expansion** **12**
3.1 Second Order Derivatives of $f : A \subset \mathbb{R}^n \to \mathbb{R}$ . . . 12
3.2 $C^k$ functions . . . 14
3.3 Taylor's Theorem . . . 15

**4 Log-linearization** **16**

**5 Implicit Function Theorem and Inverse Function Theorem** **18**

In this lecture, we review some important concepts in multivariate calculus, skipping the proofs of many of the results. You may refer to Rudin's Chapter 5 and 9 for derivatives, and Chapter 4 of FMEA for integrals.

Unless stated otherwise explicitly, we use the Euclidean distance $d_2$ in $\mathbb{R}^k$ by default when talking about openness, closedness, compactness, limit, and continuity. Also, the product of two vectors in $\mathbb{R}^k$ is the dot product, and the norm $\|\cdot\|$ of a vector is the Euclidean norm, or $L_2$ norm.

*The present lecture notes were largely based on math camp materials from César Barilla, Palaash Bhargava, Paul Koh, and Xuan Li. All errors in this document are mine. If you find a typo or an error, please send me an email at ac4790@columbia.edu.

# 1 Derivatives in one dimension

## 1.1 Definition

**Definition 1.1.** *Let $A \subset \mathbb{R}$, and $x_0 \in A \cap A'$. A function $f : A \to \mathbb{R}$ is said to be **differentiable at** $x_0$ iff the limit*

$$\lim_{x \to x_0} \frac{f(x) - f(x_0)}{x - x_0}$$

*exists. In that case, define the **derivative of** $f$ **at** $x_0$ as the limit above, denoted as $f'(x_0)$.*

*A function $f : A \to \mathbb{R}$ is said to be **differentiable** iff $A \subset A'$ and $f$ is differentiable at any $x_0 \in A$.*

*Let $\hat{A}$ be the set of points in $A \cap A'$ at which $f$ is differentiable. Then the function $f' : \hat{A} \to \mathbb{R}$ is called the **derivative (function)** of $f$.*

**Example 1.2.** *A constant function over an interval $f : I \to \mathbb{R}$ is differentiable with $f' = 0$. The identity function $Id$ over $\mathbb{R}$ is differentiable with $Id'(x) = 1$ for all $x \in \mathbb{R}$. The square root is differentiable at every $x > 0$; to see it, observe that :*

$$\frac{\sqrt{x+h} - \sqrt{x}}{h} = \frac{\sqrt{x+h} - \sqrt{x}}{x + h - x} = \frac{1}{\sqrt{x+h} + \sqrt{x}}$$

*and :*

$$\lim_{h \to 0} \frac{1}{\sqrt{x+h} + \sqrt{x}} = \frac{1}{2\sqrt{x}}$$

Observe that we can give an equivalent definition of the derivative using a change of variable :

$$f'(x_0) = \lim_{h \to 0} \frac{f(x_0 + h) - f(x_0)}{h}$$

As clearly $x_0 + h \xrightarrow[h \to 0]{} x_0$. We will use both definitions alternatively. The first one was chosen in the definition above because it provides more intuition but the second one is often more convenient to manipulate.

Clearly, if a function $f$ is differentiable at $x$, then it is continuous at $x$. This is because

$$\begin{aligned}
\lim_{x' \to x} \left[ f(x') - f(x) \right] &= \lim_{x' \to x} \left[ \frac{f(x') - f(x)}{x' - x} \cdot (x' - x) \right] \\
&= \lim_{x' \to x} \left[ \frac{f(x') - f(x)}{x' - x} \right] \cdot \lim_{x' \to x} \left[ x' - x \right] \\
&= f'(x) \cdot 0 = 0
\end{aligned}$$

**Theorem 1.3.** *If $f : A \to \mathbb{R}$ is differentiable at $x \in A$, then it is continuous at $x$.*

A function continuous at $x$ may fail to be differentiable at $x$, since the function may have a kink point. In fact, a function can be continuous everywhere, but not differentiable at a single point (e.g. Weierstrass function).

Derivatives of some commonly used functions:

$$\begin{aligned}
(x^\alpha)' &= \alpha x^{\alpha - 1} \\
(\ln x)' &= 1/x \\
(e^x)' &= e^x \\
(\sin x)' &= \cos x
\end{aligned}$$

Be aware that the formula $(x^\alpha)' = \alpha x^{\alpha-1}$ does not work at $x = 0$ if $\alpha \leq 1$.

Because a derivative is essentially the limit of the slope function $\frac{f(x+h)-f(x)}{h}$ when the deviation $h$ tends to 0, it inherits the properties of limits of functions. Especially, if $f$ and $g$ are both differentiable at $x$, then $f+g$ is also differentiable at $x$, and $(f+g)'(x) = f'(x) + g'(x)$. This is because

$$\begin{aligned}
(f+g)'(x) &= \lim_{h\to 0} \frac{(f+g)(x+h) - (f+g)(x)}{h} \\
&= \lim_{h\to 0} \frac{f(x+h) + g(x+h) - f(x) - g(x)}{h} \\
&= \lim_{h\to 0} \left[\frac{f(x+h) - f(x)}{h} + \frac{g(x+h) - g(x)}{h}\right] \\
&= \lim_{h\to 0} \frac{f(x+h) - f(x)}{h} + \lim_{h\to 0} \frac{g(x+h) - g(x)}{h} \\
&= f'(x) + g'(x)
\end{aligned}$$

Notice that the second last equality uses the property of limits of functions:

$$\lim_{x\to x_0} \left[s(x) + t(x)\right] = \lim_{x\to x_0} s(x) + \lim_{x\to x_0} t(x)$$

Also, it can be shown that

$$\begin{aligned}
(\lambda f)' &= \lambda f' \\
(fg)' &= f'g + fg' \\
\big(f/g\big)' &= \frac{f'g - fg'}{g^2}
\end{aligned}$$

## 1.2 First order expansions and affine approximations

To give some interpretation for the concept of derivatives, we introduce first order expansions.

**Definition 1.4.** *Let $f : A \to \mathbb{R}$ and $x_0 \in A \cap A'$. We say that $f$ admits a first order expansion around $x$ if there exists $a, b \in \mathbb{R}$ and a function $\varepsilon : A \to \mathbb{R}$ such that :*

$$\begin{aligned}
&\forall x \in A, f(x) = a + b(x - x_0) + (x - x_0)\varepsilon(x) \\
\textit{and } &\lim_{x\to x_0} \varepsilon(x) = 0
\end{aligned}$$

We now give a theorem relating first order expansions and derivatives

**Theorem 1.5.** *Let $f : A \to \mathbb{R}$ and $x_0 \in A \cap A'$. The following are equivalent :*

*(i) $f$ is differentiable at $x_0$*

*(ii) $f$ has a first order expansion at $x_0$*

*Furthermore the coefficients of the first order expansion when they exist are $a = f(x_0)$, $b = f'(x_0)$.*

*Proof.* First assume that f is differentiable at $x_0$. Define :

$$\varepsilon(x) := \begin{cases} \frac{f(x)-f(x_0)}{x-x_0} - f'(x_0) & \text{if } x \neq x_0 \\ 0 & \text{if } x = x_0 \end{cases}$$

$\varepsilon$ goes to zero as $x$ goes to $x_0$ and we have by construction :

$$\forall x \in A, f(x) = f(x_0) + f'(x_0)(x - x_0) + (x - x_0)\varepsilon(x)$$

Now assume conversely that $f$ has a first order expansion at $x_0$, i.e

$$\forall x \in A, f(x) = a + b(x - x_0) + (x - x_0)\varepsilon(x)$$

Since $x_0 \in A$, this implies in particular $f(x_0) = a$. Hence for $x \in A \setminus \{x_0\}$, we can write :

$$\frac{f(x) - f(x_0)}{x - x_0} = \frac{f(x) - a}{x - x_0} = b + \varepsilon(x) \xrightarrow[x \to x_0]{} b$$

Hence $f'(x_0) = b$. □

The function $x \mapsto f(x_0) + (x - x_0)f'(x_0)$ is called *the affine approximation* of $f$ at $x_0$. The affine approximation has a geometric interpretation : it is the tangent of the curve of $f$ at $x_0$.

### 1.3 L'Hospital Rule

Define the **extended real line** $\bar{\mathbb{R}} := \mathbb{R} \cup \{+\infty, -\infty\}$, where $+\infty$ and $-\infty$ are two abstract objects that are not in $\mathbb{R}$. Extend the order $\leq$ s.t. $+\infty > a$ and $-\infty < a$ for any $a \in \mathbb{R}$. Note that $\left(\bar{\mathbb{R}}, \leq\right)$ is a totally ordered set, but $\bar{\mathbb{R}}$ is not a metric space because the distance between $\pm\infty$ and real numbers cannot be well defined, since the distance between two points in a metric space cannot be $+\infty$.

It is often useful to abuse the notation lim to allow divergence to $+\infty$ or $-\infty$. For example, the notation

$$\lim_{x \to a} f(x) = -\infty$$

means that $\forall\ M \in \mathbb{R}$, $\exists\ \delta > 0$ s.t. $f(x) < M$ for any $x \in B_\delta(a)$. In this case, we say that $f(x)$ **diverges** to $-\infty$ as $x$ converges to $a \in \mathbb{R}$. In this case, we usually don't say that $f(x)$ converges to $-\infty$, because this does not fit in our definition of convergence to a limit, since the object $-\infty$ is not even in the metric space $(\mathbb{R}, d_2)$.

We also allow the argument $x$ to diverge to $+\infty$ or $-\infty$. For example, the notation

$$\lim_{x \to -\infty} f(x) = +\infty$$

means that $\forall\ M \in \mathbb{R}$, $\exists\ N \in \mathbb{R}$ s.t. $f(x) > M$ for any $x < N$. In this case, we say that $f(x)$ diverges to $-\infty$ as $x$ diverges to $+\infty$.

Using mean value theorem, it is not difficult to obtain the following result, which is known as L'Hospital rule.

**Theorem 1.6** (L'Hospital Rule)**.** *Let $-\infty \leq a < b \leq +\infty$, and $f : (a, b) \to \mathbb{R}$ and $g : (a, b) \to \mathbb{R} \setminus \{0\}$ are differentiable in $(a, b)$. If $\lim_{x \to a} f(x)$ and $\lim_{x \to a} g(x)$ are both $0$ or $\pm\infty$, and $\lim_{x \to a} f'(x)/g'(x)$ has a finite value or is $\pm\infty$, then*

$$\lim_{x \to a} \frac{f(x)}{g(x)} = \lim_{x \to a} \frac{f'(x)}{g'(x)}$$

*The statement is also true for $x \to b$.*

See Rudin's Theorem 5.13 for a proof.

L'Hospital rule is particularly useful in obtaining the limit of some particular expression. For example, it might seem difficult to determine the behavior of the function $(\ln x)/\sqrt{x}$ when $x$ diverges to $+\infty$, because both the numerator and the denominator diverge to $+\infty$. However, because

$$\frac{(\ln x)'}{\left(\sqrt{x}\right)'} = \frac{\frac{1}{x}}{\frac{1}{2\sqrt{x}}} = \frac{2}{\sqrt{x}} \to 0$$

as $x \to +\infty$, we have $\lim_{x\to+\infty} (\ln x)/\sqrt{x} = 0$ by L'Hospital rule.

According to the theorem, L'Hospital rule applies to functions with the form $0/0$ or $\infty/\infty$, i.e. both the numerator and the denominator converges/diverges to 0 or $\pm\infty$. When a function does not have this form, it must be transformed to this form before L'Hospital rule can be applied. For example, consider the limit $\lim_{x\to+\infty} \left(1+x^{-1}\right)^x$. It does not have the form $0/0$ or $\infty/\infty$, but its log

$$\ln\left(1+x^{-1}\right)^x = x\ln\left(1+x^{-1}\right) = \frac{\ln\left(1+x^{-1}\right)}{x^{-1}}$$

takes the form $0/0$, to which L'Hospital rule can be applied. This is left as an exercise.

## 1.4 Mean Value Theorem

Mean value theorem is an important result that has many useful implications.

**Theorem 1.7** (Mean Value Theorem)**.** *Let $f : [a,b] \to \mathbb{R}$, differentiable on $(a,b)$, and continuous on $[a,b]$. Then there exists $x \in (a,b)$ s.t.*

$$f'(x) = \frac{f(b) - f(a)}{b-a}$$

Notice that $\frac{f(b)-f(a)}{b-a}$ is the slope of the line connecting $\left(a, f(a)\right)$ and $\left(b, f(b)\right)$, the two end points of the graph of $f$.

*Proof.* First, let's consider the special case in which $f(a) = f(b)$, and we want to find $x \in (a,b)$ s.t. $f'(x) = 0$.

Because $[a,b]$ is compact and $f$ is continuous on $[a,b]$, the function $f$ has a maximum and a minimum. Also, there must exist $x \in (a,b)$ s.t. $f$ achieves its maximum or minimum at $x$. I want to show that $f'(x) = 0$.

Suppose $f$ achieves its maximum at $x \in (a,b)$. Arbitrarily take a sequence $(x_n)$ convergent to $x$ s.t. $x_n < x$ for any $n$. Then we have

$$f'(x) = \lim_{\tilde{x}\to x} \frac{f(\tilde{x}) - f(x)}{\tilde{x} - x} = \lim_{n\to\infty} \frac{f(x_n) - f(x)}{x_n - x} \geq 0$$

where the last inequality is because $f(x_n) - f(x) \leq 0$, $x_n - x < 0$, and $\leq$ is preserved in the limit.

Arbitrarily take a sequence $\left(x'_n\right)$ convergent to $x$ s.t. $x'_n > x$ for any $n$. Then we have

$$f'(x) = \lim_{\tilde{x}\to x} \frac{f(\tilde{x}) - f(x)}{\tilde{x} - x} = \lim_{n\to\infty} \frac{f\left(x'_n\right) - f(x)}{x'_n - x} \leq 0$$

Therefore, we must have $f'(x) = 0$ if $f$ achieves its maximum at $x \in (a,b)$. Symmetrically, we can show that $f'(x) = 0$ if $f$ achieves its minimum at $x \in (a,b)$. So we have proved the special case of the theorem where $f(a) = f(b)$.

If $f(a)$ and $f(b)$ are not necessarily equal, we can subtract the linear trend of $f$ and define a new function $g:[a,b]\to\mathbb{R}$ as

$$g(x):=f(x)-\frac{f(b)-f(a)}{b-a}x$$

By definition, we have $g(a)=g(b)$. Applying the special case we have proved, we can find $x^*\in(a,b)$ s.t. $g'(x^*)=0$. Because

$$g'(x)=f'(x)-\frac{f(b)-f(a)}{b-a}$$

for any $x\in(a,b)$, we have

$$f'\left(x^*\right)=\frac{f(b)-f(a)}{b-a}$$

□

One implication of mean value theorem is: if $f'>0$ on $(a,b)$, then $f$ is strictly increasing on $(a,b)$. To see this, take any $x_1,x_2\in(a,b)$ s.t. $x_1<x_2$. By mean value theorem, there exists $x\in(x_1,x_2)$ s.t.

$$f'(x)=\frac{f(x_2)-f(x_1)}{x_2-x_1}$$

Therefore, $f(x_2)-f(x_1)=f'(x)\cdot(x_2-x_1)>0$. Similarly, if $f'<0$ on $(a,b)$, then $f$ is strictly decreasing on $(a,b)$. If we have $f'\geq 0$ $(\leq 0)$, then $f$ is weakly increasing (decreasing) on $(a,b)$.

Finally, here is a similar theorem that does not require differentiability.

**Theorem 1.8** (Intermediate Value Theorem)**.** *Let $f:[a,b]\to\mathbb{R}$ continuous and $u$ is a number between $f(a)$ and $f(b)$, then there exists $c\in[a,b]$ s.t. $u=f(c)$.*

# 2 Derivatives in higher dimensions

## 2.1 Total derivatives

Now we generalize the notion of derivatives to multivariate functions.

**Definition 2.1.** *Let $A\subset\mathbb{R}^n$ and $x\in int(A)$. A function $f:A\to\mathbb{R}^m$ is said to be **differentiable at** $x$ iff $\exists$ an $m\times n$ real matrix $C$ s.t.*

$$\lim_{h\to 0}\frac{\left\|f(x+h)-f(x)-Ch\right\|}{\|h\|}=0$$

*In this case, define the **(total) derivative of** $f$ **at** $x$ as the matrix $C$, denoted as $f'(x)$, or $Df(x)$.*

*A function $f:A\to\mathbb{R}$ is said to be **differentiable** iff $A$ is open and $f$ is differentiable at any $x\in A$.*

*Let $A_1\subset int(A)$ be the set of points at which $f$ is differentiable. Then the function $f':A_1\to\mathbb{R}^{mn}$ is called the **derivative (function)** of $f$.*

It can be shown that the real matrix $C$ in the definition above is unique, if exists. See Rudin's Theorem 9.12 for a proof. Therefore, it is without ambiguity to talk about "the" derivative and to use the notation $f'(x)$ or $Df(x)$.

According to the definition, the derivative of function $f$ from a set in $\mathbb{R}^n$ to $\mathbb{R}^m$ is an $m\times n$ matrix $C$. This matrix $C$ should be interpreted as a linear mapping from $\mathbb{R}^n$ to $\mathbb{R}^m$, i.e. a function

that maps $h \in \mathbb{R}^n$ to $Ch \in \mathbb{R}^m$. By definition, the matrix $C$ is the derivative of $f$ iff the linear function $f(x)+Ch$ of $h$ approximates $f(x+h)$ well when $h \in \mathbb{R}^n$ is close to 0, in the sense that the approximation error is $o(\|h\|)$. If we consider $h$ as the deviation of $x'$ from $x$, clearly the function $f(x) + f'(x)(x'-x)$ in $x'$ is the linear approximation of $f(x')$ in the neighborhood of $x$.

Because the $m \times n$ real matrix $C$ can also be viewed as an $mn$-dimensional real vector, the codomain of the derivative function $f'$ can be viewed as $\mathbb{R}^{mn}$.

For a real-valued function $f$ from $A \subset \mathbb{R}^n$ to $\mathbb{R}$, its derivative $f'(x)$ at $x \in int(A)$ reduces to a $1 \times n$ row vector. In this case, the derivative is also called the **gradient** of $f$ at $x$, sometimes denoted as $\nabla f(x)$, which is essentially the same as $f'(x)$ or $Df(x)$.

Clearly, if a function $f$ from $A \subset \mathbb{R}^n$ to $\mathbb{R}^m$ is differentiable at $x \in int(A)$, then it is continuous at $x$. To see this, by triangle inequality of $\|\cdot\|$,

$$0 \leq \left\|f(x') - f(x)\right\| \leq \left\|f(x') - f(x) - f'(x)(x'-x)\right\| + \left\|f'(x)(x'-x)\right\|$$

Because the first term

$$\begin{aligned}\left\|f(x') - f(x) - f'(x)(x'-x)\right\| &= \frac{\left\|f(x') - f(x) - f'(x)(x'-x)\right\|}{\|x'-x\|}\|x'-x\| \\ &\to 0 \cdot 0 = 0\end{aligned}$$

as $x' \to x$, and the second term

$$\left\|f'(x)(x'-x)\right\| \to \|f'(x)(x-x)\| = 0$$

as $x' \to x$, we know that $\left\|f(x') - f(x)\right\| \to 0$ as $x' \to x$. Therefore, $f$ is continuous at $x$.

If two functions $f$ and $g$ from $A \subset \mathbb{R}^n$ to $\mathbb{R}^m$ are both differentiable at $x \in int(A)$, then the function $f : A \to \mathbb{R}^m$ is also differentiable at $x$, and furthermore we have $(f+g)'(x) = f'(x)+g'(x)$. To see this, observe that

$$\begin{aligned}0 &\leq \frac{\left\|(f+g)(x+h) - (f+g)(x) - \left(f'(x)+g'(x)\right)h\right\|}{\|h\|} \\ &= \frac{\left\|f(x+h) - f(x) - f'(x)h + g(x+h) - g(x) - g'(x)h\right\|}{\|h\|} \\ &\leq \frac{\left\|f(x+h) - f(x) - f'(x)h\right\|}{\|h\|} + \frac{\left\|g(x+h) - g(x) - g'(x)h\right\|}{\|h\|}\end{aligned}$$

Also, we have $\left\|f(x+h) - f(x) - f'(x)h\right\| / \|h\| \to 0$ and $\left\|g(x+h) - g(x) - g'(x)h\right\| / \|h\| \to 0$ as $h \to 0$, by definition of $f'(x)$ and $g'(x)$. Therefore,

$$\lim_{h \to 0} \frac{\left\|(f+g)(x+h) - (f+g)(x) - \left(f'(x)+g'(x)\right)h\right\|}{\|h\|} = 0$$

which means the $m \times n$ matrix $f'(x) + g'(x)$ satisfies the definition of $(f+g)'(x)$.

Similarly, we can show $(\lambda f)' = \lambda f'$. Therefore, taking derivative is a linear operator, i.e.

$$(\lambda_1 f_1 + \lambda_2 f_2)'(x) = \lambda_1 f_1'(x) + \lambda_2 f_2'(x)$$

For a function $f$ from $A \subset \mathbb{R}^n$ to $\mathbb{R}^m$, each coordinate $i \in \{1, \ldots, m\}$ of $f$ can be regarded as a function $f_i$ from $A$ to $\mathbb{R}$. By definition, it is straightforward to show that $f$ is differentiable at $x \in int(A)$ iff $f_i$ is differentiable at $x$ for each $i$, and furthermore we have

$$f'(x) = \begin{bmatrix} \nabla f_1(x) \\ \nabla f_2(x) \\ \vdots \\ \nabla f_m(x) \end{bmatrix}$$

## 2.2 Partial Derivatives

**Definition 2.2.** *Let $A \subset \mathbb{R}^n$ and $x \in int(A)$. For a function $f : A \to \mathbb{R}^m$, its **partial derivative of the $i$-th coordinate w.r.t. the $j$-th argument** at $x \in A$ is*

$$\frac{\partial f_i}{\partial x_j}(x) := \frac{d}{dt} f_i\left(x + te_j\right)\Big|_{t=0}$$

*if the right-hand side derivative exists.*

*The vector $e_j$ above is the $j$-th canonical basis of $\mathbb{R}^n$, i.e. $e_j := (0, \ldots, 1, \ldots, 0)$.*

In the expression $\frac{d}{dt} f_i\left(x + te_j\right)\Big|_{t=0}$, we fix $x$ and consider $f_i\left(x + te_j\right)$ as a single variable function in $t$, then take derivative of this single variable function, and finally evaluate the derivative at $t = 0$. In other words, the definition of this expression is

$$\frac{d}{dt} f_i\left(x + te_j\right)\Big|_{t=0} := g'(0)$$

where $g(t) := f_i\left(x + te_j\right)$.

The vector $x + te_j$ is a deviation from $x$ only in the $j$-th argument. Therefore, intuitively, the partial derivative $\frac{\partial f_i}{\partial x_j}(x)$ measures the sensitivity of the $i$-th coordinate $f_i$ of the function $f$ w.r.t. the $j$-th argument $x_j$.

Notice that the notation $\frac{\partial f_i}{\partial x_j}$ stands for a function from the set of points at which this partial derivative exists to $\mathbb{R}$, and it should be considered as an inseparable notation.

The next Theorem reveals the relation between the total derivative and the partial derivatives. Namely, the total derivative is a matrix that collects all partial derivatives as its entries.

**Theorem 2.3.** *Let $A \subset \mathbb{R}^n$ and $x \in int(A)$. If function $f : A \to \mathbb{R}^m$ is differentiable at $x$, then $\frac{\partial f_i}{\partial x_j}(x)$ exists for any $(i, j) \in \{1, \ldots m\} \times \{1, \ldots, n\}$, and furthermore we have*

$$f'(x) = \begin{bmatrix} \frac{\partial f_1}{\partial x_1}(x) & \frac{\partial f_1}{\partial x_2}(x) & \cdots & \frac{\partial f_1}{\partial x_n}(x) \\ \frac{\partial f_2}{\partial x_1}(x) & \frac{\partial f_2}{\partial x_2}(x) & \cdots & \frac{\partial f_2}{\partial x_n}(x) \\ \vdots & \vdots & & \vdots \\ \frac{\partial f_m}{\partial x_1}(x) & \frac{\partial f_m}{\partial x_2}(x) & \cdots & \frac{\partial f_m}{\partial x_n}(x) \end{bmatrix}$$

*Proof.* Let $\left(f'(x)\right)_i$ be the $i$-th row of the matrix $f'(x)$, and $\left(f'(x)\right)_{ij}$ be the $(i, j)$-th entry of $f'(x)$.

WTS: $\frac{\partial f_i}{\partial x_j}(x)$ exists and $\frac{\partial f_i}{\partial x_j}(x) = \left(f'(x)\right)_{ij}$ for any $(i, j) \in \{1, \ldots m\} \times \{1, \ldots, n\}$.

Take any $(i, j) \in \{1, \ldots m\} \times \{1, \ldots, n\}$. By definition of $f'(x)$, we have

$$\lim_{h \to 0} \frac{\left\| f(x + h) - f(x) - f'(x) h \right\|}{\|h\|} = 0$$

Because

$$0 \leq \left| f_i(x+h) - f_i(x) - \left(f'(x)\right)_i h \right| \leq \left\| f(x+h) - f(x) - f'(x)\, h \right\|$$

we have

$$\lim_{h \to 0} \frac{\left| f_i(x+h) - f_i(x) - \left(f'(x)\right)_i h \right|}{\|h\|} = 0$$

Because $te_j \to 0$ as $t \to 0$, we have

$$\lim_{t \to 0} \frac{\left| f_i\left(x + te_j\right) - f_i(x) - \left(f'(x)\right)_i \cdot te_j \right|}{\left\| te_j \right\|} = 0$$

i.e.

$$\lim_{t \to 0} \left| \frac{f_i\left(x + te_j\right) - f_i(x) - \left(f'(x)\right)_{ij} \cdot t}{t} \right| = 0$$

This implies

$$\lim_{t \to 0} \frac{f_i\left(x + te_j\right) - f_i(x)}{t} = \left(f'(x)\right)_{ij}$$

and LHS is exactly the definition of $\frac{\partial f_i}{\partial x_j}(x)$. □

Notice that the theorem above only states that existence of the total derivative implies the existence of all partial derivatives. The reverse is not true, since we can find a function $f$ s.t. $\frac{\partial f_i}{\partial x_j}(x)$ exists for all $(i,j) \in \{1, \ldots m\} \times \{1, \ldots, n\}$, but $f$ is not differentiable at $x$, i.e. its total derivative does not exist. In fact, $f$ may even be discontinuous at $x$. See the example below.

**Example 2.4.** *Consider the function $f : \mathbb{R}^2 \to \mathbb{R}$ defined as*

$$f(x,y) := \begin{cases} \frac{x^2 y}{x^4 + y^2}, \text{ if } (x,y) \neq (0,0) \\ 0, \text{ if } (x,y) = (0,0) \end{cases}$$

*By definition of partial derivatives, we have*

$$\begin{aligned} \frac{\partial f}{\partial x}(0,0) &= \left. \frac{d}{dt} f(t,0) \right|_{t=0} = \lim_{t \to 0} \frac{f(t,0) - f(0,0)}{t - 0} \\ &= \lim_{t \to 0} \frac{0 - 0}{t - 0} = 0 \end{aligned}$$

*and*

$$\begin{aligned} \frac{\partial f}{\partial y}(0,0) &= \left. \frac{d}{dt} f(0,t) \right|_{t=0} = \lim_{t \to 0} \frac{f(0,t) - f(0,0)}{t - 0} \\ &= \lim_{t \to 0} \frac{0 - 0}{t - 0} = 0 \end{aligned}$$

*So both of the partial derivatives of $f$ exist.*

*However, $f$ is not differentiable at $(0,0)$. In fact, $f$ is not even continuous at $(0,0)$. To see this, notice that $f$ constantly take the value $1/2$ along the path $y = x^2$ except for at the point $(0,0)$, where the $f$ takes the value $0$.*

As is shown in the example above, the existence of $\frac{\partial f_i}{\partial x_j}(x)$ for all $(i,j)$ does not imply differentiability of $f$ at $x$. However, if for each $(i,j)$, the partial $\frac{\partial f_i}{\partial x_j}(x)$ exists not only at $x$, but also on an open ball around $x$, and $\frac{\partial f_i}{\partial x_j}(x)$ is continuous at $x$, then $f$ is differentiable at $x$. This result is formulated by following theorem.

**Theorem 2.5.** *Let $A \subset \mathbb{R}^n$, $x \in int(A)$, and function $f : A \to \mathbb{R}^m$. Then $f$ is $C^1$ at $x$ iff $\frac{\partial f_i}{\partial x_j}(x)$ exists on an open ball around $x$ and is continuous at $x$ for any $(i,j) \in \{1,\dots m\} \times \{1,\dots,n\}$.*

The "only if" part is trivial by Theorem 2.3. The "if" part is essentially about differentiability, since once we have shown $f$ is differentiable at $x$, it is trivial to show $f'(x)$ is continuous at $x$ because each partial is continuous. See Rudin's Theorem 9.21 for a proof.

It is typically difficult to find the total derivative of a function $f$, since we need to find a $m \times n$ matrix that satisfies the limit condition specified by the definition. However, the $mn$ partial derivatives are much easier to find, since they are essentially derivatives of single variable derivatives. Therefore, to find the total derivative of a function at $x$, we usually don't directly work with the definition of total derivatives. Instead, we look at all partial derivatives of $f$ and see if all of them exist in an open ball around $x$ and are continuous at $x$. If yes, then by the theorem above we know that the total derivative exists at $x$, and is exactly the matrix of all partial derivatives at $x$.

We can generalize the mean value theorem to functions mapping from $A \subset \mathbb{R}^n$ to $\mathbb{R}$.

**Theorem 2.6.** *Let $f : A \subset \mathbb{R}^n$ is $C^1$ in an open set in $A$ which contains $[\mathbf{x},\mathbf{y}]$ ($x_i < y_i, \forall i = 1,2,...,n$). Then there exists a point $\mathbf{w}$ in $(\mathbf{x},\mathbf{y})$ (i.e. $x_i < w_i < y_i, \forall i = 1,2,...,n$) s.t.*

$$f(\mathbf{x}) - f(\mathbf{y}) = \nabla f(\mathbf{w}) \cdot (\mathbf{x} - \mathbf{y})$$

## 2.3 Directional Derivatives*

The concept of directional derivatives is a generalization of partial derivatives.

**Definition 2.7.** *Let $A \subset \mathbb{R}^n$ and $x \in int(A)$. For a function $f : A \to \mathbb{R}^m$ and a vector $z \in \mathbb{R}^n$ with $\|z\| = 1$, the **directional derivative of $f$ along the vector $z \in \mathbb{R}^n$ at $x \in A$** is*

$$f'_z(x) := \frac{d}{dt} f(x+tz)\Big|_{t=0} = \begin{bmatrix} \frac{d}{dt} f_1(x+tz)\Big|_{t=0} \\ \frac{d}{dt} f_2(x+tz)\Big|_{t=0} \\ \vdots \\ \frac{d}{dt} f_m(x+tz)\Big|_{t=0} \end{bmatrix}$$

*if the right-hand side derivative exists.*

If we let $z = e_j$, clearly by definition, we have

$$(f_i)'_{e_j}(x) = \frac{\partial f_i}{\partial x_j}(x)$$

i.e. the directional derivative of $f_i$ along the vector $e_j$ is exactly the partial derivative of $f_i$ w.r.t. $x_j$.

If $f$ is differentiable at $x$, then we know that all of its directional derivatives exist, and furthermore we have $f'_z(x) = f'(x) \cdot z$, where $z \in \mathbb{R}^n$ is considered as a column vector. To see this, consider the function $g(t) := x + tz$, we have $f(x + tz) = (f \circ g)(t)$. Then we have

$$\begin{aligned} f'_z(x) &= \frac{d}{dt} f(x + tz)\Big|_{t=0} = (f \circ g)'(0) \\ &= f'\big(g(0)\big) \cdot g'(0) = f'(x) \cdot z \end{aligned}$$

where the third equality above is because of the chain rule.

A function may not be differentiable at $x$ even if its directional derivative at $x$ exists for all directions. In fact, the function may even be discontinuous at $x$. Again consider the function $f : \mathbb{R}^2 \to \mathbb{R}$ in Example 2.4

$$f(x, y) := \begin{cases} \frac{x^2 y}{x^4 + y^2}, \text{ if } (x, y) \neq (0, 0) \\ 0, \text{ if } (x, y) = (0, 0) \end{cases}$$

For any direction $z \in \mathbb{R}^2$ with $\|z\| = 1$, we have

$$\begin{aligned} f'_z(0, 0) &= \frac{d}{dt} f(tz_1, tz_2)\Big|_{t=0} = \lim_{h \to 0} \frac{f(hz_1, hz_2) - f(0, 0)}{h} \\ &= \lim_{h \to 0} \frac{\frac{h^2 z_1^2 \cdot h z_2}{h^4 z_1^4 + h^2 z_2^2} - 0}{h} = \lim_{h \to 0} \frac{z_1^2 z_2}{h^2 z_1^4 + z_2^2} \end{aligned}$$

If $z_2 = 0$, then $f'_z(0, 0) = \lim_{h \to 0} 0 = 0$. If $z_2 \neq 0$, we have $f'_z(0, 0) = z_1^2 / z_2$. Therefore, the directional derivative $f'_z(0, 0)$ exists for all direction $z$, but $f$ is not even continuous at $(0, 0)$.

Consider a function $f$ from $A \subset \mathbb{R}^n$ to $\mathbb{R}$ and $x \in int(A)$, the gradient $\nabla f(x)$ can be interpreted as the direction in which $f$ increases the fastest at $x$. This is formulated in the proposition below.

**Proposition 2.8.** *Let $f$ be a function from $A \subset \mathbb{R}^n$ to $\mathbb{R}$ that is differentiable at $x \in int(A)$, and $\nabla f(x) \neq 0$. Then the directional derivative $f'_z(x)$ is maximized when $z = \frac{\nabla f(x)}{\|\nabla f(x)\|}$, and the maximized directional derivative is $\|\nabla f(x)\|$.*

The proof is a simple application of the dot product definition of directional derivatives.

## 2.4 Chain Rule

**Proposition 2.9** (Chain Rule)**.** *Let $S$ be a subset of $\mathbb{R}$, and $f : S \to \mathbb{R}$. Let $T$ be a set s.t. $f(S) \subset T \subset \mathbb{R}$, and $g : T \to \mathbb{R}$. If $f$ is differentiable at $x$, and $g$ is differentiable at $f(\mathbb{R})$ is differentiable at $x$, and we have*

$$(g \circ f)'(x) = g'\big(f(x)\big) \cdot f'(x)$$

The chain rule for single variable functions to can be generalized to multivariate functions. See Rudin's Theorem 9.15 for a proof.

**Proposition 2.10** (Chain Rule)**.** *Let $S \in \mathbb{R}^n$, $x \in int(S)$, and $f : S \to \mathbb{R}^m$. Let $T$ be s.t. $f(S) \subset T \subset \mathbb{R}^m$ and $f(x) \in int(T)$, and let $g : T \to \mathbb{R}^k$. If $f$ is differentiable at $x$, and $g$ is differentiable at $f(x)$, then $g \circ f : S \to \mathbb{R}^k$ is differentiable at $x$. Furthermore, we have*

$$(g \circ f)'(x) = g'\big(f(x)\big) \cdot f'(x)$$

In the equation above, the $\cdot$ on the right-hand side is the matrix multiplication. Because $g'\left(f\left(x\right)\right)$ is an $k \times m$ matrix, and $f'\left(x\right)$ is an $m \times n$ matrix, their product $g'\left(f\left(x\right)\right) \cdot f'\left(x\right)$ is a $k \times n$ matrix, which is exactly the size $\left(g \circ f\right)'\left(x\right)$ should have.

By Theorem 2.3, we can rewrite the chain rule in terms of partial derivatives to obtain more intuitions. In the equation $\left(g \circ f\right)'\left(x\right) = g'\left(f\left(x\right)\right) \cdot f'\left(x\right)$, the equality between the $(i,j)$-th entries of the matrices on two sides is

$$\frac{\partial \left(g \circ f\right)_i}{\partial x_j}\left(x\right) = \sum_{l=1}^{m}\left[\frac{\partial g_i}{\partial y_l}\left(f\left(x\right)\right) \cdot \frac{\partial f_l}{\partial x_j}\left(x\right)\right]$$

for each $(i,j) \in \{1,\ldots,k\} \times \{1,\ldots,n\}$. Intuitively, the partial $\frac{\partial (g\circ f)_i}{\partial x_j}$ measures how a change in $x_j$ will lead to a change in $\left[g\left(f\left(x\right)\right)\right]_i$. We know that a change in $x_j$ may lead to a change in $\left[g\left(f\left(x\right)\right)\right]_i$ through $f_1\left(x\right)$, $f_2\left(x\right)$, ..., and $f_m\left(x\right)$, and so the total effect $\frac{\partial (g\circ f)_i}{\partial x_j}$ should be the sum of the $m$ effects, each of which works through one $f_l\left(x\right)$. For each $l \in \{1,\ldots,m\}$, the partial $\frac{\partial f_l}{\partial x_j}\left(x\right)$ measures how sensitive $f_l\left(x\right)$ is w.r.t. $x_j$, and the partial $\frac{\partial g_i}{\partial y_l}\left(f\left(x\right)\right)$ measures how sensitive $\left[g\left(f\left(x\right)\right)\right]_i$ is w.r.t. $f_l\left(x\right)$. Therefore the product $\frac{\partial g_i}{\partial y_l}\left(f\left(x\right)\right) \cdot \frac{\partial f_l}{\partial x_j}\left(x\right)$ measures how a change in $x_j$ will lead to a change in $\left[g\left(f\left(x\right)\right)\right]_i$ through $f_l\left(x\right)$. Summing up all of the $m$ effects, we have $\frac{\partial (g\circ f)_i}{\partial x_j}\left(x\right) = \sum_{l=1}^{m}\left[\frac{\partial g_i}{\partial y_l}\left(f\left(x\right)\right) \cdot \frac{\partial f_l}{\partial x_j}\left(x\right)\right]$.

# 3 Higher Order Derivatives and Taylor Expansion

## 3.1 Second Order Derivatives of $f : A \subset \mathbb{R}^n \to \mathbb{R}$

As a special case of Theorem 2.3 when $m = 1$, for a function $f$ from $A \subset \mathbb{R}^n$ to $\mathbb{R}$, we know that its gradient at $x \in int\left(A\right)$ is equal to the vector of partial derivatives, i.e.

$$\nabla f\left(x\right) = \left(\frac{\partial f}{\partial x_1}\left(x\right), \frac{\partial f}{\partial x_2}\left(x\right), \ldots, \frac{\partial f}{\partial x_n}\left(x\right)\right)$$

The second derivative of the real-valued function $f$ at $x$ is also known as the **Hessian matrix** of $f$ at $x$, denoted as $H_f\left(x\right)$:

$$H_f\left(x\right) := f''\left(x\right) = \left(\nabla f\right)'\left(x\right) = \begin{bmatrix} \left(\nabla \frac{\partial f}{\partial x_1}\right)\left(x\right) \\ \left(\nabla \frac{\partial f}{\partial x_2}\right)\left(x\right) \\ \vdots \\ \left(\nabla \frac{\partial f}{\partial x_n}\right)\left(x\right) \end{bmatrix}$$

$$= \begin{bmatrix} \frac{\partial\left(\frac{\partial f}{\partial x_1}\right)}{\partial x_1}\left(x\right) & \frac{\partial\left(\frac{\partial f}{\partial x_1}\right)}{\partial x_2}\left(x\right) & \cdots & \frac{\partial\left(\frac{\partial f}{\partial x_1}\right)}{\partial x_n}\left(x\right) \\ \frac{\partial\left(\frac{\partial f}{\partial x_2}\right)}{\partial x_1}\left(x\right) & \frac{\partial\left(\frac{\partial f}{\partial x_2}\right)}{\partial x_2}\left(x\right) & \cdots & \frac{\partial\left(\frac{\partial f}{\partial x_2}\right)}{\partial x_n}\left(x\right) \\ \vdots & \vdots & & \vdots \\ \frac{\partial\left(\frac{\partial f}{\partial x_n}\right)}{\partial x_1}\left(x\right) & \frac{\partial\left(\frac{\partial f}{\partial x_n}\right)}{\partial x_2}\left(x\right) & \cdots & \frac{\partial\left(\frac{\partial f}{\partial x_n}\right)}{\partial x_n}\left(x\right) \end{bmatrix}$$

Notice that in the expressions above, the notation $\left(\nabla \frac{\partial f}{\partial x_i}\right)(x)$ stands for the gradient of the function $\frac{\partial f}{\partial x_i}$ at $x$. The notation $\frac{\partial\left(\frac{\partial f}{\partial x_i}\right)}{\partial x_j}(x)$ stands for the partial derivative of the function $\frac{\partial f}{\partial x_i}$ at $x$ w.r.t. the $j$-th argument, which is usually referred to as a **cross partial** at $x$. The notation for the cross partial $\frac{\partial\left(\frac{\partial f}{\partial x_i}\right)}{\partial x_j}$ is usually simplified as $\frac{\partial^2 f}{\partial x_j \partial x_i}$.

The cross partial

$$\frac{\partial^2 f}{\partial x_j \partial x_i}(x) := \frac{\partial\left(\frac{\partial f}{\partial x_i}\right)}{\partial x_j}(x)$$

and the cross partial

$$\frac{\partial^2 f}{\partial x_i \partial x_j}(x) := \frac{\partial\left(\frac{\partial f}{\partial x_j}\right)}{\partial x_i}(x)$$

are conceptually very different when $i \neq j$. However, they are equal if $f$ is twice-differentiable at $x$, and this result is usually known as Young's theorem or Schwarz's theorem.

**Theorem 3.1** (Young; Schwarz)**.** *Let $A \subset \mathbb{R}^n$ and $x \in int(A)$. If function $f : A \to \mathbb{R}$ is $C^2$ at $x$, then for any $i, j \in \{1, \ldots, n\}$ both $\frac{\partial^2 f}{\partial x_j \partial x_i}(x)$ and $\frac{\partial^2 f}{\partial x_i \partial x_j}(x)$ exists and*

$$\frac{\partial^2 f}{\partial x_j \partial x_i}(x) = \frac{\partial^2 f}{\partial x_i \partial x_j}(x)$$

See Rudin's Theorem 9.41 for a proof under a slightly different assumption.

By the theorem above, when $f$ is twice-differentiable at $x$, the Hessian matrix of $f$ at $x$

$$H_f(x) = \begin{bmatrix} \frac{\partial^2 f}{(\partial x_1)^2}(x) & \frac{\partial^2 f}{\partial x_2 \partial x_1}(x) & \cdots & \frac{\partial^2 f}{\partial x_n \partial x_1}(x) \\ \frac{\partial^2 f}{\partial x_1 \partial x_2}(x) & \frac{\partial^2 f}{(\partial x_2)^2}(x) & \cdots & \frac{\partial^2 f}{\partial x_n \partial x_2}(x) \\ \vdots & \vdots & & \vdots \\ \frac{\partial^2 f}{\partial x_1 \partial x_n}(x) & \frac{\partial^2 f}{\partial x_2 \partial x_n}(x) & \cdots & \frac{\partial^2 f}{(\partial x_n)^2}(x) \end{bmatrix}$$

is a symmetric matrix.

When $f$ is not twice-differentiable at $x$, we don't necessarily have $\frac{\partial^2 f}{\partial x_j \partial x_i}(x) = \frac{\partial^2 f}{\partial x_i \partial x_j}(x)$ even when both cross partials exist. See the example below.

**Example 3.2.** *Consider the function $f : \mathbb{R}^2 \to \mathbb{R}$ defined as*

$$f(x, y) := \begin{cases} \frac{xy\left(x^2 - y^2\right)}{x^2 + y^2}, \text{ if } (x, y) \neq (0, 0) \\ 0, \text{ if } (x, y) = (0, 0) \end{cases}$$

*It can be verified that $f$ is not $C^2$ at $(0, 0)$.*

*By definition of partial derivatives, we have*

$$\begin{aligned} \frac{\partial f}{\partial x}(0, 0) &= \frac{d}{dt} f(t, 0)\Big|_{t=0} = \lim_{t \to 0} \frac{f(t, 0) - f(0, 0)}{t - 0} \\ &= \lim_{t \to 0} \frac{0 - 0}{t - 0} = 0 \end{aligned}$$

*and at any* $(0,y)$ *with* $y \neq 0$*, we have*

$$\begin{aligned}\frac{\partial f}{\partial x}(0,y) &= \left.\frac{d}{dt} f(t,y)\right|_{t=0} = \lim_{t\to 0}\frac{f(t,y)-f(0,y)}{t-0} \\ &= \lim_{t\to 0}\frac{\frac{ty\left(t^2-y^2\right)}{t^2+y^2}-0}{t} = \lim_{t\to 0}\frac{y\left(t^2-y^2\right)}{t^2+y^2} = -y\end{aligned}$$

*and so* $\frac{\partial f}{\partial x}(0,y) = -y$ *for any* $y \in \mathbb{R}$*. Therefore, we have*

$$\frac{\partial^2 f}{\partial y \partial x}(0,0) = \left.\frac{d}{dt}\left(\frac{\partial f}{\partial x}(0,t)\right)\right|_{t=0} = \left.\frac{d}{dt}(-t)\right|_{t=0} = -1$$

*Similarly, we have*

$$\begin{aligned}\frac{\partial f}{\partial y}(0,0) &= \left.\frac{d}{dt} f(0,t)\right|_{t=0} = \lim_{t\to 0}\frac{f(0,t)-f(0,0)}{t-0} \\ &= \lim_{t\to 0}\frac{0-0}{t-0} = 0\end{aligned}$$

*and at any* $(x,0)$ *with* $x \neq 0$*, we have*

$$\begin{aligned}\frac{\partial f}{\partial y}(x,0) &= \left.\frac{d}{dt} f(x,t)\right|_{t=0} = \lim_{t\to 0}\frac{f(x,t)-f(x,0)}{t-0} \\ &= \lim_{t\to 0}\frac{\frac{xt\left(x^2-t^2\right)}{x^2+t^2}-0}{t} = \lim_{t\to 0}\frac{x\left(x^2-t^2\right)}{x^2+t^2} = x\end{aligned}$$

*and so* $\frac{\partial f}{\partial y}(0,y) = x$ *for any* $y \in \mathbb{R}$*. Therefore, we have*

$$\frac{\partial^2 f}{\partial x \partial y}(0,0) = \left.\frac{d}{dt}\left(\frac{\partial f}{\partial y}(t,0)\right)\right|_{t=0} = \left.\frac{d}{dt}(t)\right|_{t=0} = 1$$

*So we have*

$$\frac{\partial^2 f}{\partial y \partial x}(0,0) \neq \frac{\partial^2 f}{\partial x \partial y}(0,0)$$

## 3.2 $C^k$ functions

For a function $f$ from $A \subset \mathbb{R}^n$ to $\mathbb{R}^m$, the derivative $f'$ itself is a function from $A_1$ to $\mathbb{R}^{mn}$, it makes sense to talk about the derivative of $f'$.

If $f'$ is differentiable at $x \in int(A_1)$, we call the derivative of $f'$ at $x$, an $mn \times n$ real matrix, the **second derivative of** $f$ **at** $x$, and denote it as $f''(x)$. In this case, we say that $f$ is **twice differentiable at** $x$. Let $A_2 \subset int(A_1)$ be the set of points at which $f'$ is differentiable, then the derivative $f''$ is a function from $A_2$ to $\mathbb{R}^{mn^2}$. If $f''$ is differentiable at $x \in int(A_2)$, we call the derivative of $f''$ at $x$, an $mn^2 \times n$ real matrix, the **third derivative of** $f$ **at** $x$, and denote it as $f'''(x)$. In this case, we say that $f$ is **three times differentiable at** $x$. Inductively, we can define the $k$-th order derivative of $f$ at $x$, an $mn^{k-1} \times n$ real matrix, and denote it as $f^{(k)}(x)$[1].

We say that $f$ from $A \subset \mathbb{R}^n$ to $\mathbb{R}^m$ is **$k$-th continuously differentiable at** $x$ iff $x \in int(A_k)$ and $f^{(k)}(x)$ is continuous at $x$, where $A_k$ is the set of points at which $f^{(k-1)}$ is differentiable. In this case, $f$ is said to be $C^k$ at $x$. We say that $f$ is **$k$-th continuously differentiable** iff $A$ is open and $f$ is $k$-th continuously differentiable at all $x \in A$. In this case, $f$ is said to be $C^k$.

[1]The notation is not to be confused with that of the compound function when we talk about the Contraction Mapping Theorem in the Real Analysis lecture.

### 3.3 Taylor's Theorem

**Theorem 3.3** (Taylor). *Let $f:[a,b]\to\mathbb{R}$ be $C^{n-1}$ and $f^{(n)}(t)$ exists at every $t\in(a,b)$. Let $\alpha$ and $\beta$ be distinct points in $[a,b]$, and define*

$$P_{n-1}(t):=f(\alpha)+f'(\alpha)(t-\alpha)+\frac{f''(\alpha)}{2}(t-\alpha)^2 +\cdots+\frac{f^{(n-1)}(\alpha)}{(n-1)!}(t-\alpha)^{n-1}$$

*Then there exists $x$ strictly between $\alpha$ and $\beta$ s.t.*

$$f(\beta)=P_{n-1}(\beta)+\frac{f^{(n)}(x)}{n!}(\beta-\alpha)^n$$

In the theorem, $\beta$ is allowed to be greater or less than $\alpha$. See Rudin's Theorem 5.15 for a proof. Notice that Taylor's theorem reduces to the mean value theorem when $n=1$, and so Taylor's theorem can be viewed as a generalization of the mean value theorem.

This theorem states that under some differentiability and continuity conditions, $f(\beta)$ can be approximated by the polynomial

$$P_{n-1}(\beta):=f(\alpha)+f'(\alpha)(\beta-\alpha)+\frac{f''(\alpha)}{2}(\beta-\alpha)^2 +\cdots+\frac{f^{(n-1)}(\alpha)}{(n-1)!}(\beta-\alpha)^{n-1}$$

and the error is $\frac{f^{(n)}(x)}{n!}(\beta-\alpha)^n$. If we rewrite $\beta$ as $\alpha+h$, then $f(\alpha+h)$ can be approximated by the polynomial

$$f(\alpha)+f'(\alpha)h+\frac{f''(\alpha)}{2}h^2+\cdots+\frac{f^{(n-1)}(\alpha)}{(n-1)!}h^{n-1}$$

and the error is $\frac{f^{(n)}(x)}{n!}h^n$, where $x$ is some point between $x$ and $x+h$.

If we further assume that $f\in C^n$, then $f^{(n)}$ is continuous at $\alpha$, and thus

$$\frac{\frac{f^{(n)}(x)}{n!}h^n}{h^{n-1}}=\frac{f^{(n)}(x)}{n!}h\to\frac{f^{(n)}(\alpha)}{n!}0=0$$

as $h\to0$, which means that the error is small compared to $h^{n-1}$ as $h$ tends to 0.

Conventionally, the notation $o\left(f(t)\right)$ is used to denote any function $g(t)$ s.t. $\lim_{t\to0}g(t)/f(t)=0$[2]. So the error term is $o\left(h^{n-1}\right)$. Therefore, Taylor's theorem can be rewritten as

$$f(\alpha+h)=f(\alpha)+f'(\alpha)h+\frac{f''(\alpha)}{2}h^2+\cdots+\frac{f^{(n-1)}(\alpha)}{(n-1)!}h^{n-1}+o\left(h^{n-1}\right)$$

when $f$ is $C^n$, and this is sometimes known as the $(n-1)$-th **order Taylor expansion of $f$ at $\alpha$**. Notice that the correct interpretation of the equality above is

$$\lim_{h\to0}\frac{f(\alpha+h)-\left[f(\alpha)+f'(\alpha)h+\frac{f''(\alpha)}{2}h^2+\cdots+\frac{f^{(n-1)}(\alpha)}{(n-1)!}h^{n-1}\right]}{h^{n-1}}=0$$

[2]We sometimes also use the notation $O\left(f(t)\right)$ to denote any function $g(t)$ s.t. $g(t)/f(t)$ converges to some real number (may or may not be 0) as $t\to0$. Therefore, a $o\left(f(t)\right)$ is also a $O\left(f(t)\right)$. The notation $o\left(f(n)\right)$ and $O\left(f(n)\right)$, where $n\in\mathbb{N}$, are defined similarly, but the limit is taken as $n\to\infty$.

We can also write the $(n-1)$th **order Taylor approximation of $f$ at $\alpha$**:

$$f(\alpha+h) \approx f(\alpha) + f'(\alpha)h + \frac{f''(\alpha)}{2}h^2 + \cdots + \frac{f^{(n-1)}(\alpha)}{(n-1)!}h^{n-1}$$

Now let's state the (first and second order) Taylor's theorem for multivariate functions without proof.

**Theorem 3.4** (Taylor)**.** *Let $f$ be a function from $A \subset \mathbb{R}^n$ to $\mathbb{R}$, and $f$ is $C^2$ at $x \in int(A)$. Then we have*

$$f(x+h) = f(x) + \nabla f(x)h + o\left(\|h\|\right)$$

*If $f$ is $C^3$ at $x$, we have*

$$f(x+h) = f(x) + \nabla f(x)h + \frac{1}{2}h^T H_f(x)h + o\left(\|h\|^2\right)$$

Recall that the correct interpretation of the two equations above is

$$\lim_{h\to 0} \frac{f(x+h) - \left[f(x) + \nabla f(x)h\right]}{\|h\|} = 0$$

and

$$\lim_{h\to 0} \frac{f(x+h) - \left[f(x) + \nabla f(x)h + \frac{1}{2}h^T H_f(x)h\right]}{\|h\|^2} = 0$$

# 4 Log-linearization

In dynamic macro economics models, we sometimes use *log-linearization* to approximate a non-linear dynamic system using a linear dynamic system. This invokes Taylor's theorem, which tells us how to construct linear approximations of (non-linear) functions, at least near some point $\mathbf{x}^*$ (which is usually the steady state point of the system).

Consider a multivariate function $f : A \subset \mathbb{R}^n \to \mathbb{R}$, we want to approximate it around point $\mathbf{x}^* = (x_1^*, x_2^*, ..., x_n^*)$ s.t. $x_i^* \neq 0, \forall i$. For each variable $x_i$, we define $\hat{x}_i := \ln(x_i/x_i^*)$ to be its **log-deviation** [3] when $x_i$ and $x_i^*$ have the same sign (which is reasonable when $\mathbf{x}$ is "near" $\mathbf{x}^*$).

Since $x_i = x_i^* e^{\hat{x}_i}$, we can rewrite $f(\mathbf{x}) = f(x_1, x_2, ..., x_n)$ as a function $h$ of $\hat{x}_1, \hat{x}_2, ..., \hat{x}_n$:

$$h(\hat{x}_1, \hat{x}_2, ..., \hat{x}_n) = f(x_1^* e^{\hat{x}_1}, x_2^* e^{\hat{x}_2}, ..., x_n^* e^{\hat{x}_n}) = f(\mathbf{x})$$

Note that $h(\mathbf{0}) = f(\mathbf{x}^*)$ and $h_i'(\mathbf{0}) = f_i'(\mathbf{x}^*)x_i^*, \forall i = 1, 2, ..., n$.

We then take a first order Taylor expansion of $h$ around the point $\mathbf{0}$ (we replace $\approx$ with $=$ in this section, but keep in mind when involving Taylor expansion the equality is not exact and since it is first-order the approximation works well only when $\mathbf{x}$ is close to $\mathbf{x}^*$):

$$\begin{aligned} f(\mathbf{x}) = h(\hat{x}_1, \hat{x}_2, ..., \hat{x}_n) &= h(\mathbf{0}) + h_1'(\mathbf{0})\hat{x}_1 + h_2'(\mathbf{0})\hat{x}_2 + ... + h_n'(\mathbf{0})\hat{x}_n \\ &= f(\mathbf{x}^*) + f_1'(\mathbf{x}^*)x_1^*\hat{x}_1 + f_2'(\mathbf{x}^*)x_2^*\hat{x}_2 + ... + f_n'(\mathbf{x}^*)x_n^*\hat{x}_n \end{aligned}$$

[3] The reason why we use log variables is that we can consider log-deviations as percentage deviations (divided by 100) (this is because $\ln(x_i^* + h) - \ln(x_i^*) = \ln'(x_i^*)h + o(h) \approx \ln'(x_i^*)h = \frac{h}{x_i^*}$), so that the "distances" of variables in probably different absolute terms to some point are comparable to each other when we linearly approximate the function at that point.

The approximation above, in the form of $f(\mathbf{x}) = a_0 + \sum_{i=1}^{n} a_i \hat{x}_i$, is called the **log-linear approximation of function** $f$ around point $\mathbf{x}^*$.

Often, instead of log-linearizing a function, we want to log-linearize an equation (which is (a part of) the characterization of a system at its steady state):

$$f(\mathbf{x}) = f(x_1, x_2..., x_n) = 0$$

around root $\mathbf{x}^* = (x_1^*, x_2^*, ..., x_n^*)$ satisfying $f(\mathbf{x}^*) = 0$. In this case, we first write a log-linear approximation of the LHS, $f(\mathbf{x})$, then we set this log-linear approximation equal to zero. So we have

$$f_1'(\mathbf{x}^*)x_1^*\hat{x}_1 + f_2'(\mathbf{x}^*)x_2^*\hat{x}_2 + ... + f_n'(\mathbf{x}^*)x_n^*\hat{x}_n = 0$$

which, in the form of $\sum_{i=1}^{n} b_i \hat{x}_i = 0$, is called the **log-linearization of equation** $f(\mathbf{x}) = 0$ around $\mathbf{x}^*$ s.t. $f(\mathbf{x}^*) = 0$.

The discussion below is devoted to showing how to perform log-linearization of equations (faster).

If $f(\mathbf{x}^*) \neq 0$, define $\eta_i := \frac{f_i'(\mathbf{x}^*)x_i^*}{f(\mathbf{x}^*)}$ $(i = 1, 2, ..., n)$ the **elasticity of** $f$ **w.r.t** $x_i$ **at** $\mathbf{x}^*$, we can also write:

$$f(\mathbf{x}) = f(\mathbf{x}^*)[1 + \eta_1\hat{x}_1 + \eta_2\hat{x}_2... + \eta_n\hat{x}_n]$$

and therefore

$$\frac{f(\mathbf{x}) - f(\mathbf{x}^*)}{f(\mathbf{x}^*)} = \eta_1\hat{x}_1 + \eta_2\hat{x}_2... + \eta_n\hat{x}_n$$

Now we define the **log-deviation** of function $f$ around some point $\mathbf{x}^* = (x_1^*, x_2^*, ..., x_n^*)$ s.t. $f(\mathbf{x}^*) \neq 0$:

$$\widehat{f(\mathbf{x})} := \ln(f(\mathbf{x})/f(\mathbf{x}^*))$$

(when $f(\mathbf{x})$ and $f(\mathbf{x})^*$ have the same sign). Notice that $\ln(f(\mathbf{x})/f(\mathbf{x}^*)) \approx \frac{f(\mathbf{x})-f(\mathbf{x}^*)}{f(\mathbf{x})^*}$, we then have

$$\widehat{f(\mathbf{x})} = \eta_1\hat{x}_1 + \eta_2\hat{x}_2... + \eta_n\hat{x}_n$$

The following are the log-deviations of some simple functions (please verify by yourselves), which are "shortcuts" you might want to memorize (note here $x$, $x_1$, $x_2$ are scalars):

1. $\widehat{\alpha x}$
2. $\widehat{x_1 + x_2} = \frac{x_1^*}{x_1^*+x_2^*}\hat{x}_1 + \frac{x_2^*}{x_1^*+x_2^*}\hat{x}_1$
3. $\widehat{x_1 x_2} = \hat{x}_1 + \hat{x}_2$
4. $\widehat{x^\alpha} = \alpha\hat{x}$
5. $\widehat{t(x)} = \frac{t'(x^*)x^*}{t(x^*)}\widehat{x}$
6. $\widehat{c} = 0$, where c is a constant

We can get the log-deviation for even more complicated functions by treating them as compound functions and repeatedly applying these shortcuts. For example, consider the function

$$f(A, B, C, D) = \frac{(1 + \alpha C)(A + B)}{D^\alpha}$$

we have

$$\begin{aligned}
f(\widehat{A,B,C},D) &= \widehat{\left(\frac{(1+\alpha C)(A+B)}{D^{\alpha}}\right)} = \widehat{(1+\alpha C)(A+B)D^{-\alpha}} \\
&= \widehat{(1+\alpha C)} + \widehat{(A+B)} + \widehat{D^{-\alpha}} \\
&= \frac{1}{1+\alpha C^*}\hat{1} + \frac{\alpha C^*}{1+\alpha C^*}\widehat{(\alpha C)} + \frac{A^*}{A^*+B^*}\hat{A} + \frac{B^*}{A^*+B^*}\hat{B} - \alpha\hat{D} \\
&= \frac{\alpha C^*}{1+\alpha C^*}\left(\hat{\alpha}+\hat{C}\right) + \frac{A^*}{A^*+B^*}\hat{A} + \frac{B^*}{A^*+B^*}\hat{B} - \alpha\hat{D} \\
&= \frac{\alpha C^*}{1+\alpha C^*}\hat{C} + \frac{A^*}{A^*+B^*}\hat{A} + \frac{B^*}{A^*+B^*}\hat{B} - \alpha\hat{D}
\end{aligned}$$

Sometimes the function $f(\mathbf{x})$ can be written in the form of $f(\mathbf{x}) = g(\mathbf{x}) - l(\mathbf{x})$. Then the equation $f(\mathbf{x}) = 0$ can be written as $g(\mathbf{x}) = l(\mathbf{x})$. To log-linearize equation $g(\mathbf{x}) = h(\mathbf{x})$ around some $\mathbf{x}^*$ satisfying $g(\mathbf{x}^*) = l(\mathbf{x}^*)$, we can just derive the log-deviation $\widehat{g(\mathbf{x})}$ and $\widehat{h(\mathbf{x})}$ around $\mathbf{x}^*$, and set them equal to one another.

# 5 Implicit Function Theorem and Inverse Function Theorem

For a function $f$ from $A \subset \mathbb{R}^n \times \mathbb{R}^m$ to $\mathbb{R}^k$ and a point $(x_0, y_0) \in A$, the **Jacobian matrix** $f'_x(x_0, y_0)$ at $(x_0, y_0)$ is a $k \times n$ matrix defined as the derivative of $f(x, y_0)$ viewed as a function of $x$, evaluated at $x = x_0$. Similarly, the Jacobian matrix $f'_y(x_0, y_0)$ at $(x_0, y_0)$ is a $k \times m$ matrix defined as the derivative of $f(x_0, y)$ viewed as a function of $y$, evaluated at $y = y_0$.

**Theorem 5.1** (Implicit Function)**.** *Let $f$ be a function from $A \subset \mathbb{R}^n \times \mathbb{R}^m$ to $\mathbb{R}^m$. Let $(x_0, y_0) \in int(A)$ s.t. $f(x_0, y_0) = 0$. If $f$ is $C^1$ at $(x_0, y_0)$ and the $m \times m$ Jacobian matrix $f'_y(x_0, y_0)$ is invertible, then there exist an open ball $B_x$ around $x_0$ and an open ball $B_y$ around $y_0$ s.t. $\forall\, x \in B_x$ there exists a unique $y \in B_y$ s.t. $f(x, y) = 0$. Therefore, the equation $f(x, y) = 0$ implicitly defines a function $g : B_x \to B_y$ with the property*

$$f\left(x, g(x)\right) = 0$$

*for any $x \in B_x$. Furthermore, we know that the function $g$ is differentiable at any $x \in B_x$, and*

$$g'(x) = -\left[f'_y\left(x, g(x)\right)\right]^{-1} f'_x\left(x, g(x)\right)$$

Here let's admit that the implicit function $g$ is well-defined and is differentiable, and provide some intuitions only for the derivative formula $g'(x) = -\left[f'_y\left(x, g(x)\right)\right]^{-1} f'_x\left(x, g(x)\right)$ using chain rule. See Rudin's Theorem 9.28 for a complete proof.

Suppose that we can somehow show that there exists an open ball $B_x$ around $x_0$ s.t. $\forall\, x \in B_x$ there exists a unique $y \in B_y$ s.t. $f(x, y) = 0$. Then we define $g : B_x \to B_y$ as $g(x) := y$ s.t. $f(x, y) = 0$. Then we know that $f\left(x, g(x)\right) = 0$ for any $x \in B_x$. Suppose that we can somehow show that $g$ is differentiable at any $x \in B_x$. Then think of both sides of the equation $f\left(x, g(x)\right) = 0$ as a function in $x$ and take derivative, and we should have

$$\frac{d}{dx} f\left(x, g(x)\right) = \frac{d}{dx} 0$$

Clearly, the right-hand side of the equation above is 0. Consider the function $h : B_x \to B_x \times B_y$ defined as $h(x) := \big(x, g(x)\big)$ for any $x \in B_x$. Then we know that $f\big(x, g(x)\big) = f\big(h(x)\big)$, and therefore the left-hand side

$$\begin{aligned}\frac{d}{dx} f\big(x, g(x)\big) &= f'_x(x, g(x)) I_n + f'_y(x, g(x)) g'(x) \\ &= f'_x\big(x, g(x)\big) + f'_y\big(x, g(x)\big) g'(x)\end{aligned}$$

Therefore, we have $f'_x\big(x, g(x)\big) + f'_y\big(x, g(x)\big) g'(x) = 0$, i.e.

$$f'_y\big(x, g(x)\big) g'(x) = -f'_x\big(x, g(x)\big)$$

Because $f'_y(x_0, y_0) = f'_y\big(x_0, g(x_0)\big)$ is invertible, we have $\det\Big(f'_y\big(x_0, g(x_0)\big)\Big) \neq 0$. It can be shown that $\det\Big(f'_y\big(x, g(x)\big)\Big)$ is continuous in $x$, and therefore we can set $B_x$ to be small enough s.t. $\det\Big(f'_y\big(x, g(x)\big)\Big) \neq 0$ for any $x \in B_x$. Therefore, the matrix $f'_y\big(x, g(x)\big)$ is invertible for any $x \in B_x$, and left-multiplying the equation above by $\Big[f'_y\big(x, g(x)\big)\Big]^{-1}$, and we have

$$g'(x) = -\Big[f'_y\big(x, g(x)\big)\Big]^{-1} f'_x\big(x, g(x)\big)$$

The next theorem, often known as the inverse function theorem, is just a special case of the implicit function theorem.

**Theorem 5.2** (Inverse Function)**.** *Let $f$ be a function from $A \subset \mathbb{R}^n$ to $\mathbb{R}^n$. Let $x_0 \in int(A)$ and let $y_0 := f(x_0)$. If $f$ is $C^1$ at $(x_0, y_0)$ and the derivative $f'(x_0)$ is invertible, then there exists an open ball $B_y$ around $y_0$ and an open ball $B_x$ s.t. $\forall\ y \in B_y$ there exists a unique $x \in B_x$ s.t. $f(x) = y$. Therefore, the equation $f(x) = y$ implicitly defines a function $g : B_y \to B_x$ with the property*

$$f\big(g(y)\big) = y$$

*for any $y \in B_y$. Furthermore, the function $g$ is differentiable at any $y \in B_y$, and we have*

$$g'(y) = f'\big(g(y)\big)^{-1}$$

To see why the inverse function theorem is a special case of the implicit function theorem, define

$$F(y, x) := y - f(x)$$

for any $(y, x) \in \mathbb{R}^n \times A$. Clearly, $(y_0, x_0) \in int(\mathbb{R}^n \times A)$ and $F(y_0, x_0) = 0$, and $F$ is $C^1$. Furthermore $F'_x(y_0, x_0) = -f'(x_0)$ is invertible by assumption. Invoke the implicit function theorem for function $F$, and we know that $x$ is implicitly defined as a function $g$ of $y$ on an open ball $B_y$ around $y_0$, with the property $F\big(y, g(y)\big) = 0$ for any $y \in B_y$.

Furthermore, the function $g$ is differentiable at any $y \in B_y$, and

$$g'(y) = -\Big[F'_x\big(y, g(y)\big)\Big]^{-1} F'_y\big(y, g(y)\big) = -\Big[-f'\big(g(y)\big)\Big]^{-1} \cdot I_n = f'\big(g(y)\big)^{-1}$$

So we have the implicit function theorem.

Again, we can obtain some intuitions of this result using chain rule. Think of both sides of the equation $f\left(g\left(y\right)\right)=y$ as a function in $y$ and take derivative:

$$\begin{aligned}\frac{d}{dy}f\left(g\left(y\right)\right)&=\frac{d}{dy}y\\ f'\left(g\left(y\right)\right)\cdot g'\left(y\right)&=I_n\end{aligned}$$

Because $f'\left(g\left(y\right)\right)$ is invertible when $y=y_0$, and so we can set the open ball $B$ to be small enough s.t. $f'\left(g\left(y\right)\right)$ is invertible for any $y\in B_y$. Left multiplying the equation above by $f'\left(g\left(y\right)\right)^{-1}$, and we have $g'\left(y\right)=f'\left(g\left(y\right)\right)^{-1}$.