# Year 1 — Calculus of Several Variables (Honors)

Based on lectures by T. Arant
Discussions by J. Murri
Notes taken by Aryan Dalal

MATH 32AH, Fall Quarter 2024, UCLA

These notes are not endorsed by the lecturers, and I have modified them (often significantly) after lectures. They are nowhere near accurate representations of what was actually lectured, and in particular, all errors are almost surely mine. A note of acknowledgement to Dexter Chua, Ph.D. Harvard University for the template.

---

**Catalog Description**
Lecture, three hours; discussion, one hour. Enforced requisite: course 31A with grade of B or better. Honors course parallel to course 32A. P/NP or letter grading.

**Textbook Reading**
Calculus and Analysis in Euclidean Space, *Jerry Shurman*

**Contact**
This document is a summary of the notes that I have taken during lectures at UCLA; please note that this lecture note will not necessarily coincide with what you might learn. If you find any errors, don't hesitate to reach out to me below:

Aryan Dalal, aryandalal@ucla.edu

# Contents

# 1 Euclidean Spaces

**Definition 1.** Space
A set of elements *plus* a structure to these elements.

The set, $\mathbb{R}$ is the set of *real numbers.* Let $n$ be a positive integer, (i.e. $n$ is an element of the set $\{1, 2, 3, \dots\}$), $n$-dimensional euclidean space is the set $\mathbb{R}^n$.

$$\mathbb{R}^n = \{(x_1, x_2, \dots) : x_i \in \mathbb{R} \text{ for } i = 1, 2, \dots, n\}$$

i.e., the set of *n-tuples* where each coordinate $x_i$ is a real.

**Note:** The notation, ":" reads, 'such that'. Alternatively, | can also be used to express *such that.*

A quick note on **Set Notation**,

1. $a \in A$ (*a is a member of* set $A$)
2. $a \notin A$ (*a is not a member of* set $A$)

**Remark.** $(x_1, x_2, \dots, x_n) = (y_1, y_2, \dots, y_n)$ if and only if (iff) $x_i = y_i$ $\forall$ (for all) $i = 1, 2, \dots, n$.

Notice that the order of coordinates matter in tuples, that is to say, the order in which elements appear in a tuple is significant and can't be ignored. This is **not** the case for sets. Members of a set are not ordered.

$$\{1, 2, 3\} = \{3, 2, 1\}$$

Sets also don't have *multiplicity*, that is, the number of occurrences of particular elements within a set is not significant.

$$\{1, 1, 1, 2, 2, 3\} = \{1, 2, 3\}$$

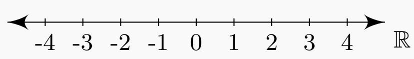
It's possible to conclude that while tuples are ordered and have multiplicity, sets in comparison are **not** ordered and do **not** have multiplicity.

## 1.1 Visualizing Euclidean Spaces

$\mathbb{R}^1$ is technically the set of 1-tuples. $(x_1)$ where $x \in \mathbb{R}$. However, we usually conflate $\mathbb{R}^1$ with $\mathbb{R}$.

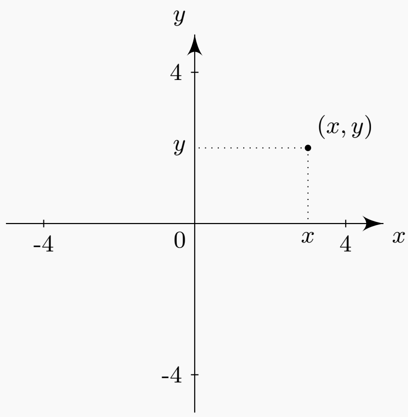
**Visualizing 1-dimensional Euclidean Space**

$\mathbb{R}$ is 1-dimensional euclidean space. Geometrically, it's the real line.



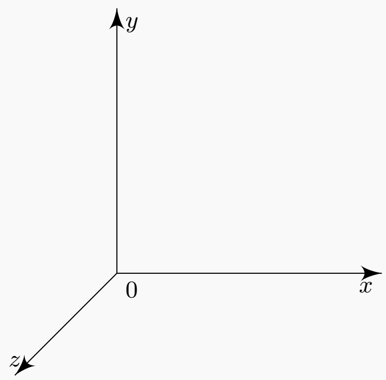
**Visualizing 2-dimensional Euclidean Space**

$\mathbb{R}^2$ is 2-dimensional euclidean space. Geometrically, it's the real plane.




**Visualizing 3-dimensional Euclidean Space**

$\mathbb{R}^3$ is 3-dimensional euclidean space. Geometrically, it's the 3-space.



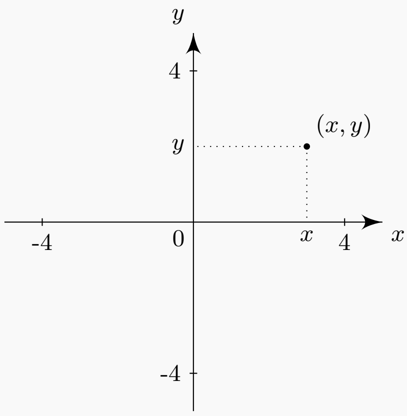

$\mathbb{R}^4$ is also a 4-dimensional euclidean space. However, we can't visually discern $\mathbb{R}^4$ but it is still important. $\mathbb{R}^4$ can be construed as *spacetime*. $(x, y, z, t)$ are the coordinates used to express $\mathbb{R}^4$. In *spacetime*, the coordinates, $(x, y, z)$ express coordinates of position in space, while $t$ expresses the time coordinate. Furthermore, the aforementioned is only a suggested notation, $t$ doesn't necessarily need to be "time" in actuality.

## 1.2 Vectors

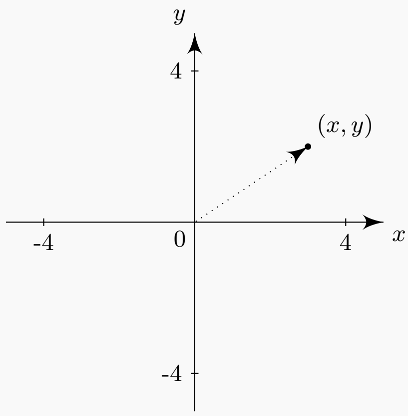
Elements of $\mathbb{R}^n$ are often called (real) vectors. There are two ways to think about vectors,

**Vectors**

**A Point in Space** (E.g. in $\mathbb{R}^2$)




**Arrow with Tail and Head**




**Remark.** Every $\mathbb{R}^n$ has a zero vector or *origin.* $\mathbf{0} = \underbrace{(0, 0, \ldots, 0)}_{n \text{ times}}$

Let's elaborate more on the term, "space". Space doesn't just mean a set, a space, as mentioned before, is a set *and* structure. What is this structure? This structure can be focused on two algebraic operations.

### 1.2.1 Vector Addition

Vector addition, in function notation, is expressed as,

$$\underbrace{+}_{+\text{ is a function}} : \underbrace{\mathbb{R}^n \times \mathbb{R}^n}_{\text{input values, e.g. } x,\ y} \rightarrow \underbrace{\mathbb{R}^n}_{\text{output value, e.g. } x+y}$$

**Remark.** Both of the $\mathbb{R}^n$ in the above function expression are two vectors in $\mathbb{R}^n$.

Vector Addition is defined point-wise (or coordinate-wise).

$$(x_1, \ldots, x_n) \underbrace{+}_{\text{Vector Addition}} (y_1, \ldots, y_n) = \left( x_1 \underbrace{+}_{+\text{ on } \mathbb{R}} y_1, \ldots, x_n + y_n \right)$$

Note that the two '+' symbols in the aforementioned expression are **not** equivalent; they mean different things.

### 1.2.2 Scalar Multplication

The 2nd algebraic operation is *scalar multiplication*, formally written as '$\cdot$' (center dot). In function notation, this is expressed as,

$$\cdot : \underbrace{\mathbb{R}^n}_{\text{scalar value}} \times \underbrace{\mathbb{R}^n}_{\text{vector value}} \rightarrow \mathbb{R}^n$$

If $a$ is a scalar, and $x$ is a vector,

$$a \cdot (x_1, \ldots, x_n) = (ax_1, \ldots, ax_n)$$

**Note:** We don't multiply vectors in $\mathbb{R}^n$ when $n \geq 2$. Expressions such as $xy$ and $\dfrac{y}{x}$ are not possible and should not be written at all. We can only multiply vectors in $\mathbb{R}^1$, not $\mathbb{R}^2$ or above. When we have these two algebraic operations, there are some properties that show how these operations work together.

## 1.3 Vector Space Axioms

The structure, $(\mathbb{R}^n, +, \cdot)$ satisfies the following properties:

(A1) Addition is ***associative***. $\forall\ x, y, z \in \mathbb{R}^n$, $(x + y) + z = x + (y + z)$

(A2) $\mathbf{0}$ (Zero vector) $= \underbrace{(0, \ldots, 0)}_{n \text{ times}}$ is the ***additive identity***. $\forall\ x \in \mathbb{R}^n$, $x + 0 = x$

(A3) Existence of ***additive inverse***. $\forall\ x \in \mathbb{R}^n$, there exists $y \in \mathbb{R}^n$ such that $x + y = 0$ (This $y$ is unique and we write $y = -x$)

(A4) Addition is ***commutative***. $\forall\ x, y \in \mathbb{R}^n$, $x + y = y + x$

(M1) Scalar Multiplication is ***associative***. $\forall\ a, b \in \mathbb{R}$, $x \in \mathbb{R}^n$, $(a \cdot b)\, x = a \cdot (bx)$.

(M2) $\mathbf{1} \in \mathbb{R}$ is the scalar ***multiplicative identity*** $\forall\ x \in \mathbb{R}^n$, $1 \cdot x = x$

(D1) Scalar Multiplication *distributes* over ***scalar addition***. $\forall\ a, b \in \mathbb{R}$, $x \in \mathbb{R}^n$, $(a + b) \cdot x = ax + bx$

(D2) Scalar Multiplication ***distributes*** over *vector addition*. $\forall\ a \in \mathbb{R}$, $x, y \in \mathbb{R}^n$, $a \cdot (x + y) = ax + ay$

**Remark.** If we replace $\mathbb{R}^n$ with an arbitrary set $V \neq \emptyset$ and $(V, +, \cdot)$ satisfy these properties, then $(V, +, \cdot)$ is called a *real vector space*.

**Theorem 1.** *For any $x \in \mathbb{R}^n$, $(-1)\, x = -x$, i.e. $(-1)\, x$ is the additive inverse of $x$.*

*Proof.* We want to check $(-1)\, x + x = \mathbf{0}$

$$\begin{aligned}
(-1)\, x + x &= (-1)\, x + (1)\, x \\
&= (-1 + 1)\, x = 0x \,\text{[Distributing]} \\
&= 0
\end{aligned}$$

□

**Lemma.** For all $x \in \mathbb{R}^n$, $0x = 0$

*Proof.*

$$\begin{aligned}
0x &= (0 + 0)\, x = 0x + 0x \\
0x - 0x &= 0x + 0x - 0x \\
&\implies 0 = 0x
\end{aligned} \tag{1}$$

□

**Remark.** This is just algebra.

Why are (A1) to (A3), (M1), (M2), (D1), (D2) true? That vector addition is commutative easily follows from commutativty of $+$ in $\mathbb{R}$.

**Theorem 2.** *(A4) is true.*

*Proof.* Let $(x_1, \ldots, x_n)$, $(y_1, \ldots, y_n) \in \mathbb{R}^{\ltimes}$

$$\begin{aligned}
\text{By Definition of Vector Addition}\, (x_1, \ldots, x_n) + (y_1, \ldots, y_n) &= (x_1, \ldots, x_n + y_n) \\
\text{By Commutativty} &= (y_1 + x_1, \ldots, y_n + x_n) \\
\text{By Definition of Vector Addition}\, (y_1, \ldots, y_n) + (x_1, \ldots, x_n)
\end{aligned}$$

∎

## 1.4 Basis of $\mathbb{R}^n$

**Definition 2.** A set of $n$ vectors $\{v_1, \ldots, v_n\}$ in $\mathbb{R}^n$ is a *basis* if for every vector $x \in \mathbb{R}^n$, there exists *unique* scalars $a_1, \ldots, a_n \in \mathbb{R}$ such that $x = a_1 v_1 + \cdots + a_n v_n$.

**Remark.** $a_1 v_1 + \cdots + a_n v_n$ is the linear combination of $v_1, \ldots, v_n$.

By uniqueness, we mean that for any $b_1, \ldots, b_n \in \mathbb{R}$,

$$a_1 v_1 + \ldots a_n v_n = b_1 v_1 + \ldots b_n v_n \implies a_i = b_i$$

for $i = 1, \ldots, n$.

### 1.4.1 The Standard Basis of $\mathbb{R}^n$

**Definition 3.** The standard basis of $\mathbb{R}^n$ is the set $\{e_1, \ldots, e_n\}$ where

$$e_i = \left( 0, \ldots, 0, \overbrace{1}^{i\text{-th component}}, 0, \ldots, 0 \right) \in \mathbb{R}^n$$

In $\mathbb{R}^2$, $e_1 = (1, 0), e_2 = (0, 1)$, in $\mathbb{R}^3, e_1 = (1, 0, 0), e_2 = (0, 1, 0), e_3 = (0, 0, 1)$.

**Remark.** Standard basis is not the only basis.

If $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$,

$$\begin{aligned}
x = (x_1, \ldots, x_n) &= (x_1, 0, \ldots, 0) + (0, x_2, \ldots, 0) + \cdots + (0, 0, \ldots, x_n) \\
&= x_1 (1, 0, \ldots, 0) + x_2 (0, 1, \ldots, 0) + \cdots + x_n (0, 0, \ldots, 1) \\
&= x_1 e_1 + x_2 e_2 + \cdots + x_n e_n
\end{aligned}$$

**Example 1.** Show $\{(1, 0), (1, 1)\}$ is a basis of $\mathbb{R}^2$.

*Proof.* Let $x \in \mathbb{R}^2$. We want $a, b$ which satisfy

$$\begin{aligned}
&x = (x_1, x_2) = a(1, 0) + b(1, 1) = (a + b, b) \\
&\iff \begin{cases} x_1 = a + b \\ x_2 = b \end{cases} \iff \begin{cases} x_1 - b = a \\ x_2 = b \end{cases} \iff \begin{cases} a = x_1 - x_2 \\ b = x_2 \end{cases}
\end{aligned}$$

Thus, $x = (x_1 - x_2)(1, 0) + x_2 (1, 1)$

For every $x$, we've shown that there is some linear combination determined solely by the vectors of $x$. ∎

**Remark.** Uniqueness also follows from our above work since we showed

$$x = a\,(1,0) + b\,(1,1) \iff a = x_1 - x_2, b = x_2$$

# 2 Geometry: Length & Angle

## 2.1 Length and The Inner Product

**Definition 4.** *The Inner product.*
The inner product (or the dot product) of vectors $x, y \in \mathbb{R}^n$ is $\langle x, y\rangle$ such that

$$\begin{aligned}\langle x, y\rangle &= \langle (x_1, \ldots, x_n), (y_1, \ldots, y_n)\rangle = \sum_{i=1}^{n} x_i y_i \\ &= x_1y_1 + x_2y_2 + \cdots + x_ny_n\end{aligned}$$

**Example 2.**

$$\begin{aligned}\langle (1,3,-1), (2,0,4)\rangle &= 1\cdot 2 + 3\cdot 0 + (-1)\cdot 4 \\ &= 2 + 0 - 4 = -2\end{aligned}$$

**Remark.** The context of the result from the inner product will be discussed later.

For $x \in \mathbb{R}^n$,

$$\langle x, e_i\rangle = \langle (x_1, \ldots, x_n), \left(0, \ldots, 0, \overbrace{1}^{i\text{-th component}}, 0, \ldots, 0\right)\rangle = x_i$$

**Definition 5.** Vectors $x, y \in \mathbb{R}^n$ are called *orthogonal* if $\langle x, y\rangle = 0$

Note that $\mathbf{0}$ is orthogonal to every vector,

**Example 3.** $(1,1,1,1)$ and $(-1,1,-1,1)$ are orthogonal.

**Theorem 3.** *(Inner Product Properties)*

*(IP1) Positive Definite:* $\forall\ x \in \mathbb{R}^n$, $\langle x, x\rangle \geq 0$
*and only one case* $\langle x, x\rangle = 0 \iff x = 0$

*(IP2) Symmetry:* $\forall\ x, y \in \mathbb{R}^n, \langle x, y\rangle = \langle y, x\rangle$

*(IP3) Bilinearity:* $\forall\ x, x', y, y' \in \mathbb{R}^n, a, b \in \mathbb{R}$

$$\begin{cases}\textit{Linear in 1st component} & \langle x + x', y\rangle = \langle x, y\rangle + \langle x', y\rangle, \langle ax, y\rangle = a\langle x, y\rangle \\ \textit{Linear in 2nd component} & \langle x, y + y'\rangle = \langle x, y\rangle + \langle x, y'\rangle, \langle x, ay\rangle = a\langle x, y\rangle\end{cases}$$

Let's prove (IP1),

*Proof.* $\langle x, x\rangle = x_1^2 + \cdots + x_n^2 \geq 0$. Since $x^2$ is always positve, $\langle x, x\rangle \geq 0$ □

**Remark.** Only case where $\langle x, x\rangle = 0$ is if each $x_i$ for $i = 1, \ldots, n$ is also 0.

**Definition 6.** The modulus (or absolute value or length) of $x \in \mathbb{R}^n$ is

$$|x| = \sqrt{\langle x, x\rangle}$$

In $\mathbb{R}$, $|a| = \sqrt{a^2}$. We are only taking the principle branch, the positive square root. The modulus of $x$ is always real by (IP1).

**Theorem 4.** *(Modulus Properties)*
*(Mod 1)* $\forall\ x \in \mathbb{R}^n,\ |x| \geq 0, |x| = 0 \iff x = 0$
*(Mod 2)* $\forall\ x \in \mathbb{R}^n, a \in \mathbb{R},\ |ax| = |a|\,|x|$

**Theorem 5.** *Cauchy-Bunyakovsky-Schwarz Inequality*

$$\forall x, y \in \mathbb{R}^n, |\langle x, y\rangle| \leq |x| \cdot |y|$$

*Proof.* When $x = 0$, the result is trivial since both sides $= 0$. So we may assume $x \neq 0$. Then, for any $a \in \mathbb{R}$,

$$\begin{aligned}
0 \leq \langle ax - y, ax - y\rangle &= \langle ax, ax - y\rangle + \langle -y, ax - y\rangle \\
&= \langle ax, ax\rangle + \langle ax, -y\rangle + \langle -y, ax\rangle + \langle -y, -y\rangle \\
&= a\langle x, ax\rangle + a\langle x, -y\rangle + \langle -y, ax\rangle + \langle -y, -y\rangle \\
&= a^2\langle x, x\rangle + a\langle x, -y\rangle + a\langle -y, x\rangle + \langle -y, -y\rangle \\
&= a^2\langle x, x\rangle + a\langle x, -y\rangle - a\langle y, x\rangle - \langle y, -y\rangle \\
&= a^2\langle x, x\rangle - a\langle x, y\rangle - a\langle y, x\rangle + \langle y, y\rangle \\
&= a^2\,|x|^2 - 2a\langle x, y\rangle + |y|^2
\end{aligned}$$

It follows that some quadratic polynomial in $a$, $f(a) \geq 0$. Therefore, this polynomial can at most have one real root (or no real roots). This implies that its discriminant is non-positive.

$$\begin{aligned}
D \leq 0 &\implies 4\langle x, y\rangle^2 - 4\,|x|^2\,|y|^2 \leq 0 \\
\langle x, y\rangle^2 - |x|^2\,|y|^2 \leq 0 &\implies \langle x, y\rangle^2 \leq |x|^2\,|y|^2 \\
&\implies |\langle x, y\rangle| \leq |x| \cdot |y|
\end{aligned}$$

□

**Remark.** The Cauchy-Schwarz Inequality is equivalent to,

$$\underbrace{-\,|x| \cdot |y|}_{\text{Lower bound}} \leq \langle x, y\rangle \leq \overbrace{|x| \cdot |y|}^{\text{Upper bound}}$$

The $\leq$ is actually $=$ if and only if one of $x, y$ is a scalar multiple of the other, i.e., $x, y$ are parallel.

**Theorem 6.** *Triangle Inequality*

$$\forall x, y, \in \mathbb{R}^n, |x + y| \leq |x| + |y|$$

*Proof.*

$$\begin{aligned}
|x + y|^2 &= \langle x + y, x + y\rangle \\
\text{(By Linearity in 1st Variable)} &= \langle x, x + y\rangle + \langle y, x + y\rangle > \\
\text{(By Linearity in 2nd Variable)} &= \langle x, x\rangle + \langle x, y\rangle + \langle y, x\rangle + \langle y, y\rangle \\
\text{(By Symmetry)} &= |x|^2 + 2\langle x, y\rangle + |(|\,y)^2 \\
\text{(By Cauchy-Schwarz)} &\leq |x|^2 + 2\,|x| \cdot |y| + |y|^2 = (|x| + |y|)^2 \\
|x + y|^2 &\leq (|x| + |y|)^2
\end{aligned}$$

$$\therefore |x + y| \leq |x| + |y|$$

□

**Remark.** For any $x_1, x_2, \ldots, x_k \in \mathbb{R}^n$ for $1 \leq k \leq n$,

$$|x_1 + (x_2 + \cdots + x_k)|$$

$$\text{Induction} \begin{cases} \leq |x_1| + |x_2 + (\cdots + x_k)| \\ \leq |x_1| + |x_2| + |x_3 + (\cdots + x_k)| \\ \leq \cdots \leq |x_1| + |x_2| + \cdots + |x_k| \end{cases}$$

$\therefore$

$$|x_1 + \ldots + x_k| \leq |x_1| + \ldots + |x_k|$$

**Remark.** There are important variants of the triangle inequality,

$$||x| + |y|| \leq |x - y| \leq |x| + |y|$$

## 2.2 Angles

By Cauchy-Schwarz,

$$-|x| \cdot |y| \leq \langle x, y \rangle \leq |x| \cdot |y|$$

when $x \neq \mathbf{0}$ & $y \neq \mathbf{0}$ (so $|x| \neq 0$, $|y| \neq 0$)

$$\frac{-|x| \cdot |y|}{|x| \cdot |y|} \leq \frac{\langle x, y \rangle}{|x| \cdot |y|} \leq \frac{|x| \cdot |y|}{|x| \cdot |y|} \iff -1 \leq \frac{\langle x, y \rangle}{|x| \cdot |y|} \leq 1$$

There's a unique angle $0 \leq \theta_{x,y} \leq \pi$ such that $\cos \theta_{x,y} = \dfrac{\langle x, y \rangle}{|x| \cdot |y|}$.

By definition, $\theta_{x,y} =$ angle between $x, y$ or, $\theta_{x,y} = \arccos \left( \dfrac{\langle x, y \rangle}{|x| \cdot |y|} \right)$

In $\mathbb{R}^2$, $\mathbb{R}^3$, $\theta_{x,y}$ is the "correct angle", [diagram of acute angle between vectors] **Note:** By rearranging, we get a geometric interpretation of

$$\langle x, y \rangle = (\cos \theta_{x,y}) \, |x| \cdot |y| \, , x \neq \mathbf{0}, y \neq \mathbf{0}$$

# 3 Analysis: Convergence of Sequences

We start by reviewing convergence in $\mathbb{R}$:

Let $\{a_\nu\}$ be a sequence in $\mathbb{R}$. The sequence is really just an infinite list of numbers $a_1, a_2, a_3, \ldots$. The greek letter $\nu$ is used so as to avoide using $n, m$, etc. which is usually used as dimensions in Euclidean space—avoid overloading notation.

A more precise notation for a sequence in $\mathbb{R}$ would be:

$$\{a_\nu\}_{v=1}^{\infty}$$

Recall what it means for $\{a_\nu\}$ to converge to some $b \in \mathbb{R}$,

**Notation:** *Converge to* can be written as

$$\lim a_\nu = b \quad \text{or} \quad \lim_{\nu \to \infty} = b \text{ (more precise)} \quad \text{or} \quad a_\nu \to b \text{ (as } \nu \to \infty)$$

**Definition 7.** $a_\nu \to b$ means $\forall \epsilon > 0$, there is $\nu_0 > 0$ such that $\nu > \nu_0 \implies |a_\nu - b| < \epsilon$.

[diagram about epsilon-delta error tolerance]

Here are some important facts about convergence in $\mathbb{R}$,

1) Limit laws: Suppose $\{a_\nu\}, \{b_\nu\}$ are sequences in $\mathbb{R}$ such that $a_\nu \to a$ and $b_\nu \to b$. Then, $a_\nu + b_\nu \to a + b$ and $a_\nu b_\nu \to ab$.
2) Squeeze Theorem (for sequences): $\{a_\nu\}, \{b_\nu\}, \{c_\nu\}$ are sequences in $\mathbb{R}$ and $a_\nu \to l, c_\nu \to l$ and $a_\nu \le b_\nu \le c_\nu$ $\forall \nu$, then $b_\nu \to l$.

**Special Case.** If $|b_\nu| \le c_\nu$ for all $\nu$ and $c_\nu \to 0$, then $b_\nu \to 0$. Also, it goes without saying, $|b_v| \to 0$. This follows because $|b_\nu| \le c_\nu \iff -c_\nu \le b_\nu \le c_\nu$

**Remark.** Sequences are **not** sets. Their elements are ordered.

$$\{a_\nu\}_{\nu=1}^{\infty} = (0, 0, 0, \ldots), \{a_\nu : \nu = 1, 2, 3, \ldots\} = \{0\}$$

The goal for this section is to define convergence of sequences in $\mathbb{R}^n, n \ge 2$

Let $\{x_\nu\}$ be a sequence of vectors in $\mathbb{R}^n$ where 'sequence of vectors' really just means an infinite list of vectors $x_1, x_2, \ldots \in \mathbb{R}^n$

**Notation:** $x_\nu \in \mathbb{R}^n$, to list the coordinates, we write

$$x_\nu = (x_{1,\nu}), x_{2,\nu}, \ldots, x_{n,\nu}$$

We can also use the notation $x_\nu = x^{(\nu)} = \left(x_1^{(\nu)}, x_2^{(\nu)}, \ldots, x_n^{(\nu)}\right)$

**Example 4.** $\left\{\left(\frac{1}{\nu}, 1 + \nu\right)\right\}_{\nu=1}^{\infty}$. As a list:

$$(1, 2), \left(\frac{1}{2}, 3\right), \left(\frac{1}{3}, 4\right), \ldots$$

**Definition 8.** Let $\{x_\nu\}$ be a sequence in $\mathbb{R}^n$, let $p \in \mathbb{R}^n$.

$\{x_\nu\}$ $\underbrace{\text{converges}}_{x_\nu \to \infty}$ to $p$ if the sequence $\{|x_\nu - p|\}$ $\overbrace{\text{converges to}}^{\text{in } \mathbb{R}}$ $0 \in \mathbb{R}$.

In terms of the epsilon definition for a limit, $\{x_\nu\}$ converges to $p$ if for every $\epsilon > 0$, there is $\nu_0$ such that $\nu > \nu_0 \implies |x_\nu - p| < \epsilon$.

What does this mean? It means for every $x, y \in \mathbb{R}^n, |x - y|$ = distance between $x, y$. When $|x_\nu - p|$ is very small, it means $x_\nu$ is very close to $p$.

**Theorem 7.** *Let $\{x_\nu\}$ be a sequence in $\mathbb{R}^n$, let $p \in \mathbb{R}^n$*

$$x_\nu \to p \textit{ if and only if for each } j = 1, 2, \ldots, n, \lim_{\nu \to \infty} x_{j,\nu} \to p_j$$

Each coordinate of $x_\nu$ can be written as

$$x_\nu = \left( \underbrace{x_{1,\nu}}_{\{x_{i,\nu}\}_{\nu=1}^{\infty}}, x_{2,\nu}, \ldots, \underbrace{x_{n,\nu}}_{\{x_{n,\nu}\}_{\nu=1}^{\infty}} \right)$$

**Example 5.** $\left\{\left(\frac{1}{2^v}, e^{\frac{1}{v}}\right)\right\}_{\nu=1}^{\infty}$ (sequence in $\mathbb{R}^2$)

First coordinate sequence converges: $\dfrac{1}{2^v} \to 0$
Second coordinate sequence converges: $e^{\frac{1}{v}} \to e^0 = 1$

Them, by the thm., $\left(\dfrac{1}{2^v}, e^{\frac{1}{v}}\right) \to (0,1)$

**Example 6.** $\left\{\left(\dfrac{1}{v}, 1+v\right)\right\}$ does not converge since the second coordinate sequence $\{1+v\}$ diverges to $+\infty$.

*Proof.* ($\Rightarrow$) Suppose $x_\nu \to p$. Fix some $j = 1, 2, \ldots, n$.
...
($\Leftarrow$) Suppose $x_{j,\nu} \to p_j$ for $j = 1, 2, \ldots, n$ (i.e. $|x_{j,v} - p_j| \to 0$)

We want to show $x_\nu \to p$ (in $\mathbb{R}^n$)
...

$$|x_\nu - p| \leq \sum_{j=1}^{\infty} |x_{j,\nu} - p_j| \to 0$$

$\sum_{j=1}^{\infty} |x_{j,\nu} - p_j|$ is a fine sum of sequence which all $\to 0$. By limit law in $\mathbb{R}$, the right hand side $\to 0$.

By squeeze, $|x_\nu - p| \to 0$. ☐

## 3.1 Componentwise Nature of Convergence

$\big(\text{In } \mathbb{R}^3 \text{ case}\big)$ $\{x_\nu\}$ converges to $p \iff \begin{cases} x_{1,\nu} \to p_1 \\ x_{2,\nu} \to p_2 \\ x_{3,\nu} \to p_3 \end{cases}$

**Theorem 8.** *Linearity of Convergence.*

*(1) If $\{x_\nu\}$, $\{y_\nu\}$ are sequences in $\mathbb{R}^n$ and $x_\nu \to p, y_\nu \to p$, then $x_\nu + y_\nu \to p + q$.*

*(2) If $\{x_\nu\}$ is a sequence in $\mathbb{R}^n$, $x_\nu \to p, c \in \mathbb{R}$, then $cx_\nu \to cp$.*

*Proof.* (1) Suppose $x_\nu \to p, y_\nu \to q$,

By componentwise nature of convergence, we have,

$$x_{j,\nu \to p_j} \text{ and } y_{j,\nu} \to q_j \text{ for } j \in \{1, \ldots, n\}$$

$(x_\nu + y_\nu)_j = x_{j,\nu} + y_{j,\nu} \to p_j + q_j = (p+q)_j$ $\forall j \in \{1, \ldots, n\}$ by limit law in $\mathbb{R}^n$.

By componentwise nature again,

$$x_\nu + y_\nu \to p + q$$

☐

## 3.2 Mappings Between Euclidean Spaces

Note that we don't always require a function or map to be defined on all of a euclidean space.

E.g. $f(x) = \dfrac{1}{x}$ $(x \in \mathbb{R})$

$$\operatorname{dom}(f) = \text{ domain of } f = \{x \in \mathbb{R} : x \neq 0\}$$

$A \subseteq \mathbb{R}^n$ means $A$ is a subset of $\mathbb{R}^n$. $A$ could be equal to $\mathbb{R}^n$, it could possibly be not $\mathbb{R}^n$. If we want $A$ to not be $\mathbb{R}^n$, then we use $A \subsetneq \mathbb{R}^n$ which means $A \subseteq \mathbb{R}^n, A \neq \mathbb{R}^n, A$ is a proper subset.

$A \subseteq \mathbb{R}^n, f : A \to \mathbb{R}^n$ means $f$ is a function with $\operatorname{dom}(f) = A$ and $f(x) \in \mathbb{R}^m$ $\forall x \in A$.

$f(x) \in \mathbb{R}^m$ so it has $m$ coordinates. We write,

$$f(x) = (f_1(x), f_2(x), \ldots, f_m(x)).$$

Each $f_j : A \to \mathbb{R}$ is a real-valued function with domain $A$ called $j$-th coordinate function of $f$.

**Example 7.** $f : \mathbb{R} \to \mathbb{R}^2, f(x) = \left(x^2 + 1, 3x\right)$.

Then, $f_1(x) = x^2 + 1, f_2(x) = 3x$.

## 3.3 Function Limits

**Definition 9.** $A \subseteq \mathbb{R}^n, f : A \to \mathbb{R}^m, p \in \mathbb{R}^n, L \in \mathbb{R}^m$

$$\lim_{x \to p} f(x) = L$$

when

(1) There is at least one sequence $\overbrace{\{x_\nu\} \text{ in A}}^{x_\nu \in A \text{ for all } \nu}$ such that $x_\nu \to p$

(2) For all sequences $\{x_\nu\}$ in $A$, $x_\nu \to p$ implies $f(x_\nu) \to L$. e

## 3.4 Point-Set Topology: Compact Sets and Continuity

We will now lay the topological foundations required to defining the derivative. Point-set topology is the study of continuity on spaces. We will only be using a level of topology that is required for this course.

**Definition 10. (ball)** A ball in $\mathbb{R}^n$ is a set of the form

$$\mathcal{B}(p, r) = \{x \in \mathbb{R}^n : |x - p| < r\}$$

for $p \in \mathbb{R}^n, r > 0$ where $p$ is the center of the ball and $r$ is the radius.

We think of balls in euclidean spaces as a region that surrounds a defined point. When we define balls in $\mathbb{R}^n$ for any $n$-dimensions, interesting results follow. In $\mathbb{R}^2$, the 'ball' is really just a circular region centered a point $p$. In $\mathbb{R}^3$, it's a sphere. What would it be in $\mathbb{R}^1$? It's an *open* interval. I have italicized 'open' since this will be defined rigorously later.

It's important to note that the boundary of a ball as defined above is not included in the set since $|x - p| < r$. This follows for all $n$.

### 3.4.1 Bounded Set

**Definition 11. (Bounded Sets)** A set $A \subseteq \mathbb{R}^n$ is *bounded* if there is a ball $\mathcal{B}(p, r)$ such that $A \subseteq \mathcal{B}(p, r)$.

A bounded set can be thought of as a region $A \subseteq \mathbb{R}^n$ that's completely enclosed by a ball $\mathcal{B}(p, r)$ such that all points in $A$ are also included in $\mathcal{B}(p, r)$ but all points in $\mathcal{B}(p, r)$ may or may not be included in $A$.

There are some interesting examples for bounded sets. Consider the following,

(1) All balls are bounded.

(2) For any $p \in \mathbb{R}^n$, any $r > 0$,

$$\mathcal{B}(p, r) = \{x \in \mathbb{R}^n : |x - p| \leq r\}$$

is bounded. Proof by $\overline{\mathcal{B}}(p, r) \subseteq \mathcal{B}(p, r + 1)$.

(3) $\mathbb{R}^n$ is not bounded.

(4) $(0, +\infty)$ is not a bounded subset of $\mathbb{R}$.

(5) A sequence $\{x_\nu\}$ in $\mathbb{R}^n$ is bounded when its set of terms $\{x_1, x_2, \ldots, x_n\}$ is a bounded set.

**Theorem 9.** *If $\{x_\nu\}$ is convergent, then it's bounded.*

*Proof.* Suppose $x_\nu \longrightarrow p$. Recall, this means for every $\epsilon > 0$, $\exists \nu_o$ s.t. $\nu > \nu_0 \implies |x_\nu - p| < \epsilon$. Apply the definition of convergence with $\epsilon = 1$ to get a $\nu_o$ s.t. $\nu > \nu_0 \implies |x_\nu - p| < 1$.

Let $M = \max\left(\{|x_\nu - p| : \nu = 1, 2, \ldots, \nu_0\} \cup \{1\}\right)$

Then, simple to check that

$$\{x_1, x_2, \ldots, x_{\nu_0}, x_{\nu_0+1,\ldots}\} \subseteq \mathcal{B}(p, m + 1)$$

$\square$

### 3.4.2 Closed Set

**Definition 12. (Closet Sets)** A set $F \subseteq \mathbb{R}^n$ is *closed* if for every sequence $\{x_\nu\}$ in $F$ which happens to converge in $\mathbb{R}^n$, in fact converges in $F$, that is, $\lim_{\nu \to \infty} x_\nu \in F$.

It can also be said that $F$ is a closed set if $F$ containts all of its limit points. Let's consider some examples of closed sets.

(1) Closed intervals in $\mathbb{R}$ are closed such as $[1, 2]$. $[1, +\infty)$ is also a closed set since its endpoints are real numbers.

(2) $(0, 1]$ is not closed. *Proof.* Consider the sequence $\left\{\frac{1}{\nu}\right\}$ in $A$ which converges to 0 but $\lim \frac{1}{\nu} \neq 0 \in A$.

(3) Let $n = 1, 2, \ldots$. $S^{n-1} \subseteq \mathbb{R}^n$ , called the $(n-1)$th dimensional sphere is defined by

$$S^{n-1} = \{x \in \mathbb{R}^n : |x| = 1\}$$

(4) $\mathbb{R}^n$ is a closed set.

Example (3) is important. Intuitively, $S^0$ is two points $-1$ and 1; $S^1$ is a unit circle and $S^2$ is a unit sphere.

**Claim.** For each $n = 1, 2, \ldots$, $S^{n-1}$ is closed.

*Proof.* Let $\{x_\nu\}$ be a sequence in $S^{n-1}$, and suppose $x_\nu \longrightarrow p$. We want to show $p \in S^{n-1}$, i.e., $|p| = 1$.

Since $|\cdot|$ is continuous, $|x_\nu| \longrightarrow |p|$. But $\{|x_\nu|\}$ is a constant sequence with value 1, hence $1 \longrightarrow |p| \implies |p| = 1$. $\square$

A similar proof show that for any $p \in \mathbb{R}^n$ and $r > 0$,

$$\overline{\mathcal{B}}\left(p, r\right) = \{x \in \mathbb{R}^n : |x - p| \leq r\}$$

is closed (we call it a closed ball).

### 3.4.3 Compact Set

**Definition 13. (Compact Sets)** A set $K \subseteq \mathbb{R}^n$ is *compact* if it is both *closed* and *bounded.*

Following this definition, we will understand the Extreme Value Theorem in the context of several variables. Consider the following examples,

(1) $S^{n-1}$, $n = 1, 2, \ldots$ are all compact.

(2) Bounded, closed intervals $[a, b]$ for $(a, b) \in \mathbb{R}$ are compact.

(3) $R^n$ is closed but not bounded, so it's not compact.

**Theorem 10.** ***(Extreme Value Theorem)*** *Let $K \subseteq A \subseteq \mathbb{R}^n, K \neq \varnothing$ and $K$ is compact. $f\colon A \longrightarrow \mathbb{R}$ (real valued) continuous on $K$, then, $\exists x_{max}, x_{min}$ s.t.*

$$f\left(x_{min}\right) \leq f\left(x\right) \leq f\left(x_{max}\right)$$

*for any $x \in K$.*

**Remark.** The single-variable version of the extreme value theorem suggests $K = [a, b]$ which is closed, bounded. Using single-variable functions, we can see that $K$ needs to be both closed and bounded (i.e., compact).

Let's explore an application of the Extreme Value Theorem. Let $K \subseteq \mathbb{R}^n$ be non-empty and compact. Then, there are vectors in $K$ which are longest and shortest with respect to vectors in $K$.

*Proof.* $|\cdot|$ is a continuous function. By extreme value theorem (EVT) applied to $|\cdot|$ on $K$, there are $x_{\max}, x_{\min} \in K$ such that $|x_{\min}| \leq |x| \leq |x_{\max}|$ for all $x \in K$. $\square$

# 4 Linear Algebra

We will now explore the idea behind linear transformations and move towards defining the derivative in several variables.

## 4.1 Matrices

**Definition 14. (Matrix)** A matrix $A$ is a rectangular array of numbers:

$$A = \left[a_{ij}\right]_{m \times n} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}$$

where $m$ is the number of rows and $n$ is the number of columns.

Consider the example,

$$A = \begin{bmatrix} 1 & 0 & -1 \\ 2 & 3 & 6 \end{bmatrix}$$

is a $2 \times 3$ matrix.

Note that it's somewhat tedious to manually write matrix entries both by hand and on LaTeX and so it's preferable to refer to some matrix $A$ as just $[a_{ij}]_{m\times n}$.

**Remark.** When all the entries of $A$ are real, then $A$ is called a real matrix.

**Example 8.** The $m \times n$ matrix $E_{i_0 j_0}$ $(1 \le i_0 \le m, 1 \le j_0 \le n)$ is

$$E_{ij} = \begin{bmatrix} \delta_{ii_0} & \delta_{jj_0} \end{bmatrix}_{m\times n}$$

where $\delta_{kl} = \begin{cases} 1 & \text{if } k = l \\ 0 & \text{otherwise} \end{cases}$

The $3 \times 3$ matrix,

$$E_{21} = \begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

### 4.1.1 Operations with Matrices

Matrix addition is entrywise.

$$A + B = \begin{bmatrix} a_{ij} \end{bmatrix}_{m\times n} + \begin{bmatrix} b_{ij} \end{bmatrix}_{m\times n} = \begin{bmatrix} a_{ij} + b_{ij} \end{bmatrix}_{m\times n}$$

Scalar multiplication of matrices is entrywise.

$$cA = c\begin{bmatrix} a_{ij} \end{bmatrix}_{m\times n} = \begin{bmatrix} ca_{ij} \end{bmatrix}_{m\times n}$$

**Example 9.**

$$2\begin{bmatrix} 1 & -1 \\ 0 & 3 \end{bmatrix} + \begin{bmatrix} 0 & -1 \\ 2 & 4 \end{bmatrix} = \begin{bmatrix} 2 & -2 \\ 0 & 6 \end{bmatrix} + \begin{bmatrix} 0 & -1 \\ 2 & 4 \end{bmatrix} = \begin{bmatrix} 2 & -3 \\ 2 & 10 \end{bmatrix}$$

**Remark.** Matrix Addition and Scalar Multiplication of Matrices have all the same nice algebraic properties as vector addition and scalar multiplication. E.g., $A + B = B + A, (A + B) + C = A + (B + C), c(dA) = (cd)A$, etc.

### 4.1.2 Matrix by Vector Multiplication

If $A$ is a $m \times n$ matrix and $x$ is an $n$-dimensional (column) vector,

$$x = (x_1, \ldots, x_n) \implies \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}$$

We can think of Matrix by Vector multiplication in multiple ways,

(1) By column-wise approach,

$$A_{m\times n} = \begin{bmatrix} v_1 & \ldots & v_n \end{bmatrix}$$

where $v_1, \dots, v_n$ are the $m$-dimensional columns of $A$.

$$Ax = \begin{bmatrix} v_1 & \dots & v_n \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} = x_1 v_1 + x_2 v_2 + \cdots + x_n v_n$$

This approach of Matrix by Vector multiplication is nice theoretically.

(2) By row-wise approach,

$$A_{m \times n} = \begin{bmatrix} w_1 \\ \vdots \\ w_m \end{bmatrix}$$

where $w_1, \dots, w_m$ are the $n$-dimensional rows of $A$.

$$Ax = \begin{bmatrix} w_1 \\ \vdots \\ w_m \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} \langle w_1, x \rangle \\ \vdots \\ \langle w_n, x \rangle \end{bmatrix}$$

This approach of Matrix by Vector multiplication is nice computationally.

**Example 10.**

$$\begin{bmatrix} 1 & -1 \\ 2 & 3 \end{bmatrix} \begin{bmatrix} 2 \\ 4 \end{bmatrix} = \begin{bmatrix} 1 \cdot 2 + (-1) \cdot 4 \\ 2 \cdot 2 + 3 \cdot 4 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \end{bmatrix} 2 + \begin{bmatrix} -1 \\ 3 \end{bmatrix} 4$$

We can generalize Matrix by Vector multiplication by a nice formula,

$$Ax = \left[a_{ij}\right]_{m \times n} \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} \sum_{j=1}^{n} a_{1j} x_j \\ \vdots \\ \sum_{j=1}^{n} a_{mj} x_j \end{bmatrix}$$

**Remark.** $Ax$ is a $m$-dimensional column vector. If $A$ is $m \times n$ and $x$ is $n \times 1$, then we can think of cancelling the common $n$ and the resultant vector is $m \times 1 = m$.

## 4.2 Linear Mappings: Linear Transformations

**Proposition.** Matrix by Vector multiplication is linear, i.e., if $A$ is $m \times n$ matrix, then,

(1) $\forall x, y \in \mathbb{R}^n, A(x + y) = Ax + Ay$

(2) $\forall c \in \mathbb{R}, x \in \mathbb{R}^n, A(cx) = c(Ax)$

**Definition 15. (Linear Transformation)** A *linear transformation* is a map $T \colon \mathbb{R}^n \longrightarrow \mathbb{R}^m$, such that,

(1) For every $x, y \in \mathbb{R}^n, T(x + y) = T(x) + T(y)$

(2) For every $c \in \mathbb{R}$, $x \in \mathbb{R}^n$, $T(cx) = cT(x)$

When we say that $T \colon \mathbb{R}^n \longrightarrow \mathbb{R}^m$ is linear, it means that $T$ is a linear transformation. Under this definition, we note that for $\mathbb{R}^n, \mathbb{R}^m$, the zero map $T_0 \colon \mathbb{R}^n \longrightarrow \mathbb{R}^m$ defined by $T_0(x) = 0 \; \forall x \in \mathbb{R}^n$ is a linear transformation.

A real-valued linear transformation is such that for any $y \in \mathbb{R}^n$, the function $T_y \colon \mathbb{R}^n \longrightarrow \mathbb{R}^m$,

$$T_y(x) = \langle x, y \rangle$$

is linear. *Proof* by properties of $\langle \cdot, \cdot \rangle$.

In fact, functions of this form are the only real-valued linear transformations.

*Proof.* Suppose $T\colon \mathbb{R}^n \longrightarrow \mathbb{R}$ is linear. Define $y = (T(e_1), T(e_2), \ldots, T(e_n)) \in \mathbb{R}^n$ Then, for any $x \in \mathbb{R}^n$,

$$\begin{aligned} T(x) &= T(x_1 e_1 + \ldots + x_n) \\ &= x_1 T(e_1) + \ldots + x_n T(e_n) && \text{(By Linearity of } T\text{)} \\ &= \langle x, y \rangle = T_y(x) \implies T = T_y \end{aligned}$$

□

Linear transformations are closely related to matrices. For a $m \times n$ matrix $A$, $T_A\colon \mathbb{R}^n \longrightarrow \mathbb{R}^m$ defined by

$$T_A(x) = Ax$$

is linear (by properties of matrix by vector multiplication). In fact, every linear $T\colon \mathbb{R}^n \longrightarrow \mathbb{R}^m$ is of the form $T(x) = Ax$ for some $m \times n$ matrix $A$. This matrix is unique and is called **the matrix of** $T$.

*Proof.* Let $T\colon \mathbb{R}^n \longrightarrow \mathbb{R}^m$ be linear, set

$$A = \begin{bmatrix} T(e_1) & T(e_2) & \ldots & T(e_n) \end{bmatrix}$$

Then, for any $x \in \mathbb{R}^n$,

$$\begin{aligned} T(x) &= T(x_1 e_1 + \cdots + x_n e_n) \\ &= x_1 T(e_1) + \cdots + x_n T(e_n) \\ &= \begin{bmatrix} T(e_1) & \ldots & T(e_n) \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} = Ax \end{aligned}$$

□

Let's see a linear transformation in action.

**Example 11.** Rotation about the origin by $\frac{\pi}{2}$ radians counterclockwise in $\mathbb{R}^2$ is a linear transformation. Let $T_{\frac{\pi}{2}}\colon \mathbb{R}^2 \longrightarrow \mathbb{R}^2$ be linear. Its matrix is

$$A = \begin{bmatrix} T_{\frac{\pi}{2}}(e_1) & T_{\frac{\pi}{2}}(e_2) \end{bmatrix} = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}$$

We can conclude that every linear transformation $T\colon \mathbb{R}^n \longrightarrow \mathbb{R}^m$ is represented by a unique matrix $A$ in the sense that $T(x) = Ax \forall x \in \mathbb{R}^n$. This unique matrix $A$

$$A = \begin{bmatrix} T_1(e_1) & \ldots T_n(e_n) \end{bmatrix}$$

by applying $T$ to all the standard basis vectors in order. $(e_1, \ldots, e_n)$ is the standard basis of $\mathbb{R}^n$.

**Example 12.** $T_0\colon \mathbb{R}^n \longrightarrow \mathbb{R}^m$ is the zero transformation. Then, the matrix representation of $T_0$ is $[0]_{m \times n}$.

**Example 13.** $\mathrm{id}\colon \mathbb{R}^n \longrightarrow \mathbb{R}^n$ is the identity map on $\mathbb{R}^n$, $\mathrm{id}\,(x) = x$. id is linear and its matrix representation is,

$$A = \begin{bmatrix} \mathrm{id}\,(e_1) & \dots & \mathrm{id}\,(e_n) \end{bmatrix} = \begin{bmatrix} e_1 & \dots & e_n \end{bmatrix} = \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{bmatrix}$$
$$= \begin{bmatrix} \delta_{ij} \end{bmatrix}_{m\times n} = \mathrm{I}_n$$

**Example 14.** Let $T\colon \mathbb{R}^2 \longrightarrow \mathbb{R}^2$ be a reflection across the line $y = x$ in $\mathbb{R}^2$. Its matrix representation is

$$A = \begin{bmatrix} T\,(e_1) & T\,(e_2) \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$$

**Theorem 11.** *Every linear transformation is continuous on its domain.*

*Proof.* Let $T\colon \mathbb{R}^n \longrightarrow \mathbb{R}^m$ be linear, let $A = [v_1 \cdots v_n]$ be the matrix of $T$. For any $x \in \mathbb{R}^n$,

$$T\,(x) = Ax = \begin{bmatrix} v_1 & \dots & v_n \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} = x_1 v_1 + \cdots + x_n v_n$$

So $T$ is obtained from a composition of scalar multiplication and vector addition. Then, $T$ is continuous since composition of continuous functions. $\square$

**Theorem 12.** *Let $T, S\colon \mathbb{R}^n \longrightarrow \mathbb{R}^m$ be linear, let $c \in \mathbb{R}$. Then, $T + S$ and $cT$ are also linear.*

The proof is mostly skipped, however, it can be noted that,

$$\begin{aligned} (T+S)\,(x+y) &= T\,(x+y) + S\,(x+y) \\ &= (T\,(x) + T\,(y)) + (S\,(x) + S\,(y)) \\ &= (T\,(x) + S\,(x)) + (T\,(y) + S\,(y)) \\ &= (T+S)\,(x) + (T+S)\,(y) \end{aligned}$$

**Remark.** The matrix of $T$ is $A$ and the matrix of $S$ is $B$, then, it follows that the matrix of $T + S$ is $A + B$ and the matrix of $cT$ is $cA$.

### 4.2.1 Matrix by Matrix Multiplication

To mulyiply two matrices $A$ and $B$, we need the dimenstions to satisfy

$$\underbrace{A}_{m\times n}\underbrace{B}_{n\times l} = \underbrace{AB}_{m\times l}$$

**Definition 16.** $A$ is $m \times n$, $B$ is $n \times l$, where

$$B = \begin{bmatrix} v_1 & \cdots & v_l \end{bmatrix}$$

then,
$$AB = A \begin{bmatrix} v_1 & \cdots & v_l \end{bmatrix} = \begin{bmatrix} Av_1 & \cdots & Av_l \end{bmatrix}$$
where $v_1, \dots, v_l$ each are an $n$-dimensional column vector.

There is a nice formula that expresses the $i, j$ coordinate as,

$$(AB)_{ij} = \sum_{k=1}^{n} a_{ik} b_{kj} = \langle i^{\text{th}} \text{ row of } A, j^{\text{th}} \text{ column of } B\rangle$$

**Example 15.**

$$\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} \begin{bmatrix} 0 & 2 & 1 \\ -1 & 1 & 0 \end{bmatrix} = \begin{bmatrix} -2 & 4 & 1 \\ -4 & 10 & 3 \end{bmatrix}$$

Assuming the dimensions make sense, consider the following properties for matrices,

(1) $A(BC) = (AB)C$

(2) $A(B+C) = AB + AC$

(3) $(A+B)C = AC + BC$

It seems tempting to also say that matrix multiplication is commutative since it looks very similar to the properties of vector spaces, however, matrix multiplication is **not** commutative. That is to say,

$$AB \neq BA$$

**Example 16.** $A = \left[\begin{smallmatrix} 0 & -1 \\ 1 & 0 \end{smallmatrix}\right]$, $B = \left[\begin{smallmatrix} 1 & 0 \\ 0 & 0 \end{smallmatrix}\right]$, then,

$$AB = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix}$$

$$BA = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} = \begin{bmatrix} 0 & -1 \\ 0 & 0 \end{bmatrix}$$

It follows that $AB \neq BA$.

**Theorem 13.** *$T\colon \mathbb{R}^n \longrightarrow \mathbb{R}^m, S\colon \mathbb{R}^m \longrightarrow \mathbb{R}^l$ both $T, S$ linear. Let $A$ be matrix of $T$, let $B$ be matrix of $S$. Then, the matrix of $S \circ T$ is $\underbrace{B}_{l\times m}\underbrace{A}_{m\times n} = \underbrace{BA}_{l\times n}$ (makes sense because $S \circ T\colon \mathbb{R}^n \longrightarrow \mathbb{R}^l$)*

*Proof.* Trivial.

$$\begin{aligned} (S \circ T)(x) &= S(T(x)) \\ &= S(Ax) \\ &= B(Ax) \\ &= (BA)x \end{aligned}$$

□

**Definition 17. (Matrix Inverses)** An $n \times n$ matrix $A$ is invertible if there exists a $n \times n$ matrix $B$ such that

$$AB = I_n = BA$$

The notation is $B = A^{-1}$ that is to say, when $B = A^{-1}$, then $A = B^{-1}$.

**Remark.** It does not make sense to talk about a non-square matrix being invertible. If $A$ is not square, then there is no $B$ such that $AB$ and $BA$ both defined.

## 4.3 Matrix Norms

Let $A$ be an $m \times n$ matrix, let $T\colon \mathbb{R}^n \longrightarrow \mathbb{R}^m$, $T(x) = Ax$. Recall that $|T(x)|$ is continuous and $S^{n-1} = \{x \in \mathbb{R}^n : |x| = 1\}$. By the *extreme value theorem*, $\exists x_0 \in S^{n-1}$ s.t.

$$|Ax| = |T(x)| = \max\left\{|T(x)| : x \in S^{n-1}\right\} = \max\left\{|Ax| : x \in S^{n-1}\right\}$$

We define the norm of $A$ to be

$$\|A\| = \max\left\{|Ax| : x \in S^{n-1}\right\}$$

For the linear transformation $T(x) = Ax$, this quantity is also defined to be the norm of

$$\|T\| = \|A\| = \max\left\{|T(x) : x \in S^{n-1}\right\}$$

The following are examples of matrix norms,

(1) $T_0 \colon \mathbb{R}^n \longrightarrow \mathbb{R}^m$ is zero transformation, $\|T\| = 0$ since $|T(x)| = 0\ \forall x \in \mathbb{R}^n$. Similarly, $\|0_{m\times n}\| = 0$

(2) Let $R_\theta \colon \mathbb{R}^2 \longrightarrow \mathbb{R}^2$ be counterclockwise rotation by $\theta$. Then $\|R_\theta\| = 1$ since $|R_\theta(x)| = |x| = 1\ \forall x \in S^{n-1}$

In general, to compute $\|T\|$, $\|A\|$, we have to solve a multivariable optimization problem. We're not interested in that, we're interested in the theoretical applications of matrix norms, for example,

**Theorem 14.** *Let $A$ be an $m \times n$ matrix. Then, for any $x \in \mathbb{R}^n$, we have*

$$|Ax| \le \|A\| \cdot |x|$$

*(also, if $T \colon \mathbb{R}^n \longrightarrow \mathbb{R}^m$ is linear, then $|T(x)| \le \|T\|\,|x|\ \forall x \in \mathbb{R}^n$)*

*Proof.* Let $x \in \mathbb{R}^n$. When $x = 0$, result is trivial, so we may assume $x \neq 0$, i.e., $|x| > 0$.

Then, we have $\frac{1}{|x|}x \in S^{n-1}$ since $\left|\frac{1}{|x|}\right| = \frac{1}{|x|}|x| = 1$.

Then,

$$|Ax| = \left|A\left(\frac{|x|}{|x|}x\right)\right| = |x|\left|A\left(\frac{x}{|x|}\right)\right| \le |x|\,\|A\|$$

by the definition $\|A\| = \max\left\{|Ay| \colon y \in S^{n-1}\right\}$ □

**Proposition.** Let $A$ be a $m \times n$ matrix. Then,

$$\|A\| = 0 \iff A = 0_{m\times n}$$

*Proof.* ( $\Longleftarrow$ ) trivial.

( $\Longrightarrow$ ) Suppose $\|A\| = 0$. By the theorem, we have $|Ax| \le \|A\| \cdot |x| = 0\ \forall x \in \mathbb{R}^n$. Thus, $|Ax| = 0$, hence $Ax = 0\ \forall x \in \mathbb{R}^n$.

Then, $A = 0_{m\times n}$. □

## 4.4 Towards Defining the Derivative

In single-variable calculus, $f \colon \mathbb{R} \longrightarrow \mathbb{R}, a \in \mathbb{R}$,

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

when the limit actually exists. We will now rewrite $f'(a)$ as

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

We say that $f'(a) = \lim_{h\to 0} \frac{f(a+h) - f(a)}{h} \rightsquigarrow f'(a) \approx \frac{f(a+h) - f(a)}{h}$ when $h$ is small enough and $\rightsquigarrow f(a+h) \approx f(a) + f'(a)h$ which is the linear approximation of $f$ near $a$.

We introduce an error term $\varepsilon(h) = f(a+h) - (f(a) + f'(a)h)$. It's intuitive to see that

$$\varepsilon(h) \longrightarrow 0 \text{ as } h \longrightarrow 0$$

But this is not a strong enough statement. A stronger statement is to say that

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

This says that not only is $\varepsilon(h) \longrightarrow 0$ as $h \longrightarrow 0$ but in fact $\varepsilon(h)$ is going to 0 much faster than $h$ goes to 0. In general, what we're trying to achieve is to minimize a bounded area between the function and the derivative at that point which is $\varepsilon(h)$.

### 4.4.1 Generalizing to Several Variables

In single-variable, $T_a(h) = f'(a)h$ is a linear transformation $\mathbb{R} \longrightarrow \mathbb{R}$. So, for multivariable functions, the derivative at a vector will not be a vector, but a linear transformation

$$Df_a \colon \mathbb{R}^n \longrightarrow \mathbb{R}^n$$

where $f \colon \mathbb{R}^n \longrightarrow \mathbb{R}^m$. The matrix of $Df_a$ is called a **Jacobian Matrix**.

We also have to be careful in saying "goes to 0 faster than $h \longrightarrow 0$". Let $f \colon \mathbb{R}^n \longrightarrow \mathbb{R}^m$,

$$\lim_{h\to 0_n} \frac{|f(a+h) - f(a) - Df_a(h)|}{|h|} = 0$$

A nice way to think about this is,

$$f(a+h) = f(a) + Df_a(h) + o(h)$$

where instead of $\varepsilon(h)$, we write $o(h)$ (*little oh of* $h$) to express the error terms where $\dfrac{|o(h)|}{|h|} \longrightarrow 0$ as $h \longrightarrow 0$. We will study the behavior of functions through $o(h)$ and other forms that also trace functional behavior through the *Bachmann-Landau Notation* in the next section.

# 5 Derivatives in Several Variables

## 5.1 Bachmann-Landau Notation

**Definition 18.** Consider $\varphi \colon \mathcal{B}(0_n, \varepsilon) \longrightarrow \mathbb{R}^m$

(1) $\varphi$ is $o(1)$ ("little oh of 1" or "smaller than order 1") if for every $c > 0$, $|\varphi(h)| \le c$ for all small enough $h$.

$|\varphi(h)| \le c$ is an abbreviation for 'there is $\delta > 0$ s,t, $|h| < \delta \implies |\varphi(h)| \le c$'. Note that $\delta$ depends on $c$. For example, a smaller $c$ will require a smaller $\delta$, generally. Equivalently, $\varphi(h) \longrightarrow 0$ as $h \longrightarrow 0$.

(2) $\varphi$ is $o(h)$ ("little oh of $h$" or "smaller than order $h$") if for every $c > 0$, $|\varphi(c)| \le c|h|$ for all small enough $h \implies \frac{|\varphi(h)|}{|h|} \le c$.

Equivalently, $\lim_{h\longrightarrow 0} \frac{|\varphi(h)|}{|h|} = 0$ and $\varphi(h) = 0$. The idea is that $\varphi$ is $o(h)$ means $\varphi$ goes to 0 faster than $h$ goes to 0.

(i) (3) $\varphi$ is $O(h)$ ("big oh of $h$" or "of order $h$") *if there is a* $c > 0$ such that $|\varphi(h)| \le c|h|$ for all small enough $h$.

This is a much weaker claim than $o(h)$ since it asks for only *one instance* but $o(h)$ asks *for every* $c$. Eauivalently, there is a bound on $\frac{|\varphi(h)|}{|h|}$ for all small enough $h$. The idea is that $\varphi$ is $O(h)$ means $\varphi$ goes to 0 *at least* as fast as $h$ goes to 0.

**Proposition.** $o(h) \subsetneq O(h) \subsetneq o(1)$. $\varphi$ is $o(h) \implies \varphi$ is $O(h) \implies \varphi$ is $o(1)$.

**Theorem 15.** *Every linear* $T\colon \mathbb{R}^n \longrightarrow \mathbb{R}^m$ *is* $O(h)$.

*Proof.* Set $c = \|T\|$. Then,

$$|T(h)| \le \|T\| \cdot |h| = c|h| \forall h \in \mathbb{R}^n$$

□

**Theorem 16.** *A linear* $T$ *is* $o(h) \iff T$ *is the zero transformation.*

*Proof.* ( $\Longleftarrow$ ) trivial.

( $\implies$ ) Suppose $T\colon \mathbb{R}^n \longrightarrow \mathbb{R}^m$ is linear and $o(h)$. $\|T\| = \max\left\{|T(x)|\colon x \in S^{n-1}\right\}$, so by Extreme Value Theorem, $\exists x_0 \in S^{n-1}$ s.t. $\|T\| = |T(x_0)|$.

Set $h = sx_0, s \in (0,1)$.

$$\begin{aligned} 0 &= \lim_{h\to 0} \frac{|T(h)|}{|h|} = \lim_{s\to 0} \frac{|T(sx_0)|}{|sx_0|} = \lim_{s\to 0} \frac{s|T(x_0)|}{s} \\ &= \lim_{s\to 0} |T(x_0)| = |T(x_0)| = \|T\| \end{aligned}$$

$\implies 0 = \|T\| \implies T$ is the zero transformation. □

Keep in mind the notation. $o(h)$ is often used to denote some unspecified $o(h)$ function. For example, $f(h) = g(h) + o(h) \iff f(h) - g(h) = o(h) \iff f(h) - g(h)$ is $o(h)$.

**Proposition.** $o(h) + o(h) = o(h)$.

In other words, the sum of any two $o(h)$ functions is also $o(h)$. We will follow this with saying, $c \cdot o(h) = o(h)$ for any $c \in \mathbb{R}$, that is to say, the persistence of $o(h)$ under scalar multiplication.

*Proof.* $o(h) + o(h) = o(h)$

Let $\varphi, \psi\colon \mathcal{B}(0_n, \varepsilon) \longrightarrow \mathbb{R}^m$ both be $o(h)$. Then,

$$\lim_{h\to 0} \frac{|\varphi(h) + \psi(h)|}{|h|} \le \lim_{h\to 0} \frac{|\varphi(h)| + |\psi(h)|}{|h|} = \lim_{h\to 0} \frac{|\varphi(h)|}{|h|} + \lim_{h\to 0} \frac{|\psi(h)|}{|h|} = 0 + 0 = 0$$

Also, $(\varphi + \psi)(0) = \varphi(0) + \psi(0) = 0 + 0 = 0$.

Thus, $\varphi + \psi$ is $o(h)$. □

**Proposition.** $o(1) + o(1) = o(1), c \cdot o(1) = o(1)$ and $O(h) + O(h) = O(h), c \cdot O(h) = O(h)$ for any $c \in \mathbb{R}$.

**Proposition.** $o(1) \cdot O(h) = o(h)$. i.e., if $\varphi$ is $o(1)$, $\psi$ is $O(h)$, then, $|\varphi(h)| \cdot |\psi(h)|$ is $o(h)$.

The proof will not be stated, however, the sketch of the proof is,

$$\lim_{h\to 0} \frac{|\varphi(h)| \cdot |\psi(h)|}{|h|} = \lim_{h\to 0} |\varphi(h)| \left( \frac{|\psi(h)|}{|h|} \right)$$

$\frac{|\psi(h)|}{|h|}$ is bounded since $\psi$ is $O(h)$. It follows that,

$$\lim_{h\to 0} |\varphi(h)| \left( \frac{|\psi(h)|}{|h|} \right) \leq \lim_{h\to 0} |\varphi(h)| \cdot M = M \lim_{h\to 0} |\varphi(h)| = M \cdot 0 = 0$$

Note that $\varphi(h)$ is $o(1)$.

**Corollary.** $o(h) \cdot o(h) = o(h)$

## 5.2 Defining the Derivative

Recall that

$$\begin{cases} |o(1)| \longrightarrow 0 \\ \dfrac{|O(h)|}{|h|} \text{ is bounded} \\ \dfrac{|o(h)|}{|h|} \to 0 \end{cases}$$

as $h \to 0$ and $o(h) \subsetneq O(h) \subsetneq o(1)$. Furthermore, $o(h) + o(h) = o(h), O(h) + O(h) = O(h)$, all linear transformations are $O(h)$ and the only $o(h)$ linear transformation is the zero map.

**Definition 19. (Interior point)** Let $A \subseteq \mathbb{R}^n$, then $a \in A$ is an *interior point* of $A$ if $\exists \varepsilon > 0$ s.t. $\mathcal{B}(a, \varepsilon) \subseteq A$.

**Remark.** Points on the boundary of $A$ are *not* interior points.

**Definition 20.** Let $A \subseteq \mathbb{R}^n$, $a \in A$ be an interior point. $f \colon A \longrightarrow \mathbb{R}^m$ is differentiable at $a$ if there is a linear transformation $T_a \colon \mathbb{R}^n \longrightarrow \mathbb{R}^m$ s.t.

$$f(a+h) - f(a) - T_a(h) = o(h)$$

**Remark.** $f(a+h) = f(a) + T_a(h) + o(h)$. $T_a$ is a linear approximation of $f$ near $a$ whose error term is $o(h)$.

**Proposition.** If $f$ is differentiable at $a$, then $T_a$ is *unique.*

We will now use the notation $T_a = Df_a$. The matrix of $Df_a$ is denoted $f'(a)$ and is called the *Jacobian matrix* of $f$ at a.

$$Df_a = f'(a)\, h$$

**Remark.** This proposition means that when $Df_a$ exists, it is the single best linear approximation.

*Proof.* Suppose $T_a, S_a \colon \mathbb{R}^n \longrightarrow \mathbb{R}^m$ both satisfy the definition of derivative. We want to show $T_a = S_a$. It suffices to show $T_a - S_a$ is the zero transformation.

$$\begin{aligned} T_a(h) - S_a(h) &= (f(a+h) - f(a) + o(h)) - (f(a+h) - f(a) + o(h)) \\ &= o(h) + o(h) = o(h) \end{aligned}$$

The only $o(h)$ linear transformation is the zero mao. Therefore, $T_a - S_a$ is the zero transformation $\implies T_a = S_a$. ☐

We must be careful with the Bachmann-Landau notation and avoid writing $o(h) - o(h) = 0$ since $o(h)$ is not the typical kind of function we deal with, it's simply a general statement regarding the error terms for a function.

**Example 17.** Let $T\colon \mathbb{R}^n \longrightarrow \mathbb{R}^m$ be linear. Then, for any $a \in \mathbb{R}^n, T$ is differentiable at $a$ and $DT_a = T$.

*Proof.* We just need to check $T(a+h) - T(a) - DT_a(h) = o(h)$.

$$\begin{aligned} T(a+h) &= T(a+h) - T(a) - T(h) \\ &= T(a) + T(h) - T(a) - T(h) \\ &= 0 = o(h) \end{aligned}$$

The zero function is $o(h)$. ☐

**Example 18.** Constant maps are differentiable at every interior point of their domain and their derivative is the zero transformation.

*Proof.* Trivial. ☐

**Example 19.** $f\colon \mathbb{R}^2 \longrightarrow \mathbb{R}, f(x,y) = x^2 + y^2$. $f$ is differentiable at every $(a,b) \in \mathbb{R}^2$ and

$$Df_{(a,b)}(h,k) = 2ah + 2bk$$

where $(h,k) \in \mathbb{R}^2$.

*Proof.* Easy to check that $T(h,k) = 2ah + 2bk$ is linear. It's enough to show that $f(a+h, b+h) - f(a,b) - T(h,k)$ is $o(h,k)$.

$$\begin{aligned} f(a+h, b+h) &= (a+h)^2 + (b+k)^2 - \left(a^2 + b^2\right) - (2ah + 2bk) \\ &= a^2 + 2ah + h^2 + b^2 + 2bh + h^2 - a^2 - b^2 - 2ah - 2bk \\ &= h^2 + k^2 = |(h,k)|^2 \end{aligned}$$

It's easy to check that $|(h,k)|^2$ is $o(h,k)$ since $\lim \frac{|(h,k)|^2}{|(h,k)|} = 0$ ☐

The *Jacobian matrix* at $(a,b)$ where $Df_{(a,b)}\colon \mathbb{R}^2 \longrightarrow \mathbb{R}$ is $f'(a,b) = \begin{bmatrix} 2a & 2b \end{bmatrix}$

**Theorem 17.** *$f\colon A \longrightarrow \mathbb{R}^m$, $A \subseteq \mathbb{R}^n, a \in A$ interior point. If $f$ is differentiable at $a$, then $f$ is continuous at $a$.*

*Proof.*

$$\begin{aligned} f(a+h) - f(a) &= (f(a+h) - f(a) - Df_a(h)) + Df_a(h) \\ &= O(h) + O(h) \\ &= O(h) = o(1) \end{aligned}$$

Thus, $f(a+h) - f(a) \longrightarrow 0$ as $h \longrightarrow 0$, hence, $f(a+h) \longrightarrow f(a)$ as $h \longrightarrow 0$, so $f$ is continuous at $a$. ☐

**Theorem 18.** ***(Linearity of Differentiation)*** *$A \subseteq \mathbb{R}^n, f\colon A \longrightarrow \mathbb{R}^m, a \in A$ interior point. If $f, g$ both differentiable at $a$, then,*

*(1) $f+g$ is differentiable at $a$ and $D\left(f+g\right)_a = Df_a + Dg_a$*

*(2) For any $c \in \mathbb{R}$, $cf$ is differentiable at $a$ and $D\left(cf\right)_a = c\left(Df_a\right)$*

*Proof.* (1) $T\left(h\right) = Df_a\left(h\right) + Dg_a\left(h\right)$ is linear since it is the sum of two linear maps.

$$
\begin{aligned}
&\left(f+g\right)\left(a+h\right) - \left(f+g\right)\left(a\right) - T\left(h\right) \\
&\qquad = f\left(a+h\right) - g\left(a+h\right) - \left(f\left(a\right) + g\left(a\right)\right) - \left(Df_a\left(h\right) + Dg_a\left(h\right)\right) \\
&\qquad = \left(f\left(a+h\right) - f\left(a\right) - Df_a\left(h\right)\right) + \left(g\left(a+h\right) - g\left(a\right) - Dg_a\left(h\right)\right)
\end{aligned}
$$

By definition of the derivative,

$$= o\left(h\right) + o\left(h\right) = o\left(h\right)$$

□

## 5.3 Derivative Rules

The goal of this subsection is to derive the chain rule using the *Bachmann-Landau Notation* and subsequently, derive the quotient/product rules for serveral variables. In single-variable calculus, it's common to see the chain rule to be introduced after the quotient/product rule, however, will see that we require the chain rule to derive the product rule.

*Proof.* $o\left(O\left(h\right)\right) = o\left(h\right)$

Suppose $\psi\left(h\right)$ is $O\left(h\right)$, $\varphi\left(k\right)$ is $o\left(k\right)$, $\psi \circ \varphi$ makes sense (i.e., composable) defined on some $\mathcal{B}\left(0_n, \varepsilon\right)$. We want to show that $\varphi\left(\psi\left(k\right)\right)$ is $o\left(h\right)$.

We know that $|\psi\left(h\right)| \leq d|h|$ since $\psi$ is $O\left(h\right)$. Let $c > 0, |\varphi\underbrace{\left(\psi\left(h\right)\right)}_{k}| \leq a|\underbrace{\psi\left(h\right)}_{k}| = ad|h|$.

We want to show that $\exists \delta > 0$ s.t. $|h| < \delta \implies |\varphi\left(\psi\left(h\right)\right) \leq c|h|$.

Since $\psi\left(h\right)$ is $O\left(h\right)$, $\exists d > 0$ and $\delta_1 > 0$ s.t. $|h| < \delta_1 \implies |\psi\left(h\right)| \leq d|h|$.

Since $\varphi\left(k\right)$ is $o\left(k\right)$, $\exists \delta_2 > 0$ s.t. $|k| < \delta_2 \implies |\varphi\left(k\right)| \leq \dfrac{c}{d}|k|$.

Since $\psi\left(h\right)$ is also $o\left(1\right)$, $\exists \delta_3 > 0$ s.t. $|h| < \delta_3 \implies |\psi\left(h\right)| < \delta_2$.

Set $\delta = \min\left(\delta_1, \delta_3\right)$, then, for any $h$ with $|h| < \delta$,

$$|\varphi\left(\psi\left(h\right)\right)| \underbrace{\leq}_{k=\psi(h), |h|<\delta_3 \implies |k|=|\psi(h|<\delta_2)} \frac{c}{d}|\psi\left(h\right)| \overbrace{\leq}^{|h|<\delta_1} \frac{c}{d} \cdot d|h| = c|h|$$

□

**Theorem 19.** ***(Chain Rule.)*** *$A \subseteq \mathbb{R}^n, B \subseteq \mathbb{R}^m, f\colon A \longrightarrow \mathbb{R}^m, f\left(A\right) \subseteq B, g\colon B \longrightarrow \mathbb{R}^l$ (so $g \circ f$ makes sense), $a \in A$ interior point of $A$, $f\left(a\right) \in B$ interior point of $B$. If $f$ is differentiable at $a$ and $g$ is differentiable at $f\left(a\right)$, then $g \circ f$ is differentiable at $a$, and*

$$D\left(g \circ f\right)_a = Dg_{f(a)} \circ Df_a$$

Note that this is matrix multiplication so the order matters.

*Proof.* Clearly, $Dg_{f(a)} \circ Df_a$ is linear (by composition of linear transformations is a linear transformation). It follows that,

$$\begin{aligned}(g \circ f)(a+h) = g(f(a+h)) = g\left(f(a) + \underbrace{Df_a(h) + o(h)}_{k}\right) \\ &= g(f(a)) + Dg_{f(a)}(Df_a(h) + o(h)) + o(Df_a(h) + o(h)) \\ &= (g \circ f)(a) + Dg_{f(a)}\left(Df'_a(h)\right) + \underbrace{Dg_{f(a)}(o(h)) + o(Df_a(h) + o(h))}_{\text{show that this is } o(h)}\end{aligned}$$

$Dg_{f(a)}(o(h))$ is simply $O(o(h))$. Thus, it follows that,

$$\begin{aligned}Dg_{f(a)}(o(h)) + o(Df_a(h) + o(h)) &= O(o(h)) + o(O(h) + o(h)) && \text{(Lin. Tra. are } O(h)) \\ &= o(h) + o(O(h)) && (O(o(h)) = o(h)) \\ &= o(h) + o(h) = o(h) && (o(O(h)) = o(h))\end{aligned}$$

□

**Lemma.** It follows,

(1) $p\colon \mathbb{R}^2 \longrightarrow \mathbb{R}, p(x,y) = xy$ is differentiable on $\mathbb{R}^2$ and $Dp_{(a,b)}(h,k) = bh + ak, (h,k) \in \mathbb{R}^2$.

(2) $r\colon \mathbb{R}-\{0\} \longrightarrow \mathbb{R}, r(x) = \dfrac{1}{x}$ is differentiable on its domain, $r'(a) = \dfrac{-1}{a^2}, Dr_a(h) = \dfrac{-h}{a^2}$.

*Proof.* (1) $p(a+h, b+k) - p(a,b) - (bh + ak)$. We want to show that the whole expression is $o((h,k))$.

$$\begin{aligned}&= (a+h)(b+k) - abh - bh - ak \\ &= ab + ak + bh + hk - ab - bh - ak \\ &= hk\end{aligned}$$

Claim that $hk$ is $o((h,k))$.

$$|hk| = |h||k| \leq |(h,k)| \cdot |(h,k)| = \underbrace{|(h,k)|}_{o(h,k)}{}^2$$

So, done by a simple argument using the next lemma. □

**Lemma.** $(|\varphi(h)| \leq o(h) \,\forall$ small enough $h) \implies \varphi(h)$ is $o(h)$. Same for $O(h), o(1)$.

*Proof.* Simple. □

**Theorem 20.** *Let $A \subseteq \mathbb{R}^n, f, g\colon A \longrightarrow \mathbb{R}$, both differentiable at an interior point $a \in A$.*

*(1) $fg\colon A \longrightarrow \mathbb{R}$ is differentiable at $a$ and $D(fg)_a = g(a) \cdot Df_a + f(a) \cdot Dg_a$*

*(2) If $g(a) \neq 0$, then $f/g$ is differentiable at $a$ and*

$$D\left(\frac{f}{g}\right)_a = \frac{g(a)\,Df_a - f(a)\,Dg_a}{g(a)^2}$$

**Theorem 21** (Componentwise Nature of Differentiability). *$A \subseteq \mathbb{R}^n, a \in A$ interior point of $A$, $f\colon A \longrightarrow \mathbb{R}^n$, $f\colon A \longrightarrow \mathbb{R}^m$. $f = (f_1, \cdots, f_m)$ component functions of $f$, $f_i\colon A \longrightarrow \mathbb{R}$.*

*$f$ is differentiable at $a \iff$ each $f_i$ for $i \in \{1, 2, \cdots, m\}$ is differentiable at $a$.*

*Proof.* ( $\Longrightarrow$ ) Assume $f$ is differentiable at $a$, so,

$$f(a+h) - f(a) - Df_a(h) \qquad \text{is } o(h)$$

By size bounds,

$$|f_i(a+h) - f_i(a) - (Df_a(h))_i| \leq |f(a+h) - f(a) - Df_a(h)| \qquad = o(h)$$

If a function above is bounded by $o(h)$, then the function also must be $o(h)$. It follows that $f_i(a+h) - f_i(a) - (Df_a(h))_i$ is $o(h)$. Hence, $D(f_i)_a$ exists and

$$D(fi)_a(h) = (Df_a(h))_i$$

( $\Longleftarrow$ ) Suppose $|f_i(a+h) - f_i(a) - (Df_a(h))|$ is $o(h)$ for each $i = 1, \cdots, m$. Then,

$$|f(a+h) - f(a) - Df_a(h) \leq \sum_{i=1}^{m} |f_i(a+h) - f_i(a) - (Df_a(h))_i|$$

Notice that the right hand side is simply sum of $o(h)$, thus, $f(a+h) - f(a) - Df_a(h)$ is $o(h)$. ☐

## 5.4 Partial Derivatives and the Jacobian

The *Jacobian* is the matrix of $Df_a$.

**Definition 21** (Partial Derivative). $A \subseteq \mathbb{R}^n, a \in A$ interior point of $A$. $f\colon A \longrightarrow \mathbb{R}$. For any $j \in \{1, \cdots, n\}$ define $\varphi_j\colon \mathbb{R} \longrightarrow \mathbb{R}, \varphi_j(t) = f(a_1, \cdots, a_{j-1}, t, a_j, \cdots, a_n)$. The $j^{\text{th}}$ partial derivative of $f$ at $a$ is

$$D_j f(a) = \varphi_j'(a)$$

This is under under the assumption that the right side derivative exists. Equivalently, we can write the $j^{\text{th}}$ partial derivative as

$$D_j f(a) = \lim_{t \to 0} \frac{f(a + te_j) - f(a)}{t}, t \in \mathbb{R}$$

Equivalently,

$$f(a + te_j) + D_j f(a) \cdot t + o(t)$$

Computing partial derivatives is relatively easy. Just pretend all other variables are constant.

**Example 20.** $f(x, y, z) = e^y \cos x + z$.

First, pretend $y, z$ are constants, $D_1 f(x, y, z) = -e^y \sin x$. Next, pretend $x, z$ are constants, $D_2 f(x, y, z) = e^y \cos x$. Last, pretend $x, y$ are constants, $D_3 f(x, y, z) = 1$.

Partial derivatives may be more commonly seen in the classical partial derivative notation. See below

$$D_1 f(x,y,z) = \frac{\partial f}{\partial x}(x,y,z)$$
$$D_2 f(x,y,z) = \frac{\partial f}{\partial y}(x,y,z)$$
$$D_2 f(x,y,z) = \frac{\partial f}{\partial z}(x,y,z)$$

$x, y, z$ are overloaded here in the classical notation but it's good to be aware of this notation.

## 5.5 Necessary & Sufficient Condition Theorem

**Theorem 22** (Necessary Condition)**.** *$A \subseteq \mathbb{R}^n, a \in A$ interior point of $A, f\colon A \to \mathbb{R}^m$ (vector-valued). If $f$ is differentiable at $a$, then every partial derivative $D_j f_i(a)$ exists $i \leq k \leq n, 1 \leq i \leq m$. Morover $D_j f_i(a)$ is the $(i,j)$ entry of the Jacobian matrix $f'(a)$, so*

$$f'(a) = [D_j f_i(a)]_{m\times n} = \begin{bmatrix} D_1 f_1(a) & D_2 f_1(a) & \cdots & D_n f_i(a) \\ \vdots & \vdots & \ddots & \vdots \\ D_1 f_m(a) & D_2 f_m(a) & \cdots & D_n f_m(a) \end{bmatrix}$$

*Proof.* We want to show that $D_j f_i(a)$ exists.

$$f'(a)\, e_j = j^{\text{th}} \text{ column of } f'(a) \implies \left(f'(a)\right)_{ij} = \left(f'(a)\, e_j\right)_i$$

Since $f$ is differentiable at $a$, $f(a+h) - f(a) - Df_a(h)$ is $o(h)$. Specialize $h = te_j, f(a+te_j) - f(a) - Df_a(te_j)$ is $o(te_j) - o(t)$. Every coordinate function is also $o(t)$. It follows that

$$\implies f_i(a+te_j) - f_i(a) - t\left(f'(a)\, e_j\right)_i \text{ is } o(t)$$

i.e., $\left(f'(a)\, e_j\right)_i$ satisfies the definition of $D_j f_i(a)$. □

**Remark.** Converse of the theorem is not true: each partial exists at $a \not\rightarrow$ differentiable at $a$.

**Example 21.**

$$f(x,y) = \begin{cases} \dfrac{2xy}{x^2+y^2} & \text{if } (x,y) \neq (0,0) \\ 0 & \text{if } (x,y) = (0,0) \end{cases}$$

$D_1 f(0,0), D_2 f(0,0)$ both exist and $= 0$. (on coordinate axes, $f$ is constant $= 0$.) However, the function is not even continuous at $(0,0)$ so it's not differentiable at $(0,0)$. When $Df_a$ exists, we must have $f'(a) = [D_j f_i]_{m\times n}$ i.e., we know its Jacobian matrix. This is the *necessary condition.* What about sufficient condition? That is to say, what is sufficient to show that $Df_a$ exists.

**Theorem 23** (Sufficient Condition)**.** *$A \subseteq \mathbb{R}^n, a \in A$ interior point of $A, f\colon A \longrightarrow \mathbb{R}^m$. Suppose every partial derivative of $f$ exists on some ball around $a$ and each is continuous at $a$. Then, $f$ is differentiable at $a$.*

*Proof.* By componentwise nature of differentiability, we may assume $m = 1$ (so $f$ is real-valued without any loss of generality).

For simplicity of notation, we assume $n = 2$. (This does lose generality but the argument for general cases is very similar).

Interior point $(a,b) \in A \subseteq \mathbb{R}^2$. By the necessary condition theorem, we know that our only candidate for $f'(a,b)$ is

$$\begin{bmatrix} D_1 f(a,b) & D_2 f(a,b) \end{bmatrix}$$

(simplified the case to 2 variables and real valued). I.e., we have

$$Df_{(a,b)}(h,k) = \begin{bmatrix} D_1 f(a,b) & D_2 f(a,b) \end{bmatrix} \begin{bmatrix} h \\ k \end{bmatrix} = hD_1 f(a,b) + kD_2 f(a,b)$$

$$\begin{aligned}
f(a+h,b+k) - f(a,b) &= (f(a+h,b+k) - f(a,b+k)) + (f(a,b+k) - f(a,b)) \\
&= (D_1 f(a,b+k) \cdot h + o(h)) + (D_2 f(a,b) \cdot h + o(k)) \\
&= D_1 f(a,b+k) \cdot h + D_2 f(a,b) \cdot k + o(h) + o(k) \\
&= (D_1 f(a,b) + o(1)) \cdot h + D_2 f(a,b) \cdot k + o(h,k) \\
&= D_1 f(a,b) \cdot h + o(1) \cdot h + D_2 f(a,b) \cdot k + o(h,k) \\
&= D_1 f(a,b) \cdot h + o(1) \cdot o(h) + D_2 f(a,b) \cdot k + o(h,k) \\
&= D_1 f(a,b) \cdot h + o(h,k) + D_2 f(a,b) \cdot k + o(h,k) \\
&= D_1 f(a,b) \cdot h + D_2 f(a,b) \cdot k + o(h,k) \\
&= \begin{bmatrix} D_1 f(a,b) & D_2 f(a,b) \end{bmatrix} \begin{bmatrix} h \\ k \end{bmatrix} + o(h,k)
\end{aligned}$$

□

To summarize, the *necessary condition* theorem suggests

$(f$ differentiable at $a) \implies \big($ all partial derivatives exist at $a$ and $f'(a) = [D_i f_i(a)]\big)$

Converse false. The *sufficient condition* theorem suggests

(All partial derivatives at $a$/near $a$ and are continuous at $a$) $\implies$ ($f$ differentiable at $a$)

Converse false. See next example.

**Example 22.**

$$f(x,y) = \begin{cases} (x^2+y^2)\sin\left(\frac{1}{\sqrt{x^2+y^2}}\right) & \text{if } (x,y) \neq (0,0) \\ 0 & \text{otherwise} \end{cases}$$

Easy to show that $f$ is differentiable at the origin and $D_f(0,09)$ is the zero transformation.

$$\begin{aligned}
&|f(0+h,0+k) - f(0,0) - 0| = |f(h,k)| \\
&\qquad = |h^2+k^2| |\sin\left(\frac{1}{\sqrt{h^2+k^2}}\right)| \leq 1 (h,k)|^2 \cdot 1 = o(h,k)
\end{aligned}$$

So, $f$ is differentiable at $(0,0)$ but the partials are at continuous at $(0,0)$. We will show for $D_1 f$. $f(x,0) = x^2 \sin\left(\frac{1}{|x|}\right) = x^2 \sin\left(\frac{1}{x}\right)$ when $x > 0$. Here, the partial derivatives are

$$\begin{aligned} D_1 f(x,0) &= 2x \sin\left(\frac{1}{x}\right) + x^2 \cos\left(\frac{1}{x}\right)\left(-\frac{1}{x}\right) \\ &= 2x \sin\left(\frac{1}{x}\right) - \cos\left(\frac{1}{x}\right) \end{aligned}$$

What you will notice is that $2x \sin\left(\frac{1}{x}\right) \to 0$ as $x \to 0^+$ while $\cos\left(\frac{1}{x}\right)$ oscillates wildly as $x \to 0$. The limit does not exist as $x \to 0^+$.

**Example 23.**

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

Away from $(0,0)$,

$$D_1 f(x,y) = \frac{\left(x^2+y^2\right)(2xy) - x^2 y (2x)}{\left(x^2+y^2\right)^2} = \frac{2xy^3}{\left(x^2+y^2\right)^2}$$

$$D_2 f(x,y) = \frac{\left(x^2+y^2\right)x^2 - x^2 y (2y)}{\left(x^2+y^2\right)^2} = \frac{x^2\left(x^2-y^2\right)}{\left(x^2+y^2\right)^2}$$

So, partials exist on $\mathbb{R}^2 - \{(0,0)\}$ and are continuous (rational) functions on $\mathbb{R}^2 - \{(0,0)\}$. Thus, by the sufficient condition theorem, $f$ is differentiable at every $(a,b) \neq (0,0)$, and

$$f'(a,b) = \begin{bmatrix} \frac{2ab^3}{\left(a^2+b^2\right)^2} & \frac{a^2\left(a^2-b^2\right)}{\left(a^2+b^2\right)^2} \end{bmatrix}$$

The case for $(0,0)$: partials are not continuous at $(0,0)$.

$$|D_1 f(h,0) - f(0,0) - 0| = |0-0-0| = 0 = o(h)$$

$$|D_2 f(0,k) - f(0,0) - 0| = |0-0-0| = 0 = o(k)$$

Thus, $D_1 f(0,0) = D_2 f(0,0) = 0$. So, if $f$ is differentiable at 0, we must have

$$f'(0,0) = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$$

i.e., $Df_{(0,0)}$ is the zero transformation. Looking at the line $h = k$, then,

$$\left|\frac{h^3}{h^2+h^2}\right| = \frac{|h|^3}{2|h|^2} = \frac{|h|}{2}$$

This is not $o(h)$ since $\frac{|h|}{2|h|} \to \frac{1}{2} \neq 0$. Thus, $f$ is not differentiable at $(0,0)$. It follows that the only $o(h)$ is possible if $Df_{(0,0)}$ is the zero transformation, however, we just showed that it is not $o(h)$.

By Chain Rule, we know

$$D\left(g\circ f\right)_a = Dg_f\left(a\right)\circ Df_a$$

We want to now know the entries of the Jacobian matrix of $\left(g\circ f\right)'\left(a\right)$. We know from the necessary condition theorem that the $(i,j)$ entry of $\left(g\circ f\right)'\left(a\right)$ is $D_j\left(g\circ f\right)_i\left(a\right)$.

**Theorem 24.** *$A\subseteq\mathbb{R}^n, a\in A$ interior point of $A, f\colon A\longrightarrow\mathbb{R}^m, f\left(a\right)\subseteq\mathcal{B}\subseteq\mathbb{R}^m, g\colon\mathcal{B}\longrightarrow\mathbb{R}^l, f\left(\right)\in\mathcal{B}$ interior of $\mathcal{B}$. If $f$ is differentiable at $a$ and $g$ is differentiable at $f\left(a\right)$, then, $g\circ f$ is differentiable at $a$, so all partials exist and*

$$D_j\left(g\circ f\right)_i = \sum D_k g_j\left(f\left(a\right)\right)\cdot D_j f_k\left(a\right)$$

*Proof.* From regular chain rule in single variable,

$$\begin{aligned}\left(g\circ f\right)'\left(a\right) &= \left[D_j\left(g\circ f\right)_i\right]_{l\times n} = g'\left(f\left(a\right)\right)f'\left(a\right)\\ &= \left[D_k g_i\left(f\left(a\right)\right)\right]_{l\times m}\left[D_j f_k\left(a\right)\right]_{m\times n}\\ &= \left[D_k g_i\left(f\left(a\right)\right)\right]_{l\times m}\left[D_j f_k\left(a\right)\right]_{m\times n}\\ &= \left[\sum_{k=1}^{m} D_k g_i\left(f\left(a\right)\right)D_j f_k\left(a\right)\right]_{l\times n}\end{aligned}$$

Notice that the above and $\left[D_j\left(g\circ f\right)_i\left(a\right)\right]_{l\times n}$ are equal entry by entry. ☐

In terms of classical notation, for $x = x\left(s,t\right), y = y\left(s,t\right), z = z\left(s,t\right)$, we have,

$$f\left(x,y,z\right) = f\left(x\left(s,t\right),y\left(s,t\right),z\left(s,t\right)\right)$$

Then,

$$\frac{\partial f}{\partial s} = \frac{\partial f}{\partial x}\times\frac{\partial x}{\partial s}+\frac{\partial f}{\partial y}\times\frac{\partial y}{\partial s}+\frac{\partial f}{\partial z}\times\frac{\partial z}{\partial s}$$

Do note that we are not dividing quantities, rather, we're applying an operator $\frac{\partial}{\partial x}$ to $f$ s.t. $\frac{\partial}{\partial x}\left(f\right) = \frac{\partial f}{\partial x}$.

## 5.6 Differential Curves

A function of the form $\gamma = \mathbb{R}\longrightarrow\mathbb{R}^m$ is called a curve in $\mathbb{R}^m$. The domain of $\gamma$ could also be the same interval in $\mathbb{R}$. As such, $D\gamma_t\colon\mathbb{R}\longrightarrow\mathbb{R}^m$, so, $\gamma'\left(t\right)$ is a $m\times 1$ matrix, i.e., a column vector. Thus,

$$\gamma'\left(t\right)\in\mathbb{R}^m, \gamma = \left(\gamma_1,\cdots,\gamma_m\right), \gamma_i\colon\mathbb{R}\longrightarrow\mathbb{R}$$

and

$$\gamma'\left(t\right) = \begin{bmatrix}\gamma_1'\left(t\right)\\ \gamma_2'\left(t\right)\\ \vdots\\ \gamma_m'\left(t\right)\end{bmatrix}$$

We also have,

$$\gamma\left(t+h\right) = \gamma\left(t\right)+\gamma'\left(t\right)h+o\left(h\right)$$

In $\mathbb{R}^2,\mathbb{R}^3$, especially, $\gamma'\left(t\right)$ is called the velocity vector at time $t$, $\left|\gamma'\left(t\right)\right|$ is called the speed.

**Example 24.** $\gamma\colon\mathbb{R}\longrightarrow\mathbb{R}^2, \gamma\left(t\right) = \left(\cos t,\sin t\right)$. $\gamma$ is called the parametrization of $S^1$ = unit circle in $\mathbb{R}^2$.

$$\gamma'\left(t\right) = \begin{bmatrix}-\sin t\\ \cos t\end{bmatrix}$$

Note $\langle \gamma(t), \gamma'(t) \rangle = 0$, always orthogonal. Thus,

$$|\gamma'(t)| = 1$$

$\gamma$ is at unit speed.

Let's see antoher parametrization of $S$,

$$\beta = \mathbb{R} \longrightarrow \mathbb{R}^2, \beta(t) = \gamma(2t) = (\cos(2t), \sin(2t))$$

$$\beta'(t) = \begin{bmatrix} -2\sin(2t) \\ 2\cos(2t) \end{bmatrix}, |\beta'(t)| = 2 \forall t$$

**Proposition.** Suppose $\gamma \colon \mathbb{R} \longrightarrow \mathbb{R}^m$ differentiable curve such that $\langle \gamma(t), \gamma'(t) \rangle = 0 \forall t \in \mathbb{R}$. Then $|\gamma(t)|$ is constant (as a function of $t$).

*Proof.* Set $f(t) = |\gamma(t)|^2$. Since $|\gamma(t)| \geq 0 \forall t$, it is enough to show that $f(t)$ is constant.

$$|\gamma(t)|^2 = \langle \gamma(t), \gamma'(t) \rangle$$

$$(\gamma, \gamma) : \mathbb{R} \longrightarrow \mathbb{R}^m, (\gamma, \gamma)(t) = (\gamma(t), \gamma(t))$$

Let $g \colon \mathbb{R}^m \times \mathbb{R}^m \longrightarrow \mathbb{R}$ be $g(x, y) = \langle x, y \rangle$. Thus, $f = g \circ (\gamma, \gamma)$. By the chain rule,

$$Df_t = \underbrace{Dg_{(\gamma(t),\gamma(t))}}_{Dg_{(\gamma(t),\gamma(t))}(h,k) = \langle h, \gamma(t) \rangle + \langle \gamma(t), k \rangle} \circ \overbrace{D(\gamma, \gamma)_t}^{D(\gamma(t),\gamma(t))_t h = (\gamma'(t), \gamma'(t))_h}$$

$$Df_t(h) = Dg_{(\gamma(t),\gamma(t))}\left(h\left(\gamma'(t), \gamma'(t)\right)\right) = h\left(2\langle \gamma(t), \gamma'(t) \rangle\right) = 0$$

$Df_t$ is zero transformation $\implies f'(t) = 0 \forall t$. $f \colon \mathbb{R} \longrightarrow \mathbb{R}$ has zero derivative everywhere. Hence, $f$ is a constant function. $\square$

### 5.6.1 Parametrization of Lines

Let $L$ be a line in $\mathbb{R}^n$. If $p, q \in L$, then $v = q - p$ is a vector parallel to $L$ (i.e., $v$ is the direction of $L$).

$$\begin{aligned} \gamma \colon \mathbb{R} \longrightarrow \mathbb{R}^n, \gamma &= p + tv \\ &= p + t(q - p) \\ &= (1 - t)p + tq \end{aligned}$$

Note that if we restrict to $\gamma$ to have domain $[0, 1]$, then $\gamma \colon [0, 1] \to \mathbb{R}^n, \gamma(t) = (1 - t)p + tq$, then this is a parametrization of the line segment starting at $p$ and ending at $q$.

## 5.7 Higher-Order Derivatives

$$\begin{aligned} f\mathbb{R}^n \longrightarrow \mathbb{R},\ & D_1 f \colon \mathbb{R}^n \longrightarrow \mathbb{R} && \text{assuming partial exists} \\ & D_2(D_1 f) : \mathbb{R}^n \longrightarrow \mathbb{R} && \text{second order partial} \\ & D_1(D_2(D_1 f)) : \mathbb{R}^n \longrightarrow \mathbb{R} && \text{third order partial} \end{aligned}$$

The point is that we can iterate partials when they exist.

**Notation.**

$$D_i(D_k f) = D_i D_k f = D_{ki} f$$

The notation suggests to first do $k$, then do $i$.

Let $w = f(x, y, z)$, then, all the following mean the same thing,

$$f_{233}, f_{yzz}, \frac{\partial}{\partial z}\left(\frac{\partial}{\partial z}\left(\frac{\partial}{\partial y}\right)\right), \frac{\partial^3}{\partial z^2 \partial y} f, w_{yzz}, \frac{\partial^3}{\partial z^2 \partial y} w$$

**Definition 22.** $f$ is $\mathcal{C}^k$ if all fo its $k^{\text{th}}$ order partial derivatives exist and are continuous at all points of domain of $g$.

**Remark.** $f$ is $\mathcal{C}^\infty$ (or *smooth*) if partial derivatives of all orders exist (and are continuous).

**Theorem 25.** *$A \subseteq \mathbb{R}^2, f\colon A \longrightarrow \mathbb{R}$ is $\mathcal{C}^2$. Then, at every interior point $a \in A$*

$$D_{12} f(a, b) = D_{21} f(a, b)$$

*Proof. (Sketch).* Since $(a, b) \in A$, there are $h, k$ small s.t.

$$C = [a, a+h] \times [b, b+k] \subseteq A$$

It's clear that $C$ is compact.

$$\begin{aligned}
\int_a^{a+h} \int_b^{b+k} D_{12} f(x, y)\; dy\, dx &= \int_a^{a+h} \int_b^{b+k} D_2 \left(D_1 f(x, y)\right)\, dy\, dx \\
&\overbrace{=}^{FTC} \int_a^{a+h} \left(D_1 f(x, b+x) - D_1 f(x, y)\right)\, dx \\
&\overbrace{=}^{FTC} f(a+h, b+k) - f(a, b+k) - f(a+h, b) + f(a+b) =: \Delta(h, k)
\end{aligned}$$

By Extreme Value Theorem, the following exist on $C$:

$$m_{h,k} = \min\left\{D_{12} f(x, y) :\ (x, y) \in C\right\}$$

or

$$C\colon m_{h,k} \leq D_{12} f(x, y) \leq M_{h,k}$$

where

$$M_{h,k} = \max\left\{D_{12} f(x, y) :\ (x, y) \in C\right\}$$

If we integrate all parts:

$$\int_a^{a+h} \int_b^{b+k} m_{h,k}\, dy\, dx \leq \int_a^{a+h} \int_b^{b+k} D_{12} f(x, y)\; dy\, dx \leq \int_a^{a+h} \int_b^{b+k} M_{h,k}\, dy\, dx$$

$$m_{h,k}(hk) \leq \Delta(h, k) \leq M_{h,k}(hk)$$

$$\Longrightarrow\ m_{h,k}(hk) \leq \frac{\Delta(h, k)}{hk} \leq M_{h,k}(hk)$$

Let $(h, k) \longrightarrow 0$. By continuity,

$$\begin{aligned}
M_{h,k} &\longrightarrow D_{12} f(a, b) \\
m_{h,k} &\longrightarrow D_{21} f(a, b)
\end{aligned}$$

Then, by squeeze,

$$\frac{D(h, k)}{hk} \longrightarrow \Delta_{12} f(a, b)$$

But, we can repeat the same argument with variables reversed to get $\frac{\Delta(h,k)}{hk} \longrightarrow$

$D_{21}f(a,b)$ as $(h,k) \longrightarrow (0,0)$. By uniqueness of limits

$$D_{12}f(a,b) = D_{21}f(a,b)$$

$\square$

**Remark.** For small enough $h, k$, as the square gets smaller and smaller,

$$\int_a^{a+h}\int_b^{b+k} g(x,y)\, dy\, dx \approx g(a,b)\cdot hk$$
$$\int_b^{b+k}\int_a^{a+h} w(x,y)\, dy\, dx \approx w(a,b)\cdot hk$$

## 5.8 Extreme Points for Multivariable Functions

Our goal in this section is to generalize the second derivative test.

Let $A \subseteq \mathbb{R}, f\colon A \longrightarrow \mathbb{R}, f$ is $\mathcal{C}^2$ (twice continuously differentiable) on the interior point of $A$. If $a \in A$ is an interior point of $A$ and $a$ is a critical point of $f$, (i.e., $f'(a) = 0$), then:

(1) If $f''(a) > 0$, then $f(a)$ is a *local minimum* for $f$.
(2) If $f''(a) < 0$, then $f(a)$ is a *local maximum* for $f$.
(3) If $f''(a) = 0$, the second derivative test is *inconclusive*.

### 5.8.1 The Hessian Matrix: Second Order Matrices

$A \subseteq \mathbb{R}^n, f\colon A \longrightarrow \mathbb{R}, a \in A$, $f$ is $\mathcal{C}^2$ on its interior points. $Df_a\colon \mathbb{R}^n \longrightarrow \mathbb{R}$ is a linear transformation and $f'(a)$ is a $1\times n$ matrix where

$$f'(a) = \begin{bmatrix} D_1f(a) & \cdots & D_nf(a)\end{bmatrix}$$

We think about $f'$ as a function from its interior points to $\mathbb{R}^n$ with component functions $D_1f, D_2f, \cdots, D_nf$. Then, $f''(a)$ is a $n\times n$ square matrix,

$$\begin{aligned} f''(a) &= \begin{bmatrix} D_1(D_1f)_a & D_2(D_1f)_a & \cdots & D_n(D_1f)_a \\ D_1(D_2f)_a & D_2(D_2f)_a & \cdots & D_n(D_1f)_a \\ \vdots & \vdots & \ddots & \vdots \\ D_1(D_nf)_a & D_2(D_nf)_a & \cdots & D_n(D_nf)_a \end{bmatrix} \\ &= \begin{bmatrix} D_{11}f(a) & D_{12}f(a) & \cdots & D_{1n}f(a) \\ D_{21}f(a) & D_{22}f(a) & \cdots & D_{2n}f(a) \\ \vdots & \vdots & \ddots & \vdots \\ D_{n1}f(a) & D_{n2}f(a) & \cdots & D_{nn}f(a) \end{bmatrix} \end{aligned}$$

Since $f$ is $\mathcal{C}^2$, equality of mixed partials, so this matrix is symmetric (across its diagonal).

**Definition 23** (Symmetric)**.** A (necessarily) square $n\times n$ matrix $A$ is *symmetric* if $A_{ij} = A_{ji}\ \forall i \leq i, j \leq n$.

We can also express symmetric matrices using *transposes*.

**Definition 24** (Transpose)**.** Let $M$ be $m\times n$ matrix. The *transpose* of $M$ is the $n\times m$ matrix, denoted $M^T$, defined

$$\left(M^T\right)_{ij} = M_{ji}$$

for $1 \leq i \leq n, 1 \leq j \leq m$.

I.e., the 1st row of $M$ is the 1st column of $M^T$, the 2nd row of $M$ is the 2nd column of $M^T$, etc. Note that a matrix $A$ is symmetric iff $A^T = A$.

**Example 25.** (1) The matrix $A = \begin{bmatrix} 1 & 2 \\ 2 & -1 \end{bmatrix}$, $A^T = A$, so $A$ is symmetric.

(2) The matrix $B = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix}$, $B^T = \begin{bmatrix} 1 & 4 \\ 2 & 5 \\ 3 & 6 \end{bmatrix}$

(3) The matrix $x = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}$, $x^T = \begin{bmatrix} 1 & 2 & 3 \end{bmatrix}$

In general, if $x$ is a column vector, $x^T$ is a row vector. For $x, y \in \mathbb{R}^n$ column vectors,

$$\begin{aligned} x^T y &= \begin{bmatrix} x_1 & \cdots & x_n \end{bmatrix} \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \\ &= x_1 y_1 + \cdots + x_n y_n \\ &= \langle x, y \rangle \end{aligned}$$

**Theorem 26.** *$A \subseteq \mathbb{R}^2, f\colon A \longrightarrow \mathbb{R}$, a $\mathcal{C}^2$ function on interior points of $A$. $(a,b) \in A$ interior point and $f'(a,b) = \begin{bmatrix} 0 & 0 \end{bmatrix}$ (i.e. $(a,b)$ is a critical point for $f$.)*

*Let $f''(a,b) = \begin{bmatrix} \alpha & \beta \\ \beta & \delta \end{bmatrix}$, then,*

*(a) If $\alpha > 0$ and $\alpha\delta - \beta^2 > 0$, then $f(a,b)$ is a local minimum.*

*(b) If $\alpha < 0$ and $\alpha\delta - \beta^2 > 0$, then $f(a,b)$ is a local maximum.*

*(c) If $\alpha\delta - \beta^2 < 0$, then $f(a,b)$ is a saddle point.*

*(d) If $\alpha\delta - \beta^2 = 0$, then no conclusion.*

**Example 26.** $f(x,y) = \sin^2 x + x^2 y + y^2$

We have $f_x(x,y) = 2\sin x + 2xy$ and $f_y(x,y) = x^2 + 2y$. We must have critical points satisfying

$$\begin{cases} 2\sin x + 2xy = 0 \\ x^2 + 2y = 0 \end{cases}$$

Note that this is not a system of linear equations. In general, it's hard to find all critical points but $(0,0)$ is one critical point so we classify it.

$$\begin{aligned} f''(x) &= \begin{bmatrix} 2\cos^2 x - 2\sin x + 2y & 2x \\ 2x & 2 \end{bmatrix} \\ f''(0,0) &= \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}. \end{aligned}$$

We note that $\alpha = 2 > 0, \alpha\delta - \beta^2 = 4 > 0$, so $(0,0)$ is a local minimum.

**Example 27.** $f(x,y) = xy(x+y-3)$ . We want to optimize $f$ on

$$T\colon \{(x,y) : x \geq 0, y \geq 0, x + y \leq 3\}$$

We begin by noting that $f$ has maximum and minimum on $T$ by extreme value theorem since $f$ is continuous and $T$ is compact. Also note that $f = 0$ on all 3 boundaries of $T$;

in interiors of $T$, $x > 0, y > 0, x + y < 3 \implies x + y - 3 < 0$. So $f(x, y) < 0$ whenever $(x, y)$ is in interior of $T$.

Thus, max value of $f$ on $T$ is 0 and it is achieved on boundary of $T$. So, we need to find the minimum by using the new theorem (just on interior points).

$$f_x(x, y) = y(2x + y - 3) = 0 \text{ and } f_y(x, y) = x(2y + x - 3) = 0 \iff \begin{cases} 2x + y - 3 = 0 \\ x + 2y - 3 = 0 \end{cases}$$

since $x > 0, y > 0$ in interior of $T$. The above is true $iff x = 1, y = 1$. So $(1, 1)$ is the only critical point in interior of $T$.

$$f''(x, y) = \begin{bmatrix} 2x & 2x + 2y - 3 \\ 2x + 2y - 3 & 2x \end{bmatrix}, f''(1, 1) = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$$

$\alpha = 2 > 0, \alpha\delta - \beta^2 = 3 > 0$, hence $f(1, 1)$ is minimum on interior of $T$.

**Theorem 27** (Critical Point Theorem)**.** *$A \subseteq \mathbb{R}^2, a \in A$ interior point, $f\colon A \longrightarrow \mathbb{R}$. If $f$ is differentiable at $a$ and $f$ has a local min/max at $a$, then $a$ is a critical point for $f$, i.e., $f''(a) =$ zero matrix.*

*Proof. (Local Min/Max Theorem in 2 Variables).* It is enough to show $D_j f(a) = 0$ for $j \in \{1, 2, \cdots, n\}$. Fix $j \in \{1, 2, \cdots, n\}$. Recall $\varphi_j(x) = f(a_1, a_2, \cdots, a_j, x, a_{j+1}, \cdots, a_n)$.

Since $a$ is a local min/max for $f$, it follows that $\varphi_j$ has a local min/max at $a_j$. By single variable calculus, $D_j f(a) = \varphi_j'(a_j) = 0$.

Proof of local mon/max theorem requires a generalization of taylor's theorem. □

### 5.8.2 Quadratic Forms

**Definition 25** (Quadratic Form)**.** Let $A$ be $n \times n$ square matrix. The *quadratic form* induced by $A$ is $\mathcal{Q}_A\colon \mathbb{R}^n \longrightarrow \mathbb{R}, \mathcal{Q}_A(h) = \langle h, Ah \rangle = h^T Ah$.

Note that $\mathcal{Q}_A(th) = \langle th, Ath \rangle = t^2 \langle h, Ah \rangle = t^2 \mathcal{Q}_a(h)$. The most important fact for us if $f\colon \mathbb{R}^2 \longrightarrow \mathbb{R}, \mathcal{Q}_{f''(a)}$, which we will denote by $\mathcal{Q}f_a, \mathcal{Q}f_a(h) = h^T f''(a) h$ (assuming $f''(a)$ exists).

**Theorem 28** (Quadratic Taylor Approximation)**.** *$I \subseteq \mathbb{R}$, open interval, $[0, 1] \subseteq I, \varphi\colon I \longrightarrow \mathbb{R}, \mathcal{C}^2$ function, then,*

$$\varphi(1) = \varphi(0) + \varphi'(0) + \frac{1}{2}\varphi''(c)$$

*for some $c \in [0, 1]$.*

We will assume this and use it to prove a multivariable version.

**Theorem 29.** *$A \subseteq \mathbb{R}^n, f\colon A \longrightarrow \mathbb{R}, \mathcal{C}^2$ function on interior points of $A$. $a \in A$ interior point, then, for small enough $h$,*

$$f(a + ch) = f(a) + Df_a(h) + \frac{1}{2}\mathcal{Q}f_{a+ch}(h)$$

*for some $c \in [0, 1]$.*

The theorem essentially shows that the expression is just an $o(h)$ term but it's a nice way to represent $o(h)$ in terms of a quadratic form.

*Proof.* Fix small enough $h$ so that $A$ contains all the points on line segment from $a$ to $a+h$. Define $\gamma\colon \mathbb{R} \longrightarrow \mathbb{R}^n, \gamma(t) = (1-t)a + t(a+h) = a+th, \gamma(0) = a, \gamma(1) = a+h$.

Define $\gamma\colon \mathbb{R} \longrightarrow \mathbb{R}, \gamma(t) = f(\gamma(t))$. By quadratic taylor in single variable, there is $c \in [0,1]$ s.t.

$$\varphi(1) = \varphi(0) + \varphi'(0) = \frac{1}{2}\varphi''(c)$$

Since

$$\begin{aligned} f(\gamma(1)) &= f(a+ch) \\ f(\gamma(0)) &= f(a) \\ \varphi'(0) = f'(\gamma(0))h &= f'(a)h = Df_a h \end{aligned}$$

We are just left to show that $\varphi''(c) = \mathcal{Q}f_{a+ch}(h)$. Using chain rule and the derivative of the inner product,

$$\begin{aligned} \varphi''(t) &= \langle (f'(\gamma(t)))', h\rangle + \langle f'(\gamma(t)), 0\rangle \\ &= \langle f''(\gamma(t))\gamma'(t), h\rangle = \langle h, f''(\gamma(t))h\rangle = \mathcal{Q}f_{\gamma(t)}(h) \end{aligned}$$

And so,

$$\varphi''(c) = \mathcal{Q}f_{a+ch}(h)$$

□

**Remark.** $f(a+h) = f(a) + Df_a(h) + \frac{1}{2}\mathcal{Q}f_{a+ch}(h)$, a critical point, then $Df_a(h) = 0$. If $a$ critical point point at $\mathcal{Q}f_{a+ch}(h) \geq 0 \forall h$, then $f(a+h) = f(a) + \frac{1}{2}\mathcal{Q}f_{a+ch}(h) \geq f(a) \implies a$ is local minimum.

**Definition 26.** Let $M$ be $n \times n$ symmetric matrix,

(1) $M$ is *positive definit* if $\mathcal{Q}_M(h) > 0$ for every non-zero $h$.
(2) $M$ is *negative definite* if $\mathcal{Q}_M(h) < 0$ for every non-zero $h$.
(3) Otherwise, $M$ is *indefinite* if $\mathcal{Q}_m(h) \geq 0$ for some $h$, $\mathcal{Q}_M(h) < 0$ for some $h$.

**Example 28.** $M = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \mathcal{Q}_M(h,k) = \begin{bmatrix} h & k \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} h \\ k \end{bmatrix} = h^2 + k^2$. $M$ is positive definite.

**Example 29.** $M = \begin{bmatrix} -2 & 0 \\ 0 & 2 \end{bmatrix}, \mathcal{Q}_M(e_1) = e_1^T \begin{bmatrix} -2 & 0 \\ 0 & 2 \end{bmatrix} e_1 = \begin{bmatrix} 1 & 0 \end{bmatrix} \begin{bmatrix} -2 \\ 0 \end{bmatrix} = -2 < 0, \mathcal{Q}_M(e_2) > 0$. $M$ is indefinite.

**Proposition.** Let $M = \begin{bmatrix} \alpha & \beta \\ \beta & \delta \end{bmatrix}$,

(1) $M$ is *positive definite* $\iff \alpha > 0, \alpha\delta - \beta^2 > 0$.
(2) $M$ is *negative definite* $\iff \alpha < 0, \alpha\delta - \beta^2 > 0$.
(3) $M$ is *indefinite* $\iff \alpha\delta - \beta^2 < 0$.

# 6 Directional Derivatives and the Gradient

$A \subseteq \mathbb{R}^n, a \in A$ interior point of $A$, $f\colon A \longrightarrow \mathbb{R}$. Let $j \in \{1, \cdots, n\}$. One definition of the $j^{\text{th}}$ partial derivative of $f$ at $a$ is

$$D_j f(a) = \lim_{t \to 0} \frac{f(a+te_j) - f(a)}{t}$$

$D_j f(a)$ signifies the rate of change of $f$ at $a$ when the input varies in the $e_j$ direction. We want to generalize this defintiion to any direction, not just in the general coordinate axes direction.

By direction, we mean a unit vector $d \in \mathbb{R}^n$, that is,

$$|d| = 1$$

**Definition 27** (Directional Derivative)**.** $A \subseteq \mathbb{R}^n, a \in A$, interior point, $f\colon A \longrightarrow \mathbb{R}, d \in \mathbb{R}^n$ a unit vector. The *directional derivative* of $f$ at $a$ in the direction $d$ is

$$D_d f(a) = \lim_{t \to 0} \frac{f(a+td) - f(a)}{t}$$

assuming this limit exists.

**Remark.** When $d$ is a standard basis vector, i.e. $d = e_j$, we use notation

$$D_j f(a) = De_j f(a)$$

and call it the $j^{\text{th}}$ partial derivative of $f$ at $a$.

Assume now that $f\colon A \longrightarrow \mathbb{R}$ is differentiable at $a$. A special case of the differentiability definition gives us

$$f(a+h) - f(a) - Df_a(h) \qquad \text{is } o(h)$$

Taking $h = td$,

$$f(a+td) - f(a) - Df_a(td) \qquad \text{is } o(td) = o(t)$$

Since $|td| = |t||d| = |t|$. It follows that,

$$\lim_{t \to 0} \frac{f(a+td) - f(a) - tDf_a(d)}{t} = 0$$

$$\lim_{t \to 0} \frac{f(a+td) - f(a)}{t} = Df_a(d) \implies D_d f(a) = Df_a(d)$$

$$\begin{aligned}
\implies D_d f(a) = f'(a)\, d &= \begin{bmatrix} D_1 f(a) & \cdots & D_n f(a) \end{bmatrix} \begin{bmatrix} d_1 \\ \vdots \\ d_n \end{bmatrix} \\
&= \sum_{j=1}^{n} D_j f(a)\, d_j \\
&= \langle \begin{bmatrix} D_1 f(a) \\ \vdots \\ D_n f(a) \end{bmatrix}, \begin{bmatrix} d_1 \\ \vdots \\ d_n \end{bmatrix} \rangle
\end{aligned}$$

**Definition 28** (Gradient)**.** The *gradient* of $f\colon A \longrightarrow \mathbb{R}, A \subseteq \mathbb{R}^n$ at $a$ is

$$\nabla f(a) = f'(a)^T$$

It follows that,

$$\begin{aligned}
D_d f(a) = Df_a(d) = f'(a)\, d &= \begin{bmatrix} D_1 f(a) & \cdots & D_n f(a) \end{bmatrix} \begin{bmatrix} d_1 \\ \vdots \\ d_n \end{bmatrix} \\
&= \sum_{j=1}^{n} D_j f(a)\, d_j \\
&= \langle \begin{bmatrix} D_1 f(a) \\ \vdots \\ D_n f(a) \end{bmatrix}, \begin{bmatrix} d_1 \\ \vdots \\ d_n \end{bmatrix} \rangle \\
&= \langle \nabla f(a), d \rangle
\end{aligned}$$

Still assuming $f\colon A \longrightarrow \mathbb{R}$ is differentiable at $a$, what are the possible values of $D_d f(a)$? (real-valued).

By the Cauchy-Schwarz Inequality,

$$\begin{aligned}
|D_d f(a)| = |\langle \nabla f(a), d \rangle| &\leq |\nabla f(a)| \cdot |d| = |\nabla f(a)| \\
\implies -|\nabla f(a)| &\leq D_d f(a) \leq |\nabla f(a)|
\end{aligned}$$

In fact, $D_d f(a) = |\nabla f(a)|$ when $d$ is parallel to $\nabla f(a)$.

When $\nabla f(a) \neq 0$, set $d = \frac{\nabla f(a)}{|\nabla f(a)|}$ (normalized gradient).

$$\begin{aligned}
D_d f(a) = \langle \nabla f(a), d \rangle &= \langle \nabla f(a), \frac{1}{|\nabla f(a)|} \nabla f(a) \rangle \\
&= \frac{1}{|\nabla f(a)|} \langle \nabla f(a), \nabla f(a) \rangle \\
&= \frac{|\nabla f(a)|^2}{|\nabla f(a)|} = |\nabla f(a)|
\end{aligned}$$

Therefore, large rate of change is achieved when travel in the direction of $\nabla f(a)$. Similar calculation shows that $D_d f(a) = -|\nabla f(a)|$ when

$$d = -\frac{1}{|\nabla f(a)|} \nabla f(a)$$

**Theorem 30.** *$A \subseteq \mathbb{R}^n, a \in A$ interior point of $A, f\colon A \longrightarrow \mathbb{R}, f$ is differentiable at $a$. Then, for every unit vector $d \in \mathbb{R}^n$, directional derivative $D_d f(a)$ exists and*

$$D_d f(a) = f'(a)\, d = \langle \nabla f(a), d \rangle = |\nabla f(a)| \cos \theta_{\nabla f(a), d}$$

Moreover,

$$-|\nabla f(a)| \leq D_d f(a) \leq |\nabla f(a)|$$

with

$$D_d f(a) = |\nabla f(a)|$$

when $d = \frac{\nabla f(a)}{|\nabla f(a)|}$ and

$$D_d f(a) = -|\nabla f(a)|$$

when $d = \frac{-\nabla f(a)}{|\nabla f(a)|}$ whenever $\nabla f(a) \neq 0$.

Furthermore, $\nabla f(a)$ points in the direction of greatest increase ($|\nabla f(a)|$ is the rate of increase) for $f$ at $a$, $-\nabla f(a)$ points in direction of greatest decrease for $fr$ at $a$ with rate $-|\nabla f(a)|$.

In addition, directions orthogonal to $\nabla f(a)$ are directions with rate of change 0 for $f$.

**Remark.** Converse of this theorem is false.

In fact, there is a real value $f$ such that $D_d f(a)$ all exist with $D_d f(a) = \langle \nabla f(a), d\rangle$ for all unit vectors $d$ but $f$ is not differentiable at $a$.

**Example 30.** $f\colon \mathbb{R}^2 \longrightarrow \mathbb{R}$; $f(x,y) = 100 - x^2 - y^2$.

We start at $(0,1)$, direction to ascend quickest

$$\nabla f(x,y) = \begin{bmatrix} D_1 f(x,y) \\ D_2 f(x,y) \end{bmatrix} = \begin{bmatrix} -2x \\ -2y \end{bmatrix}$$

$$\nabla f(0,1) = \begin{bmatrix} 0 \\ -2 \end{bmatrix}, d = \frac{1}{|\nabla f(0,1)|} = \frac{1}{2}\begin{bmatrix} 0 \\ -2 \end{bmatrix} = \begin{bmatrix} 0 \\ -1 \end{bmatrix}$$

Also note that if you travel in directions orthogonal to $\nabla f(0,1)$ (i.e., in $\begin{bmatrix} 1 \\ 0 \end{bmatrix} \begin{bmatrix} -1 \\ 0 \end{bmatrix}$), then, no immediate change in elevation will be noticed.

## 6.1 Level Sets

**Definition 29** (Level Set)**.** Let $A \subseteq \mathbb{R}^n, f\colon A \longrightarrow \mathbb{R}^m, b \in \mathbb{R}^m$. The *level set* of $f$ at $b$ is

$$L_b^f = L_b = \{x \in A\colon f(x) = b\}$$

Recall that $f\colon A \longrightarrow \mathbb{R}, A \subseteq \mathbb{R}^n, \nabla f(a) = f'(a)^T$.

**Example 31.** $f\colon \mathbb{R}^3 \longrightarrow \mathbb{R}, f(x,y,z) = 2x^2 + y^2 + 3z^2$.

graph $(f)$ is a subset of $\mathbb{R}^3 \times \mathbb{R} = \mathbb{R}^4$ (so we can't draw). What do the level sets look like? (Subsets of $\mathbb{R}^3$).

$$b \in \mathbb{R}, L_b^f = \left\{(x,y,z) : 2x^2 + y^2 + 3z^2 = b\right\}$$

For a real-valued $f$ differentiable at $a$, $\nabla f(a)$ is always orthogonal to $L_{f(a)}^f$.

**Theorem 31.** *Let $A \subseteq \mathbb{R}^n, a \in A$ interior point of $A$, $f\colon A \longrightarrow \mathbb{R}$ differentiable at $a$. $\epsilon > 0, \gamma\colon (-\varepsilon, \varepsilon) \longrightarrow L_{f(a)}, \gamma$ differentiable at $a$ and $\gamma(0) = a$. Then, $\gamma'(0)$ is orthogonal to $\nabla f(a)$.*

*Proof.* Since $\gamma(t) \in L_{f(a)} \forall t \in (-\varepsilon, \varepsilon), f \circ \gamma\colon (-\varepsilon, \varepsilon) \longrightarrow \mathbb{R}$ is a constant function $\big(f(\gamma(t)) = f(a) \text{ since } \gamma(t) \in L_{f(a)}\big)$

Moreover, by chain rule, $f \circ \gamma$ is differentiable at 0 and

$$\begin{aligned} 0 = (f \circ \gamma)'(0) &= f'(\gamma(0))\,\gamma'(0) = f'(a)\,\gamma'(0) \\ &= \nabla f(a)^T \gamma'(0) = \langle \nabla f(a), \gamma'(0)\rangle \end{aligned}$$

□

The graph of a function $f\colon A \longrightarrow \mathbb{R}^m$ $(A \subseteq \mathbb{R}^n)$ can always be thought of as a level set of some other function. The idea is that

$$(x,y) \in \text{graph}(f) \iff f(x) = y \iff f(x) - y = 0$$

So the graph $(f)$ is just $L_0^{f(x)-y}$.

Define $g\colon A\times\mathbb{R}^m \longrightarrow \mathbb{R}^m$ by $g(x,y) = f(x) - y$. Then $\text{graph}(f) = L_0^g$.

## 6.2 Tangent Hyperplanes

$A \subseteq \mathbb{R}^n, a \in A$ interior point of $A, f\colon A \longrightarrow \mathbb{R}, f$ is differentiable at $a$. We want an equation that describes the hyperplane tangent to graph$(f)$ at $(a, f(a))$.

Define $g\colon A\times\mathbb{R} \longrightarrow \mathbb{R}$, $g(x,y) = f(x) - y$ so $\text{graph}(f) = L_0^g$. Thus, $\nabla g(a, f(a))$ is normal to $L_0^g = \text{graph}(f)$ at $(a, f(a))$. Since $x \in \mathbb{R}^n, y \in \mathbb{R}$,

$$\nabla g(x,y) = \begin{bmatrix} D_1 g(x,y) \\ \vdots \\ D_n g(x,y) \\ D_{n+1} g(x,y) \end{bmatrix} = \begin{bmatrix} \frac{\partial f}{\partial x_1}(x) \\ \vdots \\ \frac{\partial f}{\partial x_n}(x) \\ -1 \end{bmatrix} = \begin{bmatrix} \nabla f(x) \\ -1 \end{bmatrix}$$

$$\nabla g(a, f(a)) = \begin{bmatrix} \nabla f(a) \\ -1 \end{bmatrix}$$

The vectors on the plane tangent to graph$(g)$ at $(a, f(a))$ are exactly the vectors so that when they are translated by $-(a, f(a))$, they become orthogonal to

$$\begin{bmatrix} \nabla f(a) \\ -1 \end{bmatrix}$$

i.e., the vectors $(x,y)$ which satisfy

$$\begin{aligned} &\langle (x,y) - (a, f(a)), \begin{bmatrix} \nabla f(a) \\ -1 \end{bmatrix} \rangle = 0 \\ \iff\ & D_1 f(a)(x_1 - a_1) + \cdots + D_n f(a)(x_n - a_n) - (y - f(a)) = 0 \\ \iff\ & y = f(a) + D_1 f(a)(x_1 - a_1) + \cdots + D_n f(a)(x_n - a_n) \end{aligned}$$

**Example 32.** $f(x,y) = e^x - y$. Find tangent plane for $f$ at $(0, 1, f(0,1))$.

$$D_1 f(x,y) = e^x, D_2 f(x,y) = -1$$

So,

$$\nabla f(0,1) = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$$

$$\begin{aligned} z &= f(0,1) + D_1 f(0,1)(x - 0) + D_2 f(0,1)(y - 1) \\ &= x - (y-1) \iff x - y - z = -1 \end{aligned}$$

# 7 Inverse & Implicit Functions

## 7.1 The Inverse Function Theorem

**Definition 30** (Open Ball)**.** A subset $A$ is open (where $A \subseteq \mathbb{R}^n$) if every $a \in A$ is an interior point of $A$, i.e., for every $a \in A$, there is $\varepsilon > 0$ s.t. $\mathcal{B}(a, \varepsilon) \subseteq A$.

**Definition 31** (Inverse Function Theorem)**.** Let $A \subseteq \mathbb{R}^n$ be open, let $f\colon A \longrightarrow \mathbb{R}^n$ be $\mathcal{C}^1$, let $a \in A$, and assume $f'(a)$ is an invertible matrix.

Then, there exists an open set $V \subseteq A$ with $a \in V$ and an open set $W \subseteq \mathbb{R}^n$ with

$f(a) \in W$ such that $f \colon V \longrightarrow W$ has a $\mathcal{C}^1$ inverse $f^{-1} \colon W \longrightarrow V$. Moreover, for each $x \in V$, $y = f(x) \in W$, we have

$$D\left(f^{-1}\right)_y = (Df_x)^{-1}, \text{ i.e., } \left(f^{-1}\right)'(y) = \left(f'(x)\right)^{-1}$$

**Remark.** An *affine function* is a function $g$ of the form $g(x) = c + T(x)$ where $c$ is a vector and $T$ is linear.

**Theorem 32.** *Given $f$, $a$ as in the inverse function theorem, $b = f(a)$, then,*

$$\begin{aligned} f^{-1}(b+h) &= f^{-1}(b) + D\left(f^{-1}\right)_b(h) \\ &= a + (Df_a)^{-1}(h) \end{aligned}$$

*is the best affine approximation to $f^{-1}$ near $b = f(a)$.*

**Remark.** We will not prove the Inverse Function Theorem. But the last claim follows easily from what's before it.

$$f^{-1} \circ f \colon V \longrightarrow V, f^{-1} \circ f = \mathrm{id}_v \implies D\left(f^{-1} \circ f\right)_a = D(\mathrm{id}_v)_x = \mathrm{id}_{\mathbb{R}^n}$$
$$\mathrm{id}_{\mathbb{R}^n} = D\left(f^{-1} \circ f\right)_x = D\left(f^{-1}\right)_{f(x)} = Df_x$$

Similarly,

$$\mathrm{id}_{\mathbb{R}_n} = Df_x \circ \left(Df^{-1}\right)_{f(x)}$$

Then,

$$D\left(f^{-1}\right)_{f(x)} = (Df_x)^{-1}$$

A quick note on notation. $f \colon V \longrightarrow W$ and $f^{-1} \colon W \longrightarrow V$ is misleading because $f \colon A \longrightarrow \mathbb{R}^n$ need not be invertible. But we write $f^{-1}$ for a local inverse. A better notation is $f|_V \colon V \longrightarrow W, (f|_V)^{-1} \colon W \longrightarrow V$ ($f$ restricted to $V$).

**Example 33.** $f \colon \mathbb{R}^2 \longrightarrow \mathbb{R}^2$ ($2 \times 2$ Jacobians), $f(x,y) = (e^x \cos y, e^x \sin y)$. Where is $f$ locally invertible?

$f$ is locally invertible wherever $f^{-1}(x,y)$ is invertible? (By the theorem, wherever the matrix is invertible).

$$f'(x,y) = \begin{bmatrix} e^x \cos y & -e^x \sin y \\ e^x \sin y & e^x \cos y \end{bmatrix}$$
$$\det\left(f^{-1}(x,y)\right) = e^{2x}\cos^2 y + e^{2x}\sin^2 y = e^2 x > 0$$

for every $(x,y)$. It follows that $f$ is locally invertible at every $(x,y) \in \mathbb{R}^2$. But $f(0,0) = (1,0) = f(0,2\pi)$ so $f$ is not one-to-one, hence, $f \colon \mathbb{R}^2 \longrightarrow \mathbb{R}^2$ is not globally invertible.

### 7.1.1 $2 \times 2$ Invertible Matrices

A matrix $A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$ is invertifle $\iff \det(A) = ad - bc \neq 0$ and

$$A^{-1} = \begin{bmatrix} a & b \\ c & d \end{bmatrix}^{-1} = \frac{1}{\det(A)} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$$

From the example, find the best affine approximation to $f^{-1}$ at $f\left(0, \frac{\pi}{4}\right) = \left(\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2}\right)$.

$$\begin{aligned} D\left(f^{-1}\right)_{\left(\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2}\right)} &= \left(Df_{\left(0, \frac{\pi}{4}\right)}\right)^{-1} = \left(f^{-1}\right)'\left(\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2}\right) = \left(f'\left(0, \frac{\pi}{4}\right)\right)^{-1} \\ &= \begin{bmatrix} \frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \end{bmatrix} = \begin{bmatrix} \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \\ -\frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \end{bmatrix} = \frac{\sqrt{2}}{2}\begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix} \end{aligned}$$

So, the best affine approximation is

$$f^{-1}\left(\left(\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2}\right) + (h,k)\right) = \begin{bmatrix} 0 \\ \frac{\pi}{4} \end{bmatrix} + \frac{\sqrt{2}}{2}\begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix}\begin{bmatrix} h \\ k \end{bmatrix}$$

## 7.2 The Implicit Function Theorem

Before introducing the Implicit Function Theorem, we want to understand what does it mean to *implicitly* define something? Consider $x + 2y = 1$, we think of this as a constraint equation on $(x, y)$. We can implicitly define $y$ in terms of $x$ by $y = \frac{1-x}{2}$. The constraint also defines $x$ as a function of $y$.

Alternately, consider the equation of a unit circle, $x^2 + y^2 = 1$. We can't globally define all $y$ as a function of $x$ but we can define it locally as so,

$$\begin{aligned} y &= \pm\sqrt{1 - x^2} \\ x &= \pm\sqrt{1 - y^2} \end{aligned}$$

**Definition 32** (Implicit Function Theorem)**.** Let $0 < c < n$ be integers and set $r = n - c$. Let $A \subseteq \mathbb{R}^n$ be open, and let $g\colon A \longrightarrow \mathbb{R}^c$ be $\mathcal{C}^1$. Let $L = \{v \in A \colon g(v) = \mathbf{0}_c\}$ . Let $p \in L$ and write $p = (a, b) \in \mathbb{R}^r \times \mathbb{R}^c$ and

$$g'(p) = \begin{bmatrix} g'_{(x)}(p) & g'_{(y)}(p) \end{bmatrix} \qquad \left(g'_{(x)}(p) \text{ is } c \times r \text{ and } g'_y(p) \text{ is } c \times c\right)$$

If $g'_y(p)$ is invertible, then $L$ is locally the graph of a function near $p$; more precisely, there is an open $B \subseteq \mathbb{R}^n$ with $p = (a, b) \in B$ and a function $\varphi$ such that

$$g(x, y) = \mathbf{0}_c \iff y = \varphi(x)\,\forall\,(x, y) \in B$$

Moreover, the function $\varphi$ is differentiable at $a$ and

$$\varphi'(a) = -\left(g'_y(p)\right)^{-1}\left(g'_x(p)\right)$$

**Example 34.** $g\colon \mathbb{R}^2 \longrightarrow \mathbb{R}, g(x, y) = x^2 + y^2 - 1, g(x, y) = 0 \iff x^2 + y^2 = 1$.

$$g'(x, y) = \begin{bmatrix} 2x & 2y \end{bmatrix}$$

$$g'(0, 1) = \begin{bmatrix} 0 & 2 \end{bmatrix}$$

The submatrix $\begin{bmatrix} 2 \end{bmatrix}$ is invertible and so, $y = \varphi(x) = \sqrt{1 - x^2}$ local to $(0, 1)$.

Putting the $y$ (dependent) variables at the end of the variable list is just a matter of convenience of stating the theorem. We can always change the order of variables without changing properties of the function.

$$g(v) = \underbrace{0_c}_{c \text{ constraint equations, } n = r + c \text{ variables}}$$

Given $p \in L$, we want to find any $c$ variables which can be written as functionsof the other $r$ variables. You can do this for any $c$ variables whose corresponding columns in

$g'(p)$ form an invertible $c \times c$ matrix.

**Example 35.** $x^2 + y^2 + z^2 - 1 = 0 \iff x^2 + y^2 + z^2 = 1$

$$g(x, y, z) = x^2 + y^2 + z^2 - 1, L_0^g = \{(x, y, z) : g(x, y, z) = 0\}$$

$$g'(x, y, z) = \begin{bmatrix} 2x & 2y & 2z \end{bmatrix}$$

For $p = (0, 01)$,

$$g'(p) = \begin{bmatrix} 0 & 0 & 2 \end{bmatrix}$$

We see that near $(0, 0, 1)$ on $L$, $z$ can be written as a function of $x, y$, that is,

$$z = \sqrt{1 - x^2 - y^2}$$

For $p = (0, -1, 0)$,

$$g'(p) = \begin{bmatrix} 0 & -2 & - \end{bmatrix}$$

We see that near $(0, -1, 0)$ on $L$, $y$ can be written as a function of $x, z$, that is

$$y = -\sqrt{1 - x^2 - z^2}$$

For $p = \left(\frac{1}{\sqrt{3}}, \frac{1}{\sqrt{3}}, \frac{1}{\sqrt{3}}\right)$,

$$g'(p) = \begin{bmatrix} \frac{2}{\sqrt{3}} & \frac{2}{\sqrt{3}} & \frac{2}{\sqrt{3}} \end{bmatrix}$$

We see that near $\left(\frac{1}{\sqrt{3}}, \frac{1}{\sqrt{3}}, \frac{1}{\sqrt{3}}\right)$, $x, y, z$ can be written as functions of the other two valuesariables since all three submatrices are invertible.

**Example 36.** $g(x, y, z) = \left(x^2 + y^2 + z^2 - 3, y + z - 2\right)$

$g(x, y, z) = (0, 0)$ defined intersection of a sphere and a plane. For $p = (1, 1, 1), g(p) = (0, 0)$.

$$g'(x, y, z) = \begin{bmatrix} 2x & 2y & 2z \\ 0 & 1 & 1 \end{bmatrix}, g'(1, 1, 1) = \begin{bmatrix} 2 & 2 & 2 \\ 0 & 1 & 1 \end{bmatrix}$$

We see that $x, y$ can be written as a function of $z$,

$$\begin{cases} x^2 + y^2 + z^2 = 3 \\ y + z = 2 \end{cases}$$

And so,

$$x = \sqrt{3 - y^2 - z^2} \implies x = \sqrt{3 - (2 - z)^2 - z^2}$$
$$y = 2 - z$$

*Proof. (Implicit Function Theorem $\implies$ Inverse Function Theorem).* We want to invert $f \colon A \longrightarrow \mathbb{R}^n$ (where $A \subseteq \mathbb{R}^n$) (local to some point), that is, solve $f(x) = y$ for $x$ in terms of $y$. Notice that $f(x) = y$ is a constraint equation so apply the implicit function theorem such that $g(x, y) = f(x) - y$. Let $f, g \in A$ be as in statement of the Implicit Function Theorem. Define $g \colon A \times \mathbb{R}^n \longrightarrow \mathbb{R}^n$ by $g(x, y) = f(x) - y$. Note $L = L_0^g = \operatorname{graph}(f)$. Set $b = f(a), g(a, b) = f(a) - b = 0$.

$$\begin{aligned} g'(a, b) &= \begin{bmatrix} g'_{(x)}(a, b) & g'_{(y)}(a, b) \end{bmatrix} \\ &= \begin{bmatrix} \underbrace{f'(a)}_{\text{invertible } n \times n block} & -I_n \end{bmatrix} \end{aligned}$$

By Implicit Function Theorem, there is an open $\mathcal{B}$ with $(a, b) \in \mathcal{B}$ and a function $\varphi$ s.t.

$$f(x) = y \iff \varphi(y) = x \forall (x, y) \in \mathcal{B}$$

In other words, $\varphi = f^{-1}$ on some suitable domain. □

*(Proof. Inverse Function Theorem $\implies$ Implicit Function Theorem).* We want to solve $g(x,y) = 0$ for $y$ variable in terms of $x$ variables. Consider $f(x,y) = (x, g(x,y))$. Suppose we can invert $f$ locally, then $g$ on any $x$

$$(x,y) \longmapsto (x,y) \text{ s.t. } g(x,y) = 0$$

$$f'(x,y) = \begin{bmatrix} I_r & 0 \\ g'_{(x)}(x,y) & g'_{(y)}(x,y) \end{bmatrix}$$

By Inverse Function Theorem, we can locally invert $f$ since the matrix is invertible. □

## 7.3 Lagrange Multipliers

To satisfy constraints is to meet a condition

$$g(x) = 0_c \begin{cases} g_1(x) = 0 \\ \vdots \\ g_c(x) = 0 \end{cases}$$

$$L = L_0^g = L_0^{g_1} \cap L_0^{g_2} \cap \cdots \cap L_0^{g_c}$$

At $p \in L$, $\nabla g_i(p)$ is normal to $L$. We also note that the linear combinations $\sum_{i=1}^{c} b_i \nabla g_i(p)$ are also normal to $L$. Do such vectors describe all vectors normal to $L$? It depends. If the Jacobian matrix $\nabla g(p)$ has *full rank*. We will cover this later.

When optimizing $f$ on $L$, we look at critical points $p$ for $f|_L$ ($f$ restricted to $L$).

If $d$ is a direction on $L$, then, $D_d f(p) = 0$. Assuming differentiability at $p$,

$$\begin{aligned} 0 &= D_d f(p) \\ 0 &= \langle d, \nabla f(p) \rangle \forall d \text{ tangent to } L \\ \implies & \nabla f(p) \perp d \\ \implies & \nabla f(p) \text{ is normal to } L \end{aligned}$$

Thus, $\nabla f(p)$ is a linear combination of $\nabla g_i f(p)$ assuming they represent all normal vectors to $L$. This gives us the Lagrange conditions for optimizer $p$,

$$\begin{cases} \nabla f(p) = \lambda_1 \nabla g_1(p) + \cdots + \lambda_c \nabla g_c(p) \\ g(p) = 0 \end{cases}$$

These provide a necessary condition for optimizer points assuming nice enough properties of $f, g$ at $p$. We can now solve optimization problems.

**Example 37.** Find the point on the intersection of $x+y+z = 1$ and $2x - y + z = -1$, which is closest to the origin?

We rewrite $\min f(x,y,z) = x^2 + y^2 + z^2$ subject to the constraint $g(x,y,z) = (g_1, g_2) = (x+y+z-1, 2x-y+z+1) = (0,0)$. Then, the Lagrange conditions are

$$\begin{cases} \nabla f(x,y,z) = \lambda_1 \nabla g_1(x,y,z) + \lambda_2 \nabla g_2(x,y,z) \\ x+y+z = 1 \\ 2x - y + z = -1 \end{cases}$$

Thus,

$$\nabla f(x,y,z) = \begin{bmatrix} 2x \\ 2y \\ 2z \end{bmatrix}, \nabla g_1(x,y,z) = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}, \nabla g_2(x,y,z) = \begin{bmatrix} 2 \\ -1 \\ 1 \end{bmatrix}$$

Hence,

$$\begin{bmatrix} 2x \\ 2y \\ 2z \end{bmatrix} = \lambda_1 \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} + \lambda_2 \begin{bmatrix} 2 \\ -1 \\ 1 \end{bmatrix}$$

The 2 gets absorbed into constants,

$$\begin{bmatrix} x \\ y \\ z \end{bmatrix} = \lambda_1 \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} + \lambda_2 \begin{bmatrix} 2 \\ -1 \\ 1 \end{bmatrix}$$

### 7.3.1 More on Lagrange Multipliers

**Definition 33** (Full Rank)**.** Let $A$ be a $c \times n, c \leq n$. $A$ has *full rank* if there are columns which form an invertible matrix.

**Example 38.** The matrix $\begin{bmatrix} 1 & 2 & 0 \\ 0 & 0 & 3 \end{bmatrix}$ has columns that form an invertible matrix, so, it has full rank.

**Example 39.** The matrix $\begin{bmatrix} 1 & 0 & 2 \\ 0 & 0 & 0 \end{bmatrix}$ doesn't have columns that form an invertible matrix, so, it's not full rank.

**Theorem 33** (Lagrange Multiplier Criterion)**.** *Let $0 < c < n$ be integers, $A \subseteq \mathbb{R}^n, g\colon A \longrightarrow \mathbb{R}^c$ be $\mathbb{C}^1$ on all interior points of $A$.*

*Let $L = L_0^g = \{x \in A\colon g(x) = c\}$. Let $f\colon A \longrightarrow \mathbb{R}$. Suppose the restriction of $f$ to $L$ has an extreme value at $p \in L$, where $p$ in interior of $A$. Suppose $f$ is differentiable at $p$ and $g'(p)$ has full rank. Then, the following conditions hold:*

$$\begin{cases} \nabla f(p) = \sum_{i=1}^{c} \lambda_i g_i(p) \\ g(p) = 0 \end{cases}$$

**Definition 34** (Regular Value)**.** 0 is a *regular value* of $f$ if $L_0^g$ contains only interior points of $A$ and $g'(p)$ has full rank for all $p \in L_0^g$.

**Corollary.** In the notation of the theorem, if $f$ is differentiable at all points $p \in L$ and if 0 is a regular value of $g$, then any $p \in L$ which has an extreme value for $f$ restricted to $L$ must satisfy the Lagrange conditions.

**Remark.** So, such points are the only candidates to solve min/max for $f(x)$ subject to $g(x) = 0$.

**Example 40.** Find point on unit circle $x^2 + y^2 = 1$ closest to the line $x + y = 2$.

We change to nomize $\min f(x,y,v,w) = (x-v)^2 + (y-w)^2$ subject to $x^2 + y^2 - 1 = 0, v + w - 2 = 0$.

$$g'(x,y,v,w) = \begin{bmatrix} 2x & 2y & 0 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix}$$

Note, as long as $(x,y) \neq (0,0)$, $g'(x,y,v,w)$ has full rank $\implies$ 0 is a regular value of $g$. ($(0,0)$ not on unit circle.)

By the corollary, we didn't miss any optimizers. Only looked at points satisgying Lagrange conditions.

**Example 41.** Minimize $f(x,y) = y$, subject to $g(x,y) = y^3 - x^4 = 0$.

Note that $(x,y) \in L = L_0^g \implies y^3 = x^4 > 0 \implies y \geq 0 \implies f(x,y) \geq 0$ on $L$. Also $(0,0) \in L$ so $f(0,0) = 0$ solves the miimization problem.

Trivial.

However, what's important to note here is that the Lagrange conditions, in fact, miss this solution. See below,

$$\begin{cases} \nabla g = \lambda \nabla g \\ y^3 - x^4 = 0 \end{cases} \implies \begin{cases} \begin{bmatrix} 0 \\ 1 \end{bmatrix} = \lambda \begin{bmatrix} -4x^3 \\ 3y^2 \end{bmatrix} \\ y^3 - x^4 = 0 \end{cases}$$

We note that $\begin{bmatrix} 0 \\ 1 \end{bmatrix} = \lambda \begin{bmatrix} -4x^2 \\ 3y^2 \end{bmatrix}$ implies $x = 0, y \neq 0, \lambda \neq 0$ however, $y^3 - x^4 = 0$ implies that $x = 0 \implies y = 0$. There is a contradiction in the value of $y$.

The Lagrange Conditions don't work since $(0,0) \in L$ but $g'(0,0) = \begin{bmatrix} 0 & 0 \end{bmatrix}$ which is not full rank so 0 is not a regular value of $g$ and so, the Lagrange conditions miss the solution.

**Example 42.** $\max f(x_1, \cdots, x_n) = (x_1 \cdots x_n)^{1/n}$ subject to $\dfrac{x_1 + \cdots + x_n}{n} = 1$ and $x_1, \cdots, x_n > 0$.

$g(x_1, \cdots, x_n) = \dfrac{x_1 + \cdots + x_n}{n} - 1$ so,

$$g'(x_1, \cdots, x_n) = \begin{bmatrix} \frac{1}{n} & \frac{1}{n} & \cdots & \frac{1}{n} \end{bmatrix} = \frac{1}{n} \begin{bmatrix} 1 & \cdots & 1 \end{bmatrix}$$

Note that 0 is a regular value so we only need to look at Lagrange conditions.

$$\begin{aligned} \nabla f(x_1, \cdots, x_n) &= \left( \frac{1}{n} (x_1 \cdots x_n)^{1/n-1} x_2 \cdots x_n, \cdots, \frac{1}{n} (x_1 \cdots x_n)^{1/n} x_1 \cdots x_{n-1} \right) \\ &= \left( \frac{f(x_1, \cdots, x_n)}{nx_1}, \cdots, \frac{f(x_1, \cdots, x_n)}{nx_n} \right) = \frac{f(x_1, \cdots, x_n)}{n} \left( \frac{1}{x_1}, \cdots, \frac{1}{x_n} \right) \end{aligned}$$

$$\begin{aligned} &\nabla f = \lambda \nabla g \\ &\implies \frac{f(x_1, \cdots, x_n)}{nx_i} = \frac{1}{n} \text{ for } i \in [1,n] \\ &\implies f(x_1, \cdots, x_n) = x_i \text{ for } i \in [1,n] \implies x_1 = x_2 = \cdots = x_n \end{aligned}$$

$$\begin{cases} x_1 = \cdots = x_n \\ \dfrac{x_1 + \cdots + x_n}{n} = 1 \end{cases}$$

All $x_i > 0 \implies x_1 = x_2 = \cdots x_n = 1$, then $f$ has max value of 1 at $(1, \cdots, 1)$.

**Corollary.** (Arithmetic-Geometric Mean Inequality) For all positive $a_1, \cdots, a_n$, $(a_1 \cdots a_n)^{1/n} \leq \frac{a_1 + \cdots + a_n}{n}$.

*Proof.* Let $a_1, \cdots, a_n > 0$. Set $a = \frac{a_1 + \cdots + a_n}{n}$. Set $x_i = \frac{a_i}{a}$ for $i = 1, \cdots, n$. Then,

$$\frac{x_1 + \cdots + x_n}{n} = \frac{\frac{a_1}{a} + \cdots + \frac{a_n}{a}}{n} = \frac{1}{a} \left( \frac{a_1 + \cdots + a_n}{n} \right) = 1$$

By previous example,

$$(x_1 \cdots x_n)^{1/n} \leq 1$$

Then,

$$\begin{aligned}(a_1 \cdots a_n)^{1/n} &= ((ax_1) \cdots (ax_n))^{1/n} = a\,(x_1 \cdots x_n)^{1/n} \\ &\leq a \cdot 1 = \frac{a_1 + \cdots + a_n}{n}\end{aligned}$$

□

## 7.4 On Determinants and Cross Product

$$\det \begin{bmatrix} a \end{bmatrix} = a$$

$$\det \begin{bmatrix} a & c \\ b & d \end{bmatrix} = ad - bc$$

We can define $n \times n$ determinant recursively. Let the matrix $A$,

$$A = \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nn} \end{bmatrix}$$

$$\det A = \sum_{i=1}^{n} (-1)^{i+1} a_{i1} \det A+^{(i1)}$$

Essentially cofactor expansion along 1st column. We can do cofactor expansion aong any row or column.

$$\begin{bmatrix} + & - & + \\ - & + & - \\ + & - & + \end{bmatrix}$$

**Theorem 34.** *For any square matrix $A$, $A$ is invertible* $\iff \det A \neq 0$.

**Definition 35** (Cross Product)**.** The cross product is a special operation on $\mathbb{R}^3$ and is defined as

$$\times : \mathbb{R}^3 \times \mathbb{R}^3 \longrightarrow \mathbb{R}^3$$

$$(a_1, a_2, a_3) \times (b_1, b_2, b_3) = \det \begin{bmatrix} a_1 & b_1 & e_1 \\ a_2 & b_2 & e_3 \\ a_3 & b_3 & e_3 \end{bmatrix}$$

Note that the abobe matrix is not a real matrix because some entries are vectors. It follows that,

$$(a_1, a_2, a_3) \times (b_1, b_2, b_3) = (a_2 b_3 - a_3 b_2, a_3 b_1 - a_1 b_3, a_1 b_2 - a_2 b_1)$$

## 7.5 Properties of Cross Product

(1) $\times$ is skew-symmetric: $(u \times v) = -(v \times u)$ (anticommutative)

(2) Bilinearity: $(au + a'u') \times v = a\,(u \times v) + a'\,(u' \times v)$ and $u \times (bv + b'v') = b\,(u \times v) + b'\,(u \times v')$

(3) $u \times v$ is orthogonal to both $u$ and $v$.

(4) $\times$ is 0 $\iff$ $u, v$ are parallel.

(5) $|u \times v|$ = area of the parallelogram spanned by $u$ and $v$.

(6) If $u, v$ not parallel, $[u, v, u \times v]$ is right-hand oriented. $(u \times v)$ is normal to the plane spanned by $u$ and $v$.