MATHEMATICS, MULTIVARIATE CALCULUS, AND LINEAR ALGEBRA

ISE – 364/464

DEPT. OF INDUSTRIAL & SYSTEMS ENGINEERING

GRIFFIN DEAN KENT

LEHIGH UNIVERSITY.

MATHEMATICS

Mathematics

Mathematics is the field of study primarily concerned with the rigorous description and manipulation of abstract objects that consist of either abstractions derived from nature or purely abstract objects that are stipulated to have certain properties, known as axioms. Further, mathematics is ultimately a collection of formal systems that, at the least, provide an extension of classical logic to include numbers and provide a framework to derive new theorems.

THE FOUNDATIONS OF MATHEMATICS

SETS

THE FUNDAMENTAL OBJECT OF MATHEMATICS

Sets

A set $\mathcal{S}$ is an unordered collection of distinct objects, which are called elements. If x is an element of a set $\mathcal{S}$, we write $x \in \mathcal{S}$. This is read as "x in $\mathcal{S}$". Set builder notation is looks like this:

$$\mathcal{S} \coloneqq \{\text{elements} : \text{conditions used to generate the elements}\}.$$

The ":" that precedes the conditions is typically read as "such that".

Classic Sets of Numbers

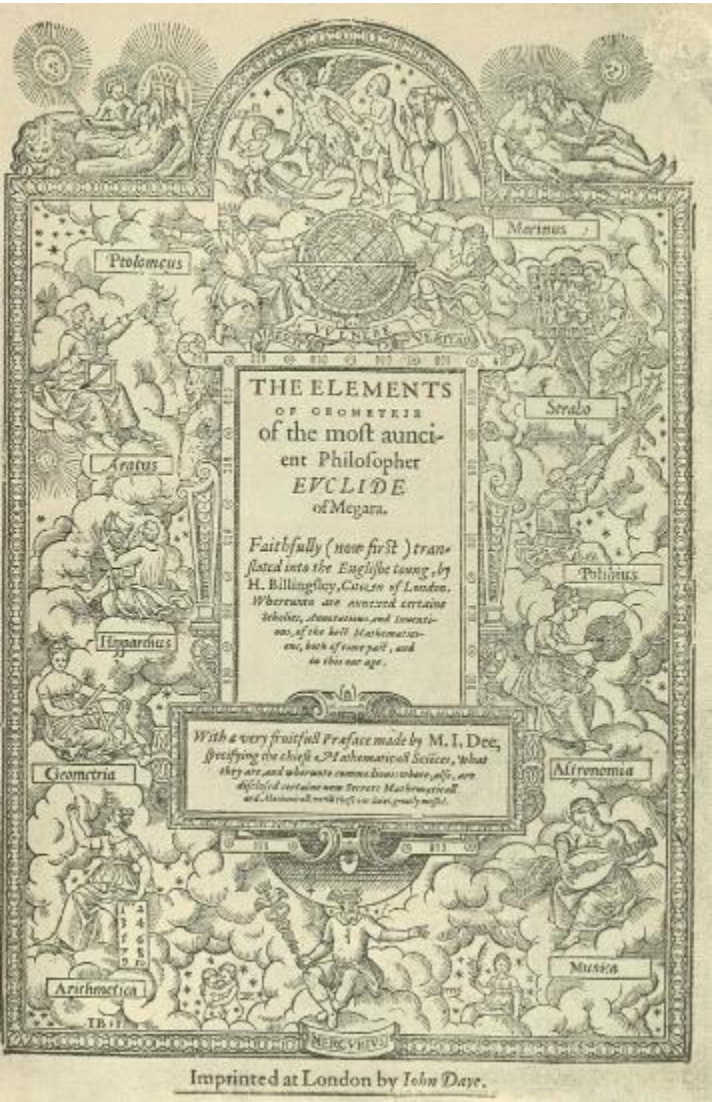
- The set of **Natural Numbers**: $\mathbb{N} = \{0,1,2,3,4,5, \dots\}$.
- The set of **Integers**: $\mathbb{Z} = \{\dots, -3, -2, -1, 0, 1, 2, 3, \dots\}$.
- The set of **Rational Numbers**: $\mathbb{Q} \coloneqq \left\{\frac{p}{q} : (p, q) \in \mathbb{Z} \times \mathbb{Z}, q \neq 0\right\}$.
- The set of **Real Numbers**: Denoted by $\mathbb{R}$ and contains all real (rational and irrational) numbers.

Some examples

- $\{1,4,9,16,25, \dots\} = \{x : x^2 \in \mathbb{N}\}$.
- $\{G, D, K\} = \{i : i \text{ is an initial of the instructor for ISE } -364/464\}$.
- $\{\text{Florida, New Mexico, Idaho, Utah, Pennsylvania, } \dots\} = \{s : s \text{ is a state in the USA}\}$.

EUCLID'S ELEMENTS

THE FIRST AXIOMATIC TREATMENT OF GEOMETRY

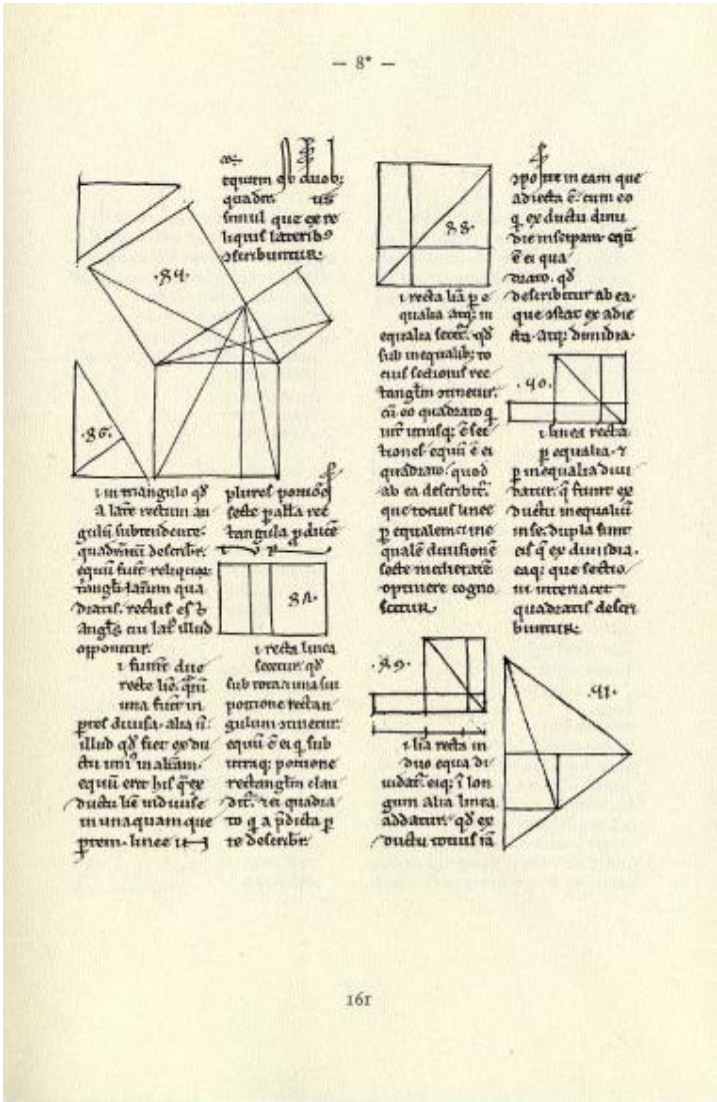


Around 300 BC, the Greek mathematician **Euclid** wrote one of the most influential works in the history of mathematics: ***The Elements***.

The Elements is a comprehensive compilation of the knowledge of geometry ad number theory at that time. Further, it was used to **teach math for 2,000 years**.

It was also the **first** work that utilized and laid the groundwork for the **axiomatization of mathematics**.

- In The Elements, Euclid starts by providing the definitions of a point, a line, and a circle.
- Next, he provides 5 postulates (axioms) relating to geometry.
- Lastly, he provides 5 other axioms (common notions) that are not necessarily specific to geometry.



THE PEANO AXIOMS:

DEFINING THE NATURAL NUMBERS

The **Peano Axioms** are a series of foundational assumptions that are used as a starting place to build out the natural numbers $\mathbb{N} = \{0,1,2,3, \dots\}$.

The first axiom establishes a "**starting point**".

- **Axiom 1)** 0 is a natural number.

The next four axioms establish the properties of the **equality relation**.

- **Axiom 2)** (Reflexive) For every natural number x, $x = x$.
- **Axiom 3)** (Symmetric) For all natural numbers x and y, if $x = y$, then $y = x$.
- **Axiom 4)** (Transitive) For all natural numbers x, y, and z, if $x = y$ and $y = z$, then $x = z$.
- **Axiom 5)** (Closure) For all x and y, if y is a natural number and $x = y$, then x is also a natural number.

The next three axioms define the **arithmetical properties** of natural numbers in terms of a "successor" function S.

- **Axiom 6)** (Closure) For every natural number x, $S(x)$ is a natural number.
- **Axiom 7)** (Injection) For all natural numbers x and y, if $S(x) = S(y)$, then $y = x$.
- **Axiom 8)** (No natural number less than 0) For every natural number x, $S(x) = 0$ is false.

The last axiom establishes the notion that each natural number can be obtain by repeatedly applying the successor function and is known as the axiom of induction.

- **Axiom 9)** (Induction) If K is a set such that $0 \in K$ and has the property that for every natural number x

THE FIELD AXIOMS:

DEFINING THE REAL NUMBERS

Although the Peano axioms can derive the natural numbers, they cannot alone derive the real numbers.

In **real analysis** (and most of applied mathematics), when it comes to deriving the set of real numbers $\mathbb{R}$, we typically utilize a mathematical construct known as a **field**. Specifically, a **field** is a **set** that **satisfies certain axiomatic properties**.

Fields

A field is a nonempty set $\mathbb{F}$, along with two binary operations, addition (+) and multiplication (·), that satisfies the following axioms:

- **Axiom 1)** (Commutativity) If $(x, y) \in \mathbb{F} \times \mathbb{F}$, then $x + y = y + x$ and $x \cdot y = y \cdot x$.
- **Axiom 2)** (Distributivity) If $(x, y, z) \in \mathbb{F} \times \mathbb{F} \times \mathbb{F}$, then $x \cdot (y + z) = x \cdot y + x \cdot z$.
- **Axiom 3)** (Associativity) If $(x, y, z) \in \mathbb{F} \times \mathbb{F} \times \mathbb{F}$, then $(x + y) + z = x + (y + z)$ and $(x \cdot y) \cdot z = x \cdot (y \cdot z)$.
- **Axiom 4)** (Identity) There are two special elements $(0,1) \in \mathbb{F} \times \mathbb{F}$, where $x + 0 = x$ and $x \cdot 1 = x$, for all $x \in \mathbb{F}$.
- **Axiom 5)** (Inverse) For all $x \in \mathbb{F}$, there is an element $-x \in \mathbb{F}$ such that $x + (-x) = 0$. If $x \neq 0$, then there is also an element $x^{-1} \in \mathbb{F}$ such that $x \cdot x^{-1} = 1$.

Notice that the natural numbers $\mathbb{N}$ don't form a field because they fail the first half of axiom 4 and all of axiom 5. However, this is still not quite enough to define the real numbers.

ORDERED FIELD AXIOMS:

DEFINING THE REAL NUMBERS

Although the Peano axioms can derive the natural numbers, they cannot alone derive the real numbers.

Ordered Fields

An Ordered field is a field $\mathbb{F}$, along with additional axiom:

- **Axiom 6)** (Order) There is a nonempty subset $P \subseteq \mathbb{F}$, called the positive elements, such that
  - If $(x, y) \in P \times P$, then $x + y \in P$ and $x \cdot y \in P$.
  - If $x \in \mathbb{F}$ and $x \neq 0$, then either $x \in P$ or $-x \in P$, but not both.

Notice that the rational numbers $\mathbb{Q} \coloneqq \left\{\frac{p}{q} : (p, q) \in \mathbb{Z} \times \mathbb{Z}, q \neq 0\right\}$ satisfy all of these requirements. However, we need one last axiom to derive the real numbers.

Completeness Axiom

- **Axiom 7)** (Completeness) Let S be an ordered field. Then, for every $A \subseteq S$, $\sup(A) \in S$. Such a set S is called complete.

Finally, all these axioms taken together are enough to fully derive the real numbers $\mathbb{R}$. Actually, there is an interesting result that can be proven relating to the real numbers.

(Theorem) Existence and Uniqueness of $\mathbb{R}$

There exists a unique complete ordered field. We call this field the real numbers $\mathbb{R}$.

CANTOR'S SET THEORY

(NAÏVE SET THEORY)

During the 1870s – 1880s, the mathematician **Georg Cantor** introduced the concept of a set and was the first to begin developing the topic of **Set Theory** (as such, he is referred to as ``The Father of Set Theory'').

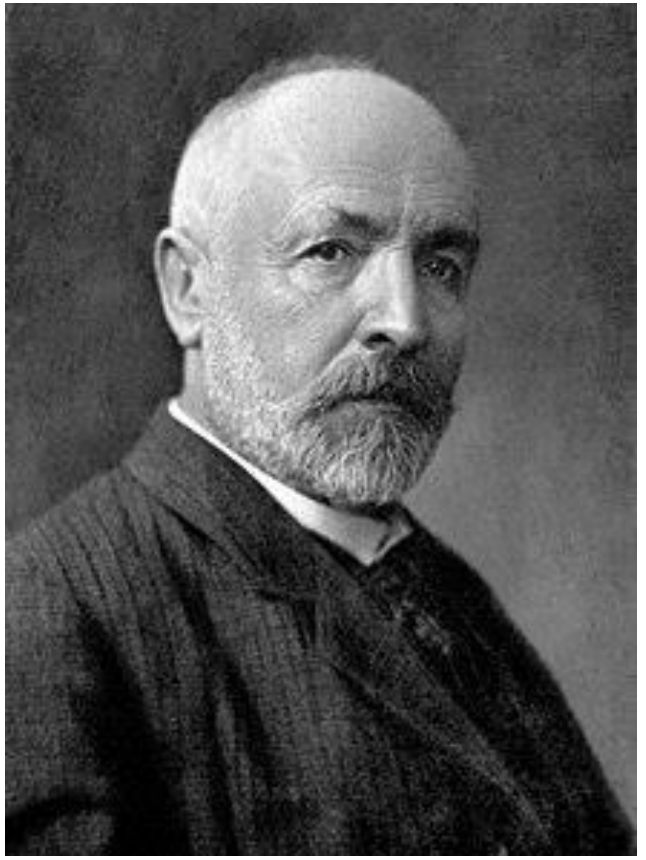
Cantor's set theory was proposed in more of an ``intuitive'' and ``informal'' sense, rather than stated a series of definite axioms that defined the properties of sets. However, his original formulation of set theory can be stated as essentially the combination of two axioms that cement his original intuitions.

Cantor's Set Theory

A set is a mathematical construct that depends on nothing more than its members. Such sets as defined by Cantor satisfy two axioms.

- **Axiom 1)** (Extensionality) Given two sets A and B. Then, $A = B$ if and only if, for all x, $x \in A$ if and only if $x \in B$.
- **Axiom 2)** (Comprehension) For any condition C, there exists a set A such that, for all x, $x \in A$ if and only if x satisfies C.

Cantor's set theory, although intuitive, is not entirely consistent as a variety of paradoxes arises out of this formulation; specifically, these paradoxes come about from Axiom 2 (which needs some restrictions on the conditions C).



Georg Cantor (1845 – 1918)

RUSSEL'S PARADOX

MATHEMATICS IN CRISIS

In 1901, the logician **Bertrand Russell** discovered a famous paradox in Cantor's set theory, which has fittingly been dubbed ``**Russell's Paradox**''.

Russell's Paradox

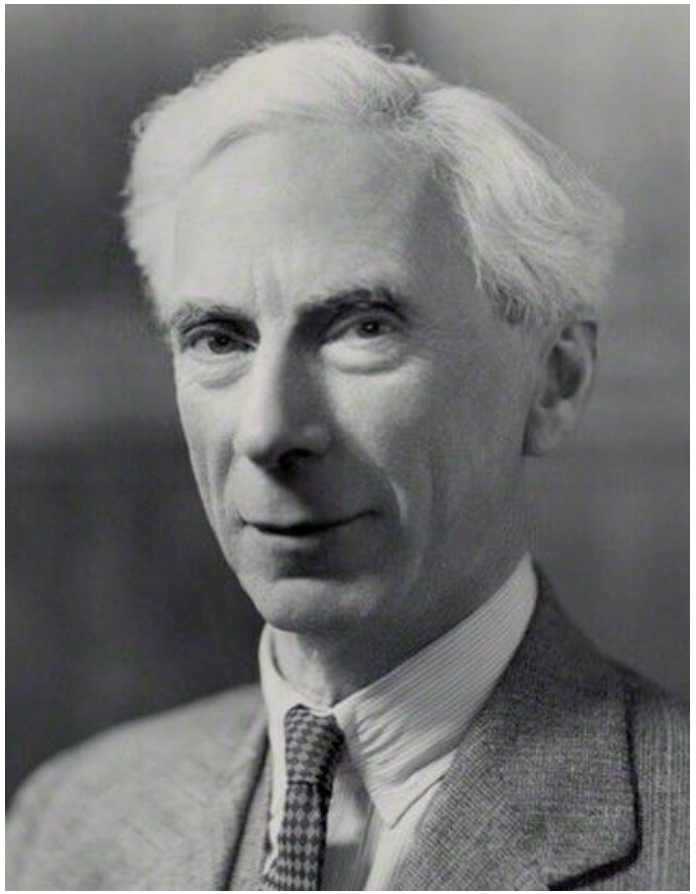
Consider the set R of all sets that are not members of themselves

(This is referred to as Russell's Set).

- Thus, if R is a set, and if R is a member of R, then it must not be a member of R by the definition of R.
- However, if R is not a member of R, then by the definition of R it must be a member of R.
- Clearly, this is a contradiction. Hence, Russell's Set illustrates a gap in Cantor's original formulation of set theory (later dubbed naïve set theory).

A classic illustration of this paradox is to consider the following example.

- There is a single barber in a secluded town who cuts the hair of all in the town who don't cut their own hair.
- Thus, since they are the only barber, they must cut their own hair.
- However, if they do cut their hair, then they don't cut their hair, because they only cut the hair of those who don't cut their own hair…
- This is a paradox, and hence, such a barber cannot exist.



Bertrand Russell (1872 – 1970)

HILBERT'S PROGRAM:

THE QUEST TO GROUND ALL OF MATHEMATICS

In 1902, the mathematician **David Hilbert** proposed a list of problems in mathematics that he deemed were the most important problems facing the field at the time. This list was referred to as ``**Hilbert's 23 Problems**'', with the second problem of ``**Demonstrate the consistency of the arithmetical axioms**''.

Further, Hilbert eventually lead a movement, termed ``**Hilbert's Program**'', in the early 1920s that was a proposed solution to the foundational crisis of mathematics during a time when the foundations of mathematics was found to suffer from paradoxes and inconsistencies.

Hilbert's Program

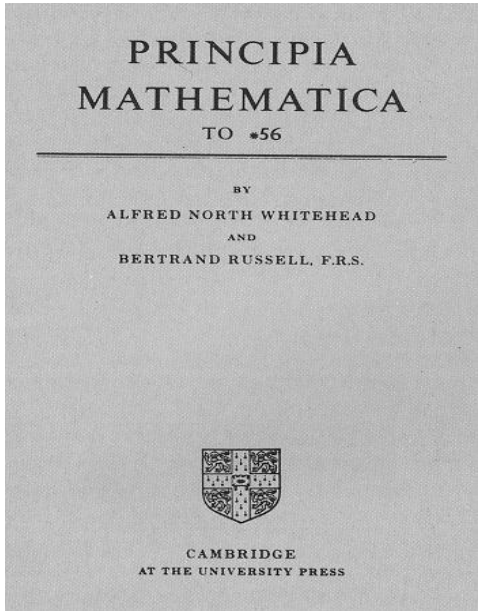
Goal: to secure the foundation of mathematics. This included:

- A precise formal language for all of mathematics.
- Proof of completeness.
- Proof of consistency.
- Proof of conservation.
- Proof of decidability.

David Hilbert (1862 – 1943)

THE PRINCIPIA MATHEMATICA

RUSSEL & WHITEHEAD'S ATTEMPT TO UNIFY MATHEMATICS AND LOGIC


PRINCIPIA MATHEMATICA
TO *56
BY
ALFRED NORTH WHITEHEAD
AND
BERTRAND RUSSELL, F.R.S.
CAMBRIDGE
AT THE UNIVERSITY PRESS

In 1910, 1912, and 1913, **Bertrand Russell** and co-author **Alfred North Whitehead** published in three volumes the monumental and landmark work in mathematical logic and philosophy collectively known as the **Principia Mathematica** (PM).

- The primary goal of PM was to demonstrate that all mathematics could be derived from a small set of logical axioms, which would also avoid paradoxes that had plagued set theory.
- This desire to **reduce arithmetic to logic** was motivated by the authors belief in Logicism. However, this goal ultimately proved to be unsuccessful, as we will see.

Logicism: The belief in the philosophical doctrine that mathematics is an extension of logic and that all mathematical truths are logical truths.

***54·43.** $\vdash :. \alpha, \beta \in 1 . \supset : \alpha \cap \beta = \Lambda . \equiv . \alpha \cup \beta \in 2$

Dem.

$\vdash . *54{\cdot}26 . \supset \vdash :. \alpha = \iota'x . \beta = \iota'y . \supset : \alpha \cup \beta \in 2 . \equiv . x \neq y .$

[*51·231] $\equiv . \iota'x \cap \iota'y = \Lambda .$

[*13·12] $\equiv . \alpha \cap \beta = \Lambda$ (1)

$\vdash . (1) . *11{\cdot}11{\cdot}35 . \supset$

$\vdash :. (\exists x, y) . \alpha = \iota'x . \beta = \iota'y . \supset : \alpha \cup \beta \in 2 . \equiv . \alpha \cap \beta = \Lambda$ (2)

$\vdash . (2) . *11{\cdot}54 . *52{\cdot}1 . \supset \vdash . \text{Prop}$

From this proposition it will follow, when arithmetical addition has been defined, that $1 + 1 = 2$.

Theorem in PM stating that ``$1 + 1 = 2$'' (350+ pages in) with the side note: ``The above proposition is occasionally useful''.

It should be mentioned that, aside from providing a paradox-free formulation of arithmetic, PM was hugely influential in the fields of mathematics and philosophy and marked a turning point in the fields. Its inception influenced great debates about the nature of mathematics, logic, and truth, as well as establishing a new standard of rigor in analytical reasoning.

PRINCIPIA MATHEMATICA

FAILURES, SHORTCOMINGS, & CRITICISMS

Although the **Principia Mathematica** of Russell and Whitehead was hugely influential and altered the landscape, it had several shortcomings…

- **Complexity and Practicality** – PM was extremely complex and cumbersome, and the difficulty in proving even basic arithmetic results highlighted this impracticality. However, this alone is not enough to disregard the work.

- **Ad Hoc & Restrictive** – PM used something called ``**the theory of types**'' to avoid paradoxes, which was criticized as being a bit arbitrary, ad hoc, and restrictive.

- **Logical & Philosophical Criticisms** – Some underlying assumptions as well as the acceptance of certain unintuitive axioms were criticized.

MODERN AXIOMATIC SET THEORY

(THE FOUNDATIONS OF MATHEMATICS)

Although PM was a valiant attempt at grounding all of mathematics, it was ultimately not widely adopted for the reasons already mentioned. However, this was remedied soon after as another system took its place.

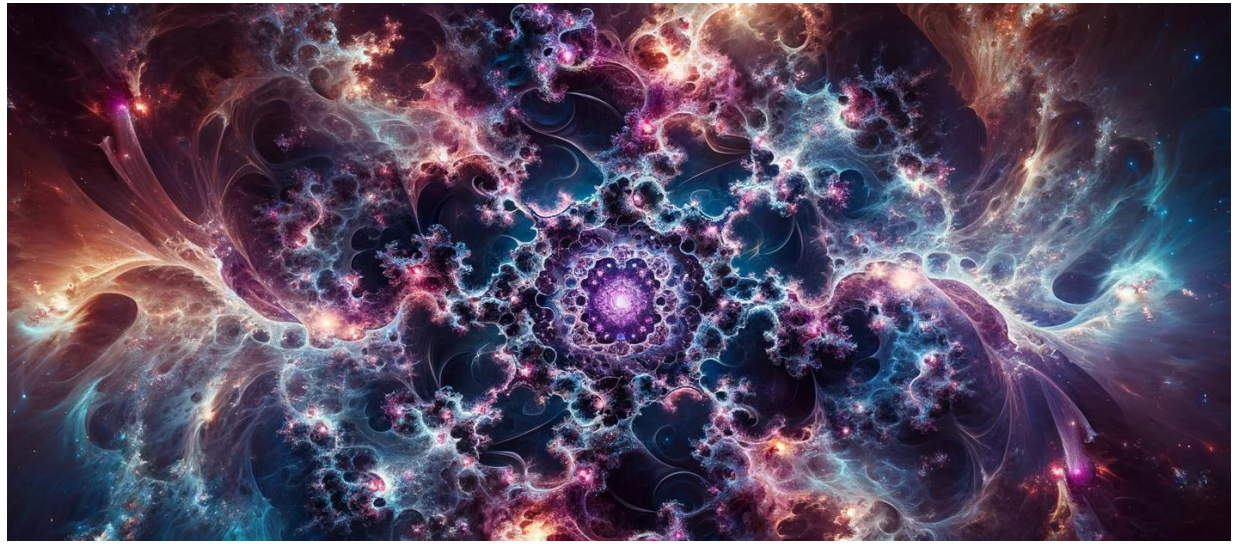
Building off Cantor's original axioms, a series of mathematicians modified and contributed additions to what has now become known as the field of **Axiomatic Set Theory**. Such contributions were made from the mid 1910s to the early 1930s from mathematicians **Ernst Zermelo**, **Abraham Fraenkel**, and **Thoralf Skolem**.

The resulting ZFC Formulation of Set Theory emerged as a solution to the problems that naïve set theory faced and became the **widely accepted foundation for ``all'' of mathematics**. In fact, ZFC is the most comprehensive foundational axiomatic system of mathematics and is one of the **greatest mathematical achievements**.

The Zermelo-Fraenkel+Choice (ZFC) Axioms

The ZFC Axioms of Set Theory are **10 Axioms** that provide a basis for essentially all of mathematics free of paradoxes. The names of these axioms are as follows.

1. Axiom of Extensionality
2. Axiom of the Empty Set
3. Axiom of Pairing
4. Axiom of Union
5. Axiom of the Power Set
6. Axiom of Infinity
7. Axiom of Replacement
8. Axiom of Specification
9. Axiom of Regularity
10. Axiom of Choice



GÖDEL

THE INCOMPLETENESS OF MATHEMATICS

In the early 1930s, the logician **Kurt Gödel** (dubbed as one of the greatest logicians of all time) published two theorems that would forever alter the position of mathematics: the **incompleteness theorems**.

Gödel's 1st Incompleteness Theorem: Any axiomatic system of logic that is consistent and is of a sufficient complexity to express the basic arithmetic of the natural numbers (number theory) is incomplete.

Gödel's 2nd Incompleteness Theorem: Any axiomatic system of logic that is of a sufficient complexity to express the basic arithmetic of the natural numbers (number theory) cannot demonstrate its own consistency.

Kurt Gödel (1906 – 1978)

At a high level, in his proof, Gödel utilized a type of encoding that assigns a unique number (a **Gödel number**; specifically, a prime number) to every ``component'' of a mathematical proposition. Thus, any statement (theorem) in mathematics can be turned into a number, and its truth value can be determined via a numerical procedure. In this way, Gödel uses mathematics to talk about mathematical statements in a very meta fashion. Finally, consider the proposition:

``This statement cannot be proven from the axioms of this system.''

Since this statement can be represented in a numerical fashion, it is a mathematical statement and has a truth value. However, the ramifications of such a statement have far-reaching implications…

PHILOSOPHY OF MATHEMATICS

A BRIEF OVERVIEW

Where is mathematics left in the wake of Gödel's Theorems?

- First, Gödel's theorems definitively put **an end to the belief of logicism**, proving Russel wrong and upending much of Hilbert's program.
- As opposed to the laws of classical logic, any axiomatic system that is sophisticated enough to derive the natural numbers (i.e., any system of **mathematics**, including PM and ZFC) **is incomplete**. That is, there will always be true mathematical statements which are true simply by virtue of the axiomatic system itself, but which can never be proven.
- **Mathematics is not able to prove its own consistency**… This is again another major difference that separates mathematics from classical logic.

How should one ultimately view mathematics?

- One should first understand where the incompleteness of mathematics really comes from… It is a result of the fact that **ANY** system that can specify numbers **WILL** inherently become **self-referential**, i.e., a type of meta-mathematical **loop** will form.
- Ultimately, we do have to assume as a new axiom that the system of mathematics (ZFC) we are utilizing is consistent.
- The next natural question becomes: **can we justify** this assumption?
- **Most** (highly) **likely**!
- Just look at the axioms of these mathematical systems (they are meant to be as self-evident as possible). Remember, the issue is that they don't prove enough truths, not that they are riddled with contradictions.

THE AXIOM OF CONSISTENCY

THE FINAL PIECE

Remember, if a system is inconsistent (it allows for contradictions), then one would be able to prove ALL possible statements (this follows from the inference rule of **Negation Elimination**, i.e., **The Principle of Explosion** – states that any proposition can be inferred from a contradiction). As is stands, there do not ``seem'' to be any inconsistencies within mathematics (this is the safest version of saying that mathematics has absolutely no contradictions). Thus, it seems reasonable to implement the axiom of consistency regarding mathematics.

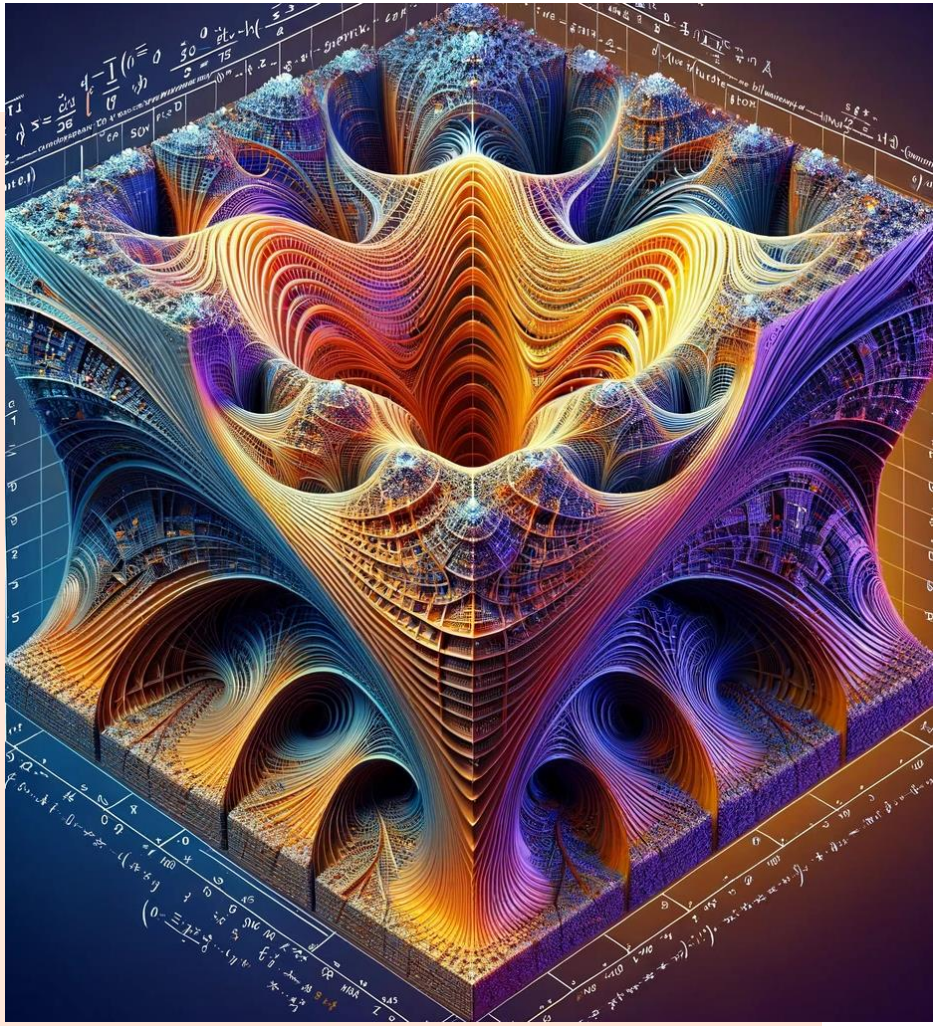
Let's expand upon this. What we are doing when we accept the belief that mathematics is consistent is the same as stating that **we believe that the system is correct**.

- Mathematics is built upon classical logic, which we know is sound.
- Thus, if the new added axioms (such as those of Peano or ZFC) are true, then the system WILL be consistent.
- This is typically why axioms are required to be so ``basic'' and almost self-evident that they are difficult to deny. Further, to question the validity of such axioms would *almost* be on par with questioning one's own sanity.

Conclusion

Accepting the truth of mathematics is *nearly* on the same level as accepting the laws of logic (in the sense that it is seemingly impossible to deny it and live consistently in that belief).

- However, this is the reason why mathematics falls below Logic in the Hierarchy of Truth.



LINEAR ALGEBRA

VECTORS

Vector:

- A quantity that has both magnitude and direction.
- Serves as a useful way of defining quantities in higher dimensions.
- An n-dimensional vector x is essentially an array that contains n elements, and whose i-th element is denoted by x_i.
- If all the elements of a vector belong to the same set of numbers (e.g., $\mathbb{N}$, $\mathbb{Z}$, or $\mathbb{R}$), then we would indicate the dimension of the vector by including an exponent on the set, i.e., the set of n-dimensional real vectors is denoted as $\mathbb{R}^n$.
- By default, we typically denote vectors in "column" form, i.e.,

$$x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \in \mathbb{R}^{n \times 1}.$$

- Conversely, we denote the transpose of vectors in "row form", i.e.,

$$x^T = [x_1 \quad x_2 \quad \cdots \quad x_n] \in \mathbb{R}^{1 \times n}.$$

VECTOR OPERATIONS

Vector Addition

$$x = \begin{bmatrix} 1 \\ 3 \end{bmatrix}$$

$$y = \begin{bmatrix} 5 \\ 2 \end{bmatrix}$$

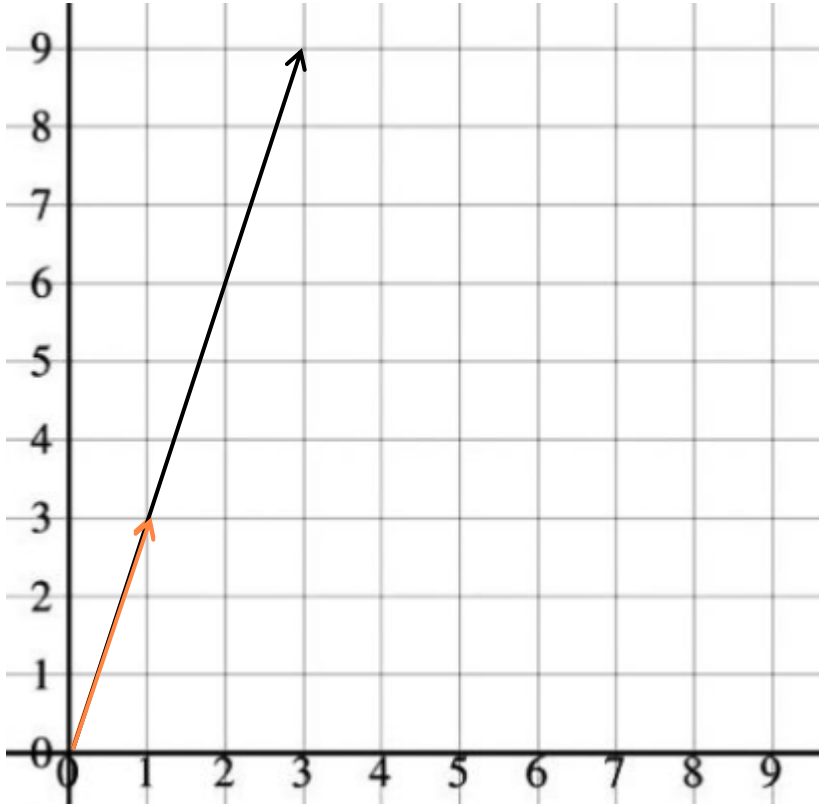
$$x + y = \begin{bmatrix} 6 \\ 5 \end{bmatrix}$$

Vector-Scalar Multiplication

$$x = \begin{bmatrix} 1 \\ 3 \end{bmatrix}$$

$$\alpha = 3$$

$$\alpha x = \begin{bmatrix} 3 \\ 9 \end{bmatrix}$$



INNER (DOT) PRODUCTS

Dot Products (Algebraic)

Given two vectors of the same dimension $x \in \mathbb{R}^n$ and $y \in \mathbb{R}^n$, the dot-product of the two vectors is defined by:

$$x^T y = \sum_{i=1}^{n} x_i y_i \,.$$

Dot Products (Geometric)

Similarly, the dot-product is also equivalent to the expression:

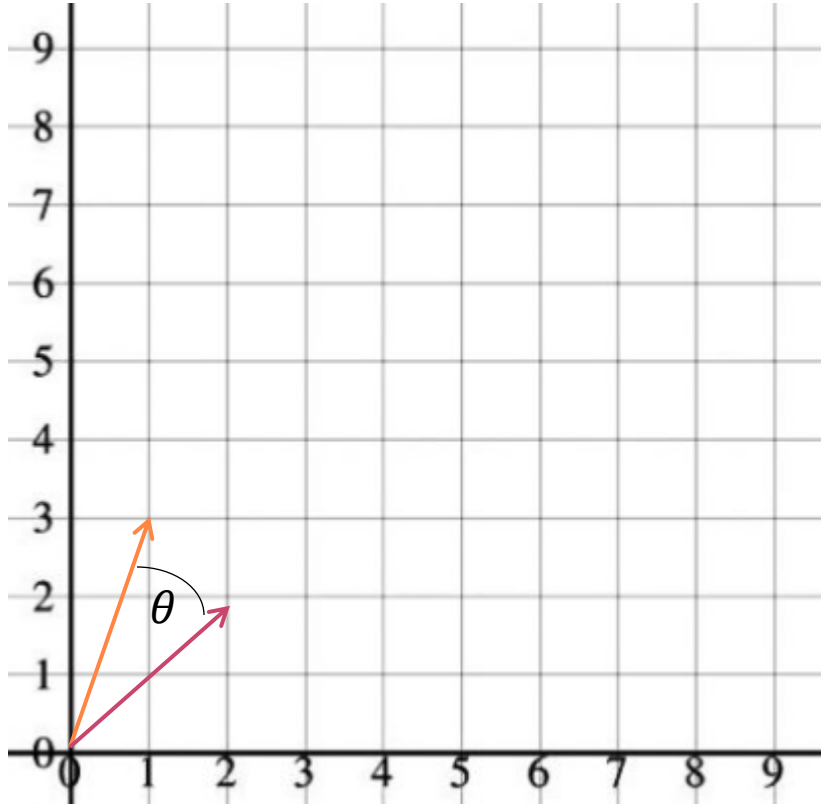
$$x^T y = \|x\| \cdot \|y\| \cos\theta \,,$$

where θ is defined as the angle between the two vectors.

$$x = \begin{bmatrix} 1 \\ 3 \end{bmatrix}$$

$$y = \begin{bmatrix} 2 \\ 2 \end{bmatrix}$$

$$x^T y = 1 \cdot 2 + 3 \cdot 2 = 8$$



VECTOR NORMS (MAGNITUDE OF A VECTOR)

Norms

A norm $\|\cdot\|: \mathbb{R}^n \to \mathbb{R}$ is a real-valued function such that the four following properties are met: Non-negativity, Definiteness, Homogeneity, and Sub-additivity.
The most common form of norms are the ℓ_p-norms (with $p \geq 0$), where for each $x \in \mathbb{R}^n$, we have

$$\|x\|_p := \left(\sum_{i=1}^{n} |x_i|^p\right)^{1/p}.$$

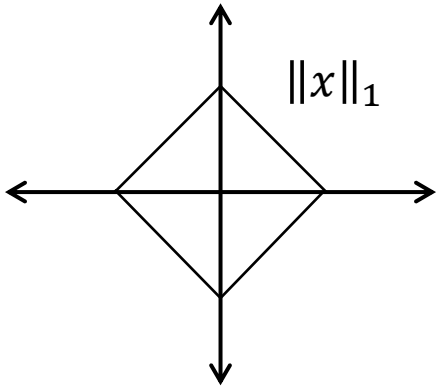
The most well-known norms:

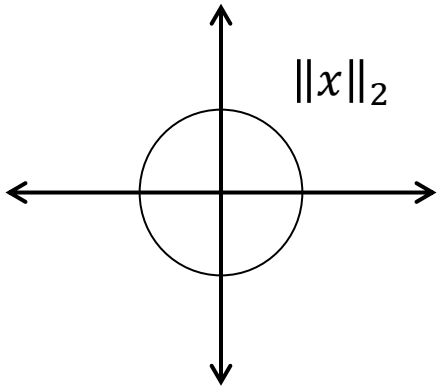
$$\|x\|_1 := \sum_{i=1}^{n} |x_i|, \qquad \|x\|_2 := \left(\sum_{i=1}^{n} |x_i|^2\right)^{\frac{1}{2}} = \sqrt{x^T x}, \qquad \|x\|_\infty := \max_{i \in \{1,2,\ldots,n\}} |x_i|.$$

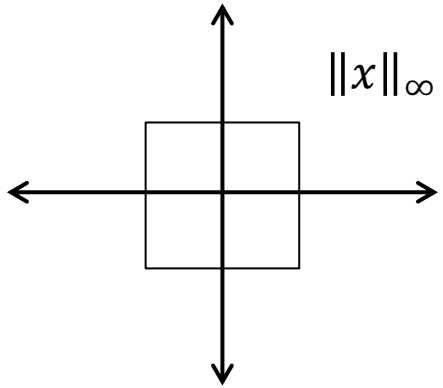
- The norm of a vector defines the magnitude (or length) of the vector.

NORM BALLS:

DEFINING DISTANCE IN DIFFERENT SPACES







Norm Ball

Given a vector $\hat{x} \in \mathbb{R}^n$, we define the set of vectors within the "ball" with a radius of $\epsilon > 0$ centered at the vector $\hat{x}$ as

$$\mathcal{B}(\hat{x}, \epsilon) \coloneqq \{x \in \mathbb{R}^n : \|\hat{x} - x\| \leq \epsilon\},$$

where $\|\cdot\|$ is any norm.

HYPERPLANES

One of the most important concepts that many others are built upon (in linear algebra, machine learning and optimization) is that of a **hyperplane**.

Hyperplane

Let $a \in \mathbb{R}^n$ such that a $\neq 0$ and let $b \in \mathbb{R}$. Given a vector of variables $x \in \mathbb{R}^n$, then in an $\mathbb{R}^{n+1}$ ambient space, a **hyperplane** is a ``flat'' affine set in $\mathbb{R}^n$ (one dimension lower than the ambient space) which splits the ambient space into two disjoint subsets (each called **half-spaces**) and can mathematically be defined by the set of all points that satisfy the equation

$$a^T x = b.$$

A hyperplane $\mathcal{H} \subseteq \mathbb{R}^n$ can also be defined by the set of points that encompass it:

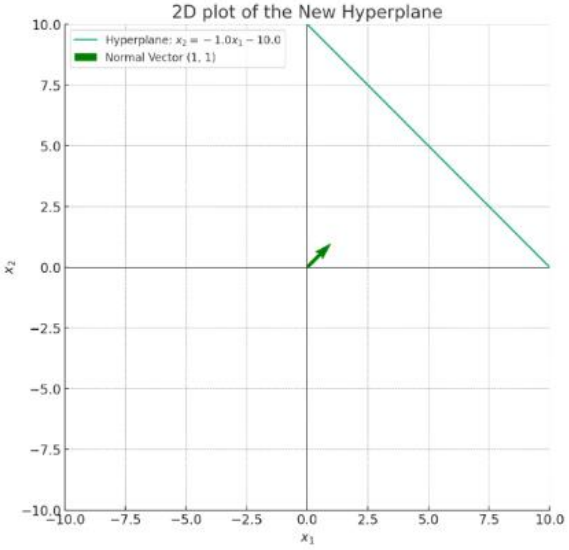
$$\mathcal{H} := \{x \in \mathbb{R}^n : a^T x = b\}.$$

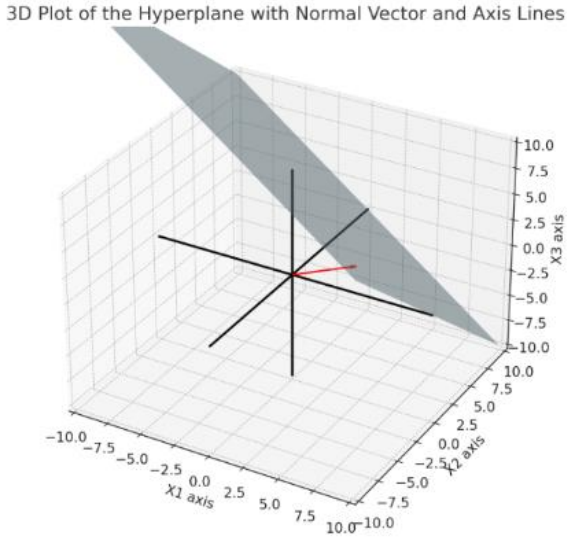
A Note on interpreting Hyperplanes

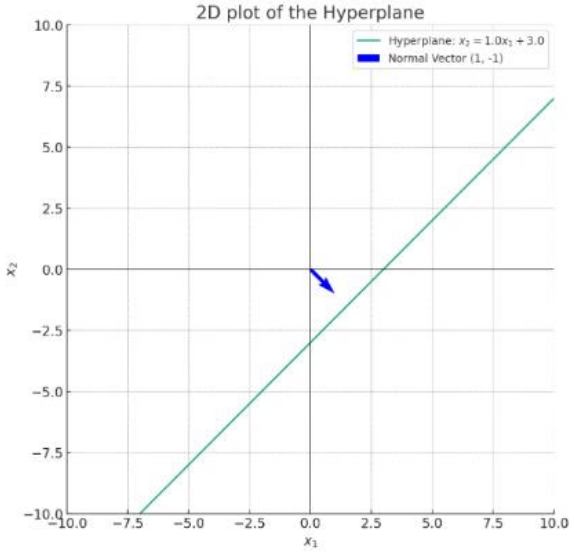
It should be mentioned that the vector $a \in \mathbb{R}^n$ is referred to as the **normal vector** to the hyperplane and dictates the orientation of the hyperplane in space. As such, the normal vector, as implied by the name, is orthogonal to the hyperplane (it forms a $90°$ angle with the plane). The scalar $b \in \mathbb{R}$ dictates the distance of the hyperplane from the origin. Further, if $b = 0$, then the hyperplane $\mathcal{H}$ will pass through the origin and will also be a subspace.

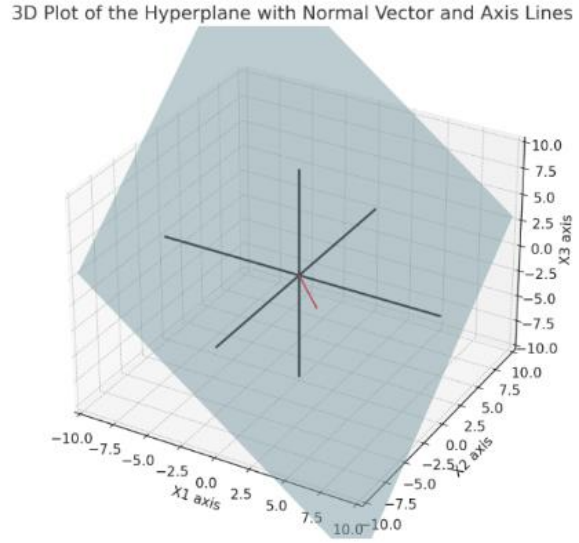
- It should be mentioned that a hyperplane $\mathcal{H}$ is completely defined by the normal vector $a \in \mathbb{R}^n$ and the scalar $b \in \mathbb{R}$.

HYPERPLANE ILLUSTRATIONS

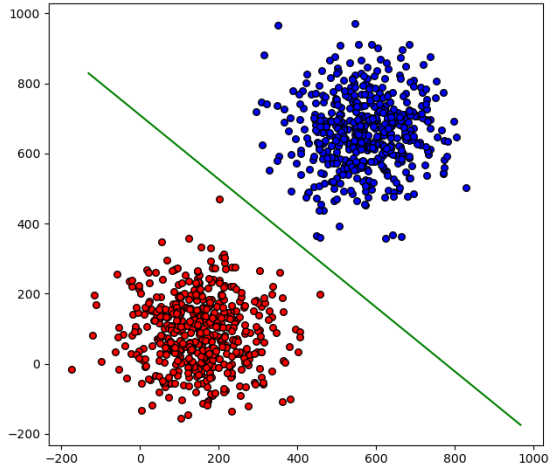




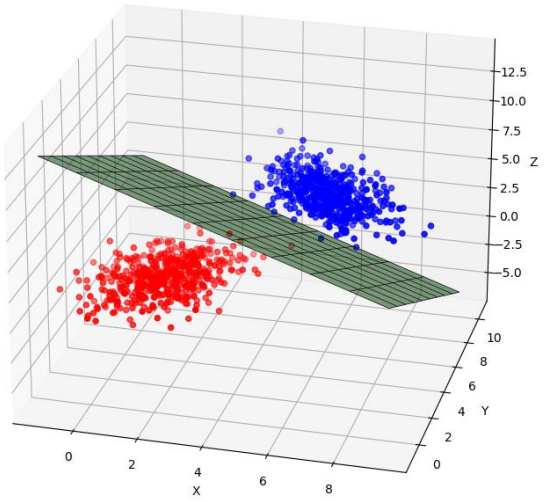




A Hyperplane in $\mathbb{R}^2$ is a Line



A Hyperplane in $\mathbb{R}^3$ is a Plane



HYPERPLANE RULES OF THUMB

Hyperplane Geometry Reference

This is a reference to help demonstrate the connection between the normal vector $a \in \mathbb{R}^n$, the scaler $b \in \mathbb{R}$, and how the unique pair (a, b) fully define a hyperplane in space along with its orientation and distance from the origin.

- There are 3 components that define a hyperplane in space:
  1. **(Orientation)** Given the normal vector $a \in \mathbb{R}^n$, the hyperplane **WILL be perpendicular** to a.
  2. **(Distance to Origin)** The distance from the hyperplane to the origin can easily be computed by as the projection of the origin 0 onto the plane $\mathcal{H}$, which will be $Proj_{\mathcal{H}}(0) = \frac{|b|}{\|a\|_2}$.
  3. **(Directions of Increase/Decrease)** If $b = 0$, the hyperplane passes through the origin. If $b \geq 0$ then the plane is offset from the origin in the direction of the normal vector a. If $b < 0$ then the hyperplane is offset from the origin in the direction opposite of the normal vector a.
     - It bears mentioning that the normal vector a points in the direction of increase of the quantity $a^T x$. In other words, a points in the direction of the side of the hyperplane such that $a^T x \geq b$, whereas the negative normal $-a$ points in the direction of the other side of the hyperplane such that $a^T x \leq b$.

PROJECTIONS ONTO A HYPERPLANE

The **projection** is a concept of huge importance in machine learning and optimization.

- At a general level, a **projection** is a function that maps some vector $y \in \mathbb{R}^n$ to some other point that is in a desired ``feasible region'' $\mathcal{S} \subseteq \mathbb{R}^n$; further, this new point (denoted as $Proj_{\mathcal{S}}(y)$) that $y$ is mapped to is the ``closest'' feasible point to y.
- As a side note, there are a variety of different projections. However, we will **only be considering orthogonal projections**, i.e., projections that are perpendicular to their original points.

Definition: Projection (Onto an Affine Set)

Let $\mathcal{S} \subseteq \mathbb{R}^n$ denote an affine set (or subspace). Then, the projection of a vector $y \in \mathbb{R}^n$ onto $\mathcal{S}$ is given by

$$Proj_{\mathcal{S}}(y) = \underset{v \in \mathcal{S}}{\operatorname{argmin}} \|v - y\|_2 .$$

Lemma: Projection to a Hyperplane

Given a vector $a \in \mathbb{R}^n$, such that $a \neq 0$, and a scalar $b \in \mathbb{R}$, let $\mathcal{H} := \{x \in \mathbb{R}^n : a^T x = b\}$ denote a hyperplane. The projection vector $v^* \in \mathbb{R}^n$that results from projecting the vector $y \in \mathbb{R}^n$ onto the hyperplane $\mathcal{H}$ is given by the equation

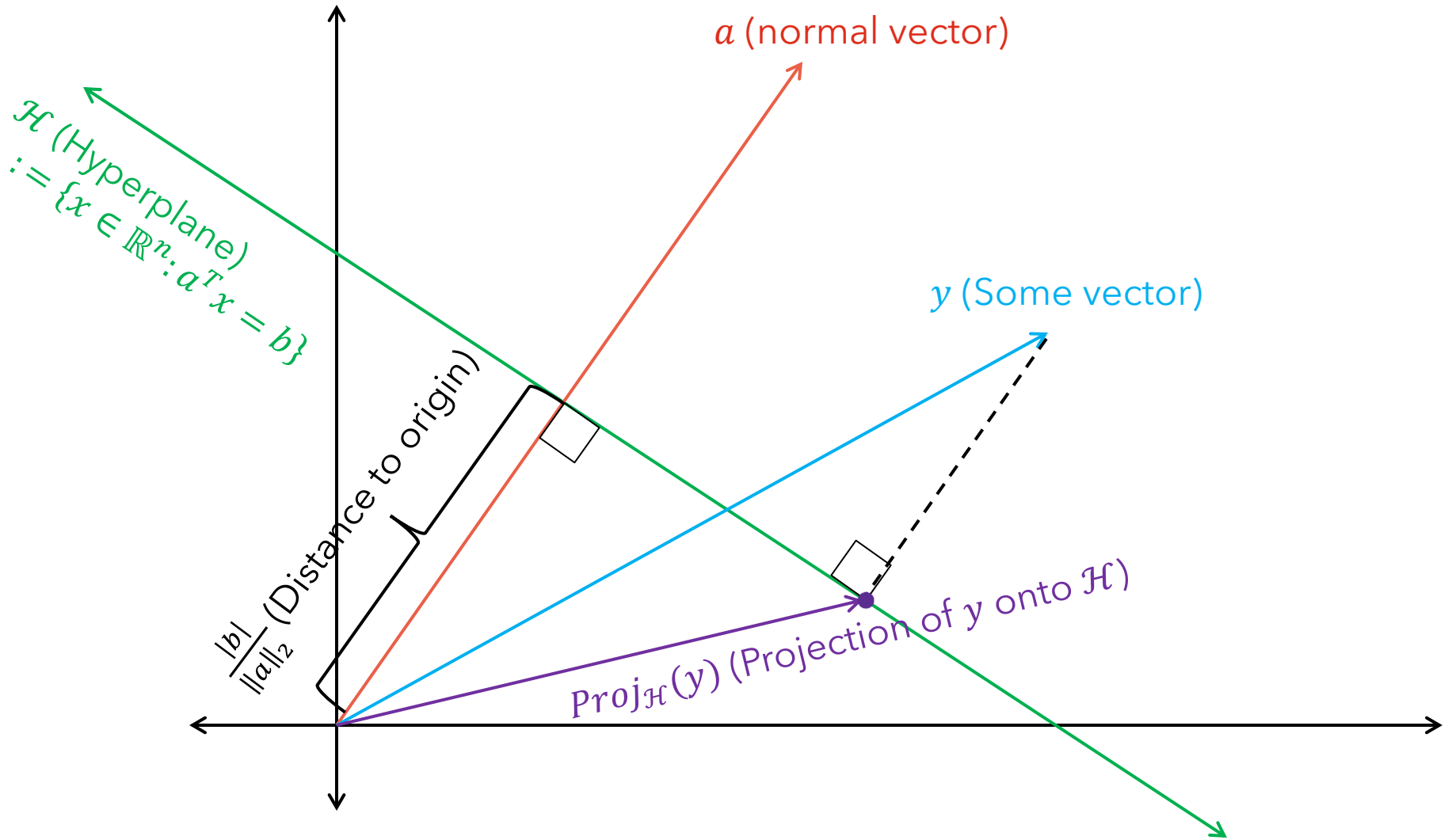
$$v^* = Proj_{\mathcal{H}}(y) = \underset{v \in \mathcal{H}}{\operatorname{argmin}} \|v - y\|_2 = y - a\frac{a^T y - b}{\|a\|_2^2}.$$

Similarly, the shortest distance from y to $\mathcal{H}$ is given by

$$\mathcal{D}(y, \mathcal{H}) = \frac{|a^T y - b|}{\|a\|_2}.$$

GEOMETRY OF PROJECTIONS

(ONTO HYPERPLANES)



MATRICES

The matrix: is a fundamental object in mathematics which stores data in a tabular form of rows and columns. It can also be viewed as the concatenation of multiple vectors into a single object upon which certain operations can be performed.

A matrix $\boldsymbol{A}$ with $\boldsymbol{m}$ **rows** and $\boldsymbol{n}$ **columns** of real-values is denoted as $\boldsymbol{A} \in \mathbb{R}^{\boldsymbol{m \times n}}$. Such a matrix can be explicitly written as

$$A = \begin{bmatrix} A_{11} & A_{12} & \dots & A_{1n} \\ A_{21} & A_{22} & \dots & A_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ A_{m1} & A_{m2} & \dots & A_{mn} \end{bmatrix}.$$

Further, the element corresponding to the $\boldsymbol{i}$**-th row** and the $\boldsymbol{j}$**-th column** is written as $\boldsymbol{A_{ij}}$.

Matrices can also be written in vector form. Letting $\boldsymbol{a_j} \in \mathbb{R}^{\boldsymbol{m}}$ denote the $\boldsymbol{j}$**-th column** vector and $\overline{\boldsymbol{a}}_{\boldsymbol{i}} \in \mathbb{R}^{\boldsymbol{n}}$ denote the $\boldsymbol{i}$**-th row** of the matrix, we have

$$A = \begin{bmatrix} | & | & & | \\ a_1 & a_2 & \dots & a_n \\ | & | & & | \end{bmatrix} \quad \text{and} \quad A^T = \begin{bmatrix} - & a_1^T & - \\ - & a_2^T & - \\ & \vdots & \\ - & a_n^T & - \end{bmatrix} \quad \text{and} \quad A^T = \begin{bmatrix} | & | & & | \\ \overline{a}_1^T & \overline{a}_2^T & \dots & \overline{a}_m^T \\ | & | & & | \end{bmatrix} \quad \text{and} \quad A = \begin{bmatrix} - & \overline{a}_1 & - \\ - & \overline{a}_2 & - \\ & \vdots & \\ - & \overline{a}_m & - \end{bmatrix}.$$

MATRIX-VECTOR PRODUCTS

Matrix-Vector Products

Given $A \in \mathbb{R}^{m \times n}$ (with rows denoted by the vectors $\overline{a}_i \in \mathbb{R}^n$) and $x \in \mathbb{R}^n$, the matrix-vector product is defined as

$$Ax = \begin{bmatrix} - & \overline{a}_1 & - \\ - & \overline{a}_2 & - \\ & \vdots & \\ - & \overline{a}_m & - \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} \overline{a}_1^T x_1 \\ \overline{a}_2^T x_2 \\ \vdots \\ \overline{a}_m^T x_n \end{bmatrix}.$$

Similarly, for columns denoted by the vectors $a_j \in \mathbb{R}^m$, we have

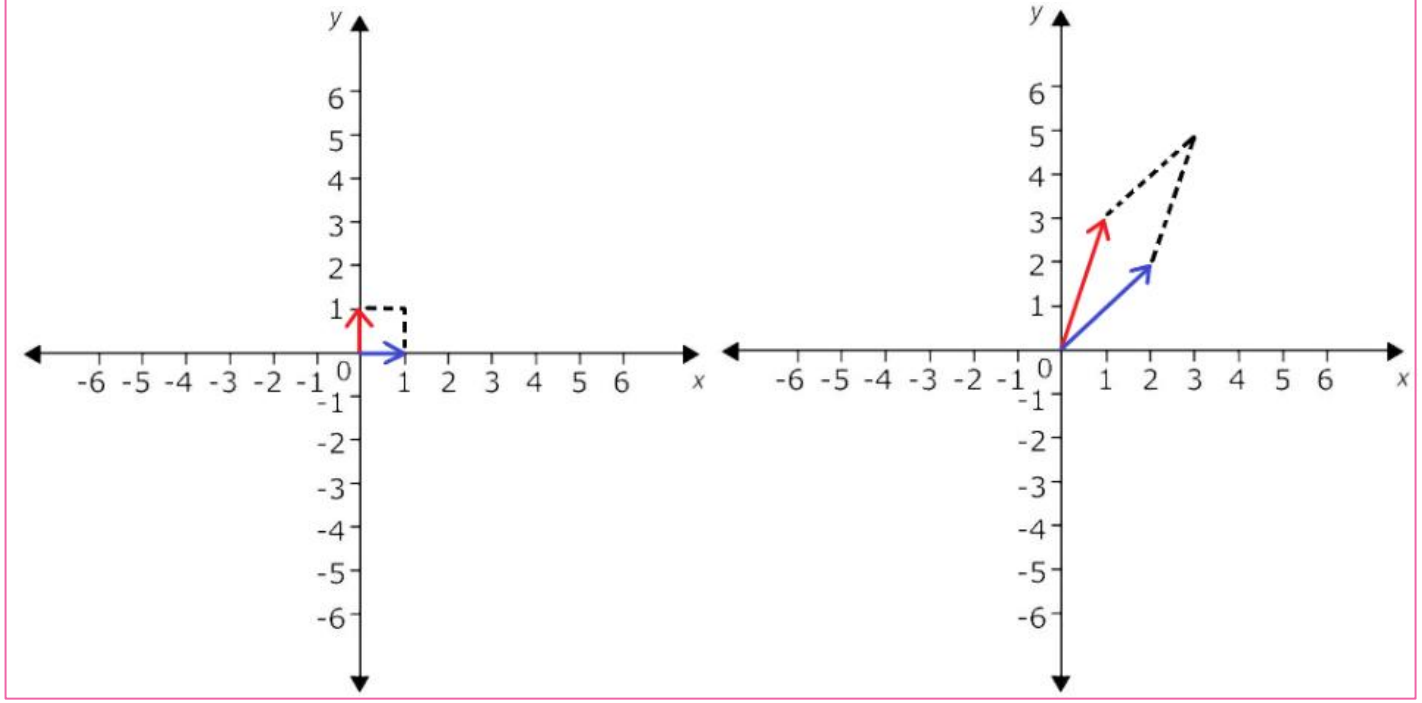
$$Ax = \begin{bmatrix} | & | & & | \\ a_1 & a_2 & \cdots & a_n \\ | & | & & | \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \sum_{j=1}^{n} a_j x_j .$$

- Matrices should be thought of as **linear transformations** (sliding / stretching / shrinking / rotating vectors) that serve as a mapping from one set of "coordinates" to another.

Example

Given $A = \begin{bmatrix} 2 & 1 \\ 2 & 3 \end{bmatrix}$ and $x = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$, we have $Ax = \begin{bmatrix} 3 \\ 5 \end{bmatrix} \coloneqq b$. This should be thought of as the following: "the linear transformation A maps the point x to the point b."

- This figure represents the unit vectors before and after applying the linear transformation A.

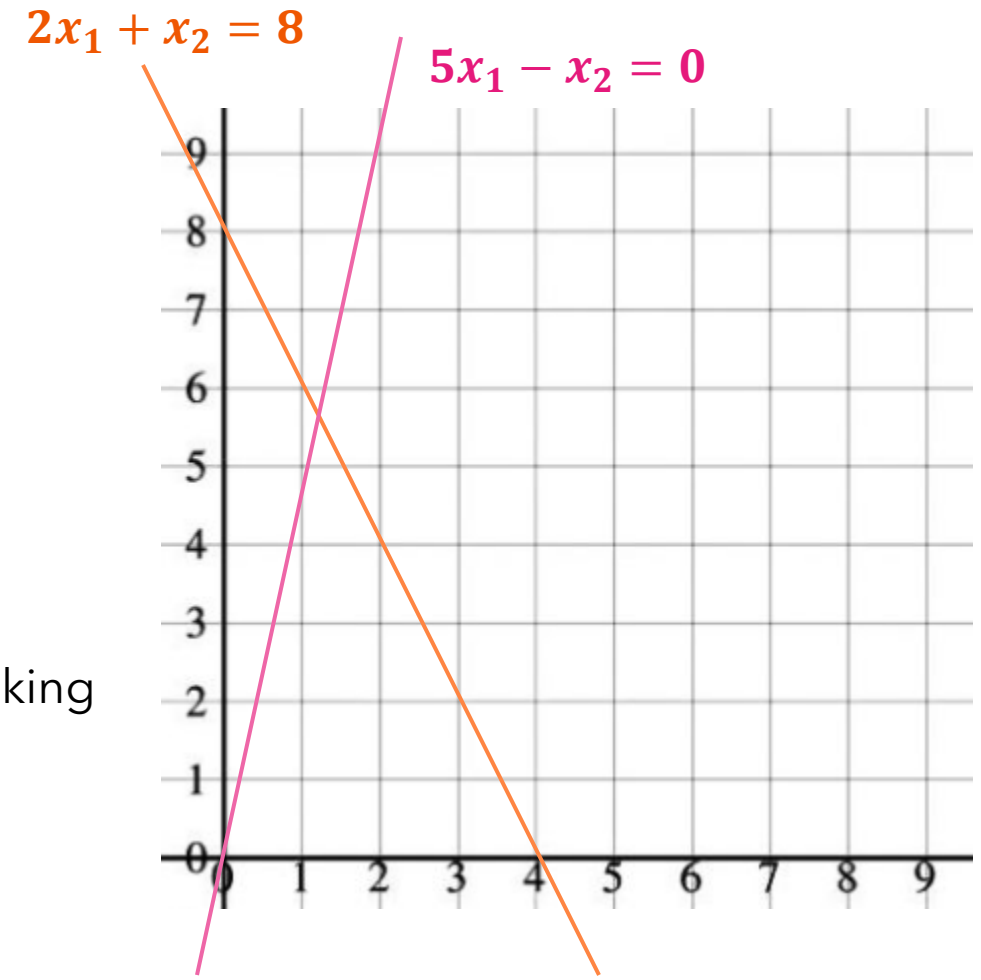


SYSTEMS OF LINEAR EQUATIONS

- Generally, when one has several equations that are related to each other through the same variables, it is often of interest to know what point(s) satisfies all the equations at once.
- Based on what we have learned, we can write each equation as a hyperplane of the form $\overline{a}_i^T x = b_i, \forall i \in \{1,2,\dots,m\}$.

$$
\begin{aligned}
a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &= b_1 \longrightarrow \text{Hyperplane } 1\\
a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n &= b_2 \longrightarrow \text{Hyperplane } 2\\
&\vdots\\
a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n &= b_m \longrightarrow \text{Hyperplane } m\\
&\downarrow\\
Ax &= b
\end{aligned}
$$

- Further, we can rewrite this system in matrix-vector notation.
- Thus, solving this system of linear equations for x is the same as asking "**what vector x lands on all the hyperplanes simultaneously**"?

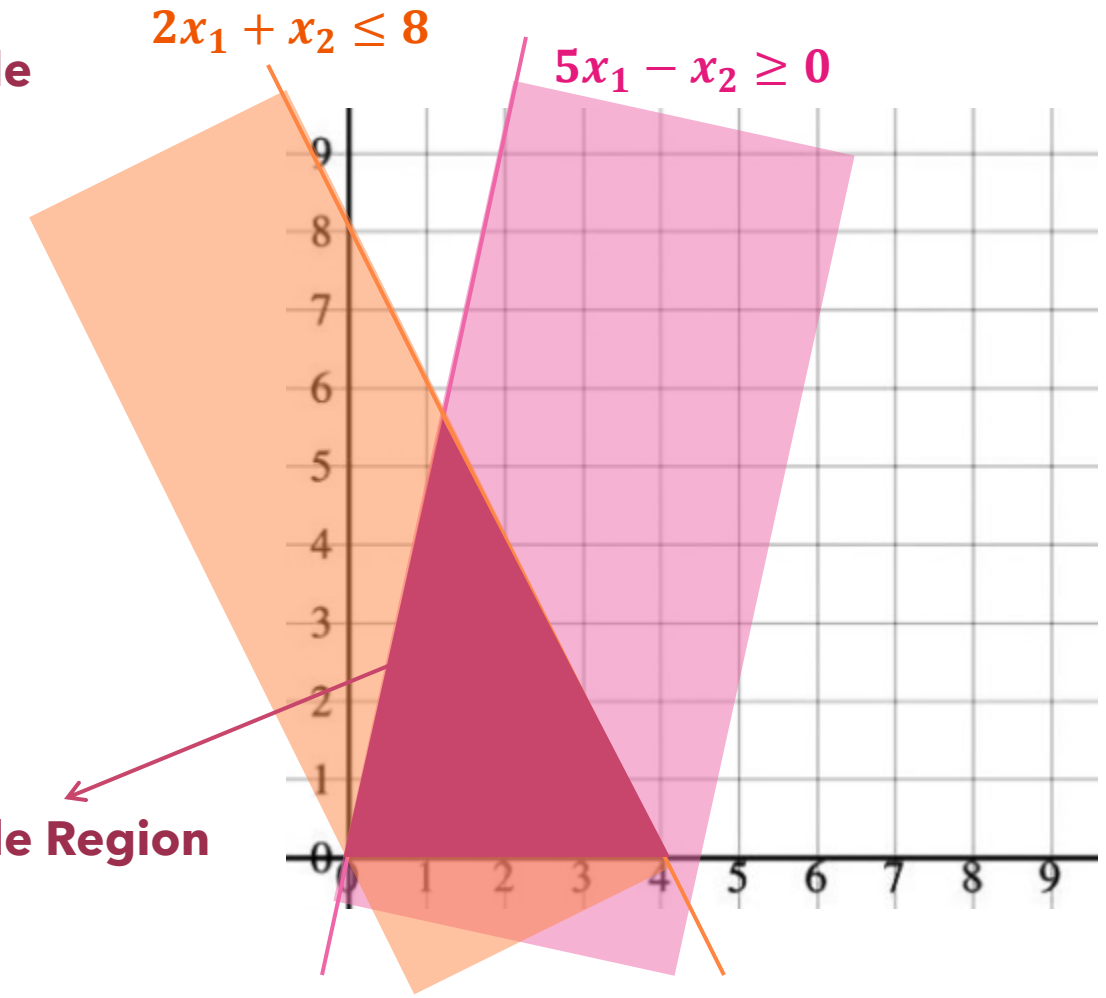


SYSTEMS OF LINEAR INEQUALITIES

- What happens when we are instead interested in in a system of linear **inequalities**?
- Assuming the inequalities overlap, there is typically a **feasible region** of solutions that will satisfy all the inequalities.
- Such a region is referred to as a **polytope**.

$$a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n \leq b_1 \longrightarrow \text{Hyperplane 1}$$
$$a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n \leq b_2 \longrightarrow \text{Hyperplane 2}$$
$$\vdots$$
$$a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n \leq b_m \longrightarrow \text{Hyperplane } m$$
$$\downarrow$$
$$Ax \leq b$$

- In this case, the hyperplanes now define **half-spaces**, which "cut" the ambient space they are in "in half".



MATRIX INVERSION

Now that we've understood how matrices can be viewed as linear transformations that rotate, stretch, shrink, or reflect a vector, a natural question is how to "undo" these transformations?

The Inverse

A square matrix $A \in \mathbb{R}^{n \times n}$ is said to be invertible if there exists a corresponding matrix (known as the inverse), denoted by $A^{-1} \in \mathbb{R}^{n \times n}$, such that

$$A^{-1}A = I = AA^{-1},$$

where $I \in \mathbb{R}^{n \times n}$ is the identity matrix.

- Intuitively, not every matrix has an inverse; clearly, non-square matrices do not have inverses. Further, even certain square matrices do not have inverses (matrices that have a **determinant** of 0).
- The **determinant** is a quantity that corresponds to the **volume** of the geometric area obtained **after applying the transformation**.
- Intuitively, what would a matrix with $det(A) = 0$ imply?
- A matrix A with $det(A) = 0$ implies that the resulting volume after applying the linear transformation is 0. This means that the matrix A maps vectors into a **lower dimension space**. The matrix A is also said to **not be full-rank** (the rows or columns are **not linearly independent**).

EIGEN-VALUES & EIGEN-VECTORS

Now we come to perhaps the pinnacle of what is taught in a linear algebra course and is probably one of the most unnecessarily ambiguous and misremembered concepts.

Eigenvalues & Eigenvectors

Let $A \in \mathbb{R}^{n \times n}$ be a square matrix. Then $x \in \mathbb{C}^n$ is an **eigenvector** and $\lambda \in \mathbb{C}$ is the corresponding **eigenvalue** if and only if

$$Ax = \lambda x, \qquad \text{for } x \neq 0.$$

What significance does an eigenvector and eigenvalue have in a **geometrical sense**?

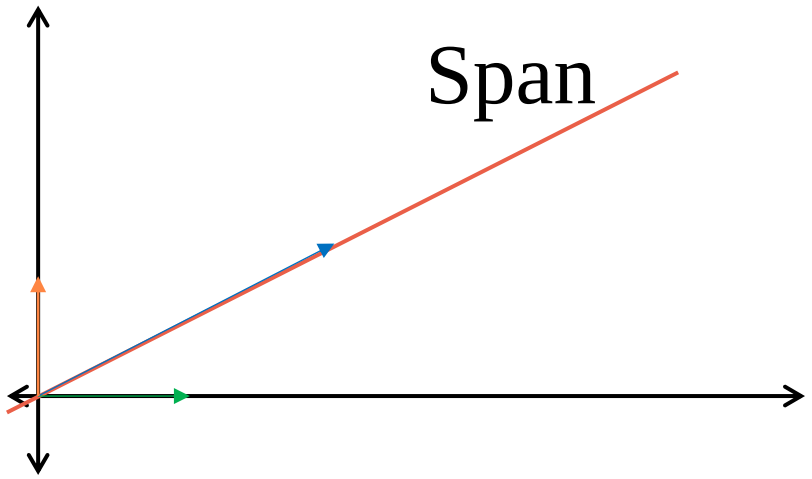
- Remember, the matrix A can be thought of as performing some linear transformation.
- As such, most original vectors $x \in \mathbb{R}^n$ have a certain direction (also known as the "**span**" of the vector when also including the opposite direction as well) before applying the linear transformation and the vector $y \in \mathbb{R}^n$ obtained after the transformation Ax will typically have a different direction.
- An **eigenvector** of A is a vector that whose **span** (before the transformation) **does not change** (after the transformation). In this way, the vector only stretches, shrinks, or reverses direction along the same line, where the amount of this change is defined by the corresponding eigenvalue.

What does it mean for an eigenvalue-eigenvector pair to be **complex** in a geometrical sense?

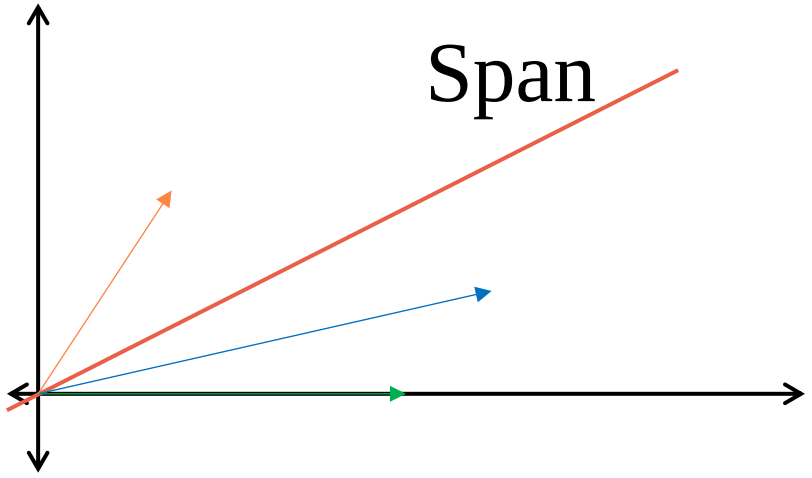
- In this case, the matrix eigenvector of A corresponds to a direction which the transformation **rotates** about. Thus, all vectors along this direction are not "knocked off" that direction (hence an eigenvector), and the amount of rotation is defined by the corresponding eigenvalue.

ILLUSTRATION OF EIGEN-VALUES & EIGEN-VECTORS

Example of a "non" eigenvector



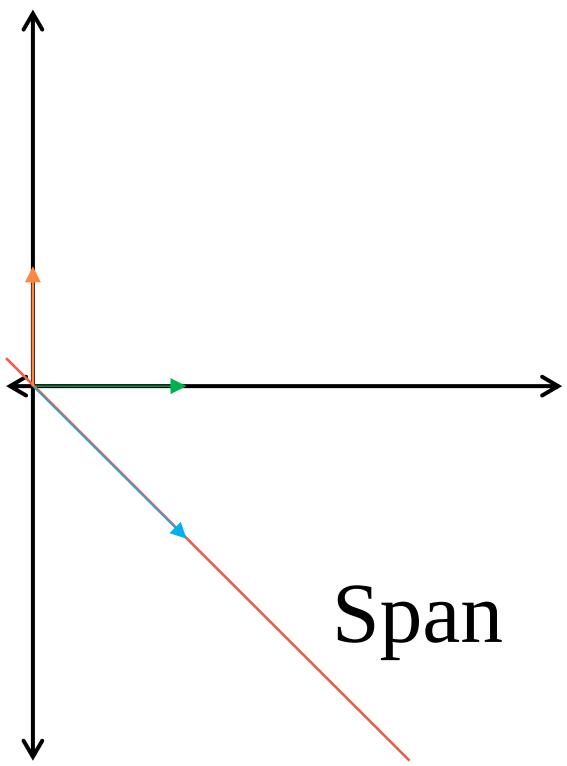
Before Transform



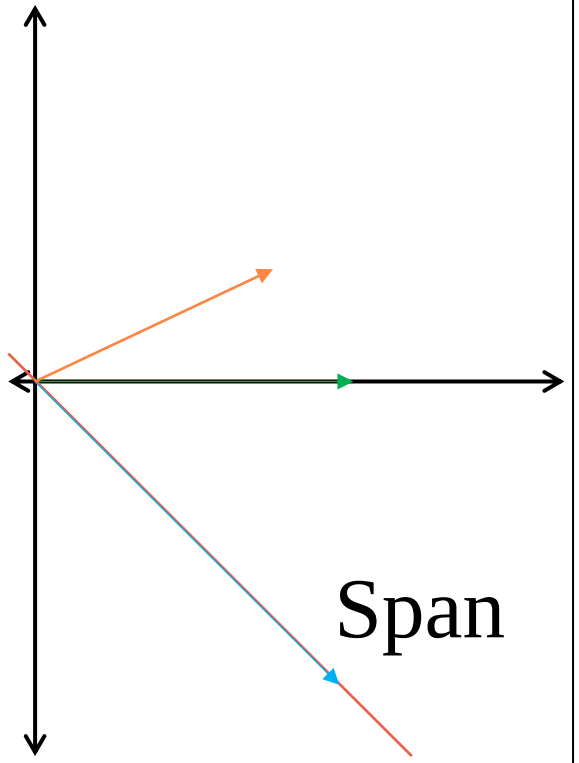
After Transform

Example of an eigenvector

Before Transform



After Transform



QUADRATIC FORMS & MATRIX DEFINITENESS

Although we have been dealing with linear systems until this point, **quadratic terms** and some of their properties are of paramount importance in for many types of problems (nonlinear optimization and ML).

- Given a square symmetric matrix $Q \in \mathbb{R}^{n \times n}$ and vector $x \in \mathbb{R}^n$, the term $\boldsymbol{x^T Q x}$ is called a **quadratic form**.

Matrix Definiteness

- A symmetric matrix $Q \in \mathbb{R}^{n \times n}$ is **positive definite** (PD) if, for all nonzero $x \in \mathbb{R}^n$, $x^T Q x > 0$. This is typically denoted $Q \succ 0$.
- A symmetric matrix $Q \in \mathbb{R}^{n \times n}$ is **positive semi-definite** (PSD) if, for all nonzero $x \in \mathbb{R}^n$, $x^T Q x \geq 0$. This is typically denoted $Q \succcurlyeq 0$.
- A symmetric matrix $Q \in \mathbb{R}^{n \times n}$ is **negative semi-definite** (NSD) if, for all nonzero $x \in \mathbb{R}^n$, $x^T Q x \leq 0$. This is typically denoted $Q \preccurlyeq 0$.
- A symmetric matrix $Q \in \mathbb{R}^{n \times n}$ is **negative definite** (ND) if, for all nonzero $x \in \mathbb{R}^n$, $x^T Q x < 0$. This is typically denoted $Q \prec 0$.

If a symmetric matrix $Q \in \mathbb{R}^{n \times n}$ doesn't meet any of these conditions, then it is **indefinite**.

The definiteness of a matrix tells us a lot of useful information.

- PD and ND matrices are always invertible.
- PD matrices have positive eigenvalues and PSD matrices have non-negative eigenvalues. Conversely, ND matrices have negative eigenvalues and PSD matrices have non-positive eigenvalues.
- The definiteness of a matrix can tell you information regarding the optimality of a function.
- Indefinite matrices can have both positive and negative eigenvalues (Could correspond to a saddle point).

PROPERTIES OF SYMMETRIC MATRICES

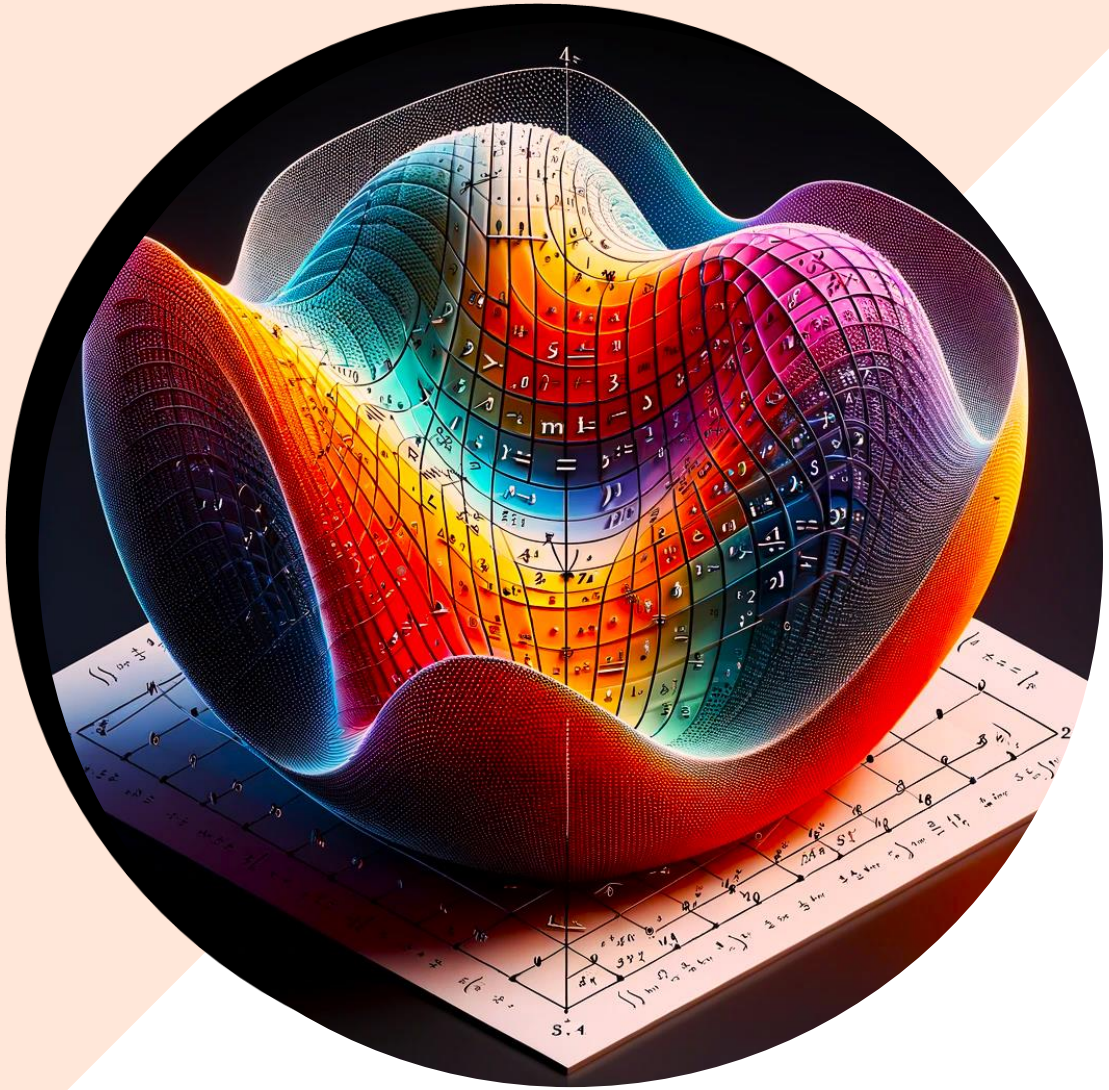
Symmetric matrices have certain useful properties and show up often in machine learning; such matrices in ML applications are typically the **Hessian matrix** of 2nd-order derivatives as well as the **covariance matrix**. As such, we will highlight some of the most relevant properties here.

Symmetric Matrix

A symmetric matrix $A \in \mathbb{R}^{n \times n}$ has the property that $A = A^T$. This means that $A_{ij} = A_{ji}$ for all i and j in $\{1,2, \dots, n\}$. Thus, symmetric matrices are "mirrored" across the diagonal.

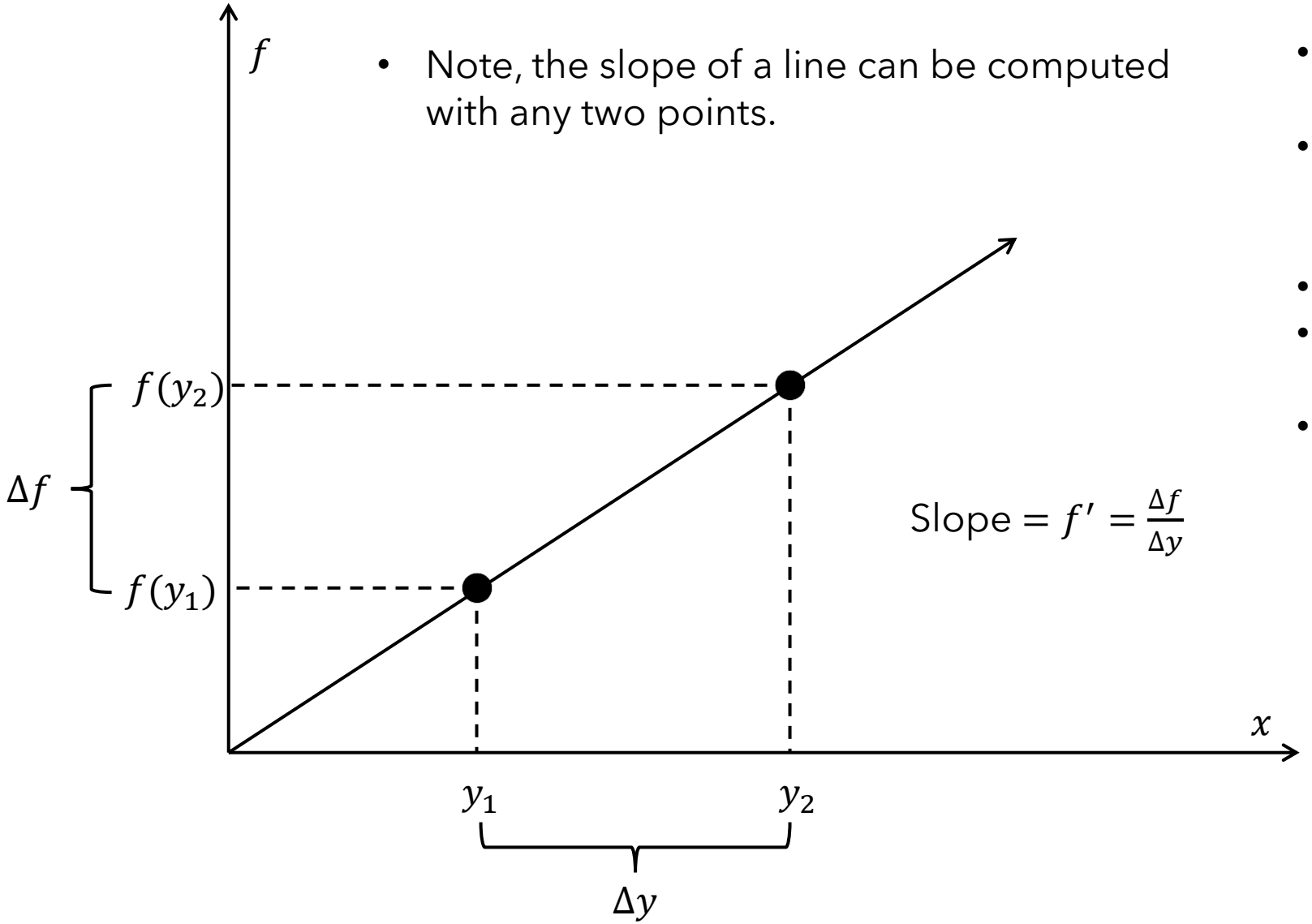
Some Properties

- All symmetric matrices must be square (have dimension $n \times n$).
- Given a symmetric matrix $A \in \mathbb{R}^{n \times n}$, all its eigenvector-eigenvalue pairs are real-valued (not complex).
- All eigenvectors of a symmetric matrix are orthogonal, i.e., $v_1^T v_2 = 0$ for any eigenvectors v_1 and v_2.
- Given that all the eigenvalues are real-valued, it is easy to determine the definiteness of a symmetric matrix.



CALCULUS

DERIVATIVE (SLOPE OF A LINE)

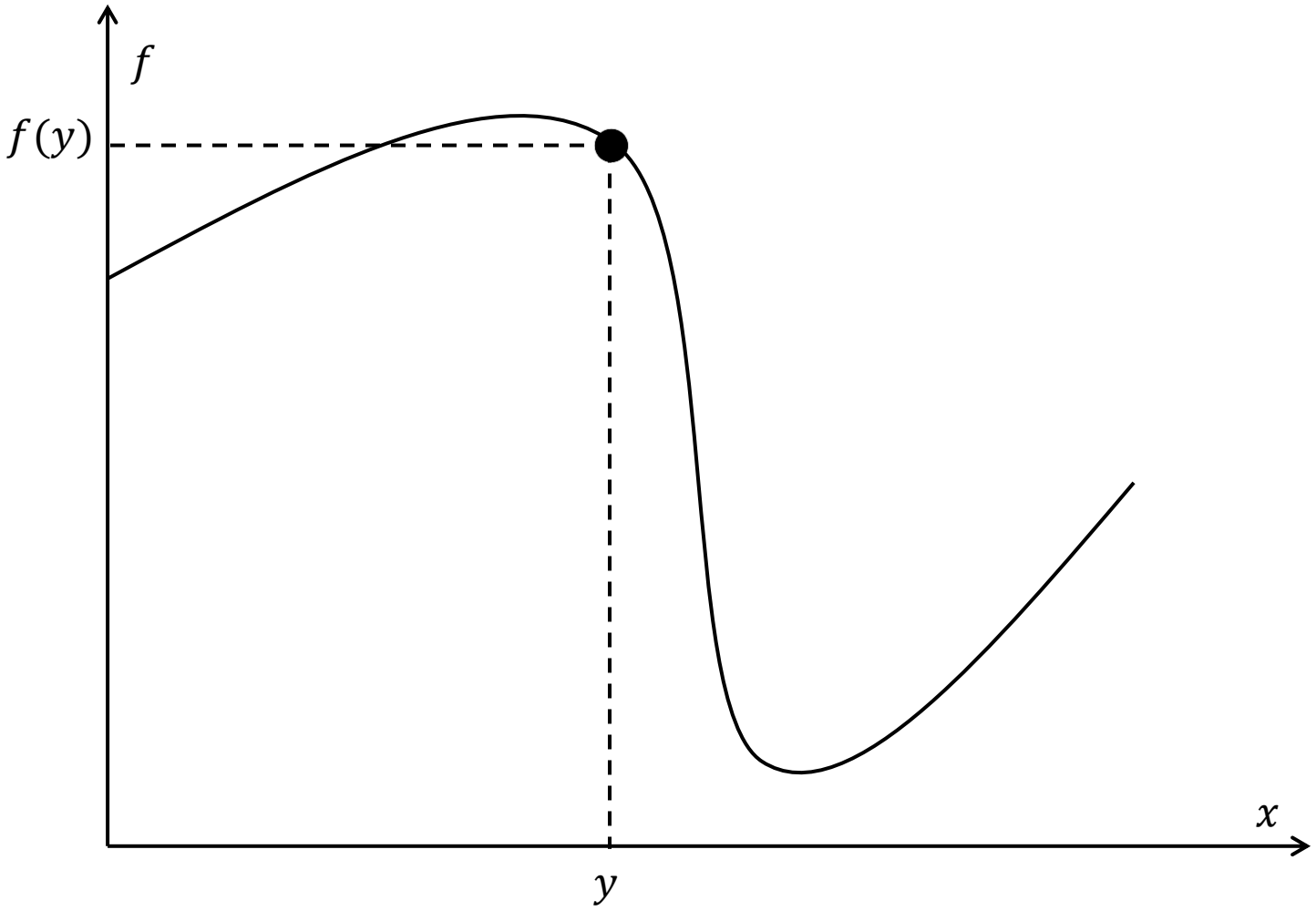


- Note, the slope of a line can be computed with any two points.

Slope $= f' = \frac{\Delta f}{\Delta y}$

- Given a line $f: \mathbb{R} \to \mathbb{R}$, what is the "rate of change" of the line?
- In other words, how does the output of the function change when changing the input?
- We call this the slope of the line.
- This is also the derivative of the line.
- Lines have a constant rate of change (because their slope is constant), and by extension, the derivative of the line is the same when evaluated at any point on the line.

DERIVATIVE (UNIVARIATE FUNCTION)

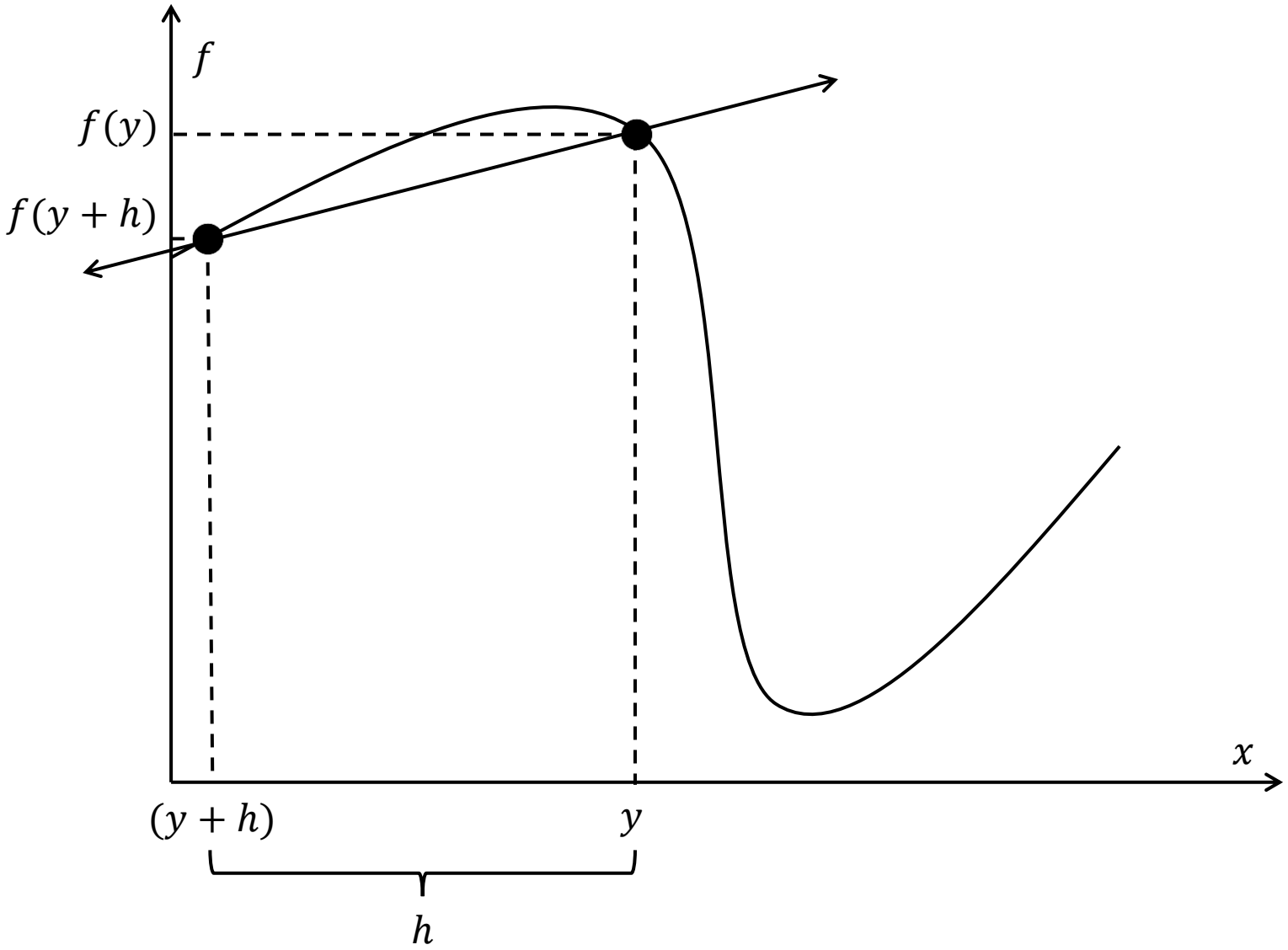


- Now, the definition of the derivative for a general univariate function $f: \mathbb{R} \to \mathbb{R}$ is given by the expression:

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

- Notice that this is essentially the same concept as computing the slope of a line.
- The difference comes in when a function is non-linear.
- The derivative is simply the measure of the "instantaneous rate-of-change" at a given point in the domain of the function.

DERIVATIVE (UNIVARIATE FUNCTION)

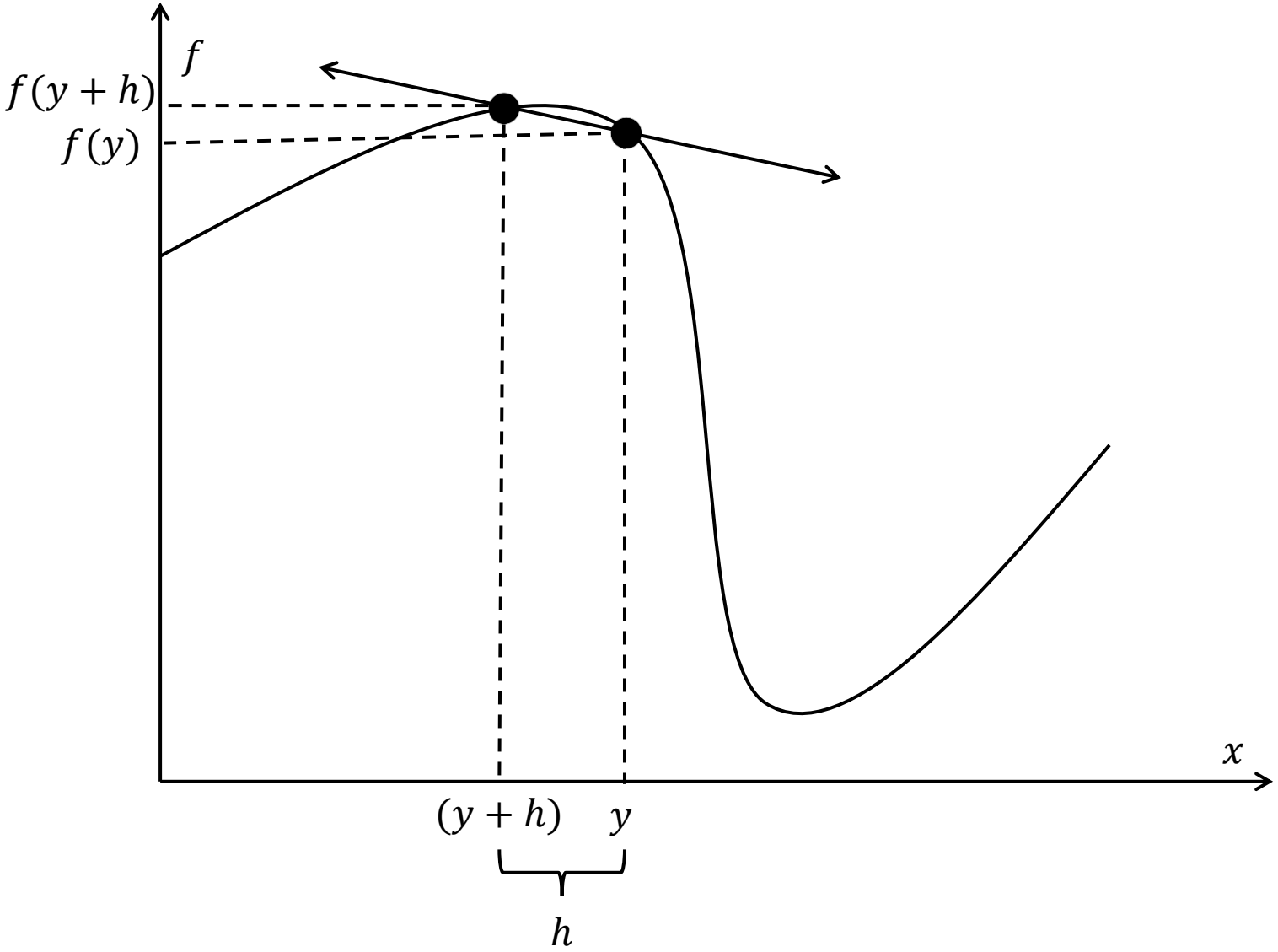


- Now, the definition of the derivative for a general univariate function $f: \mathbb{R} \to \mathbb{R}$ is given by the expression:

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

- Notice that this is essentially the same concept as computing the slope of a line.
- The difference comes in when a function is non-linear.
- The derivative is simply the measure of the "instantaneous rate-of-change" at a given point in the domain of the function.

DERIVATIVE (UNIVARIATE FUNCTION)

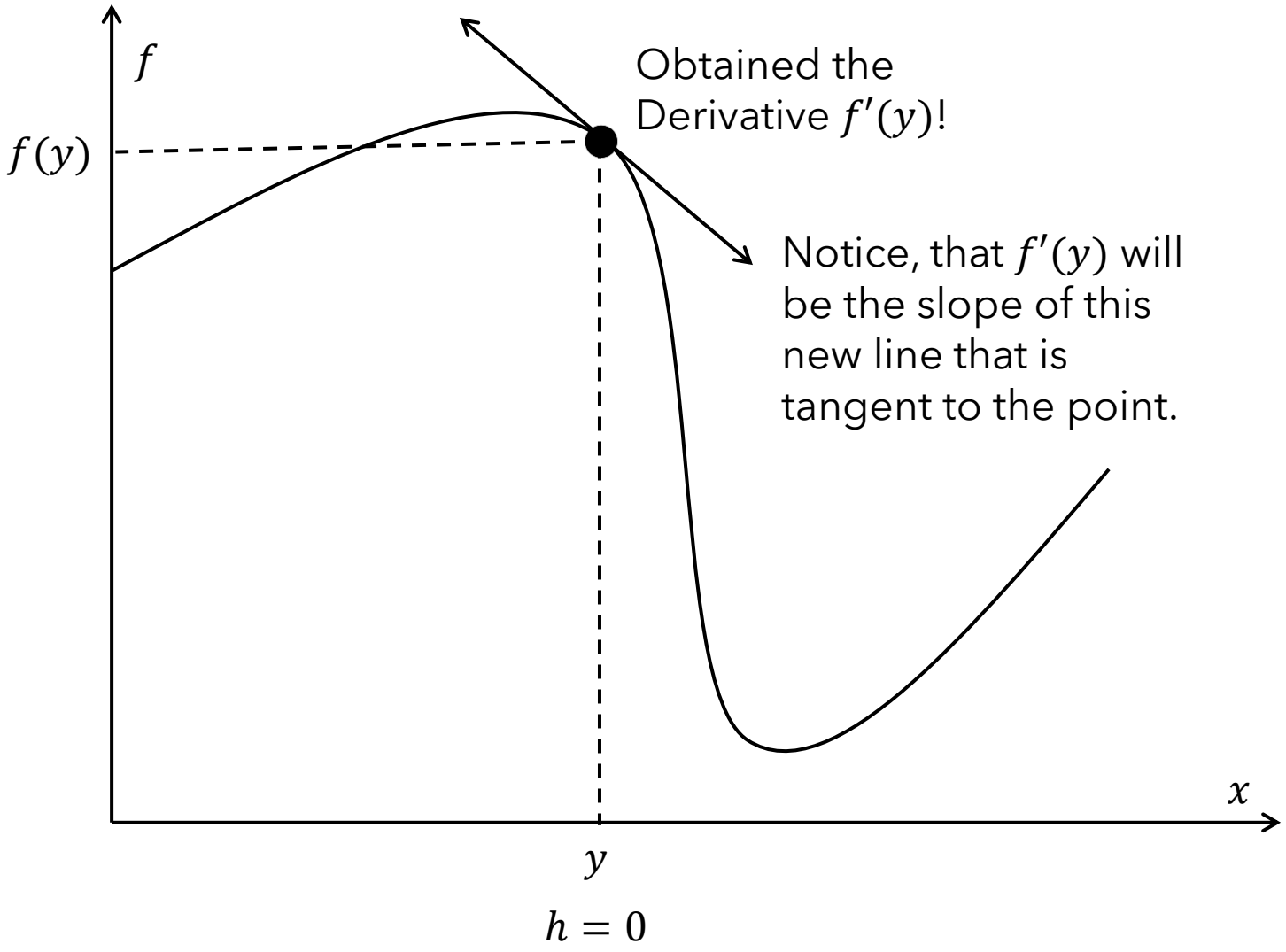


- Now, the definition of the derivative for a general univariate function $f: \mathbb{R} \to \mathbb{R}$ is given by the expression:

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

- Notice that this is essentially the same concept as computing the slope of a line.
- The difference comes in when a function is non-linear.
- The derivative is simply the measure of the "instantaneous rate-of-change" at a given point in the domain of the function.

DERIVATIVE (UNIVARIATE FUNCTION)



- Now, the definition of the derivative for a general univariate function $f: \mathbb{R} \to \mathbb{R}$ is given by the expression:

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

- Notice that this is essentially the same concept as computing the slope of a line.
- The difference comes in when a function is non-linear.
- The derivative is simply the measure of the "instantaneous rate-of-change" at a given point in the domain of the function.

DERIVATIVE (EXAMPLE)

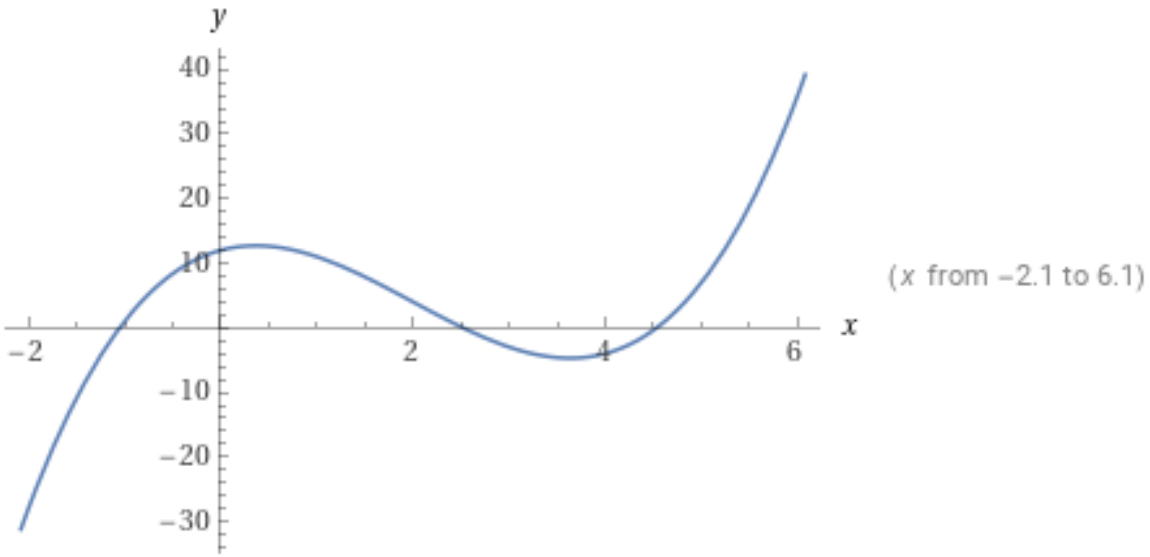
Function of interest

$$y = x^3 - 6x^2 + 4x + 12$$

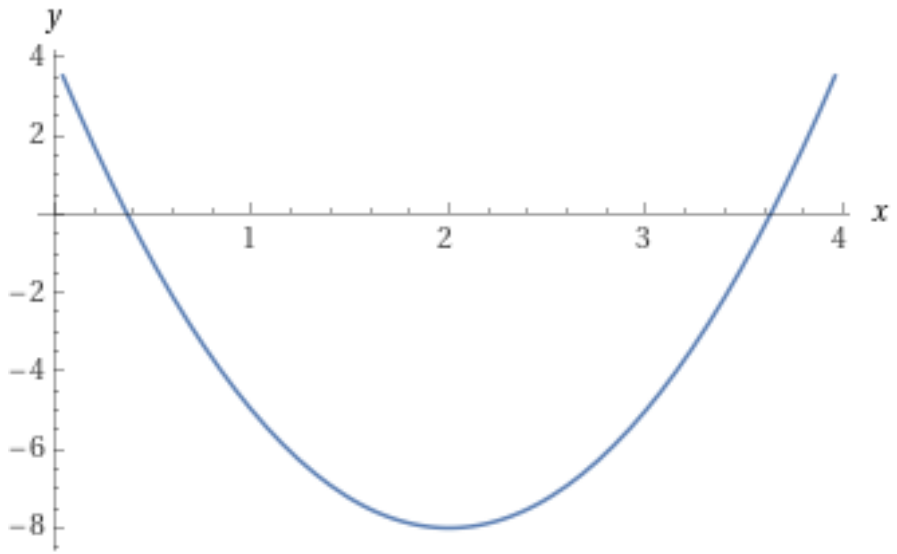
Taking Derivative of y w.r.t. x →

Corresponding derivative

$$y' = \frac{dy}{dx} = 3x^2 - 12x + 4$$



(x from −2.1 to 6.1)



(x from 0 to 4)

FIRST & SECOND DERIVATIVES

(EXAMPLE CONT.)

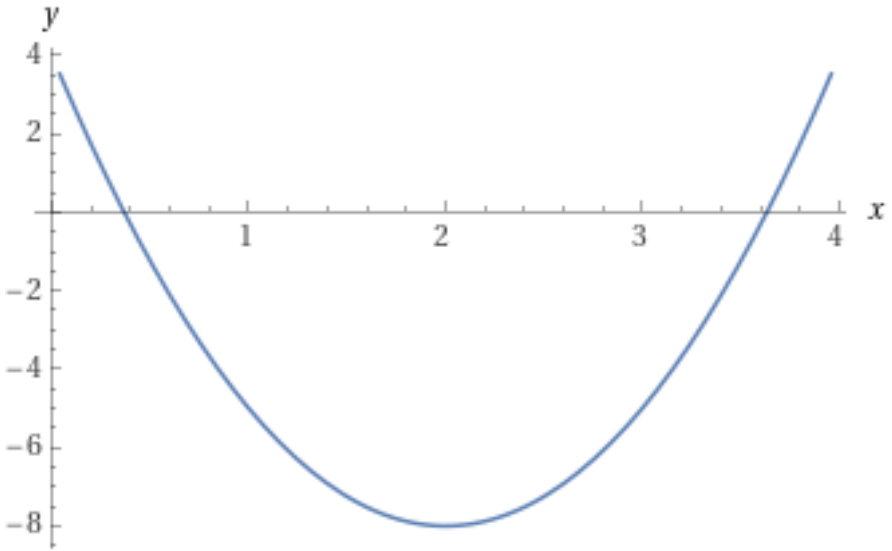
First-order derivative

$$y' = \frac{dy}{dx} = 3x^2 - 12x + 4$$

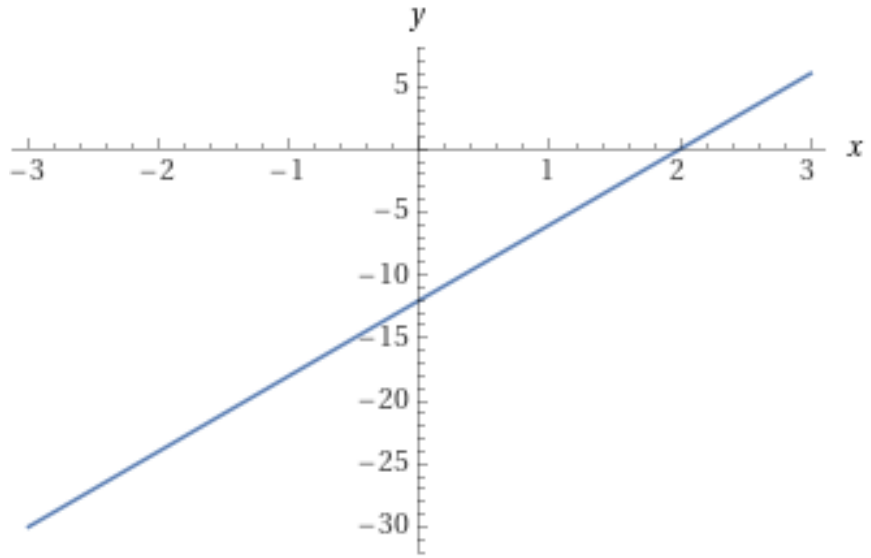
Taking Derivative of y' w.r.t. x →

Second-order derivative

$$y'' = \frac{d^2y}{d^2x} = 6x - 12$$

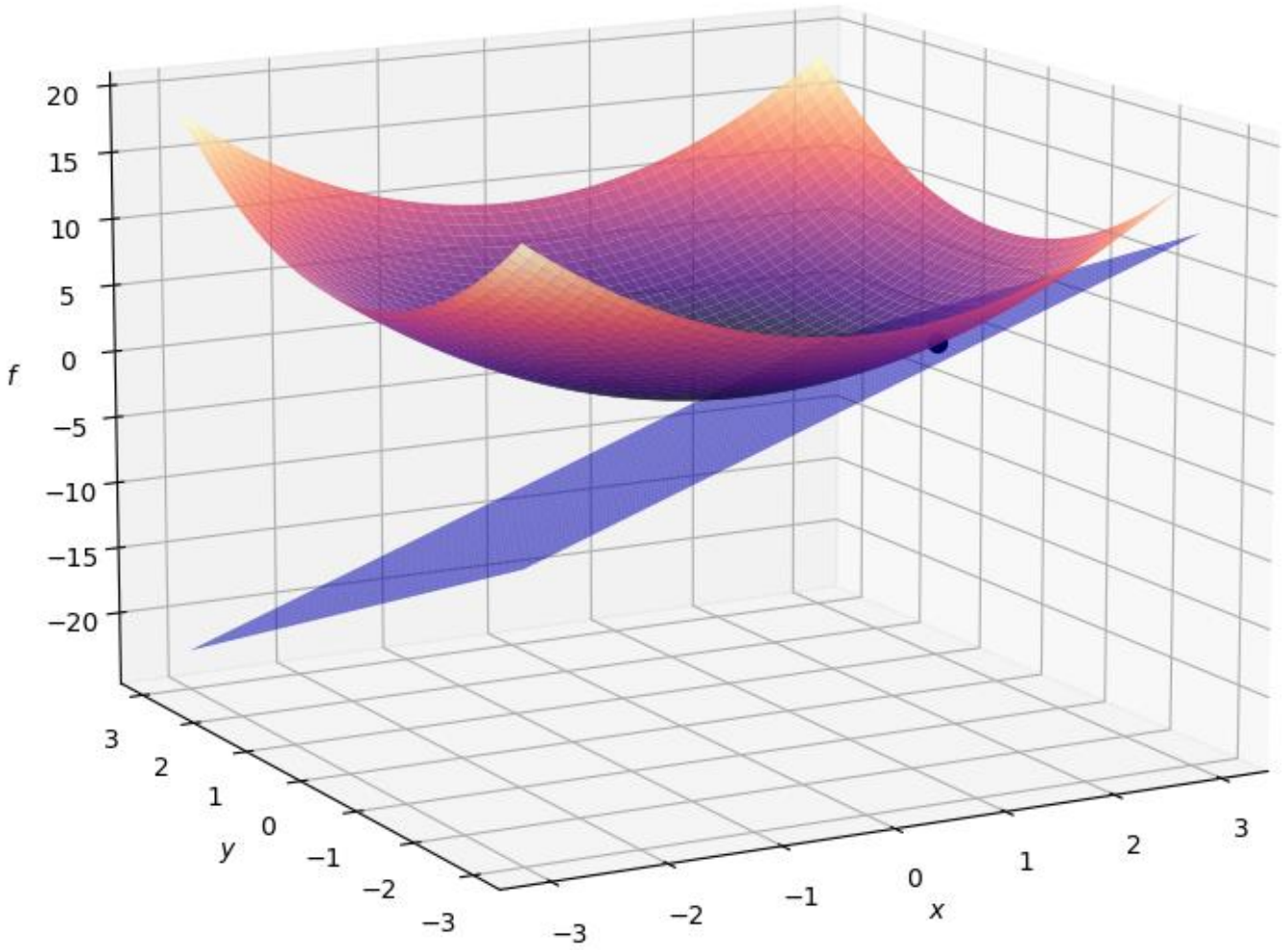


(x from 0 to 4)



(x from −3 to 3)

THE GRADIENT (MULTIVARIATE FUNCTION)



- When dealing with a function that takes multiple variables as inputs, instead of having a single derivative, the function will have a "partial derivative" corresponding to each variable.
- We define the ith partial derivative of a general multivariate function $f: \mathbb{R}^n \to \mathbb{R}$ by the expression:

$$\frac{\partial f}{\partial x_i} = \lim_{h \to 0} \frac{f(x + he_i) - f(x)}{h}$$

- Here, $x \in \mathbb{R}^n$ and e_i is the n-dimensional vector with all entries equal to 0 except for the ith component, which is 1.
- Further, we can collect all n of these partial derivatives and put them into a vector known as the gradient of f and is denoted by:

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

- The gradient vector is simply the multi-variable generalization of the derivative.

THE ``TOTAL'' DERIVATIVE

(GENERAL FORM)

Given a real-valued function $f: \mathbb{R}^n \to \mathbb{R}$, the gradient of f is also referred to as the total derivative (the total derivative is a quantity that is typically defined in terms of a general vector-valued function $f: \mathbb{R}^n \to \mathbb{R}^m$, but we will only consider the former case). The total derivative (the gradient in this case is the vector $\nabla f(x) \in \mathbb{R}^n$ of partial derivatives that has the following property.

Total Derivative

Let $U \subseteq \mathbb{R}^n$ denote an open set. Then, the function $f: U \to \mathbb{R}^m$ is differentiable at a point $x \in U$ if there exists a matrix (a vector when $f: U \to \mathbb{R}$) $Df(x) \in \mathbb{R}^{m \times n}$, referred to as the ``**total derivative**'' (which is simply called the ``**gradient**'' in the case $f: U \to \mathbb{R}$ or the ``**Jacobian**'' in the general case $f: U \to \mathbb{R}^m$) such that

$$\lim_{h \to 0} \frac{\|f(x+h) - f(x) - Df(x)h\|}{\|h\|} = 0,$$

where $h \in \mathbb{R}^n$. This definition essentially states that the term $\|f(x+h) - f(x) - Df(x)h\|$ approaches 0 faster than $\|h\|$ approaches 0. Further, it follows from this definition that (for values of h ``close'' to 0)

$$f(x+h) - f(x) \approx Df(x)h.$$

THE HESSIAN MATRIX

- Let $f: \mathbb{R}^n \to \mathbb{R}$ define a multivariate function taking the vector $x \in \mathbb{R}^n$ as input.
- The **Hessian** of f is the **symmetric matrix** of **second-order derivatives** of f evaluated at x, which we denote by $\nabla^2 f(x)$.
- The Hessian is very important in optimization as it can yield explicit information regarding the nature of the optimality of any given point.
- By the nature of second-order differentiation, the Hessian matrix dictates the **curvature** of the function f.

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

PROPERTIES OF THE HESSIAN

- **Hessian matrices are symmetric**; thus, they are endowed with all the same properties of symmetric matrices earlier in these slides.

- If the **Hessian** of a function $f(x)$, evaluated at the point x, is **positive definite**, then x is a **local minimum** of f (if the gradient evaluated at this point is also 0). This means that the function f curves upward at this point and is locally convex at the point x.

- If the **Hessian** of a function $f(x)$, evaluated at the point x, is **negative definite**, then x is a **local maximum** of f (if the gradient evaluated at this point is also 0). This means that the function f curves downward at this point and is locally concave at the point x.

REFERENCES

- Cummings. “Real Analysis – A long-Form Mathematics Textbook”. 2019.

ATTRIBUTION & LICENSE

These lecture slides are part of the "*Introduction to Machine Learning*" course materials created by **Griffin Dean Kent**.

For the latest version, updates, and additional resources, visit the GitHub repository: https://github.com/GdKent/Introduction-to-Machine-Learning.