Abstract

In this pieces of notes, we will go through the concepts related to functions and the meaning of graphs of the functions.

Contents

©Mok Owen 2024

©Mok Owen 2024

©Mok Owen 2024

1 Function and its graph

'Functions describes the world!', one Professor in Mathematics of Massachusetts Institute of Technology (a.k.a. MIT) said that. His speech was greatly influential, as I have never heard such conclusive thinking about functions. In fact, in the past few years, whenever I was studying in schools, my thought about functions is always only about projecting elements from oen set to another set, but what he said had a big impact to my knowledge about functions.

What function talks about, is a subjection of one elements to one another. It can be thought of as a pointing action started from element A to element B, which not so far away, if we could think of subjecting a lot of, a bunch of, or a list of, whatever, objects from one collection to another collection, and they can all be matched under this pointing action, then it is a so-called function.

For example, in an Indian factory, the production of food undergoes many different process. Those can all be called functions. Let say a raw material comes to the factory first, it then be put to a machine to chop into many small pieces. It is the chopping function inside the factory. Next, the chopped material will be put into a pool of yellowish-brownish liquid and be stirred by dirty hands. It is the Mixing function in the factory. After that, the liquid will be drained on the dirty floor and be stepped on by Indian workers so that they can be smell freshed. It is the flavouring function in the factory. Finally, it will be sold to stores, which is the selling function.

Another example is what our body does. We eat and drink, going down the digestive system, and we sit on a toilet. Although we stupid human knows nothing about how the digestive system works, we could still name the conversion from food to poops a digestion, which means the digestive function representing the process in our body.

So we know that function as an english word represents the naming of a process of conversion, it is the time to explore how Math functions works.

1.1 What is a Function?

A function is defined as follow:

Definition 1.1 (Function)**.** *Given an input x and an output y, a **function** is a relation between x and y so that we can write $y = f(x)$ to represent the relationship.*

Essential Practice 1.1.1. *Write down functions for the following input-output variables:*

1. *u as input and v as output;*

2. *b as input and a as output;*

©Mok Owen 2024

3. *n as input and 1 as output (which we call it a constant function);*

4. *x^2 as input and y as output;*

5. *xy as input and z as output;*

6. *2^x as input and k as output;*

7. *$\sqrt{p}$ as input and q as output;*

Remark. *It is notable that we may write functions as $y = g(x)$, $y = h(x)$, $y = d(x)$, ... as we want. The 'naming' of a function is always definitive and up to user's construction.*

We can also apply functions after functions. To do so, we have to talk about the following:

Definition 1.2 (Composite functions)**.** *Let f and g be functions such that f takes x as input and y as output, and g takes y as input and z as output. Then we can say there is a function h that takes x as input and z as output. In other words, h is a* ***composite function*** *such that $z = h(x) = g(f(x))$.*

Essential Practice 1.1.2. *Let f and g be functions such that $b = f(a), c = g(b)$. Write a function for a as input and c as output.*

Function as an input-output pair

We may now consider how functions carry things to things using arrow notations. We may use a stroked arrow $\mapsto$ to emphasis the carrying process. Let's take a look at the following examples.

Example. *Given a function that undergoes the process of adding one to the given element. We can say*

$$\ldots, 1 \mapsto 2, 2 \mapsto 3, 3 \mapsto 4, \ldots$$

In other words, we may write

$$x \mapsto x + 1$$

to generalize the process of adding one to the given element, which is how we usually write to describe any functions.

The example shows that we can generalize the process using algebraic notation, in which it is usually in the form of

$$x \mapsto f(x)$$

©Mok Owen 2024

with the function f. It is equivalent to say that

$$\ldots, 1 \mapsto f(1), 2 \mapsto f(2), \ldots$$

but what's important is how explicit the mapping process is done. More examples could be viewed to familiarize with it.

Example. *To describe a function undergoes the process 'multiplying the element by two and add one to it', we may examine that*

$$\ldots, 0 \mapsto 1, 1 \mapsto 3, 2 \mapsto 5, \ldots$$

so that it is equivalent to write

$$x \mapsto 2x + 1$$

as a generalization of the function. It is equivalent to write $f(x) := 2x + 1$ to emphasize that we shall call the function f as a naming for the given process, as long as we can write

$$x \mapsto f(x), f(x) := 2x + 1$$

Essential Practice 1.1.3. *Examine the following functions from 0 to 3, and generalize the function using algebraic notation. You may either choose writing $x \mapsto \square$ directly or $x \mapsto f(x), f(x) := \square$.*

1. *Multiply the element by 3 and then add 2 to it.*
2. *Divide the element by 10 and then Subtract 7 from it.*
3. *Multiply the element by a and then add b to it.*
4. *Squaring the element.*
5. *Multiply 3 to the square of the element, an then add 5 to it.*
6. *Add 1 to the element first, then multiply the square of the result by 6, and then add 1 to it.*
7. *Subtract 8 from the element first, then divide the square of the result by 2, and then add 4 to it.*
8. *Subtract h from the element first, then multiply the square of the result by a, and then add k to it.*

It is notable that a function can only give one output for each input, which makes the next page a fruitful discussion.

©Mok Owen 2024

Defining a function with variables

So far, we have learned how writing a generalization of a function is, and have we used algebraic notation to shorten the examination, one suggest we can always write algebraic notation for a function definition. In addition, we also find that a function cannot have more than one output, as we see functions as a pointing process from one element to another element. So we have the following definition:

Definition 1.3 (Function)**.** *A* ***function*** f ***of*** x *defines the pointing process* $x \mapsto f(x)$ *is a one-to-one pointing process, which can have only one output.*

It is important to note that $f(x)$ is a function of x if and only if one x produce one $f(x)$. The x is called a *dummy variable*, which is a variable that can be changed all the time. For example, writing $f(y)$ or $f(z)$ still makes sense to say a function f, but not of x.

From now on, we can determine whether a given relation is a function or not.

Example. *Given the relation* $y = mx + c$*. Since one* x *can produce only one* y*,* y *is a function of* x*; on the other hand, we also see one* y *produces only one* x*, so* x *is also a function of* y*.*

Example. *Given the relation* $y = x^2$*. Since one* x *can produce only one* y*,* y *is a function of* x*; however, one* y *may produce more than one* x*, say if* $y = 4$ *then* x *can be* 2 *or* -2*, so* x *is not a function of* y*.*

Essential Practice 1.1.4. *Determine whether the following given relation between* x *and* y *is a function of one another or not. Provide counterexample if it is not a function.*

1. $y = -x$;
2. $y = 4x + 3$;
3. $y = \frac{1}{x}$;
4. $y = \frac{x}{6}$;
5. $y^2 = x$;
6. $y^2 = x^2$;
7. $y^3 = 4x^2 - 3$.

©Mok Owen 2024

Domain, Co-domain and Range

For explicit definition of a function, we need the following concepts to help with: *domain*, *co-domain* and *range*.

A **domain** is where the input comes from, which is usually half-customized and half-restricted. For example, $f(x) = \frac{1}{x}$ can have input of negative real numbers, positive real numbers, any complex numbers except 0. This means that 0 is naturally restricted by the operation of $\frac{1}{x}$, but other than 0, we can choose freely our input from all complex numbers. Thus, the largest domain of $f(x) = \frac{1}{x}$ is all complex numbers except 0. However, it is not saying that the domain of $f(x) = \frac{1}{x}$ must be all complex numbers except 0, we can still put restrictions on our own, what means by customize, like all real numbers except 0 or all positive real numbers except 0 as its domain, is still a possible choice. Hence, we shall usually talk about the *greatest possible domain of a function* if we need to find the natural restrictions, and the *domain of a function* if we are going to define our source of input.

Essential Practice 1.1.5. *Find the greatest possible domain of the following functions if (i) the output is restricted to complex numbers and (ii) the output is restricted to be real numbers:*

1. $f(x) := x$;
2. $f(x) := ax + b$;
3. $f(x) := \frac{a}{x}$;
4. $f(x) := x^2$;
5. $f(x) := a(x - h)^2 + k$;
6. $f(x) := \sqrt{x}$;
7. $f(x) := \sqrt[3]{x}$;
8. $f(x) := \frac{1}{\sqrt{x}}$;

A **co-domain** is where the output can go to. It is more likely a limitation of the output of the function so that we know where our target is. Similarly, we shall usually talk about the *greatest possible co-domain of the function* if we need to find the natural restrictions, and the *domain of the function* if we are going to define our target output.

Essential Practice 1.1.6. *Find the greatest possible co-domain of the following functions if the input is unrestricted:*

1. $f(x) := x$;
2. $f(x) := ax + b$;
3. $f(x) := \frac{a}{x}$;
4. $f(x) := x^2$;
5. $f(x) := a(x-h)^2 + k$;
6. $f(x) := \sqrt{x}$;
7. $f(x) := \sqrt[3]{x}$;
8. $f(x) := \frac{1}{\sqrt{x}}$;

With domain and co-domain, we can now define a function in a more explicit manner. Writing a function with where the inputs are in and where the outputs to go, we have a nice notation - an arrow $\to$ emphasizing the direction from domain to co-domain. In general, we will write

$$f : D \to R$$

to specify the function f is a function goes from domain D to co-domain R. The formal way to define a function is like below:

$$f : \mathbb{R} \to \mathbb{R}, x \mapsto f(x)$$

which reads 'a function f sending a real number x to a real number $f(x)$'.

Example. *For a function sending a natural number n to a natural number 2^n, we may write its definition as*

$$f : \mathbb{N} \to \mathbb{N}, n \mapsto 2^n$$

Example. *For a function sending an integer n to a rational number 2^n, we may write its definition as*

$$f : \mathbb{Z} \to \mathbb{Q}, n \mapsto 2^n$$

We shall see both example shows the same function process $f(n) := 2^n$ but different domain and codomain. Therefore, we acknowledge that although both functions are having the same process, they are indeed representing different things, which yields they are in fact different functions.

From this point of view, we can further examine the so-called **range** of a function, which is the exact target region of the function under the codomain. We shall define some set notation to present its meaning.

©Mok Owen 2024

Definition 1.4 (Union and Intersection)**.** *Let A and B be sets. The **union** of A and B is defined as*

$$A \cup B := \{x : x \in A \text{ or } x \in B\}$$

*while the **intersection** of A and B is defined as*

$$A \cap B := \{x : x \in A \text{ and } x \in B\}$$

In fact, we say union is the collection of the objects which are either in set A or set B, or simply say it is the joined set of two sets. We can take a look at the following examples:

Example. *Let $A = \{1, 2, 3, 4\}, B = \{3, 4, 5\}$, then*

$$A \cup B = \{1, 2, 3, 4, 5\}$$

Example. *Let $A = \{1, 3, 5, 7, 9, \dots\}$ be the set of all positive odd numbers, $B = \{2, 4, 6, 8, 10, \dots\}$ be the set of positive even numbers, then*

$$A \cup B = \{1, 2, 3, 4, 5, 6 \dots\}$$

which is the set of all positive numbers. Sometimes, we may write the set by specifying the property of the elements as following:

$$A \cup B = \{x : x \text{ is a positive integer}\}$$

For intersection, it is generally speaking the collection of repeated elements in both sets, or we can say the sharing elements. We can take a look at the following examples:

Example. *Let $A = \{1, 2, 3, 4\}, B = \{3, 4, 5\}$, then*

$$A \cap B = \{3, 4\}$$

Example. *Let $A = \{1, 3, 5, 7, 9, \dots\}$ be the set of all positive odd numbers, $B = \{2, 4, 6, 8, 10, \dots\}$ be the set of positive even numbers, then*

$$A \cap B = \emptyset$$

which is the set with no element, an empty set.

We may represent the two definitions by shading regions in a Venn-diagram. Suppose we are calling a set by enclosing the region by a circle, then we have the following representation.

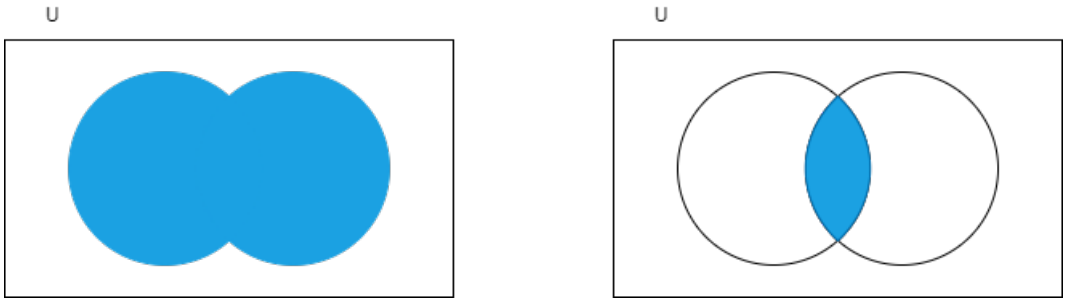


Figure 1: Union(Left) and Intersection(Right) of two sets

Essential Practice 1.1.7. *Let* $A = \{1, 3, 4, 6, 7\}, B = \{3, 4, 5, 7, 8\}$*, then find* $A \cup B$ *and* $A \cap B$*.*

Essential Practice 1.1.8. *Prove the following identities:*

1. $A \cap (B \cup C) \equiv (A \cap B) \cup (A \cap C)$;
2. $A \cup (B \cap C) \equiv (A \cup B) \cap (A \cup C)$.

Definition 1.5 (Range)**.** *The* ***range*** *of a function* $f : D \to R$*, denoted by* $\mathbf{Ran}(f)$*, is the set* $f(D) \cap R$*, where* $f(D) := \{f(x) : x \in D\}$*.*

Usually, a teacher in high school aims to discuss the range of a function rather than the co-domain of a function, as what he shall teach is the size of the output, but not where the output shall be in. However, the difference between the co-domain of a function and the range of a function is we can further define the concept of a well-defined function with the concept of range.

Example. *For a function* f *defined as*

$$f : \mathbb{N} \to \mathbb{N}, n \mapsto 2^n$$

The range of f*, denoted by* $\mathrm{Ran}(f)$*, is the set of all possible outcomes of* $f(n)$ *with natural numbers* n*. That is, the set*

$$\{2^n : n \in \mathbb{N}\} = \{2, 4, 8, 16, 32, \dots\}$$

Example. *For a function* f *defined as*

$$f : \mathbb{Z} \to \mathbb{Q}, n \mapsto 2^n$$

$\mathrm{Ran}(f)$ *is the set of all possible outcomes of* $f(n)$ *with integers* n*. That is, the set*

$$\{2^n : n \in \mathbb{Z}\} = \{\dots, \frac{1}{4}, \frac{1}{2}, 1, 2, 4, \dots\}$$

©Mok Owen 2024

We must observe that the two ranges of the same function process f are different when their co-domains are different. This shows the importance of discussion of co-domain when we are defining ranges of functions.

Essential Practice 1.1.9. *Find the range of the following functions:*

1. $f : \mathbb{N} \to \mathbb{N}, n \mapsto n$;
2. $f : \mathbb{N} \to \mathbb{N}, n \mapsto 3^n$;
3. $f : \mathbb{N} \to \mathbb{N}, n \mapsto 3^n - n^3$;
4. $f : \mathbb{Z} \to \mathbb{Z}, n \mapsto n$;
5. $f : \mathbb{Z} \to \mathbb{Z}, n \mapsto n^2$;
6. $f : \mathbb{Z} \to \mathbb{Q}, n \mapsto 5^n$;
7. $f : \mathbb{Q} \to \mathbb{Q}, n \mapsto n$;
8. $f : \mathbb{Z} \to \mathbb{Q}, n \mapsto n/2$;
9. $f : \mathbb{N} \to \mathbb{Q}, n \mapsto n/10$;
10. $f : \mathbb{Z} \to \mathbb{R}, n \mapsto \pi n^2$;

Essential Practice 1.1.10. *Are the functions in previous practice all well-defined? Which of them are not?*

1.2 Graph of a function

Graphing has been an essential skill in understanding mathematical objects. We usually say it is a mathematical modeling technique. Through graphing, we can see the relationship between the input and output clearly, whether they are related, increasing or decreasing. It is also easier to draw conclusion to estimations by valid graphs. This section aims to build the concept of representing functions by graphs.

Blobs-and-arrows diagram for discrete functions

We shall first take a step backward to a simpler function - a *discrete function* with finite inputs. This will help the construction very much.

Definition 1.6 (Discrete functions)**.** *A **discrete function** is a function with direct indication of the function process for each element in domain. That is, for each element $x \in D$, the output $f(x)$ is assigned to co-domain directly.*

A discrete function is usually with a discrete domain, and random assignment.

Example. *Let $D := \{1, 2, 3, 4, 5\}$ be the domain of a function f. Suppose f is defined by*

$$f(x) := \begin{cases} 1 & , x = 1 \\ 4 & , x = 2 \\ 3 & , x = 3 \\ 2 & , x = 4 \\ 5 & , x = 5 \end{cases}$$

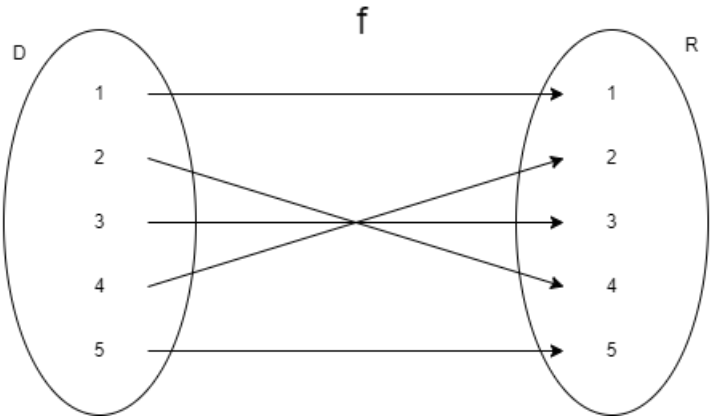
In this case, f is a discrete function.

To represent the above function using a graph, it is recommended to use a *blobs-and-arrows* diagram.

Definition 1.7 (Blobs-and-arrows diagram)**.** *A **blobs-and-arrows** diagram is a diagram representing a function by denoting the domain and co-domain by two circles, and for each element in the circle of domain, a pointer arrow over-set with f pointing to one of the element in the co-domain, to emphasize the relation between the two elements are input and output pair.*

It is undesired to read through such complicated definition of the diagram. Let's see the following example.

Example. *Recall the function in previous example, we could draw the blobs-and-arrows diagram as shown.*

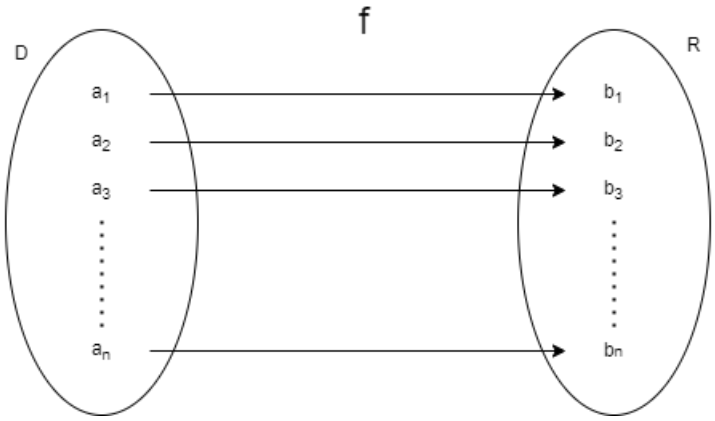


©Mok Owen 2024

In the figure above, the name of the function is hung over the main diagram, while the left ellipse denote the set of elements of domain D and the right ellipse denote the range of elements of codomain R. The letter R could also be interpreted as the range of f.

For a discrete function with finitely many inputs, where we will take them as a sequence, it can also be used to represent it efficiently. We just need to have some modification.

Axiom. *Let $f : D \to R$ be a function. Let $a_1, a_2, a_3, \ldots, a_n$ be a sequence of numbers in D and $b_1, b_2, b_3, \ldots, b_n$ be a sequence of numbers in R. Suppose $b_1 = f(a_1), b_2 = f(a_2), \ldots, b_n = f(a_n)$ defines the one-to-one correspodence from D to R. Then the following blobs-and-arrows diagram describes f formally:*



Essential Practice 1.2.1. *Draw the blobs-and-arrows diagram for the following discrete functions:*

1. $f(1) = 3, f(2) = 6, f(3) = 7, f(4) = 2, f(5) = 3.$
2. $f(x) := \begin{cases} 1 & , x = 1, 2, 3, 4, 5 \\ 2 & , x = 6, 7 \\ 3 & , x = 8, 9, 10 \\ 4 & , x = 11, 12, 13, 14, 15 \\ 5 & , x = 16, 17, 18, 19 \\ 6 & , x = 20 \end{cases}$

The pairing table and xy-coordination

To extend our concept of function from discrete version to continuous version, that is, infinitely many numbers can be plugged into the function for calculation, we may need to 'fill up' the holes of the domain.

Let say we have 1 and 2 in the domain D, then all the real numbers within 1 and 2 also in D. This is the process of **continuation** or **extension**, and the whole set of real numbers within 1 and 2 is called the **interval** between 1 and 2. We can denote the interval by $[1, 2]$ if 1 and 2 are included in the meaning, or $(1, 2)$ if 1 and 2 are excluded from the meaning. To better distinct the difference between $[1, 2]$ and $(1, 2)$, we call the previous one the **closed interval** between 1 and 2, while the latter one the **open interval** between 1 and 2.

From this point of view, a real number line is used to represent the concept of any intervals. Let's define the real number line and discuss intervals using a number line.

Definition 1.8 (Real number line)**.** *A* ***real number line*** *is a straight line with ordered property so that one of its direction is increasing, and the other direction is decreasing.*

We usually draw a horizontal real number line for convenience.

For the sake of simplicity and connected-understanding of open and closed intervals, we may put the desired parenthesis in the real number line to denote the meaning. Following that a closed interval is in a block-typed while open interval is in a round-typed, we can directly place them on the real number line.

Example. *The open interval* $(0, 1)$.

Example. *The open interval* $(-1.5, 5.5)$.

Example. *The closed interval* $[0.5, 1]$.

Example. *The closed interval* $[-100, -67]$.

Essential Practice 1.2.2. *Draw the mentioned interval on the provided real number line with suitable parenthesis.*

1. $[0, 1]$.
2. $[\alpha, \beta]$.
3. $(0, 1)$.
4. (α, β).

©Mok Owen 2024

It is of course we could draw it vertically with the same meaning, by indicating upward as positive direction and downward as negative direction. Consider a function $y = f(x)$ defines the relation between x and y coordinate, we can combine two real number lines in an orthogonal way, like a cross, to present the coordination in a sensible way. This is called the **continuous graph of a function**, or we simply call it the *graph of a function* if there's no ambiguity of what we are discussing.

In order to draw the graph of a function, we need to figure out the valid pairs of x and y first, so that the points could be addressed correctly for the function. We will make use of a **pairing table** so that each value of x in the domain is paired with a unique value of y in the co-domain. We will show its use with some examples.

Example. *Let G be the graph of $y = x + 1$. Then we have the table of testing xy-pair*

x	*0*	*1*	*2*	*3*	*4*
$y = x + 1$	*1*	*2*	*3*	*4*	*5*

and thus the graph of G be like

by linking up the points with straight lines.

Example. *Let G be the graph of $y = x^2$. Then we have the table of testing xy-pair*

x	*-2*	*-1*	*0*	*1*	*2*
$y = x^2$	*4*	*1*	*0*	*1*	*4*

and thus the graph of G be like

by linking up the points with straight lines. Moreover, by increasing the points of testing value to, say, 100 or 1000, a see-smoothing curve can be drawn by computer.

It is the continuation of the curve under testing value control.

Even higher power polynomial functions can be drawn using this strategy.

Example. *Let G be the graph of $y = x^3 + 2x^2 + x - 1$. Then we have the table of testing xy-pair*

x	*0*	*1*	*2*	*3*	*4*
$y = x^3 + 2x^2 + x - 1$	*-1*	*3*	*17*	*47*	*99*

and thus the graph of G be like, with many enough testing points,

with the continuation.

Essential Practice 1.2.3. *Configure the xy-table with testing values in between -3 to 3 and draw a suitable graph for the following functions.*

1. $y = x - 3$;
2. $y = x^2 + x + 1$;
3. $y = (x - 3)^3 + 2$.

You may also check the graph using WolframAlpha desktop to convince yourself with the plot. Try to conclude the difference between large amount and small amount of points of testing.

With the xy-coordination of a graph, many features and prediction could be done. We will try to analyze the properties that we could make use of in the next subsection.

1.3 Fundamental properties of a graph

A graph shows the relationship between two variables, namely the vertical component (in this case, it is the value of y) and the horizontal component (respectively, the value of x). We will make use of this relationship and consider different situations.

points lying on the graph

Let us reverse the sight of plotting a graph of a function. We plot the graph by linking up the points we get from inputting x into the function to get the corresponding value of y, so that each point (x, y) describes when the function has the input x. We could now see that if a point (x, y) is passed through by a curve then it is the same as the point lies on that curve. We shall say the point (x, y) **satisfies** the equation $y = f(x)$, since this is the definition of a graph.

Theorem. *For any real-valued continuous functions f, any coordinates (x, y) lies on the graph $y = f(x)$ if and only if the xy-pair satisfies the equation $y = f(x)$.*

Example. *For a given function $f(x) = 3x + 1$, the following are true:*

1. *The point $(0, 1)$ lies on $y = f(x)$;*
2. *The point $(3, 10)$ lies on $y = f(x)$;*
3. *The point $(-2, 1)$ does not lay on $y = f(x)$;*
4. *The point $(-10, 2)$ does not lay on $y = f(x)$.*

Example. *For a given function $f(x) = 3x^3 + 2x^2 + x + 1$, the following are true:*

1. *The point $(0, 1)$ lies on $y = f(x)$;*

2. *The point* $(-1,-1)$ *lies on* $y=f(x)$*;*

3. *The point* $(-2,0)$ *does not lay on* $y=f(x)$*;*

4. *The point* $(-10,2)$ *does not lay on* $y=f(x)$*.*

Essential Practice 1.3.1. *For a given function* $f(x)=x^2-1$*, show whether the following points are lying on the graph of* $y=f(x)$*:*

1. *The point* $(0,1)$*;*

2. *The point* $(-1,0)$*;*

3. *The point* $(-2,3)$*;*

4. *The set of points* $(-t,t^2-1)$ *parametrized by the variable* t*.*

We may now examine another feature of a graph, namely **the region of inequality**. In the previous paragraph we find a point satisfying the equation $y=f(x)$ has the same meaning as the point lies on the curve. If we make a small translation of the y coordinate to the upper region separate by the curve, it becomes that all the points have a larger value of y than we could calculate by the function. For that region we will call it the region for $y>f(x)$; Similarly, if we make a small translation of the y coordinate to the lower region, it becomes that all the points in that region have a smaller value of y than we could calculate by the function. For that region, we will call it the region for $y<f(x)$.

Example. *For a given function* $f(x)=3x+1$*, the following are true:*

1. *The point* $(0,2)$ *lies above* $y=f(x)$*;*

2. *The point* $(3,9)$ *lies below* $y=f(x)$*;*

3. *The point* $(-2,1)$ *lies above* $y=f(x)$*;*

4. *The point* $(-10,-120)$ *lies below* $y=f(x)$*.*

Example. *For a given function* $f(x)=3x^3+2x^2+x+1$*, the following are true:*

1. *The point* $(0,3)$ *lies above* $y=f(x)$*;*

2. *The point* $(-1,-2)$ *lies below* $y=f(x)$*;*

3. *The point* $(-2,0)$ *lies above* $y=f(x)$*;*

4. *The point* $(10, 2)$ *lies below* $y = f(x)$.

Essential Practice 1.3.2. *For a given function* $f(x) = x^2 - 1$*, determine the position of the following points with respect to the graph of* $y = f(x)$*:*

1. *The point* $(0, 9)$*;*
2. *The point* $(-1, -1)$*;*
3. *The point* $(-2, 20)$*;*
4. *The point* $(10, 10)$.

Special intersections: axis intercepts

Usually the most important element of a graph is whether it cuts the coordinating axis, namely the y-axis and the x-axis. For practical reason we value them so much and we may give them special names: the y-intercept and the x-intercept.

Definition 1.9 (y-intercept of a graph)**.** *Given a function* f*, the function has its* ***y-intercept*** *when its cuts the y-axis. In other words, the y-intercept is defined as*

$$y_0 := f(0)$$

whenever the function is well-defined at $x = 0$.

Similarly, we may want the same definition for x-intercept of a graph, however, we do not know whether the inverse function exists or not. This comes to some kind of difficulties to say directly the x-intercept. But still, we can give the possible condition for the discussion.

Definition 1.10 (x-intercept of a graph)**.** *Given a function* f*, the function has its* ***x-intercept*** *when its cuts the x-axis. In other words, the x-intercept is defined as the roots of*

$$0 = f(x)$$

whenever the given equation is solvable.

We will be familiar with them with some valid examples.

Example. *For the graph of the function* $f(x) := x + 1$*, we have:*

- *y-intercept:* $y_0 = 0 + 1 = 1$*;*

- *x-intercept: Set* $0 = x + 1$*, then* $x_0 = -1$ *is the root to the given equation, which is the x-intercept of the function.*

Example. *For the graph of the function* $f(x) := x^2 - 1$*, we have:*

- *y-intercept:* $y_0 = 0^2 - 1 = -1$*;*
- *x-intercept: Set* $0 = x^2 - 1$*, then* $x_1 = -1$ *or* $x_2 = 1$ *are the roots to the given equation, which are the x-intercepts of the function.*

Example. *For the graph of the function* $f(x) := x^2 + 1$*, we have:*

- *y-intercept:* $y_0 = 0^2 + 1 = 1$*;*
- *x-intercept: Set* $0 = x^2 + 1$*, which has no real solution. Then the graph of the function has no x-intercepts.*

We see a function has its y-intercept most of the time, but not the same case for x-intercepts. It could have either no x-intercepts, only 1 x-intercept, or more than one x-intercept. It is a consequence of the fundamental theorem of algebra, but we will dig deeper in the future to discuss simple cases we could.

As long as we could plot not only the xy-coordinate but we could pair up something weird, where they could be different functions say $v(y)$ and $u(x)$, the graphs could no longer talked about y-intercept or x-intercept - they are not exactly about x and y! Hence, we turned the naming into something general: we plot the graph by vertical axis and horizontal axis, so let us call them the **vertical intercept** and the **horizontal intercept** respectively.

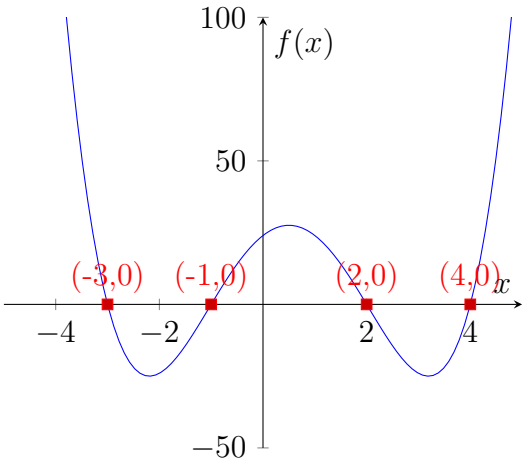
1.4 Solving equations using graphs of functions

Solving equations has been a fundamental skill to highschool students, and it has been an important concept to play across different area. One extended concept about solving equation is referred to some geometric interpretations, and we call it solving equations by graphs.

Homogeneous equations

Let us consider a simple case first. We define the term **Homogeneous** for the situation of **zero equality**, i.e. Let f be a function and $G := \{(x, y) : y = f(x), x \in \mathbb{R}\}$, the homogeneous equation of f is the equation $f(x) = 0$.

To solve a homogeneous equation, we consider the graph G in which it states the equality from where y values. We shall put $y = 0$ to see where the equality holds. In fact, we see the solution at the x-axis - they are the roots of G.



In the above example, the roots of the graph of the function $f(x)$ are the coordinates

$$\{(-3, 0), (-1, 0), (2, 0), (4, 0)\},$$

which we shall see their y-coordinates are all zero. This satisfies and due to the method we chose on the graph observation - adding the line $y = 0$, which is in fact the x-axis.

Let us proceed to higher dimension, in order to generalize the concept of homogeneity. Here we will introduce **multivariated function** $f : \mathbb{R}^n \to \mathbb{R}$ so that we have the following equation

$$f(x_1, x_2, \ldots, x_n) = 0$$

In 2-dimensional sense, we have to consider

$$f(x, y) = 0$$

which is quite confusing for newbies. One question about this type of homogeneous equation is that how to compute it solution, since most of our concepts are related to visualization. In fact, we have no simple clue to solve such abtract equation, but we should think of how can it be solved.

Let $f(x, y) := x^2 + y^2$, then the problem

$$f(x, y) = 0$$

becomes

$$x^2 + y^2 = 0$$

which has only one real solution pair $(x, y) = (0, 0)$. To show this guess is true we can do a two-step verification.

Proof of $(0,0)$ *is the only real solution to the equation* $x^2+y^2=0$.
We first claim for the existence of the solution. It is simple, as we need only to substitute the pair into the equation:

$$LHS = 0^2 + 0^2 = 0$$
$$RHS = 0$$

Then we have to show the uniqueness of the solution. By rearranging terms of the equation, we have

$$y^2 = -x^2$$

which is critical for determination. We have $x^2 \geq 0$ by real solution requirement, so as $y^2 \geq 0$. From the rearranged equation, we have also $y^2 \leq 0$. The only interception of the condition is where $y^2 = 0$, so as x^2. The result follows. □

In the example, we have let f to be equal to zero. Such solution is called a **zero level set** of f. For any other constant k such that $f(x,y) = k$, the solution to the equation will be called a k-level set of f. We will see more about it in the next part.

Non-homogeneous equations

For the case of non-homogeneous equations, we have two ways to solve the problem. One is by using the homogeneous case, another is by graph directly.

Now set the problem to be $f(x) = k$, where f is a function of x and k is a constant independent from x. The mechanism behind the nonzero equality is the manipulation of equation. We can move k to the left side so that

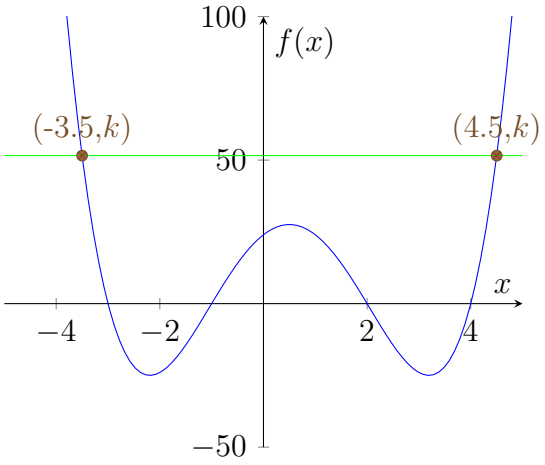
$$f(x) - k = 0$$

and we reach the homogeneous condition for

$$h(x) = 0$$

if we let $h(x) = f(x) - k$. This gives us a critical view that we can always rearrange terms to construct homogeneous problems for solving.

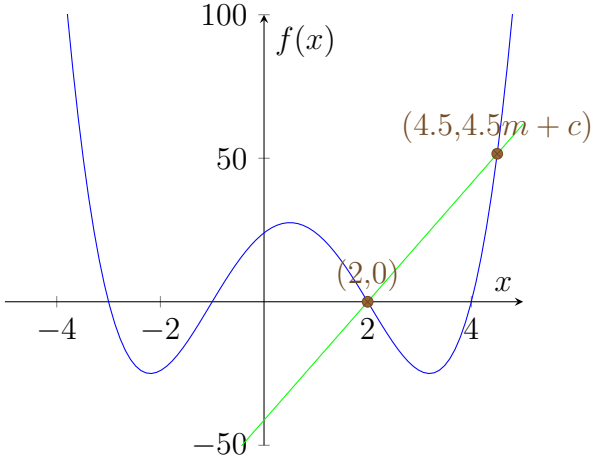
However, we are assumed to know nothing about transformation of graphs, that hold us back from this approach. Let us think of adding a line to complete the equality. For $f(x) = k$ we shall add the line $y = k$ to observe its intersection with $y = f(x)$. The graph shows the condition is in fact similar to that of homogeneous one.



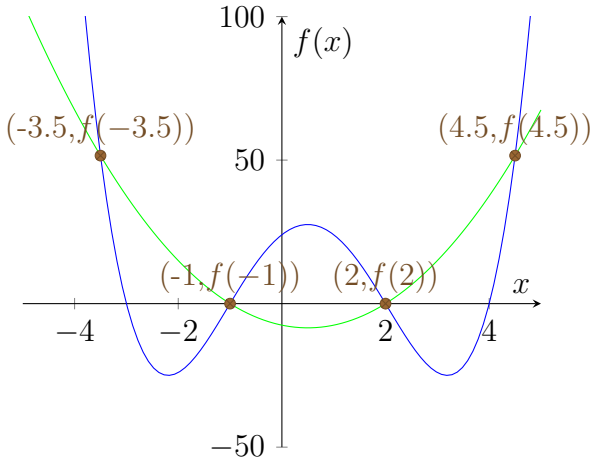
In fact, we choose to add a line rather than move the whole graph to another position as we cannot explicitly draw a new graph representing the resultant graph. We can have a new plot, but it is usually implicit due to the computation of roots.

What's more in non-homogeneous case, the homogeneity can be tuned as described above, although it can be time-consuming, the concept worth dive into. That will be a case in the next part. Now we do something more to familiarize with graphical inception.

Suppose we have another equation $f(x) = mx + c$, where m defines the slope of the straight line and c is its y-intercept. Could we still follow the graphical inception to find a solution set? The answer should be positive. Indeed, we may draw the oblique line on the graph as previous work.



For any functions f and g, if we need to find their equation, it is obvious to follow the graphing method as we have done above.



In order to extend to higher dimensional problems, we recall the concept of level sets, which we used it for homogeneous condition. In fact, if any non-homogeneous condition can be converted to homogeneous condition, the methodology holds trivially.

Consider a simple case $f(x) = k$, we know that by subtracting k on both sides, we have $f(x) - k = 0$. Let $h(x) := f(x) - k$, then the condition becomes solving

$$h(x) = 0$$

The conversion means that we may see the k-level set of f as a zero level set of h with suitable substitution. This provides a huge intuition for general situation.

Let us think of the last discussed non-homogeneous equation. Suppose $f(x)$ and $g(x)$ are functions of x and we are going to consider $f(x) = g(x)$. We rearrange them into

$$h(x) := f(x) - g(x) = 0$$

so that again a homogeneous equation appears. Note that we didn't resrtict f and g to be any fixed type function, that means the methodology is widely applicable.

Moreover, if we are giong to higher dimensional problem, where we have to solve for at least two-dimensional homogeneous equation $f(x, y) = 0$, a powerful tool can be imagined by reversing the thought we discussed.

Theorem (Implicit function theorem)**.** *Let $f : \mathbb{R}^n \to \mathbb{R}$ be an n-dimensional function and suppose $f(\mathbf{x}_0, y) = 0$. Consider a smooth neighbourhood $N_\varepsilon(\mathbf{x}_0)$ with $\varepsilon > 0$ such that $\forall \mathbf{x} \in N_\varepsilon(\mathbf{x}_0)$, $|\mathbf{x} - \mathbf{x}_0| < \varepsilon$. Then $\exists \varphi : N_\varepsilon(\mathbf{x}_0) \to \mathbb{R}$ such that $y = \varphi(\mathbf{x})$ for $\mathbf{x} \in N_\varepsilon(\mathbf{x}_0)$.*

The proof requires higher level mathematics that inherits the concepts of differentiability. Let me write it down as understandable as I could for interested readers. You may also believe in it without knowing the proof, since it is the level of pure mathematcian.

©Mok Owen 2024

Proof.
Define $\dfrac{\partial}{\partial x_i}$ to be the partial derivative operator with respect to x_i only, i.e. Any other variables that is not equivalent to x_i is seen constant. From this if we let $\{x_1, x_2, \ldots, x_n\}$ be a set of orthogonal basis, we may see

$$\frac{\partial x_j}{\partial x_i} = \delta_{ij}$$

and by $f(\mathbf{x}_0, y) = 0$ such that

$$0 = \mathrm{d}f = (\sum \frac{\partial f}{\partial x_i}\mathrm{d}x_i) + \frac{\partial f}{\partial y}\mathrm{d}y$$
$$\frac{\mathrm{d}y}{\mathrm{d}x_i} \neq 0 \qquad \forall i$$

Then There is a linearization for y in terms of x_i within $N_\varepsilon(\mathbf{x}_0)$ such that

$$\varphi(\mathbf{x}) = \sum \frac{\mathrm{d}y}{\mathrm{d}x_i} x_i$$

For a small enough ε, such function $\varphi(\mathbf{x})$ can be computed that

$$|f(\mathbf{x}, y) - f(\mathbf{x}, \varphi(\mathbf{x}))| \approx 0$$

□

The implicit function theorem (where I stated a simpllified version) provides us a convincing basis of giving a 'good approximation' on a local region of a surface. That means whenever we need to consider higher dimensional eqiuations, we have another view for solving level sets with partial coordinates. It also provides a valid transformation between geometric problem and algebraic problems.

Simultaneous equations

The most general case of equation is usually called a system of equations. A system is usually identified when there are correlation between equations, i.e. they are required to be solved at the same time. A textbfsimultaneous equation is a specific form of a system that inherits only a pair of equations, and that provided a fundamental case for the development of solving skills.

Let us review a first degree duo variable simultaneous equation that we have learnt in junior secondary.

Let a, b, c, d, m, n be some known constants and x, y be unknown variables. The first degree

duo variable simultaneous equations is in the form of

$$\begin{cases} ax + by &= m \\ cx + dy &= n \end{cases}$$

and indeed with solutions

$$\begin{cases} x &= \dfrac{dm - bn}{ad - bc} \\ y &= \dfrac{cm - an}{bc - ad} \end{cases}$$

If we generalize the situation to any degree of duo variables, we may consider

$$\begin{cases} f(x, y) &= m \\ g(x, y) &= n \end{cases}$$

where f, g are functions of x adn y, and m, n are constants.

1.5 Transformation of functions

Let us revisit the fundamental of transformation, pointwise transformation, so that we have the solid foundation for what transformation of function does.

Point Translation

Pick a point (a, b) on the rectangular coordinate plane, we want to perform different kinds of movement as a simplified information of a map. We call a movement a **translation** if it is a directed displacement. Moreover, if a translation is parallel to the vertical axis of the plane, i.e.

$$(a, b) \to (a, b + k)$$

for some constant k, we say that such translation is a **vertical translation**; while the one parallel to the horizontal line is a **horizontal translation**, i.e.

$$(a, b) \to (a + h, b)$$

for some constant h. They are just some specified names for certain types of fundamental intuition.

Example. *Let a point* $(1, 2)$ *on the rectangular plane. If we first translate the point vertically upward by 5 units, then horizontally rightward by 2 units, the resulting coordinates will be* $(3, 7)$*; if we further translate the point vertically downward by 2 units, then horizontally leftward by 4 units, then vertically downward by 3 units, and finally horizontally rightward by 2 units, the resulting coordinates will be at* $(1, 2)$*, which is the starting point.*

The example shows that translation is invertible, and an useful observation on translation can be drawn:

Theorem. *The order of translation is neglegible. In formal words, translation is commutative.*

We see that the order of translation does not affect the resulting coordinates of translation, that means it is essential to combine many steps of ranslation into one step. We call such technique a **vector operation**.

Definition 1.11 (Vector)**.** *The ordered pair (x, y) defines a pair of translation horizontally by x units and vertically by y units respectively. Such pair is called a vector and vector addition defines the manipulation of translation along both axes.*

It is easy to check that moving along x-axis (resp. y-axis) by h units (resp. k units) is equivalent to writing $+(h, 0)$ (resp. $+(0, k)$). The addition

$$+(h, 0) + (0, k) = +(h, k)$$

shows the commutativity of translation. Moreover, a vector can be an extended coordination method by its format: we may see the notation of vector as the traslation from origin, so that coordinates agrees with vector. We define vector addition by

$$(a, b) + (h, k) = (a + h, b + k)$$

which is a bi-linear operation.

Point Rotation

A point rotation is usually defined by a planar rotation about the origin anti-clockwisely. Why is it so? It is due to a convention to polar coordinates.

©Mok Owen 2024

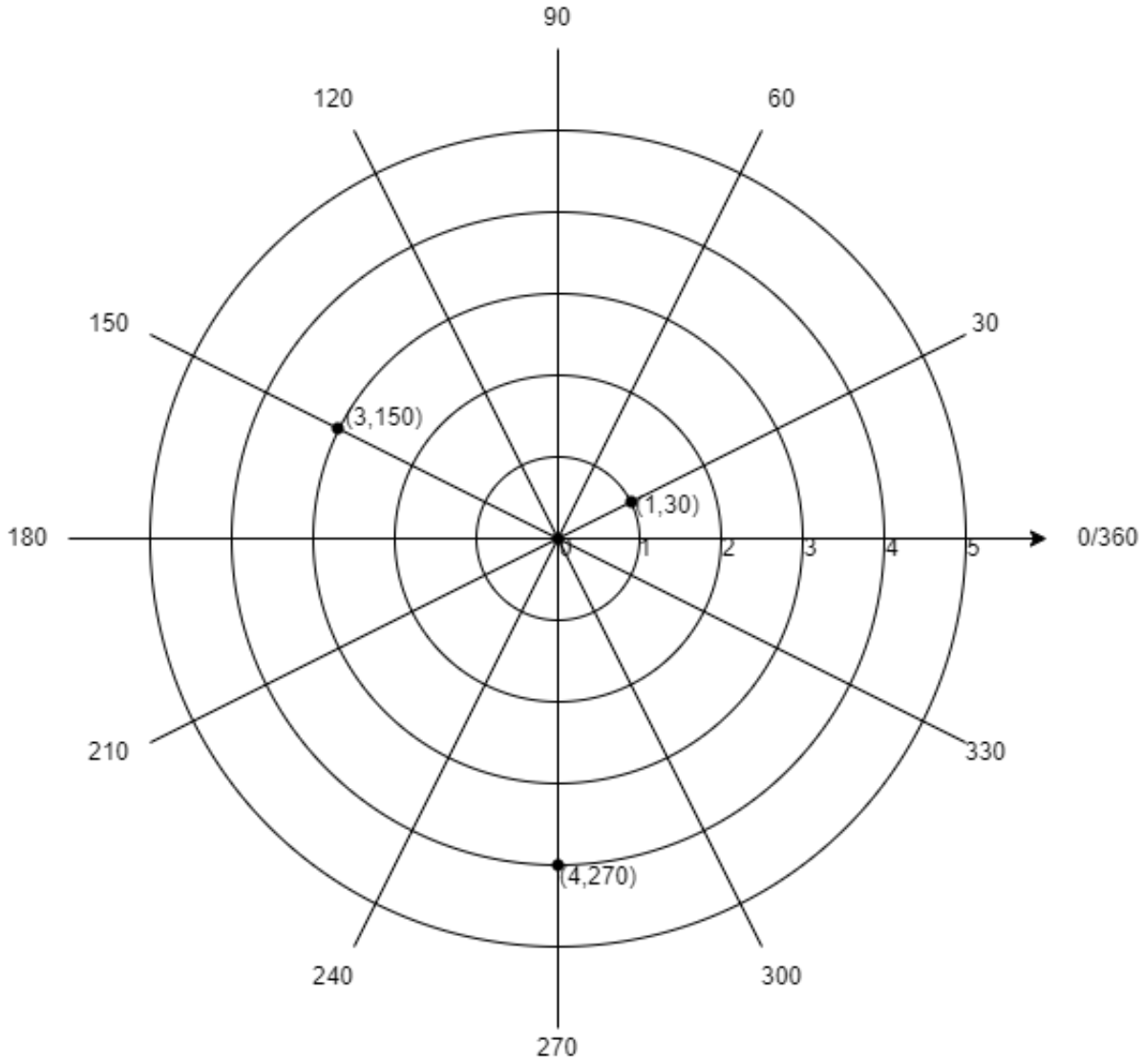


Figure 2: Polar plane with coordinates (r, θ)

As in the figure, a pair of polar coordinates consists of two parts, the length and the degree of rotation. The degree of rotation is accounted anti-clockwisely. Therefore, we made the convention in this way.

So, for a rotation of 90 degrees about origin anti-clockwisely, we see the following:

1. the positive x-axis part are rotated and moved to the positive y-axis part;
2. the positive y-axis part are rotated and moved to the negative x-axis part;
3. the negative x-axis part are rotated and moved to the negative y-axis part, while;

©Mok Owen 2024

4. the negative y-axis part are rotated and moved to the positive x-axis part.

Our recitation on anti-clockwise rotation can be done by observing the general pattern:

$$(x, y) \mapsto (-y, x)$$

for every 90 degree anti-clockwise rotation. Define the rotation function $\mathcal{R}$ such that $\mathcal{R}(x, y) = (-y, x)$, we shall call $\mathcal{R}$ a **rotor**.

On the other hand, considering a clockwise rotation by 90 degrees about origin as an anti-clockwise rotation by 270 degrees about origin, we can write

$$(x, y) \mapsto (y, -x)$$

to be a clockwise rotation by 90 degrees.

Not only rotation of specific angles can be done. If you wish, a rotation with any degrees can be represented by a specific operation. We now observe that for any rotation, the length between the rotated coordinates and the origin is exactly the same as that before the rotation, which is due to the definition of rotation. It follows the discussion on trigonometry that we can generalize the condition as below:

Let (x, y) be the original coordinates and assume the coordinates is equivalent to (r, θ) in polar coordinates. A **φ-rotation** is defined by applying $\mathcal{R}_\varphi : (r, \theta) \mapsto (r, \theta + \varphi)$ and we can write the rotation result as follows:

$$\begin{aligned}
\mathcal{R}_\varphi(x, y) &= \mathcal{R}_\varphi(r\cos\theta, r\sin\theta) \\
&= (r\cos(\theta + \varphi), r\sin(\theta + \varphi)) \\
&= (r\cos\theta\cos\varphi - r\sin\theta\sin\varphi, r\sin\theta\cos\varphi + r\cos\theta\sin\varphi) \\
&= (x\cos\varphi - y\sin\varphi, x\sin\varphi + y\cos\varphi)
\end{aligned}$$

Such $\mathcal{R}_\varphi$ is called a φ-rotor. It should have no ambiguities that a rotor means an operator of rotation, while a rotation is a noun for such performance. For some pure mathematicians, they had a tastier version of revision through matrix

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} \cos\varphi & -\sin\varphi \\ \sin\varphi & \cos\varphi \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}$$

Function Translation

We comes to the place of functions. Defining function with a graph seems abstract, but the representation makes sense and behave well during discussions. Let us consider the graph G stands

for the picture of the function $y = f(x)$, so whenever we need to discuss the properties of the graph, we can write the definition as

$$G := \{y = f(x) : x \in \mathbb{R}\}$$

For convenience, we translate our words using coordinates so that the application of transformation of graphs can be analyze through coordinates. Redefining G by the following set:

$$G := \{(x, y) : y = f(x), x \in \mathbb{R}\}$$

It seems wasting logic, but the underlying concept of such definition is stronger than writing an equivalence. We will see that every point on the graph shares similarities.

To construct a useful theorem for translation operation, we start by looking at every coordinates on the graph. We put our definition of translation down here.

Definition 1.12 (Function graph translation)**.** *Let G be a graph of the function $y = f(x)$. A* ***function graph translation*** *means a general point translation on all points in the graph.*

This means that, if we have to move the graph around, it is equivalent to saying moving each point on the graph around.

Consider a horizontal translation first, as in point translation. A horizontal translation of the function graph by h units is preformed as

$$(x, y) \mapsto (x + h, y).$$

Since we have $y = f(x)$ to define the graph, each coordinates can be rewritten as

$$(x + h, f(x)) = (x_0, f(x_0 - h))$$

by substituting $x_0 = x + h$. It follows that a horizontal translation by h units is by transforming $y = f(x)$ into $y = f(x - h)$.

Similarly, a horizontal translation of the function graph by k units is preformed as

$$(x, y) \mapsto (x, y + k).$$

Since, again, we have $y = f(x)$ to define the graph, each coordinates can be rewritten as

$$(x, f(x) + k).$$

It follows that a vertical translation by k units is by transforming $y = f(x)$ into $y = f(x) + k$.

Now we draw a partial conclusion to summarize translation of graphs.

Theorem. *Given $G := \{(x, y) : y = f(x), x \in \mathbb{R}\}$ be a function graph.*

- *A horizontal translation by h units is performed by $y = f(x) \mapsto f(x - h)$;*
- *A vertical translation by k units is performed by $y = f(x) \mapsto f(x) + k$.*

Furthermore, translation is commutative and compositive. A composite translation by (h, k) defines the simultaneous translation in both direction is performed by

$$y = f(x) \mapsto f(x - h) + k$$

Function Dilation

A dilation means we can enlarge or diminish the graph, at the same time perform rotation if needed. It is important to notice that dilation and translation are not commutative to each other, so we separate their section of discussion here.

Let us focus on the cases centered at origin first. Again, we will define the graph G of the function f using the set

$$G := \{(x, y) : y = f(x), x \in \mathbb{R}\}$$

Definition 1.13 (Vertical dilation centered at origin)**.** *A vertical dilation is a dilation along y-axis. Given the dilation factor a of vertical dilation, we have*

$$(x, y) \mapsto (x, ay)$$

The main idea here is to rescale the vertical displacement between the point (x, y) and the origin. We then apply $y = f(x)$ to the presentation to yield

$$y = f(x) \mapsto af(x)$$

Similarly, we have the version for horizontal dilation.

Definition 1.14 (Horizontal dilation centered at origin)**.** *A horizontal dilation is a dilation along x-axis. Given the dilation factor b of vertical dilation, we have*

$$(x, y) \mapsto (bx, y)$$

With the same strategy of modification, we have

$$y = f(x) \mapsto f(\frac{x}{b})$$

And what's the purpose to call it 'about origin'? If we apply vertical and horozontal dilation simultaneously, it is natural to ask to questions:

1. How about when we do not scale the graph about origin? and;

2. What is the total scaling?

We shall answer the first question first. Recall in previous part we have been discussing translation thoroughly. Let us define translation operator $T_{(h,k)}$ to perform the action 'translating by (h,k)' and its inverse $T^{-1}_{(h,k)}$ to return the original position. It is not hard to see $T^{-1}_{(h,k)} = T_{-(h,k)}$. On the other hand, pick D^b_x to be the dilation operator by factor b along x-axis, and D^a_y for that of y-axis.

Observe that a dilation centered at some point (h,k) is difficult to imagine when no clues can be drawn in terms of direction (assuming we need to understand the process without vectors). We shall translate the thought to centering at origin first so that we can apply what we have investigate for the case centered at origin. Then we translate back to (h,k). Follow the thought, we have the following idea.

Theorem (Vertical dilation centered at (h,k))**.** *Given the dilation factor a of vertical dilation, for a vertical dilation centered at (h,k), we perform*

$$T_{(h,k)} \circ D^a_y \circ T_{-(h,k)}(x,y)$$

Following the idea of composition of translation and dilation, we can write down the conclusion without operators.

$$\begin{aligned} T_{(h,k)} \circ D^a_y \circ T_{-(h,k)}(x,y) &= T_{(h,k)} \circ D^a_y(x-h, y-k) \\ &= T_{(h,k)}(x-h, a(y-k)) \\ &= (x, ay - ak + k) \end{aligned}$$

In terms of function, it is equivalent to

$$y = f(x) \mapsto af(x) - ak + k$$

Similarly, we construct the Horizontal dilation centered at (h,k) by the following composition.

Theorem (Horizontal dilation centered at (h,k))**.** *Given the horizontal factor b of vertical dilation, for a horizontal dilation centered at (h,k), we perform*

$$T_{(h,k)} \circ D^b_x \circ T_{-(h,k)}(x,y)$$

$$\begin{aligned} T_{(h,k)} \circ D_x^b \circ T_{-(h,k)}(x,y) &= T_{(h,k)} \circ D_x^b(x-h, y-k) \\ &= T_{(h,k)}(b(x-h), y-k) \\ &= (bx - bh + h, y) \end{aligned}$$

Now that we can rewrite in terms of function

$$y = f(x) \mapsto f(\frac{1}{b}x + \frac{bh-h}{b})$$

In order to simplify the theory, let us combine all situations to close the file of dilation in highschool level.

Theorem. *Let f be a function and G be the graph of f. If the graph is dilated horizontally by b centered at (x_1, y_1) and vertically by a centered at (x_2, y_2), the function is transformed as follows:*

$$y = f(x) \mapsto af(\frac{1}{b}x + \frac{bx_2 - x_2}{b}) - ay_1 + y_1$$

We stop here for highschool part. We are going to answer the second question with some view points in higher Mathematics.

Let us define the **total scaling** by observing that under the dilation centered at origin, the displacement from origin will be (i) multiplied by b horizontally with D_x^b and (ii) multiplied by a vertically with D_y^a. We define the operator norm $\|f\|$ to be the maximal scaling factor of applying the function f, i.e. there exists an integer M such that

$$M = \|f\| = \sup\{|f(x)| : |x| = 1\}$$

Then a total scaling is exactly the norm of a dilation operator. To compute, we have

$$\begin{aligned} \|D_y^a \circ D_x^b\| &= \|(x,y) \mapsto (bx, ay)\| \\ &= \sup \sqrt{a^2 + b^2} \\ &= \sqrt{a^2 + b^2} \end{aligned}$$

It is complicated to check that any dilation follows the above total scaling under Euclidean gemetry. We shall believe it, otherwise, more nuclear tools are needed to prove the result rigorously.

1.6 Challenging questions

2 Trigonometric functions

A right-angled triangle is usually a pure triangle we want to study the beauty of geometry. In this section, we are going to examine the usage of trigonometric functions as far as we could.

2.1 Why are we calling it 'trigonometry' ?

The word 'trigonometry' can be divided into two parts - 'trigono-' and '-metry'. The word 'trigono-' comes from the word 'trigonal', meaning 'the things of trigon'. What trigon means is a triangle: in English, the word 'tri-' means the number 3, and what counts to 3 can be named 'tri-' things, like tripoid, triangle, tricycle, trisep, etc. So a trigon is equal to a triangle, just a difference in perspective: one tries to describe the number of vertices in a polygon, while one makes a point to number of angles in it. Another word '-metry' comes from the word 'metric', which means 'the method of comparison'. In complete word, 'trigonometry' means 'the method of comparison of triangle'. Therefore, when we learn trigonometry, we will compare many attributes: the length of sides in one triangle, the shapes and sizes of different triangles.

2.2 The relation of sides of triangle

We shall first declare what is a triangle, how it forms and what is special about it.

Triangle inequality and Pythagoras theorem

Given a triangle, if we could draw it on a plane, we must have it with following axiomatic structures:

- There are 3 vertices in a triangle.
- There are 3 edges connecting 3 vertices.
- Given a vertex and its adjacent edges, an angle is formed. There are 3 angles in total.

To emphasize the structure inheriting 3 angles, we call it a **triangle**, with the prefix *tri-* standing for the number 3.

We shall immediate find a property of any given triangle that any edge in a triangle must be shorter than the sum of length of the other two edges. Here we call it **the triangle inequality**.

Axiom (Triangle inequality)**.** *Let a, b, c be some positive numbers. If they denote the side lengths of a triangle, then we have the following inequalities:*

- $a + b \geq c$;
- $a + c \geq b$;
- $b + c \geq a$.

We shall take them as axioms, since we do not need to dig too deep to the roots of such geometric constructions. We shall take it to be true with our intuition.

Theorem (Pythagoras theorem)**.** *Let a, b, c be positive numbers. Let a, b be the adjacent sides of a right-angled triangle and c be the hypotenuse of the right-angled triangle. Then*

$$a^2 + b^2 = c^2$$

Corollary. *The following holds for a right-angled triangle with adjacent sides a, b and hypotenuse c.*

- $a < c$;
- $b < c$.

The Sine function

Historically, the sine function is defined for the relation between the opposite side of a given acute angle in a triangle and the hypotenuse of the triangle.

Definition 2.1 (The Sine function)**.** *Let a, b, c be sides of a given right-angled triangle and θ be an acute angle inside the triangle as shown.*

*Then the **sine function of** θ, $\sin\theta$, is defined by*

$$\sin\theta := \frac{a}{c}$$

In some sense, we call the side a the opposite side of θ and c the hypotenuse. We so shorthand opposite side by 'o' and hypotenuse by 'h'. This produces a well known memorization technique of the relation described by sine function, read as 'soh' relation.

Theorem. *The range of the sine function in a right-angled triangle is constrained to be*

$$0 < \sin\theta < 1$$

©Mok Owen 2024

Proof.
By definition, $a < c$, so $\frac{a}{c} < 1$; at the same time, $a > 0$ and $c > 0$, hence $\frac{a}{c} > 0$. □

The Cosine function

The cosine function is defined for the relation between the adjacent side of a given acute angle in a triangle and the hypotenuse of the triangle.

Definition 2.2 (The Cosine function)**.** *Let a, b, c be sides of a given right-angled triangle and θ be an acute angle inside the triangle as shown.*

*Then the **cosine function of** θ, $\cos\theta$, is defined by*

$$\cos\theta := \frac{b}{c}$$

In some sense, we call the side b the adjacent side of θ and c the hypotenuse. We so shorthand adjacent side by 'a' and hypotenuse by 'h'. This produces a well known memorization technique of the relation described by cosine function, read as 'cah' relation.

Theorem. *The range of the sine function in a right-angled triangle is constrained to be*

$$0 < \cos\theta < 1$$

Proof.
By definition, $b < c$, so $\frac{b}{c} < 1$; at the same time, $b > 0$ and $c > 0$, hence $\frac{b}{c} > 0$. □

The tangent function

The tangent function is defined for the relation between the opposite side and the adjacent side of a given acute angle in a triangle.

Definition 2.3 (The Tangent function)**.** *Let a, b, c be sides of a given right-angled triangle and θ be an acute angle inside the triangle as shown.*

*Then the **tangent function of** θ, $\tan\theta$, is defined by*

$$\tan\theta := \frac{a}{b}$$

In some sense, we call the side a the opposite side of θ and b the opposite side of θ. We so shorthand opposite side by 'o' and adjacent side by 'a'. This produces a well known memorization technique of the relation described by tangent function, read as 'toa' relation.

Theorem. *The range of the sine function in a right-angled triangle is constrained to be*

$$0 < \sin\theta$$

Proof.
By definition, $a > 0$ and $b > 0$, hence $\frac{a}{b} > 0$. □

Fundamental identities for trigonometric functions

Theorem. $\sin(90^\circ - \theta) = \cos\theta$

Proof.
Suppose a, b, c be the opposite side, the adjecent side and the hypotenuse to the acute angle θ in a right-angled triangle respectively. Note that by the law of angle sum of triangle, if we denote the remaining angle other than θ and 90° by ϕ, we get $\phi = 90^\circ - \theta$. Then

$$\begin{aligned}\sin(90^\circ - \theta) &= \sin\phi \\ &= \frac{b}{c} \\ &= \cos\theta\end{aligned}$$

□

Corollary. $\cos(90^\circ - \theta) = \cos\theta$

Theorem. $\sin^2\theta + \cos^2\theta = 1$.

Proof.
Suppose a, b, c be the opposite side, the adjecent side and the hypotenuse to the acute angle θ in a right-angled triangle respectively. Then

$$\begin{aligned}\sin^2\theta + \cos^2\theta &= (\frac{a}{c})^2 + (\frac{b}{c})^2 \\ &= \frac{a^2}{c^2} + \frac{b^2}{c^2} \\ &= \frac{a^2 + b^2}{c^2} \\ &= \frac{c^2}{c^2} = 1\end{aligned}$$

□

Theorem. $\tan\theta = \dfrac{\sin\theta}{\cos\theta}$.

Proof.

Suppose a, b, c be the opposite side, the adjecent side and the hypotenuse to the acute angle θ in a right-angled triangle respectively. Then

$$\begin{aligned}\frac{\sin\theta}{\cos\theta} &= \sin\theta/\cos\theta \\ &= \frac{a}{c}/\frac{b}{c} \\ &= \frac{a}{c}\cdot\frac{c}{b} \\ &= \frac{a}{b} \\ &= \tan\theta\end{aligned}$$

□

Example. *Simplify the expression* $\dfrac{[\sin(90^\circ-\theta)-\cos(90^\circ-\theta)]^2}{1-(\tan\theta)(2\cos^2\theta)}$.

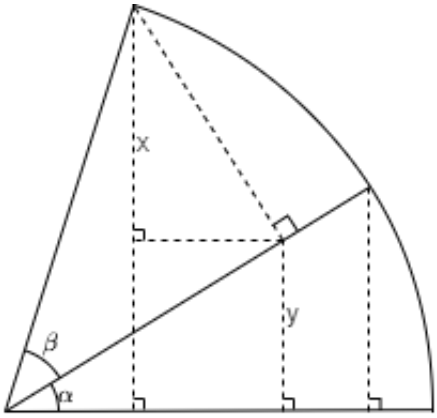
Sol. By proved theorem, we have

$$\begin{aligned}\frac{[\sin(90^\circ-\theta)-\cos(90^\circ-\theta)]^2}{1-(\tan\theta)(2\cos^2\theta)} &= \frac{(\cos\theta-\sin\theta)^2}{1-(\frac{\sin\theta}{\cos\theta})(\cos^2\theta)} \\ &= \frac{\cos^2\theta-2\sin\theta\cos\theta+\sin^2\theta}{1-2\sin\theta\cos\theta} \\ &= \frac{1-2\sin\theta\cos\theta}{1-2\sin\theta\cos\theta} \\ &= 1\end{aligned}$$

2.3 The compound angle formulae

Someone might be interested in computing compound angle formulae, which are important in developing different results of trigonometric relations. We will assume that α and β are both acute angles, so that all the results could be applied to whenever $0 < \alpha+\beta < 180^\circ$.

Consider the following sector with radius 1:



Theorem. $\sin(\alpha+\beta) = \sin\alpha\cos\beta + \sin\beta\cos\alpha$.

Proof.

Referring to the figure, we have $\sin(\alpha+\beta) = x + y$.

On the other hand, we have

$$\begin{cases} x = \sin\beta\cos\alpha \\ y = \cos\beta\sin\alpha \end{cases}$$

Therefore, $\sin(\alpha+\beta) = \sin\alpha\cos\beta + \sin\beta\cos\alpha$. □

For now we have the result for $\sin(\alpha+\beta)$. Then the remaining results could be proven under substitution of variables:

Theorem. $\sin(\alpha-\beta) = \sin\alpha\cos\beta - \sin\beta\cos\alpha$.

Proof.

$$\begin{aligned} \sin(\alpha-\beta) &= \sin(\alpha+(-\beta)) \\ &= \sin\alpha\cos(-\beta) + \sin(-\beta)\cos\alpha \\ &= \sin\alpha\cos\beta - \sin\beta\cos\alpha \end{aligned}$$

□

Theorem. $\cos(\alpha+\beta) = \cos\alpha\cos\beta - \sin\alpha\sin\beta$.

Proof.

$$\begin{aligned}\cos(\alpha+\beta) &= \sin\Big(\frac{\pi}{2}-(\alpha+\beta)\Big)\\ &= \sin\Big(\frac{\pi}{2}-\alpha\Big)\cos(-\beta)+\sin(-\beta)\cos\Big(\frac{\pi}{2}-\alpha\Big)\\ &= \cos\alpha\cos\beta-\sin\alpha\sin\beta\end{aligned}$$

□

Theorem. $\cos(\alpha-\beta)=\cos\alpha\cos\beta+\sin\alpha\sin\beta$.

Proof.

$$\begin{aligned}\cos(\alpha-\beta) &= \cos(\alpha+(-\beta))\\ &= \cos\alpha\cos(-\beta)-\sin\alpha\sin(-\beta)\\ &= \cos\alpha\cos\beta+\sin\alpha\sin\beta\end{aligned}$$

□

Theorem. $\tan(\alpha+\beta)=\frac{\tan\alpha+\tan\beta}{1-\tan\alpha\tan\beta}$.

Proof.

$$\begin{aligned}\tan(\alpha+\beta) &= \frac{\sin(\alpha+\beta)}{\cos(\alpha+\beta)}\\ &= \frac{\sin\alpha\cos\beta+\sin\beta\cos\alpha}{\cos\alpha\cos\beta-\sin\alpha\sin\beta}\\ &= \frac{\frac{\sin\alpha}{\cos\alpha}+\frac{\sin\beta}{\cos\beta}}{1-\frac{\sin\alpha}{\cos\alpha}\frac{\sin\beta}{\cos\beta}}\\ &= \frac{\tan\alpha+\tan\beta}{1-\tan\alpha\tan\beta}\end{aligned}$$

□

Theorem. $\tan(\alpha-\beta)=\frac{\tan\alpha-\tan\beta}{1+\tan\alpha\tan\beta}$

Proof.

$$\begin{aligned}\tan(\alpha-\beta) &= \frac{\tan\alpha+\tan(-\beta)}{1-\tan\alpha\tan(-\beta)}\\ &= \frac{\tan\alpha-\tan\beta}{1+\tan\alpha\tan\beta}\end{aligned}$$

□

All the results shown are based on $\alpha+\beta<90°$, so the case where $\alpha+\beta>90°$ are left to students.

2.4 The trigonometric laws

To understand the trigonometric functions better, we will try to extend the meaning of the trigonometric functions. The first step is to construct some laws with them so that the meaning could be seen.

In our discussion, we will look at two different triangles: one with acute angles only, while another with one obtuse angle. We will refer to the following triangles by T_1 and T_2 at anytime.

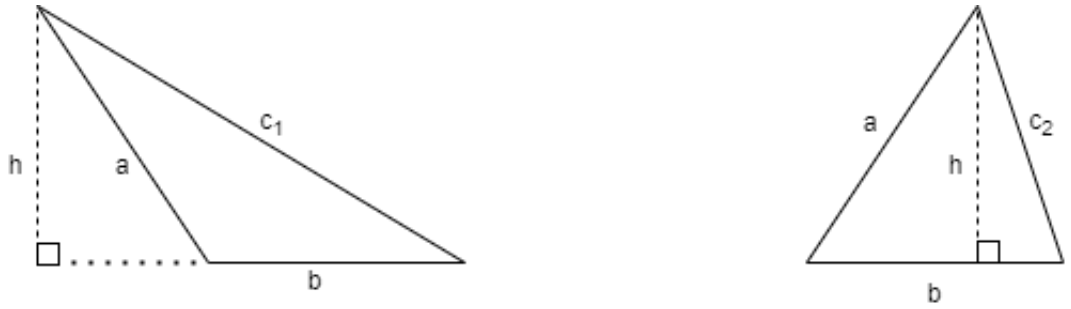


Figure 3: T_1 (left) and T_2 (right)

In convenience, we will denote the angle opposite to c_i to be γ_i. α to a and β to b.

The area formula of a triangle

Consider the area of T_1 and T_2, if we are taken their height to be the same, we will have the following interesting results:

Consider the area of T_2, we have $h=a\sin(\gamma_2)$, so its area is

$$\frac{1}{2}ab\sin(\gamma_2)$$

Consider the area of T_1, we have $h = a\sin(180^\circ - \gamma_1)$. The value of $\sin(180^\circ - \gamma_1)$ can be obtained by consiering

$$\begin{aligned}
\gamma_1 &= 180^\circ - \alpha - \beta \\
\sin\gamma_1 &= \sin(180^\circ - \alpha - \beta) \\
&= \sin((90^\circ - \alpha) + (90^\circ - \beta)) \\
&= \sin(90^\circ - \alpha)\cos(90^\circ - \beta) + \cos(90^\circ - \alpha)\sin(90^\circ - \beta) \\
&= \cos\alpha\sin\beta + \sin\alpha\cos\beta \\
&= \sin(\alpha + \beta) \\
&= \sin(180^\circ - \gamma_1)
\end{aligned}$$

Hence, we have its area to be

$$\frac{1}{2}ab\sin(180^\circ - \gamma_1) = \frac{1}{2}ab\sin(\gamma_1)$$

In conclusion, the area of a triangle with given angle θ and adjacent sides can be computed as

Theorem. *Given a triangle with a given angle θ in which $0 \leq \theta \leq 180^\circ$ and corresponding adjacent sides of length a and b, the area of the triangle is*

$$\frac{1}{2}ab\sin\theta$$

The Sine Law

Our first law in discussion of trigonometric function is the **sine law** - the law inheriting the charateristics of sine function.

Theorem (Sine Law)**.** $\dfrac{a}{\sin\alpha} = \dfrac{b}{\sin\beta} = \dfrac{c_i}{\sin\gamma_i}$.

Proof by extended right-angle triangle.
We first complete the case in T_2. It is not hard to see that

$$\begin{aligned}
c_2\sin\alpha &= a\sin\gamma_2 \\
\frac{c_2}{\sin\gamma_2} &= \frac{a}{\sin\alpha}
\end{aligned}$$

By rotation, similar results can be proven for other equality. We are done in the case of T_2.

We then look at the case of T_1. As in area computation, we shall see that

$$\sin\gamma_1 = \sin(180^\circ - \gamma_1)$$

Then

$$c_1 \sin\alpha = a\sin(180^\circ - \gamma_1)$$
$$\frac{c_1}{\sin\gamma_1} = \frac{a}{\sin\alpha}$$

□

Proof by area formula.
Consider the area of the triangle, we obtain the following equality:

$$\frac{1}{2}ab\sin\gamma = \frac{1}{2}bc\sin\alpha = \frac{1}{2}ac\sin\beta$$

which yields the following result when dividing each expression by $\frac{abc}{2}$:

$$\frac{\sin\gamma}{c} = \frac{\sin\alpha}{a} = \frac{\sin\beta}{b}$$

□

The law connects an angle with its opposite sides by the ratio of sine function. This is a good perspective to apply the formula to solve any triangle with given two sides and an angle.

Example. *Suppose a triangle $\triangle ABC$ with $AB = 3, AC = 5, \angle B = 60^\circ$. Then it is not hard to see*

$$\begin{aligned}
\frac{AC}{\sin\angle B} &= \frac{AB}{\sin\angle C} = \frac{BC}{\sin\angle A} \\
\sin\angle C &= \frac{3\sin 60^\circ}{5} \\
&= \frac{3\sqrt{3}}{10} \\
\sin\angle A &= \sin\angle B\cos\angle C + \sin\angle C\cos\angle B \\
&= \frac{\sqrt{3}}{2}\cdot\frac{\sqrt{73}}{10} + \frac{3\sqrt{3}}{10}\cdot\frac{1}{2} \\
&= \frac{\sqrt{219}+3\sqrt{3}}{20} \\
BC &= \frac{10\sqrt{3}}{3}\cdot\frac{\sqrt{219}+3\sqrt{3}}{20} \\
&= \frac{3+\sqrt{73}}{2}
\end{aligned}$$

Remark. *It is the usual sine function when one of the angle becomes 90°.*

The Cosine Law

The second law in trigonometry to be discussed will be the **cosine law** - the law inheriting cosine function.

Theorem (Cosine Law)**.** *The following formulae holds:*

- $c^2 = a^2 + b^2 - 2ab\cos\gamma \iff \cos\gamma = \dfrac{a^2+b^2-c^2}{2ab}$.
- $b^2 = a^2 + c^2 - 2ac\cos\beta \iff \cos\beta = \dfrac{a^2+c^2-b^2}{2ac}$.
- $a^2 = b^2 + c^2 - 2bc\cos\alpha \iff \cos\alpha = \dfrac{b^2+c^2-a^2}{2bc}$.

Standard proof.

We first see the case in T_2. It is not hard to see if we partition b into $x+y$ according to their order, we could construct the relations

$$\begin{cases} x+y & = b \\ a^2-x^2 & = h^2 = c_2^2 - y^2 \\ x & = a\cos\gamma_2 \end{cases} \implies \begin{cases} a^2-x^2 & = c_2^2-(b-x)^2 \\ x & = a\cos\gamma_2 \end{cases}$$

$$\implies \begin{cases} a^2+b^2-2bx & = c_2^2 \\ x & = a\cos\gamma_2 \end{cases}$$

$$\implies c_2^2 = a^2+b^2-2ab\cos\gamma_2$$

On the other hand, for T_1, we have to tackle particularly the obtused angle case. Note that if T_1 is divided into two right-angled triangle, then γ_1 can be divided into $\theta_1+\theta_2$. It is equivalent to say

$$\begin{aligned} \cos\gamma_1 &= \cos\theta_1\cos\theta_2 - \sin\theta_1\sin\theta_2 \\ &= \sin(90^\circ-\theta_1)\sin(90^\circ-\theta_2) - \cos(90^\circ-\theta_1)\cos(90^\circ-\theta_2) \\ &= -\cos(180^\circ-\gamma_1) \end{aligned}$$

Therefore, we pick to consider $x = a\cos(180^\circ-\gamma_1)$, then

$$\begin{aligned} c_1^2-(b+x)^2 &= a^2-x^2 \\ c_1^2-b^2-2bx &= a^2 \\ c_1^2 &= a^2+b^2+2ab\cos(180^\circ-\gamma_1) \\ c_1^2 &= a^2+b^2-2ab\cos\gamma_1 \end{aligned}$$

This concludes that in any case, the cosine law holds. □

The best part of cosine law is that it generalizes the Pythagoras theorem to aritrary triangles. It is also more applicable than Sine law to different situations.

Example. *Refer to the same example in sine law. We can now directly find the length of BC.*

$$\begin{aligned} AC^2 &= AB^2 + BC^2 - 2 \cdot AB \cdot BC \cos \angle B \\ 5^2 &= 3^2 + BC^2 - 2 \cdot 3 \cdot BC \cos 60^\circ \\ 0 &= BC^2 - 3BC - 16 \end{aligned}$$

Shall we learn to solve quadratic equations, the solution to the equation is $\dfrac{3 \pm \sqrt{3^2 - 4(-16)}}{2} = \dfrac{3 \pm \sqrt{73}}{2}$. *Since* $\sqrt{73} > 3$, *we need* $BC > 0$, *so* $BC = \dfrac{3 + \sqrt{73}}{2}$.

The Tangent Law

It is the third law in trigonometry, which uses less but still understandable to write down. It is the **tangent law** - a law inheriting tangent function as its argument.

Theorem (Tangent Law)**.** *The following formulae holds:*

- $\dfrac{a-b}{a+b} = \dfrac{\tan \frac{\alpha-\beta}{2}}{\tan \frac{\alpha+\beta}{2}}$.
- $\dfrac{b-c}{b+c} = \dfrac{\tan \frac{\beta-\gamma}{2}}{\tan \frac{\beta+\gamma}{2}}$.
- $\dfrac{c-a}{c+a} = \dfrac{\tan \frac{\gamma-\alpha}{2}}{\tan \frac{\gamma+\alpha}{2}}$.

Algebraic proof.
Consider $\dfrac{\tan \frac{\alpha-\beta}{2}}{\tan \frac{\alpha+\beta}{2}}$. Denote $t_\alpha = \tan \frac{\alpha}{2}$ and $t_\beta = \tan \frac{\beta}{2}$.

$$\begin{aligned}
\frac{\tan\frac{\alpha-\beta}{2}}{\tan\frac{\alpha+\beta}{2}} &= \frac{\tan\frac{\alpha}{2}-\tan\frac{\beta}{2}}{1+\tan\frac{\alpha}{2}\tan\frac{\beta}{2}}\frac{1-\tan\frac{\alpha}{2}\tan\frac{\beta}{2}}{\tan\frac{\alpha}{2}+\tan\frac{\beta}{2}} \\
&= \frac{t_\alpha - t_\beta - t_\alpha^2 t_\beta + t_\alpha t_\beta^2}{t_\alpha + t_\beta + t_\alpha^2 t_\beta + t_\alpha t_\beta^2} \\
&= \frac{t_\alpha(\sec^2\frac{\beta}{2}) - t_\beta(\sec^2\frac{\alpha}{2})}{t_\alpha(\sec^2\frac{\beta}{2}) + t_\beta(\sec^2\frac{\alpha}{2})} \\
&= \frac{t_\alpha(\cos^2\frac{\alpha}{2}) - t_\beta(\cos^2\frac{\beta}{2})}{t_\alpha(\cos^2\frac{\alpha}{2}) + t_\beta(\cos^2\frac{\beta}{2})} \\
&= \frac{\sin\alpha - \sin\beta}{\sin\alpha + \sin\beta} \\
&= \frac{ab\sin\alpha - ab\sin\beta}{ab\sin\alpha + ab\sin\beta} \\
&= \frac{a-b}{a+b}
\end{aligned}$$

□

The tangent law relates two angles with two corresponing opposite sides. The application makes an agreement with the sine law, however, without advantages to apply the tangent law because of its complexity. We just do it a favor to learn a complete view to all three trigonometric functions.

In advance, let's introduce one more result related to the tangent law to close this extra section.

Theorem (Mollweide's Equations)**.** *The following holds:*

$$\frac{a-b}{c} = \frac{\sin\frac{\alpha-\beta}{2}}{\cos\frac{\gamma}{2}} \tag{1}$$

$$\frac{a+b}{c} = \frac{\cos\frac{\alpha-\beta}{2}}{\sin\frac{\gamma}{2}} \tag{2}$$

Proof.

As α, β, γ represents 3 angles in a triangle, their sum agrees with $\alpha+\beta+\gamma = 180°$, therefore $\frac{\gamma}{2} = 90° - \frac{\alpha+\beta}{2}$. We have the equality

$$\begin{aligned}
\frac{\sin\frac{\gamma}{2}}{\cos\frac{\alpha-\beta}{2}} &= \frac{2\sin\frac{\gamma}{2}\cos\frac{\gamma}{2}}{2\cos\frac{\alpha-\beta}{2}\cos\frac{\gamma}{2}} && (multiply\,by\,\frac{2\cos\frac{\gamma}{2}}{2\cos\frac{\gamma}{2}}) \\
&= \frac{\sin\gamma}{\sin\alpha+\sin\beta} && (compound\,angle\,formula) \\
&= \frac{abc\sin\gamma}{abc\sin\alpha+abc\sin\beta} && (multiply\,by\,\frac{abc}{abc}) \\
&= \frac{c(ab\sin\gamma)}{a(bc\sin\alpha)+b(ac\sin\beta)} \\
&= \frac{c}{a+b} && (Sine\,Law)
\end{aligned}$$

This proves the first equation.

For the second equation, consider

$$\begin{aligned}
\frac{a-b}{c} &= \frac{a-b}{a+b}\cdot\frac{a+b}{c} \\
&= \frac{\tan\frac{\alpha-\beta}{2}}{\tan\frac{\alpha+\beta}{2}}\frac{\cos\frac{\alpha+\beta}{2}}{\sin\frac{\gamma}{2}} \\
&= \frac{\sin\frac{\alpha-\beta}{2}}{\cos\frac{\gamma}{2}}
\end{aligned}$$

Then the Mollweide's equations are proved. □

Heron's formula

In order to compute formula from all perspective, we now look at a well-known formula for computing area of an arbitrary triangle using side length only. This is the Heron's formula.

Theorem (Heron's formula)**.** *Let a, b, c be the side length of 3 sides of a triangle. Then the area of the triangle is computed by*

$$\sqrt{s(s-a)(s-b)(s-c)}$$

where $s = \dfrac{a+b+c}{2}$.

Geometric proof of Heron's formula.

Let r be the radius of the inscribe circle. Then the area of the triangle can be computed by

$$\frac{ra+rb+rc}{2} = rs$$

where $s = \frac{a+b+c}{2}$.

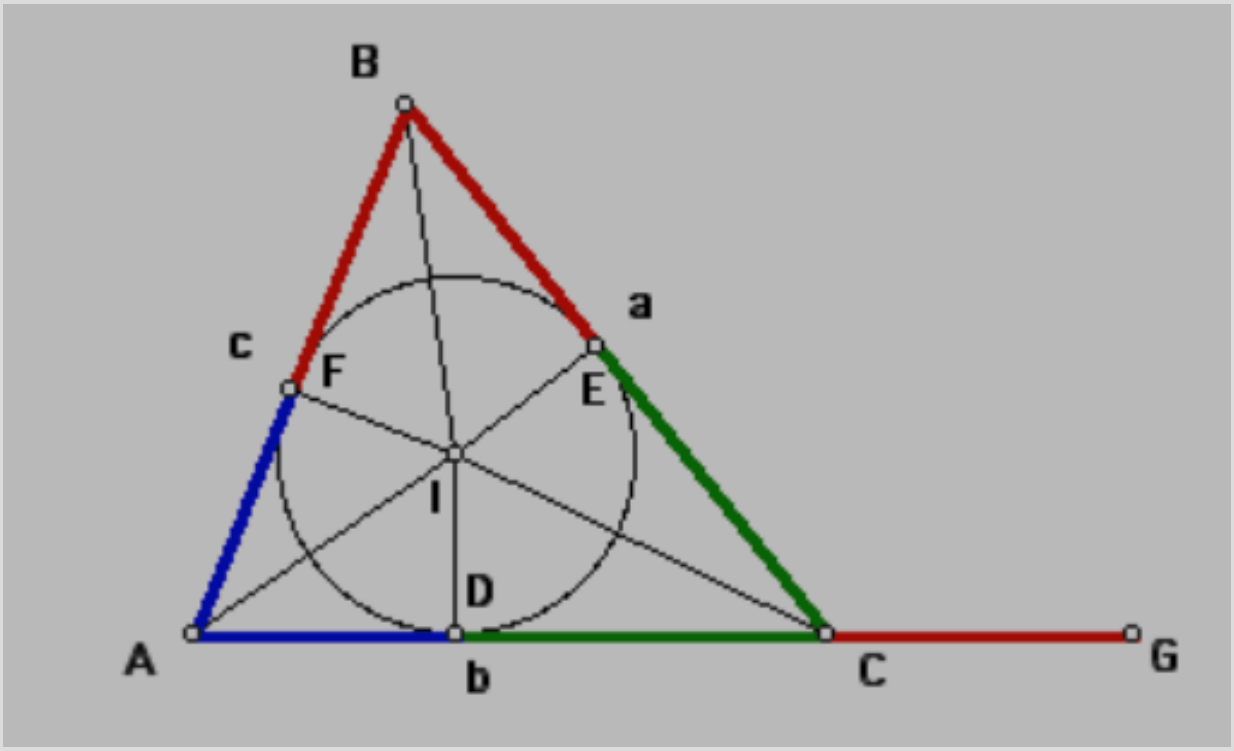


Figure 4: Reference: http://jwilson.coe.uga.edu/EMT668/EMAT6680.2000/Umberger/MATH7200/HeronFormulaProject/GeometricProof/geoproof.html

The following steps will try to resolve rs from other perspective. Here we orientate the side AC to be the base of the triangle (parallel to horizon), and extend the segment to G as in the figure, where the extended part CG is equal to BF.

Note that by construction, our line segments AB, BC, CA are all tangent to the inscribed circle, which are all perpendicular to the radius of touch to each of themselves. It is easy to check the following fact:

$$\begin{aligned} \triangle IFB &\cong \triangle IEB && (RHS) \\ \triangle IFA &\cong \triangle IDA && (RHS) \\ \triangle IDC &\cong \triangle IEC && (RHS) \end{aligned}$$

That means $AG = \dfrac{a+b+c}{2} = s$ The area becomes $r \cdot AG$.

Next we adjoin the figure with a line from I so that $\angle AIJ = 90°$ as follows.

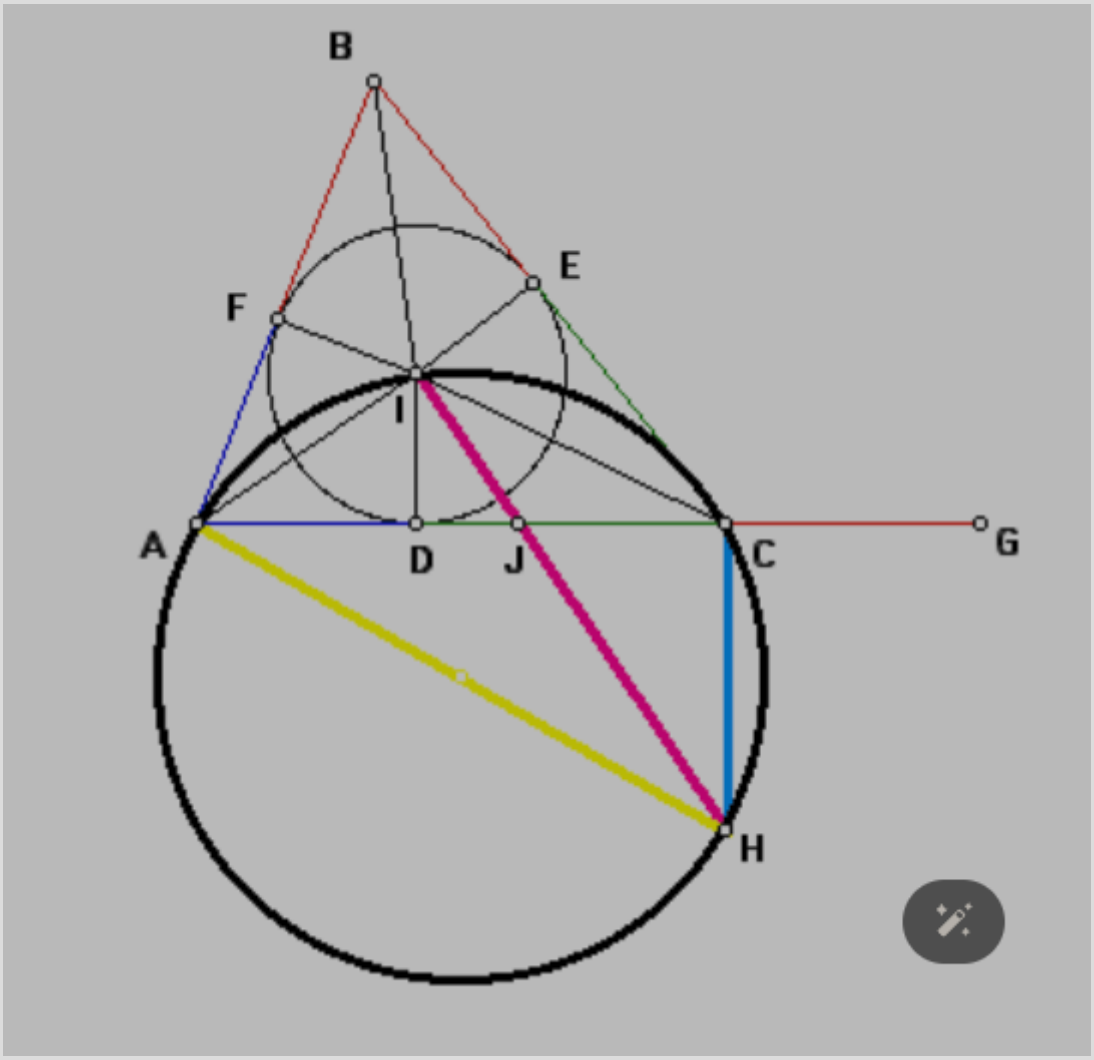


Figure 5: Reference: http://jwilson.coe.uga.edu/EMT668/EMAT6680.2000/Umberger/MATH7200/HeronFormulaProject/GeometricProof/geoproof.html

We try to figure out some similar triangles.

$$\begin{aligned} \angle AHC + \angle AIC &= 180^\circ && (concyclic, \angle\, supp.) \\ \angle BIE + \angle AIC &= 180^\circ && (\angle s\, at\, a\, pt.) \\ \therefore \angle AHC &= \angle BIE \\ \angle BEI = \angle ACH &= 90^\circ && (By\, construction) \\ \therefore \triangle ACH &\sim \triangle BEI \end{aligned}$$

On the other hand, it is not hard to see that

$$\triangle IDJ \sim \triangle HCJ$$

Then the following relation holds:

$$\begin{aligned}
\frac{AC}{CH} &= \frac{BE}{IE} && (\triangle ACH \sim \triangle BEI)\\
\frac{AC}{CG} &= \frac{CH}{ID} && (BE = CG, IE = ID)\\
&= \frac{CJ}{DJ} && (\triangle IDJ \sim \triangle HCJ)\\
\frac{AG}{CG} &= \frac{CD}{DJ} && (+1\, on\, both\, sides)\\
AG \cdot ID &= \frac{CD}{DJ} CG \cdot ID\\
&= CD \cdot CG \cdot \frac{AD}{ID} && (\frac{ID}{DJ} = \frac{AD}{ID}, \triangle ADI \sim \triangle IDJ)\\
&= CD \cdot CG \cdot \frac{AD \cdot AG}{ID \cdot AG}\\
AG \cdot ID &= \sqrt{AD \cdot DC \cdot CG \cdot AG} && (rearrangement)
\end{aligned}$$

Which, the area of $\triangle ABC$ is exactly

$$AG \cdot ID = \sqrt{AD \cdot DC \cdot CG \cdot AG}$$

and the proof is done. □

We have also the algebraic proof of heron's formula which is much sensible.

Algebraic proof of Heron's formula.
Consider the sine law and cosine law where we have proven to be true in any case. We have the area formula

$$A = \frac{1}{2} ab \sin \gamma$$

and thus if we do a little trick, square it, we have

$$\begin{aligned}
A^2 &= \frac{1}{4}a^2b^2\sin^2\gamma \\
&= \frac{1}{4}a^2b^2(1-\cos^2\gamma) \\
&= \frac{1}{4}a^2b^2(1-(\frac{a^2+b^2-c^2}{2ab})^2) \\
&= \frac{1}{4}a^2b^2(1+\frac{a^2+b^2-c^2}{2ab})(1-\frac{a^2+b^2-c^2}{2ab}) \\
&= \frac{1}{16}(a^2+2ab+b^2-c^2)(c^2-a^2+2ab-b^2) \\
&= \frac{1}{16}[(a+b)^2-c^2][c^2-(a-b)^2] \\
&= \frac{1}{16}(a+b+c)(a+b-c)(-a+b+c)(a-b+c) \\
&= s(s-a)(s-b)(s-c)
\end{aligned}$$

Thus, by the positivity of area, we have

$$A = \sqrt{s(s-a)(s-b)(s-c)}$$

where $s = \dfrac{a+b+c}{2}$ □

Example. *Let $\triangle ABC$ be with sides $7, 10, 13$. Define $s = \dfrac{7+10+13}{2} = 15$. Then the area of the triangle is*

$$S_{\triangle ABC} = \sqrt{15(15-7)(15-10)(15-13)} = \sqrt{1200} = 20\sqrt{3}$$

Essential Practice 2.4.1. *Compute the area of $\triangle ABC$ if the side lengths are:*

1. *1,2,3.*
2. *3,4,5.*
3. *2,6,7.*
4. *13,15,17.*
5. *9,9,6.*

2.5 Extending trigonometric functions to describing circles

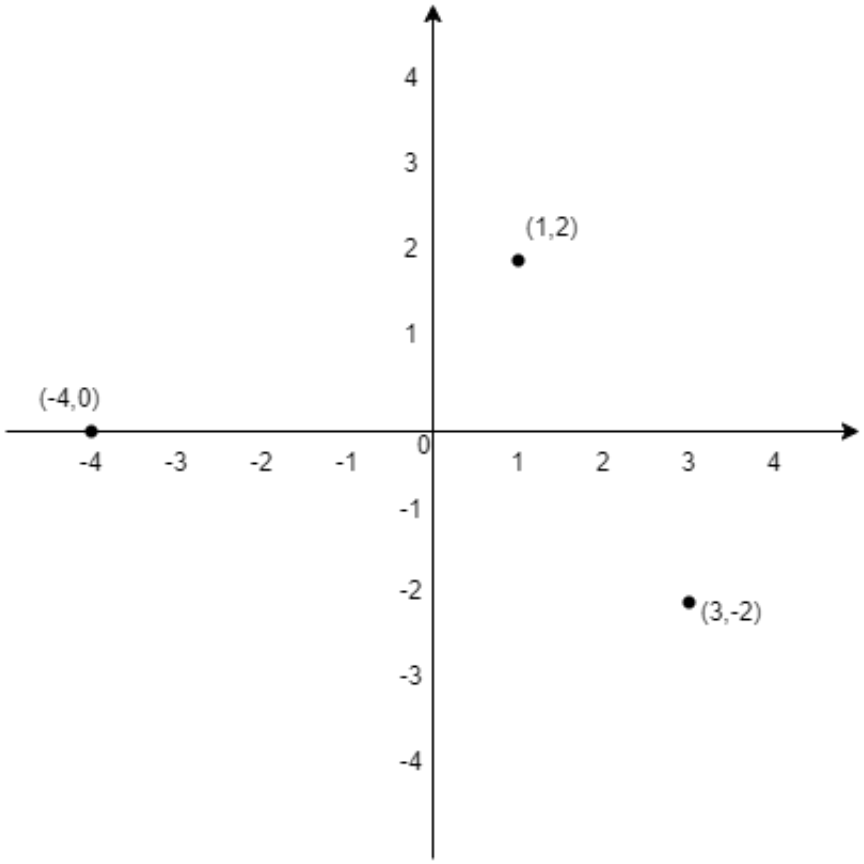
Trigonometric functions are not bounded by a triangle. In fact, if we want to, and trust to, extend trigonometric functions to any degree, we could still give a definition.

The relation between rectangular coordinates and polar coordinates

Let us recall the definitions of coordinate systems and their notations.

Definition 2.4 (Rectangular coordinates)**.** *A rectangular coordinates is a pair of numbers determining the position with respect to a pair of orthogonal axes, which is also called the rectangular coordinates plane.*

Example. *The positions* $(1,2),(3,-2),(-4,0)$ *are shown in the following rectangular coordinates plane.*



The coordinates are written in an ordered pair with horizontal position first then followed by vertical position.

It is important to note and remember that coordinates must be written in a certain order - horzontal position first, then vertical position. Otherwise, it lost its meaning in comparison.

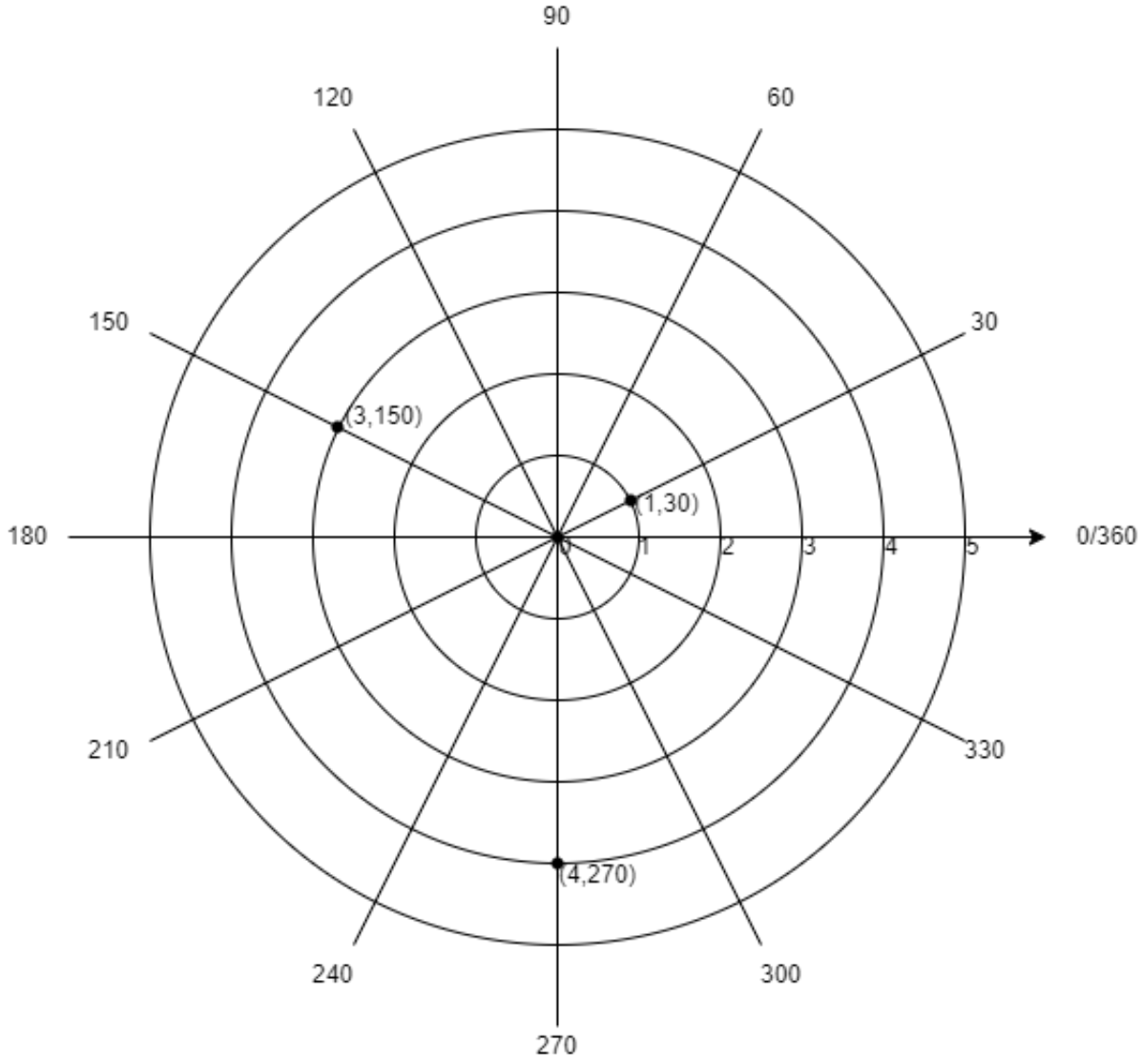
The point $(0,0)$ is usually denoted by O and named the **origin** of the plane. An origin is usually the respective 'center' of a plane although the plane could be infinitely extensive, so that

comparison maintain its stability. From this point of view, we may think of the coordinates (x, y) as the unital displacement from the center in horizontal and vertical position respectively. This is a concept of component-wise thinking.

Definition 2.5 (Polar coordinates)**.** *A polar coordinates is a pair of numbers determining the position with magnitude and direction.*

We usually define the magnitude by the shortest displacement from origin, again, the center of the plane, the position $(0, 0)$. But this time, $(0, 0)$ is just a convention to write down the position of center, since we cannot determine its direction of pointing.

Example. *The positions* $(1, 30°), (3, 150°), (4, 270°)$ *are shown in the following polar coordinates plane.*

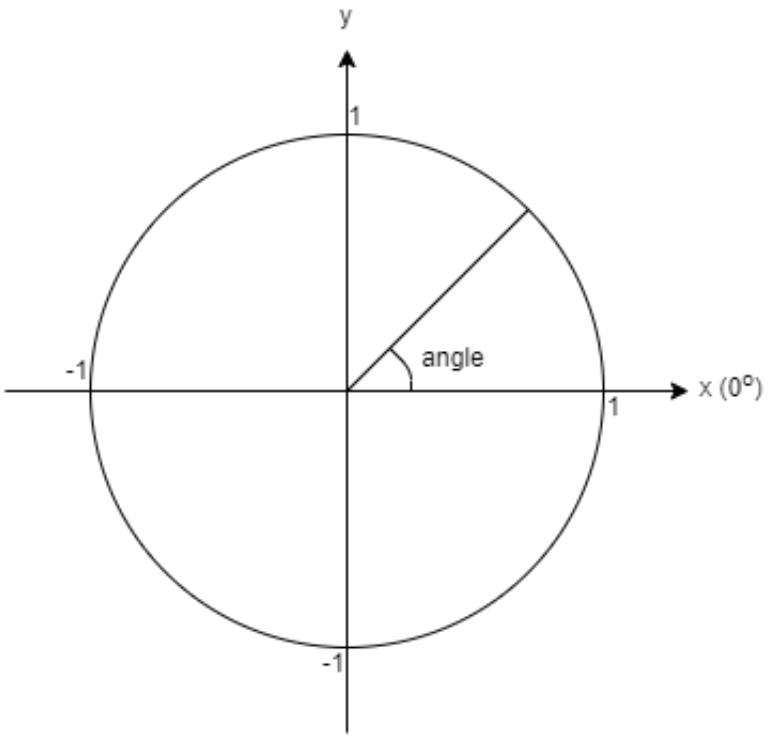


Therefore, we shall ask a question: could we interpolate the rectangular coordinates and polar coordinates with suitable functions?

To answer the above quesiton, we need to clarify what our desired interpolate strategy does and so we could choose a suitable set of functions to complete the interpolation.

The interpolation

Let us declare the relation between rectangular plane and polar coordinates by merging them together. Observe they both have an origin (which is important to show they both have a point of respect) and thus we can stash them into one picture. We will see the following result:



Placing a unit circle (a circle of radius 1) on the rectangular plane, we could see that if we identify the positive x-direction as the direction of 0° in polar coordinates, we can clearly merge two plane together. For each particular angle, we can draw a straight line to an intersection on the unit circle. Such point can be represented by a pair of rectangular coordinates, at the same time it can also be represented by a pair of polar coordinates by construction. We shall see the one-to-one correspondence between (x, y) and (r, θ) if we call the angle θ and restrict its domain to $0 \leq \theta < 360^\circ$. We restrict the domain to be less than 360° because of the duplication at 0°.

We can conclude what we need for such interpolation: some functions that relates the following

pair of numbers

$$(x, y) \sim (r, \theta)$$

and indeed, we can further write

$$x = f_1(r, \theta),\ y = f_2(r, \theta)$$

for some functions f_1 and f_2.

Defining trigonometry as coordinate function

We come to the part of configuring trigonometric functions as coordinate functions. Let us not use the notation of sine and cosine too early, since I will convince you that trigonometric functions will no longer be about triangle only with some reason behind.

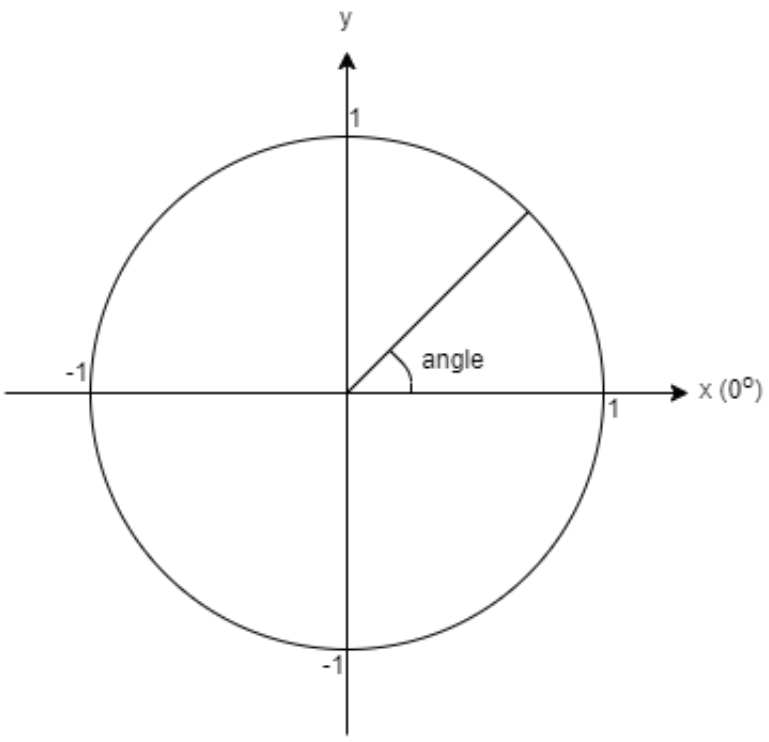
Let the discussed f_1 and f_2 be replaced by C and S for convenience. Then, we have to write

$$x = C(r, \theta),\ y = S(r, \theta)$$

One property of C and S that can easily be observed from similar triangle is that

$$C(r, \theta) = rC(1, \theta)$$
$$S(r, \theta) = rS(1, \theta)$$

and so we return to the unit circle for simpler observations.



As $C(1,\theta)$ defines the x-coordinate at a certain degree in a unit circle, we observe the following:

- When $\theta = 0°$, $C(1,\theta) = 1$;
- When $\theta = 90°$, $C(1,\theta) = 0$;
- When $\theta = 180°$, $C(1,\theta) = -1$;
- When $\theta = 270°$, $C(1,\theta) = 0$.

In particular, if we observe for $0° \leq \theta \leq 90°$, the values of $C(1,\theta)$ is equivalent to seeing the length of adjeacent side of a right-angled triangle with hypotenuse 1, by definition of trigonometric functions, is exactly the value of $\cos\theta$. Hence, when $0° \leq \theta \leq 90°$, we can write $C(1,\theta) = \cos\theta$.

We then say the function C defined for coordinates is an extension of the standard cosine function, and we will write $\cos\theta$ to simplify the writings, at the same time admit the meaning of $\cos\theta$ when $\theta > 90°$ is about the x-coordinate on a unit circle. Hence, we will write the following

$$x = r\cos\theta$$

for $0° \leq \theta < 360°$.

Similarly, we also observe that, as $S(1,\theta)$ defines the y-coordinate at a certain degree in a unit circle, we have the following:

- When $\theta = 0°$, $S(1,\theta) = 0$;
- When $\theta = 90°$, $S(1,\theta) = 1$;
- When $\theta = 180°$, $S(1,\theta) = 0$;
- When $\theta = 270°$, $S(1,\theta) = -1$.

In particular, if we observe for $0° \leq \theta \leq 90°$, the values of $S(1,\theta)$ is equivalent to seeing the length of opposite side of a right-angled triangle with hypotenuse 1, by definition of trigonometric functions, is exactly the value of $\sin\theta$. Hence, when $0° \leq \theta \leq 90°$, we can write $S(1,\theta) = \sin\theta$.

We then say the function S defined for coordinates is an extension of the standard sine function, and we will write $\sin\theta$ to simplify the writings, at the same time admit the meaning of $\sin\theta$ is about the y-coordinate on a unit circle. Hence, we will write the following

$$y = r\sin\theta$$

for $0° \leq \theta < 360°$.

Conclude the one-way implication of coordinates functions by the following relation:

©Mok Owen 2024

Proposition. *Let (r, θ) be a pair of polar coordinates, with $r > 0$ and $0° \leq \theta < 360°$. We can compute the corresponding rectangular coordinates by the formula*

$$(x, y) = (r\cos\theta, r\sin\theta)$$

Then, how about the reverse side of implication? We shall consider the reverse side of relation

$$(r, \theta) \sim (x, y)$$

which can be explicitly written as

$$r = g_1(x, y),\ \theta = g_2(x, y)$$

for some functions g_1 and g_2.

Let again write g_1 and g_2 into R and T for convenience. From the plot, it is easy to see for every pair of x and y, we can consider the distance between the circum-point and the center of the unit circle (the origin) to be $\sqrt{(x-0)^2 + (y-0)^2} = \sqrt{x^2 + y^2}$. From the formula it is easy to check that $R(x, y) = \sqrt{x^2 + y^2} \leq 0$ for any (x, y) in the plane. We so defined

$$r = R(x, y) = \sqrt{x^2 + y^2}$$

For the function T, it is a bit tricky. Let us consider the geometric meaning of the angle θ so that we could find a way to define the function T. One critical observation is the slope of the line connecting the origin and the coordinates (x, y). Let the slope be $\mathfrak{M}$. Then we can write

$$\mathfrak{M} = \frac{y-0}{x-0} = \frac{y}{x}$$

Therefore, the function T should be defined by another fucntion Q so that $\theta = T(x, y) = Q(\frac{y}{x})$, and if there is a function P such that $P \circ Q = \texttt{id}$, the identity function, we can write

$$P(\theta) = \frac{y}{x}$$

to observe the relation between θ and x, y.

Once again, when $0° \leq \theta < 90°$, the value of $P(\theta)$ is exactly describing the ratio of the length of opposite side over the length of adjacent side of a right-angled triangle, which by definition is $\tan\theta$. Hence, we conclude when $0° \leq \theta < 90°$, $P(\theta) = \tan\theta$. As in previous deduction, we will take the extension of $\tan\theta$ to be describing the slope to origin on rectangular plane. Hence, we conclude the relation by taking the inverse of tangent function:

$$\theta = \arctan\left(\frac{y}{x}\right)$$

Therefore, the implication from rectangular plane to polar plane, call it **polarization**, is concluded by

Proposition (Polarization)**.** *Let (x, y) be a pair of rectangular coordinates. We can compute the corresponding polar coordinates by the formula*

$$(r, \theta) = (\sqrt{x^2 + y^2}, \arctan\left(\frac{y}{x}\right))$$

One more hint to condense the name trigonometry to describing coordinates is the consistency of trigonometric identity:

Proposition. *Let $(x, y) \sim (r, \theta)$ be a pair of coordinate relation with discussed relating functions. The following identity holds well:*

1. $\sin^2\theta + \cos^2\theta \equiv 1$;
2. $\tan\theta \equiv \dfrac{\sin\theta}{\cos\theta}$.

Proof.

1. Since $x = r\cos\theta$ and $y = r\sin\theta$, following the definition of r,

$$\begin{aligned} x^2 + y^2 &= r^2 \\ (\frac{x}{r})^2 + (\frac{y}{r})^2 &= 1 \\ \cos^2\theta + \sin^2\theta &= 1 \end{aligned}$$

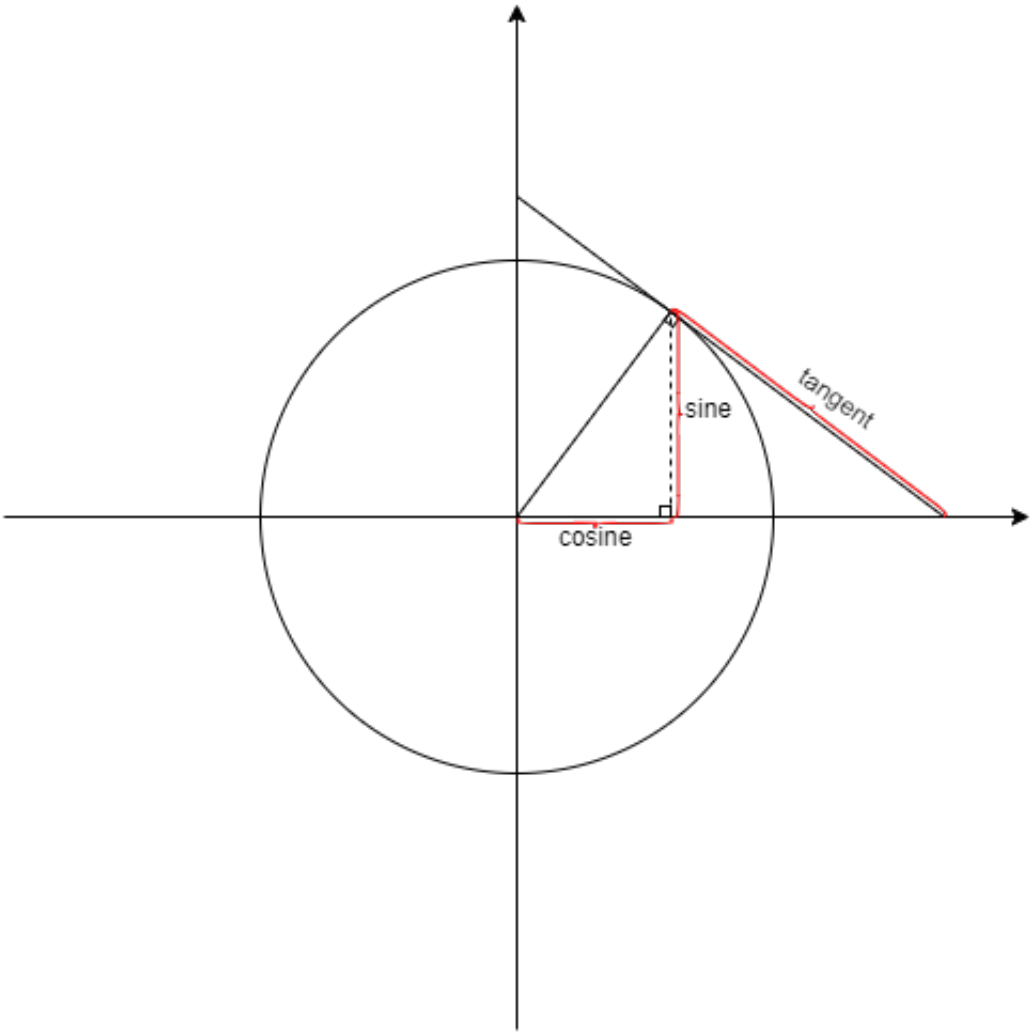
2. It is not hard to follow the equation deduced:

$$\begin{aligned} RHS &= \frac{\sin\theta}{\cos\theta} \\ &= \frac{y/r}{x/r} \\ &= \frac{y}{x} \\ &= \tan\theta \\ &= LHS \end{aligned}$$

□

But there's a lot more to discuss about the name 'tangent' - it is not a simple discussion to explain why we call the slope of the line connecting the point and the origin by tangent function, in fact, more intuition is needed to explain that naming.

For if any clear explanation is required, one popular saying is using the unit circle centered at origin.



Let us revise the meaning of $\tan\theta$. We usually call the straight line touches the circle at point (x, y) to be the **tangent line** to the circle at (x, y). Note that if we would like to find the **length of tangent captured by** θ, we would see the mentioned tangent as in the figure. With respect to the drawn unit circle, we could see a pair of similar triangle, which stands

$$\frac{\sin\theta}{\cos\theta} = \frac{tan\theta}{1}$$

and such that we have the relation $\tan\theta = \dfrac{\sin\theta}{\cos\theta}$ same as in trigonometry. The relation accidentally computes the same result as the slope of the radius, so we take in advance the slope equals to $\tan\theta$.

Using the concept of slope, we can easily describe the geometric meaning of negativity of tangent - negative slope. Therefore, this part of configuration is done.

©Mok Owen 2024

Reduction formula for trigonometric functions

According to the defined 'brand new' trigonometry (coordinating functions), we may now observe the symmetric properties over those trigonometric functions. We first read the symmetry of the sine function.

Theorem. $\sin(90^\circ - \theta) = \sin(90^\circ + \theta)$ *and* $\sin(270^\circ - \theta) = \sin(270^\circ + \theta)$ *for* $0^\circ \leq \theta \leq 90^\circ$.

Proof.
For each part of the theorem, note that we may place the y-axis the common side of the triangle drawn with both radius. The congruence is trivial and hence the equality holds. □

Corollary. $\sin(180^\circ - \theta) = \sin\theta$.

Theorem. $\cos(360^\circ - \theta) = \cos\theta$ *and* $\cos(180^\circ - \theta) = \cos(180^\circ + \theta)$ *for* $0^\circ \leq \theta \leq 90^\circ$.

Proof.
For each part of the theorem, note that we may place the x-axis the common side of the triangle drawn with both radius. The congruence is trivial and hence the equality holds. □

With the collection from fundamental trigonometry, we have the following collection of Reduction formulae:

Theorem. *Let* θ *be such that* $0^\circ \leq \theta \leq 90^\circ$*, the following holds:*

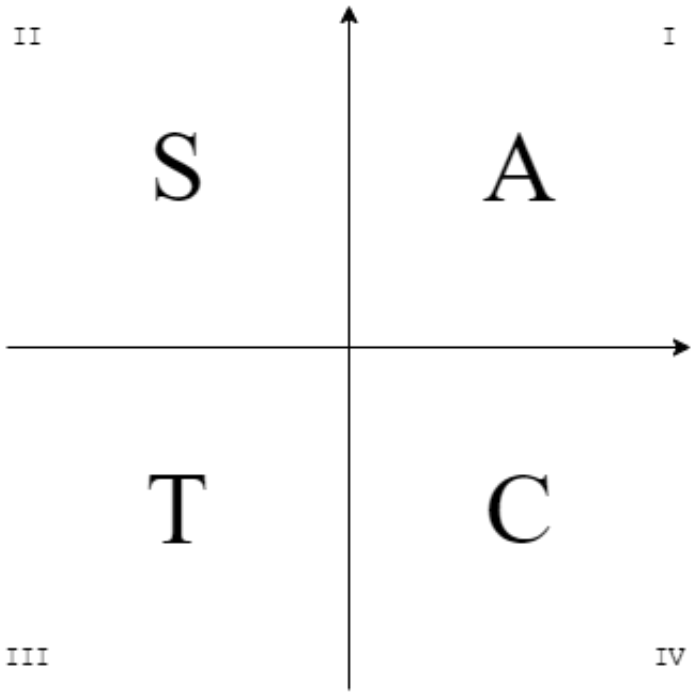
$$\begin{aligned}
&\sin(90^\circ \pm \theta) = \cos\theta, && \cos(90^\circ \pm \theta) = \mp\sin\theta, && \tan(90^\circ \pm \theta) = \mp\frac{1}{\tan\theta}\\
&\sin(180^\circ \pm \theta) = \mp\sin\theta, && \cos(180^\circ \pm \theta) = -\cos\theta, && \tan(180^\circ \pm \theta) = \pm\tan\theta\\
&\sin(270^\circ \pm \theta) = -\cos\theta, && \cos(270^\circ \pm \theta) = \pm\sin\theta, && \tan(270^\circ \pm \theta) = \mp\frac{1}{\tan\theta}\\
&\sin(360^\circ - \theta) = -\sin\theta, && \cos(360^\circ - \theta) = \cos\theta, && \tan(360^\circ - \theta) = -\tan\theta
\end{aligned}$$

However, it was too terrifying if we need to remember all the above equalities. Is there any quick method to do with?

The C-A-S-T graph and reduction strategy

Let us conclude the positivity of trigonomertric functions and see whether we can generate a useful memorization technique in order to remember the discussed reduction formulae.

By categorizing the domain of θ into different quadrant, we could see the following picture:



From the point of view of polar coordinates, we count for the first 90 degree as quadrant I; the second 90 degree as quadrant II; the third 90 degree as quadrant III; the fourth 90 degree as quadrant IV.

The discussion on different quadrant provides different sign of axis. We concludes the following.

Quadrant I is where both x and y coordinates being positive, and we can observe that for every trigonometric function in quadrant I, they are all positive. That is, for $0° < \theta < 90°$,

$$\sin\theta > 0$$
$$\cos\theta > 0$$
$$\tan\theta > 0$$

Hence, the results show that ALL trigonometric functions are positive, and we name quadrant I by 'A' to amplify the phenomenon.

Quadrant II is where only y coordinates being positive, and we can observe that for trigonometric functions in quadrant I, only sine function is positive. That is, for $90° < \theta < 180°$,

$$\sin\theta > 0$$
$$\cos\theta < 0$$
$$\tan\theta < 0$$

©Mok Owen 2024

Hence, the results show that only Sine is positive, and we name quadrant II by 'S' to amplify the phenomenon.

Quadrant III is where no coordinate is positive, and we can observe that for trigonometric functions in quadrant III, only tangent function is positive. That is, for $180° < \theta < 270°$,

$$\sin\theta < 0$$
$$\cos\theta < 0$$
$$\tan\theta > 0$$

Hence, the results show that only Tangent is positive, and we name quadrant III by 'T' to amplify the phenomenon.

Quadrant IV is where only x coordinates being positive, and we can observe that for trigonometric functions in quadrant IV, only cosine function is positive. That is, for $270° < \theta < 360°$,

$$\sin\theta < 0$$
$$\cos\theta > 0$$
$$\tan\theta < 0$$

Hence, the results show that only Cosine is positive, and we name quadrant IV by 'C' to amplify the phenomenon.

What's special about this naming technique is that we can memorize the quadrants of positivity by spelling C-A-S-T, the CAST graph, to simply guess the required sign of reduction.

Another important observation to fast-guess the reduction formula is paying attention to previous theorem, which, said

$$\sin(180° \pm \theta) = \mp\sin\theta, \qquad \cos(180° \pm \theta) = -\cos\theta, \qquad \tan(180° \pm \theta) = \pm\tan\theta$$
$$\sin(360° - \theta) = -\sin\theta, \qquad \cos(360° - \theta) = \cos\theta, \qquad \tan(360° - \theta) = -\tan\theta$$

It is important to notice that we don't need to change the functions into another functions when the leading angle is 180° or 360°. On the other hand, when the angle is not a multiple of 180, the conversion behave like

$$\sin \mapsto \cos, \cos \mapsto \sin, \tan \mapsto \frac{1}{\tan}$$

Therefore, the strategy of reduction can be summarized by following:

1. Sign-deduction: by CAST;

2. Function-deduction: by multiple of 180°.

2.6 Extending trigonometry to any degree

We known from definition of polar coordinates that we can reduce angle over 360° to less than it. The rotation deforms by subtracting or adding whole circles to the angle according to its positivity and magnitudes. We therefore access to any degree with trigonometric functions.

periodic function

Taking advantages from polar plane, we can write down an equivalence class of angle:

$$360^\circ + \theta \equiv \theta$$

and so we can write the following extension of angle formulae:

$$\sin(360^\circ + \theta) = \sin\theta$$
$$\cos(360^\circ + \theta) = \cos\theta$$
$$\tan(360^\circ + \theta) = \tan\theta$$

For the above situation, we name the 360° as the **period** of polar coordinates, and also the period of trigonometric functions sin, cos and tan. Alternatively, we say those trigonometric functions are **periodic functions** with period 360°, with following definition of periodic functions.

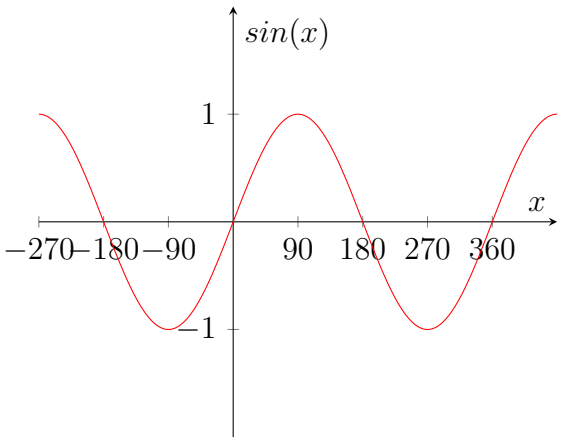
Definition 2.6 (periodic function)**.** *Let T be a number and f be a function. If for any real number x, the function f always satisfy the equation*

$$f(x+T) = f(x)$$

then we call f a periodic function with period T.

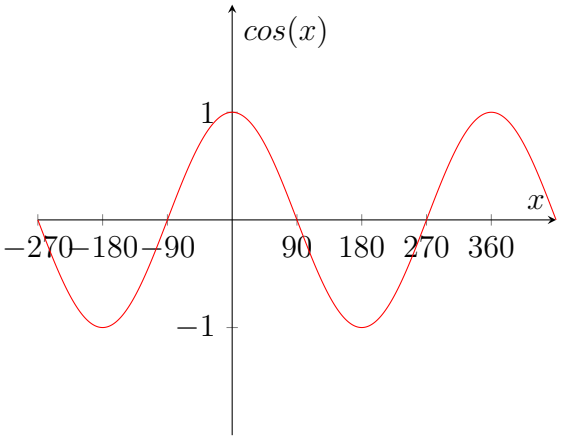
To visualize the periodicity of trigonometric functions, we plot the graphs out.

Example. *Graph of sine function:*



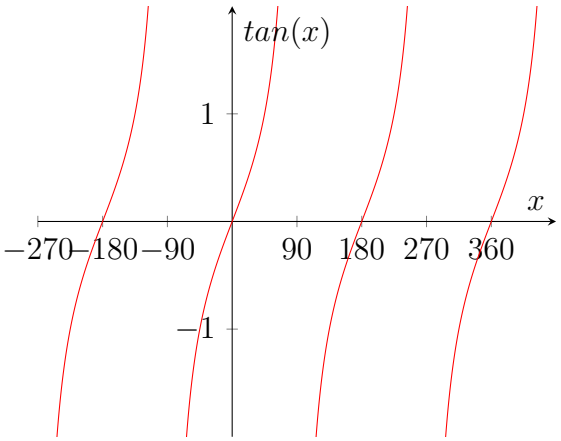
We shall observe that the sine function starts at 0, maximized and minimized at 1 *and* -1 *respectively. The value of* $\sin x$ *at* $x = 0°$ *is equal to at* $x = 360°$*, that at* $x = 90°$ *is equal to at* $x = 450°$*, etc. It looks like a copy-and-paste action to draw the curve after* $360°$*.*

Example. *Graph of cosine function:*



We shall observe that the cosine function starts at 1, maximized and minimized at 1 *and* -1 *respectively. The value of* $\cos x$ *at* $x = 0°$ *is equal to at* $x = 360°$*, that at* $x = 90°$ *is equal to at* $x = 450°$*, etc. It looks like a copy-and-paste action to draw the curve after* $360°$*.*

Example. *Graph of tangent function:*



We shall observe that the tangent function starts at 0, increasing all the time. It has no maxima or minima, however, still periodic. The value of $\tan x$ *at* $x = 0°$ *is equal to at* $x = 180°$*, and to at* $x = 360°$*, etc. Moreover, the periodicity of tangent function can be written* $180°$ *instead of* $360°$*, as it looks like a copy-and-paste action to draw the curve after* $180°$*.*

©Mok Owen 2024

Extending domain and bounded range

So far, we have periodicity of trigonometric functions, and therefore, we want to discuss one more property of sine and cosine - its boundedness.

Let us do an abstract proof for boundedness first, and read the geometric interpretation of such boundedness.

Definition 2.7 (Norm)**.** *Let f be a function. Define the norm $\|\cdot\| : C(\mathbb{R}) \to \mathbb{R}$ by the maximal absolute value of the real-valued function among its domain, i.e.*

$$\|f\| := \max\{|f(x)| : x \in D\}$$

Definition 2.8 (Bounded function)**.** *A function f is called a bounded function if its norm is finite, i.e.*

$$\|f\| \leq M$$

for some number $M > 0$.

Theorem. *Let $f : D \to \mathbb{R}$ be a periodic function with period T. Let f be bounded on some interval $a \leq x < b$ of D with $b - a = T$. Then f is a bounded function on D.*

Proof.
If f is a periodic function with period T, then the following equality holds for any $x \in D$:

$$f(x + T) = f(x)$$

Since f is bounded on the interval $a \leq x < b$, we can write

$$\|f\|_{a \leq x < b} \leq M$$

for some constant $M > 0$.

Let $x \in D$, by division algorithm, we can always write $x = nT + y$ for some $n \in \mathbb{Z}$ and $a \leq y < b$, such that

$$|f(x)| = |f(y + nT)| = |f(y)| \leq \|f\|_{a \leq x < b} \leq M$$

for any $x \in D$.

Hence, f is bounded on D. □

The theorem shows a nice generalization for boundedness of periodic functions. By the theorem, if we would like to say any periodic function is unbounded, we can check if there exist an interval I such that f is unbounded on I.

Hence we may see the boundedness of sine and cosine function:

Example. *The periodicity of sine function is by* $360°$*, so if we let* $f(x) := \sin x$*, we can see that*

$$\|f\| = \|f\|_{0° \leq x < 360°} = 1$$

Example. *The periodicity of cosine function is by* $360°$*, so if we let* $f(x) := \sin x$*, we can see that*

$$\|f\| = \|f\|_{0° \leq x < 360°} = 1$$

The above conclusion can be drawn by graphical observation, or if we take from advantage of geometric interpretation, we can perform the following deduction:

1. Both sine and cosine functions are deduced to be the representation of coordinates of a unit circle for any given angle θ;
2. By definition of a unit circle, its radius is 1;
3. That means the greatest distance in both coordinates are bounded by the value 1;
4. Hence sine and cosine function must be bounded by 1 for any given angles.

2.7 Challenging Questions

The following problems are challenging that you must have patience and guts to think about the solution for a few days. You are welcome to invite friends and teachers to discuss the validity of your solution.

1. Define $\{\theta_n\}_{n=1}^{\infty}$ to be a sequence of acute angles defined by $\theta_n := \tan^{-1} n$, where $\tan^{-1} x$ denotes the angle θ such that $\tan\theta = x$. Define another sequence of acute angles $\{\phi_n\}_{n=1}^{\infty}$ by $\phi_n := \cot^{-1} n$, where $\cot^{-1} y$ denotes the angle ϕ such that $\dfrac{1}{\tan\phi} = y$.
   (a) Prove that $\theta_n = 90° - (\frac{1}{2})^n \cdot 90°$ for all $n = 1, 2, 3, \ldots$
   (b) Hence, or otherwise, prove that $\phi_n = (\frac{1}{2})^n \cdot 90°$ for all $n = 1, 2, 3, \ldots$
   (c) Define three more sequences by following definitions:

$$\begin{cases} a_n & := \sin\theta_n, \\ b_n & := \sin\theta_n \cos\theta_n, \\ c_n & := \cos\theta_n \end{cases}$$

   i. Does there exist a positive integer n such that the equation $a_n^2 - c_n^2 = b_n$? Justify your answer.

ii. Simplify $\dfrac{a_1 + a_2 + \cdots + a_n}{c_1 + c_2 + \cdots + c_n}$.

2. Define $i \notin \mathbb{R}$ in which $i^2 = -1$. Let $\text{cis}(x) := \cos x + i \sin x : \mathbb{R} \to C$ be a function composed by sine, cosine and i. $\mathbb{R}$ denotes the real number set and C denotes the range of $\text{cis}(x)$

   (a) In rigorous mathematics, a function should be indeed satisfying existence and uniqueness. Check $\text{cis}(x)$ satisfies the mentioned properties with provided definitions:

   - Existence: For every x in the domain, there is a y in the codomain defined such that $x \mapsto y$.
   - Uniqueness: For any x in the domain, there is only one y in the codomain such that $x \mapsto y$, i.e. if $x \mapsto y_1$ and $x \mapsto y_2$ simultaneously, then $y_1 = y_2$.

   (b) Prove that $\text{cis}(x + y) = \text{cis}(x)\text{cis}(y)$. Hence, prove $(\text{cis}(x))^n = \text{cis}(nx)$.

©Mok Owen 2024

3 Linear functions

In this section, we focus on a specific form of function, which is called **Linear functions**, a type of functions that generate straight lines in the Euclidean space, or we may call it the xy-plane, denoted by $\mathbb{R}^2$. It is the simplest form of function, and has many properties we are interested in.

3.1 Fundamental concepts of points

Revisiting the xy-plane, we may define some useful tools for measurement and ratio properties, since we shall always be equipped with the sense of measuring and performing comparisons.

Distance between two points

One important practical measurement in the xy-plane is the so-called **distance between two points**. We recall that given a right-angled triangle $\triangle ABC$ with $\angle ABC = 90°$, the following relation on the side lengths holds:

$$|\overline{AB}|^2 + |\overline{BC}|^2 = |\overline{AC}|^2$$

where the stroke sign $|\cdot|$ denotes the length of a certain line and each line is mentioned explicitly using the over-lined notation $\overline{XY}$ to indicate the line connecting point X and point Y. As we learned the relation is mentioned by *Pythagoras Theorem* or *Pythagorean Theorem*.

Following from the Pythagoras Theorem, we shall examine the distance between any two points on the xy-plane by the length of the straight line connecting them. With the help of the following figure, let us carry out our definition of distance between two points.

Definition 3.1 (Distance between two points)**.** *Let $A(x_1, y_1)$ and $B(x_2, y_2)$ be two points on $\mathbb{R}^2$-plane. The distance between A and B is computed by the formula*

$$\text{dist}(A, B) := \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

We may think of the following construction for distance between two points:

Taking the right-angled triangle as distance reference, the vertical displacement becomes $|y_1 - y_2|$ and the horizontal displacement becomes $|x_1 - x_2|$. Putting the oblique displacement $\text{dist}(A, B)$ gives the Pythagorean relation, and we could get the desired formula.

Example. *Let $A(1, 2)$ and $B(4, -2)$ be two points on the rectangular coordinate plane, i.e. the xy-plane. Then,*

$$\text{dist}(A, B) = \sqrt{(1 - 4)^2 + (2 - (-2))^2} = \sqrt{3^2 + 4^2} = \sqrt{25} = 5$$

Example. *Let* $A(1,2)$ *and* $B(k,k-1)$ *be two points on the rectangular coordinate plane such that* $\text{dist}(A,B)=2$. *Then,*

$$\begin{aligned}\sqrt{(k-1)^2+(k-3)^2}&=2\\(k-1)^2+(k-3)^2&=4\\k^2-2k+1+k^2-6k+9&=4\\2k^2-8k+6&=0\\k^2-4k+3&=0\\(k-1)(k-3)&=0\end{aligned}$$

Then $k=1$ *or* $k=3$, *i.e. the possible coordinates of* B *are* $(1,0)$ *or* $(3,2)$.

Example. *Given two trajectories* $T_1:(x_1(t),y_1(t))$ *and* $T_2:(x_2(t),y_2(t))$, *we say the distance between* T_1 *and* T_2 *is*

$$\text{dist}(T_1,T_2):=\min_{t>0}\sqrt{(x_1(t)-x_2(t))^2+(y_1(t)-y_2(t))^2}$$

It is quite abstract to talk about the distance between two trajectories in this stage. Just keep in mind that this is a truth, or say you may believe in it.

Essential Practice 3.1.1. *Compute the distance between following pair of points:*

1. $(1,0)$ *and* $(0,1)$;
2. $(1,-1)$ *and* $(-1,1)$;
3. $(3,4)$ *and* $(12,-13)$;
4. $(7,-8)$ *and* $(-3,0)$;

Essential Practice 3.1.2. *Given the distance between two points, find the possible value(s) of the unknowns:*

1. $A(0,0),B(3,k)$; $\text{dist}(A,B)=3$.
2. $A(2,0),B(3,k)$; $\text{dist}(A,B)=1$.
3. $A(k,0),B(0,k)$; $\text{dist}(A,B)=5$.
4. $A(3+k,1-k),B(3,-3)$; $\text{dist}(A,B)=8$.

Mid-point and Division points

Definition 3.2 (Mid-way of a trajectory)**.** *Given a trajectory defined by $(x(t), y(t))$ where $t \in [0, r]$ with $r > 0$ a real number, the **midway** (or **half-way**) of the trajectory is given by the point M such that*

$$\text{arclength}((x(0), y(0)), M) = \text{arclength}(M, (x(r), y(r)))$$

In particular, if we are talking about a mid-point, we usually pick the shortest trajectory between the two points. This is due to the uniqueness of the shortest trajectory if our distance is well-defined. In the Euclidean space, the shortest trajectory is usually the straight line segment connecting the two points.

Theorem (Mid-point of two points)**.** *Given $A(x_1, y_1)$ and $B(x_2, y_2)$ be two points on the Euclidean space, the **mid-point** M of A and B is defined by*

$$M := (\frac{x_1 + x_2}{2}, \frac{y_1 + y_2}{2})$$

Proof.
Let L be the straight line connecting A and B. Assume that $M := (\frac{x_1 + x_2}{2}, \frac{y_1 + y_2}{2})$, then

$$\begin{aligned}
\text{dist}(A, M) &= \sqrt{(x_1 - \frac{x_1 + x_2}{2})^2 + (y_1 - \frac{y_1 + y_2}{2})^2} \\
&= \frac{1}{2}\sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2} \\
&= \frac{1}{2}\text{dist}(A, B) \\
\text{dist}(M, B) &= \sqrt{(\frac{x_1 + x_2}{2} - x_2)^2 + (\frac{y_1 + y_2}{2} - x_2)^2} \\
&= \frac{1}{2}\sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2} \\
&= \frac{1}{2}\text{dist}(A, B)
\end{aligned}$$

In particular,

$$\text{dist}(A, M) + \text{dist}(M, B) = \text{dist}(A, B)$$

Thus M preserve a point on the straight line connecting A and B and in fact the mid-point of A and B. □

Remark. *The memorization of the computation can be written as $M = \frac{A+B}{2}$ to abstractize the similarity in both coordinates.*

Example. *Given $A(100, 20)$ and $B(30, 70)$, their mid-point is the point $M(65, 45)$.*

Essential Practice 3.1.3. *For the following given pairs of points, find their mid-points:*

1. $(2, 2)$ *and* $(0, 0)$*;*
2. $(2, 1)$ *and* $(-2, -1)$.

Other than the mid-point, we also consider the general version of line separation, which is the point of division.

Definition 3.3 (Point of division)**.** *Given two points A and B, any point P lies in between A and B is called* ***the point of division with ratio*** $\text{dist}(A, P) : \text{dist}(P, B)$.

Theorem (Formula for point of division)**.** *Let $A(x_1, y_1)$ and $B(x_2, y_2)$ be a point on the Euclidean space. Let P be the point of division of A and B with ratio $r : s$. Then the coordinates of P is given by*

$$\frac{sA + rB}{r + s} = (\frac{sx_1 + rx_2}{r + s}, \frac{sy_1 + ry_2}{r + s})$$

In particular, the mid-point of A and B is given by setting $s = r = 1$.

Proof.
Similar to the proof of mid-point. Let $P := (\frac{sx_1 + rx_2}{r + s}, \frac{sy_1 + ry_2}{r + s})$, then

$$\begin{aligned}
\text{dist}(A, P) &= \sqrt{(x_1 - \frac{sx_1 + rx_2}{r + s})^2 + (y_1 - \frac{sy_1 + ry_2}{r + s})^2} \\
&= \frac{1}{r + s}\sqrt{(rx_1 - rx_2)^2 + (ry_1 - ry_2)^2} \\
&= \frac{r}{r + s}\text{dist}(A, B) \\
\text{dist}(P, B) &= \sqrt{(\frac{sx_1 + rx_2}{r + s} - x_2)^2 + (\frac{sy_1 + ry_2}{r + s} - x_2)^2} \\
&= \frac{1}{r + s}\sqrt{(sx_1 - sx_2)^2 + (sy_1 - sy_2)^2} \\
&= \frac{s}{r + s}\text{dist}(A, B)
\end{aligned}$$

In particular,

$$\text{dist}(A, P) + \text{dist}(P, B) = \text{dist}(A, B)$$

Thus P preserve a point on the straight line connecting A and B and satisfies $\text{dist}(A, P) : \text{dist}(P, B) = \frac{r}{r + s} : \frac{s}{r + s} = r : s$. □

Example. *Let $A(2,4)$ and $B(-4,7)$. To trisect the line connecting A and B, we need two points on the line where they divide the straight line in the ratio of $1:1:1$. In particular, let the point closer to A be X and the other be Y, we have X the division point of ratio $1:2$ and Y the division point of ratio $2:1$. Therefore,*

$$X := (\frac{2(2)+1(-4)}{1+2}, \frac{2(4)+1(7)}{1+2}) = (0,5)$$
$$Y := (\frac{1(2)+2(-4)}{1+2}, \frac{1(4)+2(7)}{1+2}) = (-2,6)$$

Essential Practice 3.1.4. *Given $A(-100,-10)$ and $B(100,100)$. Find the point of division for the following given ratio $r:s$:*

1. *$r:s=1:1$;*
2. *$r:s=1:3$;*
3. *$r:s=3:7$;*
4. *$r:s=19:91$.*

Slope of two points

For two points, we want to know not only how far away they are, but also how steep they see each other, i.e. the direction of proceeding. We call it the slope of the two points.

To be professional, we say a slope of the two points is how line goes from the left point to the right point; that is, whether y increase with x or y decrease with x.

Definition 3.4 (Slope)**.** *Let $A(x_1,y_1)$ and $B(x_2,y_2)$ be two points on the Euclidean space. Then the **slope** of A and B, denoted by $\mathfrak{M}_{AB}$, is given by*

$$\mathfrak{M}_{AB} := \frac{y_2-y_1}{x_2-x_1}$$

Example. *The slope of $A(1,2)$ and $B(4,1)$ is $\mathfrak{M}_{AB} = \dfrac{1-2}{4-1} = -\dfrac{1}{3}$.*

We could observe the following fact. For if two points are aligned horizontally on the plane, then their slope is computed equal to 0; if the right point is higher than the left, then the slope is positive, meaning the corresponding line is increasing with x; otherwise the slope is negative, meaning the corresponding line is decreasing with x. For if two points are aligned vertically on the

plane, then the slope is undefined, since we could not explicitly tell whether the corresponding line is increasing or decreasing - it could be both. Due to the uncertainty, we identify it as undefined slope. However, we can still define the slope of vertical lines for consistent usage, but that's another page of discussion.

Essential Practice 3.1.5. *Compute the slope of the following pairs of points:*

1. $(1, 0)$ *and* $(0, 1)$*;*
2. $(1, -1)$ *and* $(-1, 1)$*;*
3. $(3, 4)$ *and* $(12, -13)$*;*
4. $(7, -8)$ *and* $(-3, 0)$*;*

Definition 3.5 (Collinear)**.** *3 points are called* ***collinear*** *if there is a straight line passes through 3 points at the same time.*

In fact, collinearity is somehow the converse of computing points of division, by the definition of point of division. We will make use of the equivalence in proving the following theorem:

Theorem. *3 points on the plane are collinear if and only if the slope of any two of the three points are the same.*

Proof.
For the direction from collinear to slope equality, we have if 3 points are collinear, then they can be written as

$$A(x_1, x_2), B(\frac{sx_1 + rx_2}{r+s}, \frac{sy_1 + ry_2}{r+s}), C(x_2, y_2)$$

We then have

$$\begin{aligned}
\mathfrak{M}_{AB} &= \frac{\frac{sy_1+ry_2}{r+s} - y_1}{\frac{sx_1+rx_2}{r+s} - x_1} \\
&= \frac{sy_1 + ry_2 - (r+s)y_1}{sx_1 + rx_2 - (r+s)x_1} \\
&= \frac{ry_2 - ry_1}{rx_2 - rx_1} \\
&= \frac{y_2 - y_1}{x_2 - x_1} \\
\mathfrak{M}_{BC} &= \frac{y_2 - \frac{sy_1+ry_2}{r+s}}{x_2 - \frac{sx_1+rx_2}{r+s}} \\
&= \frac{(r+s)y_2 - sy_1 + ry_2}{(r+s)x_2 - sx_1 + rx_2} \\
&= \frac{sy_2 - sy_1}{sx_2 - sx_1} \\
&= \frac{y_2 - y_1}{x_2 - x_1} \\
\mathfrak{M}_{AC} &= \frac{y_2 - y_1}{x_2 - x_1}
\end{aligned}$$

Then collinear implies the equality of the slopes.

Conversely, suppose 3 points A, B, C such that $\mathfrak{M}_{AB} = \mathfrak{M}_{BC} = \mathfrak{M}_{AC} = m$. We may compute that

$$\begin{aligned}
\text{dist}(A, B) &= \sqrt{(x_A - x_B)^2 + (y_A - y_B)^2} \\
&= |x_A - x_B|\sqrt{1 + \mathfrak{M}_{AB}^2}
\end{aligned}$$

Similarly, we have

$$\begin{aligned}
\text{dist}(B, C) &= |x_B - x_C|\sqrt{1 + \mathfrak{M}_{BC}^2} \\
\text{dist}(A, C) &= |x_A - x_C|\sqrt{1 + \mathfrak{M}_{AC}^2}
\end{aligned}$$

Without loss of generality, we may assume their x-coordinates are in fixed order, where I will set $x_A < x_B < x_C$. Then

$$\begin{aligned}
\text{dist}(A, B) + \text{dist}(B, C) &= (|x_A - x_B| + |x_B - x_C|)\sqrt{1 + m^2} \\
&= (x_B - x_A + x_C - x_B)\sqrt{1 + m^2} \\
&= |x_A - x_C|\sqrt{1 + \mathfrak{M}_{AC}^2} \\
&= \text{dist}(A, C)
\end{aligned}$$

This shows that B is a point of division of A and C, and hence A, B, C are collinear. □

Corollary. *3 points on the plane are not collinear if and only if the slope of any two of the three points are not the same.*

By the theorem, we could easily show the collinearity of any 3 points, and of course we may push a little forward from it.

Example. *Determine the collinearity of following sets of points:*

1. $(0,0),(1,2),(2,4)$.
   Sol. Denote $(0,0),(1,2),(2,4)$ *by* A,B,C *respectively. Then*
   $$\mathfrak{M}_{AB} = \frac{2-0}{1-0} = 2$$
   $$\mathfrak{M}_{BC} = \frac{4-2}{2-1} = 2$$
   Then, by the previous theorem, A,B,C *are collinear.*

2. $(0,1),(1,2),(2,2)$.
   Sol. Denote $(0,0),(1,2),(2,4)$ *by* A,B,C *respectively. Then*
   $$\mathfrak{M}_{AB} = \frac{2-1}{1-0} = 2$$
   $$\mathfrak{M}_{BC} = \frac{2-2}{2-1} = 0$$
   Then, by the previous theorem, A,B,C *are not collinear.*

Essential Practice 3.1.6. *Determine the collinearity of the following sets of points:*

1. $(0,0),(1,3),(2,6)$.
2. $(1,0),(0,-1),(-1,-2)$.
3. $(0,2),(1,-3),(-2,6)$.
4. $(k,4),(4+k,4+k),(4,k)$.

This measurement provides another way of line construction. In fact, by one of its powerful property, there is a generalized version of slope with various application in higher mathematics and other fields including computer science.

Theorem (Invariant properties of slope of two points)**.** *Let* A *and* B *be two points on the Euclidean space, and* $\mathfrak{M}_{AB}$ *be the slope of* A *and* B*. Under any simultaneous translation to* A *and* B*,* $\mathfrak{M}_{AB}$ *remains unchanged.*

Proof.
Let $A(x_1, y_1)$ and $B(x_2, y_2)$ be two points. Suppose A' and B' be the resulting coordinates after simultaneously translating A and B, we may write $A'(x_1+\Delta x, y_1+\Delta y)$ and $B'(x_2+\Delta x, y_2+\Delta y)$. Then the slope after simultaneous translation is

$$\mathfrak{M}_{A'B'} = \frac{(y_2+\Delta y)-(y_1+\Delta y)}{(x_2+\Delta x)-(x_1+\Delta x)} = \frac{y_2-y_1}{x_2-x_1} = \mathfrak{M}_{AB}$$

□

3.2 Different forms of a linear function

Following the concepts of distance, point of division and slope, we can now discuss the form of a linear function.

Segment

We may first consider a segment in the plane, which is the straight line connecting any two points in the Euclidean plane, ends at that two points.

Definition 3.6 (Segment)**.** *A **segment** with two endpoints A and B is the collection of all division points between A and B. In other words, it is a straight line with endpoints A and B.*

By the definition of division point, we may let $A(x_1, y_1)$ and $B(x_2, y_2)$ be two points, and two numbers r and s such that every division point of ratio $r : s$ could be represented in the following form

$$(\frac{sx_1+rx_2}{r+s}, \frac{sy_1+ry_2}{r+s})$$

By simple algebra, we can see that $\dfrac{s}{r+s} = 1 - \dfrac{r}{r+s}$, so we may let $t := \dfrac{r}{r+s}$ be a number in between 0 and 1 such that every division point can be written as

$$((1-t)x_1+tx_2, (1-t)y_1+ty_2) = (x_1+t(x_2-x_1), y_1+t(y_2-y_1))$$

with $0 \leq t \leq 1$. Now, we could see the relationship between x- and y- coordinates in the following

derivation:

$$\begin{aligned} y :&= (1-t)y_1 + ty_2 \\ &= y_1 + t(y_2 - y_1) \\ &= y_1 + t\frac{y_2 - y_1}{x_2 - x_1}(x_2 - x_1) \\ &= y_1 + t\mathfrak{M}_{AB}(x_2 - x_1) \\ &= y_1 + \mathfrak{M}_{AB}(x - x_1) \end{aligned}$$

Line as extension of a segment

We have established that a segment can be represented by the equation

$$y = y_1 + \mathfrak{M}_{AB}(x - x_1)$$

where it is interesting to see the equation involves no variable t. In fact, it is independent of t and could be extended to out of the segment. We call this extension a **straight line** passes through A and B.

The above equation looks simple, but let us confirm that the derived equation is indeed satisfying the extension of a segment. We shall return to the fundamental form of a segment but revising the domain of t:

$$((1-t)x_1 + tx_2, (1-t)y_1 + ty_2)$$

where $-\infty < t < \infty$. In fact, a line should be proven to be this form rigorously. Let us dive into it.

Theorem. *A straight line passes through (x_1, y_1) and (x_2, y_2) is in the form of*

$$((1-t)x_1 + tx_2, (1-t)y_1 + ty_2)$$

where $-\infty < t < \infty$.

Proof.
The major discussion is the domain of t diverges to infinity. Let us prove first that for every real value t, the point $((1-t)x_1 + tx_2, (1-t)y_1 + ty_2)$ is collinear with $A(x_1, y_1)$ and $B(x_2, y_2)$. In fact,

$$\mathfrak{M} = \frac{(1-t)y_1 + ty_2 - y_1}{(1-t)x_1 + tx_2 - x_1} = \frac{y_2 - y_1}{x_2 - x_1}$$

is true for all $t \neq 0$. At the same time, the case $t = 0$ is just considering (x_1, y_1), which is immediately true. Hence, every point in that form is collinear with A and B.

Next, we shall show that t diverges at both ends. Suppose on contrary that t is bounded above by M, then the longest segment ends at $V_M((1-M)x_1+Mx_2,(1-M)y_1+My_2)$, but letting $V_{M+1}((-M)x_1+(M+1)x_2,(-M)y_1+(M+1)y_2)$

$$\text{dist}(A,V_{M+1})=(M+1)\text{dist}(A,B)>M\text{dist}(A,B)=\text{dist}(A,V_M)$$

This contradicts that AV_M is the longest segment we have. Thus t is unbounded. The argument is similar for t's lower bound and hence $-\infty<t<\infty$. □

Point-slope form

Observe that

$$y=y_1+\mathfrak{M}_{AB}(x-x_1)\iff y-y_1=\mathfrak{M}_{AB}(x-x_1)\iff\frac{y-y_1}{x-x_1}=\mathfrak{M}_{AB}$$

for given points $A(x_1,y_1)$ and slope $\mathfrak{M}_{AB}$, we may give the following result.

Theorem. *If a line L passes through point $A(x_1,y_1)$ and having slope $\mathfrak{M}$, then L can be constructed from writing*

$$L:\frac{y-y_1}{x-x_1}=\mathfrak{M}$$

The meaning of this form can be thought of as: Given any point (x,y) on the line L, its slope with A is the same as the given slope $\mathfrak{M}$.

Example. *Suppose L is a straight line passes through $(0,0)$ with slope being equal to 3. Then L could be modified as*

$$L:\frac{y-0}{x-0}=3\iff y=3x$$

Example. *Suppose L is a straight line passes through $(1,2)$ with slope being equal to 4. Then L could be modified as*

$$L:\frac{y-2}{x-1}=4\iff y=4x-2$$

Two-point form

With any two given points on the Euclidean plane, we could rewrite the slope $\mathfrak{M}$ into $\dfrac{y_2-y_1}{x_2-x_1}$. This means the line passes through any two given points could be modelled by

Theorem. *If a line L passes through point $A(x_1,y_1)$ and $B(x_2,y_2)$, then L can be constructed from writing*

$$L:\frac{y-y_1}{x-x_1}=\frac{y_2-y_1}{x_2-x_1}$$

©Mok Owen 2024

Example. *Suppose L is a straight line passes through $(0,0)$ and $(1,2)$. Then L could be modified as*

$$L : \frac{y-0}{x-0} = \frac{2-0}{1-0} \iff y = 2x$$

Example. *Suppose L is a straight line passes through $(1,2)$ and $(10,3)$. Then L could be modified as*

$$L : \frac{y-2}{x-1} = \frac{3-1}{10-2} \iff y = \frac{1}{4}x + \frac{7}{4}$$

Slope-intercept form

Given the slope $\mathfrak{M}$ and the y-intercept c, we could take the y-intercept as one particular coordinate $(0,c)$. This means the line with given information is

Theorem. *If a line L has slope $\mathfrak{M}$ and y-intercept c, then L can be constructed from writing*

$$L : \frac{y-c}{x-0} = \mathfrak{M}$$

Example. *Suppose L is a straight line with slope 5 and y-intercept 0. Then L could be modified as*

$$L : \frac{y-0}{x-0} = 5 \iff y = 5x$$

Example. *Suppose L is a straight line with slope $\frac{1}{10}$ and y-intercept 5. Then L could be modified as*

$$L : \frac{y-5}{x-0} = \frac{1}{10} \iff y = \frac{1}{10}x + 5$$

Intercepts form

Except the previous forms, an intercepts form is a form purely taken from the two intercepts. Let a be the x-intercept and b be the y-intercept of a line L. We see from the view of division point, the line shall be formed by writing

$$((1-t)a, tb)$$

for all the points on L, where $-\infty < t < \infty$. We could derive that

$$\frac{x}{a} + \frac{y}{b} = 1$$

for any value of t.

Theorem. *Let L be having x-intercept a and y-intercept b. Then L could be modelled by*

$$L : \frac{x}{a} + \frac{y}{b} = 1$$

©Mok Owen 2024

Example. *Suppose L has x-intercept 5 and y-intercept 5. Then L could be modified as*

$$L : \frac{x}{5} + \frac{y}{5} = 1 \iff y = 5 - x$$

Example. *Suppose L is a straight line with x-intercept k and no y-intercept. Then L could be modified as*

$$L : \frac{x}{k} = 1 \iff x = k$$

Example. *Suppose L is a straight line with y-intercept k and no x-intercept. Then L could be modified as*

$$L : \frac{y}{k} = 1 \iff y = k$$

Example. *Suppose L is a straight line completely lies on x-axis. This means x-intercept is undefined. Then L could be modified as*

$$L : y = 0$$

Example. *Suppose L is a straight line completely lies on y-axis. This means y-intercept is undefined. Then L could be modified as*

$$L : x = 0$$

General form

In convenience to solving linear equations, we aim to set a form that is always constructive. We put our attention to what a general form is.

Recall that a straight line could be either a vertical line, an oblique line, or a horizontal line. We shall introduce a form that is generally true to write a straight line. This thought should eliminate the presence of slope. Perhaps the intercepts form provides a good start to write a general form. Indeed,

$$\frac{x}{a} + \frac{y}{b} = 1 \iff bx + ay = ab$$

gives us the form purely by multiplication and addition. This can so be rearranged such that

$$bx + ay = ab \iff Ax + By + C = 0$$

with $A = b$, $B = a$ and $C = -ab$. Of its elimination of slope, this is the desired form of a straight line.

Theorem. *Any straight line L can be written in the form of*

$$Ax + By + C = 0$$

where A, B, C are constants. Furthermore, the slope of L is determined $-\frac{A}{B}$, the x-intercept $-\frac{C}{A}$ and y-intercept $-\frac{C}{B}$.

Proof.
The first part of the proof is to determine the connection between different forms of equation of straight line. Indeed, the point-slope form connects with two-point form, while the point-slope form also connects with slope intercept form. The general form comes form intercepts form, so it remains to show the connection between slope intercept form and general form. In fact,

$$\begin{aligned} y = \mathfrak{M}x + c &\iff \mathfrak{M}x - y + c = 0 \\ &\iff Ax + By + C = 0 \end{aligned}$$

shows the connection between slope-intercept form and general form. This completes the first part.

The second part of the proof should show the "furthermore" implication. Consider the definitions of x-intercept and y-intercept, we will see if x_0 denotes the x-intercept, then $Ax_0+B(0)+C = 0$ implies $x_0 = -\frac{C}{A}$; Similarly, denoting y_0 the y-intercept, we have $y_0 = -\frac{C}{B}$. For the slope we take the provided intercepts and write

$$\mathfrak{M} = \frac{y_0 - 0}{0 - x_0} = -\frac{A}{B}$$

concludes the proof. □

It is the best to write the general form with integers instead of fractions or decimal numbers, since it is more convenient and intuitive to solve equations using integers. In general, if we have to handle irrational numbers, we prefer to write everything without quotient and shall be simplest.

Example. *It is better to write* $x + 2y + 3 = 0$ *instead of* $\frac{1}{2}x + y + \frac{3}{2} = 0$.

Example. *If we need to handle irrationals, say* $\frac{\sqrt{3}}{2} + \sqrt{5}y + \frac{3}{\sqrt{7}} = 0$*, it is more desired to write* $\sqrt{21}x + 2\sqrt{35}y + 6 = 0$.

Essential Practice 3.2.1. *Find the straight line in general form with given conditions.*

1. *Of slope* 2 *and passes through* $(1, 2)$.
2. *Of slope* $-\frac{2}{3}$ *and passes through* $(1, -1)$.
3. *Passes through* $(1, 3)$ *and* $(-2, 9)$.
4. *Passes through* $(1, 0)$ *and* $(-2, 0)$.
5. *Of slope* 1 *and y-intercept* 3.

6. *Of slope* $\frac{5}{7}$ *and y-intercept* $\frac{2}{3}$.

7. *x-intercept* 1 *and y-intercept* -2.

8. *x-intercept* $-\frac{1}{2}$ *and y-intercept* $\frac{1}{3}$.

Essential Practice 3.2.2. *Determine the slope, the x-intercept and the y-intercept for the following straight lines.*

1. $x + y + 1 = 0$.

2. $3x + 4y + 9 = 0$.

3. $2x = 4$.

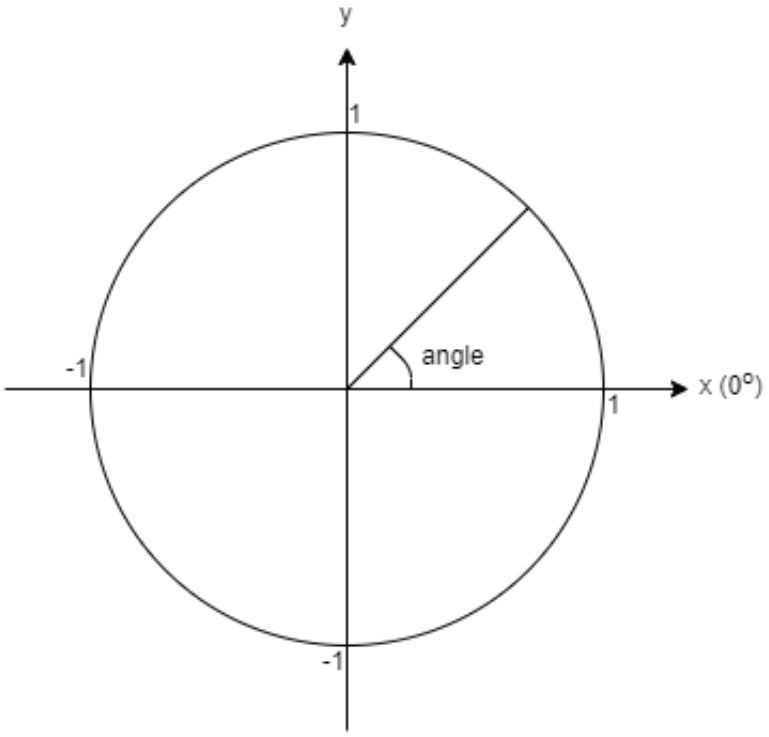
4. $7y = 1$.

3.3 Rotational geometry

So far, we have determined the slope of a straight line which relates to the direction of prosession very much, and we aim to discuss more about the real-world application of slope of straight line. We will take inspiration from polar coordinate and circular trigonometry.

Any straight line has a unique x-intercept, and we will make use of the angle at that point. Recall from trigonometry, the value of the tangent function at some point agrees with the slope from origin to that point. Let $\mathfrak{M}$ be the slope of the line L which *passes through origin*. Then it is not hard to see

$$\tan\theta = \mathfrak{M} = \frac{y}{x}$$



If we see $\tan\theta$ as a function of x and y, let us write $\theta(x,y) = \tan^{-1}(\frac{y}{x})$ to define the method of finding the angle between the horizontal line and L. If we need to consider any line on the plane, we may simply do a shift on L to find the required angle, but the slope function of a line is in fact invariant under translation, i.e. does not change its value when translation is done. Hence, it is satisfying to see the following result

Theorem. *The angle between the line L passes through any two points (a_1, b_1) and (a_2, b_2) is given by*

$$\theta = |\arctan\frac{b_1 - b_2}{a_1 - a_2}|$$

Angle of elevation from horizon

Angle of depression from horizon

Angle between two arbitrary lines

Rotation of points and lines

3.4 Relation between lines

Point of intersection

Parallel lines

Perpendicular lines

Number of intersections

Angle bisector

3.5 Additional content: Point-line distance

By definition, the distance between two points on a $\mathbb{R}^2$ plane is

Definition 3.7 (Distance between two points)**.** *Let $A(x_1, y_1)$ and $B(x_2, y_2)$ be two points on $\mathbb{R}^2$-plane. The distance between A and B is computed by the formula*

$$\text{dist}(A, B) := \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

Now let $L : ax + by + c = 0$ be a straight line and $P(x_0, y_0)$ be a point on $\mathbb{R}^2$-plane. Note that with the following axiom

Axiom. *The distance between a point P and a line L is defined by the shortest distance between P and a point on L*

$$\text{dist}(P, L) := \inf\{\text{dist}(P, Q) : Q \in L\}$$

we can choose the perpendicular displacement of P from L to define the distance. Using the line Γ perpendicular to L passing through P, we can find such Q by following computation:

$$\begin{cases} L : ax + by + c = 0 \\ \Gamma : bx - ay + (ay_0 - bx_0) = 0 \end{cases} \implies Q(\frac{b^2x_0 - aby_0 - ac}{a^2 + b^2}, \frac{a^2y_0 - abx_0 - bc}{a^2 + b^2})$$

Therefore, the distance between P an Q is

$$\begin{aligned}
\text{dist}(P,Q) &= \frac{1}{a^2+b^2}\sqrt{[(a^2+b^2)x_0-(b^2x_0-aby_0-ac)]^2+[(a^2+b^2)y_0-(a^2y_0-abx_0-bc)]^2} \\
&= \frac{1}{a^2+b^2}\sqrt{(a^2x_0+aby_0+ac)^2+(b^2y_0+abx_0+bc)^2} \\
&= \frac{\sqrt{a^2+b^2}\sqrt{(ax_0+by_0+c)^2}}{a^2+b^2} \\
&= \frac{\sqrt{(ax_0+by_0+c)^2}}{\sqrt{a^2+b^2}} \\
&= \left|\frac{ax_0+by_0+c}{\sqrt{a^2+b^2}}\right|
\end{aligned}$$

3.6 Linear inequalities

One variable inequality

Two variable inequality

Linear programming

3.7 Challenging questions

4 Quadratic functions

In ancient times, someone asked about 'solving the side length of a square for which area equals 2'. Recall the formula for computing area of a square is: given the side length of a square to be ℓ, then the area A of the square is

$$A = \ell^2$$

To tackle such problem, the ancient greeks tries the process of reversing the square - the square root of some numbers. They created a symbol Γ for halving the square multiplication, as the meaning of a square root. They first observed for integers

$$\begin{aligned} 1^2 = 1 &\implies \sqrt{1} = 1 \\ 2^2 = 4 &\implies \sqrt{4} = 2 \\ 3^2 = 9 &\implies \sqrt{9} = 3 \\ &\vdots \end{aligned}$$

And for the rest of the integral square they call them **perfect sqaures**. However, 2 is not a perfect square, and they struggled for years which number can be used to represent the value of $\sqrt{2}$. For historical and religous reason, they could only think of rational numbers to tryout. By many trials, decimal values can be approximated for $\sqrt{2}$, like 1.41, 1.4142, etc. But the first one who claimed $\sqrt{2}$ is irrational, he slept silently in the ground by many nice faith.

Today, we know that there are many numbers that could not be expressed by a fraction, which are called irrational numbers.

4.1 Solving Quadratic Equations

Let us once again introduce the concept of taking sqaure root formally.

Definition 4.1 (Surds)**.** *The* ***radical sign*** $\sqrt{}$ *is a function that helps defines the positive solution to sqaure equation. If*

$$x^2 = a$$

then $x = \sqrt{a}$ *is a positive solution to the equation.*

Example. *If* $x^2 = 6$*, then* $x = \sqrt{6}$ *is a solution to the equation.*

Example. *If* $y^2 = 11$*, then* $y = \sqrt{11}$ *is a solution to the equation.*

We shall notice the definition only talks about it is a solution, but not the only solution. In fact, we have to take care of the negatives. We notice by some fundamental result in Algebra, $(-1)(-1) = 1$. This means -1 is also a solution to the equation $x^2 = 1$.

Therefore, it is needed to say when $x^2 = a > 0$, the solutions can be $x = \sqrt{a}$ or $x = -\sqrt{a}$. The equation contains at least two results. But is it possible to have more?

Proposition. *For the equation $x^2 = a > 0$, we have at most 2 distinct solutions to the equation.*

Proof.
Assume α, β, γ are all distinct solutions to the equation $x^2 = a$. Then $\alpha^2 = \beta^2 = \gamma^2$. By their distinction and the completeness of real numbers, we can write without loss of generality that $\alpha < \beta < \gamma$.

When $0 < \alpha < \beta < \gamma$, $\alpha^2 < \alpha\beta < \beta^2$ leads to contradiction of the equality.

When $\alpha < 0 < \beta < \gamma$, the relation between β and γ contradicts the equality as in previous case.

When $\alpha < \beta < 0 < \gamma$, the relation between β and α contradicts the equality as in previous case.

When $\alpha < \beta < \gamma < 0$, the relation between β and γ contradicts the equality as in previous case.

Therefore, it is solid to conclude by contradiction, that $x^2 = a > 0$ cannot have 3 distinct solutions, and it restricts to at most 2 distinct solutions. □

To conclude, we could now say for the equation $x^2 = a > 0$, either $x = \sqrt{a}$ or $x = -\sqrt{a}$. Moreover, the stated solutions are the only possibilities.

We write for convenience concatenating plus sign $+$ and minus sign $-$ to become the plus-or-minus sign $\pm$, so that we can shoerten our writings:

If $x^2 = a > 0$, then $x = \pm\sqrt{a}$.

Example. *If $x^2 = 16$, then $x = \pm 4$.*

Example. *If $y^2 = 11$, then $x = \pm\sqrt{11}$.*

We also deserve simplicity in writing surds so that computation can be simpler. We return to its meaning, that $\sqrt{a}$ means it is the positive solution to the equation $x^2 = a$. If a is a composite number so that part of its factor is a perfect square, then we may observe the following:

$$x^2 = k^2 p$$

Let's take a look at $x^2 = ab$ first. Note that by commutativity of multiplication $(\sqrt{a}\sqrt{b})^2 = \sqrt{a}\sqrt{b}\sqrt{a}\sqrt{b} = ab$. Then $\pm\sqrt{a}\sqrt{b}$ are two possible solutions to the equation; on the other hand, the direct square root $\pm\sqrt{ab}$ also satisfies. We have proved earlier that the equation must have at most 2 distinct solutions, so these 4 solutions must be pairwise identical. Note that partial square-root $(\sqrt{a}\sqrt{b})$ and direct square-root $(\sqrt{ab})$ are different strategies, but both provide 2 solutions following the plus-or-minus distinction. By the positivity of radical sign, we need to have the following idetification:

$$\sqrt{ab} = \sqrt{a}\sqrt{b}$$

Returning to the original problem, when a number is composite and part of its factor is a perfect square, we can choose to factor out the perfect square part and complete the simplification as

$$\sqrt{k^2 p} = \sqrt{k^2}\sqrt{p} = k\sqrt{p}$$

Example. $\sqrt{12} = \sqrt{4}\sqrt{3} = 2\sqrt{3}$.

Example. $\sqrt{18} = \sqrt{9}\sqrt{2} = 3\sqrt{2}$

The main reason is to facilitate the simplicity in computation. Consider

$$\sqrt{75} + \sqrt{12} = 5\sqrt{3} + 2\sqrt{3} = 7\sqrt{3} = \sqrt{147}$$

and

$$\sqrt{75} \cdot \sqrt{12} = 5\sqrt{3} \cdot 2\sqrt{3} = 30$$

which are seemingly impossible but indeed true, the result are amazing with this methodology. It turns out simplification is always the required steps to clean things up.

Notice it is unclear to read a fraction when the denominator is not an integer, or say it is hard to compare the values between two fractions when the situation was like that. Instead, we aim to write every fraction with integral denominator, so that the comparison is more convenient and direct to intuition.

We shall here introduce a method of fraction modification called **rationalization**.

Definition 4.2 (Rationalization)**.** *A category of method to modify fractions with non-integral denominator.*

There should be at least 3 situations to consider rationalization.

When the denominator includes only a surd, the rationalization be like

Example. *Rationalize* $\frac{1}{\sqrt{2}}$.

Solution.

$$\begin{aligned}\frac{1}{\sqrt{2}} &= \frac{1}{\sqrt{2}} \cdot \frac{\sqrt{2}}{\sqrt{2}} \\ &= \frac{\sqrt{2}}{2}\end{aligned}$$

. . . end of solution

When the denominator includes two terms, we need to recall the identity

$$a^2 - b^2 \equiv (a-b)(a+b)$$

where it turns each terms into terms of sqaure. The application be like

Example. *Rationalize* $\frac{\sqrt{2}}{\sqrt{3}-1}$.

Solution.

$$\begin{aligned}\frac{\sqrt{2}}{\sqrt{3}-1} &= \frac{\sqrt{2}}{\sqrt{3}-1} \cdot \frac{\sqrt{3}+1}{\sqrt{3}+1} \\ &= \frac{\sqrt{6}+\sqrt{2}}{2}\end{aligned}$$

. . . end of solution

Example. *Rationalize* $\frac{1+\sqrt{2}}{\sqrt{3}+\sqrt{2}-1}$.

Solution.

$$\begin{aligned}\frac{1+\sqrt{2}}{\sqrt{3}+\sqrt{2}-1} &= \frac{1+\sqrt{2}}{\sqrt{3}+\sqrt{2}-1} \cdot \frac{\sqrt{3}-(\sqrt{2}-1)}{\sqrt{3}-(\sqrt{2}-1)} \\ &= \frac{\sqrt{6}+\sqrt{3}-1}{2\sqrt{2}} \\ &= \frac{\sqrt{6}+\sqrt{3}-1}{2\sqrt{2}} \cdot \frac{\sqrt{2}}{\sqrt{2}} \\ &= \frac{\sqrt{6}+2\sqrt{3}-\sqrt{2}}{4}\end{aligned}$$

. . . end of solution

The square-root method

From hereon we try to investigate some critical example step by step to broaden the view on the square root method.

Example. *Solve* $x^4 = 16$.

Solution.

Observe $(x^2)^2 = x^4$. *Then*

$$\begin{aligned}(x^2)^2 &= 16\\ x^2 &= 4, -4\end{aligned}$$

Since $x^2 \geq 0$, $x^2 \neq -4$. *Thus,*

$$\begin{aligned}x^2 &= 4\\ x &= 2, -2\end{aligned}$$

. . . end of solution

Example. *Solve* $(x-3)^2 = 1$.

Solution.

$$\begin{aligned}(x-3)^2 &= 1\\ x-3 &= 1, -1\\ x &= 4, 2\end{aligned}$$

. . . end of solution

Example. *Solve* $2(x-5)^2 - 7 = 0$.

Solution.

$$\begin{aligned}2(x-5)^2 &= 7\\ (x-5)^2 &= \frac{7}{2}\\ x-5 &= \pm\frac{\sqrt{14}}{2}\\ x &= 5 \pm \frac{\sqrt{14}}{2}\end{aligned}$$

. . . end of solution

©Mok Owen 2024

General form and Completing the square method

From previous examples, we have developed a useful skill for solving simple quadratic formula. Let us consider one form of quadratic functions that usually shows up in writing equations.

Let a, b, c be constants and $a \neq 0$, the **general form of a quadratic equation** in x is written in the form of

$$ax^2 + bx + c = 0$$

The quadratic formula

4.2 Solvability of Quadratic Equations

Discriminant: the factor affecting solvability

Application of solvability

4.3 Relation between coefficients and roots

What does it mean by a root?

The Vieta formula

Forming quadratic equations with given roots

4.4 Vertex form of a quadratic function

Obtaining vertex form using Transformation

Relation between vertex form and general form

Opening direction

The axis of symmetry

The extremum of quadratic functions

4.5 Quadratic inequalities

Boolean algebra

Solving quadratic inequalities

4.6 Challenging questions

©Mok Owen 2024

5 Polynomial functions

Polynomial functions worth investigation due to its general properties for any formula, as long as it approximates arbitrary functions in an acceptable precision when the number of terms increased. In this section, we will try to understand what polynomial functions brought to us in many situations.

5.1 Polynomials

The origination of polynomial comes from the generalization of thoughts, like if we want to discuss the general result of following multiplications

$$1 \cdot 2, 2 \cdot 3, 3 \cdot 4, \dots$$

We can write $n(n+1)$ or $(n-1)n$ to conclude the pattern instead of writing all results out. Many questions about number also tackled by considering polynomial funcitons, such as the following:

Is the product of two consecutive integer always even?

To answer the above question, we can write if we choose an even number then the next number is odd, so $2n(2n+1)$ is an even number; if we choose an odd number then the next number is even, so $(2n+1)(2n+2)$ is an even number. This shows all possible cases and the result is concluded to be true.

Of course, in nowadays mathematics, many fields of research requires polynomial functions as a model for observations. We shall now dive into the knowledge.

Algebraic expressions

What algebra means is the art of symbolic representation of some astract concepts, or we may think of the abstraction itself as algebra. Given a concept A and another concept B, if we want to put a relation between A and B we say there is a relation R relating A and B such that we write ARB to abstractize the saying.

Make it simple, we do abstraction on number first. We call a number that can be control by us to be a **variable**, and for those cannot be a variable we call them **constant**. If a variable is being multiplied by a constant, to differentiate such constant from the usual-defined constant, we call such number a **coefficient**.

Example. *Given an algebraic expression* $3x + 4$*, the variable in this expression is* x*, while the constant is* 4 *and the coefficient is* 3*.*

©Mok Owen 2024

Example. *Given a number* 8*, we can say this is also an algebraic expression with no variable, coefficient being equal to* 0*, and constant is* 8*.*

Example. *Given an algebraic expression* $3x+5y$*, there are two variables in this expression, namely* x *and* y*, and to differentiate the coefficients of different variables, we call* 3 *to be the* ***coefficient of*** x *and* 5 *to be the* ***coefficient of*** y*.*

Example. *Given an algebraic expression* x^2+3x+4*, there are one variable of two forms in this expression, namely* x^2 *and* x*. Similar to previous example, we need to differentiate the coefficients of different forms of variables, so we call* 1 *to be the* ***coefficient of*** x^2 *and* 3 *to be the* ***coefficient of*** x*.*

Monomial, binomial, trinomial and multinomials

In order to clarify different situations, we have different names for different number of terms - terms mean to be the block combined by multiplication.

Example. *The expression* $3x+4$ *has two terms. The first term is* $3x$ *and the second term is* 4*.*

Example. *The expression* 8 *has one term only.*

Example. *The expression* x^2+3x+4 *has three terms.*

Example. *The expression* $x+2y+3z+4$ *has four terms.*

By different number of terms, we have different names for them. If an expression has only one term, we call it a **monomial**; for an expression having two terms, we call it a **binomial**; for an expression having three terms, we call it a **trinomial**; and for an expression having four or more terms, we call it a **multinomial**.

Example. *The following expressions are all monomials:* 1*,* x*,* $2x$*,* $3x^2$*,* $7y$*, etc.*

Example. *The following expressions are all binomials:* $x+1$*,* $x+y$*,* $2x-3x^2$*,* $3x^2+5y^2$*,* $7y^3-1$*, etc.*

Example. *The following expressions are all trinomials:* x^2+x+1*,* $x+y+z$*,* $y^2-2yx-3x^2$*,* $3x^2+5y^2+7z^2$*, etc.*

Example. *The following expressions are all multinomials:* x^3+x^2+x+1*,* $x^2+x+y^2+y+z^2+z$*,* $y^2-2y-3x-z$*,* $3x^2+5y^2+7z^2+1+2x^3+3xy$*, etc.*

©Mok Owen 2024

Polynomial and multinomial

A smaller group of multinomial plays a fundamental role in structural modelling. That category of multinomial is called **ploynomial**, which are multinomials with less than or only one variable.

Definition 5.1 (Polynomial)**.** *Let n be a non-negative integer. Suppose a_n is a non-zero real number and $a_0, a_1, a_2, \ldots, a_{n-1}$ be any real numbers. If any algebraic expression satisfy the form*

$$a_0 + a_1 x + a_2 x^2 + \cdots + a_n x^n$$

then such algeraic expression is called a ***polynomial of degree*** *n.*

The **degree** of a polynomial is determined by the leading term of the plynomial after *rearrangement of terms in decending order*, which we can say is the highest power of a polynomial.

Example. *The following expressions are all polynomials with different degrees:*

1. *1 is a polynomial of degree 0.*
2. *x and $2x + 1$ are polynomials of degree 1.*
3. *$3x^2$ and $x^2 + 2x + 1$ are polynomials of degree 2.*
4. *x^3 and $x^3 - x + 1$ are polynomials of degree 3.*

So on and so forth.

Following polynomials, we might also determine the 'degree' of a multinomial. One way to generalize the concept of a degree is to determine, similar to polynomial, the leading term of a multinomial in decending order. The problem is, what does it mean by 'larger power' when more than one variable is being discussed.

Consider the following monomial:

$$x^2 y^3 z^4$$

To determine the degree of this multinomial, we ask a question: how does it grows 'on average'? The problem becomes a statistical measurement, and what mathematicians see the solution to the problem is to identify every variables in the multinomial to be the same - in other words, focusing on the monomial we are discussing, it becomes the situation when $x = y = z$. Letting another variable t so that $x = y = z = t$ gives us the identification

$$x^2 y^3 z^4 = t^9$$

and we are now able to determine the degree of any multinomial fairly under this strategy. As exmaplified, the degree of $x^2y^3z^4$ is equal to 9. For the sake of presentation, we may write $\deg(x^2y^3z^4) = 9$ for simplicity.

Hence, we may conclude the degree of any multinomial by the following:

Definition 5.2 (Degree of multinomial)**.** *Let n be a positive integer and $x_1, x_2, \ldots, x_n$ be distinct variables. The degree of a multinomial in the form of*

$$\sum x_1^{p_1} x_2^{p_2} \cdots x_n^{p_n}$$

is equal to the maximum of the sum of the powers in one term, i.e. $\max(p_1 + p_2 + \cdots + p_n)$.

5.2 Arithmetic rules for polynomials

Polynomial plays a fundamental role in Mathematics investiagtion, and similar to numbers, it satisfies the basic rules of arithmetic.

Addition, Subtraction and Multiplication

As numbers, polynomials can be added according to the terms. Some might misunderstand the rules of addition and subtraction.

©Mok Owen 2024

Division

5.3 Divisibility of polynomials

Division algorithm

Remainder Theorem

Factor Theorem and General Vieta formula

5.4 G.C.D. and L.C.M.

Greatest Common Divisor

Least Common Multiple

5.5 Rational Function

Simplification of Rational Functions

Extended: Partial Fractions

5.6 Positional notation

Binary, Octal, Decimal and Hexadecimal

Conversion between different bases

5.7 Challenging questions

©Mok Owen 2024

6 Exponential and Logarithmic functions

The evolution of computational concepts comes from shortening the notation of recursive action. One has the '+' notation for shortening process of counting, and '×' notation for shortening the process of adding. Here we have a new notation, call it **exponentiation** for shortening the process of multiplying.

6.1 Index notation

We first revisit the concept of self-multiplication.

From multiplication to power

Recall that multiplication was invented from repeated addition, we so defined:

$$3 \times 2 = 3 + 3 = 6$$
$$4 \times 4 = 4 + 4 + 4 + 4 = 16$$

Alternatively, we can define multiplication by a recursive definition:

Definition 6.1 (Multiplication)**.** *Let m, n be positive integers. Then the multiplication among m and n can be defined as*

$$m \times n = m \times (n - 1) + m$$

*where the notation $\times$ names times and the result of $m \times n$ names the **product** of m and n.*

A good table can be used to remember the results of digital product:

×	1	2	3	4	5	6	7	8	9
1	1	2	3	4	5	6	7	8	9
2	2	4	6	8	10	12	14	16	18
3	3	6	9	12	15	18	21	24	27
4	4	8	12	16	20	24	28	32	36
5	5	10	15	20	25	30	35	40	45
6	6	12	18	24	30	36	42	48	54
7	7	14	21	28	35	42	49	56	63
8	8	16	24	32	40	48	56	64	72
9	9	18	27	36	45	54	63	72	81

©Mok Owen 2024

And the product of more digit numbers can be computed as follows:

$$11 \times 11 = 10 \times 11 + 1 \times 11 = 110 + 11 = 121$$
$$23 \times 13 = 20 \times 13 + 3 \times 13 = 260 + 39 = 299$$

The above calculation method is called separation of position.

Proceeding to the definition of power, what we could observe from the product table is the specialty of its diagonal. We saw

$$1 \times 1 = 1$$
$$2 \times 2 = 4$$
$$3 \times 3 = 9$$
$$4 \times 4 = 16$$
$$\vdots$$

which are all self-multiplication. We found this specialty occured frequently in our daily computation, so we call it **square**, following the calculation of the area of a square. However, we also have volume of a cube, so when the self multiplication counts up to 3 selves, we call it **cube**. But how about more selves are going to be multiplied?

Hence, we develop a new notation, the index notation, to shorten our writing.

Definition 6.2 (Exponentiation)**.** *Let a, n be positive integers. Then the* ***exponentiation*** *of a by n, or call it a to the power of n, can be defined as*

$$a^n := a \cdot a^{n-1}$$

Example. $1^1 = 1$.

Example. $2^3 = 8$.

Example. $10^2 = 100$.

From the definition, we can write when $n = 1$, the definition exploit that

$$a = a \cdot a^0$$

In order to survive from all possible conclusion, we need to define $a^0 = 1$ for any positive integer a. This extends our thought on exponentiation.

negative integral power

We will then try to extend the index notation to include all possible computation with multiplication. Observe the connection between multiplication and quotient: they are the inverse relation to each other! We so can define the following:

Given a positive integer r, following the definition of exponentiation, if $r \cdot r^{-1} = r^0 = 1$, we shall write $r^{-1} = \dfrac{1}{r}$ to survive the condition.

Definition 6.3 (Multiplicative inverse)**.** *Let a, n be positive integers, the negative power is defined as*

$$a^{-n} := \frac{1}{a^n}$$

Example. $1^{-1} = 1$.

Example. $2^{-3} = \frac{1}{8}$.

Example. $10^{-2} = \frac{1}{100}$.

We have to pay attention to the condition of negative powers. The meaning of negative powers is to extend the power notation to covering the law of quotient, but when 0 is involved in the calculation, one should be careful that the experssion

$$0^n$$

can only be well-defined when $n > 0$. We have no clues for computing $\dfrac{1}{0}$.

Law of integral index

So far, we invented the basic action by exponentiation from the definition of multiplication. We shall see more rules about index notations.

Recall the definition of power indices, we can observe the first law of index

$$a^{m+n} = a^m \cdot a^n$$

which can be proven by basically repeating the process of addition. We immediately have the second law of index:

$$\begin{aligned} a^{m-n} &= a^m \cdot a^{-n} \\ &= \frac{a^m}{a^n} \end{aligned}$$

The third law is a bit tricky. Recall the definition of multiplication is repeated addition, we could write

$$\begin{aligned}(a^m)^n &= a^m \cdot a^m \cdots a^m \\ &= a^{m+m+\cdot+m} \\ &= a^{mn}\end{aligned}$$

Therefore, we have for one base a, the law of multiplication and division, with the law of repeated power developed. Consider also the negative power index, we conclude 5 basic laws of exponentials.

Theorem (Monobased)**.** *Let a, m, n be positive integers. Then*

1. $a^{m+n} = a^m \cdot a^n$;
2. $a^{m-n} = \dfrac{a^m}{a^n}$;
3. $(a^m)^n = a^{mn}$;
4. $a^0 = 1$ *for* $a \neq 0$;
5. $a^{-n} = \dfrac{1}{a^n}$ *for* $a \neq 0$.

In order to manipulate with different bases, we need the law of distribution. That is, when we have a product ab as the base of exponents, we shall consider

$$\begin{aligned}(ab)^n &= ab \cdot ab \cdots ab \\ &= (aa \cdots a)(bb \cdots b) \\ &= a^n b^n\end{aligned}$$

by the commutativity of multiplication. Similarly, the version for fractions does hold:

$$\begin{aligned}(\frac{a}{b})^n &= \frac{aa \cdots a}{bb \cdots b} \\ &= \frac{a^n}{b^n}\end{aligned}$$

This allows us to discuss further for even fractional bases. Since fractions are quotient of integral divisions.

This then complete the integral index laws.

Theorem (Multibased)**.** *Let a, b be rational numbers and m, n be positive integers. Then*

1. $a^{m+n} = a^m \cdot a^n$;
2. $a^{m-n} = \dfrac{a^m}{a^n}$;
3. $(a^m)^n = a^{mn}$;
4. $a^0 = 1$ *for* $a \neq 0$;
5. $a^{-n} = \dfrac{1}{a^n}$ *for* $a \neq 0$;
6. $(ab)^n = a^n b^n$;
7. $(\dfrac{a}{b})^n = \dfrac{a^n}{b^n}$.

6.2 Rational power

The n-th root notation

To solve quadratic eqaution (equation of power 2), we used to have the sqaure-root notation $\sqrt{}$ to write down exact solutions. How about higher powers like $x^3 = a$ or $x^5 = a$?

We could introduce a small number on the top-left-corner of the surd symbol so that the corresponding eqautions are successful to be solved. It looks like

$$\begin{aligned}
x^2 = a &\implies x = \pm\sqrt{a} \\
x^3 = a &\implies x = \sqrt[3]{a} \\
x^4 = a &\implies x = \pm\sqrt[4]{a} \\
x^5 = a &\implies x = \sqrt[5]{a} \\
&\vdots \\
x^{2n} = a &\implies x = \pm\sqrt[2n]{a} \\
x^{2n+1} = a &\implies x = \sqrt[2n+1]{a} \\
&\vdots
\end{aligned}$$

To understand surds in terms of index notation, one could observe that following the index law, we have

$$(x^2)^{\frac{1}{2}} = x$$

which turns out the result of applying square-root is nearly the half power. However, we should carefully differentiate two methods.

For $\sqrt{x}$, we mean to care about the positive real solution only. That is, no matter which domain we are seeing, we need and only need to take care of positive real number solution. For instance, $x^2 = 1 \implies x = \pm 1$. The functional notation can be written $\sqrt{} : D \to \mathbb{R}_{\geq 0}$.

Definition 6.4 (Surds)**.** *The* ***n-th root of*** *a, written $\sqrt[n]{a}$, denotes the positive real solution to the equation $x^n = a$, where a is a positive real number.*

It is noticeable that the number inside the n-th root notation **must be** a positive real number: it could not be well-written if the content is negative since $x^2 = -1$ has no real solution, while $x^3 = -1$ has only negative solution, so on and so forth.

On the other hand, for $x^{\frac{1}{2}}$, we mean to care all solution to the equation $x^2 = 1$. This helps us generalize the notation to any higher dimension in mathematics, and is unrestricted to any particular domain. Hence, $1^{\frac{1}{2}} = \pm 1$ if the codomain is not specified, while $1^{\frac{1}{4}} = \{\pm 1. \pm i\}$.

Definition 6.5 (Power of reciprocal)**.** *The $\dfrac{1}{n}$* ***power of*** *a denotes the set of* ***all*** *solutions to the equation $x^n = a$ in all fields, when no other specifications.*

Remark. *While mathematicians and other scientific subjects needs the usage of fractional power, one will always abuse notations of power of reciprocals to denote the result in the simplest form. The order of simplicity is usually $\mathbb{Z} \to \mathbb{Q} \to \mathbb{R} \to \mathbb{C} \to \cdots$ following the historic investigation and positive argument first.*

Example. *Let $x^2 = 1$, then $x \in 1^{\frac{1}{2}} = \{1, -1\}$. By abusing notation, we usually denote $1^{\frac{1}{2}} = 1$ to simplify computation.*

Example. *Let $x^3 = -1$, then $x \in (-1)^{\frac{1}{3}} = \{-1, \frac{1}{2} - \frac{\sqrt{3}}{2}i, \frac{1}{2} + \frac{\sqrt{3}}{2}i\}$. By abusing notation, we usually denote $(-1)^{\frac{1}{3}} = 1$ to simplify computation.*

Example. *Let $x^2 = -1$, then $x \in (-1)^{\frac{1}{2}} = \{i, -i\}$. By abusing notation, we usually denote $(-1)^{\frac{1}{2}} = i$ to simplify computation.*

Of course, we shall discuss only the positive real solution with respect to the syllabus. We so restrict our codomain to positive real numbers only, so that $\sqrt[n]{x}$ is equivalent to $x^{\frac{1}{n}}$ without any ambiguities.

Definition 6.6. *Let n be a positive integer and x be a positive real number. Then*

$$x^{\frac{1}{n}} = \sqrt[n]{x}$$

when the codomain is restricted to be positve real numbers.

Fractional power

From the discussed definition of fractional index with surds, we can now determine the expression

$$x^{\frac{m}{n}}$$

By extending the index law to fractional power, as we did in defining $x^{\frac{1}{n}}$, we have

$$x^{\frac{m}{n}} = \sqrt[n]{x^m}$$

However, we could also define the root first such that

$$x^{\frac{m}{n}} = (\sqrt[n]{x})^m$$

The question is, is there any differences with different order of computation? Indeed, for positive number x, it doesn't; for negative x, it does. For if m and n are both even numbers, the order of computation is problematic.

1. If the exponential goes first, we have x^m being a positive number. Of course, the result $\sqrt[n]{x^m}$ must be positive, by the definition of the root notation.

2. If, however, the root goes first, we have $\sqrt[n]{x}$ possibly complex. Then $(\sqrt[n]{x})^m$ may or may not be positive.

The cure to the problem should be a simplification on the fractional index. Let us dive into it.

Proposition. *Let m, n be positive integers and x be a real number. Suppose m and n is not coprime, i.e. $HCF(m, n) = \ell$. Then*

$$x^{\frac{m}{n}} = x^{\frac{h}{k}}$$

where $m = h\ell$, $n = k\ell$ and h coprime with k, i.e. $HCF(h, k) = 1$.

Proof.
The part of simplification is trivial. We will prove the representing solution set are indeed equivalent.

Suppose $x^{\frac{1}{n}}$ defines a set of n complex values

$$\{r, rz, rz^2, \ldots, rz^{n-1}\}$$

such that all numbers in the set are solution to the equation $x^n = r^n$ with r a real constant and $z = e^{i\theta}$ being a rotation factor in the complex plane. It is easy to check no other complex numbers are in the set except those in the form of rz^d. Hence,

$$(x^{\frac{1}{n}})^m = \{r^m, r^m z^m, r^m z^{2m}, \ldots, r^m z^{m(n-1)}\}$$

Similarly, the required equivalence is the set

$$(x^{\frac{1}{k}})^h = \{s^h, s^h y^h, s^h y^{2h}, \dots, s^h y^{h(k-1)}\}$$

It suffices to show the surjection from $(x^{\frac{1}{n}})^m$ to $(x^{\frac{1}{k}})^h$. For each $s^h y^{h\kappa}$, defining $s = r^\ell$ satisfy the equality $s^h = r^m$; For $y^{h\kappa}$ we can choose $y = z^{\gamma\ell}$ such that $y^h = z^{\gamma m}$. We now have $\gamma\kappa \equiv i \mod n$. As $\kappa < k < n$, the relation $\gamma \equiv i\kappa^{-1} \mod n$ is bijective.

Alternatively, suppose it is the case of $(x^m)^{\frac{1}{n}}$, we have the solution set

$$\{r, rz, rz^2, \dots, rz^{n-1}\}$$

for the equation $x^m = r^n$. Since $m = h\ell$ and $n = k\ell$, we have the set satisfies also the equation $x^h = r^k$ where $r = s^\ell$ and $z = y^\ell$ constructs the required set in the form of sy^c. As in previous case, the set are bijective.

Therefore, reduction formula $x^{\frac{m}{n}} = x^{\frac{h}{k}}$ holds in any case. □

As the equality is bi-directional, we can perform fractional operation for index, i.e.

$$x^{\frac{m}{n}+\frac{p}{q}} = x^{\frac{mq+pn}{nq}}$$

Corollary. *Let m, n be positive integers such that $HCF(m, n) = \ell$ and x be a real number. Then*

$$x^{\frac{m}{n}} = \sqrt[k]{x^h} = (\sqrt[k]{x})^h$$

where $m = h\ell$, $n = k\ell$ and h coprime with k, i.e. $HCF(h, k) = 1$.

Law of fractional index

Following the proved result in previous part, we have the following laws of fractional index. They are similar to that for integral powers, but their condition are much stronger than that of integer case.

Theorem (Fractional index law)**.** *Let a, b, p, q be non-zero rational numbers. Then*

1. $a^{p+q} = a^p \cdot a^q$;
2. $a^{p-q} = \dfrac{a^p}{a^q}$;
3. $(a^p)^q = a^{pq}$;
4. $a^0 = 1$ *for* $a \neq 0$;

5. $a^{-p} = \frac{1}{a^p}$ *for* $a \neq 0$;

6. $(ab)^p = a^p b^p$;

7. $(\frac{a}{b})^p = \frac{a^p}{b^p}$.

The major evolution of the theorem is where we successfully observe the same properties from integral case to fractional case. One may notice that rational number covers almost real numbers, so the next question is, could we advance the result to that level.

6.3 Exponential functions

From here we will investigate a cool result for power index. That is, we can always write real functions using exponents in the form

$$f(x) = a^x$$

where a is a positive real constant and x a real variable.

Extending to real powers

It is not hard to extend the base to irrational number. The main reason is even the base is irrational number, the process of self-nultiplication and the concept of solving equation still holds. That means, we have no reason to reject the expression

$$a^p$$

with real number a and rational number p.

However, the concept becomes unclear when the index becomes irrational. We need to show that some values could be uniquely determined for irrational power. In fact, if we define irrational power only on codomain of real numbers, by the completion from rational number to real number, we can always find for any irrational number s, there exist two rational numbers $p < q$ that are close enough to s, say $q - s < \frac{1}{n}$ and $s - p < \frac{1}{n}$ so that

$$s - \frac{1}{n} < p < s < q < s + \frac{1}{n}$$

Raising such numbers to the power index of a positive real base a, we have the inequality follows

$$a^{s-\frac{1}{n}} < a^p < a^s < a^q < a^{s+\frac{1}{n}}$$

As n approaches infinity, the approximated approaching value becomes unique, which we see the error of approximation is small enough in the presentation. This shows the concept of irrational number is solid under restriction.

©Mok Owen 2024

Definition 6.7 (Exponential function of real numbers)**.** *Define a function $f : \mathbb{R} \to \mathbb{R}_{>0}$ by $x \mapsto a^x$ with $a > 0$ a real number. The function f is called* ***exponential function of base*** *a.*

Example. *The function $f(x) = \pi^x$ is an exponential function of base π. $f(0) = \pi^0 = 1$.*

Example. *The function $f(x) = 2^x$ is an exponential function of base 2. $f(\pi) = 2^\pi \approx 8.825$.*

Example. *The function $f(x) = e^x$ is an exponential function of base e. $f(-\pi) = e^{-\pi} \approx 0.0432$.*

Graph of exponential functions

Law of Exponential algebra

Similar to that of fractional index, when we restrict the codomain to real numbers only, the laws are well-defined.

Theorem (Law of exponential algebra)**.** *Let a, b be positive real numbers and x, y be non-zero real numbers. Then*

1. $a^{x+y} = a^x \cdot a^y$;
2. $a^{x-y} = \dfrac{a^x}{a^y}$;
3. $(a^x)^y = a^{xy}$;
4. $a^0 = 1$ *for* $a \neq 0$;
5. $a^{-x} = \dfrac{1}{a^x}$ *for* $a \neq 0$;
6. $(ab)^x = a^x b^x$;
7. $(\dfrac{a}{b})^x = \dfrac{a^x}{b^x}$.

Solving equations using exponetiation method

6.4 Logarithmic functions

What is Logarithm?

Law of Logarithmic algebra

Solving equations using Logarithmic method

6.5 Challenging questions

7 Sequence as a function of natural numbers

7.1 What is a sequence?

7.2 Arithmetic sequence

General form

Summation of arithmetic sequence

7.3 Geometric sequence

General form

Summation on Geometric sequence

Infinite sum of Geometric sequence

7.4 Additional content: Arithmetic-Geometric sequence

7.5 Challenging questions

©Mok Owen 2024

8 Probability functions

8.1 Counting Principle

Let's introduce one notation for sum and one notation for product. We will see their convenience during conceptualisation.

Definition 8.1 (Summation notation)**.** *We use the symbol sigma Σ to emphasis the process of addition. Usually we define the addition process using indexing format. The notation*

$$\sum_{i=1}^{n} x_i$$

means adding up from x_1 to x_n. In other words, it is saying

$$\sum_{i=1}^{n} x_i = x_1 + x_2 + \cdots + x_n$$

Example. $\sum_{i=1}^{3} 1 = 1 + 1 + 1 = 3.$

Example. $\sum_{i=1}^{4} i = 1 + 2 + 3 + 4 = 10.$

Example. $\sum_{i=2}^{5} \frac{i^2}{2} = 27.$

Example. $\sum_{i=2}^{5} \frac{i^2}{2} + 2 = 29.$

Example. $\sum_{i=2}^{5} (\frac{i^2}{2} + 2) = 35.$

Remark. *The summation notation can usually be a misconception to many students, but bear in mind that the sum is computed by adding up terms.*

Theorem (Properties of summation notation)**.** *The following properties holds for summation notation:*

1. $\sum_{i=1}^{n} a_i + \sum_{i=1}^{n} b_i = \sum_{i=1}^{n} (a_i + b_i)$;
2. $c \sum_{i=1}^{n} a_i = \sum_{i=1}^{n} c a_i$;

3. $(\sum_{i=1}^{n} a_i)(\sum_{j=1}^{m} b_j) = \sum_{i=1}^{n}\sum_{j=1}^{m} a_i b_j$;

4. $\sum_{i=1}^{n}\sum_{j=1}^{m} a_i b_j = \sum_{j=1}^{m}\sum_{i=1}^{n} a_i b_j$.

Definition 8.2 (Product notation)**.** *We use the symbol pi* Π *to emphasis the process of multiplication. Usually we define the multiplication process using indexing format. The notation*

$$\prod_{i=1}^{n} x_i$$

means multiplying up from x_1 *to* x_n*. In other words, it is saying*

$$\prod_{i=1}^{n} x_i = x_1 x_2 \cdots x_n$$

Example. $\prod_{i=1}^{3} 1 = 1 \cdot 1 \cdot 1 = 1.$

Example. $\prod_{i=1}^{4} 2 = 2 \cdot 2 \cdot 2 \cdot 2 = 16.$

Example. $\prod_{i=1}^{4} i = 1 \cdot 2 \cdot 3 \cdot 4 = 24.$

Addition rule

The addition rule starts by finding the total number of things. We count by doing adding values.

Definition 8.3 (Addition rule)**.** *Let* A *and* B *be two collections with* m *and* n *objects respectively. Then the total number of objects in* A *and in* B *is* $m + n$*.*

Let us introduce the notation $|\cdot|$ to present the menaing of counts: If A represents a collection of n objects, then $|A| = n$.

Definition 8.4 (Rewriting Addition rule)**.** *Let* A *and* B *be two distinct collections. Then the total number of objects among them is* $|A| + |B|$*.*

Example. *Let Alice has 3 apples and Bob has 4 apples. Then they have a total of 7 apples.*

Definition 8.5 (General version of addition rule)**.** *Let $A_1, A_2, \ldots, A_n$ be n collections. Then the total number of objects among all these collections is* $\sum_{i=1}^{n} |A_i|$.

Example. *Let Alice has 2 apples, Bob has 5 apples and Chris has 7 apples. Then they have a total of 14 apples.*

Multiplication rule

The multiplication rule comes from finding the total number of combinations. We count by spanning a table.

Definition 8.6 (Multiplication rule)**.** *Let A and B be two collections. Then the total number of combining objects in A and in B is* $|A| \cdot |B|$.

	1	2	$\cdots$	m
1	(1,1)	(2,1)	$\cdots$	(m,1)
2	(1,2)	(2,2)	$\cdots$	(m,2)
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
n	(1,n)	(2,n)	$\cdots$	(m,n)

Example. *Let Alice has 3 different T-shirts and Bob has 4 different pants. Then the number of costume combination is* 12.

Example. *Suppose there are 15 boys and 18 girls in a class. Then the number of possible couples in this class is 270.*

Definition 8.7 (General version of multiplication rule)**.** *Let $A_1, A_2, \ldots, A_n$ be n collections with $x_1, x_2, \ldots, x_n$ objects respectively. Then the total number of n-tuples among all these collections is* $\prod_{i=1}^{n} x_i$.

Example. *Let class A has 2 class members, class A has 5 class members and class C has 7 class members. To form a committee by selecting one students from each class, 70 different committee can be formed.*

8.2 Pigeonhole Principle

One should identify the constraints to manage how distributions are formed, and so does probability. We will see one question to facilitate the mind on distribution.

Let us see we have for instance 4 pigeonholes. If we have 1 pigeon to choose where to live in, it will fill one of the 4 holes; if we have 2 pigeons, 2 out of 4 will be filled; if 3, then 3 out of 4 will be filled; and 4 out of 4 for the case of 4. However, what happened when 5 pigeons are there?

Assume they do not fight to each other so that no dead bodies are there, and they agreed to live together in some pigeonhole. Alright, no extra space for building new holes either. Then we could see when 4 holes are all filled with 1 pigeon, 1 is left behind to choose which pigeonhole to live in. In this stage, we observe no matter which hole it chose, **at least one of the 4 holes have to be with 2 or more pigeons**. We then extend this concept to a general case of $n+1$ pigeons with n pigeonholes. This is called the **Pigeonhole principle**.

Fundamental version

Theorem (Baby pigeon)**.** *If there are n containers with $n+1$ objects, then at least one of the containers must have 2 or more objects in it.*

Example. *Consider 31 books are going to be distributed to 30 students, then at least one of the students has 2 or more books.*

Essential Practice 8.2.1. *To ensure 10 children are all having at least 1 slice of pizza, at least how many slices should the pizza be sliced into?*

General version

A general version of pigeonhole principle can be introduced from the perspective of division.

Theorem (Adult pigeon)**.** *If there are n containers with $mn+1$ objects, then at least one of the containers must have $m+1$ or more objects in it.*

Proof.
We may consider $mn+1$ as division algorithm and decide the worst situation. Indeed, it is the worst when every container is containing the least amount of objects in it. This is equivalent to saying the distribution is uniform, i.e. averagely distributed. We then conclude after the worst case, every container is containing m objects with 1 left behind. The principle is deduced. □

Corollary (Pigeonhole principle)**.** *Suppose t objects are going to be distributed to n containers, where $mn+1 \leq t \leq (m+1)n$. Then at least one of the containers must be having $m+1$ objects.*

Example. *Suppose 99 balls are stored in the storage room. If there are 5 basket ball players, then at least one of the players could have 20 balls.*

This is why we usually use division to measure the worst case of distribution.

8.3 Counting functions

Based on the rules of counting, we establish the following method of counting measure. Our perspective of counting starts from permutation.

Factorial

Suppose we are arranging some people in a queue. Let say there are 5 people to be arranged. Then how many different queues could be formed?

Assume 5 placement should be filled by 5 people, then we have no duplicated choice for 2 different placement. We call the selection be without replacement.

For the first placement, we have 5 choice; for the second placement, as we have already select 1 person out of 5, we can only choose from the remaining 4 person, so the secnod placement has 4 choices; for the third placement, we have a remaining of 3 person, so 3 choices for the third placement; so on and so forth, the fourth and the fifth placement has 2 and 1 choice respectively. As long as a queue can be seen as an ordered combination, we shall apply multiplication rule to compute the number of permutation:

$$5 \times 4 \times 3 \times 2 \times 1 = 120$$

To shorten the multiplication (Yes, we are lazy), we define the following notation:

Definition 8.8 (Factorial)**.** *Let n be a non-negative integer, then the **factorial** notation is defined as recursive definition:*

$$n! = n(n-1)!,\ 0! = 1$$

In some sense, if we see the factorial notation as some function, we have

$$n! := \prod_{i=1}^{n} i$$

However, how was 0! be defined?

Let's consider the reverse view of factorial. We note that factorial is built from multiplication, and so can be extended using division. Consider the definition

$$n! = n(n-1)! \implies (n-1)! = \frac{n!}{n}$$

which means 0! can be computed as

$$0! = \frac{1!}{1} = 1$$

Example. *For a queue of 7 people, we have $7! = 5040$ different permutations.*

Example. *For a queue of 0 people, we have* $0! = 1$ *permutation only. That is, arranging nothing. It is so defined because this is also a valid situation to be counted!*

Essential Practice 8.3.1. *Compute the number of distinct queues can be formed if there are (a)6 people; (b)9 people; (c)12 people.*

Permutation

Somehow, we have another problem of arranging not all people but part of them. We shall now consider if **not all people are selected in the arrangement**.

Suppose now we have 6 people in the selection zone, and we need to arrange any 3 of them in order. It would be easier to think of counting championships, but never mind. Similar to the part of factorial, we just need to reduce the outstanding arrangement. Therefore, we have in total

$$6 \times 5 \times 4 = 120$$

numbers of arrangement for 3 out of 6 people. It is interesting to do an algebra trick: we complete the factorials by filling up the gaps so that

$$6 \times 5 \times 4 = \frac{6 \times 5 \times 4 \times 3 \times 2 \times 1}{3 \times 2 \times 1} = \frac{6!}{3!} = \frac{6!}{(6-3)!}$$

The reason to write the last equality is we shall observe and fulfill the generality of the writings. Consider we need to arrange 4 out of 6, the computation becomes

$$6 \times 5 \times 4 \times 3 = \frac{6 \times 5 \times 4 \times 3 \times 2 \times 1}{2 \times 1} = \frac{6!}{2!} = \frac{6!}{(6-4)!}$$

Therefore, in general, we find the permutation of r out of n objects is equal to $\dfrac{n!}{(n-r)!}$, and one denote the concept by taking P as the notation standing for permutation. The formal definition says:

Definition 8.9 (Permutation)**.** *The amount of permutation for arranging* r *objects out of* n *objects is defined as*

$$P_r^n := \frac{n!}{(n-r)!}$$

Example. *There are* $9 \times P_4^9$ *5-digit numbers without repeated digits.*

Essential Practice 8.3.2. *How many strings (alphabet permutations) of length 6 can be formed if no alphabet is used repeatedly?*

©Mok Owen 2024

Combination

We may now consider another case of counting. For who focus on only the category of selection but not the difference in order, we name the method of finding combinations.

To find the number of combinations, we have to look at permutation first. Should we compute the number of combinations by dividing the number of global permutations by local permutation. We shall identify what should be considered as global permutations and local permutations.

Let there be 5 objects, name them A, B, C, D, E, and we are going to permute them in different orders. The whole set of possible arrangement of any 3 objects can be concluded by row and column format:

ABC	ABD	ABE	ACD	ACE	ADE	BCD	BCE	BDE	CDE
ACB	ADB	AEB	ADC	AEC	AED	BDC	BEC	BED	CED
BAC	BAD	BAE	CAD	CAE	DAE	CBD	CBE	DBE	DCE
BCA	BDA	BEA	CDA	CEA	DEA	CDB	CEB	DEB	DEC
CAB	DAB	EAB	DAC	EAC	EAD	DBC	EBC	EBD	ECD
CBA	DBA	EBA	DCA	ECA	EDA	DCB	ECB	EDB	EDC

If we consider the table column-by-column, we found that each column are of the same set of objects, which are what we called the same combination for one column. To express them explicitly, we can write them into

$$\begin{array}{ll} \{A,B,C\} & \{A,B,D\} \\ \{A,B,E\} & \{A,C,D\} \\ \{A,C,E\} & \{A,D,E\} \\ \{B,C,D\} & \{B,C,E\} \\ \{B,D,E\} & \{C,D,E\} \end{array}$$

That means, there are in total 10 combinations for if we choose 3 objects out of 5 objects. By observation, we found the total number of permutations involves all objects in the set we considered. We call it the **global permutation**. Meanwhile, for each column, the permutations are involving the same objects chosen from the set. We shall call it **local permutation**. Therefore, the mechanism can be expressed by the name defined: Dividing the number of global permutation by the number of local permutation, we will get the number of combination.

So the final judgement to the number of combination is the formula

$$\frac{P_3^5}{3!} = 10$$

©Mok Owen 2024

and the general formula shall be

$$\frac{P_r^n}{r!} = \frac{n!}{(n-r)!r!}$$

We will denote the combination symbol using the first alphabet C. Definition as follows:

Definition 8.10 (Combination)**.** *The amount of combination for collecting r objects out of n objects is defined as*

$$C_r^n := \frac{n!}{(n-r)!r!}$$

Example. *To form a group of 5 people from a class of 31 person, C_5^{31} different groups can be formed.*

Essential Practice 8.3.3. *How many ways are there if we need to form a group of 4 people from 100 people?*

Example. *To form 5 groups of equal size among 30 people, the first group has C_5^{30} different formation; the second group has C_5^{25} different formation; the third, the fourth, the fifth and the sixth has $C_5^{20}, C_5^{15}, C_5^{10}$ and C_5^5 respectively. But the order of formation should be ignored, so the total number of formation should be*

$$\frac{C_5^{30} \cdot C_5^{25} \cdot C_5^{20} \cdot C_5^{15} \cdot C_5^{10} \cdot C_5^5}{6!}$$

Essential Practice 8.3.4. *How many different formations are there if we need to form 7 groups of equal size among 49 people?*

Essential Practice 8.3.5 (Hard)**.** *How many different formations are there if we need to form 7 groups of at least 2 people among 15 people?*

8.4 Fundamental Probability

When we talked about probability, the first come to our mind should be the word 'possibility', whether it could be or could not be. But there's a misconception between possibility and probability. The word possibility means if there exist such event to be happened, while the word probability is a measure of possibility: how often and how large the chance it tends to be happening.

Fair dice intuition

Let us consider a fair dice of 6 faces. We can easily say there are six faces, so six events can be happening: the 1-point face, the 2-point face,..., the 6-point face. There are six events in total. And if we consider it is a fair dice, which means all six faces have equal chance to happen, then each face have the probability of $\frac{1}{6}$ to happen. From this, we invest the thought of probability measurement:

Definition 8.11 (Fair Dice Probability)**.** *Given a fair die of six faces, the probability on each face is defined as* $\frac{1}{6}$.

From this intuition, if we could observe that

Example. *The probability of getting the face of 1-point is* $\frac{1}{6}$.

Example. *The probability of getting the faces of (less than 3)-point is* $2 \times \frac{1}{6} = \frac{1}{3}$.

Example. *The probability of getting the faces of (even)-point is* $3 \times \frac{1}{6} = \frac{1}{2}$.

Of course, we also see that

Example. *The probability of getting any faces is* $6 \times \frac{1}{6} = 1$.

Example. *The probability of getting the faces of (less than 1)-point or (more than 6)-point is* $0 \times \frac{1}{6} = 0$.

So we understand probability as follows:

Theorem. *Let* P *be the function of probability, and let* X *to be the type of face we need in a die.*

1. *If* X *declare all types of faces that can be found in a die, then* $P(X) = 1$.
2. *If* X *cannot be found in a die, then* $P(X) = 0$.
3. *Since all or nothing defined the bounds of possibility, for any* X, $0 \leq P(X) \leq 1$.

Since probability follows the counting of faces on a die, all computation should be following the computation in counting functions.

Events and probability

Seeing the type of faces on a die as an event, and generalize the concept of a face on a die as an event in a situation, we may now write down the definition of general probability.

Definition 8.12 (Events)**.** *An* ***event*** *is a collection of objects that satisfying some conditions.*

Example. *Let say we have a fair die. Then* ***getting even number of dots after rolling the die*** *is an event, which objects to be collected are faces with dots on the die, and condition for the collection is when the number of dots on that face is an even number.*

Definition 8.13 (Probability)**.** *Suppose U be the universal set of events and X be the targetted set of events within U. Let P be the probability function on U so that $P : U \to [0, 1]$. Then we shall write*

$$P(X|U) := \frac{|X|}{|U|}$$

If U is explicitly defined and well-understood for all situations throughout the discussion, then we can write $P(X)$ instead of $P(X|U)$.

Example. *If it is understood the discussion is based on a fair die, we claim that $P(getting\,even - point\,face) = \frac{1}{2}$.*

Example. *If it is understood the discussion is not based on a fair die, and each face occurs the number of times equal to their point over 21 rolls, we claim that $P(getting\,even - point\,face) = \frac{12}{21} = \frac{4}{7}$.*

To distinguish the above examples, we shall write $P(getting\,even - point\,face|fair\,die) = \frac{1}{2}$ and $P(getting\,even - point\,face|times\,equal\,to\,point) = \frac{4}{7}$ so that no ambiguities should be done.

8.5 Probability of independent events

Independent events

When the discussion becomes complex and uniting different events in one situation, we have to perform distinction of events: some events are not affecting one another. For example, the probability of giving shits to this notes in the morning does not affect the probability of taking a shower in the evening, or the probability of buying a Mathematics book does not affect the probability in winning a Jackpot. We then say if events A and B does not affect each other, A is independent from B and B is independent from A, we call A and B are independent from each other and they are **independent events** in the situation.

Definition 8.14 (Independent Events)**.** *Let A and B be two events in the universal set U. If $P(A|B) = P(A)$ and $P(B|A) = P(B)$, then A and B are called independent to each other.*

We will derive suitable formula for independent events.

Consider the 'must go on' events, that is, no matter what is happening, the probability of such event being happened will never be changed due to any situation, and we will call that event stable. One stable event commits to any situation, its probability of happening will not be changed,

meaning a stable event must be independent from everything. Thus we will have the following relation:

Given a stable event A and a stable event B, When A happens, B's probability of happening is still $P(B)$. By the rules of counting, we have the probability for both A and B happening at the same time is $P(A)P(B)$.

Theorem. *Given A and B are independent from each other, the following holds:*

$$P(A \cap B) = P(A)P(B)$$

Complementary event

A **complementary event** means it covers all the outstanding events of a particular event. We will define a new symbol for this.

Definition 8.15 (Complementary event)**.** *Given a set of events A under a universal set of events U, we define the set A^c or A' to be the **complementary set** of A, which contains all the events in U but not in A. In symbolic contexts, it is*

$$A^c = U \setminus A$$

It is not difficult to see that A and A^c has no intersectional events. That is, if an event x is in set A then it is not in set A^c, and vice versa. Therefore, we can easily deduce that

Theorem. *If A is a set of events and A^c is the complementary set of events, their probability satisfies the equations*

$$P(A) + P(A^c) = 1,\ P(A \cap A^c) = 0$$

Rules of coherence

Let A and B be two independent events. We denote $A \cap B$, the intersection of the two independent events, as the event of having A and B simultaneously; $A \cup B$, the union of the two independent events, as the event of having either A or B at once. The probability of $A \cap B$, denoted by $P(A \cap B)$, and the probability of $A \cup B$, denoted by $P(A \cup B)$, has the following relationship

Theorem. *One of the rules of coherence in probability is stated as follows*

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

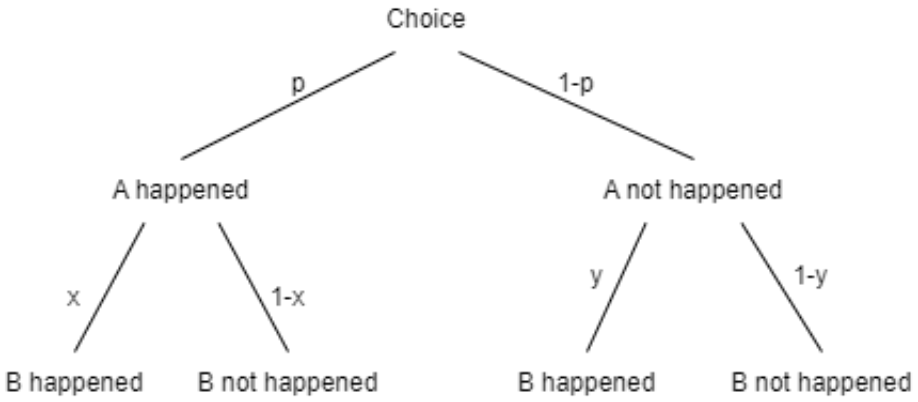
8.6 Probability of dependent events

The case of independent events has been discussed, and we will move to the case of dependent events. What dependent events brought us is the simplification of complex cases, so that we can always extract useful information for complex probability.

Conditional proability

We should acknowledge that we are unable to handle dependent events directly, say if the probability of A being happened depends on B, we should see the difference between when and when not B is happened.

Let say we have a complex situation inheriting the existence of event A and event B. Given the probability of A being happened equals p. If A is happened, the probability of B being happened is x, else the probability of B being happened is y. The graph describing the above situation can be illustrated as follows:



We define the paths of the graph as the independence decomposition of a complex events, providing a focus on particular situations. The probability value x and y are called conditional proability, which is a particular probaility of a given condition. We denote $P(B|A) = x$ to be the probability of B when A is given happening. Similarly, $P(B'|A) = 1 - x$ to be the probability of B does not happen when A is given happening. $P(B|A') = y$ and $P(B'|A') = 1 - y$ for similar reasoning.

In particular, we may observe that $px = P(A \cap B)$, and that becomes one of the formula defining conditional proability:

Theorem. *Given A and B be two events, the conditional probability of B given A is happened is*

$$P(B|A) = \frac{P(A \cap B)}{P(A)}$$

The formula provides an interesting connection for the bi-directional conditional probability formula:

Corollary. *Given A and B be two events, the following holds:*

$$P(A)P(B|A) = P(B)P(A|B)$$

To resolve the proability of B, we shall see the collection of all event B is from both A happens or not, hence it is natural to write $P(B) = px+(1-p)y$, which is in other words the total probability formula.

Theorem. *The total probability of event B is given by, when A and B are dependent on each other,*

$$P(B) = \sum P(A_n \cap B) = \sum P(A_n)P(B|A_n)$$

where $A = \bigcup A_n$.

8.7 Discrete random variables

Finite domain and infinite domain

Mean and expected value

Standard deviation and variance

8.8 Distributions

Binomial distribution

Geometric distribution

Poisson distribution

Normal distribution and Standard distribution

8.9 Sampling

Point sampling

Central limit theorem

Confidence interval

8.10 Challenging questions